Quantitative quantum soundness for all multipartite compiled nonlocal games

Matilde Baroni[1†], Igor Klep[2,3,4], Dominik Leichtle[5], Marc-Olivier Renou[6,7,8], Ivan Šupić[9], Lucas Tendick[6,7,8], Xiangling Xu[6,7,8*]

[1]Sorbonne Université, CNRS, LIP6, 4 place Jussieu, 75005 Paris, France
[2]Faculty of Mathematics and Physics, University of Ljubljana, Slovenia
[3]Faculty of Mathematics, Natural Sciences and Information Technologies, University of Primorska, Slovenia
[4]Institute of Mathematics, Physics and Mechanics, Ljubljana, Slovenia
[5]School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, United Kingdom
[6]Inria Paris-Saclay, Bâtiment Alan Turing, 1 rue Honoré d'Estienne d'Orves, 91120 Palaiseau, France
[7]CPHT, Ecole polytechnique, Institut Polytechnique de Paris, Route de Saclay, 91128 Palaiseau, France
[8]LIX, Ecole polytechnique, Institut Polytechnique de Paris, Route de Saclay, 91128 Palaiseau, France
[9]Université Grenoble Aples, CNRS, Grenoble INP, LIG, 38000 Grenoble, France

[†]matilde.baroni@lip6.fr

[*]xu.xiangling@inria.fr

Abstract

Compiled nonlocal games transfer the power of Bell-type multi-prover tests into a single-device setting by replacing spatial separation with cryptography. Concretely, the KLVY compiler (STOC'23) maps any multi-prover game to an interactive single-prover protocol, using quantum homomorphic encryption. A crucial security property of such compilers is quantum soundness, which ensures that a dishonest quantum prover cannot exceed the original game's quantum value. For practical cryptographic implementations, this soundness must be quantitative, providing concrete bounds, rather than merely asymptotic. While quantitative quantum soundness has been established for the KLVY compiler in the bipartite case, it has only been shown asymptotically for multipartite games. This is a significant gap, as multipartite nonlocality exhibits phenomena with no bipartite analogue, and the difficulty of enforcing space-like separation makes single-device compilation especially compelling. This work closes this gap by showing the *quantitative quantum soundness* of the KLVY compiler for *all multipartite* nonlocal games. On the way, we introduce an *NPA-like hierarchy for quantum instruments* and prove its completeness, thereby characterizing correlations from operationally-non-signaling sequential strategies. We further develop novel geometric arguments for the decomposition of sequential strategies into their signaling and non-signaling parts, which might be of independent interest.

1 Introduction

Nonlocal games are cooperative tasks involving multiple, non-communicating players (provers) and a referee (verifier), see Fig. 1(a), originally designed to test the foundational limits of classical physics in Bell's seminal work [Bel64]. In this setting, provers who share quantum entanglement can coordinate their answers to the verifier's questions in ways that are provably impossible using only classical resources. This "quantum advantage" makes nonlocal games powerful tools for the device-independent (i.e., black-box) certification of quantum properties, such as entanglement, which finds application, for instance, in the generation of genuine randomness [Bru+14; Sca12]. The security of these protocols, however, relies on a crucial assumption: that the provers cannot communicate. Typically, this is enforced by physical (space-like) separation, an approach that is experimentally demanding [Giu+15; Hen+15] and scales poorly, especially as the number of provers

increases [HR19]. This raises a fundamental question in both theory and practice: can we leverage the power of nonlocality within a protocol involving only a *single*, untrusted quantum device?

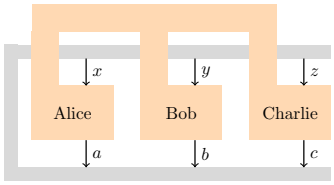
The naive approach of having one prover who plays all the roles fails heavily, as the prover would see all questions and trivially bypass the no-communication constraint. Cryptography provides an alternative solution. Kalai et al. [Kal+23] introduced a procedure (the KLVY compiler) to transform any k-player nonlocal game into a sequential, multi-round protocol between a verifier and a single prover, using quantum homomorphic encryption (QHE) [Bra18; Mah20]; see Fig. 1(c). Intuitively, the QHE allows the prover to compute on encrypted questions, preventing them from learning the questions for early players while generating a valid answer for a later player, up to negligible probability under standard cryptographic assumptions. The KLVY compiler [Kal+23] is known to be (i) *classically sound* (a classical prover cannot perform better than in the original game) and (ii) *quantum complete* (an honest quantum prover can achieve the optimal quantum score). The central, unresolved security question is that of *quantum soundness*: can a dishonest quantum prover exploit the compiled protocol to exceed the score achievable in the original, spatially-separated game?

Progress on this question has been incremental. Initial results established soundness for specific bipartite (two-player) games [Bar+24; Cui+24; MPW24; NZ23]. Subsequently, Kulpe et al. [Kul+25] proved asymptotic soundness for all bipartite games, showing that the compiled game's score approaches the true quantum score in the limit of $\lambda \to \infty$ for the cryptographic security parameter. While theoretically significant, this guarantee is insufficient for practical cryptography, which requires *quantitative* bounds for a finite, fixed security parameter λ. Furthermore, these bipartite frameworks did not extend to the multipartite setting, which provides a much richer structure of quantum advantages, leaving a crucial gap in our understanding.

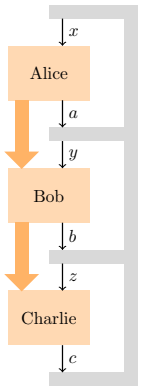
Two recent and independent works began to close these gaps. Klep et al. [Kle+25] established the first quantitative soundness bounds for all bipartite games by introducing a sequential variant of the Navascués-Pironio-Acín (NPA) hierarchy [NPA08; PNA10]—a powerful tool for bounding quantum correlations based on semidefinite programming. (See also [Cui+25] for a parallel independent work that approaches the same problem from the dual sums-of-squares perspective.) In parallel, Baroni et al. [Bar+25] proved asymptotic quantum soundness for all multipartite games, developing composable tools that fully generalize the algebraic approach of [Kul+25], setting up a strong mathematical foundation for future work.

However, a unified solution for the general multipartite case remains challenging, as both previous results use different (a priori not compatible) techniques, leaving the question of whether quantitative quantum soundness for all multipartite games can be achieved widely open. More precisely, the sequential NPA hierarchy of [Kle+25] is tailored to two-player sequential games and does not naturally accommodate the complex algebraic structure of multipartite interactions of [Bar+25]. Moreover, its core analytical tool—a signaling vs. non-signaling decomposition—is intrinsically bipartite and does not generalize. This is a significant obstacle, as multipartite nonlocality exhibits richer quantum phenomena than its bipartite counterpart, and the practical difficulty of enforcing physical separation among many parties makes a robust cryptographic compilation especially desirable.

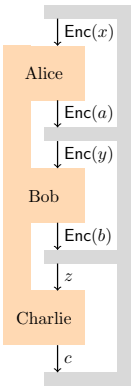
This work overcomes the above obstacles. That is, we answer the question of whether the quantitative quantum soundness for all multipartite compiled nonlocal games can be achieved in the affirmative. Our main tool is a novel and composable NPA-like hierarchy designed to model the sequential application of quantum instruments. This framework, combined with new geometric proof techniques for the weak signaling decomposition, is rich enough to capture general multipartite scenarios in the compiled setting and may be of independent interest for the study of complex quantum protocols.



(a) Three spatially separated provers, Alice A, Bob B, and Charlie C, receive questions x, y, z and return answers a, b, c. The players' shared strategy is defined by the correlations $p(abc|xyz)$ with the corresponding score (winning probability) given by $\sum_{a,b,c,x,y,z} \beta_{abcxyz} p(abc|xyz)$, where β_{abcxyz} is the payoff tensor associated with the rule of the game $\mathcal{G}$. By using a shared entangled state, the players can generate quantum correlations leading to scores that provably exceed classical limitations. The maximum achievable quantum score is denoted by $\omega_{\mathrm{q}}(\mathcal{G})$ (the tensor-product quantum score).



(b) The game is played in sequence: Alice acts first with (x, a), then Bob with (y, b), then Charlie with (z, c). In the Heisenberg algebraic picture, a state σ is fixed. Actions of A, B are described by quantum instruments (completely positive maps) $\{\mathsf{T}_{a|x}\}_a$ and $\{\mathsf{T}_{b|y}\}_b$, while C measures with POVM effects $\{f_{c|z}\}_c$ as usual. The resulting correlations are $p(abc|xyz) = \mathrm{Tr}\big(\sigma \cdot \mathsf{T}_{a|x} \circ \mathsf{T}_{b|y}(f_{c|z})\big)$. *Operational-non-signaling* requires $\sum_a \mathsf{T}_{a|x} = \sum_a \mathsf{T}_{a|x'}$ for all x, x' and $\sum_b \mathsf{T}_{b|y} = \sum_b \mathsf{T}_{b|y'}$ for all y, y'.



(c) A single prover P plays all roles sequentially. The pairs (x, a) and (y, b) are sent and returned in encrypted form $\mathrm{Enc}(x), \mathrm{Enc}(a)$ and $\mathrm{Enc}(y), \mathrm{Enc}(b)$ (the prover computes on the encrypted data homomorphically), while (z, c) is sent in the clear. Security of the QHE scheme enforces computational non-signaling between the A-, B-, and C-interfaces.

Figure 1: Nonlocal game variants: (a) standard, (b) sequential, (c) compiled. Time flows from top to bottom.

1.1 Main results

Our first main result is a quantitative quantum soundness theorem for all multipartite compiled nonlocal games, providing bounds for finite security levels. For a given security parameter λ, we define an *efficient* prover as one implementable in quantum-polynomial-time (QPT), i.e., by a quantum circuit of size $\mathrm{poly}(\lambda)$, and we call a function *negligible*, denoted $\mathrm{negl}(\lambda)$, if it vanishes faster than the reciprocal of any polynomial in λ.

