MIP* = RE: A NEGATIVE RESOLUTION TO CONNES' EMBEDDING PROBLEM AND TSIRELSON'S PROBLEM

THOMAS VIDICK

ABSTRACT

Connes' Embedding Problem is a deep question on approximability of certain tracial von Neumann algebras by finite-dimensional matrix algebras. We survey the connections between operator algebras, quantum information and theoretical computer science that enabled the recent resolution of this problem. The resolution goes through an equivalent formulation, known as Tsirelson's problem, in terms of separating convex sets whose definition is motivated by the study of nonlocality in quantum mechanics. We construct an explicit separating hyperplane using the theory of two-player games from complexity theory.

MATHEMATICS SUBJECT CLASSIFICATION 2020

Primary 68A12; Secondary 46L10, 68Q15

KEYWORDS

Quantum interactive proofs, Tsirelson's problem, entanglement

ICM INTERNATIONAL CONGRESS OF MATHEMATICIANS 2022 JULY 6—14

Proc. Int. Cong. Math. 2022, Vol. 6, pp. 4996–5025
DOI 10.4171/ICM2022/71

1. INTRODUCTION

In the 1930s [58] von Neumann laid the foundations for the theory of (what are now known as) von Neumann algebras, with the explicit goal of establishing Heisenberg's matrix-based formulation of quantum mechanics on a rigorous footing. Following the initial explorations of Murray and von Neumann, the new theory progressively took on a life of its own, and von Neumann algebras now routinely make their apparition in areas as diverse as geometry, representation theory, free probability, statistical mechanics, and many others. In his 1976 paper completing the classification of injective von Neumann algebras [14], for which he received the 1982 Fields medal, Alain Connes made a casual remark that has become a central problem in the theory of operator algebras. Paraphrasing, Connes' remark was that any finite von Neumann algebra, i.e., one that has a finite trace, "ought to" be well approximated by finite-dimensional matrix algebras. Thanks to the work of other mathematicians, including Kirchberg and Voiculescu, the remark, now known as *Connes' Embedding Problem* (CEP), rose to prominence as one of the most important open questions in operator algebras. Quoting Vern Paulsen, "The reason that so many operator algebraists care about this conjecture is that it plays much the same role in operator algebras as is played by the Riemann hypothesis in number theory. There are many problems that we would know the answer to, if only Connes were true." For example, Kirchberg showed that CEP is equivalent to the *QWEP conjecture* about the equivalence of the minimal and maximal tensor products on the full group C^* algebra of a nonabelian free group [33]. Voiculescu gave a reformulation in terms of the existence of matrix microstates in free probability [56]. Rădulescu showed that a group is hyperlinear if and only if its group von Neumann algebra satisfies CEP [48]. Goldbring and Hart showed that CEP holds if and only if every type II_1 tracial von Neumann algebra has a computable universal theory [23]. Many more equivalent formulations are known (see, e.g., [11] for a survey).

In these notes we give an overview of an approach to CEP that arose from the study of the nonlocal effects of entanglement in quantum mechanics, and recently led a negative answer to the problem [27]. In the 1980s Boris Tsirelson was placing the study of quantum correlations, i.e., those families of distributions that can be generated from local measurements on a bipartite physical system, on a rigorous mathematical footing. In his work Tsirelson discovered that there was a freedom in deciding how locality should be reflected in the mathematical formalism, and asked if that freedom had observable consequences. Namely, Tsirelson realized that "locality" of measurements could be modeled either by requiring that the Hilbert space associated with the entire system factors as $\mathcal{H} = \mathcal{H}_A \otimes \mathcal{H}_B$, with observables on either system being localized to the corresponding subspace as $A \otimes \mathrm{Id}$ and $\mathrm{Id} \otimes B$, respectively, *or* by allowing the Hilbert space to remain arbitrary but requiring that observables associated with each system mutually commute, i.e., $[A, B] = 0$. While the two models are clearly different from an algebraic point of view, Tsirelson's Problem (TP) asks whether they lead to the same families of distributions, i.e., whether the algebraic distinction has any observable consequence.

Tsirelson's problem rose to prominence due to its relevance for a purely computational task: as we will see later, were TP to have a positive answer then the "largest quantum violation of a Bell inequality," a quantity of much interest to experimentalists, would be computable. (This "largest violation" determines how conclusive an experiment demonstrating the nonlocal effects of quantum mechanics may be.) This realization led to a further study of the problem and a proof of its equivalence with CEP, thus elevating it to the same status as the multitude of other equivalent formulations already mentioned. Moreover, it also led to a potential approach to a negative answer, by showing that the largest quantum violation of a Bell inequality is in fact *not* computable.

The goal of these notes is to explain the relation between CEP and TP, describe the approach to Tsirelson's problem through computability theory, and sketch how that approach eventually led to a resolution of the problem. Our main conceptual tool will be the theory of two-player games, a construct from classical complexity theory which rose to prominence in the 1990s through its connection with the PCP theorem, a sweeping reformulation of the complexity class NP, and applications to hardness of approximation for constraint satisfaction problems. Techniques developed in this study, entirely independent from quantum information, play an essential role in the resolution of CEP.

We start by giving a precise formulation of the two (equivalent) problems which we are concerned with, Connes' Embedding Problem (CEP) and Tsirelson's Problem (TP), in Section 2. In Section 3 we give a first hint of our approach to resolving these problems, which proceeds by constructing hyperplanes separating two convex sets introduced by Tsirelson. This will lead us to introduce nonlocal games as a rich class of hyperplanes to work with. We end the section by sketching a proof of the equivalence between CEP and TP that goes through nonlocal games and an algebra associated to them. In Section 4 we get to the heart of the matter, which is the construction of interesting two-player games and concrete requirements on them that suffice to answer our algebraic problems. It is in this section that complexity theory makes its apparition, as our requirements will push us into the design of very efficient "compression" procedures that find their inspiration in the efficient "proof checking" revolution that led to the PCP theorem in complexity theory. In Section 5 we explain how the complexity-theoretic techniques are combined with ideas from self-testing in quantum information and stability in group representation theory to complete the argument. We end with a brief outlook in Section 6.

2. PROBLEM STATEMENT(S)

We start by reviewing two equivalent, but rather distinct in flavor, formulations of the problem that is the focus of this article. The first formulation is due to Connes [14] and known as *Connes' Embedding Problem* (CEP). The second formulation is due to Tsirelson [53], and we will refer to it as *Tsirelson's Problem* (TP).

2.1. Connes' embedding problem

The standard formulation of CEP states that "every separable type II_1 von Neumann algebra has an approximate embedding into the hyperfinite factor $\mathcal{R}$." Shortly we reformulate this statement using more elementary language. Before doing so we clarify the terms used in Connes' formulation.

A (separable) von Neumann algebra $\mathcal{M}$ is a subalgebra of $B(\mathcal{H})$, the bounded linear operators on a (separable) Hilbert space $\mathcal{H}$, that contains the identity, is closed under taking adjoints (an operation which we denote *), and is closed in the strong operator topology.[1] A *state* τ on $\mathcal{M}$ is a positive linear functional such that $\tau(1) = 1$. A state τ is *tracial* if $\tau(xy) = \tau(yx)$ for all $x, y \in \mathcal{M}$. It is *normal* if the restriction of τ to the unit ball of $\mathcal{M}$ is continuous with respect to the strong operator topology. A *tracial* von Neumann algebra $(\mathcal{M}, \tau)$ is a von Neumann algebra $\mathcal{M}$ equipped with a faithful normal tracial state τ.

A *commutative* von Neumann algebra is isomorphic to $L^\infty(X, \mu)$ for some probability measure space (X, μ). For this reason tracial von Neumann algebras are often thought of as *noncommutative probability spaces*. A von Neumann algebra is a *factor* if it has a trivial center. von Neumann factors are classified in *types*. In their pioneering work on von Neumann algebras, Murray and von Neumann showed that every tracial von Neumann algebra decomposes as a product of type I_n factors, for $1 \leq n < \infty$, and a type II_1 factor. While for any $1 \leq n \leq \infty$, a type I_n factor is always isomorphic to $B(\mathcal{H})$ for some separable Hilbert space $\mathcal{H}$ of dimension n, type II_1 factors are much harder to classify; in fact, there cannot be a classification up to isormorphism by countable structures [49], rendering the problem all but hopeless. (Connes received the Fields medal in 1982 for his work on the classification of type III factors, which are not tracial.)

Murray and von Neumann introduced a specific II_1 factor denoted $\mathcal{R}$ and referred to as *the hyperfinite factor*. Here the use of "the" is justified by the fact that $\mathcal{R}$ is characterized up to isomorphism as the unique separable II_1 factor that satisfies a strong form of approximability by matrix algebras. Namely, $(\mathcal{M}, \tau)$ is said to be *approximately finite-dimensional* (AFD) if for every finite subset F of $\mathcal{M}$ and every $\varepsilon > 0$ there is a $*$-subalgebra $Q \subset M$ such that $Q \simeq M_n(\mathbb{C})$ for some n and for every $x \in F$ there is $y \in Q$ such that $\|x - y\|_2 \leq \varepsilon$.[2] It can be shown that there is a unique AFD II_1 factor, which is referred to as "the hyperfinite factor $\mathcal{R}$" when the specific isomorphism does not matter. Concretely, there are many possible definitions of $\mathcal{R}$. The most straightforward definition, which is also the original one, is as the completion of the algebra $\bigcup_{n \geq 1} M_{2^n}(\mathbb{C})$, where each $M_{2^n}(\mathbb{C})$ isometrically embeds in $M_{2^{n+1}}(\mathbb{C})$ using diagonal blocks. The trace on M is the natural extension of the (dimension-normalized) matrix trace on each M_{2^n}, which we write as $\mathrm{tr}(\cdot)$. With this definition it is immediate that $\mathcal{R}$ is AFD.

There exist some nonhyperfinite tracial von Neumann algebras (we give an example below). CEP is the statement that every such algebra, nevertheless, has some form of weak

1 This is the topology generated by the seminorms $x \mapsto \|xv\|$ for $v \in \mathcal{H}$, with $\|\cdot\|$ the operator norm on $\mathcal{H}$.

2 The norm is given by $\|x\|_2 = \tau(x^*x)^{1/2}$.

approximation by finite-dimensional matrix algebras. The meaning of the second half of the statement of CEP, "has an approximate embedding into the hyperfinite factor $\mathcal{R}$," can be formalized by requiring a trace-preserving embedding into an ultrapower $\mathcal{R}^\omega$. Rather than defining ultrapowers, we give an equivalent formulation due to Voiculescu [57]. For $(\mathcal{M}, \tau)$ a tracial von Neumann algebra and $x_1, \ldots, x_n$ Hermitian elements of $\mathcal{M}$, we say that $(x_1, \ldots, x_n)$ *has matricial microstates* if for every $\varepsilon > 0$ and $N \geq 1$, there is an integer $d \geq 1$ and $A_1, \ldots, A_n \in M_d(\mathbb{C})$ self-adjoint such that for all $p \leq N$ and $i_1, \ldots, i_p \in \{1, \ldots, n\}$,

$$\left|\operatorname{tr}(A_{i_1} \cdots A_{i_p}) - \tau(x_{i_1} \cdots x_{i_p})\right| < \varepsilon.$$

Then CEP is the statement that for any tracial von Neumann algebra $(\mathcal{M}, \tau)$, every tuple (x_i) of self-adjoint elements in $\mathcal{M}$ has matricial microstates. With more work, Kirchberg [33] (see also [16]) showed using the theory of Jordan algebras that CEP is equivalent to the statement that for every tracial von Neumann algebra $(\mathcal{M}, \tau)$, every finite sequence of unitaries $u_1, \ldots, u_n$ in $\mathcal{M}$ and every $\varepsilon > 0$ there are an integer $d \geq 1$ and $U_1, \ldots, U_n$ unitaries in $M_d(\mathbb{C})$ such that for all $i, j \in \{1, \ldots, n\}$,

$$\left|\operatorname{tr}(U_i^* U_j) - \tau(u_i^* u_j)\right| < \varepsilon. \tag{2.1}$$

This last formulation may be appealing to the computer scientist as it states that every finite subset of the unitary group of M approximately embeds into a finite-dimensional matrix unitary group—a form of infinite-dimensional, nonquantitative Johnson–Lindenstrauss lemma [29] for operators.

