Influence in Completely Bounded Block-Multilinear Forms and Classical Simulation of Quantum Algorithms

Nikhil Bansal
University of Michigan, Ann Arbor, MI, USA

Makrand Sinha
Simons Institute, Berkeley, CA, USA
University of California Berkeley, CA, USA

Ronald de Wolf
QuSoft, CWI, Amsterdam, The Netherlands
University of Amsterdam, The Netherlands

Abstract

The Aaronson-Ambainis conjecture (Theory of Computing '14) says that every low-degree bounded polynomial on the Boolean hypercube has an influential variable. This conjecture, if true, would imply that the acceptance probability of every d-query quantum algorithm can be well-approximated almost everywhere (i.e., on almost all inputs) by a poly(d)-query classical algorithm. We prove a special case of the conjecture: in every completely bounded degree-d block-multilinear form with constant variance, there always exists a variable with influence at least $1/\text{poly}(d)$. In a certain sense, such polynomials characterize the acceptance probability of quantum query algorithms, as shown by Arunachalam, Briët and Palazuelos (SICOMP '19). As a corollary we obtain efficient classical almost-everywhere simulation for a particular class of quantum algorithms that includes for instance k-fold Forrelation. Our main technical result relies on connections to free probability theory.

2012 ACM Subject Classification Theory of computation → Quantum query complexity; Theory of computation → Oracles and decision trees

Keywords and phrases Aaronson-Ambainis conjecture, Quantum query complexity, Classical query complexity, Free probability, Completely bounded norm, Analysis of Boolean functions, Influence

Digital Object Identifier 10.4230/LIPIcs.CCC.2022.28

Related Version *arXiv Version*: `https://arxiv.org/abs/2203.00212`

Funding *Nikhil Bansal*: Supported in part by the NWO VICI grant 639.023.812.
Makrand Sinha: Supported by a Simons-Berkeley postdoctoral fellowship.
Ronald de Wolf: Partially supported by the Dutch Research Council (NWO/OCW), as part of the Quantum Software Consortium programme (project number 024.003.037), and through QuantERA ERA-NET Cofund project QuantAlgo (680-91-034).

Acknowledgements We thank Scott Aaronson, Srinivasan Arunachalam, Jop Briët, Shachar Lovett and Ryan O'Donnell for helpful comments and pointers to the literature.

1 Introduction

This paper is motivated by quantum query complexity and its relation to classical query complexity. Query complexity has been the context in which many of the main quantum algorithms have been developed, including Shor's [32] (building on [33]) and Grover's [21]. It has the added advantage that we actually know how to prove good lower bounds on query complexity, in contrast to a setting like circuit complexity.

Quantum query complexity is closely connected to the study of bounded polynomials (or forms) on the Boolean hypercube. The key to this connection is that the amplitudes of the final state of a d-query quantum algorithm are polynomials of degree at most d in the bits of the input x, and therefore its acceptance probability $p(x)$ is a polynomial of degree at most $2d$. This observation was made by Beals, Buhrman, Cleve, Mosca and de Wolf [12], who used it to show that the bounded-error quantum query complexity and classical query complexity are polynomially related for any *total* Boolean function. Since then a long line of research [7, 3, 13, 8, 34, 4, 10, 31] has tried to pinpoint the exact polynomial dependence as well as studied the relationship with other measures of complexity of a Boolean function (e.g., sensitivity, certificate complexity, and others [27, 15, 11]).

On the other hand, quantum algorithms can offer a huge (superexponential) advantage for partial functions, which are only defined on a subset of the Boolean hypercube; there are many known examples of partial functions whose classical query complexity is much larger than their quantum query complexity, for instance k-fold Forrelation and its variants [2, 34, 10, 31]. This means that the acceptance probability $p(x)$ of a quantum algorithm cannot always be efficiently approximated by a classical algorithm, since otherwise quantum algorithms could offer only a polynomial speedup for any function, be it total or partial.

However, we can set our sights lower, and ask whether it is possible to classically efficiently approximate $p(x)$ on *almost all* inputs. The following conjecture, which first appeared in [1] and is attributed there to folklore, says that we can.

▶ **Conjecture 1** (Folklore). *The acceptance probability of any d-query quantum algorithm on n-bit inputs can be estimated up to additive error ϵ on a $1-\delta$ fraction of the inputs by a classical query algorithm making* $\mathrm{poly}(d, 1/\epsilon, 1/\delta)$ *queries.*

This conjecture is one expression of the general idea that quantum computers can only give significant speedup (in terms of queries, circuit complexity, or other things) on very structured problems, i.e., when the input to the problem has a particular structure, for instance some periodicity or specific correlations between different parts of the input. For generic unstructured inputs, the conjecture says that only a limited quantum speedup can be expected. This conjecture motivates and is implied by the following conjecture due to Aaronson and Ambainis [1]:

▶ **Conjecture 2** (Aaronson-Ambainis conjecture). *Let $f : \{\pm 1\}^n \to [0,1]$ be a degree-d multilinear polynomial. Then, the maximum influence among all variables in f is at least* $\mathrm{poly}(\mathrm{Var}[f], 1/d)$.

The above conjecture poses a fundamental structural question about bounded polynomials on the hypercube and is a notable open problem in the analysis of Boolean functions. Conjecture 2 is known to hold if the function is Boolean-valued (this follows from [24, 29]). For bounded polynomials, [1] observed that the results of Dinur, Friedgut, Kindler and O'Donnell [19] imply that the conjecture holds with at least an exponential dependence in d. Montanaro [25] proved a special case of the conjecture for *block-multilinear forms* where all coefficients have the same magnitude[1]. Defant, Mastyło and Perez [18] generalized this to bounded polynomials where all Fourier coefficients have the same magnitude and showed that the conjecture holds with an $\exp(\sqrt{d \log d})$ dependence. O'Donnell and Zhao [30] showed that it is sufficient to prove the conjecture for so-called *one-block decoupled* polynomials.

[1] This argument can be generalized to the case when $n^{\Omega(d)}$ coefficients have the same magnitude and the rest are zero, as noted in [25] where the observation is attributed to Ambainis.

In this work, our motivation is to study Conjecture 2 for polynomials that represent the acceptance probability of quantum algorithms. Such polynomials have a lot more structure – as shown by Arunachalam, Briët and Palazuelos [9], they can be represented in terms of *completely bounded block-multilinear forms* (as described in the next section) and conversely, such forms even characterize quantum algorithms in a certain sense (see Section 4). As such here we focus on understanding influences in such polynomials.

1.1 Our results

A degree-d block-multilinear form $f(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_d)$ mapping $\{\pm 1\}^{n\times d}$ to $\mathbb{R}$ is a polynomial where the variables are partitioned into d blocks of n variables each, and each monomial contains at most one variable from each block. Formally, $\boldsymbol{x}_b = (x_b(1),\ldots,x_b(n)) \in \{\pm 1\}^n$ constitutes the b^{th} block of variables and

$$f(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_d) = \mathbb{E}f + \sum_{m=1}^{d} \sum_{\substack{|\boldsymbol{b}|=m \\ |\boldsymbol{i}|=m}} \widehat{f}_{\boldsymbol{b},\boldsymbol{i}} \cdot\ x_{b_1}(i_1)x_{b_2}(i_2)\cdots x_{b_m}(i_m), \tag{1}$$

where the tuple $\boldsymbol{b} = (b_1,\ldots,b_m)$ satisfies $1 \le b_1 < \ldots < b_m \le d$ and $\boldsymbol{i} \in [n]^m$ is an m-tuple. Note that m is determined from the size of the tuple $\boldsymbol{b}$, so we just write $\widehat{f}_{\boldsymbol{b},\boldsymbol{i}}$ above.

Since each non-constant monomial contains at most one variable from each block and the ordering of the blocks is fixed, a degree-d block-multilinear form $f : \{\pm 1\}^{n\times d} \to \mathbb{R}$ can be naturally viewed as a non-commutative polynomial in matrix variables with the constant term replaced with $\mathbb{E}f$ times the identity. Denoting the non-commutative polynomial as $f(\boldsymbol{U}_1,\ldots,\boldsymbol{U}_d)$ where each $\boldsymbol{U}_b = (U_b(1),\ldots,U_b(n))$ is a block of non-commutative variables, the completely bounded norm[2] $\|f\|_{\text{cb}}$ of the form f is defined as

$$\|f\|_{\text{cb}} = \sup\left\{ \|f(\boldsymbol{U}_1,\ldots,\boldsymbol{U}_d)\|_{\text{op}} \;\middle|\; N \in \mathbb{N}, U_b(i) \in \mathbb{C}^{N\times N}, \|U_b(i)\|_{\text{op}} \le 1, b \in [d], i \in [n] \right\}.$$

The supremum above is always attained and can be computed by solving a semidefinite program as shown by Gribling and Laurent [20]. One can also equivalently restrict the supremum in the definition above to unitary matrices since the convex hull of the set of unitary matrices is the unit operator-norm ball. Moreover, $\|T\|_\infty \le \|T\|_{\text{cb}}$ where $\|T\|_\infty = \max_{x\in\{\pm 1\}^{n\times d}} |T(x)|$, so forms that are completely bounded are also bounded on the hypercube.

Our main result is a proof of the Aaronson-Ambainis conjecture for block-multilinear forms that are completely bounded. To state our result, we recall that the influence of a variable $x_b(i)$ on f is

$$\text{Inf}_{b,i}(f) = \mathbb{E}\left|\partial_{b,i} f(X)\right|^2,$$

where X is uniform in $\{\pm 1\}^{n\times d}$ and $\partial_{b,i} f(x)$ is the discrete derivative (see Section 2). Denoting by $\text{MaxInf}(f) = \max_{b\in[d],i\in[n]} \text{Inf}_{b,i}(f)$ the maximum influence of any variable in f and by $\text{Var}[f]$ the variance of f on the hypercube, we show:

[2] The completely bounded norm originates in the theory of operator algebras. In the literature, this norm is sometimes defined for homogeneous block-multilinear forms only, but here we extend the definition to non-homogeneous block-multilinear forms.

▶ **Theorem 3.** *Let f be a degree-d block-multilinear form with $\|f\|_{\mathrm{cb}} \leq 1$. Then, we have*

$$\mathrm{MaxInf}(f) \geq \frac{(\mathrm{Var}[f])^2}{e(d+1)^4}.$$

The main technical ingredient in the proof of Theorem 3 is a new influence inequality for *homogeneous* block-multilinear forms that relates the completely bounded norm to the influences.