Theorem A (Theorem 4.4)**.** *Let $k, \lambda \in \mathbb{N}$, let $\mathcal{G}$ be a k-partite nonlocal game with optimal commuting-operator quantum score $\omega_{\mathrm{qc}}(\mathcal{G})$, and let $\mathcal{G}_{\mathrm{comp}}$ be its compiled version. Let $S = (S_\lambda)_\lambda$ be any strategy employed by an efficient prover, and denote by $\omega_\lambda(\mathcal{G}_{\mathrm{comp}}, S)$ its compiled Bell score.*

If $\mathcal{G}$ admits a finite-dimensional optimal quantum strategy (hence $\omega_{\mathrm{qc}}(\mathcal{G}) = \omega_{\mathrm{q}}(\mathcal{G})$, where $\omega_{\mathrm{q}}(\mathcal{G})$ is the optimal tensor-product quantum score), then there exists a negligible function $\mathrm{negl}_S(\lambda)$ (depending on the QHE scheme and on S) such that

$$\omega_\lambda(\mathcal{G}_{\mathrm{comp}}, S) \leq \omega_{\mathrm{q}}(\mathcal{G}) + \mathrm{negl}_S(\lambda). \tag{1}$$

More generally, for every k-partite nonlocal game $\mathcal{G}$ and for every $n \in \mathbb{N}$, there exists a negligible function $\mathrm{negl}_{S,n}(\lambda)$ (depending on the QHE scheme, the strategy S, and n) such that

$$\omega_\lambda(\mathcal{G}_{\mathrm{comp}}, S) \leq \omega^n_{k\mathrm{seqNPA}}(\mathcal{G}) + \mathrm{negl}_{S,n}(\lambda), \tag{2}$$

where $\omega^n_{k\text{seqNPA}}(\mathcal{G})$ *is the level-*n *value of the* k*-partite sequential NPA hierarchy (see Eq.* (10) *for the full definition), such that* $\omega^n_{k\text{seqNPA}}(\mathcal{G}) \searrow \omega_{\text{qc}}(\mathcal{G})$ *as* $n \to \infty$*.*

This theorem recovers the quantitative bipartite result of [Kle+25] for $k = 2$. In the limit where $n, \lambda \to \infty$, it reproduces the asymptotic multipartite result of [Bar+25]. Thus, Theorem A unifies all prior work, delivering both quantitative guarantees for finite-dimensional games and asymptotic soundness in the general case, definitively settling the problem of quantum soundness for KLVY compilers.

The central technical tool enabling this result is a novel sequential NPA hierarchy, for which we show completeness, strict feasibility, and a stopping criterion for finite convergence.

Theorem B (Informal, Theorems 3.6, 3.8 and 3.9)**.** *For any* k*, the* k*-partite sequential NPA hierarchy (Eq.* (10)*) is strictly feasible, and is complete with respect to* k*-partite commuting-observable strategies (and thus* k*-partite operationally-non-signaling sequential strategies), i.e., the sequence of its finite-level values satisfies*

$$\omega^n_{k\text{seqNPA}}(\mathcal{G}) \searrow \omega_{\text{qc}}(\mathcal{G}) \quad \text{as } n \to \infty.$$

Moreover, a game $\mathcal{G}$ *admits a flat optimal solution to Eq.* (10) *if and only if it admits a finite-dimensional optimal quantum strategy.*

1.2 Methods and techniques

As in Theorem A, our primary goal is to characterize and compute upper bounds on the scores of compiled nonlocal games. Recent works [Bar+25; Kul+25] established a crucial connection: in the asymptotic limit of perfect cryptographic security, compiled strategies (Fig. 1(c)) are equivalent to sequential strategies (Fig. 1(b)) where players are constrained to be operationally-non-signaling—meaning their quantum output, averaged over all classical outcomes, reveals no information about their classical input. This provides a path to bounding compiled scores by analyzing their sequential counterparts.

However, for any real-world implementation with a finite security parameter, this correspondence is not exact, and a quantitative analysis is required. The framework of [Kle+25] provided the first quantitative bounds for bipartite nonlocal games as follows: they introduce a sequential NPA hierarchy that is complete to the bipartite quantum scenarios. Then, they show that any efficient strategy for a compiled game is necessarily "close" to a feasible solution of this hierarchy, using a signaling decomposition argument. By proving this, the score of the compiled strategy becomes bounded by the value of the sequential hierarchy, thus proving soundness due to its convergence.

This bipartite solution, however, faces two fundamental obstacles that prevent its generalization to multipartite cases:

1. *An unextendable hierarchy:* The bipartite hierarchy in [Kle+25] is built on the Schrödinger picture, modeling the post-measurement state passed between players. While sufficient for two players, this approach loses critical algebraic information and is known to fail multipartite scenarios [Sai+15]. Multipartite sequential protocols require a more robust model that can handle a chain of quantum instruments (CP maps).

2. *A non-scalable proof technique:* The signaling decomposition argument in [Kle+25] relies on a Gelfand-Naimark-Segal (GNS) representation of the underlying algebra. This argument is intrinsically bipartite and does not generalize to the multipartite setting, necessitating a completely new approach.

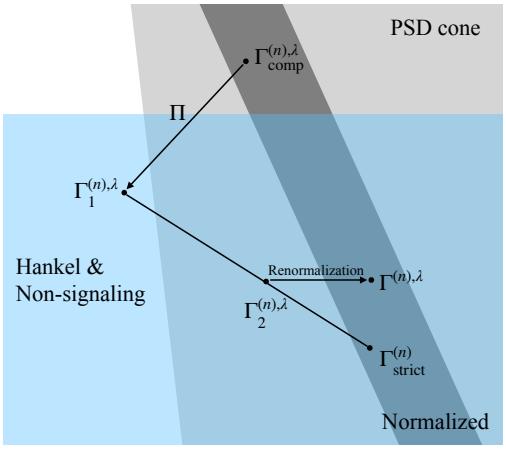


Figure 2: Geometric sketch of proof for Theorem A, detailed in Section 4.2. The gray region is the PSD cone; the blue slice encodes the Hankel condition and operationally-non-signaling constraints of our k-partite sequential NPA hierarchy Eq. (10). The thick dark line indicates normalized moment matrices. From a compiled strategy we extract $\Gamma^{(n),\lambda}_{\text{comp}}$. Applying the projector Π (constructed based on Eq. (12)) yields $\Gamma^{(n),\lambda}_{1}$ that satisfies the affine constraints but may fail PSD and normalization. Positivity is restored by convexly mixing with a strictly feasible point $\Gamma^{(n)}_{\text{strict}}$, giving $\Gamma^{(n),\lambda}_{2}$, which is then rescaled to a normalized $\Gamma^{(n),\lambda}$ that is a feasible solution to our multipartite NPA hierarchy.

This paper overcomes both obstacles by adopting the same high-level approach while introducing entirely new, scalable tools. First, to replace the unextendable bipartite model, we introduce a composable multipartite sequential NPA hierarchy based on the Heisenberg picture, which is guaranteed to converge for any number of players. Second, to replace the non-scalable proof technique, we develop a novel, concise, and scalable geometric proof illustrated by Fig. 2. This new argument establishes the crucial closeness result between a near-solution from a compiled strategy and a true feasible solution in our hierarchy. Together, these contributions prove quantitative soundness for compiled games with any number of players. We now detail our techniques in two parts.

A sequential NPA hierarchy for quantum instruments. The Navascués-Pironio-Acín (NPA) hierarchy [NPA08; PNA10], a noncommutative generalization of the Lasserre-Parrilo hierarchy [Las01; Par03], is a cornerstone for analyzing nonlocal games. It provides a sequence of increasingly tight semidefinite programming (SDP) relaxations that systematically compute upper bounds on the optimal quantum score. However, the standard NPA hierarchy is designed for the standard Bell scenario (Fig. 1(a)), where all players' actions are modeled as terminal measurement POVMs that commute. Specifically, consider a game $\mathcal{G}$ with questions $x, y, z, \ldots$ and answers $a, b, c \ldots$, the POVMs of all players are modeled with commuting letters $\{f_{a|x}\}, \{f_{b|y}\}, \{f_{c|z}\}$, etc. These letters form words that index a moment matrix, which encodes the expectation values of all such operator words with respect to the underlying quantum states and thus characterizes the commuting observable quantum strategies. However, it cannot describe sequential protocols (Fig. 1(b)) or their compiled counterparts (Fig. 1(c)) that are central to our work, where the actions of earlier players are quantum instruments that transform the state for subsequent players.

This necessitates a fundamental generalization of the NPA framework for the sequential settings. To this end, we introduce a composable, multipartite sequential NPA hierarchy (Eq. (10)) that is complete for both the quantum sequential and nonlocal scenarios (Theorem B). We now sketch its construction, focusing on the tripartite case for demonstration purposes.

We use notations of Fig. 1(b), based on the Heisenberg algebraic picture of a sequential game, dual to the standard Schrödinger's picture in which maps act on states. More precisely, the quantum instruments modeling Alice's and Bob's actions do not act on the state σ which is fixed, but on Charlie's measurement operators (see Eq. (3) below). Alice's and Bob's actions are described by completely positive (CP) maps (quantum instruments) $\{\mathsf{T}_{a|x}\}_a$ and $\{\mathsf{T}_{b|y}\}_b$, while the final player, Charlie, performs a standard measurement with POVM effects $\{f_{c|z}\}_c$. With these objects, the corresponding correlations are

$$p(abc|xyz) = \mathrm{Tr}\big(\sigma \cdot \mathsf{T}_{a|x} \circ \mathsf{T}_{b|y}(f_{c|z})\big),$$

achieving the score

$$\sum_{a,b,x,y} \beta_{abcxyz} p(abc|xyz),$$

where β_{abcxyz} is the payoff tensor associated with the rule of the game $\mathcal{G}$. Diagrammatically, the sequentiality can be represented by

$$f_{c|z} \text{ of } C \xrightarrow{\mathsf{T}_{b|y}} \mathsf{T}_{b|y}(f_{c|z}) \text{ of } B \xrightarrow{\mathsf{T}_{a|x}} \mathsf{T}_{a|x} \circ \mathsf{T}_{b|y}(f_{c|z}) \text{ of } A. \tag{3}$$

The final pieces of the sequential scenarios are the *operationally-non-signaling constraints* of Alice's and Bob's actions, which are, respectively,

$$\sum_a \mathsf{T}_{a|x} = \sum_a \mathsf{T}_{a|x'}, \quad \sum_b \mathsf{T}_{b|y} = \sum_b \mathsf{T}_{b|y'},$$

for all x, x', y, y'.