The versatility of CEP arises from the many examples of tracial von Neumann algebras that are known. We give some examples coming from groups; for many more, see, e.g., [1]. We restrict our attention to discrete, countable groups. For G a countable discrete group, let λ be the left regular representation of G in $\ell^2(G)$. Then the strong operator closure of the linear span of $\lambda(G)$ in $B(\ell^2(G))$ is a von Neumann algebra called the *group von Neumann algebra* of G and denoted $L(G)$. Letting $(\delta_g)_{g \in G}$ be the natural orthonormal basis of $\ell^2(G)$ and $e \in G$ the unit, there is a natural trace φ on $L(G)$ given by $\varphi(x) = \langle \delta_e, x\delta_e \rangle$. One can check that this is a normal faithful tracial state, hence $(L(G), \varphi)$ is a tracial von Neumann algebra. Moreover, $L(G)$ is a factor if and only if G has the i.c.c. property, namely every nontrivial conjugacy class is infinite. Thus the group von Neumann algebra of an infinite i.c.c. group G is a II_1 factor. Some examples are $L(S_\infty)$, where S_∞ is the group of finitely supported permutations of the natural numbers, and $L(\mathbb{F}_n)$ for $n \geq 2$, with $\mathbb{F}_n$ the free group on n generators. It can be shown that $L(G)$ is isomorphic to $\mathcal{R}$ if and only if G is an i.c.c. amenable group. Thus $L(S_\infty)$ is isomorphic to $\mathbb{R}$, whereas $L(\mathbb{F}_n)$ for $n \geq 2$ is not. Connes [14] showed that $L(\mathbb{F}_n)$ satisfies CEP, i.e., it embeds in R^ω, and this discovery prompted his remark about all type II_1 factors.

A group G is *hyperlinear* if and only if for every finite $F \subseteq G$ and $\varepsilon > 0$ there are a $d \geq 1$ and a map $\theta : F \to U_d(\mathbb{C})$ that is an (F, ε)-almost homomorphism. Namely, if $g, h \in F$ are such that $gh \in F$ then $\|\theta(g)\theta(h) - \theta(gh)\|_2 < \varepsilon$, if $e \in F$ then $\|\theta(e) - \mathrm{Id}\,\|_2 < \varepsilon$, and if $x \neq y \in F$ then $\|\theta(x) - \theta(y)\|_2 \geq 1/4$. This formulation is due to Rădulescu [47] who introduced the terminology "hyperlinear." Later, Elek and Szabó [18] showed that the notion

of soficity introduced by Gromov can be characterized in an equivalent manner, requiring θ to map to the symmetric group S_d. Radulescu showed that a countable group G is hyperlinear if and only if $L(G)$ embeds into R^ω, and he gave an example of G, different from $\mathbb{F}_n$, such that $L(G)$ is not hyperfinite but embeds into R^ω, thus giving another example of a nonhyperfinite II_1 factor that satisfies CEP. The conjecture whether every countable group is hyperlinear remains open (as does the stronger conjecture whether every countable group is sofic).

2.2. Tsirelson's problem

In the early 1980s Boris Tsirelson [53] wrote a series of papers laying out the mathematical formalism for the systematic study of the nonlocal properties of quantum mechanics. In quantum mechanics, the state of a physical system is represented by a unit vector $|\psi\rangle$ in a separable Hilbert space $\mathcal{H}$.[3] A measurement (or PVM, for projective-valued measure) is represented by a finite collection $\{P_1, \dots, P_k\}$ of projections on $\mathcal{H}$ such that $\sum_i P_i = \mathrm{Id}$. Here k is the number of outcomes that the measurement can have; according to the Born rule, the probability that the ith outcome is obtained when a system in state $|\psi\rangle$ is measured according to $\{P_i\}$ is given by $\langle\psi|P_i|\psi\rangle$.

Tsirelson was interested in modeling situations in which a physical system is composed of two isolated parts that can be measured independently, by observers present in separated locations.[4] Let us imagine that each observer can make one out of n possible measurements, each with k possible outcomes, on their share of the system. To model the statistical behavior that such an experiment might have, Tsirelson introduced the following subset of $[0,1]^{n^2k^2}$:

$$\begin{aligned} C_{qs}(n,k) = \Big\{ \big(\langle\psi|A_a^x \otimes B_b^y|\psi\rangle\big)_{x,y,a,b} : \mathcal{H}_A, \mathcal{H}_B \text{ Hilbert spaces}, |\psi\rangle \in \mathcal{H}_A \otimes \mathcal{H}_B, \\ \big\||\psi\rangle\big\| = 1, \forall (x,y) \in \{1,\dots,n\}^2, \{A_a^x\}_{a\in\{1,\dots,k\}}, \\ \{B_b^y\}_{b\in\{1,\dots,k\}} \text{ PVM on } \mathcal{H}_A, \mathcal{H}_B \text{ resp.} \Big\}. \end{aligned} \tag{2.2}$$

Here the subscript qs stands for *quantum spatial* and refers to the presence of a tensor product in the expression $\langle\psi|A_a^x \otimes B_b^y|\psi\rangle$. This tensor product is natural if one accepts the rule for associating a Hilbert space to composite systems in nonrelativistic quantum mechanics, which proceeds by tensoring. Thus in the definition of C_{qs} it is understood that observer A's system is modeled using a Hilbert space $\mathcal{H}_A$, observer B's using $\mathcal{H}_B$, and the Hilbert space associated with them jointly is $\mathcal{H}_A \otimes \mathcal{H}_B$, the space in which the system state vector $|\psi\rangle$ lives. Continuing, Tsirelson observed that one could consider an a priori more general

3 We adopt Schödinger's bra-ket notation: a ket $|\psi\rangle$ is used to denote a vector $|\psi\rangle \in \mathcal{H}$, whereas a bra $\langle\psi|$ is used to denote a linear form $\langle\psi| : |\varphi\rangle \in \mathcal{H} \mapsto \langle\psi|\varphi\rangle = \langle\psi,\varphi\rangle \in \mathbb{C}$.

4 We do not make the notion of "separated locations" precise other than through the upcoming formalism; indeed, finding a formalization of it is the entire point of Tsirelson's work. For the moment, the reader can consider that we are only interested in the nonrelativistic scenario.

definition,

$$C_{qc}(n,k) = \Big\{\big(\langle\psi|A_a^x B_b^y|\psi\rangle\big)_{x,y,a,b} : \mathcal{H} \text{ Hilbert space}, |\psi\rangle \in \mathcal{H}, \||\psi\rangle\| = 1, \\ \forall (x,y) \in \{1,\dots,n\}^2, \{A_a^x\}_{a\in\{1,\dots,k\}}, \{B_b^y\}_{b\in\{1,\dots,k\}} \text{ PVM on } \mathcal{H} \\ \text{such that } \big[A_a^x, B_b^y\big] = 0 \forall (a,b) \in \{1,\dots,k\}^2\Big\}. \tag{2.3}$$

Here the subscript qc stands for "quantum commuting" and refers to the fact that in this definition spatial isolation is modeled by the constraint that measurement operators should commute, a condition which also allows for their joint measurability. This definition is more natural from a relativistic viewpoint, e.g., in algebraic quantum field theory, observables associated with space-time isolated regions are required to commute, but there is no a priori separation of the global Hilbert space into tensor products.

Each definition gives rise to a family of convex sets (convexity is easily verified by taking direct sums of PVMs and scaled vectors). Both provide reasonable models for the distributions, sometimes also referred to as *correlations*, that can be generated by an experiment of the form that Tsirelson envisioned. Moreover, in the case all Hilbert spaces are taken to be finite-dimensional, it is an exercise to show that the two sets coincide.[5] Possibly due to this observation, Tsirelson initially assumed that the sets coincide in general, and went on to prove results about the sets C_{qs}; in particular, he introduced techniques to bound certain facets of it. When asked for a proof of the equality, however, Tsirelson realized that it eluded him and posed the question as an open problem.[6]

Tsirelson's problem has two variants. The first, referred to as *Tsirelson's strong problem*, asks about strict equality between the two sets. This problem was answered in 2019 in a beautiful work by Slofstra [51], who showed that the set $C_{qs}(n,k)$ is not closed for all large enough n,k. Since $C_{qc}(n,k)$ is easily verified to be closed, the sets cannot always be equal. Slofstra proved this result by introducing novel techniques relating approximation properties for groups to the suprema of linear functionals on these sets through the language of two-player games, which we will introduce in the next section. In his formulation of the problem, Tsirelson indicated that, if the sets were shown distinct, then an "even more important" problem would arise, which is referred to as the *weak Tsirelson's problem*: does $\overline{C_{qs}(n,k)} = C_{qc}(n,k)$ for all n,k? Here we will refer to this formulation directly as *Tsirelson's Problem* (TP).

While Tsirelson's problem may at first glance look like an arcane question in the foundations of quantum mechanics, there is a good reason why the authors of [41] asked Tsirelson for a proof of his claim regarding equality of the two sets. To explain their motivation, one should bear in mind that the problem of optimizing a linear functional over $C_{qs}(n,k)$ is of primary importance for experiments demonstrating the nonlocality of quantum mechanics, a key feature of the theory that has puzzled physicists and philosophers alike ever since the EPR thought experiment brought it to the fore. Unfortunately, even for small,

5 A slightly more difficult exercise is to show that they always coincide when $n = k = 2$.

6 See "Bell inequalities and operator algebras", available at https://www.tau.ac.il/~tsirel/download/bellopalg.pdf.

fixed n, k, direct optimization over $C_{qs}(n,k)$ seems intractable, as one has no a priori bound on the dimension of the space $\mathcal{H}$ that will lead to an (even approximately) optimal correlation. In their paper, Navascues et al. introduce a decreasing family of outer approximations of the set $C_{qs}(n,k)$ that are each represented as a positive semidefinite set, which implies that optimization over each set can be performed in time commensurate with its description size using semidefinite programming, an extension of linear programming. However, Navascues et al. were only able to show that their outer approximations converge to the set $C_{qc}(n,k)$, instead of $C_{qs}(n,k)$. If Tsirelson's (weak) problem had an affirmative answer, their work would lead to an algorithm for computing the supremum of a linear function over $C_{qs}(n,k)$, or equivalently, computing the largest quantum violation of a Bell inequality. Thus the original motivation for solving Tsirelson's problem is purely computational, and as we will see later, it is surprising also how the problem was eventually resolved.

Further motivation for resolving Tsirelson's problem arose when Fritz [20] and Junge et al. [30] independently showed that Tsirelson's problem follows from Kirchberg's QWEP conjecture, itself shown equivalent to CEP by Kirchberg. Later, Ozawa [43] established the equivalence between the three conjectures, thus tying TP to CEP and the many equivalent formulations of it. In Section 3.2 below we will sketch a different proof of the equivalence between TP and CEP that does not go through the QWEP conjecture.