▶ **Theorem 4.** (Non-commutative root-influence inequality). *Let f be a homogeneous degree-d block-multilinear form. Then, for blocks $b \in \{1, d\}$,*

$$\|f\|_{\mathrm{cb}} \geq \frac{1}{\sqrt{e(d+1)}} \sum_{i=1}^{n} \sqrt{\mathrm{Inf}_{b,i}(f)}.$$

In general, the completely bounded norm can change if we permute the blocks, and the theorem above only gives a bound in terms of the influences of the variables in the leftmost and rightmost blocks.

The inequality also easily implies the special case of Theorem 3 for homogeneous forms, with a better dependence on d, as follows:

$$\begin{aligned}\|f\|_{\mathrm{cb}} &\geq \frac{1}{\sqrt{e(d+1)}} \sum_{i=1}^{n} \sqrt{\mathrm{Inf}_{b,i}(f)} \\ &\geq \frac{1}{\sqrt{e(d+1)}} \sum_{i=1}^{n} \frac{\mathrm{Inf}_{b,i}(f)}{\sqrt{\mathrm{MaxInf}(f)}} \geq \frac{\mathrm{Var}[f]}{\sqrt{e(d+1)\cdot \mathrm{MaxInf}(f)}},\end{aligned}$$

where the last inequality follows since for any homogeneous block-multilinear form the sum of influences of variables in *any* one block equals $\mathrm{Var}[f]$ (see (7) in the preliminaries). Then, if $\|f\|_{\mathrm{cb}} \leq 1$, it follows that

$$\mathrm{MaxInf}(f) \geq \frac{(\mathrm{Var}[f])^2}{e(d+1)}.$$

The non-homogeneous case (Theorem 3) requires a bit more care and we use the inequality as an intermediate step to prove Theorem 3 with a worse polynomial dependence on d.

Combined with the results of [1], we obtain that completely bounded forms can be well-approximated by classical query algorithms (decision trees) on most inputs.

▶ **Corollary 5.** *Let $\epsilon, \delta > 0$ and let $f : \{\pm 1\}^{n\times d} \to \mathbb{R}$ be a degree-d block-multilinear form with $\|f\|_{\mathrm{cb}} \leq 1$. Then, there is a deterministic classical algorithm that makes $O(d^5\epsilon^{-8}\delta^{-5})$ queries and approximates $f(x)$ up to an additive error ϵ on $1-\delta$ fraction of the inputs $x \in \{\pm 1\}^{n\times d}$.*

1.1.1 Application to quantum algorithms

We consider quantum query algorithms of the type shown in Figure 1. Any such algorithm has black-box access to the input $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$ where $\boldsymbol{x}_b \in \{\pm 1\}^n$ for each $b \in [d]$, via a phase oracle. In other words, the algorithm can apply the unitary $O_{\boldsymbol{x}_b} = \mathrm{Diag}(\boldsymbol{x}_b)$ for each $b \in [d]$.

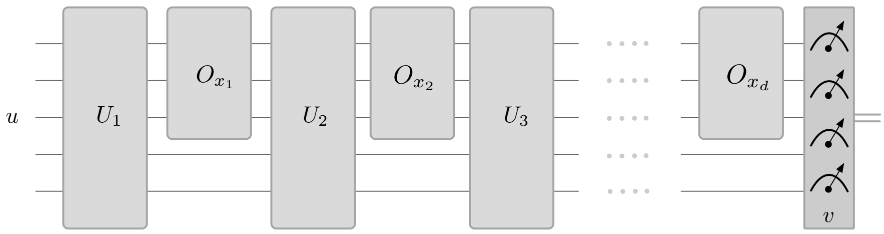

Figure 1 Quantum algorithms considered in Section 1.1.1.

The algorithm starts in some arbitrary quantum state[3] $u \in \mathbb{R}^n$, makes d quantum (phase) queries to oracles $O_{\boldsymbol{x}_b}$ for each $b \in [d]$, and succeeds according to a projective measurement that measures the projection of the final state onto some fixed state $v \in \mathbb{R}^n$. The algorithm is restricted to use each oracle $O_{\boldsymbol{x}_b}$ at most once (which means that each query is made to a disjoint input block not queried previously). The inner product of the state v with the final state at the end of the algorithm is given by the following degree-d block-multilinear form $T : \{\pm 1\}^{n\times d} \to \mathbb{R}$,

$$T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d) = uU_1(O_{\boldsymbol{x}_1} \otimes I_s)U_2(O_{\boldsymbol{x}_2} \otimes I_s)U_3 \cdots (O_{\boldsymbol{x}_d} \otimes I_s)v, \tag{2}$$

and the acceptance probability of the algorithm on input $x = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d)$ is $T(x)^2$.

The connection between such algorithms and completely bounded norm comes from the following proposition in [9].

▶ **Proposition 6** ([9], Theorem 3.2)**.** *Let $T : \{\pm 1\}^{n\times d} \to \mathbb{R}$ be a degree-d block-multilinear form given by (2). Then, $\|T\|_{\mathrm{cb}} \leq 1$.*

Using this connection, applying Corollary 5 to T implies the following almost-everywhere simulation result for quantum algorithms of the type mentioned above.

▶ **Corollary 7.** *The acceptance probability of any d-query quantum algorithm of the type shown in Figure 1 can be estimated up to an additive error ϵ on $1 - \delta$ fraction of the inputs in $\{\pm 1\}^{n\times d}$ by a classical query algorithm making $O(d^5\epsilon^{-8}\delta^{-5})$ queries.*

Note that quantum algorithms of the type considered in the above theorem can already exhibit super-exponential separation over classical algorithms in the query complexity model. For instance, problems like k-fold Forrelation (for $k = O(1)$) or its variants exhibit a $O(1)$ vs $n^{1-1/k}$ separation [10, 31] between the quantum and classical query complexities. In contrast however, Corollary 7 implies that such quantum algorithms can always be simulated almost-everywhere with only a polynomial overhead.

We remark that a general quantum algorithm can query the same block multiple times. Such an algorithm can be converted to an algorithm of the type given in Figure 1 by taking all the blocks $\boldsymbol{x}_b$ to be the same (and allowing some bits to be fixed to constants). However, the corresponding polynomial (2) that results from such an algorithm is not block-multilinear and thus handling this more general case remains an interesting open problem. We discuss the challenges involved in more detail in Section 4.

[3] Throughout this paper, we will assume that all unitaries and states used in the quantum algorithm are real, which one may assume without loss of generality (see e.g. [6]).

1.2 Proof overview

We first consider the case of homogeneous forms and explain the key ideas that go towards proving Theorem 4. We can write a homogeneous block-multilinear form in the following way,

$$\begin{aligned} f(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_d) &= \sum_{i_1,\ldots,i_d\in[n]} \widehat{f}_{i_1,\ldots,i_d}\; x_1(i_1)x_2(i_2)\cdots x_d(i_d) \\ &= \sum_{i=1}^{n} x_1(i) \underbrace{\left(\sum_{i_2,\ldots,i_d\in[n]} \widehat{f}_{i,\ldots,i_d}\; x_2(i_2)\cdots x_d(i_d)\right)}_{:=\, f_i(\boldsymbol{x}_2,\ldots,\boldsymbol{x}_d)}. \end{aligned}$$

For a first attempt, let us try to show that $\|f\|_{\mathrm{cb}}$ must be large by picking x from the discrete cube $\{\pm1\}^{n\times d}$ as follows: for each block b except the first block, we choose $\boldsymbol{x}_b$ uniformly and independently from $\{\pm1\}^n$, and for the first block we take $x_1(i) = \mathrm{sign}(f_i(\boldsymbol{x}_2,\ldots,\boldsymbol{x}_d))$. Taking expectation, this gives us that

$$\|f\|_{\mathrm{cb}} \geq \sum_{i=1}^{n} \mathbb{E}|f_i| \geq 2^{-d/2}\sum_{i=1}^{n}\|f_i\|_2,$$

where the second inequality follows from the multilinear Khintchine inequality[4] which gives us an exponential dependence in d. Note that $\|f_i\|_2^2 = \mathrm{Inf}_{1,i}(f)$ for each i, thus we get that

$$\|f\|_{\mathrm{cb}} \geq 2^{-d/2}\sum_{i=1}^{n}\sqrt{\mathrm{Inf}_{1,i}(f)}. \tag{3}$$

The above also gives a lower bound on $\|f\|_\infty$ which is also a lower bound on $\|f\|_{\mathrm{cb}}$. However, the exponential dependence in d is necessary for the sup-norm as the following example shows.

Example. Consider the following block-multilinear form closely related to the address function. Let $n = 2^d$ and for $a = (a_1,\ldots,a_d)\in\{0,1\}^d$, let $\mathrm{addr}(a)$ denote the unique integer in $[n]$ whose binary expansion equals a. Define the degree-$(d+1)$ homogeneous block-multilinear form $f : \{\pm1\}^{n\times(d+1)} \to \{\pm1\}$ as follows,

$$f(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_d,\boldsymbol{x}_{d+1}) = \sum_{a\in\{0,1\}^d} g_a(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_d)\cdot x_{d+1}(\mathrm{addr}(a)), \tag{4}$$

where $g_a(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_d) : \{\pm1\}^{n\times d} \to \{-1,0,1\}$ is defined as

$$g_a(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_d) = \left(\frac{x_1(1) + (-1)^{a_1}x_1(2)}{2}\right)\cdot\ldots\cdot\left(\frac{x_d(1)+(-1)^{a_d}x_d(2)}{2}\right).$$

Note that f only depends on the first two variables in the blocks $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_d$ (which we refer to as the address blocks) and all the variables in the last block $\boldsymbol{x}_{d+1}$ (which we

---

[4] The multilinear Khintchine inequality states that $\mathbb{E}|f| \geq 2^{-d/2}\|f\|_2$ for a homogeneous degree-d block-multilinear form f. A similar conclusion $\mathbb{E}|f| \geq 3^{-d}\|f\|_2$ holds for any degree-d polynomial f on the hypercube and can be derived from the $(4,2)$-hypercontractive inequality $(\mathbb{E}f^4)^{1/4} \leq 3^{d/2}\|f\|_2$ and using that $\mathbb{E}[|S|] \geq \mathbb{E}[S^2]^{3/2}/\mathbb{E}[S^4]^{1/2}$ for any random variable S by Hölder's inequality.

refer to as the data block). Moreover, $g_a(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d) \in \{\pm 1\}$ iff the parity of bits in the address blocks matches with a, that is $x_b(1)x_b(2) = (-1)^{a_b}$ for every $b \in [d]$, and $g_a(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d) = 0$ otherwise.
It follows that $\|f\|_\infty = 1$, as for any setting of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d$ exactly one term in the summation in (4) survives. However, for $i = \mathrm{addr}(a) \in [n]$,

$$\mathrm{Inf}_{d+1,i}(f) = \mathbb{E}[|g_a|^2] = 2^{-d} = \frac{1}{n},$$

thus $\sum_{i=1}^n \sqrt{\mathrm{Inf}_{d+1,i}(f)} = \sqrt{n} = 2^{d/2}$.