Our sequential variant must therefore algebraically encode both the ordered action of these instruments and their CP-map-level operationally-non-signaling constraints. The core of our solution is a level-n moment matrix, $\Gamma^{(n)}$, whose entries correspond to the expectation values of different operators with respect to the shared quantum state σ. That is, for any two operator words w and v, the corresponding matrix entry is:

$$\Gamma^{(n)}_{w,v} = L^{2n}(w^* v) = \mathrm{Tr}(\sigma \cdot w^* v),$$

where L^{2n} is the associated linear map modeling $\mathrm{Tr}(\sigma\,\cdot)$. The fundamental challenge is to (1) define a set of operator words that captures the sequential structure of the game and (2) implement the appropriate constraints on $\Gamma^{(n)}$ to model the sequential scenarios.

For (1), the solution is a sequential construction of the "words" (operator monomials) that index our $\Gamma^{(n)}$, which we start from the last player as in Eq. (3):

(a) *The final player Charlie*: We begin with Charlie, whose action is a standard POVM. Similarly to the standard NPA hierarchy, we describe his measurements with a set of operator letters $\{f_{c|z}\}$. These letters form words (monomials) such as $u = f_{c_1|z_1} f_{c_2|z_2} \cdots f_{c_m|z_m}$, which have a degree m, denoted by $\deg u = m$. The set of all such words with $\deg u \leq n$ forms our base word set, $\mathcal{W}^n_C$.

(b) *The preceding player Bob*: To model Bob's quantum instrument $\{\mathsf{T}_{b|y}\}$, we introduce a set of $*$-homomorphisms $\{T_{b|y}\}$. (This representation is justified by the Stinespring dilation theorem, which guarantees that any quantum instrument (a CP map) can be "lifted" to a $*$-homomorphism on a larger space.) These are maps that act on Charlie's operators to generate

new letters, $f_{bc|yz} := T_{b|y}(f_{c|z})$, which algebraically capture the effect of Bob's instrument on the subsequent system. The set of all words formed from these new letters up to degree n is denoted $\mathcal{W}^n_{BC}$.

(c) *The first player Alice*: We proceed sequentially for Alice, defining her corresponding Stinespring dilated $*$-homomorphisms $\{T_{a|x}\}$ which act on the words of Bob and Charlie. This generates the final set of letters $f_{abc|xyz} = T_{a|x}(f_{bc|yz})$ and the degree $\leq n$ word set $\mathcal{W}^n_{ABC}$. This layered construction directly models the sequentiality of the protocol analogous to Eq. (3):

$$f_{c|z}\ \text{(Charlie)} \xrightarrow{T_{b|y}} f_{bc|yz}\ \text{(Bob)} \xrightarrow{T_{a|x}} f_{abc|xyz}\ \text{(Alice)}.$$

(d) *The moment matrix*: This sequential process defines the full set of words $\mathcal{W}^n_{ABC}$ that index our moment matrix $\Gamma^{(n)}$. The entry $\Gamma^{(n)}_{\mathbb{1},f_{abc|xyz}}$ represents the correlation $p(abc|xyz)$, and consequently corresponding score in game $\mathcal{G}$ with the payoff tensor β_{abcxyz} is

$$\sum_{a,b,c,x,y,z} \beta_{abcxyz}\Gamma^{(n)}_{\mathbb{1},f_{abc|xyz}}.$$

With the indices of $\Gamma^{(n)}$ sequentially defined, we now address (2) by identifying the appropriate constraints to recover a sequential quantum strategy as the NPA level $n \to \infty$.

(A) *Standard constraints*: Since $\Gamma^{(n)}$ and L^{2n} are to model the expectation map $\mathrm{Tr}(\sigma\,\cdot)$, this physical interpretation leads directly to three fundamental constraints on $\Gamma^{(n)}$. First, for L^{2n} to be a well-defined linear map on the word sets, the matrix $\Gamma^{(n)}$ must be symmetric, and satisfy a Hankel-like condition. Second, since $\mathrm{Tr}(\sigma\,\cdot)$ is a positive map for any state σ, the functional L^{2n} must also be positive. This is precisely equivalent to the constraint that $\Gamma^{(n)}$ must be positive semidefinite (PSD). Third, the state normalization $\mathrm{Tr}(\sigma) = 1$ corresponds to $L^{2n}(\mathbb{1}) = 1$, which implies the matrix normalization condition $\Gamma^{(n)}_{\mathbb{1},\mathbb{1}} = 1$.

(B) *Alice's operationally-non-signaling constraint*: A crucial innovation of our hierarchy are *CP map level constraints* that formalize the notion of operationally-non-signaling, meaning that the marginal maps $\sum_a T_{a|x}$ and $\sum_a T_{a|x'}$ are the same if only tested against polynomials up to degree n; as n goes to infinity, perfect operationally-non-signaling constraints are retrieved. The constraints for Alice's instrument read:

$$L^{2n}\Big(w^* \sum_a \big(T_{a|x}(r^*s) - T_{a|x'}(r^*s)\big)v\Big) = 0$$

for all x, x', and for all words r, s built from $\{f_{bc|yz}\}$ and w, v from $\{f_{abc|xyz}\}$ up to a total degree $\leq 2n$. Here, the central term $\sum_a \big(T_{a|x}(r^*s) - T_{a|x'}(r^*s)\big)$ tests Alice's non-signaling condition on an arbitrary operator r^*s from Bob's algebra. This test is then embedded within the context of arbitrary actions from the full sequential strategy that Alice can perform (words w, v), ensuring the condition holds universally.

(C) The constraint for Bob's instrument is more involved, but demonstrates the composable structure of our method:

$$L^{2n}\Big(T_{a|x}\Big(r^* \sum_b \big(T_{b|y}(t^*u) - T_{b|y'}(t^*u)\big)s\Big)\Big) = 0,$$

for all a, x, y, y', and for all words r, s built from $\{f_{bc|yz}\}$ and t, u from $\{f_{c|z}\}$ up to a total degree $\leq 2n$. Similarly, the central term, $\sum_b \big(T_{b|y}(t^*u) - T_{b|y'}(t^*u)\big)$, tests Bob's non-signaling condition on an arbitrary operator t^*u from Charlie's algebra. This core expression is embedded within the context of arbitrary operators from Bob's layer (words r, s) and placed under every possible action by Alice ($T_{a|x}$, $\forall a, x$). In this way, the SDP enforces Bob's non-signaling constraints across every possible algebraic context created by the preceding players.

Following (a)-(d) and (A)-(C), we reach the subsequent definition of the tripartite sequential NPA hierarchy that maximizes the score achievable by all sequential quantum strategies (Eq. (7)):

$$
\begin{aligned}
\omega^n_{3\text{seqNPA}}(\mathcal{G}) \;:=\; \max \quad & \sum_{a,b,c,x,y,z} \beta_{abcxyz}\, \Gamma^{(n)}_{\mathbb{1}, f_{abc|xyz}} \\
\text{s.t.} \quad & \Gamma^{(n)} \succeq 0, \quad \Gamma^{(n)}_{\mathbb{1},\mathbb{1}} = 1, \quad \text{(PSD and normalization)} \\
& \Gamma^{(n)}_{w,w'} = \Gamma^{(n)}_{v,v'} \quad \text{whenever } w^*w' = v^*v', \quad \text{(Hankel condition)} \\
& L^{2n}\Big(w^* \sum_a \big(T_{a|x}(r^*s) - T_{a|x'}(r^*s)\big)v\Big) \;= 0 \\
& \qquad\qquad \substack{\forall x,x',\, \forall w,v \in \mathcal{W}^n_{ABC},\\ r,s \in \mathcal{W}^n_{BC},\\ \deg w + \deg v + \deg r + \deg s \leq 2n} \quad \text{(Alice operationally-non-signaling)} \\
& L^{2n}\Big(T_{a|x}\Big(r^* \sum_b \big(T_{b|y}(t^*u) - T_{b|y'}(t^*u)\big)s\Big)\Big) \;= 0 \\
& \qquad\qquad \substack{\forall\, a,x,y,y',\\ r,s \in \mathcal{W}^n_{BC},\\ t,u \in \mathcal{W}^n_{C},\\ \deg r + \deg s + \deg t + \deg u \leq 2n} \quad \text{(Bob operationally-non-signaling).}
\end{aligned}
\tag{4}
$$

As showcased in the sequential construction (a)-(d), it is straightforward to obtain the word sets to any k-partite scenarios. Analogously, the operationally-non-signaling constraints in (B) and (C) can be generalized to k-partite scenarios in an inductive way. Indeed, by induction, we formulate in Eq. (10) the general k-partite sequential NPA hierarchy. As stated in Theorem B, this hierarchy is complete to both the standard multipartite quantum strategies (Fig. 1(a)) and the multipartite quantum sequential strategies (Fig. 1(b)).

In contrast to the bipartite hierarchy of [Kle+25] formulated in the Schrödinger picture, our novel sequential construction operates in the Heisenberg picture. Instead of tracking the evolving state, we model the transformations themselves using $*$-homomorphisms. This preserves the complete algebraic structure of each quantum instrument in the sequence. This perspective provides a clean and rigorous way to encode the CP-map-level non-signaling constraints, leading to a hierarchy that is powerful enough to converge for any number of players (as shown in Theorem B). We expect this technique to be of independent interest for the computational analysis of multi-round quantum protocols and in fields where CP maps are fundamental [Pau02; Rag03].

A scalable geometric argument to decompose compiled correlations. To prove the quantitative soundness for compiled games using our new hierarchy, we must connect the two. The high-level strategy, mirroring that of [Kle+25], is to show that any efficient compiled strategy is "close" to a genuinely feasible solution within our multipartite sequential NPA hierarchy. However, as previously discussed, the representation-based signaling decomposition argument used in [Kle+25] is intrinsically bipartite and fails here.

We therefore introduce a novel, scalable geometric proof that achieves this closeness result (illustrated in Fig. 2 and fully presented in Section 4). The core idea is as follows:

1. As a consequence of [Bar+25], an efficient compiled strategy, due to finite cryptographic security λ, produces a moment matrix $\Gamma^{(n),\lambda}_{\text{comp}}$ that acts as a "pseudo-solution" for our hierarchy. It is positive and normalized, but it weakly violates the operationally-non-signaling constraints. Based on $\Gamma^{(n),\lambda}_{\text{comp}}$, our proof then geometrically constructs a nearby, genuinely feasible solution.