3. SEPARATING HYPERPLANES AS NONLOCAL GAMES

The formulation of Tsirelson's problem as a question about equality of two convex sets provides a natural geometric approach to its resolution. For $n, k \geq 1$ and $\lambda \in (\mathbb{R}^{n^2k^2})^*$, a linear functional on $\mathbb{R}^{n^2k^2}$, we introduce the quantities (see also Figure 1)

$$\omega_{qa}(\lambda) = \sup_{p \in C_{qs}(n,k)} |\lambda \cdot p| \quad \text{and} \quad \omega_{qc}(\lambda) = \sup_{p \in C_{qc}(n,k)} |\lambda \cdot p|. \tag{3.1}$$

Here the subscript qa stands for "quantum approximate"; we write $C_{qa}(n,k)$ for the closure $C_{qa}(n,k) = \overline{C_{qs}(n,k)}$. We also define a quantity $\omega_{\mathrm{loc}}(\lambda)$, where the supremum is taken over "local" correlations p (this is the case where all PVMs in (2.2) mutually commute, see (3.3) below for a precise definition and a justification of the term "local").

To give a negative answer to Tsirelson's problem, it suffices to find n, k and a λ such that $\omega_{qa}(\lambda) < \omega_{qc}(\lambda)$. In the foundations of quantum mechanics, an inequality of the form $\omega_{\mathrm{loc}}(\lambda) \leq \alpha$ is called a *Bell inequality*, and an inequality of the form $\omega_{qa}(\lambda) \leq \beta$ is called a *Tsirelson inequality*. The best right-hand side in a Tsirelson's inequality is referred to as the "largest quantum violation" of the corresponding optimal Bell inequality. The design of functionals λ such that $\omega_{\mathrm{loc}}(\lambda) < \omega_{qa}(\lambda)$ is relevant to the design of experiments witnessing the "nonlocality" of quantum correlations. Because of this, many functionals have been studied, such as the famous CHSH inequality $\omega_{\mathrm{loc}}(\lambda_{\mathrm{CHSH}}) < 2\sqrt{2}$ where $\lambda_{\mathrm{CHSH}} \in (\mathbb{R}^{2^2 2^2})^*$ is a specific functional named after its inventors, who also showed that it satisfies $\omega_{qa}(\lambda_{\mathrm{CHSH}}) \geq 4$ (Tsirelson later showed that this bound is tight [52]). How does one go about finding interesting λ? One can use guessing and physical intuition for how special quantum phenomena

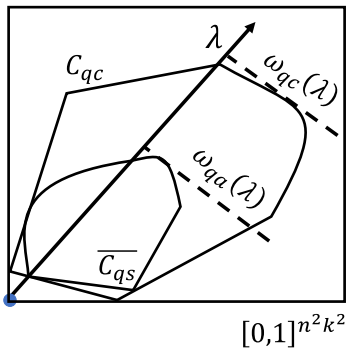

FIGURE 1
Separating convex sets

such as mutual incompatibility of observables might be "detected" by some λ. This, however, can be rather tedious due to the infinite search space: essentially no better algorithm for approaching ω_{qa} from below is known other than enumerating over progressively finer nets in increasing dimensions for the Hilbert space; for approaching it from above, slightly better candidate algorithms are known [41] that work well in practice but, as mentioned earlier, are not even known to converge to the right value—indeed, showing that they do led to formulating Tsirelson's problem, and it follows from the refutation of it that they do not.

In the 1990s, emerging collaborations between physicists and computer scientists stimulated by the nascent field of quantum computation led to the study of a subclass of functionals termed "nonlocal games" which we now introduce.

3.1. Nonlocal games

The idea for a nonlocal game is to interpret the supremum in (3.1) as the optimal winning probability in a certain cooperative two-player game. Let us start with an example of such a game. Fix an n-vertex graph H, as well as a target number of colors $k \geq 1$. The "coloring game" associated with H is played as follows. In the game, two cooperating, but noncommunicating, *players* (traditionally referred to as "Alice" and "Bob") interact with a *referee* as follows. The referee first selects a pair of questions by sampling two vertices of G, x and y, independently and uniformly at random. The referee sends the label x to Alice, and y to Bob. Each player is required to reply with a "color" represented by an integer $a, b \in \{1, \ldots, k\}$, respectively. The referee declares this run of the game as a win for the players if and only if whenever $x = y$ then $a = b$ and whenever (x, y) is an edge in H then $a \neq b$. (If $x \neq y$ is not an edge in H then all answers are accepted.) The players' goal is to maximize their winning probability, taken over the referee's choice of questions, in the game; they are allowed to coordinate their choice of strategy but not to communicate once the game starts.

This last sentence is rather informal; let us make it more precise. What is a valid strategy? For each pair of questions (x, y), the players provide answers according to some

distribution $p(a,b|x,y)$. So a strategy specifies a correlation in the sense of Section 2.2. Physical restrictions on the players' actions translate into restrictions on the class of correlations that are allowed. The informal restriction here is that the players "cannot communicate" with each other. The most natural formalization of this requirement is that players are constrained to compute their answers "locally", using functions $f_A, f_B : \{1,\dots,n\} \to \{1,\dots,k\}$, respectively. For two players determining their answers in this way, the success probability is precisely

$$p_{\text{succ}} = \frac{1}{n^2} \sum_{x,y=1}^{n} (1_{x=y} 1_{f_A(x)=f_B(y)} + 1_{\{x,y\}\in E} 1_{f_A(x)\neq f_B(y)}),$$

where 1_S denotes the characteristic function of a set S and E is the edge set of the graph H. Clearly, this expression is 1 if and only if $f_A = f_B$ is a proper coloring of the graph, i.e., adjacent vertices never get assigned the same color. Thus the game has a local strategy which wins with probability 1 if and only if the chromatic number of H is at most k. This relation, between success probability in a game and a natural graph parameter, hints at rich connections between games and combinatorial optimization, with games providing a conceptual framework in which to study specific questions about combinatorial optimization such as *hardness of approximation*.[7]

Generalizing the preceding example, a (two-player, one-round) game is specified by integers n, k, the number of questions and answers per player in the game, respectively, a distribution π on $\{1,\dots,n\}^2$ according to which questions are chosen, and a decision predicate $V : \{1,\dots,n\}^2 \times \{1,\dots,k\}^2 \to \{0,1\}$ which identifies correct question–answer tuples. With this notation the maximum success probability of a local strategy, which we refer to as the "local value" of the game, is

$$\omega_{\text{loc}}(G) = \sup_{f_A, f_B} \sum_{x,y} \pi(x,y) \sum_{a,b} V(x,y,a,b) 1_{f_A(x)=a} 1_{f_B(y)=b}. \tag{3.2}$$

Defining $\lambda_G \in (\mathbb{R}^{n^2k^2})^*$ by $(\lambda_G)_{x,y,a,b} = \pi(x,y)V(x,y,a,b)$ and introducing the polytope

$$C_{\text{loc}}(n,k) = \text{Conv}\big\{(1_{f_A(x)=a} 1_{f_B(y)=b})_{x,y,a,b} : f_A, f_B : \{1,\dots,n\} \to \{1,\dots,k\}\big\} \tag{3.3}$$

we have that

$$\omega_{\text{loc}}(G) = \sup_{p \in C_{\text{loc}}(n,k)} |\lambda_G \cdot p| = \omega_{\text{loc}}(\lambda_G),$$

justifying our abuse of the notation $\omega_{\text{loc}}(\cdot)$ in (3.2). To summarize, the maximum success probability of local strategies in a game G with n questions and k answers per player can be identified with the supremum of a certain linear functional derived from G over the convex set $C_{\text{loc}}(n,k)$. This connection having been made, a natural question arises: why not consider *quantum* strategies, in which the players would make local measurements on a shared quantum state in order to determine their answers? Instead of a pair of functions, a strategy

7 We emphasize that the games discussed here are entirely distinct from the games considered in the "game theory" of Nash equilibria, where there are two players playing against each other. There is little or no connection between the two areas.

is now modeled as a tuple $\mathscr{S} = (\{A_a^x\}, \{B_b^y\}, |\psi\rangle)$ of measurement operators (PVM) for each player and a shared state $|\psi\rangle$. The no-communication assumption can be implemented by requiring that the tuple satisfies the conditions introduced in the definition of $C_{qs}(n,k)$ in (2.2) (in which case we qualify the strategy as "quantum spatial") or of $C_{qc}(n,k)$ in (2.3) (in which case we qualify it as "quantum commuting").[8] This leads us to define

$$\omega_{qa}(G) = \sup_{p \in C_{qs}(n,k)} |\lambda_G \cdot p| \quad \text{and} \quad \omega_{qc}(G) = \sup_{p \in C_{qc}(n,k)} |\lambda_G \cdot p|. \tag{3.4}$$

Beyond a mere reformulation of the optimization problems (3.1), the framing of linear functionals as two-player (also called "nonlocal" to emphasize their use as witnesses of quantum "nonlocality") games suggests a particular mode of thinking about them, e.g., we can now use intuition about player strategies, questions and answers as opposed to arguably much dryer doubly-indexed families of PVMs.

Going back to the example of the coloring game, each of the quantities in (3.4) leads us to a variant of the chromatic number: for H a graph and G_H the coloring game associated to it, we define the *quantum spatial* (resp. *quantum commuting*) *chromatic number* of H as the smallest k such that $\omega_{qa}(G_H) = 1$ (resp. $\omega_{qc}(G_H) = 1$). Examples of graphs whose quantum spatial chromatic number is strictly smaller than their chromatic number have long been known [10,21]. The possible relevance of the study of the new chromatic numbers to TP and CEP is pointed out in [45], who formulate some related quantities in terms of operator systems; multiple works have since explored further variants of the chromatic number [44,50] and introduced other classes of games that are connected to combinatorial parameters. For example, the coloring game was generalized in [42] to a *graph homomorphism game* whose study led the authors to associate a C^*-algebra with a game; we describe this algebra in the next section. In [4] the authors introduced a *quantum isomorphism game* and a related notion of "quantum isomorphism" of two graphs, and showed that there exist graphs that are quantum isomorphic, but not isomorphic. Further study of this notion led to connections with quantum groups [38] and a surprising characterization of quantum isomorphism in terms of homomorphism counts from *planar graphs* [34] (in contrast, Lovász characterized "classical" graph isomorphism in terms of homomorphism counts from any graph). To summarize, we find that the study of quantum strategies in two-player games has provided a rich framework in which to connect combinatorics and functional analysis, leading to valuable insights in both areas.

3.2. The game algebra

The connection between TP and CEP made in [20,30,43] goes through Kirchberg's QWEP conjecture. An arguably more direct route has more recently been found using nonlocal games. Rather informally, the idea is that a quantum strategy for the players in a game G, i.e., a collection of PVM operators, can be thought of as a certain kind of representation for

8 To show formally that both types of strategies do not imply communication, we compute the marginal distribution on one player's answers and observe that it is independent of the question to the other player.

an abstract algebra $\mathcal{A} = \mathcal{A}(G)$ associated with the game, whose generators are labeled by (question, answer) pairs and whose relations express the game constraints. The (non)existence of different types of successful strategies (quantum spatial, quantum commuting) in the game corresponds to the (non)existence of different kinds of representations for the algebra, thus tying a statement such as $\omega_{qa}(G) < 1 = \omega_{qc}(G)$ to representability properties of $\mathcal{A}$.