On the other hand, $\|f\|_{\mathrm{cb}} \geq 2^{d/2}$ in the example above – this can be checked by plugging in the following values on the complex unit circle (one-dimensional unitaries): $x_b(1) = 1, x_b(2) = i$ for each $b \in [d]$ and choosing the data block $\boldsymbol{x}_{d+1}$ so that all the magnitudes add up. Thus, one can hope that the freedom to choose large matrices can still allow us to show something like inequality (3) for the completely bounded norm with a polynomial dependence on d, instead of exponential.

Lower bounding $\|f\|_{\mathrm{cb}}$ using Haar random unitaries

Our key observation is that a non-commutative analog of the above strategy works very well. In particular, substituting $N \times N$ Haar random unitaries $U_2(1), \ldots, U_2(n), \ldots, U_d(1), \ldots, U_d(n)$ for the blocks $x_2, \ldots, x_d$ and choosing the block x_1 depending on the polar decomposition of $f_i(\boldsymbol{U}_2, \ldots, \boldsymbol{U}_d)$ allows one to obtain a much larger lower bound on the completely bounded norm $\|f\|_{\mathrm{cb}}$, losing only a polynomial rather than an exponential factor in d.

To obtain quantitative bounds, we need to understand the operator norm of low-degree polynomials of Haar random unitaries. A standard way to upper bound the expected operator norm of random matrices is via the trace method: computing the expected (normalized) trace of the matrix $(AA^*)^m$ for large enough m, and then taking mth root, gives a good control of the operator norm $\|A\|_{\mathrm{op}}$. Since the entries of a Haar random unitary are not independent of one another, it is hard to get a handle on the expected trace directly. A powerful method to understand such quantities is via free probability theory, which considers what happens when the dimension of the matrices $N \to \infty$. In this case, large random matrices behave like free operators, which live on an infinite-dimensional space with a corresponding "trace". We rely on a limiting theorem of Collins and Male [16] who, by strengthening a result of Voiculescu [35], show that the operator norm of a polynomial of Haar random unitaries converges to the operator norm of the polynomial of certain infinite-dimensional operators, called *free Haar unitaries*; thus it suffices to study such free operators.

In free probability theory, such quantities have been studied for a long time (since the work of Haagerup [22]), and we rely on a result of Kemp and Speicher [23] who generalized Haagerup's inequality and showed that for free Haar unitaries one can obtain much better bounds for the operator norm using the usual trace method. In particular, one gets that almost surely as $N \to \infty$, we have

$$\|f_i(\boldsymbol{U}_2, \ldots, \boldsymbol{U}_d)\|_{\mathrm{op}} \leq \mathrm{poly}(d)\|f_i\|_2,$$

in the non-commutative setting. Crucially, the improvement comes because free operators are much more constrained, and many terms that arise while looking at higher moments using the trace method in free probability are zero. One can keep close track of the non-zero terms by using careful combinatorial counting involving what are called *non-crossing partitions*.

Using the above, one can obtain Theorem 4 with the strategy described above using the polar decomposition. The non-homogeneous case requires a bit more technical care, but the key underlying idea is the same.

2 Preliminaries

Notation. Throughout this paper, $[d]$ denotes the set $\{1, 2, \dots, d\}$. For a random vector (or bit-string) z in $\mathbb{R}^n$, we will use z_i or $z(i)$ to denote the i-th coordinate of z, depending on whether we need to use the subscript for another index. We shall use $\boldsymbol{i} = (i_1, \dots, i_d)$ for a d-tuple of indices. For a d-tuple $\boldsymbol{i}$, we write $|\boldsymbol{i}| = d$ to denote the size of the tuple.

For a matrix $M \in \mathbb{C}^{N\times N}$, we denote by M^* its conjugate transpose. Given a string $x \in \mathbb{R}^N$, the $N \times N$ diagonal matrix with x on the diagonal is denoted by $\mathrm{Diag}(x)$. The *normalized trace* of an $N\times N$ matrix M is defined as $\mathrm{tr}_N(M) = \frac{1}{N}\left(\sum_{i=1}^N M_{ii}\right)$. The operator norm of a matrix M is denoted by $\|M\|_{\mathrm{op}}$. The left (resp. right) polar decomposition (V, P) of a square matrix M is a factorization of the form $M = VP$ (resp. $M = PV$) where V is a unitary matrix and P is a positive semidefinite matrix – such a factorization always exists for any square matrix, and can be obtained easily from the singular-value decomposition of M. An $N \times N$ matrix U is called a Haar random unitary if it is distributed according to the Haar measure on the Unitary group $\mathbb{U}(N)$.

Random variables are typically denoted by capital letters (e.g., X). We write $\mathbb{E}[f(X)]$ and $\mathrm{Var}[f(X)]$ to denote the expectation and variance of the random variable $f(X)$; if $f : \{\pm 1\}^m \to \mathbb{R}$ then we abbreviate it to $\mathbb{E}f$ and $\mathrm{Var}[f]$, where the expectation and variance are taken with respect to the uniform measure on the discrete cube $\{\pm 1\}^m$.

Fourier Analysis on the Discrete Cube

We give some basic facts about Fourier analysis on the discrete cube and refer to the book [28] for more details. Every function $f : \{\pm 1\}^m \to \mathbb{R}$ can be written uniquely as a sum of monomials $\chi_S(x) = \prod_{i\in S} x_i$,

$$f(x) = \sum_{S\subseteq[m]} \widehat{f}(S)\chi_S(x), \tag{5}$$

where $\widehat{f}(S) = \mathbb{E}[f(X)\chi_S(X)]$ is the Fourier coefficient with respect to the uniform $X \in \{\pm 1\}^m$. The monomials $\chi_S(x) = \prod_{i\in S} x_i$ form an orthonormal basis for real-valued functions on $\{\pm 1\}^m$, called the *Fourier basis*. Parseval's identity implies that for uniform $X \in \{\pm 1\}^m$,

$$\mathbb{E}f^2 = \sum_{S\subseteq[m]} \widehat{f}(S)^2, \text{ and } \mathrm{Var}[f] = \sum_{S\subseteq[m]:S\neq\emptyset} \widehat{f}(S)^2.$$

For a function on the hypercube, we define $\|f\|_2^2 := \mathbb{E}f^2$ which can also be viewed as the sum of squared Fourier coefficients because of Parseval's identity.

The discrete derivative of a function on the hypercube $\{\pm 1\}^m$ is given by

$$\partial_i f(x) = \frac{1}{2}(f(x^{i\to 1}) - f(x^{i\to -1})),$$

where $x^{i\to b}$ is the same as x except that the i-th coordinate is set to b. It is easily checked that $\partial_i f(x)$ coincides with the real partial derivative $\dfrac{\partial}{\partial x_i}$ of the real multilinear polynomial given by (5).

For a real-valued function $f : \{\pm 1\}^m \to \mathbb{R}$, the influence of a variable x_i on f is defined as

$$\mathrm{Inf}_i(f) = \mathbb{E}|\partial_i f|^2 = \sum_{S \subseteq [m]: i \in S} \widehat{f}(S)^2.$$

Block-multilinear Forms

A degree-d block-multilinear form $f : \{\pm 1\}^{n \times d} \to \mathbb{R}$ is given by

$$f(\boldsymbol{x}_1, \dots, \boldsymbol{x}_d) = \mathbb{E} f + \sum_{m=1}^{d} \sum_{\substack{|\boldsymbol{b}|=m \\ |\boldsymbol{i}|=m}} \widehat{f}_{\boldsymbol{b},\boldsymbol{i}} \cdot \; x_{b_1}(i_1) x_{b_2}(i_2) \cdots x_{b_m}(i_m), \tag{6}$$

where the tuple $\boldsymbol{b} = (b_1, \dots, b_m)$ satisfies $1 \leq b_1 < \dots < b_m \leq d$ and $\boldsymbol{i} \in [n]^m$ is an m-tuple. The expectation that yields the constant term is uniform over $\{\pm 1\}^{n \times d}$. Note that m is determined from the size of the tuple $\boldsymbol{b}$, so we just write $\widehat{f}_{\boldsymbol{b},\boldsymbol{i}}$ above.

From Parseval's identity, the variance of f and the influence of $x_b(i)$ on f (where $b \in [d]$ and $i \in [n]$) are respectively given by

$$\mathrm{Var}[f] = \sum_{m=1}^{d} \sum_{\substack{|\boldsymbol{b}|=m \\ |\boldsymbol{i}|=m}} \widehat{f}^2_{\boldsymbol{b},\boldsymbol{i}} \quad \text{and} \quad \mathrm{Inf}_{b,i}(f) = \sum_{m=1}^{d} \sum_{\substack{|\boldsymbol{b}|=m: b \in \boldsymbol{b} \\ |\boldsymbol{i}|=m: i_b = i}} \widehat{f}^2_{\boldsymbol{b},\boldsymbol{i}}.$$

From the above, it follows that for any block $b \in [d]$,

$$\sum_{i \in [n]} \mathrm{Inf}_{b,i}(f)^2 \leq \mathrm{Var}[f] \leq \sum_{b \in [d], i \in [n]} \mathrm{Inf}_{b,i}(f)^2 \tag{7}$$

where the first inequality is an equality if f is a homogeneous degree-d block-multilinear form. For any $b \in [d]$, we write $\mathrm{MaxInf}_b(T) = \max\{\mathrm{Inf}_{b,i}(T) \mid i \in [n]\}$ to denote the maximum influence of any variable in the block $\boldsymbol{x}_b$.

Note that if f is a degree-d block-multilinear form and if we fix some of the input bits to ± 1, then the resulting function g is also a degree-d block-multilinear form with the same blocks $\boldsymbol{x}_1, \dots, \boldsymbol{x}_d$, but it does not depend on the variables that were fixed. It is also easy to see that $\|g\|_{\mathrm{cb}} \leq \|f\|_{\mathrm{cb}}$ because while computing $\|g\|_{\mathrm{cb}}$ we may restrict the matrix variables to $\pm I$ if that particular variable was set to ± 1. In other words, completely bounded norm does not increase under restrictions.

3 Influence in Completely Bounded Block-multilinear Forms

In this section we prove the non-commutative root-influence inequality (Theorem 4), the special case of the Aaronson-Ambainis conjecture given in Theorem 3, and also briefly mention how the simulation result in Corollary 5 follows from Theorem 3 and the results in [1]. We first need some preliminaries from free probability theory.