2. We first project $\Gamma^{(n),\lambda}_{\text{comp}}$ onto the affine subspace of matrices that perfectly satisfy the non-signaling and Hankel constraints. However, this projected matrix might no longer be PSD and normalized.

3. We then restore positivity by taking a slight convex combination with a known, strictly feasible solution $\Gamma^{(n)}_{\text{strict}}$ (Theorem 3.9). Finally, we re-normalize the result to obtain a genuine, feasible solution $\Gamma^{(n),\lambda}$ that satisfies all constraints of our hierarchy.

4. We prove that $\Gamma^{(n),\lambda}$ is negligibly close to the original $\Gamma^{(n),\lambda}_{\text{comp}}$ in operator norm (Theorem 4.3). The quantitative soundness theorem is then a direct corollary of this closeness.

The key advantage of this geometric approach is its scalability. It provides a powerful template for establishing quantitative control in complex quantum scenarios where representation-based algebraic arguments are insufficient.

1.3 Further discussions

Our results provide a definitive resolution to the question of quantitative quantum soundness for all compiled nonlocal games and introduce a powerful new hierarchy for analyzing sequential quantum protocols. We briefly discuss the implications of both contributions.

Resolving quantum soundness for multipartite games. Our main result (Theorem A) establishes the quantitative soundness of multipartite compiled games, elevating them to a provably secure framework. This is a critical step because many unique quantum phenomena, such as the GHZ paradox and genuine multipartite nonlocality, have no bipartite analogue. Our work makes it possible to explore and certify these genuinely multipartite phenomena, such as self-testing multipartite states, within the practical single-device model, opening a new perspectives for both theoretical and experimental research.

With quantitative soundness established, a natural next step is to optimize the resulting protocols for efficiency. Compiling a game with k players produces a protocol consisting of k rounds. An interesting open question is whether the number of rounds can be reduced without jeopardizing observable classical-quantum and quantum-post quantum separations. In particular, for strategies that give rise to correlations violating Svetlichny-type inequalities (genuine multipartite nonlocality) [Sve87], it is plausible that merging two rounds into a single one might still yield an observable classical–quantum separation. Furthermore, adapting techniques used in self-testing allowing communication among certain parties [Mey+24; MB23] might lead to retain quantum soundness of compiled games with a reduced number of rounds. Exploring such round-compression techniques could lead to more efficient compiled protocols.

Future directions for the multipartite sequential NPA hierarchy. Beyond its immediate application in this work, our multipartite sequential NPA hierarchy for quantum instruments is a technical contribution that could be of independent interest. It provides the first systematic, convergent, and

composable tool for computing bounds on the capabilities of sequential quantum strategies. This opens up several avenues for future research:

1. *Numerical performance and convergence.* Our sequential NPA hierarchy involves smaller moment matrices at each level compared to the standard NPA hierarchy, but incorporates more complex constraints. Understanding this trade-off between the rate of convergence and the moment matrix size can be interesting both theoretically and important for practical numerical implementation. Such insight can have further implications to related SDP hierarchies, such as the sparse SOS hierarchies [KMP22; MW23] and the bipartite sequential NPA hierarchies [Kle+25].

2. *Extending device-independent certification.* The standard NPA hierarchy is often the foundation for the device-independent certification of properties like randomness and entanglement. Our hierarchy extends this capability to a broader class of sequential and multi-round protocols, enabling the certification of tasks where players act one after another.

3. *Beyond operationally-non-signaling constraints.* The algebraic nature of our framework is not limited to non-signaling. It can be adapted to enforce other constraints on quantum instruments to describe different scenarios. For instance, one could require a specific instrument outcome ($a = 0$) to correspond to an identity channel, a common feature in error correction or gate-based protocols. Moreover, it can be adapted to model new constraints due to cryptographic primitives beyond the homomorphic encryption.

4. *Analyzing quantum interactive proofs.* Our current work models a linear sequence of players. However, its composable, algebraic nature suggests it could be extended to analyze protocols with more complex interaction structures. It could provide a framework for understanding more general interactive quantum protocols (where a quantum prover with memory interacts with a verifier over multiple rounds), a central scenario in quantum complexity theory and quantum cryptography.

In summary, by providing both a complete solution to the soundness problem and a novel analytical tool, our work strengthens the foundations of single-device quantum cryptography and offers new capabilities for the broader study of quantum information protocols.

1.4 Outline of the manuscript

The remainder of this paper is organized as follows. Section 2 presents necessary preliminaries: We review the necessary background, defining nonlocal and compiled games (Section 2.1), and the standard NPA hierarchy (Section 2.2). Section 3 then presents the novel multipartite sequential NPA hierarchy, where we introduce the construction progressively: Section 3.1 first revisits the bipartite case to provide a proof of concept and contrast our Heisenberg-picture approach with the Schrödinger-picture method of [Kle+25]. Section 3.2 then uses the tripartite case as a pedagogical bridge to the general construction. Section 3.3 presents the full, general k-partite hierarchy and proves its key properties—completeness, the flatness condition, and strict feasibility. Then, Section 4 works up to the main quantitative quantum soundness result: We leverage our new hierarchy to prove the main result of this paper: quantitative quantum soundness for all multipartite compiled nonlocal games (Theorem A). After establishing notation in Section 4.1, Section 4.2 develops the novel geometric proof, detailing the projection, regularization, and normalization argument that overcomes the limitations of prior techniques. The manuscript finishes with Section A on a brief comparison on [Bar+25; Kle+25].

2 Preliminaries

We introduce the notations of nonlocal games and compiled games in Section 2.1, followed by a brief introduction of the NPA hierarchy in Section 2.2.

2.1 Nonlocal games and compiled games

A nonlocal game $\mathcal{G}$ is an interaction involving a referee and multiple players (provers) who cannot communicate with each other during the game; see Fig. 1(a) for a tripartite example. The referee sends each player a question x_i (drawn from some specified distribution $\mu(\vec{x})$), and each player must respond with an answer a_i. Whether the players win is decided by a publicly known rule (a predicate depending on all questions and answers) $V(\vec{a},\vec{x})$. No communication means each player must base their answer only on their own question and a prior agreed-upon strategy among the players, but not on the other players' questions or answers. The players' goal is to maximize their winning probability $\sum_{\vec{x}\vec{a}} \mu(\vec{x})V(\vec{a},\vec{x})p(\vec{a}|\vec{x})$ by coordinating a strategy ahead of time (they know the game's description in advance).

The *classical value* (or score) of a nonlocal game is the maximum success probability achievable when the players share only classical resources, e.g., a pre-shared random string (common randomness) or any predetermined classical strategy. Denoting with f_i an arbitrary deterministic function mapping input x_i to output a_i, the classical value takes the form

$$\omega_c(\mathcal{G}) = \max_{f_1,\cdots,f_k} \sum_{\vec{x}} \mu(\vec{x})V(x_1,\cdots,x_k,f_1(x_1),\cdots,f_k(x_k)).$$

In contrast, the *quantum value* (or score) is the maximum winning probability when players can use quantum resources such as shared entangled states ρ and quantum measurements $M_{a_i|x_i}$:

$$\omega_q(\mathcal{G}) = \sup_{\rho,M_{a_1|x_1},\cdots,M_{a_k|x_k}} \sum_{\vec{a},\vec{x}} \mu(\vec{x})V(\vec{x},\vec{a}) \operatorname{Tr}\left(\bigotimes_{i=1}^{k} M_{a_i|x_i}\rho\right).$$

Quantum strategies can outperform classical ones in certain games, a phenomenon known as a nonlocal *quantum advantage*. For example, in the famous CHSH game [Cla+69], the best classical strategy wins with probability 0.75, whereas players sharing an entangled qubit pair can win with probability about 0.85. This higher success rate ($\approx 85\%$ vs 75%) is a quantum advantage that cannot be achieved by any classical means. Such nonlocal games (also called Bell games in physics) thus highlight the stronger-than-classical correlations allowed by entanglement.

It is worth noting a subtle point about the definition of "quantum value". In finite-dimensional quantum systems, insisting that players cannot communicate is equivalent to requiring that they act on separate Hilbert spaces (tensor factors) of a joint entangled state. More generally, one can impose that all of one player's measurement operators commute with all of the other player's operators, which is a formal way to enforce a non-signaling condition in possibly infinite-dimensional or arbitrary systems. While the tensor-product and commuting definitions coincide for finite dimensions [SW08], a result now known to be robust [Oza13; XRK25], they differ in general [Ji+22]. This leads to the definition of the *quantum commuting operator value* of a game, which is the supremum win probability attained by any strategy where the players' operations commute:

$$\omega_{\mathrm{qc}}(\mathcal{G}) = \sup_{\rho,N_{a_1|x_1},\cdots,N_{a_k|x_k}} \sum_{\vec{a},\vec{x}} \mu(\vec{x})V(\vec{x},\vec{a}) \operatorname{Tr}\left(\prod_{i=1}^{k} N_{a_i|x_i}\rho\right).$$

While nonlocal games traditionally require multiple spatially isolated devices, the recent line of work initiated by [Kal+23] shows that, under computational assumptions, one can compile a k-prover game $\mathcal{G}$ to produce a new protocol between a single QPT prover and a classical verifier, which we call a compiled nonlocal game. Intuitively, the compilation procedure simulates space-like separation using cryptography and a specific sequential structure as in Fig. 1(c): the verifier sends an encrypted question, and always wait for an encrypted answer before sending the following question; the last question-answer round doesn't need to be encrypted.