To introduce the game algebra more formally, we first describe the class of *synchronous* games to which the construction applies. A game is synchronous if $X = Y$, $A = B$, and for all x and $a \neq b$, $V(x, x, a, b) = 0$, i.e., identical questions always require identical answers. Informally, the synchronicity condition enables to "factor out" the bipartite structure of a game and focus on representing the strategy for a single player.

Definition 3.1. Let $G = (X, A, \pi, V)$ be a synchronous game. The game algebra $\mathcal{A}(G)$ is the abstract unital $*$-algebra generated by elements $\{e_{x,a}\}_{x,a \in X \times A}$ such that for all $x, y \in X$ and $a, b \in A$,

$$e_{x,a}^* = e_{x,a}, \quad e_{x,a}^2 = e_{x,a}, \quad \sum_a e_{x,a} = 1, \quad \text{and} \quad V(x, y, a, b) = 0 \implies e_{x,a} e_{y,b} = 0.^{9}$$

Note that the game algebra may be trivial; for example, if $V(x, y, a, b) = 0$ always then the constraints cannot be satisfied. To see the connection between representations of the game algebra and perfect strategies in G (we call a strategy *perfect* for a certain game if it leads to a winning probability of 1 in the game), as a first exercise one may verify that $\omega_{\text{loc}}(G) = 1$ (i.e., there exists a perfect local strategy for G) if and only if there is a unital $*$-homomorphism from $\mathcal{A}(G)$ into $\mathbb{C}$. (The "if" direction is easier; the synchronicity condition on the game is used for the "only if" direction.) This observation can be generalized as follows.

Theorem 3.2. *Let G be a synchronous game. Then*

(i) **[32, COROLLARY 3.7]** *$\omega_{qa}(G) = 1$ if and only if there is a unital *-representation of $\mathcal{A}(G)$ into $\mathcal{R}^\omega$;*

(ii) **[44, COROLLARY 5.6]** *$\omega_{qc}(G) = 1$ if and only if there is a *-representation of $\mathcal{A}(G)$ into a C^*-algebra with a tracial state.*

Similarly to Voiculescu's reformulation of CEP in terms of microstates or Radulescu's definition of hyperlinearity the condition (i) is equivalent to the existence of approximate representations of $\mathcal{A}(G)$ in finite-dimensional matrix algebras. The theorem implies that the existence of a synchronous game G such that $\omega_{qa}(G) < 1 = \omega_{qc}(G)$ is equivalent to the existence of a tracial C^*-algebra that does not embed into $\mathcal{R}^\omega$; the latter statement is easily seen to be equivalent to the negation of CEP.

We say a few words about the proof of Theorem 3.2. To show the "only if" direction for the second claim, given a commuting strategy $(\{A_a^x\}, \{B_b^y\}, |\psi\rangle)$ there is a natural state on $\mathcal{A}(G)$ given by $\tau(W) = \langle\psi|\varphi(W)|\psi\rangle$ where W is a polynomial in the $e_{x,a}$ and

9 The algebra does not depend on the question distribution π.

$\varphi(W)$ replaces $e_{x,a}$ by A_a^x in W. It is immediate that this is a state; that it is tracial follows (with some work) from the synchronicity condition. For the first claim, a priori the condition $\omega_{qa}(G) = 1$ only gives a sequence of finite-dimensional strategies whose success probability approaches 1. One can turn each such strategy in an approximate representation of $\mathcal{A}(G)$ into finite matrix algebras, eventually leading to a representation into some ultrapower of $\mathcal{R}$.

To show the "if" direction for the second claim, applying the GNS construction, we get PVMs for the first player from any tracial state on $\mathcal{A}(G)$. Constructing appropriate PVMs for the second player requires a little more work; essentially, one uses the trace to construct commuting left and right representations of the game algebra. For the first claim, our starting point is a sequence of approximate representations in finite dimensions. From this we immediately get a sequence of families of PVM for the first player. There is a natural definition for PVM elements for the second player which guarantees that PVM elements associated with different players commute. To conclude, the player's PVMs can be put into the required tensor-product form by appealing to the equivalence between spatial and commuting strategies in finite dimensions.

4. CONSTRUCTING NONLOCAL GAMES

To build intuition about nonlocal games and the associated game algebra, we first review a fundamental example, the "Mermin–Peres Magic Square game." In Section 4.2 we build on this example to construct a family of games whose game algebra has approximate representations into matrix algebras of increasing minimal dimension. In Section 4.3 we outline our approach for turning this family of games into a counterexample to TP. This forces us into complexity-theoretic considerations which we explore in Section 4.4.

4.1. The Magic Square game

We start with a classic example, the *Magic Square game* G_{MS} due to Mermin and Peres [36, 46]. This game is a synchronous game with $n = 6$ questions, which are best visualized as the three rows and three columns of a 3×3 square that can be pictured as follows:

$$\begin{array}{ccccc} y_1 & y_2 & y_3 & & +1 \\ y_4 & y_5 & y_6 & & +1 \\ y_7 & y_8 & y_9 & & +1 \\ & & & & \\ -1 & -1 & -1 & & \end{array}$$

In the game, each of the 6 questions has $k = 4$ possible answers, which are identified with the four possible $\{\pm 1\}$ assignments to the entries of the three squares in the row or column indicated by the question such that the product of the entries is as labeled on the picture, $+1$ for a row and -1 for a column. For example, possible answers to the question associated with the first row are $\{(1, 1, 1), (1, -1, -1), (-1, 1, -1), (-1, -1, 1)\}$, which are identified with the answer set $\{1, \ldots, 4\}$ in some arbitrary way. The game decision predicate V_{MS} enforces the constraint that, whenever the players are asked a row and column that intersect, the values that their respective answers assign to the intersection square(s) should be

identical. For example, if x is associated with the first row and y with the first column then $V_{\mathrm{MS}}(x, y, (1,1,1), (1,1,-1)) = 1$ whereas $V_{\mathrm{MS}}(x, y, (1,1,1), (-1,1,1)) = 0$. Note that this constraint implies that whenever the players are asked the same question then their answers should be identical, hence G_{MS} is a synchronous game.

A local strategy for this game is a pair of functions $f_A, f_B : \{1, \ldots, 6\} \to \{1, \ldots, 4\}$; its success probability is the probability over $x, y \in \{1, \ldots, 6\}$ chosen uniformly at random that $V_{\mathrm{MS}}(x, y, f_A(x), f_B(y)) = 1$. As an exercise, the reader may use the fact that not all constraints in the square can be simultaneously satisfied to show that $\omega_{\mathrm{loc}}(G_{\mathrm{MS}}) = 34/36$. This example illustrates the connection between games and constraint satisfaction problems that has proved so fruitful in complexity theory.

What is the game algebra $\mathcal{A}_{\mathrm{MS}} = \mathcal{A}(G_{\mathrm{MS}})$? Generators for $\mathcal{A}_{\mathrm{MS}}$ are six PVM with four elements each, $\{e_{x,a}\}_{a \in \{1,\ldots,4\}}$ such that $\sum_a e_{x,a} = 1$ for all $x \in \{1, \ldots, 6\}$. An equivalent presentation in terms of self-adjoint operators that square to identity can be found as follows. Let $\{e_{x,a}\}_a$ be the four orthogonal projections associated with the first row. Let $y_1 = e_{x,(1,1,1)} + e_{x,(1,-1,-1)} - e_{x,(-1,1,-1)} - e_{x,(-1,-1,1)}$ and similarly define y_2 and y_3. Then y_1, y_2, y_3 square to 1, pairwise commute, and satisfy $y_1 y_2 y_3 = 1$. Conversely, to any such triple, it is straightforward to associate a four-outcome PVM $\{e_{x,a}\}_a$. A similar construction can be employed for each row and column, a priori leading to 18 y_i operators. However, using the condition that $V(a, b, x, y) = 0 \implies e_{x,a} e_{y,b} = 0$ and the consistency condition enforced in G_{MS}, we get that y_i defined in this way from the PVM associated with the corresponding row must equal to y_i defined from the PVM associated with the column that y_i appears in.

To summarize, $\mathcal{A}_{\mathrm{MS}}$ is generated by elements $y_1, \ldots, y_9$ such that $y_i^* = y_i$, $y_i^2 = 1$, any two y_i appearing in the same row or column of the magic square commute, and the y_i satisfy the magic square row and column constraints, e.g., $y_1 y_4 y_7 = -1$. Our observation that G_{MS} does not have a local strategy that succeeds with probability 1 implies that $\mathcal{A}_{\mathcal{S}}$ has no unital $*$-homomorphism into $\mathbb{C}$. What about homomorphisms in higher-dimensional algebras? With a little work, it is possible to show that there is no such (unital) homomorphism into $M_2(\mathbb{C})$ or $M_3(\mathbb{C})$, but there is one into $M_4(\mathbb{C})$ given by the following operators:

$$\begin{array}{ccc} I \otimes \sigma_Z, & \sigma_Z \otimes I, & \sigma_Z \otimes \sigma_Z, \\ \sigma_X \otimes I, & I \otimes \sigma_X, & \sigma_X \otimes \sigma_X, \\ -\sigma_X \otimes \sigma_Z, & -\sigma_Z \otimes \sigma_X, & \sigma_Y \otimes \sigma_Y, \end{array} \tag{4.1}$$

where

$$\sigma_X = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_Z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad \text{and} \quad \sigma_Y = i\sigma_X\sigma_Z$$

are the Pauli matrices. Moreover, homomorphisms from $\mathcal{A}_{\mathrm{MS}}$ into $M_d(\mathbb{C})$ for some d obey an interesting "rigidity" phenomenon. Let $Y_1, \ldots, Y_9$ be the image of the generators under any such homomorphism. Then it is easy to verify that the row and column constraints imply that

$$\{Y_1, Y_5\} = \{Y_2, Y_4\} = 0, \quad \text{and} \quad [Y_1, Y_2] = [Y_1, Y_4] = [Y_5, Y_2] = [Y_5, Y_4] = 0, \tag{4.2}$$

where $\{A, B\} = AB + BA$ is the anticommutator. Conversely, any four self-adjoint matrices that square to identity and satisfy (4.2) can be extended to a $*$-homomorphism of $\mathcal{A}_{\mathrm{MS}}$. The rigidity phenomenon referred to above is that the algebra generated by any such (finite-dimensional) Y_1, Y_2, Y_4, Y_5 is isomorphic to $M_4(\mathbb{C}) \otimes M_{d'}(\mathbb{C})$ for some d', with $Y_1 \mapsto I \otimes \sigma_Z$, etc. (The other operators, Y_3, Y_6, Y_7, Y_8, Y_9 are uniquely defined from those four by the row & column constraints.) Thus *any* finite-dimensional unital $*$-representation of $\mathcal{A}_{\mathrm{MS}}$ is isomorphic to the representation given in (4.1), possibly tensored with the identity. This very special property allows us to use the fact that a correlation achieves a high success probability in a game, $\lambda_{\mathrm{MS}} \cdot p = 1$, to conclude that any realization of this correlation using PVMs acting on a Hilbert space must satisfy specific algebraic relations; this fact will be crucial to the eventual resolution of TP.

To summarize, the example of the Magic Square helps us demonstrate two important points. Firstly, it is possible to design a game such that $\omega_{qa}(G) = 1$ and, moreover, any strategy that witnesses this is of a certain minimal dimension—here, 4. Secondly, it is possible to force such strategies to have a certain rigid structure—here, the operators used as part of the strategy must contain two pairs of mutually anticommuting operators, such that operators from different pairs commute.

4.2. The Pauli braiding game

To bound $\omega_{qa}(G)$ for some game G it is useful to understand the structure of *approximately* optimal strategies in G. This is because, due to the nonclosure of $C_{qs}(n, k)$, we can have $\omega_{qa}(G) = 1$ without there being any perfect quantum spatial strategy for G, and so it will be convenient to develop techniques that are able to rule out the existence of not only perfect but also near-perfect strategies.