3.1 Low-degree polynomials of Haar random unitaries

As discussed in the proof overview, we require bounds on the operator norm (as well as normalized trace) of low-degree polynomials of random unitaries and these follow from known results in free probability theory. Here we explain these connections and also prove some auxillary lemmas needed for the proof of Theorem 4 and Theorem 3.

Let $z_{\boldsymbol{i}}$ denote the non-commutative monomial $z_{i_1} z_{i_2} \cdots z_{i_d}$ for a d-tuple $\boldsymbol{i} = (i_1, \ldots, i_d) \in [t]^d$ and let $p(z_1, \ldots, z_t)$ be a non-commutative polynomial in the variables $z_1, \ldots, z_t$. We are interested in computing the operator norm $\|\cdot\|_{\mathrm{op}}$ and the normalized trace tr_N of the polynomial $p(z_1, \ldots, z_t)$ (or its higher moments) when substituting $N \times N$ Haar random unitaries for the variables z_i.

As explained previously, the theory of free probability gives us tools that allow us to compute the above in the limit $N \to \infty$. In particular, Voiculescu [35] showed that the (normalized) trace of polynomials in Haar random unitaries and their conjugates converges to the trace of the same polynomial evaluated on certain infinite-dimensional operators called *Haar unitaries* that satisfy a non-commutative notion of independence called *free independence*. This was strengthened by Collins and Male [16] who showed that such convergence also holds for the operator norm. A short primer on free probability is given in Appendix A.1, but for now one can think of $\mathcal{A}$ as a self-adjoint algebra of bounded linear operators on a Hilbert space and φ as a trace functional for such operators in the statement given below.

▶ **Theorem 8** ([35, 16]). *Let $p(z_1, \ldots, z_{2t})$ be a non-commutative polynomial in $\mathbb{R}\langle z_1, \ldots, z_{2t}\rangle$. If $U_1, \ldots, U_t$ are $N \times N$ Haar random unitaries, then almost surely,*

$$\begin{aligned} \lim_{N\to\infty} \mathrm{tr}_N[p(U_1, \ldots, U_t, U_1^*, \ldots, U_t^*)] &= \varphi[p(u_1, \ldots, u_t, u_1^*, \ldots, u_t^*)], \\ \lim_{N\to\infty} \|p(U_1, \ldots, U_t, U_1^*, \ldots, U_t^*)\|_{\mathrm{op}} &= \|p(u_1, \ldots, u_t, u_1^*, \ldots, u_t^*)\|, \end{aligned}$$

where $u_1, \ldots, u_t$ are free Haar unitaries in a C^*-probability space $(\mathcal{A}, \varphi)$ and $\|\cdot\|$ is the norm for the underlying C^*-algebra.

Using the above result it suffices to consider free Haar unitaries in a C^*-probability space to compute the operator norm and trace of polynomials of random unitaries. For a non-commutative polynomial $p(z_1, \ldots, z_t) = \sum_{|\boldsymbol{i}|\le d} c_{\boldsymbol{i}} z_{\boldsymbol{i}}$, denoting by $\|p\|_2 = \left(\sum_{|\boldsymbol{i}|\le d} |c_{\boldsymbol{i}}|^2\right)^{1/2}$, one can show the following easily using techniques from free probability.

▶ **Lemma 9.** *Let $p(z_1, \ldots, z_t) = \sum_{|\boldsymbol{i}|\le d} c_{\boldsymbol{i}} z_{\boldsymbol{i}}$ be a non-commutative degree-d polynomial in $\mathbb{R}\langle z_1, \ldots, z_t\rangle$ and $u_1, \ldots, u_t$ be free Haar unitaries in a C^*-probability space $(\mathcal{A}, \varphi)$. Then,*

$$\varphi[p(u_1, \ldots, u_t)(p(u_1, \ldots, u_t))^*] = \|p\|_2^2.$$

The above implies that $\mathrm{tr}_N[p(U_1, \ldots, U_t)(p(U_1, \ldots, U_t))^*]$ converges to $\|p\|_2^2$ almost surely as $N \to \infty$. We shall defer the proof of Lemma 9 to Appendix A, but to aid our intuition we note here that since the U_i's are independent $N \times N$ Haar random unitaries, the expected value

$$\mathbb{E}\left[\mathrm{tr}_N[p(U_1, \ldots, U_t)(p(U_1, \ldots, U_t))^*\right] = \|p\|_2^2,$$

and from concentration of measure, it is natural to expect that it converges to the expected value.

Similarly, to compute the operator norm of $p(U_1, \ldots, U_t)$ for Haar random unitaries one can instead study the norm of the polynomial evaluated on free Haar unitaries. Such bounds are easier to prove using the trace method since free independence imposes strong restrictions on the non-commutative moments. For instance, if U_1 and U_2 are independent $N \times N$ Haar random matrices, then $\mathbb{E}[\mathrm{tr}_N(U_1 U_2 U_1^* U_2^*)]$ is non-zero (albeit quite small), while the corresponding trace evaluated on free Haar unitaries u_1 and u_2 is zero, that is $\varphi(u_1 u_2 u_1^* u_2^*) = 0$. Thus, computing the trace $\varphi[p(u_1, \ldots, u_t, u_1^*, \ldots, u_t^*)]$ reduces to handling the combinatorics of the patterns of u_i's and u_i^*'s.

In particular, we will rely on the following result that follows from the work of Kemp and Speicher [23] who consider the operator norm of homogeneous polynomials evaluated on free R-diagonal operators, a class that includes free Haar unitaries as well. We also remark that a bound where the right-hand side below is worse by a multiplicative $O(d^{1/2})$ factor also follows from the work of Haagerup[5][22] who proved it in another context, predating even the introduction of free probability theory.

▶ **Theorem 10** ([23]). *Let $p(z_1,\dots,z_t) = \sum_{|\boldsymbol{i}|=d} c_{\boldsymbol{i}} z_{\boldsymbol{i}}$ be a homogeneous non-commutative degree-d polynomial in $\mathbb{R}\langle z_1,\dots,z_t\rangle$ and $u_1,\dots,u_t$ be free Haar unitaries in a C^*-probability space. Then,*

$$\|p(u_1,\dots,u_t)\| \le \sqrt{e(d+1)} \cdot \|p\|_2,$$

where the left-hand side denotes the norm in the underlying C^*-algebra.

For completeness, we introduce the necessary free probability background and some combinatorial details in Appendix A, and we present the fairly short proof of Theorem 10 (from [23]) there in a self-contained way. We shall need to extend the above bound to non-homogeneous polynomials. Let $p(z_1,\dots,z_t) = \sum_{|\boldsymbol{i}|\le d} c_{\boldsymbol{i}} z_{\boldsymbol{i}}$ and let $p_k(z_1,\dots,z_t) = \sum_{|\boldsymbol{i}|=k} c_{\boldsymbol{i}} z_{\boldsymbol{i}}$ denote the degree-k homogeneous part of p. Writing $p_k = p_k(u_1,\dots,u_t)$ for $0 \le k \le d$ and $p = p(u_1,\dots,u_t)$, it follows from the triangle inequality, Theorem 10, and Cauchy-Schwarz, that

$$\begin{aligned}\|p\| &\le \sum_{k=0}^{d} \|p_k\| \le \sum_{k=0}^{d} \sqrt{e(k+1)}\|p_k\|_2 \\ &\le \sqrt{e}\left(\sum_{k=0}^{d}(k+1)\right)^{1/2}\left(\sum_{k=0}^{d}\|p_k\|_2^2\right)^{1/2} \le \sqrt{e}(d+1)\cdot\|p\|_2.\end{aligned}$$

Thus, we essentially get the same bound as in the homogeneous case, at the expense of an additional $O(d^{1/2})$ factor.

Collecting all the above we have the following as a direct consequence:

▶ **Theorem 11.** *Let $p(z_1,\dots,z_t) = \sum_{|\boldsymbol{i}|\le d} c_{\boldsymbol{i}} z_{\boldsymbol{i}}$ be a non-commutative degree-d polynomial in $\mathbb{R}\langle z_1,\dots,z_t\rangle$ and $U_1,\dots,U_t$ be independent $N\times N$ Haar random unitaries. Then, the following holds almost surely,*

$$\lim_{N\to\infty} \operatorname{tr}_N[p(U_1,\dots,U_t)(p(U_1,\dots,U_t))^*] = \|p\|_2^2,$$

and

$$\lim_{N\to\infty} \|p(U_1,\dots,U_t)\|_{\mathrm{op}} \le \sqrt{e}(d+1)\cdot\|p\|_2,$$

Moreover, the factor $(d+1)$ in the operator norm bound can be improved to $\sqrt{d+1}$ if additionally p is also assumed to be homogeneous.

Based on the above theorem, we prove the following key lemma which captures the polar decomposition strategy mentioned in the earlier proof overview (Section 1.2). This will serve as the key ingredient in the proof of Theorem 3 and Theorem 4.

[5] We note that Haagerup considered the more general case of polynomials in both u_i's and u_i^*'s.

▶ **Lemma 12.** *Let p be a non-commutative degree-d polynomial in $\mathbb{R}\langle y_1,\dots,y_m,z_1,\dots,z_t\rangle$ given by*

$$p(y_1,\dots,y_m,z_1,\dots,z_t)=\sum_{i=1}^m y_iq_i(z_1,\dots,z_t)+q_0(z_1,\dots,z_t).$$

Then, for every $\delta>0$, there exist an integer N and $N\times N$ unitaries $V_1,\dots,V_m,W_1,\dots,W_t$ such that

$$\|p(V_1,\dots,V_m,W_1,\dots,W_t)\|_{\mathrm{op}}\geq\frac{1}{\sqrt{e}(d+1)}\sum_{i=1}^m\|q_i\|_2-\delta.$$

Moreover, the factor in front can be improved to $(e(d+1))^{-1/2}$ if p is homogeneous as well.

We note that the definition of completely bounded norm allows us to plug in matrices of arbitrarily large dimensions, so the exact dependence on δ is not relevant for our purposes since we can always make it arbitrarily small.

Proof. For an arbitrary integer N, let us pick independent $N\times N$ Haar random unitaries $W_1,\dots,W_t$ which we substitute for the variables $z_1,\dots,z_t$, respectively, and let $M_i=q_i(W_1,\dots,W_t)$ be the corresponding random matrices. Then, for any tuple of matrices $V_1,\dots,V_m$ that we could substitute for the variables $y_1,\dots,y_m$, we have that

$$p(V_1,\dots,V_m,W_1,\dots,W_t)=\sum_{i=1}^m V_iM_i+M_0.$$

Theorem 11 and union bound imply that as $N\to\infty$, with probability 1 all the following events simultaneously hold:

- $\|M_i\|_{\mathrm{op}}\leq\sqrt{e}(d+1)\cdot\|q_i\|_2$ for each i,
- $\mathrm{tr}_N(M_i^*M_i)=\|q_i\|_2^2$ for each i, where $\mathrm{tr}_N(M)$ is the normalized trace.