Definition 2.1 (Compiled k-partite nonlocal game)**.** *Let $\mathcal{G}$ be a k-partite nonlocal game with input alphabets X_i, answer alphabets A_i, input distribution μ on $X := X_1 \times \cdots \times X_k$, and predicate $V(\vec{x}, \vec{a})$, where $\vec{x} = (x_1, \ldots, x_k)$ and $\vec{a} = (a_1, \ldots, a_k)$. Fix a quantum homomorphic encryption scheme $\mathsf{QHE} = (\mathrm{Gen}, \mathrm{Enc}, \mathrm{Eval}, \mathrm{Dec})$.*

For a security parameter λ, the KLVY compilation outputs a sequential protocol $\mathcal{G}^{\lambda}_{\mathrm{comp}}$ with k rounds. Define by $\mathcal{G}_{\mathrm{comp}} := (\mathcal{G}^{\lambda}_{\mathrm{comp}})_{\lambda}$ the compiled nonlocal game *of $\mathcal{G}$. At the beginning of the protocol, the verifier runs $\mathrm{Gen}(1^{\lambda})$ to generate a secret key sk for QHE, and samples inputs $\vec{x}$ from the distribution μ. In rounds $i \in \{1, \ldots, k-1\}$, the verifier then uses their key to compute and send encryptions $\mathrm{Enc}^{\mathsf{sk}}(x_i)$ and receives as responses a_i; in the last round k the verifier sends x_k in the clear and receives a_k. The verifier decrypts the responses to obtain $a_i := \mathrm{Dec}^{\mathsf{sk}}(\mathrm{a}_i)$ for all $i \in \{1, \ldots, k-1\}$, forming the transcript $(\vec{x}, \vec{a})$, which is evaluated using the same predicate V as in $\mathcal{G}$.*

Let $S = (S_\lambda)_\lambda$ be any efficient prover strategy for $\mathcal{G}_{\mathrm{comp}}$ (i.e., S_λ implementable by quantum circuits of size $\mathrm{poly}(\lambda)$). Denote by $p^{\lambda}(\vec{a}|\vec{x})$ the correlation realized by S_λ in $\mathcal{G}^{\lambda}_{\mathrm{comp}}$ at security parameter λ. The compiled Bell score *of S is*

$$\omega_\lambda(\mathcal{G}_{\mathrm{comp}}, S) := \sum_{\vec{x},\vec{a}} \beta_{\vec{x},\vec{a}}\, p^{\lambda}(\vec{a}|\vec{x}),$$

where we write the payoff tensor $\beta_{\vec{x},\vec{a}} := \mu(\vec{x})V(\vec{x},\vec{a})$ to simplify the notation for the remainder of this manuscript.

Despite the added difficulty of encryption, the compiled game is designed so that an honest quantum prover can still play optimally. In fact, KLVY [Kal+23] proved two key properties of their compiler, *classical soundness* and *quantum completeness*, for all k-partite games. Classical soundness refers to the fact that any efficient classical prover (one who does not use entanglement or quantum memory) with strategy $S_{\mathrm{classical}}$ cannot win the compiled game with probability higher than the original game's classical value:

$$\omega_\lambda(\mathcal{G}_{\mathrm{comp}}, S_{\mathrm{classical}}) \leq \omega_c(\mathcal{G}) + \mathrm{negl}(\lambda),$$

where $\mathrm{negl}(\lambda)$ is a negligible function that goes to zero faster than the reciprocal of any polynomial in λ, which is also dependent on the QHE scheme for the compilation. Quantum completeness refers to the fact that there exists an efficient quantum strategy S_{complete} for the compiled game whose success probability approaches the original game's quantum value as the security parameter grows:

$$\lim_{\lambda\to\infty} \omega_\lambda(\mathcal{G}_{\mathrm{comp}}, S_{\mathrm{complete}}) = \omega_q(\mathcal{G}).$$

In particular, if the original nonlocal game exhibits a quantum advantage, then the compiled single-prover protocol also exhibits a classical-quantum gap, at least as big as the nonlocal one.

With the above setup, *quantum soundness* refers to the property that even a malicious quantum

prover cannot win with probability exceeding the optimal quantum value of the original k-player nonlocal game.

Definition 2.2 (Quantum soundness for compiled k-partite games)**.** *Let $\mathcal{G}$ be a k-partite nonlocal game with quantum value $\omega_q(\mathcal{G})$ and commuting-operator value $\omega_{\mathrm{qc}}(\mathcal{G})$. Let $\mathcal{G}_{\mathrm{comp}} := (\mathcal{G}^{\lambda}_{\mathrm{comp}})_{\lambda}$ denote the family of compiled games produced by the KLVY compilation at security parameter λ.*

Let $S = (S_\lambda)_\lambda$ be any efficient (QPT) prover strategy for $\mathcal{G}_{\mathrm{comp}}$, and let $p^{\lambda}(\vec{a}|\vec{x})$ be the correlation realized by S_λ in $\mathcal{G}^{\lambda}_{\mathrm{comp}}$. We say that the compiler is quantum sound *(against QPT strategies) if there exists a value $B(\mathcal{G}) \in \{\omega_q(\mathcal{G}), \omega_{\mathrm{qc}}(\mathcal{G})\}$ such that for every efficient strategy S there exists a negligible function $\mathrm{negl}_S(\lambda)$ with*

$$\omega_\lambda(\mathcal{G}_{\mathrm{comp}}, S) \leq B(\mathcal{G}) + \mathrm{negl}_S(\lambda) \quad \textit{for all } \lambda \in \mathbb{N}.$$

Our focus here is precisely on establishing quantitative quantum soundness for *all* compiled *k-partite* games: an analytical upper bound on the compiled protocol's quantum winning probability as a function of λ, showing it stays within a negligible distance of the appropriate benchmark from the original k-partite game.

2.2 The NPA hierarchy for bounding quantum correlations

To upper-bound the winning probabilities of quantum strategies in a k-partite nonlocal game $\mathcal{G}$, a powerful tool is the Navascués-Pironio-Acín (NPA) hierarchy [NPA08; PNA10]. The hierarchy is a sequence of increasingly tight semidefinite programming (SDP) relaxations that characterize the set of quantum-achievable correlations. In broad terms, NPA provides a family of efficiently checkable necessary conditions for a conditional distribution $p(\vec{a}|\vec{x})$ to arise from some quantum strategy. By optimizing the game's payoff over these conditions, one obtains an upper bound on the quantum value. If, at some finite level, a distribution violates an NPA constraint, then it cannot come from any measurement on any shared quantum state. Conversely, as one increases the level (adding higher-degree algebraic constraints), the feasible set converges to the set of commuting-operator quantum correlations, and thus the bounds converge to the commuting-operator value $\omega_{\mathrm{qc}}(\mathcal{G})$. In practice, low levels already yield sharp bounds for many games.

To explain the construction in the k-partite setting, it is convenient to use an abstract algebraic presentation.

Definition 2.3 (Measurement symbols and relations)**.** *For each player $i \in [k]$, question $x_i \in X_i$, and answer $a_i \in A_i$, introduce a symbol $f^{(i)}_{a_i|x_i}$ that represents a measurement effect. Without loss of generality the measurements can be assumed to be projective*

$$f^{(i)}_{a_i|x_i}{}^{*} = f^{(i)}_{a_i|x_i}, \qquad f^{(i)}_{a_i|x_i} f^{(i)}_{a_i'|x_i} = \delta_{a_i,a_i'} f^{(i)}_{a_i|x_i}, \qquad \sum_{a_i} f^{(i)}_{a_i|x_i} = \mathbb{1}$$

and we impose inter-party commutation

$$[f^{(i)}_{a_i|x_i}, f^{(j)}_{a_j|x_j}] = 0 \qquad (i \neq j).$$

Any commuting-operator strategy is a $*$-representation of these symbols on a Hilbert space with a state ρ.

The NPA hierarchy is phrased in terms of words (monomials) in these symbols and a corresponding moment matrix.

Definition 2.4 (Words and level-n moment matrix)**.** *Fix a truncation level $n \in \mathbb{N}, n \geq 1$. Let $\mathcal{W}^n$ be the set of words of length $\leq n$ in the symbols $f_{a|x}^{(i)}$ (and $\mathbb{1}$), modulo the relations above; because different players' symbols commute, words admit a canonical normal order. Given a commuting-operator realization $\left(\rho, \{f_{a|x}^{(i)}\}\right)$ and a purification $|\psi\rangle$ of ρ, the level-n moment matrix is the Hermitian matrix*

$$\Gamma_{w,v}^{(n)} = \langle\psi| w^* v |\psi\rangle, \qquad \forall w, v \in \mathcal{W}^n.$$

Then $\Gamma^{(n)} \succeq 0$ (it is a Gram matrix), and it satisfies all linear identities implied by Theorem 2.3. The degree-1 block reads off the correlation:

$$p(\vec{a}|\vec{x}) = \Gamma^{(n)}_{\mathbb{1}, \Pi_{i=1}^k f_{a_i|x_i}^{(i)}},$$

whenever $\Pi_{i=1}^k f_{a_i|x_i}^{(i)} \in \mathcal{W}^{(n)}$.

With these ingredients, the level-n NPA relaxation is the SDP that maximizes the game's payoff over all PSD matrices $\Gamma^{(n)}$ obeying the linear identities.

Definition 2.5 (Level-n NPA upper bound for a k-partite game)**.** *Let $\beta_{\vec{a},\vec{x}}$ be the payoff tensor (in predicate form $\beta_{\vec{a},\vec{x}} = \mu(\vec{x})V(\vec{x},\vec{a})$) from the nonlocal game $\mathcal{G}$. The level-n NPA bound is*

$$\begin{aligned} \omega_{\mathrm{NPA}}^n(\mathcal{G}) := \max_{\Gamma^n} & \sum_{\vec{a},\vec{x}} \beta_{\vec{a},\vec{x}} \Gamma^{(n)}_{\mathbb{1}, \Pi_{i=1}^k f_{a_i|x_i}^{(i)}} \\ s.t. \quad & \Gamma^{(n)} \succeq 0, \\ & \textit{all linear identities induced by the relations on } \mathcal{W}^n. \end{aligned}$$

Two basic properties encapsulate the usefulness of the hierarchy:

(a) *Soundness.* Any commuting-operator strategy produces a feasible $\Gamma^{(n)}$, so for all n

$$\omega_{\mathrm{qc}}(\mathcal{G}) \leq \omega_{\mathrm{NPA}}^n(\mathcal{G}).$$

(b) *Monotone convergence.* The sequence $\omega_{\mathrm{NPA}}^n(\mathcal{G})$ is nonincreasing and converges to the commuting-operator value:

$$\omega_{\mathrm{NPA}}^1(\mathcal{G}) \geq \omega_{\mathrm{NPA}}^2(\mathcal{G}) \geq \cdots \searrow \omega_{\mathrm{qc}}(\mathcal{G}).$$

Thus, optimizing at higher levels tightens the upper bound, and in the limit (including moments of all lengths) one recovers the exact commuting-operator quantum value.

3 Generalized NPA hierarchy for multipartite quantum sequential scenarios

In this section, we introduce a generalized NPA hierarchy that converges to multipartite sequential setups as in Fig. 1(b). The main novelty of our multipartite generalization is the use of $*$-homomorphisms with appropriate constraints, which allows us to model quantum instruments that are essential in the description of multipartite quantum sequential scenarios.