To get us started we state an important tool in the study of approximate group representations.

Theorem 4.1 ([24]). *Let G be a finite group and $f : G \to U_d(\mathbb{C})$ such that*

$$\mathop{E}_{x,y\in G} \operatorname{tr}\big(f(y)^* f(x) f(x^{-1}y)\big) \geq 1 - \varepsilon,$$

for some $\varepsilon \geq 0$ and where the expectation is taken over the choice of a uniformly random pair of elements from G. Then there is a representation $g : G \to U_{d'}(\mathbb{C})$ and an isometry $V : \mathbb{C}^d \to \mathbb{C}^{d'}$ such that

$$\mathop{E}_{x\in G} \big\| f(x) - V^* g(x) V \big\|_2^2 \leq 2\varepsilon.$$

A map f as in the theorem is called an *approximate representation* of the group G; indeed, the condition with $\varepsilon = 0$ is equivalent to that of being a representation. The theorem is an example of a *stability* result, stating that approximate representations are close to exact representations. Here, the measure of "approximate representation" is rather loose, since group relations are only required to hold on average and under the ℓ_2-norm $\|X\|_2 = \operatorname{tr}(X^*X)^{1/2}$ (as opposed to, say, for all relations and under the operator norm). The use of the ℓ_2-norm requires us to allow $d' > d$ in the conclusion of the theorem; that this is necessary is easy to see by "cutting off a corner" from a high-dimensional representation. We

remark that Theorem 4.1 has been extended to the case of amenable groups, with appropriate modifications to allow for infinite-dimensional representations, in [15].

In some cases, such as the Magic Square game studied in the previous section, we can observe that the game algebra is "almost" a group algebra—in fact, it is isomorphic to a quotient of a group C^*-algebra. Namely, if we let $\mathcal{P}_2$ be the group generated by X_1, Z_1, X_2, Z_2, J satisfying $X_1^2 = Z_1^2 = X_2^2 = Z_2^2 = J^2 = 1, [J, X_1] = [J, Z_1] = [J, X_2] = [J, Z_2] = 1$, and $[X_1, X_2] = [X_1, Z_2] = [Z_1, X_2] = [Z_1, Z_2] = 1, [X_1, Z_1] = [X_2, Z_2] = J$ (where now $[a, b] = aba^{-1}b^{-1}$ denotes the group commutator) then it can be verified that $\mathcal{A}(G) \simeq \mathbb{C}(\mathcal{P}_2)/\langle J + 1\rangle$.[10] In particular, any (approximate) representation of $\mathcal{A}(G)$ "descends" to an (approximate) representation of $\mathcal{P}_2$ that (approximately) sends J to -1. Since $\mathcal{P}_2$ has a single (exact) representation that sends J to -1, Theorem 4.1 can be applied to deduce that near-perfect strategies in G_{MS}, i.e., strategies whose winning probability is close but not necessarily equal to 1, must be proportionately close to optimal strategies. In particular, it implies the existence of a constant $\varepsilon_0 > 0$ such that any quantum spatial strategy that succeeds with probability larger than $1 - \varepsilon_0$ in G_{MS} makes use of a Hilbert space for each player that has dimension at least 4; moreover, the algebra generated by the strategy's PVMs contains operators that are close, in the norm $\|\cdot\|_2$, to a representation of the group $\mathcal{P}_2$.

The connection between game algebra and quotient of a group C^*-algebra is quite general and extends to a large class of synchronous games introduced in [12,32] and referred to as *linear constraint system games*; this was shown in [22]. The tools introduced so far suggest the possibility of designing games whose game algebra is isomorphic to quotients of larger groups, such as, for example, the group $\mathcal{P}_N$ which is defined as $\mathcal{P}_2$ but with N pairs of anticommuting generators; this group has a unique irreducible representation sending J to -1, of dimension 2^N. Working out the rules for such a game leads to the following.

Theorem 4.2 ([39]). *There is an $\varepsilon_0 > 0$ and for every $N \geq 2$ a synchronous game $G_{\text{PBT}}^{(N)}$ with $2^{O(N)}$ questions and $O(1)$ answers such that any quantum (spatial or commuting) strategy which succeeds in $G_{\text{PBT}}^{(N)}$ with probability at least $1 - \varepsilon_0$ induces an approximate representation of $\mathcal{P}_N$ sending J to -1 and must have dimension at least 2^N.*[11]

The game from Theorem 4.2 is called the "Pauli braiding game," referring to how the defining (anti)commutation relations "braid" the group generators together. For succinctness, we do not describe this game in its entirety here. To design it, it suffices to find an appropriate decision predicate function that will enforce the group relations. The simpler case of $\mathbb{Z}_2^n$ is known in complexity theory as the "linearity test" of Blum, Luby, and Rubinfeld [8]. This test amounts to verifying that the players' answers a, b, and c to questions $x, y \in \mathbb{Z}_2^n$ and

10 The algebraic relations obtained from the stated relations by sending $J \mapsto -1$ are known as the Weyl–Heisenberg relations.

11 One might worry that Theorem 4.1 only guarantees closeness to a representation up to an isometry, which can change the dimension of the underlying space. This is true, and a little extra work which we skip here is needed to obtain the strict dimension bound mentioned in the theorem.

$x + y$, respectively, are related as $c = a + b$.[12] Blum et al. show that near-perfect *local* strategies are close to homomorphisms from $\mathbb{Z}_2^n$ to $\{-1, 1\}$, and this is extended to finite-dimensional matrix representations in [54]. For the case of $\mathcal{P}_N$, we combine the linearity test for testing the product rule between commuting elements in $\mathcal{P}_N$ and the Magic Square game for testing anticommuting elements. The stated number of questions, $2^{O(N)}$, follows from the number of group elements, which is $2 \cdot 4^N$, and is about quadratically larger due to the use of auxiliary questions that are associated with, e.g., a pair of group elements.

4.3. A fixed-point argument

At this point we have designed an infinite family of games $(G_{\mathrm{PBT}}^{(N)})_{N\geq 1}$ such that for all $N \geq 1$, $\omega_{qa}(G_{\mathrm{PBT}}^{(N)}) = \omega_{qc}(G_{\mathrm{PBT}}^{(N)}) = 1$. While this clearly does not provide a separation, there is more that we may hope to use. In particular, thanks to the rigidity (stability) arguments exposed in the previous section, we know that there is an $\varepsilon_0 > 0$ such that, for any $N \geq 2$ and any quantum spatial strategy for $G_{\mathrm{PBT}}^{(N)}$ that succeeds with probability at least $1 - \varepsilon_0$, the Hilbert space underlying the strategy must have dimension at least the dimension of the smallest representation of $\mathcal{P}_N$ that sends J to -1, i.e., 2^N. For a game G and $p \in [0, 1]$, we let $\mathcal{E}(G; p)$ be the smallest dimension of a strategy that succeeds in G with probability at least p; then, according to Theorem 4.2, we have that

$$\forall N \geq 1, \quad \mathcal{E}(G_{\mathrm{PBT}}^{(N)}; 1 - \varepsilon_0) \geq 2^N. \tag{4.3}$$

Equation (4.3) shows that any quantum strategy of dimension $< 2^N$ has success probability *bounded away from* 1 in $G_{\mathrm{PBT}}^{(N)}$. To complete our goal, it would suffice to create a *single* game G that satisfies this property *for every* $N \geq 1$. Indeed, if $\mathcal{E}(G; 1 - \varepsilon_0) \geq 2^N$ for all N then it follows that $\omega_{qa}(G) < 1$, because the optimal success probability of a quantum spatial strategy in G can be arbitrarily well approximated by finite-dimensional strategies. If, in addition, we are able to guarantee that $\omega_{qc}(G) = 1$ then we will have completed our negative resolution of TP, separating $C_{qa}(n, k)$ from $C_{qc}(n, k)$ for n and k being the number of questions and answers in G, respectively.

The key idea is to define the game G as the *fixed point* of a certain *compression procedure* that transforms families of games such as $(G_{\mathrm{PBT}}^{(N)})_{N\geq 1}$ into other families with comparable size but increased requirements in terms of the minimal dimension of near-optimal strategies. To make this precise, we first need a means of representing infinite families of games. Recall that a *computable function* is $f : \mathbb{N} \to \mathbb{N} \cup \{\perp\}$ such that, informally, there is an algorithm that on input $n \in \mathbb{N}$ returns $f(n)$ if $f(n) \in \mathbb{N}$; if $f(n) = \perp$ the algorithm does not terminate. A computable function is *total* if $f(n) \in \mathbb{N}$ for all n.[13] Computable functions are enumerable and can thus themselves be encoded as integers in a natural way (e.g., via some unambiguous encoding of a Turing machine that computes the function).

---

12 Here it seems like there are three players; a small variant of the test works with two players.

13 We wrote "roughly speaking" because we are not making the notion of algorithm precise. It is a major success of computability theory that essentially any reasonable notion of computability that has been formalized has been shown equivalent to the other notions. For concreteness, one can replace "algorithm" by "Turing machine."

Fix a canonical encoding of games as natural numbers; since the collection of all games (with, say, question distribution that has rational coefficients) is countable, this can be done in a straightforward manner. We say that a function $\mathcal{G} : \mathbb{N} \to \mathbb{N}$ *succinctly represents* the family $(G_N)_{N\geq 1}$ if $\mathcal{G}$ is computable and for every $N \geq 1$, $\mathcal{G}(N)$ is the representation of G_N. Now suppose that there exists a total computable function Compress that, given as input a succinct representation $\mathcal{G}$ for a family of games $(G_N)_{N\geq 1}$, returns a succinct representation $\mathcal{G}'$ for a family of games $(G'_N)_{N\geq 1}$ such that the following conditions hold for all $N \geq 1$:

(C.1) If $\omega_{qa}(G_{N+1}) = 1$ then $\omega_{qa}(G'_N) = 1$;

(C.2) $\mathcal{E}(G'_N; \frac{1}{2}) \geq \min\{\mathcal{E}(G_{N+1}; \frac{1}{2}), N\}$.[14]

In the next section we argue that the existence of such a "compression" procedure is fairly natural once one is familiar with the use of nonlocal games in complexity and cryptography, and in particular with the design of delegated computation protocols using the PCP theorem—buzzwords that will be explained later.[15] For the time being, let us assume that the map Compress exists. We will make use of an additional ingredient in the form of a *refutation procedure* NPA for the quantum commuting value. NPA is an algorithm that takes as input the integer representation of a (single) game G and halts if and only if $\omega_{qc}(G) < 1$. (If $\omega_{qc}(G) = 1$, then NPA(G) runs forever.) The existence of such a procedure follows from the results of Navascues et al. [41] that were already mentioned in Section 2.2, and we take it for granted.

Using these two procedures, Compress and NPA, let us define another function, call it F, that takes as input (the integer representation of) a succinct representation $\mathcal{G}$ for a family of games $(G_N)_{N\geq 1}$ and returns a succinct representation $\mathcal{G}'$ that is defined as follows. (We specify $\mathcal{G}'$ as an algorithm expressed in high-level language, which can ultimately be implemented by some computable function.) On input N, $\mathcal{G}'$ does the following:

(1) It computes the description of $G_1 = \mathcal{G}(1)$.

(2) It runs NPA on G_1 for N steps. If NPA halts, then it returns the description of a trivial game that always accepts.

(3) It computes $\mathcal{T} = \mathsf{Compress}(\mathcal{G})$.