To show that the operator norm must be large, let us fix a sufficiently large N and a choice of $N\times N$ unitaries $W_1,\dots,W_t$ such that M_i satisfies $\|M_i\|_{\mathrm{op}}\leq\sqrt{e}(d+1)\cdot\|q_i\|_2+\epsilon$ and $\mathrm{tr}_N(M_i^*M_i)\geq\|q_i\|_2^2-\epsilon$ for each $0\leq i\leq m$, where ϵ can be made arbitrarily small by increasing N. For $0\leq i\leq m$, let $M_i=U_iP_i$ be the left polar decomposition of M_i, where U_i is a unitary matrix and P_i is a positive semidefinite matrix.

We select the tuple of unitary matrices $V_1,\dots,V_m$ that we substitute for the variables $y_1,\dots,y_m$ to be $V_i=U_0U_i^*$ for $i\in[m]$. With this we have that $\|p(V_1,\dots,V_m,W_1,\dots,W_t)\|_{\mathrm{op}}$ is at least

$$\begin{aligned}\Big\|M_0+\sum_{i=1}^m V_iM_i\Big\|_{\mathrm{op}}&=\Big\|U_0P_0+\sum_{i=1}^m U_0U_i^*U_iP_i\Big\|_{\mathrm{op}}\\&=\Big\|U_0P_0+\sum_{i=1}^m U_0P_i\Big\|_{\mathrm{op}}\\&=\Big\|P_0+\sum_{i=1}^m P_i\Big\|_{\mathrm{op}}\geq\mathrm{tr}_N\Big(P_0+\sum_{i=1}^m P_i\Big)\geq\mathrm{tr}_N\Big(\sum_{i=1}^m P_i\Big),\end{aligned}$$

where the last equality follows since the operator norm is unitarily invariant and the last two inequalities follow from the positive semidefiniteness of the P_i's.

For every positive semidefinite matrix P, we have that $\mathrm{tr}_N(P)\geq\mathrm{tr}_N(P^2)/\|P\|_{\mathrm{op}}$. Hence,

$$\|p(V_1,\dots,V_m,W_1,\dots,W_t)\|_{\mathrm{op}}\geq\sum_{i=1}^m\frac{\mathrm{tr}_N(P_i^2)}{\|P_i\|_{\mathrm{op}}}.$$

By our choice of M_i, we have that $\mathrm{tr}_N(P_i^2) = \mathrm{tr}_N(M_i^* M_i) \geq \|q_i\|_2^2 - \epsilon$ and $\|P_i\|_{\mathrm{op}} = \|M_i\|_{\mathrm{op}} \leq \sqrt{e}(d+1)\|q_i\|_2 + \epsilon$. Since ϵ can be made arbitrarily small by increasing N, it follows that

$$\|p(V_1, \ldots, V_m, W_1, \ldots, W_t)\|_{\mathrm{op}} \geq \frac{1}{\sqrt{e(d+1)}} \sum_{i=1}^{m} \|q_i\|_2 - \delta,$$

for large enough N. The improved bound for the homogeneous case follows directly by plugging the bound of Theorem 11 into the above proof. ◀

3.2 Non-commutative root-influence inequality

For clarity in the proofs below, we remind our convention that all tuples or blocks are denoted with boldface fonts (e.g. $\boldsymbol{U}_1$ or $\mathbf{A}$), while a single element is denoted without boldface (e.g. $U_1(i)$ or A_i or A). Before proceeding with the proof, we restate the statement for convenience.

▶ **Theorem 4.** (Non-commutative root-influence inequality). *Let f be a homogeneous degree-d block-multilinear form. Then, for blocks $b \in \{1, d\}$,*

$$\|f\|_{\mathrm{cb}} \geq \frac{1}{\sqrt{e(d+1)}} \sum_{i=1}^{n} \sqrt{\mathrm{Inf}_{b,i}(f)}.$$

Proof of Theorem 4. Since f is homogeneous, we can write

$$\begin{aligned} f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d) &= \sum_{i_1, \ldots, i_d \in [n]} \widehat{f}_{i_1, \ldots, i_d}\, x_1(i_1) x_2(i_2) \cdots x_d(i_d) \\ &= \sum_{i=1}^{n} x_1(i) \underbrace{\left(\sum_{i_2, \ldots, i_d \in [n]} \widehat{f}_{i, \ldots, i_d}\, x_2(i_2) \cdots x_d(i_d) \right)}_{:= f_i(\boldsymbol{x}_2, \ldots, \boldsymbol{x}_d)}. \end{aligned}$$

In this case, it follows from (7) that for each $i \in [n]$, we have

$$\mathrm{Var}[f_i] = \|f_i\|_2^2 = \mathrm{Inf}_{1,i}(f) \text{ and } \mathrm{Var}[f] = \sum_{i=1}^{n} \mathrm{Inf}_{1,i}(f). \tag{8}$$

Let us denote the corresponding non-commutative block-multilinear polynomials by $f(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_d)$ and $f_i(\boldsymbol{U}_2, \ldots, \boldsymbol{U}_d)$ where $\boldsymbol{U}_b = (U_b(1), \ldots, U_b(n))$ denotes the b^{th} block of non-commutative variables. To show a lower bound on $\|f\|_{\mathrm{cb}}$ it suffices to exhibit a collection of square matrices $\{U_b(i)\}_{b \in [d], i \in [n]}$ with operator norm at most 1, such that $\|f(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_d)\|_{\mathrm{op}}$ is large.

Applying Lemma 12 for the homogeneous case (with $p = f$, $q_i = f_i$ for $i \in [n]$, and $q_0 = 0$), it follows that for every $\delta > 0$ there exists an integer N and a choice of tuples $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_d$ of $N \times N$ unitaries such that

$$\|f\|_{\mathrm{cb}} \geq \|f(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_d)\|_{\mathrm{op}} \geq \frac{1}{\sqrt{e(d+1)}} \sum_{i \in [n]} \|f_i\|_2 - \delta \overset{(8)}{=} \frac{1}{\sqrt{e(d+1)}} \left(\sum_{i=1}^{n} \sqrt{\mathrm{Inf}_{1,i}(f)} \right) - \delta.$$

Taking $\delta \to 0$, we get the statement of the lemma. The proof for the inequality when $b = d$ is the last block follows similarly by using the right polar decomposition analogue of Lemma 12. ◀

3.3 Aaronson-Ambainis Conjecture for non-homogeneous forms

In this section, we prove Theorem 3, which requires handling non-homogeneous forms. The proof will be similar to the proof of Theorem 4 but we will need to be careful about certain details.

Proof of Theorem 3. Any block-multilinear polynomial $f(x_1, \ldots, x_d)$ can be written as

$$f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d) = \mathbb{E}f + \sum_{b \in [d]} f_b(\boldsymbol{x}_b, \boldsymbol{x}_{b+1}, \ldots, \boldsymbol{x}_d),$$

where f_b consists of all monomials of f that start with a variable in the b^{th} block $\boldsymbol{x}_b$. Note that f_b depends only on the variables in blocks $\boldsymbol{x}_b, \boldsymbol{x}_{b+1}, \ldots, \boldsymbol{x}_d$. Moreover, it follows from Parseval that

$$\operatorname{Var}[f] = \sum_{b \in [d]} \|f_b\|_2^2 = \sum_{b \in [d]} \operatorname{Var}[f_b], \tag{9}$$

so there exists a block $\beta \in [d]$ such that $\operatorname{Var}[f_\beta] \geq \frac{1}{d}\operatorname{Var}[f]$.

Since f_β contributes a lot to the variance, it is natural to try to find an influential variable in the block $\boldsymbol{x}_\beta$. Towards this end, we pull out the variables $x_\beta(i)$ and write

$$f_\beta(\boldsymbol{x}_\beta, \ldots, \boldsymbol{x}_d) = \sum_{i \in [n]} x_\beta(i) f_{\beta,i}(\boldsymbol{x}_{\beta+1}, \ldots, \boldsymbol{x}_d),$$

for block-multilinear polynomials $f_{\beta,i}(\boldsymbol{x}_{\beta+1}, \ldots, \boldsymbol{x}_d)$. Note that some of the $f_{\beta,i}$'s could be identically zero, so let us define S to be the set of those i such that $f_{\beta,i}$ is non-zero. We note that

$$\|f_{\beta,i}\|_2^2 = \operatorname{Inf}_{\beta,i}(f_\beta) \leq \operatorname{Inf}_{\beta,i}(f), \tag{10}$$

which implies (using Parseval's identity) that

$$\frac{1}{d}\operatorname{Var}[f] \leq \operatorname{Var}[f_\beta] = \sum_{i \in S} \|f_{\beta,i}\|_2^2 = \sum_{i \in S} \operatorname{Inf}_{\beta,i}(f_\beta). \tag{11}$$

Denote the corresponding non-commutative block-multilinear polynomials by $f(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_d)$, $f_b(\boldsymbol{U}_b, \ldots, \boldsymbol{U}_d)$, and $f_\beta(\boldsymbol{U}_{\beta+1}, \ldots, \boldsymbol{U}_d)$ where $\boldsymbol{U}_b = (U_b(1), \ldots, U_b(n))$ denotes the b^{th} block of non-commutative variables. To show a lower bound on $\|f\|_{\text{cb}}$ it suffices to exhibit a collection of square matrices $\{U_b(i)\}_{b \in [d], i \in [n]}$ with operator norm at most 1 such that $\|f(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_d)\|_{\text{op}}$ is large.