We begin with the bipartite sequential scenarios in Section 3.1 to set a foundation for the more complex multipartite scenarios, and to contrast the subnormalized-moment-matrix method of [Kle+25]. In Section 3.2, we then discuss in detail the construction of the tripartite case for pedagogical purposes and finally introduce the k-partite sequential NPA hierarchy in Section 3.3.

3.1 Bipartite case revisited

As a proof of concept, we first consider the well-understood bipartite case and introduce a sequential generalization of the NPA hierarchy that is more composable than that of [Kle+25] (see Section A.1 for a quick review).

In a sequential bipartite game $\mathcal{G}$, Alice receives some state σ, applies some quantum instrument $A_{a|x}$, and passes it to Bob for another measurement $B_{b|y}$. Using the notation from [Bar+25, Lemma 12] and thinking in the Heisenberg picture, we first consider generators $\{f_{b|y} \mid \forall b, y\}$ satisfying the relations $\mathcal{R}_B$:

$$f_{b|y}^* = f_{b|y}, \quad f_{b|y} f_{b'|y} = \delta_{b,b'} f_{b|y}, \quad \sum_b f_{b|y} = \mathbb{1}.$$

Define Bob's algebra by the universal PVM C^*-algebra as

$$\mathcal{A}_B = C^*(\{f_{b|y}\}_{b,y} \mid \mathcal{R}_B).$$

In addition, consider generators $\{f_{ab|xy} \mid \forall a, b, x, y\}$ satisfying the relations $\mathcal{R}_{AB}$:

$$f_{ab|xy}^* = f_{ab|xy}, \quad f_{ab|xy} f_{a'b'|xy} = \delta_{a,a'} \delta_{b,b'} f_{ab|xy},$$
$$\sum_b f_{ab|xy} = \sum_b f_{ab|xy'} \ \forall y, y', \quad \sum_{a,b} f_{ab|xy} = \mathbb{1}.$$

Define the post-Alice-measurement algebra of Bob by

$$\mathcal{A}_{AB} = C^*(\{f_{ab|xy}\}_{a,b,x,y} \mid \mathcal{R}_{AB}),$$

denoted as $\mathcal{A}_{A\to B}$ in [Bar+25]. For every a, x, it is shown in [Bar+25] that there exists a (not necessarily unital) *-homomorphism that maps generators to generators

$$T_{a|x} : \mathcal{A}_B \to \mathcal{A}_{AB}, \quad f_{b|y} \mapsto f_{ab|xy},$$

which shall be central to our analysis. A simple –but very effective– change of perspective with respect to all previous works, is to not model Alice's action as post-measurement states $\varphi_{a|x}$, but through CP maps $T_{a|x}$.

For the game $\mathcal{G}$ and a correlation $p(ab|xy)$, the associated score is $\sum_{a,b,x,y} \beta_{abxy} p(ab|xy)$. Denote the objective Bell polynomial by

$$\beta = \sum_{a,b,x,y} \beta_{abxy} f_{ab|xy} \in \sum_{a,x} T_{a|x}(\mathcal{A}_B) \subset \mathcal{A}_{AB}.$$

We now construct a variant of the NPA hierarchy with words $f_{ab|xy}$ and leverage the equality $f_{ab|xy} = T_{a|x}(f_{b|y})$ to encode the sequential information, such that it converges to the algebraic bipartite sequential strategy.

To this end, define the word sets at level n by

$$\mathcal{W}_{AB}^n := \left\{ w = f_{a_1 b_1|x_1 y_1} \cdots f_{a_k b_k|x_k y_k} \;\middle|\; 0 \le k \le n \right\} \subset \mathcal{A}_{AB},$$
$$\mathcal{W}_B^n := \left\{ w = f_{b_1|y_1} \cdots f_{b_k|y_k} \;\middle|\; 0 \le k \le n \right\} \subset \mathcal{A}_B,$$
$$T_{a|x}\big(\mathcal{W}_B^n\big) = \left\{ f_{ab_1|xy_1} \cdots f_{ab_k|xy_k} \;\middle|\; 0 \le k \le n \right\} \subset T_{a|x}(\mathcal{A}_B) \subset \mathcal{A}_{AB}.$$

Clearly, $\beta \in \mathrm{span}(T_{a|x}(\mathcal{W}_B^n)) \subset \mathrm{span}(\mathcal{W}_{AB}^n)$ for $n \geq 1$. Note that for any polynomial $s \in \mathrm{span}(\mathcal{W}_B^n)$, the polynomial $T_{a|x}(s) \in \mathcal{W}_{AB}^n$ has the same degree as s, i.e., the homomorphism $T_{a|x}$ does not increase the degree.

Consider the matrix $\Gamma^{(n)}$ indexed by the monomials/words in $\mathcal{W}_{AB}^n$ with the associated functional $L^{2n} : \mathcal{W}_{AB}^{2n} \to \mathbb{C}$ defined by

$$L^{2n}(w^* v) := \Gamma^{(n)}_{w,v}, \quad \forall w, v \in \mathcal{W}_{AB}^n.$$

For any $P \in \mathrm{span}(\mathcal{W}_{AB}^{2n})$ with $\deg P = k$, denote by

$$\Gamma^{(n)}(P)_{w,v} = L^{2n}(w^* P v), \quad \forall w, v \in \mathcal{W}_{AB}^{n-\lceil k/2 \rceil}$$

the localizing matrix of $\Gamma^{(n)}$ at P. This leads to the following NPA-like hierarchy at level n:

$$\begin{aligned}
\omega^n_{\text{2seqNPA}}(\mathcal{G}) := \max \ & \sum_{a,b,x,y} \beta_{abxy} \, \Gamma^{(n)}_{\mathbb{1}, f_{ab|xy}} \\
\text{s.t.} \ & \Gamma^{(n)} \succeq 0, \\
& \Gamma^{(n)}_{\mathbb{1},\mathbb{1}} = 1, \\
& \Gamma^{(n)}_{w,v} = \Gamma^{(n)}_{w',v'} \quad \big(\text{when } w^* v = w'^* v'\big), \\
& M^{x,x'}_{(w,r),(s,v)} := \Gamma^{(n)}\Big(\sum_a \big(T_{a|x}(r^* s) - T_{a|x'}(r^* s)\big)\Big)_{w,v} = 0 \\
& \qquad\qquad \substack{\forall x, x', \\ \forall w,v \in \mathcal{W}_{AB}^n, r,s \in \mathcal{W}_B^n, \\ \deg w + \deg v + \deg r + \deg s \leq 2n} \quad \text{(operationally-non-signaling)}.
\end{aligned} \tag{5}$$

The last operationally-non-signaling constraint is equivalent to

$$\sum_a L^{2n}(P^* T_{a|x}(S) Q) - \sum_a L^{2n}(P^* T_{a|x'}(S) Q) = 0$$

for all x, x', $P, Q \in \mathrm{span}(W_{AB}^n)$, and $S \in \mathrm{span}(W_B^n)$ such that $\deg P + \deg Q + \deg S \leq 2n$. A moment matrix $\Gamma^{(n)}$ is said to be a feasible solution of Eq. (5) if it satisfies all the constraints but does not necessarily maximize the score, and is said to be an optimal feasible solution of Eq. (5) if it satisfies all constraints and maximizes the score.

The physical submatrices of $\Gamma^{(n)}$ include the block indexed by $T_{a|x}(\mathcal{W}_B^n) \times T_{a|x}(\mathcal{W}_B^n)$ corresponding to Bob's measurement subject to output-input (a, x) by Alice's measurement. Thus, the hierarchy Eq. (5) is stricter than the bipartite sequential NPA hierarchy (Eq. (20)) by identifying this submatrix with its Riesz functional $\sigma_{a|x}$.

We present the following convergence theorem on the hierarchy Eq. (5).

Theorem 3.1. *Let $p(ab|xy)$ be a correlation for the nonlocal game $\mathcal{G}$. The following statements are equivalent:*

(i) The correlation $p(ab|xy)$ arises from a bipartite sequential operationally-non-signaling strategy.

(ii) The correlation $p(ab|xy)$ arises from a bipartite commuting operator strategy.

(iii) There exists a family of $\{\Gamma^{(n)}\}_n$ of feasible solutions to Eq. (5) such that $p(ab|xy) = \Gamma^{(n)}_{\mathbb{1}, f_{ab|xy}}$ for all n.

Consequently, $\omega^n_{2\text{seqNPA}}(\mathcal{G}) \searrow \omega_{\text{qc}}(\mathcal{G})$ *monotonically as* $n \to \infty$.

Proof. It is clear that both *(i)* and *(ii)* imply *(iii)*. The equivalence between *(i)* and *(ii)* is well-established in, e.g., [Kul+25].

The statement *(iii)* implies *(i)* is a consequence of the paragraph proceeding the theorem, combined with the convergence property of the sequential NPA hierarchy of [Kle+25]. Nonetheless, we do another proof that connects to [Bar+25] better.

To this end, one has $\text{span}(\mathcal{W}^n_{AB}) \to \mathcal{A}_{AB}$ as $n \to \infty$. Since all generators $f_{ab|xy}$ are projective, each entry of $\Gamma^{(n)}$ is necessarily bounded for all n. Then the standard Banach-Alaoglu argument implies the existence of a convergent subsequence $\Gamma^{(n_k)}$ and an infinite moment matrix Γ, such that

$$\Gamma^{(n_k)}_{w,v} \to \Gamma_{w,v}$$

for all w, v and all $n_k \geq \max\{\deg(w), \deg(v)\}$. That is, we have recovered a state σ on $\mathcal{A}_{AB}$ via

$$\begin{aligned} \sigma : \mathcal{A}_{AB} &\to \mathbb{C}, \\ w^* v &\mapsto \Gamma_{w,v}, \end{aligned}$$

such that

$$p(ab|xy) = \sigma(T_{a|x}(f_{b|y}))$$

and

$$\sum_a \sigma(P^* T_{a|x}(S) Q) - \sigma(P^* T_{a|x'}(S) Q) = 0$$

for all $S \in \mathcal{A}_B$ and $P, Q \in \mathcal{A}_{AB}$. We are done by identifying σ with an asymptotically-secured C^*-algebraic compiled strategy for two players [Bar+25, Definition 17 and Theorem 14].