(4) It returns a description of the game $G'_N = \mathcal{T}(N)$.

We observe that, provided $\mathcal{G}$ and Compress are total computable functions, and Compress returns a total computable function when given one as input, F is also a total computable function. Applying a fundamental result in the theory of computable functions, Rogers' fixed point theorem, the map F has a fixed point, call it $\mathcal{G}_\infty$, that is a computable function. Let

14 Here the last "N" can be replaced by any unbounded function of N for the ensuing argument to work.

15 This is not to say that it is straightforward—indeed, the two conditions together already imply that executing Compress once on a trivial family of games that always accept yields an infinite family of games with increasing dimension requirement.

$G_\infty = \mathcal{G}_\infty(1)$ (more precisely, the game whose integer representation is $\mathcal{G}_\infty(1)$). We claim that $\omega_{qa}(G_\infty) < \omega_{qc}(G_\infty) = 1$, thus providing us with the desired separation. To show this, suppose first that $\omega_{qc}(G_\infty) < 1$. Then since $\mathsf{F}(\mathcal{G}_\infty) = \mathcal{G}_\infty$, and since NPA must halt on G_∞ after some number N_∞ of steps, for $N \geq N_\infty$ the game $G_\infty(N)$ is a trivial game that always accepts. By a straightforward induction using property (C.1) of Compress, it follows that $\omega_{qa}(G_\infty) = 1$, hence $\omega_{qc}(G_\infty) = 1$ as well, a contradiction. So $\omega_{qc}(G_\infty) = 1$ and at step 2 NPA never halts. We then get by induction using property (C.2) of Compress that $\mathcal{E}(G_\infty; \frac{1}{2}) \geq N$ for all N. This implies that $\omega_{qa}(G_\infty) \leq \frac{1}{2}$, because no sequence of finite-dimensional strategies can ever get a success probability larger than $\frac{1}{2}$.

The preceding argument shows that to refute TP it "only" remains to design the Compress procedure. This, of course, is the hard part. Before we tackle this task, we discuss a subtle point about the preceding argument.

4.4. Enter complexity

In the analysis of the fixed point $\mathcal{G}_\infty$ of the map F, we implicitly assumed that the game $G_\infty = \mathcal{G}_\infty(1)$ is well defined. What if $\mathcal{G}_\infty$ never halts on input 1? Rogers' fixed point theorem does not guarantee that the fixed point itself is a total function, and it need not be defined on all inputs. Even if it were a total function, there would not be an a priori guarantee that $\mathcal{G}_\infty$ returns well-formed outputs on every input—in general, it will return integers which, depending on our encoding procedure, may not all correspond to well-defined games. Indeed, we should detect that there is something suspicious in the entire setup. A function Compress satisfying all the requirements we have listed is easy to design; for example, $\mathcal{G}' = \mathsf{Compress}(\mathcal{G})$ could on input N return a game that is a mixture of G_{N+1} and $G_{\mathrm{PBT}}^{(N)}$,[16] and this would easily satisfy both (C.1) and (C.2) (indeed, with a stronger bound of 2^N instead of N in (C.2)).

Observe that by virtue of being a fixed point of F, $\mathcal{G}_\infty$, as a family of games, has a size (as a function of N) that is at least that of $\mathsf{Compress}(\mathcal{G}_\infty)$, which has a size that is at least that of $\mathsf{Compress}(\mathsf{Compress}(\mathcal{G}_\infty))$, etc. Therefore, for F to have a fixed point that is a well-defined family of games, it is *necessary* that the procedure Compress lives up to its name, i.e., satisfies the following additional requirement:

(C.3) The size of the game G'_N is smaller than the size of the game G_N.

Here, by "size" we mean the size of an explicit representation of the game, which we can approximate by the total number of questions and answers. In the next section we will see that a more refined notion of size, in terms of the running time of an algorithm computing the referee's questions and its decision, is needed.

While it may not be immediately clear at the level of the discussion, a proper formalization of (C.3), together with small modifications to the description of F (e.g., the intro-

16 What we mean is that the referee would flip a coin to decide which game is played, and inform the players of their decision; for both conditions to hold we'd place a higher probability on $G_{\mathrm{PBT}}^{(N)}$ being played.

duction of a "time-out" condition that ensures that the output of F is always a well-defined family of games, whatever its input), leads to a procedure such that we are able to guarantee that any fixed point is a valid description of a family of games. Thus complexity-theoretic requirements on $\mathsf{Compress}$ arise naturally from our strategy based on identifying $\mathcal{G}_\infty$ as a fixed point, and this beyond the most elementary requirement that the map be computable. In the next section we give some of the main ideas that go in the design of $\mathsf{Compress}$; as we do so, we will discover that more refined complexity-theoretic requirements are required for us to proceed with the construction.

5. COMPRESSION

So how do we implement a "compression" procedure such that (C.1), (C.2), *and* (C.3) hold? Although it has well-established parallels in classical complexity and cryptography, this is a relatively new question in the study of nonlocal games and comparatively few techniques are known for it [19,26,37]. Two main ideas have been used. The first is the idea of *efficient verification of computations*, which takes its origin in classical complexity theory in the 1980s (where it was studied under the name of "program checking" [7]) and received a huge boost when probabilistically checkable proofs (PCP) were discovered in the 1990s [2,3]. The second is the idea of *rigidity*, which we already encountered when analyzing the Magic Square game in Section 4.1 and whose relevance to quantum information and cryptography was first made explicit in work by Mayers and Yao who coined the term "self-testing" for it [35].

In this section we aim to give a flavor of both techniques and how they come together to implement compression. In the process we will see that more refined arguments about complexity make their apparition. As observed in Section 4.4, the design of a procedure which satisfies both (C.1) and (C.2) is relatively straightforward if one does not impose any requirement on how the size of family of games $\mathcal{G}' = \mathsf{Compress}(\mathcal{G})$ depends on that of $\mathcal{G}$. This leads us to reframe the question of implementing compression into one of reducing the size of a game given as input—given a game G (which we think of as $\mathcal{G}(N+1)$), how do we design G' (which we think of as $\mathcal{G}'(N)$) that has similar properties (same $\omega_{qa}(G)$, same dimension requirements) but a smaller number of questions and answers—since we are now talking about the Nth game in the family, and not the $(N+1)$th? We first discuss the problem of reducing the number of answers in a game, and then that of reducing the number of questions.

5.1. The PCP theorem and answer reduction

The colloquial formulation of the PCP theorem is that mathematical proofs can be written in a format such that the validity of the entire proof can be verified by looking only at a few randomly chosen locations of it. It will be useful to express this slightly more formally. First, we fix a *language*, which in general is a subset $L \subseteq \{0,1\}^*$ of strings of bits of any length, and for the example could be the set of all valid statements in, say, Peano arithmetic. Second, we fix a proof verification procedure D that takes as input a statement x and a proof

$\Pi \in \{0,1\}^*$ and returns a bit $D(x,\Pi) \in \{0,1\}$, with 1 indicating that the proof is valid. In the example, D would check that all the claimed steps in Π follow from an axiom and that the proof indeed establishes the statement x.[17] The PCP theorem states that from D it is possible to compute a D' such that D' takes inputs x and Π', is allowed to toss some random coins, but can only look at 10 bits of Π' and then returns a decision in $\{0,1\}$. It should be that (i) for any (x,Π) that D accepts there is a Π', which can be computed from Π, such that D' accepts (x,Π') (this is usually referred to as the "completeness" property) and (ii) for any x such that there is no Π such that D accepts (x,Π) there is also no Π' such that D' accepts (x,Π') with probability larger than $1/3$ (this is referred to as "soundness"—note the apparition of a small probability of error, which can be made arbitrarily small by allowing D' to make more queries to Π', but cannot in general be driven to zero).

There is a crucial requirement for the PCP theorem to apply that is worth emphasizing: the transformation described above is only possible in the case where D is *efficient*, in the sense that the time it takes to evaluate an input (x,Π) is a fixed polynomial in the length of x (in particular, it can only ever access polynomially many bits of Π, which can thus be truncated without loss of generality). This efficiency requirement is crucial to the proof of the PCP theorem, which first represents the entire computation done by D as a "tableau" with intermediate variables associated to each computation step, before finding an encoding of it that can be checked very efficiently; this last step uses techniques from the theory of error-correcting codes. The PCP theorem thus states that proofs that can be verified efficiently, in a number of computation steps polynomial in their length, can be encoded in such a way that verification can also be sparse—only a few bits need to be accessed in order to make a high-confidence decision.[18]

Why is this relevant to our task? Recall that, given a game G, our goal is to find a game G' that is smaller than G and such that (C.1) and (C.2) hold. Here we are concerned with reducing the size of answers in G'; we will address the size of questions in the next section. Fix a pair of questions (x,y) for G. We can think of the referee's task in the game as verifying the claim that "there exists a pair of answers (a,b), *that can be locally produced from* (x,y), such that $V(x,y,a,b)=1$." Setting aside the italicized part, the referee's task amounts to verifying the existence of a proof, the pair (a,b), that passes some verification procedure, $V(x,y,\cdot,\cdot)$. The PCP theorem indicates that there is some "encoding" of (a,b), call it (Π^a,Π^b), that can be verified by only examining a few locations of it. We could then devise another verification procedure V', for the game G', that samples a pair of questions (x,y) as in G, as well as a few locations $(i_1,\dots,i_5)$ that it needs to see in Π^a, $(j_1,\dots,j_5)$ that it needs to see in Π^b, and would send $x'=(x,i_1,\dots,i_5)$ and $y'=(y,j_1,\dots,j_5)$ as its questions. The players would locally compute Π^a and Π^B, respectively, and respond with the requested locations. The PCP theorem would guarantee that this verification procedure

17 Note that here D should accept statements and proofs of any length. Formally, it could be modeled as a Turing machine with two input tapes.

18 "Sparse" is often called "local" in the literature. We use a different word to make the distinction with the notion of "locality" associated with the players in a two-player game, which in our context is distinct.

is essentially equivalent to the original one; however, now the length of answers has been reduced down to a constant.

While this is a plausible sketch for how answer reduction may be achieved, there are a number of major caveats that need to be addressed. Firstly, we implicitly assumed that the PCP encoding of the "proof" (a, b) would naturally take the form (Π^a, Π^b). However, in general a PCP encoding is calculated globally, and such a nice bipartitioning may not (in fact, cannot) hold. Secondly, it is essential for the soundness of the argument that we certify that (a, b) not only exist but can be produced locally from (x, y). Thirdly, again for soundness Π^a should not be allowed to depend on the queries $(i_1, \ldots, i_5)$ that are being made to it: we need to find a mechanism that forces the player to fix it independently of them, even though the referee will never see the entirety of it. Finally and crucially, as already mentioned the PCP theorem only applies to efficient verification procedures. To make use of it here, it is therefore essential that the verification predicate V used in G can be implemented by an algorithm that runs in time polynomial in the length of (x, y).

The last point forces us to rethink our approach. While we initially thought of games as some mildly restrictive formulation of linear forms, the desire to "compress" games puts us face to face with a new algorithmic requirement: we now have to keep track of the complexity of the verification predicate. As long as we do so, however, we have a plan for reducing the size of answers. While this plan raises specific challenges, all of them can be addressed using variations of techniques that have been developed in the decades-long history of using the PCP theorem to implement efficient proof verification in a variety of settings. In the next section we will see how reducing the size of questions prompts us to impose similar efficiency requirements on the procedure used to sample questions $(x, y) \sim \pi$ in G.