We set the matrices in blocks $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_{\beta-1}$ to be zero (that is, the all-zero matrix $\mathbf{0}$). Note that with this choice all polynomials $f_b(\boldsymbol{U}_b, \ldots, \boldsymbol{U}_d)$ where $b < \beta$ vanish and the non-commutative polynomial becomes

$$\begin{aligned} &f(\mathbf{0}, \ldots, \mathbf{0}, \boldsymbol{U}_\beta, \boldsymbol{U}_{\beta+1}, \ldots, \boldsymbol{U}_d) \\ &\quad = \sum_{i \in S} U_\beta(i) f_{\beta,i}(\boldsymbol{U}_{\beta+1}, \ldots, \boldsymbol{U}_d) + \sum_{b=\beta+1}^{d} f_b(\boldsymbol{U}_b, \boldsymbol{U}_{b+1}, \ldots, \boldsymbol{U}_d) + \mathbb{E}f, \end{aligned}$$

which is a non-commutative polynomial of the form considered in Lemma 12 (with $m = |S|$, $q_i = f_{\beta,i}$ and $q_0 = \sum_{b=\beta+1}^{d} f_b + \mathbb{E}f$). Thus, by Lemma 12 for every small $\delta > 0$ there exists an integer N and a choice of $N \times N$ matrices for the blocks $\boldsymbol{U}_\beta, \ldots, \boldsymbol{U}_d$ such that

$$\begin{aligned}\|f\|_{\mathrm{cb}} &\geq \|f(\mathbf{0},\ldots,\mathbf{0},\boldsymbol{U}_\beta,\boldsymbol{U}_{\beta+1},\ldots,\boldsymbol{U}_d)\|_{\mathrm{op}}\\ &\geq \frac{1}{\sqrt{e}(d+1)}\sum_{i\in S}\|f_{\beta,i}\|_2-\delta \overset{(10)}{=} \frac{1}{\sqrt{e}(d+1)}\left(\sum_{i\in S}\sqrt{\mathrm{Inf}_{\beta,i}(f_\beta)}\right)-\delta\\ &\overset{(11)}{\geq} \frac{1}{\sqrt{e}(d+1)}\left(\frac{\sum_{i\in S}\mathrm{Inf}_{\beta,i}(f_\beta)}{\sqrt{\mathrm{MaxInf}(f)}}\right)-\delta \overset{(10)}{\geq} \frac{1}{\sqrt{e}(d+1)^2}\left(\frac{\mathrm{Var}[f]}{\sqrt{\mathrm{MaxInf}(f)}}\right)-\delta\end{aligned}$$

Taking $\delta \to 0$ and using the assumption that $\|f\|_{\mathrm{cb}} \leq 1$, we obtain the statement of the theorem:

$$1 \geq \|f\|_{\mathrm{cb}} \geq \frac{1}{\sqrt{e}(d+1)^2}\cdot\frac{\mathrm{Var}[f]}{\sqrt{\mathrm{MaxInf}(f)}} \implies \mathrm{MaxInf}(f) \geq \frac{(\mathrm{Var}[f])^2}{e(d+1)^4}. \qquad ◀$$

3.4 Approximating completely bounded forms with decision trees

In this section, we briefly mention how to obtain Corollary 5. Aaronson and Ambainis [1, Theorem 3.3] showed that querying the most influential variable reduces the variance of the function f, and if that influence is lower bounded by a polynomial in $\mathrm{Var}[f]/d$, then after $\mathrm{poly}(d)$ queries (the exact quantitative dependence can be read off from their proof), the variance of the function becomes small enough so that it can be approximated almost-everywhere by its expectation. Since the family of degree-d block-multilinear forms with completely bounded norm at most one is closed under restrictions, one can apply Theorem 3 repeatedly. This gives us Corollary 5.

4 Discussion and Open Problems

To prove Conjecture 1 in full generality, one would need to consider arbitrary quantum query algorithms: such an algorithm operating on an input $z \in \{\pm 1\}^m$ always makes queries to the same oracle O_z (with a control qubit possibly). One can always convert any such algorithm to the type given in Figure 1 by replacing the oracle O_z used at each step b with a new oracle $O_{\boldsymbol{x}_b}$ where $\boldsymbol{x}_b \in \{\pm 1\}^{m+1}$. The execution of the original algorithm can then be recovered by substituting $\boldsymbol{x}_b = (z,1)$ for every $b \in [d]$. As such one can always obtain a completely bounded block-multilinear form associated with any quantum query algorithm. Conversely, the work [9] shows that the existence of a degree-$2d$ homogeneous block-multilinear form $F : \{\pm 1\}^{(m+1)\times 2d}$ with completely bounded norm at most one also implies the existence of a d-query quantum algorithm whose bias is given by $F((z,1),\ldots,(z,1))$ on every input $z \in \{\pm 1\}^m$.[6] Thus, completely bounded homogeneous block-multilinear forms fully characterize quantum query algorithms in this sense.

In many works in quantum query complexity that concern worst-case complexity, understanding completely bounded or bounded block-multilinear polynomials is sufficient to prove lower bounds as well as give worst-case classical simulation results (i.e. for all inputs), see for instance [2, 14]. However, a transformation that converts a general quantum query algorithm to the type shown in Figure 1 is not conducive to the almost-everywhere results considered in this paper, as the size of the input domain increases exponentially and the number of relevant inputs (i.e. where each $\boldsymbol{x}_b$ is set to the same $(z,1)$) becomes an exponentially small fraction of the new domain.

[6] Note, however, that the polynomial in the variables z obtained after this substitution may not be block-multilinear.

It thus remains an intriguing open problem to see if the characterization of [9] can be used to make further progress on Conjecture 1. One can also hope to make progress on Conjecture 1 without relying on the connection via influences – recently, Aaronson, Ingram and Kretschmer [5] managed to directly prove Conjecture 1 for the special case where the quantum algorithm queries a sparse oracle, without first proving a special case of Conjecture 2.

Another interesting direction is to show that the Aaronson-Ambainis conjecture holds for bounded block-multilinear polynomials, that is, polynomials whose sup-norm on the Boolean hypercube is at most one. While this by itself does not suffice for the application to quantum algorithms as explained above, it might pave the way towards Conjecture 2 in full generality. Lastly, the free-probability toolbox has already found several applications in quantum information theory (see e.g. [36, 17]), and we hope this work will stimulate more applications elsewhere as well.

References

1 Scott Aaronson and Andris Ambainis. The need for structure in quantum speedups. *Theory of Computing*, 10(6):133–166, 2014.

2 Scott Aaronson and Andris Ambainis. Forrelation: A problem that optimally separates quantum from classical computing. *SIAM Journal on Computing*, 47(3):982–1038, 2018.

3 Scott Aaronson, Shalev Ben-David, and Robin Kothari. Separations in query complexity using cheat sheets. In *Proceedings of the 48th Annual ACM Symposium on Theory of Computing*, pages 863–876, 2016.

4 Scott Aaronson, Shalev Ben-David, Robin Kothari, Shravas Rao, and Avishay Tal. Degree vs. approximate degree and quantum implications of Huang's sensitivity theorem. In *Proceedings of the 53rd Annual ACM Symposium on Theory of Computing*, pages 1330–1342, 2021. arXiv:2010.12629.

5 Scott Aaronson, DeVon Ingram, and William Kretschmer. The acrobatics of BQP. *CoRR*, abs/2111.10409, 2021. arXiv:2111.10409.

6 Leonard M. Adleman, Jonathan Demarrais, and Ming-Deh A. Huang. Quantum computability. *SIAM Journal on Computing*, 26(5):1524–1540, 1997.

7 Andris Ambainis. Polynomial degree vs. quantum query complexity. *Journal of Computer and System Sciences*, 72(2):220–238, 2006.

8 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.

9 Srinivasan Arunachalam, Jop Briët, and Carlos Palazuelos. Quantum query algorithms are completely bounded forms. *SIAM Journal on Computing*, 48(3):903–925, 2019.

10 Nikhil Bansal and Makrand Sinha. k-Forrelation optimally separates quantum and classical query complexity. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1303–1316, 2021.

11 Howard Barnum, Michael E. Saks, and Mario Szegedy. Quantum query complexity and semi-definite programming. In *Proceedings of 18th Annual IEEE Conference on Computational Complexity*, pages 179–193, 2003.

12 Robert Beals, Harry Buhrman, Richard Cleve, Michele Mosca, and Ronald de Wolf. Quantum lower bounds by polynomials. *Journal of the ACM*, 48(4):778–797, 2001.

13 Shalev Ben-David, Pooya Hatami, and Avishay Tal. Low-sensitivity functions from unambiguous certificates. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, volume 67 of *LIPIcs*, pages 28:1–28:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

14 Sergey Bravyi, David Gosset, Daniel Grier, and Luke Schaeffer. Classical algorithms for forrelation, 2021. arXiv:2102.06963.

15 Harry Buhrman and Ronald de Wolf. Complexity measures and decision tree complexity: a survey. *Theoretical Computer Science*, 288(1):21–43, 2002.

16 Benoît Collins and Camille Male. The strong asymptotic freeness of Haar and deterministic matrices. *Annales Scientifiques de l'ENS*, (4) 47, fascicule 1:147–163, 2014. `arXiv:1105.4345`.

17 Benoît Collins and Ion Nechita. Random matrix techniques in quantum information theory. *Journal of Mathematical Physics*, 57(1):015215, 2016.

18 Andreas Defant, Mieczyslaw Mastylo, and Antonio Pérez. On the Fourier spectrum of functions on Boolean cubes. *Mathematische Annalen*, 374:653–680, 2018.

19 Irit Dinur, Ehud Friedgut, Guy Kindler, and Ryan O'Donnell. On the Fourier tails of bounded functions over the discrete cube. In *Proceedings of the 38th Annual ACM Symposium on Theory of Computing*, pages 437–446, 2006.

20 Sander Gribling and Monique Laurent. Semidefinite programming formulations for the completely bounded norm of a tensor, 2019. arXiv:1901.04921.

21 Lov K. Grover. A fast quantum mechanical algorithm for database search. In *Proceedings of the 28th Annual ACM Symposium on Theory of Computing*, pages 212–219, 1996.

22 Uffe Haagerup. An example of a non nuclear C^*-algebra, which has the metric approximation property. *Inventiones Mathematicae*, 50:279–293, 1978/79.

23 Todd Kemp and Roland Speicher. Strong Haagerup inequalities for free R-diagonal elements. *Journal of Functional Analysis*, 251(1):141–173, 2007. `arXiv:math/0512481`.

24 Gatis Midrijanis. On randomized and quantum query complexities, 2005. `arXiv:quant-ph/0501142`.

25 Ashley Montanaro. Some applications of hypercontractive inequalities in quantum information theory. *Journal of Mathematical Physics*, 53(12):122206, 2012.

26 Alexandru Nica and Roland Speicher. *Lectures on the Combinatorics of Free Probability*. London Mathematical Society Lecture Note Series. Cambridge University Press, 2006.

27 Noam Nisan and Mario Szegedy. On the degree of Boolean functions as real polynomials. *Computational Complexity*, 4(4):301–313, 1994.

28 Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014.

29 Ryan O'Donnell, Michael E. Saks, Oded Schramm, and Rocco A. Servedio. Every decision tree has an influential variable. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 31–39, 2005.