For completeness, let us give an explicit proof. Consider the GNS representation $(\mathcal{H}, \pi, |\Omega\rangle)$ of σ, where $\mathcal{H}$ is a Hilbert space, $\pi : \mathcal{A}_{AB} \to B(\mathcal{H})$ is a $*$-representation, and $|\Omega\rangle \in \mathcal{H}$ such that $\sigma(P) = \langle \Omega | \pi(P) | \Omega \rangle$ for all $P \in \mathcal{A}_{AB}$. It follows that, for any x, x', $S \in \mathcal{A}_B$, and $P, Q \in \mathcal{A}_{AB}$,

$$0 = \langle \Omega | \pi(P)^* \pi(\sum_a T_{a|x}(S) - \sum_a T_{a|x'}(S)) \pi(Q) | \Omega \rangle.$$

By cyclicity, this implies that

$$\sum_a \pi \circ T_{a|x} = \sum_a \pi \circ T_{a|x'} := \bar{T} : \mathcal{A}_B \to B(\mathcal{H}),$$

where $\pi \circ T_{a|x}$ is completely positive map $\mathcal{A}_B \to B(\mathcal{H})$ (due to *-homomorphisms being completely positive) dominated by the completely positive map $\bar{T}$. Thus, $(|\Omega\rangle, \pi \circ T_{a|x})$ defines a C^*-algebraic sequential correlation as in [Bar+25, Definition 22], finishing *(iii)* $\implies$ *(i)*.

Let us show *(iii)* $\implies$ *(ii)*, analogously to [Bar+25, Theorem 16] for further demonstration. Let $(\mathcal{K}, \pi_{\mathcal{K}}, V)$ be the minimal Stinespring dilation of $\bar{T}$ such that, for all $S \in \mathcal{A}_B$,

$$\bar{T}(S) = \sum_a \pi \circ T_{a|x}(S) = V^* \pi_{\mathcal{K}}(S) V.$$

Arveson's Radon-Nikodym derivative [Arv69, Theorem 1.4.2], see also [Rag03], then implies the

existence of unique positive operator $F_{a|x} \in \pi_{\mathcal{K}}(\mathcal{A}_B)'$ such that

$$\pi \circ T_{a|x}(S) = V^* F_{a|x} \pi_{\mathcal{K}}(S) V.$$

The uniqueness of the minimal dilation with $\sum_a T_{a|x} = \bar{T}$ further imposes $\sum_a F_{a|x} = \mathbb{1}$, i.e., $F_{a|x}$ form POVMs for every x. We are done by identifying the state σ with $V|\Omega\rangle$. □

We now analyze some properties of the novel sequential hierarchy that will be useful later; more precisely, we discuss the stopping criterion and strict feasibility.

Definition 3.2. *For $n \in \mathbb{N}$, let $\Gamma^{(n)}$ be the solution for Eq. (5) at level n for some nonlocal game $\mathcal{G}$. Consider its block form*

$$\Gamma^{(n)} = \begin{pmatrix} \Gamma^{(n-1)} & M \\ M^* & N \end{pmatrix},$$

where $\Gamma^{(n-1)}$ is the principal block indexed by words in $\mathcal{W}_{AB}^{n-1}$. We say that the solution $\Gamma^{(n)}$ is flat (or has a rank-loop) if

$$\operatorname{rank}(\Gamma^{(n)}) = \operatorname{rank}(\Gamma^{(n-1)}) < \infty.$$

Proposition 3.3. *The hierarchy of Eq. (5) for $\mathcal{G}$ admits a flat optimal solution at some finite level n if and only if the $\mathcal{G}$ admits a finite-dimensional optimal quantum strategy.*

Proof. This follows from the standard flatness argument, see e.g., [NPA08, Theorem 10]. Note that the equality $\operatorname{rank}(\Gamma^{(n)}) = \operatorname{rank}(\Gamma^{(n-1)})$ is sufficient since this already enforces the operationally-non-signaling condition on the generator $f_{ab|xy}$ in the resulting finite-dimensional representation and thereby can propagate to higher degree words. □

Note that the existence of flat solutions does not guarantee that numerical algorithms will find it in practice. In fact, it is possible that there exist infinitely many inequivalent finite-dimensional optimal strategies, leading the SDP solver to return any convex mixture of them.

Proposition 3.4. *For every level n, the SDP in Eq. (5) is strictly feasible. That is, there exists a moment matrix $\Gamma_{\text{strict}}^{(n)} \succ 0$ that satisfies all the linear constraints of Eq. (5).*

Proof. The unconstrained NPA hierarchy with commuting PVMs $\{A_{a|x}\}, \{B_{b|y}\}$ admits a strictly feasible moment matrix at every level. This follows from the faithfulness of the left-regular GNS representation of the universal $*$-algebra generated by $A_{a|x}$ and $B_{b|y}$. Thus, for each n there exists a full-rank moment matrix $\Gamma_{\text{NPA}}^{(n)} \succ 0$ for the standard NPA hierarchy. (See [Tav+24, Appendix C] for a more explicit argument.)

Define $f_{ab|xy} := A_{a|x} B_{b|y}$. Then every word in the sequential hierarchy of Eq. (5) is also a word in the larger algebra generated by $A_{a|x}$, $B_{b|y}$. Therefore, we let $\Gamma_{\text{strict}}^{(n)}$ be the principle submatrix of $\Gamma_{\text{NPA}}^{(n)}$ corresponding to all the $f_{ab|xy}$-words, which can be straightforwardly checked satisfy all constraints of Eq. (5). Finally, since a principal submatrix of a positive definite matrix is itself positive definite, $\Gamma_{\text{strict}}^{(n)} \succ 0$. □

3.2 Tripartite case

Now, we showcase the construction of the tripartite sequential NPA hierarchy in detail to inspire the general multipartite case. As in Fig. 1(b), the game begins with Alice A receiving and answering

with the pair (x, a), followed by Bob B with the pair (y, b), and ends with Charlie C with the pair (z, c), during which the operationally non-signaling condition is respected by Alice and Bob.

To describe this scenario, consider generators $\{f_{c|z} \mid \forall c, z\}$ satisfying the relation $\mathcal{R}_C$:

$$f_{c|z}^* = f_{c|z}, \quad f_{c|z} f_{c'|z} = \delta_{c,c'} f_{c|z}, \quad \sum_c f_{c|z} = \mathbb{1},$$

and define Charlie's algebra by the universal PVM C^*-algebra

$$\mathcal{A}_C = C^*(\{f_{c|z}\}_{c,z} \mid \mathcal{R}_C).$$

Next, consider generators $\{f_{bc|yz} \mid \forall b, c, y, z\}$ satisfying the relation $\mathcal{R}_{BC}$:

$$\begin{aligned} f_{bc|yz}^* = f_{bc|yz}, \quad f_{bc|yz} f_{b'c'|yz} = \delta_{b,b'} \delta_{c,c'} f_{bc|yz}, \\ \sum_c f_{bc|yz} = \sum_c f_{bc|yz'} \ \forall z, z', \quad \sum_{b,c} f_{bc|yz} = \mathbb{1}. \end{aligned} \tag{6}$$

Define post-Bob-measurement algebra of Charlie by

$$\mathcal{A}_{BC} = C^*(\{f_{bc|yz}\}_{b,c,y,z} \mid \mathcal{R}_{BC}).$$

denoted as $\mathcal{A}_{B\to C}$ [Bar+25]. Then, consider the generators $\{f_{abc|xyz} \mid \forall a, b, c, x, y, z\}$ satisfying the relation $\mathcal{R}_{ABC}$:

$$\begin{aligned} f_{abc|xyz}^* = f_{abc|xyz}, \quad f_{abc|xyz} f_{a'b'c'|xyz} = \delta_{a,a'} \delta_{b,b'} \delta_{c,c'} f_{abc|xyz}, \\ \sum_c f_{abc|xyz} = \sum_c f_{abc|xyz'}, \quad \sum_{b,c} f_{abc|xyz} = f_{abc|xy'z'}, \quad \sum_{a,b,c} f_{abc|xyz} = \mathbb{1}, \end{aligned}$$

and define the post-Alice-Bob-measurement algebra of Charlie by

$$\mathcal{A}_{ABC} = C^*(\{f_{abc|xyz}\}_{a,b,c,x,y,z} \mid \mathcal{R}_{ABC}),$$

denoted as $\mathcal{A}_{A\to B\to C}$ in [Bar+25]. Finally, define the natural *-homomorphisms for a, b, x, y by

$$\begin{aligned} T_{b|y} : \mathcal{A}_C \to \mathcal{A}_{BC}, \quad f_{c|z} \mapsto f_{bc|yz}, \\ T_{a|x} : \mathcal{A}_{BC} \to \mathcal{A}_{ABC}, \quad f_{bc|yz} \mapsto f_{abc|xyz}. \end{aligned}$$

Similarly, for a tripartite nonlocal game $\mathcal{G}$ and a correlation $p(abc|xyz)$, the associated score $\sum_{a,b,c,x,y,z} \beta_{abcxyz} p(abc|xyz)$ results in the objective polynomial

$$\beta = \sum_{a,b,c,x,y,z} \beta_{abcxyz} f_{abc|xyz} \in \sum_{a,b,x,y} T_{a|x} T_{b|y}(\mathcal{A}_C) \subset \mathcal{A}_{ABC}.$$

Analogous to Section 3.1, define the word sets at level n:

$$\begin{aligned} \mathcal{W}_{ABC}^n &:= \left\{ f_{a_1 b_1 c_1|x_1 y_1 z_1} \cdots f_{a_k b_k c_k|x_k y_k z_k} \mid 0 \le k \le n \right\} \subset \mathcal{A}_{ABC}, \\ \mathcal{W}_{BC}^n &:= \left\{ f_{b_1 c_1|y_1 z_1} \cdots f_{b_k c_k|y_k z_k} \mid 0 \le k \le n \right\} \subset \mathcal{A}_{BC}, \\ \mathcal{W}_{C}^n &:= \left\{ f_{c_1|z_1} \cdots f_{c_k|z_k} \mid 0 \le k \le n \right\} \subset \mathcal{A}_{C}, \end{aligned}$$

along with $T_{b|y}(\mathcal{W}_C^n) \subset \mathcal{A}_{BC}$ and $T_{a|x}T_{b|y}(\mathcal{W}_C^n) \subset \mathcal{A}_{ABC}$. Again, $\beta \in \mathrm{span}(\mathcal{W}_{ABC}^n)$ for $n \geq 1$.