5.2. Rigidity and question reduction

In the previous section we saw how techniques developed for the study of PCPs could be leveraged to implement savings in the length of answers in a nonlocal game (at the cost of a small increase in the question length). The idea for reducing the length of questions appears in [40], where it is referred to as "introspection." While the PCP theorem takes its full meaning in a classical context, the idea of introspection makes essential use of quantum-mechanical features, and in particular the possibility to test that incompatible measurements have been made on a shared quantum state.

To explain the idea suppose first that the distribution π on questions in the game G is uniform over $\{(x, x) : x \in \{0, 1\}^N\}$. Suppose that G is modified into a game G' such that with probability $1/2$ the players are asked to play the game $G^{(N)}_{\mathrm{PBT}}$ introduced in Section 4.2 (and with probability $1/2$ they play the original G). Let us see how introducing the Pauli braiding game can be used to force the players to locally generate their own questions in exactly the same way as the referee would have.

For simplicity, let us assume that the players' strategy succeeds with probability 1 in $G^{(N)}_{\mathrm{PBT}}$, when it is played. Again for simplicity let us assume that the state $|\psi\rangle \in \mathcal{H} \otimes \mathcal{H}$ used as part of the players' strategy is a "maximally entangled" state, i.e., it satisfies

$\langle\psi|A\otimes B|\psi\rangle = \mathrm{tr}(AB^T)$ for any $A, B \in B(\mathcal{H})$.[19] In the game $G^{(N)}_{\mathrm{PBT}}$, there is a question associated with each element of $\mathcal{P}_N$, to which the answer is a single bit.[20] For each such question, in the strategy there is a two-outcome PVM $\{A^x_0, A^x_1\}$ that the player applies when receiving the question. We can write each such PVM as an observable $A^x = A^x_0 - A^x_1$. Adapting our notation to the present situation, we conclude that as part of the strategy for every $a, b \in \{0,1\}^N$ there is an observable $A(a,b)$ that corresponds to the player's measurement on the question associated with the group element $X_1^{a_1}\cdots X_N^{a_N} Z_1^{a_1}\cdots Z_N^{a_N} \in \mathcal{P}_N$. Moreover, whenever the strategy has a success probability sufficiently close to 1 in the game, there is an isometry V such that $A(a,b) \simeq V^*\sigma_X(a)\sigma_Z(b)V$, where the approximation is meant in the sense of Theorem 4.1 and we introduced the shorthand $\sigma_X(a) = \sigma_X^{a_1}\otimes\cdots\otimes\sigma_X^{a_n}$ and similarly for $\sigma_Z(b)$.

Consider the following modification to $G^{(N)}_{\mathrm{PBT}}$. Introduce a pair of additional questions, labeled X resp. Z, on which the player is expected to perform the 2^N-outcome PVM that corresponds to a joint measurement of all observables $\{A(a,0^N,0)\}$ resp. $\{A(0^N,b,0)\}$, which is possible since they commute (in a perfect strategy).[21] These two questions have much longer answers which can be used for "randomness generation," in the following sense. Using that $A(0^N,b,0) \simeq V^*\sigma_Z(b)V$, it follows that the PVM applied on question Z is isometric to a rank-1 measurement in the joint eigenbasis of all $\sigma_Z(b)$ (possibly tensored with an irrelevant identity). Since all rank-1 projections have the same trace (recall that here we are working in a finite-dimensional matrix algebras, whose trace is unique) it follows that each answer is obtained with the same probability, $1/2^N$. Furthermore, using that the joint eigenbases of $\{\sigma_X(a)\}$ and of $\{\sigma_Z(b)\}$ are mutually unbiased, it follows that if both players are sent the same question (Z,Z) then they must provide the same uniformly distributed answer, while if the question is (X,Z) they must each provide uncorrelated uniformly random answers. Geometrically this observation corresponds to the statement that any p such that $\lambda_{G^{(N)}_{\mathrm{PBT}}}\cdot p = 1$ has a certain projection (e.g., to $x = Z$, $y = X$) that is proportional to the all-1 vector.

Thus using $G^{(N)}_{\mathrm{PBT}}$, suitably modified by the introduction of additional questions as described above, we can guarantee that any strategy which succeeds with sufficiently high probability in this part of the game must return uniformly random identical answers $(a,a) \in (\{0,1\}^N)^2$ to the question (Z,Z). In this way it is possible to enforce that the players locally generate a pair (a,a) that is distributed exactly as the questions that the referee would send them in the game G. Moreover, the "effort" in doing so is virtually trivial: each player was sent a single question that essentially reads "generate the same random value as your partner!"

---

19 This can be shown to hold without loss of generality whenever the game is a synchronous game, which is the case for all games considered here.

20 There are more questions, which are used to test the group relations; this explains the "$2^{O(N)}$" in Theorem 4.2.

21 The game should *enforce* that such a measurement is being made in any (near)-optimal strategy; this is not hard to achieve.

While this constitutes the main idea—using rigidity to force players to locally generate their own questions—there are many issues to address. Firstly, we need the player not to report the question that they generated, but use it in order to compute an answer that it then sends to the referee. For this there is a simple workaround applying ideas from the previous section: we can in a first step ask the player to report the generated question, as well as its answer, and in a second step to perform answer reduction. Secondly, the distribution of questions in the game need not be uniform over identical (a, a) or independent (a, b) answers. This requires extending the method described above. In particular, "complicated" distributions are likely to be harder to enforce, requiring more questions to a point where no question reduction is achieved—thus another complexity requirement creeps in, that the families of games we consider should have question distributions of bounded complexity, a requirement that should of course be formalized in an appropriate manner. Finally and most importantly, while the "useful" question (Z, Z) in the game is short, ensuring that the player performs the right action on it requires the referee to implement the entire game $G_{\mathrm{PBT}}^{(N)}$. As mentioned in Theorem 4.2, this game has $2^{O(N)}$ questions, which in general will be far larger than the number of questions in G! In the next section we address this issue by discussing a game that has similar guarantees in terms of testing but much smaller question size.

5.3. The quantum low-degree test

In the previous section we sketched how the task of question reduction can be completed, *provided* there is an analogue of the game $G_{\mathrm{PBT}}^{(N)}$ for testing the group $\mathcal{P}_N$ but with a reduced number of questions. Such a result is shown in [27], building on [39].

Theorem 5.1 ([28]). *There are an $\varepsilon_1 > 0$ and for every $N \geq 2$ a synchronous game $G_{\mathrm{LDT}}^{(N)}$ with $2^{\mathrm{poly}\log(N)}$ questions and answers such that any quantum (spatial or commuting) strategy which succeeds with probability at least $1 - \varepsilon_0$ in $G_{\mathrm{LDT}}(N)$ induces an approximate representation of $\mathcal{P}_N$ that sends J to -1 and must have dimension at least 2^N.*

The only difference with Theorem 4.2 is the number of questions and answers, which is now quasipolynomial instead of exponential. This difference hides a deeper difference in terms of how the game is structured. Recall that the Pauli braiding test is built on the linearity test of [8], interpreting the latter as a test for the group $\mathbb{Z}_2^N$ and extending it to a test for $\mathcal{P}_N$ whose analysis could be performed based on Theorem 4.1. With a much smaller number of questions it is no longer possible to have a question associated with each element of the group. Since $\mathcal{P}_N$ can be generated by $2N$ elements it is still possible to have a question per generator, and plausible to show that in optimal strategies the observables associated with each of these questions generate a group isomorphic to $\mathcal{P}_N$. This task, however, will clearly be rather arduous in the case of near-perfect strategies. This is because near-perfect strategies provide a presentation (observables associated to the generators) which satisfy, at best, a certain set of relations on the average to some small constant error (in the norm $\|\cdot\|_2$). Extending the generators to the entire group by taking products will quickly build up the error in a way that, if the only available tool is the triangle inequality, is likely to become unmanageable.

What is needed for Theorem 5.1 is an *efficient* stability result: a small (quasipolynomially many generators and relations) presentation of the group $\mathcal{P}_N$ such that any collection of operators that approximately satisfies the defining relations (in a similar sense as Theorem 4.1) is close to an exact representation—where crucially the closeness should depend on the initial approximation quality but not, or only very mildly, on the size of the group.

Theorem 5.1 is obtained as a quantum extension of the PCP theorem. The generating set it is based on is defined using the Reed–Muller error-correcting code, in a way that we do not have space to detail here. As far as we are aware, it is the only "efficient" (small number of generators) and "robust" (approximate representations are close to exact ones) stability result of its kind and may be a tool of independent interest in other areas.[22] Interestingly, a pared-down version of the result for the group $\mathbb{Z}_2^N$ is used in the analysis of the answer reduction procedure from Section 5.1. Indeed, while the use of the PCP theorem made in that section is a priori entirely classical the analysis needs to take into account quantum strategies for the players, and the classical soundness analysis is not sufficient.

5.4. MIP* = RE

The previous sections complete our sketch of the design of the compression procedure, and following the argument from Section 4.3 of the construction of a correlation separating C_{qc} from $\overline{C_{qs}}$. While we started off without making considerations of complexity, we were led to introduce such considerations due to (1) the requirements for applying the fixed-point argument, and, more crucially, (2) the necessity of using tools such as the PCP theorem to implement the game compression procedure.

A small modification of the definition of the fixed-point $\mathcal{G}_\infty$ leads to an interesting consequence in complexity theory itself. In the definition of F, replace the algorithm NPA used at step 2 by the execution of an arbitrary Turing machine M, i.e., replace the step by "Run M for N steps. If M halts then return the description of a trivial game that always halts." We claim that with this modification the game $G_\infty = \mathcal{G}_\infty(1)$ satisfies $\omega_{qa}(G_\infty) = 1$ if and only if M halts (in some finite number of steps), and $\omega_{qa}(G_\infty) < 1$ otherwise. This can be shown using very similar reasoning to that employed in Section 4.3. Suppose first that M halts. Then step 2 detects this for some large enough N, and, as in Section 4.3, we conclude that $\omega_{qa}(G_\infty) = 1$. If, however, M never halts then step 2 never completes, and, again as in Section 4.3, we quickly see that $\omega_{qa}(G_\infty) \leq \frac{1}{2}$.

Furthermore, it can be verified that the procedure which to M associates the corresponding game $G_\infty = G_\infty(M)$ can be implemented in time polynomial in the description of M. That is, to any Turing machine M we are able to associate a game $G_\infty(M)$ that has a perfect quantum spatial strategy if M halts, and no near-perfect quantum spatial strategy in case M does not halt. In complexity-theoretic terms this establishes a reduction from the halting problem to the problem of deciding between $\omega_{qa}(G) = 1$ and $\omega_{qa}(G) \leq \frac{1}{2}$ (here $\frac{1}{2}$ is an arbitrary positive quantity < 1). The halting problem is a complete problem for the class of recursively enumerable languages RE, while the latter problem is (once properly formu-

22 See, e.g., [6] for a discussion of some group stability results.

lated) complete for the class MIP* of languages that have "quantum multiprover interactive proof systems." Thus the argument establishes the equality MIP* = RE, which gives its title to [27]. From a purely complexity-theoretic standpoint, this equality is interesting because it relates two classes that are a priori defined in very different terms, and it is surprising because the class RE is very large and makes no reference to time complexity at all (the definition of the halting problem does not refer to how much time is allowed for the Turing machine to halt) while the class MIP* does impose efficiency requirements on the verification time, i.e., the time it takes for the referee in the game to generate questions and verify answers to them. It is notable that the equality parallels a celebrated result MIP = NEXP [5], a major stepping stone on the way to the proof of the PCP theorem which is now given a form of "quantum" or "noncommutative" extension. On a more philosophical note, the equality MIP* = RE vindicates the long-witnessed hardness of designing and analyzing interesting Bell inequalities, showing that the optimal quantum bound is in general an uncomputable function of the coefficients of the Bell functional.