30 Ryan O'Donnell and Yu Zhao. Polynomial bounds for decoupling, with applications. In *Proceedings of 31st Conference on Computational Complexity*, volume 50 of *LIPIcs*, pages 24:1–24:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2016.

31 Alexander A. Sherstov, Andrey A. Storozhenko, and Pei Wu. An optimal separation of randomized and quantum query complexity. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 1289–1302, 2021.

32 Peter W. Shor. Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. *SIAM Journal on Computing*, 26(5):1484–1509, 1997.

33 Daniel R. Simon. On the power of quantum computation. *SIAM Journal on Computing*, 26(5):1474–1483, 1997.

34 Avishay Tal. Towards optimal separations between quantum and randomized query complexities. In *Proceedings of the 61st IEEE Annual Symposium on Foundations of Computer Science*, pages 228–239, 2020.

35 Dan Voiculescu. A strengthened asymptotic freeness result for random matrices with applications to free entropy. *International Mathematics Research Notices*, 1998:41–63, 1998.

36 Z. Yin, A. W. Harrow, M. Horodecki, M. Marciniak, and A. Rutkowski. Random and free observables saturate the Tsirelson bound for CHSH inequality. *Physical Review A*, 95(032101), 2017. `arXiv:1512.00223`.

A Free Probability Primer

There are many excellent books on free probability theory. In particular, we refer to the book [26] for more details than the brief introduction given here.

A.1 Preliminaries

C^*-algebras

Let $\mathcal{A}$ be a unital C^*-algebra. For our purposes, we can think of this as an algebra of bounded operators on a complex Hilbert space which is self-adjoint ($a \in \mathcal{A}$ implies $a^* \in \mathcal{A}$), closed in the operator norm $\|\cdot\|$, and contains the identity ($\mathbf{1} \in \mathcal{A}$). A faithful trace φ on $\mathcal{A}$ is a continuous linear functional $\varphi : \mathcal{A} \to \mathbb{C}$ that is unital ($\varphi(\mathbf{1}) = 1$), positive $\varphi(aa^*) \geq 0$, and $\varphi(aa^*) = 0$ iff $a = 0$.

The pair $(\mathcal{A}, \varphi)$ where $\mathcal{A}$ is a unital C^*-algebra and φ is a faithful trace on $\mathcal{A}$ is called a *C^*-probability space*. Elements of $\mathcal{A}$ are called non-commutative random variables. An example of a finite-dimensional C^*-probability space is the class $(M_n(\mathbb{C}), \mathrm{tr}_n)$, which is the class of $n \times n$ complex matrices with the normalized trace functional defined as $\mathrm{tr}_n(M) = \frac{1}{n}\sum_{i=1}^n M_{ii}$. General C^*-probability spaces allow us to extend these definitions to infinite-dimensional operators, which are needed to define a non-commutative analog of independence called *free independence*. Faithfulness of the trace φ then ensures that $\|a\| = \lim_{m\to\infty} \varphi((aa^*)^m)^{1/2m}$ (see [26, Proposition 3.17]). In particular, this allows one to compute the norm $\|\cdot\|$ by using the trace method and taking higher powers of the trace functional φ, as we will see later.

An illustrative example of an infinite-dimensional C^*-algebra is the following: let G be an infinite discrete group with the identity element $\mathbf{1}$ and let $\ell^2(G)$ be the associated complex Hilbert space. The canonical orthonormal basis for $\ell^2(G)$ is given by the standard basis vectors $\{e_g \mid g \in G\}$. The left-regular representation of G maps elements of G to bounded linear operators on $\ell^2(G)$ and is defined by its action on the standard basis vectors,

$$\rho(g)e_h = e_{gh}, \ \forall h \in G.$$

The operator norm closure of the linear span of $\{\rho(g) \mid g \in G\}$ forms a unital algebra $\mathcal{A}$ (called the reduced C^*-algebra of G) and together with the faithful trace φ defined as

$$\varphi\left(\sum_g \alpha_g \rho(g)\right) = \alpha_{\mathbf{1}}$$

gives a C^*-probability space.

Free Independence

Let $(\mathcal{A}, \varphi)$ be a C^*-probability space and let $\{\mathcal{A}_i\}_{i=1}^n$ be unital $*$-subalgebras of $\mathcal{A}$. They are said to be *free* (or *freely independent*) if for all $k \in [n]$, for all indices $i_1, \dots, i_k \in [n]$, and for all $a_1 \in \mathcal{A}_{i_1}, \dots, a_k \in \mathcal{A}_{i_k}$ satisfying $\varphi(a_1) = \ldots = \varphi(a_k) = 0$, the joint *free moment*,

$$\varphi(a_1 \cdots a_k) = 0$$

whenever $j_1 \neq j_2, j_2 \neq j_3, \dots, j_{k-1} \neq j_k$, that is, the free moments vanish when all the neighboring elements in the sequence $a_1, \dots, a_k$ come from subalgebras with distinct indices, for example, $\varphi(a_1a_2a_1^*a_2^*a_3a_2) = 0$.

Non-commutative random variables $a_1, \dots, a_n \in (\mathcal{A}, \varphi)$ are said to be free if the subalgebras $\{\mathcal{A}_i\}_{i=1}^n$ are free, where $\mathcal{A}_i$ is the unital $*$-subalgebra generated by a_i (the linear span of all monomials $a_i^{\epsilon_1} a_i^{\epsilon_2} \cdots a_i^{\epsilon_r}$ where $\epsilon_1, \dots, \epsilon_r \in \{1, *\}$ and $r \in \mathbb{N} \cup \{0\}$). Note that the corresponding unital C^*-subalgebras obtained by taking the norm closure of each $\mathcal{A}_i$ are also freely independent in this case (see [26, Exercise 5.23]).

We remark that the set of free non-commutative random variables is an empty set if the underlying C^*-probability space is finite (for instance $(M_n(\mathbb{C}), \mathrm{tr}_n)$), so to find non-trivial examples one needs to work with infinite-dimensional C^*-probability spaces.

Free Haar Unitaries and Free Groups

Let $(\mathcal{A}, \varphi)$ be a C^*-probability space. An element $u \in \mathcal{A}$ is a *Haar unitary* if it is a unitary, i.e. $uu^* = u^*u = \mathbf{1}$, and if $\varphi(u^k) = 0$ for all non-zero integers k. A family $S = \{u_1, \dots, u_n\} \in \mathcal{A}$ in a C^*-probability space $(\mathcal{A}, \varphi)$ is called a *free Haar unitary family* if each $u \in S$ is a Haar unitary and if $u_1, \dots, u_n$ are free. For notational convenience, let us define $S^* = \{u_1^*, \dots, u_n^*\}$ to be the set of corresponding adjoints.

One can give a concrete construction of free Haar unitaries in terms of the free group. The *free group* F_n with a generating set S of size n is an infinite discrete group constructed as follows: a word is defined to be product of elements of $S \cup S^*$ with $\perp$ denoting the empty word that contains no symbols. A word is called reduced if it does not contain a sub-word of the form gg^* or g^*g for $g \in S$. Given a word that is not reduced, the process of repeatedly removing such sub-words until it becomes reduced is called reduction. The free group F_n consists of all reduced words that can be built from the symbols in $S \cup S^*$ with the group operation being a product of words followed by reduction. The identity is the empty word $\perp$.

Let ρ denote the left-regular representation of the free group F_n. Then, writing $S = \{s_1, \cdots, s_n\}$, the family $u_i = \rho(s_i)$ for $i \in [n]$, is a free Haar unitary family in the C^*-probability space given by the reduced C^*-algebra of the group F_n. This connection to the free group also allows us to give a very precise condition when the trace φ evaluated on a non-commutative monomial in the u_i's vanishes. For a d-tuple $\boldsymbol{i} = (i_1, \dots, i_d) \in [m]^d$, let $u_{\boldsymbol{i}}$ denote the non-commutative monomial $u_{i_1} \cdots u_{i_d}$ and write $u_{\boldsymbol{i}}^* = (u_{\boldsymbol{i}})^* = u_{i_d}^* \cdots u_{i_1}^*$. Let $\boldsymbol{i}_1, \dots, \boldsymbol{i}_t, \boldsymbol{j}_1, \dots, \boldsymbol{j}_t$ each be a d-tuple in $[m]^d$ and consider the degree-$2td$ non-commutative monomial $w = u_{\boldsymbol{i}_1} u_{\boldsymbol{j}_1}^* u_{\boldsymbol{i}_2} u_{\boldsymbol{j}_2}^* \cdots u_{\boldsymbol{i}_t} u_{\boldsymbol{j}_t}^*$. Note that a degree-$2td$ monomial w corresponds to an ordered $2td$-tuple of variables. To illustrate, if $t = 1, m = 3$ and $\boldsymbol{i}_1 = (1, 2, 3)$ and $\boldsymbol{j}_1 = (2, 2, 1)$, then $w = u_1 u_2 u_3 (u_2 u_2 u_1)^* = u_1 u_2 u_3 u_1^* u_2^* u_2^*$ and corresponds to the ordered tuple $(u_1, u_2, u_3, u_1^*, u_2^*, u_2^*)$. We can also interpret w as a word in the free group by applying the reduction rules. Then the next proposition follows from the definitions of free independence and Haar unitaries.

▶ **Proposition 13.** *$\varphi(w) = 1$ iff w reduces to identity in the free group F_n, and $\varphi(w) = 0$ otherwise.*

For a monomial w that reduces to identity in the free group, the procedure for reducing a monomial w as above first removes some adjacent pair u_k (at index i) and u_k^* (at index j), then removes another adjacent pair u_l and u_l^* in the resulting word and so on and so forth until we reach the empty word. In particular, this reduction procedure produces a pairing of the set $[2td]$ where the index i and j are paired up iff the variables at indices i and j in the monomial w are u_k and u_k^* (for some k). Moreover, this pairing is what is called a *non-crossing* pairing defined below (see Figure 2). Note that a monomial could be reduced to identity in different ways, so there could be many such non-crossing pairings for a given monomial w.

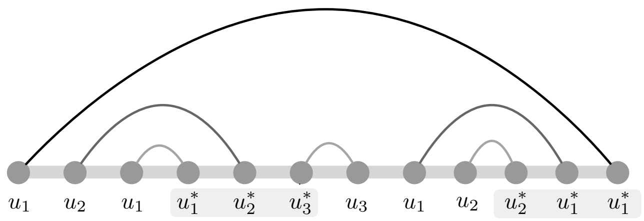

Figure 2 A non-crossing $*$-pairing resulting from the reduction of a word to identity in the free group.

Non-crossing Pairings

For any even integer n, let $\mathcal{P}_2(n)$ denote the set of all pairings of n, that is, the set of all partitions of $[n]$ where each block is of size two. Let $\mathcal{NC}_2(n) \subseteq \mathcal{P}_2(n)$ denote the set of all pairings of $[n]$ that are non-crossing, *i.e.* pairings which do not contain blocks $\{i_1, i_3\}, \{i_2, i_4\}$ such that $i_1 < i_2 < i_3 < i_4$.

For integers d, m, we divide the set $[2dm]$ into $2m$ consecutive blocks of d elements each and color consecutive blocks alternatively with red and blue. Formally, for $i \in [2m]$, the elements $\{(i-1)d+1, \ldots, id\}$ are colored red if i is odd and blue if i is even. We define $\mathcal{NC}_2^*(d,m) \subseteq \mathcal{NC}_2(2dm)$ to be the set of those non-crossing pairings of $[2dm]$ which only pair up elements of different colors. We call any pairing in $\mathcal{NC}_2^*(d,m)$ a $*$-pairing.

We shall need the following combinatorial fact about the number of $*$-pairings (see [23, Corollary 3.2]).

▶ **Lemma 14.** *For all d, m, the number of $*$-pairings $|\mathcal{NC}_2^*(d,m)|$ equals the Fuss-Catalan number*

$$C_{d,m} = \frac{1}{m}\binom{m(d+1)}{m-1} = O\left(\frac{(d+1)^{m(d+1)}}{\left(d+\frac{1}{m}\right)^{md+1}}\right).$$

A.2 Proofs of Lemma 9 and Theorem 11

Proof of Lemma 9. Writing $u_{\boldsymbol{i}}^* = (u_{\boldsymbol{i}})^*$ for a tuple $\boldsymbol{i}$ and using linearity of φ, we have that

$$\varphi[p(u_1, \ldots, u_t)(p(u_1, \ldots, u_t))^*] = \sum_{|\boldsymbol{i}|,|\boldsymbol{j}| \le d} c_{\boldsymbol{i}} c_{\boldsymbol{j}} \varphi(u_{\boldsymbol{i}} u_{\boldsymbol{j}}^*).$$

From Proposition 13, the term $\varphi(u_{\boldsymbol{i}} u_{\boldsymbol{j}}^*)$ is 1 iff $u_{\boldsymbol{i}} u_{\boldsymbol{j}}^*$ reduces to identity in the free group F_t with generators $u_1, \ldots, u_t$. For the right-hand side above, this only happens when $\boldsymbol{i} = \boldsymbol{j}$ and thus these are the only non-zero terms. Thus,

$$\varphi[p(u_1, \ldots, u_t)(p(u_1, \ldots, u_t))^*] = \sum_{|\boldsymbol{i}| \le d} |c_{\boldsymbol{i}}|^2. \qquad ◀$$

Below we present the argument of Kemp and Speicher [23]. Our exposition follows their proof closely but we adapt it to our context.

Proof of Theorem 10. We have that $\|p\| = \lim_{m \to \infty} \left(\varphi((pp^*)^m)\right)^{1/(2m)}$ by the faithfulness of the trace φ. Writing $u_{\boldsymbol{j}}^* = (u_{\boldsymbol{j}})^*$ for a tuple $\boldsymbol{j}$, we can compute

$$\varphi((pp^*)^m) = \sum_{\substack{|\boldsymbol{i}_1| = \ldots = |\boldsymbol{i}_m| = d \\ |\boldsymbol{j}_1| = \ldots = |\boldsymbol{j}_m| = d}} c_{\boldsymbol{i}_1} \cdots c_{\boldsymbol{i}_m} c_{\boldsymbol{j}_1} \cdots c_{\boldsymbol{j}_m} \varphi(u_{\boldsymbol{i}_1} u_{\boldsymbol{j}_1}^* \cdots u_{\boldsymbol{i}_m} u_{\boldsymbol{j}_m}^*).$$

Since $u_1, \ldots, u_t$ are free Haar unitaries, Proposition 13 implies that $\varphi(u_{\boldsymbol{i}_1} u^*_{\boldsymbol{j}_1} \cdots u_{\boldsymbol{i}_m} u^*_{\boldsymbol{j}_m})$ is 1 iff the word $u_{\boldsymbol{i}_1} u^*_{\boldsymbol{j}_1} \cdots u_{\boldsymbol{i}_m} u^*_{\boldsymbol{j}_m}$ reduces to identity in the free group F_t, and is 0 otherwise. Moreover, if the word corresponding to the index $(\boldsymbol{i}_1, \boldsymbol{j}_1, \ldots, \boldsymbol{i}_m, \boldsymbol{j}_m)$ reduces to identity, then there exists a $*$-pairing $\pi \in \mathcal{NC}^*_2(d,m)$ which matches only variables with the same indices. We call any such $*$-pairing π consistent with the $2dm$-tuple $(\boldsymbol{i}_1, \boldsymbol{j}_1, \ldots, \boldsymbol{i}_m, \boldsymbol{j}_m)$ and denote this by the indicator function $\mathbf{1}[\pi, \boldsymbol{i}_1, \boldsymbol{j}_1, \ldots, \boldsymbol{i}_m, \boldsymbol{j}_m]$.

The above implies that we may bound

$$\varphi(u_{\boldsymbol{i}_1} u^*_{\boldsymbol{j}_1} \cdots u_{\boldsymbol{i}_m} u^*_{\boldsymbol{j}_m}) \le \sum_{\pi \in \mathcal{NC}^*_2(d,m)} \mathbf{1}[\pi, \boldsymbol{i}_1, \boldsymbol{j}_1, \ldots, \boldsymbol{i}_m, \boldsymbol{j}_m],$$

where the inequality occurs because there could be multiple $*$-pairings consistent with a tuple. We thus have that

$$\begin{aligned}\varphi((pp^*)^m) &\le \sum_{\substack{|\boldsymbol{i}_1|=\ldots=|\boldsymbol{i}_m|=d \\ |\boldsymbol{j}_1|=\ldots=|\boldsymbol{j}_m|=d}} |c_{\boldsymbol{i}_1} \cdots c_{\boldsymbol{i}_m} c_{\boldsymbol{j}_1} \cdots c_{\boldsymbol{j}_m}| \sum_{\pi \in \mathcal{NC}^*_2(d,m)} \mathbf{1}[\pi, \boldsymbol{i}_1, \boldsymbol{j}_1, \cdots, \boldsymbol{i}_m, \boldsymbol{j}_m] \\ &= \sum_{\pi \in \mathcal{NC}^*_2(d,m)} \sum_{\substack{|\boldsymbol{i}_1|=\ldots=|\boldsymbol{i}_m|=d \\ |\boldsymbol{j}_1|=\ldots=|\boldsymbol{j}_m|=d}} |c_{\boldsymbol{i}_1} \cdots c_{\boldsymbol{i}_m} c_{\boldsymbol{j}_1} \cdots c_{\boldsymbol{j}_m}| \mathbf{1}[\pi, \boldsymbol{i}_1, \boldsymbol{j}_1, \cdots, \boldsymbol{i}_m, \boldsymbol{j}_m].\end{aligned}$$

If a term corresponding to a fixed $*$-pairing π is non-zero, then the list of indices $(\boldsymbol{i}_1, \ldots, \boldsymbol{i}_m)$ is the same as $(\boldsymbol{j}_1, \ldots, \boldsymbol{j}_m)$ up to the exact ordering. Let us relabel $(\boldsymbol{i}_1, \ldots, \boldsymbol{i}_m) = (a_1, \ldots, a_{dm})$ and $(\boldsymbol{j}_1, \ldots, \boldsymbol{j}_m) = (b_1, \ldots, b_{dm})$ and let $c_{a_1,\ldots,a_{dm}} = c_{\boldsymbol{i}_1} \ldots c_{\boldsymbol{i}_m}$ and $c_{b_1,\ldots,b_{dm}} = c_{\boldsymbol{j}_1} \cdots c_{\boldsymbol{j}_m}$. Since π gives a non-crossing bijection between the two lists $(a_1, \ldots, a_{dm})$ and $(b_1, \ldots, b_{dm})$, it holds that $c_{b_1,\ldots,b_{dm}} = c_{\pi(a_1),\ldots,\pi(a_{dm})}$. Thus, the above sum is

$$\begin{aligned}\varphi((pp^*)^m) &\le \sum_{\pi \in \mathcal{NC}^*_2(d,m)} \sum_{a_1,\ldots,a_{dm}} |c_{a_1,\ldots,a_{dm}}| \cdot |c_{\pi(a_1),\ldots,\pi(a_{dm})}| \\ &\le \sum_{\pi \in \mathcal{NC}^*_2(d,m)} \left(\sum_{a_1,\ldots,a_{dm}} |c_{a_1,\ldots,a_{dm}}|^2 \right)^{1/2} \left(\sum_{a_1,\ldots,a_{dm}} |c_{\pi(a_1),\ldots,\pi(a_{dm})}|^2 \right)^{1/2},\end{aligned}$$

where the inequality follows from Cauchy-Schwarz. The two internal summations are exactly the same since the summation is over all dm tuples of indices and π is a bijection. Switching back to the old indexing scheme, the internal summation then equals

$$\sum_{a_1,\ldots,a_{dm}} |c_{a_1,\ldots,a_{dm}}|^2 = \sum_{|\boldsymbol{i}_1|=\ldots=|\boldsymbol{i}_m|=d} |c_{\boldsymbol{i}_1} \cdots c_{\boldsymbol{i}_m}|^2 = \left(\sum_{|\boldsymbol{i}|=d} |c_{\boldsymbol{i}}|^2 \right)^m.$$

Overall, we have

$$\varphi((pp^*)^m) \le |\mathcal{NC}^*_2(d,m)| \left(\sum_{|\boldsymbol{i}|=d} |c_{\boldsymbol{i}}|^2 \right)^m.$$

Using Lemma 14 to bound the number of $*$-pairings,

$$|\mathcal{NC}^*_2(d,m)| = C_{d,m} = \frac{1}{m} \binom{m(d+1)}{m-1} = O\left(\frac{(d+1)^{m(d+1)}}{(d+\frac{1}{m})^{md+1}} \right).$$

Thus, taking the m-th root in the limit $m \to \infty$ yields

$$\|p\|^2 = \lim_{m\to\infty} \varphi((pp^*)^m)^{1/m} = \frac{(d+1)^{d+1}}{d^d} \left(\sum_{|\boldsymbol{i}|=d} |c_{\boldsymbol{i}}|^2 \right) \le e(d+1) \left(\sum_{|\boldsymbol{i}|=d} |c_{\boldsymbol{i}}|^2 \right).$$

This completes the proof of the theorem. ◀