6. OUTLOOK

We end with some brief remarks on future work. While in this document we have insisted on the role played by complexity theory in the design of a separating correlation, and hence indirectly in the design of an algebra that refutes Connes' Embedding Problem, we are not aware of a metaargument that would require this. In particular, while it can be formally shown that the complexity-theoretic equality MIP* = RE directly implies a refutation of Tsirelson's problem, the converse is not known to hold. It would be very interesting if a more direct argument, without making any reference to even computability theory, could be found. This has previously been the case, when Slofstra's proof that C_{qs} is not closed [51] (which was closely tied to a proof of undecidability) was later greatly simplified [13, 17], removing all references to computability.

The particular proof technique described here leads to some interesting follow-up questions. In the realm of complexity theory, it is interesting to study variants of complexity classes associated with quantum correlations and characterize their complexity; see [37] for recent work in this direction. In terms of group theory, we believe that the notion of efficient stability put forward in Section 5.3 deserves further study, as stability questions already have a rich history [9, 24, 25, 31]. Of course, an important open question is that of proving the existence of a nonhyperlinear or even nonsofic group; the work outlined in Section 3.2 provides a promising avenue towards this.

ACKNOWLEDGMENTS

I thank Henry Yuen and Vern Paulsen for comments. A small portion of the text of these notes is repurposed from an article by the same author published in the Notices of the AMS [55].

FUNDING

This work is supported by NSF CAREER Grant CCF-1553477, AFOSR YIP award number FA9550-16-1-0495, MURI Grant FA9550-18-1-0161 and the IQIM, an NSF Physics Frontiers Center (NSF Grant PHY-1125565) with support of the Gordon and Betty Moore Foundation (GBMF-12500028).

REFERENCES

[1] C. Anantharaman and S. Popa, An introduction to II_1 factors. Draft book available at https://www.math.ucla.edu/~popa/Books/IIun.pdf, 2017.

[2] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy, Proof verification and the hardness of approximation problems. *J. ACM* **45** (1998), no. 3, 501–555.

[3] S. Arora and S. Safra, Probabilistic checking of proofs: a new characterization of np. *J. ACM* **45** (1998), no. 1, 70–122.

[4] A. Atserias, L. Mančinska, D. E. Roberson, R. Šámal, S. Severini, and A. Varvitsiotis, Quantum and non-signalling graph isomorphisms. *J. Combin. Theory Ser. B* **136** (2019), 289–328.

[5] L. Babai, L. Fortnow, and C. Lund, Non-deterministic exponential time has two-prover interactive protocols. *Comput. Complexity* **1** (1991), no. 1, 3–40.

[6] O. Becker and M. Chapman, Stability of approximate group actions: uniform and probabilistic. 2020, arXiv:2005.06652.

[7] M. Blum and S. Kannan, Designing programs that check their work. *J. ACM* **42** (1995), no. 1, 269–291.

[8] M. Blum, M. Luby, and R. Rubinfeld, Self-testing/correcting with applications to numerical problems. *J. Comput. System Sci.* **47** (1993), no. 3, 549–595.

[9] M. Burger, N. Ozawa, and A. Thom, On Ulam stability. *Israel J. Math.* **1** (2013), no. 193, 109–129.

[10] P. J. Cameron, A. Montanaro, M. W. Newman, S. Severini, and A. Winter, On the quantum chromatic number of a graph. *Electron. J. Combin.* **14** (2007), no. R81, 1.

[11] V. Capraro, In *Connes' embedding conjecture*, pp. 73–107, Springer, Cham, 2015.

[12] R. Cleve and R. Mittal, Characterization of binary constraint system games. In *International colloquium on automata, languages, and programming*, pp. 320–331, Springer, 2014.

[13] A. Coladangelo, A two-player dimension witness based on embezzlement, and an elementary proof of the non-closure of the set of quantum correlations. *Quantum* **4** (2020), 282.

[14] A. Connes, Classification of injective factors. *Ann. of Math.* (1976), 73–115.

[15] M. De Chiffre, N. Ozawa, and A. Thom, Operator algebraic approach to inverse and stability theorems for amenable groups. *Mathematika* **65** (2019), no. 1, 98–118.

[16] K. Dykema and K. Juschenko, Matrices of unitary moments. *Math. Scand.* (2011), 225–239.

[17] K. Dykema, V. Paulsen, and J. Prakash, Non-closure of the set of quantum correlations via graphs. *Comm. Math. Phys.* **365** (2019), no. 3, 1125–1142.

[18] G. Elek and E. Szabó, Hyperlinearity, essentially free actions and l_2-invariants. the sofic property. *Math. Ann.* **332** (2005), no. 2, 421–441.

[19] J. Fitzsimons, Z. Ji, T. Vidick, and H. Yuen, Quantum proof systems for iterated exponential time, and beyond. In *Proceedings of the 51st annual ACM SIGACT symposium on theory of computing*, pp. 473–480, ACM, 2019.

[20] T. Fritz, Tsirelson's problem and Kirchberg's conjecture. *Rev. Math. Phys.* **24** (2012), no. 05, 1250012.

[21] V. Galliard, A. Tapp, and S. Wolf, The impossibility of pseudo-telepathy without quantum entanglement. 2002, arXiv:quant-ph/0211011.

[22] A. Goldberg, Synchronous linear constraint system games. *J. Math. Phys.* **62** (2021), no. 3, 032201.

[23] I. Goldbring and B. Hart, Computability and the Connes embedding problem. *Bull. Symbolic Logic* **22** (2016), no. 2, 238–248.

[24] W. T. Gowers and O. Hatami, Inverse and stability theorems for approximate representations of finite groups. *Sb. Math.* **208** (2017), no. 12, 1784.

[25] D. H. Hyers, On the stability of the linear functional equation. *Proc. Natl. Acad. Sci. USA* **27** (1941), no. 4, 222.

[26] Z. Ji, Compression of quantum multi-prover interactive proofs. In *Proceedings of the 49th annual ACM SIGACT symposium on theory of computing*, pp. 289–302, ACM, 2017.

[27] Z. Ji, A. Natarajan, T. Vidick, J. Wright, and H. Yuen, MIP* = RE. 2020, arXiv:2001.04383.

[28] Z. Ji, A. Natarajan, T. Vidick, J. Wright, and H. Yuen, Quantum soundness of the classical low individual degree test. 2020, arXiv:2009.12982.

[29] W. B. Johnson and J. Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert space. *Contemp. Math.* **26** (1984).

[30] M. Junge, M. Navascues, C. Palazuelos, D. Perez-Garcia, V. B. Scholz, and R. F. Werner, Connes' embedding problem and Tsirelson's problem. *J. Math. Phys.* **52** (2011), no. 1, 012102.

[31] D. Kazhdan, On ϵ-representations. *Israel J. Math.* **43** (1982), no. 4, 315–323.

[32] S.-J. Kim, V. Paulsen, and C. Schafhauser, A synchronous game for binary constraint systems. *J. Math. Phys.* **59** (2018), no. 3, 032201.

[33] E. Kirchberg, On non-semisplit extensions, tensor products and exactness of group C*-algebras. *Invent. Math.* **112** (1993), no. 1, 449–489.

[34] L. Mančinska and D. E. Roberson, Quantum isomorphism is equivalent to equality of homomorphism counts from planar graphs. In *2020 IEEE 61st annual symposium on foundations of computer science (FOCS)*, pp. 661–672, IEEE, 2020.

[35] D. Mayers and A. Yao, Self testing quantum apparatus. *Quantum Inf. Comput.* **4** (2004), no. 4, 273–286.

[36] D. Mermin, Simple unified form for the major no-hidden-variables theorems. *Phys. Rev. Lett.* **65** (1990), no. 27, 3373.

[37] H. Mousavi, S. S. Nezhadi, and H. Yuen, On the complexity of zero gap MIP. 2020, arXiv:2002.10490.

[38] B. Musto, D. Reutter, and D. Verdon, The Morita theory of quantum graph isomorphisms. *Comm. Math. Phys.* **365** (2019), no. 2, 797–845.

[39] A. Natarajan and T. Vidick, A quantum linearity test for robustly verifying entanglement. In *Proceedings of the 49th annual ACM SIGACT symposium on theory of computing*, pp. 1003–1015, ACM, 2017.

[40] A. Natarajan and J. Wright, NEEXP is contained in MIP*. In *2019 IEEE 60th annual symposium on foundations of computer science (FOCS)*, pp. 510–518, IEEE, 2019.

[41] M. Navascués, S. Pironio, and A. Acín, Bounding the set of quantum correlations. *Phys. Rev. Lett.* **98** (2007), no. 1, 010401.

[42] C. M. Ortiz and V. I. Paulsen, Quantum graph homomorphisms via operator systems. *Linear Algebra Appl.* **497** (2016), 23–43.

[43] N. Ozawa, About the Connes embedding conjecture. *Jpn. J. Math.* **8** (2013), no. 1, 147–183.

[44] V. I. Paulsen, S. Severini, D. Stahlke, I. G. Todorov, and A. Winter, Estimating quantum chromatic numbers. *J. Funct. Anal.* **270** (2016), no. 6, 2188–2222.

[45] V. I. Paulsen and I. G. Todorov, Quantum chromatic numbers via operator systems. *Q. J. Math.* **66** (2015), no. 2, 677–692.

[46] A. Peres, Incompatible results of quantum measurements. *Phys. Lett. A* **151** (1990), no. 3–4, 107–108.

[47] F. Rădulescu, The von Neumann algebra of the non-residually finite Baumslag group $\langle a, b | ab^3a^{-1} = b^2 \rangle$ embeds into $\mathbb{R}^\omega$. 2000, arXiv:math/0004172.

[48] F. Rădulescu, A non-commutative, analytic version of Hilbert's 17th problem in type II_1 von Neumann algebras. *Theta Ser. Adv. Math.* (2008), 93–101.

[49] R. Sasyk and A. Törnquist, The classification problem for von Neumann factors. *J. Funct. Anal.* **256** (2009), no. 8, 2710–2724.

[50] G. Scarpa and S. Severini, Kochen–Specker sets and the rank-1 quantum chromatic number. *IEEE Trans. Inf. Theory* **58** (2011), no. 4, 2524–2529.

[51] W. Slofstra, The set of quantum correlations is not closed. *Forum Math. Pi* **7** (2019).

[52] B. Tsirelson, Quantum generalizations of Bell's inequality. *Lett. Math. Phys.* **4** (1980), no. 2, 93–100.

[53] B. S. Tsirelson, Some results and problems on quantum Bell-type inequalities. *Hadronic J. Suppl.* **8** (1993), no. 4, 329–345.

[54] T. Vidick, *The complexity of entangled games*. Ph.D. thesis, UC Berkeley, 2011.

[55] T. Vidick, From operator algebras to complexity theory and back. *Not. Amer. Math. Soc.* **66** (2019), no. 10, 1618–1627.

[56] D. Voiculescu, The analogues of entropy and of Fisher's information measure in free probability theory, II. *Invent. Math.* **118** (1994), no. 1, 411–440.

[57] D. Voiculescu, Free entropy. *Bull. Lond. Math. Soc.* **34** (2002), no. 3, 257–278.

[58] J. von Neumann. *Mathematische Grundlagen der Quantenmechanik*. Springer, Berlin, 1932.

THOMAS VIDICK

California Institute of Technology, Pasadena, CA, USA, vidick@caltech.edu