Consider the matrix $\Gamma^{(n)}$ indexed by monomials/words in $\mathcal{W}_{ABC}^n$. Using the same notation as in Section 3.1, we define the following tripartite sequential NPA-like hierarchy at level n:

$$
\begin{aligned}
\omega_{3\text{seqNPA}}^n(\mathcal{G}) \;:=\; \max \quad & \sum_{a,b,c,x,y,z} \beta_{abcxyz}\, \Gamma^{(n)}_{\mathbb{1},f_{abc|xyz}} \\
\text{s.t.} \quad & \Gamma^{(n)} \succeq 0, \\
& \Gamma^{(n)}_{\mathbb{1},\mathbb{1}} = 1, \\
& \Gamma^{(n)}_{w,w'} = \Gamma^{(n)}_{v,v'} \quad \text{whenever } w^*w' = v^*v', \\
& M^{x,x'}_{(w,r),(s,v)} \;:=\; \Gamma^{(n)}\Big(\big(\sum_a T_{a|x}(r^*s) - T_{a|x'}(r^*s)\big)\Big)_{w,v} = 0 \\
& \qquad\qquad \substack{\forall x,x', \forall w,v \in \mathcal{W}_{ABC}^n, \\ r,s \in \mathcal{W}_{BC}^n, \\ \deg w + \deg v + \deg r + \deg s \leq 2n} \quad \text{(Alice operationally-non-signaling)} \\
& N^{y,y'}_{(r,t),(s,u)} \;:=\; \Gamma^{(n)}\Big(T_{a|x}\big(\sum_b T_{b|y}(t^*u) - T_{b|y'}(t^*u)\big)\Big)_{T_{a|x}(r),T_{a|x}(s)} = 0 \\
& \qquad\qquad \substack{\forall\, a,x,y,y', \\ r,s \in \mathcal{W}_{BC}^n, \\ t,u \in \mathcal{W}_C^n, \\ \deg r + \deg s + \deg t + \deg u \leq 2n} \quad \text{(Bob operationally-non-signaling)}
\end{aligned}
\tag{7}
$$

Here, let L^{2n} be the normalized positive linear map associated with $\Gamma^{(n)}$, then Alice operationally-non-signaling constraint is equivalent to

$$\sum_a L^{2n}(P^*T_{a|x}(S)Q) - \sum_a L^{2n}(P^*T_{a|x'}(S)Q) = 0$$

for all x, x', $P, Q \in \mathrm{span}(\mathcal{W}_{ABC}^n)$, and $S \in \mathrm{span}(\mathcal{W}_{BC}^n)$ such that $\deg P + \deg Q + \deg S \leq 2n$. In addition, Bob operationally-non-signaling constraint is equivalent to

$$\sum_b L^{2n} \circ T_{a|x}(R^*T_{b|y}(O)S) - \sum_b L^{2n} \circ T_{a|x}(R^*T_{b|y'}(O)S) = 0$$

for all a, x, y, y', $R, S \in \mathrm{span}(\mathcal{W}_{BC}^n)$, and $O \in \mathrm{span}(\mathcal{W}_C^n)$ such that $\deg R + \deg S + \deg O \leq 2n$.

A moment matrix is said to be a feasible solution of Eq. (7) if it satisfies all the constraints but does not necessarily maximize the score, and is said to be an optimal feasible solution of Eq. (7) if it satisfies all constraints and maximizes the score.

Similarly to the bipartite case, $\mathcal{W}_C^n \times \mathcal{W}_C^n$ plays the role of Charlie's measurement, the block of $\mathcal{W}_{b|y}^n \times \mathcal{W}_{b|y}^n$ corresponds to the case when Bob measures afterwards with (b, y), and the block $\mathcal{W}_{ab|xy}^n \times \mathcal{W}_{ab|xy}^n$ represents the case of Alice measuring with (a, x) after Bob and Charlie's measurement. We analogously have the following theorem.

Theorem 3.5. *Let $p(abc|xyz)$ be a correlation for the tripartite nonlocal game $\mathcal{G}$. The following statements are equivalent:*

(i) The correlation $p(abc|xyz)$ arises from a tripartite sequential operationally-non-signaling strategy.

(ii) The correlation $p(abc|xyz)$ arises from a tripartite commuting operator strategy.

(iii) There exists a family of $\{\Gamma^{(n)}\}_n$ *of feasible solutions to Eq.* (7) *such that* $p(abc|xyz) = \Gamma^{(n)}_{\mathbb{1},f_{abc|xyz}}$ *for all* n.

Consequently, $\omega^n_{3\text{seqNPA}}(\mathcal{G}) \searrow \omega_{\text{qc}}(\mathcal{G})$ *monotonically as* $n \to \infty$.

Proof. It is straightforward to check that both *(i)* and *(ii)* imply *(iii)*. The equivalence between *(i)* and *(ii)* is shown by [Bar+25] thanks to the new chain rule for Arveson's Radon-Nikodym derivatives.

The direction *(iii)* $\implies$ *(i)* is almost the same as the proof of Theorem 3.1. Again, $\text{span}(\mathcal{W}^n_{ABC}) \to \mathcal{A}_{ABC}$ as $n \to \infty$ with projective generators. Then by the Banach-Alaoglu Theorem there exists a weak-* convergent subsequence $\Gamma^{(n_k)} \to \Gamma$. That is, we have obtained a state on $\mathcal{A}_{ABC}$ via

$$\begin{aligned} \sigma : \mathcal{A}_{ABC} &\to \mathbb{C}, \\ w^*v &\mapsto \Gamma_{w,v} \end{aligned}$$

such that

$$p(abc|xyz) = \sigma(f_{abc|xyz}) = \sigma(T_{a|x}T_{b|y}(f_{c|z})).$$

Furthermore, we have the operational-non-signaling for Alice

$$\sum_a \sigma(P^*T_{a|x}(S)Q) - \sum_a \sigma(P^*T_{a|x'}(S)Q) = 0$$

for all $P, Q \in \mathcal{A}_{ABC}$ and $S \in \mathcal{A}_{BC}$; and for Bob

$$\sum_b \sigma \circ T_{a|x}(R^*T_{b|y}(O)S) - \sum_b \sigma \circ T_{a|x}(R^*T_{b|y'}(O)S) = 0$$

for all $R, S \in \mathcal{A}_{BC}$ and $O \in \mathcal{A}^n_C$.

We are done by identifying σ with a asymptotically-secured C^*-algebraic compiled strategy for three players as in [Bar+25, Definition 17 and Theorem 14] and then invoking [Bar+25, Theorem 16 or 17]. □

We omit the flatness condition and the strict feasibility for the tripartite hierarchy since it is straightforward, and instead present the k-partite variant directly in the next subsection.

3.3 General multipartite case

We are ready to tackle the general k-partite sequential quantum scenarios for any $k \geq 2$.

First, let us develop a notation for any k-partite sequential algebras and the corresponding NPA hierarchy. In our sequential convention, we begin with the kth party and then $(k-1)$-th until the 1-st party, i.e., $k \to k-1 \to \cdots \to 1$. Write $[j] := \{1, 2, \ldots, j\}$ and denote $a_1 \cdots a_j, a'_1 \cdots a'_j, x_1 \cdots x_j$ by $a_{[j]}, a'_{[j]}, x_{[j]}$, respectively, when there is no ambiguity.

For any $j \in [k]$, consider the generators/letters $\{f_{a_{[j]}|x_{[j]}} \mid \forall a_{[j]}, x_{[j]}\}$ satisfying the relation $\mathcal{R}_{[j]}$:

$$f^*_{a_{[j]}|x_{[j]}} = f_{a_{[j]}|x_{[j]}}, \quad f_{a_{[j]}|x_{[j]}} f_{a'_{[j]}|x_{[j]}} = \delta_{a_{[j]},a'_{[j]}} f_{a_{[j]}|x_{[j]}}, \quad \sum_{a_{[j]}} f_{a_{[j]}|x_{[j]}} = \mathbb{1},$$

$$\sum_{a_1,\ldots a_l} f_{a_1\cdots a_l a_{l+1}\cdots a_j|x_1\cdots x_l x_{l+1}\cdots x_j} = \sum_{a_1,\ldots a_l} f_{a_1\cdots a_l a_{l+1}\cdots a_j|x'_1\cdots x'_l x_{l+1}\cdots x_j}, \ \forall l \in [j].$$

The corresponding universal C^*-algebra is defined as

$$A_{[j]} = C^*(\{f_{a_{[j]}|x_{[j]}}\}_{a_{[j]},x_{[j]}} \mid \mathcal{R}_{[j]}). \tag{8}$$

For every j such that $1 < j \leq k$, there exists a natural *-homomorphism for a_k, x_k such that

$$T^j_{a_j|x_j} : \mathcal{A}_{[j-1]} \to \mathcal{A}_{[j]}, \quad f_{a_1\cdots a_{j-1}|x_1\cdots x_{j-1}} \mapsto f_{a_1\cdots a_{j-1}a_j|x_1\cdots x_{j-1}x_j}.$$

That is, in the Heisenberg picture, the sequentiality is captured as

$$\mathcal{A}_{[1]} \xrightarrow{T^2_{a_2|x_2}} \mathcal{A}_{[2]} \xrightarrow{T^3_{a_3|x_3}} \cdots \xrightarrow{T^{k-1}_{a_{k-1}|x_{k-1}}} \mathcal{A}_{[k-1]} \xrightarrow{T^k_{a_k|x_k}} \mathcal{A}_{[k]}$$

with $\mathcal{A}_{[1]} = C^*(\{f_{a_1|x_1}\}_{a_1,x_1} \mid \mathcal{R}_{[1]})$ as the end party (e.g., Bob in the bipartite case and Charlie in the tripartite case).

For a k-partite nonlocal game $\mathcal{G}$ with the correlation $p(a_{[k]}|x_{[k]})$, the associated score objective Bell polynomial is

$$\beta = \sum_{a_{[k]},x_{[k]}} \beta_{a_{[k]}x_{[k]}} f_{a_{[k]}|x_{[k]}} \in \sum_{a_{[k]},x_{[k]}} T^k_{a_k|x_k} \cdots T^2_{a_2|x_2}(\mathcal{A}_{[1]}) \subset \mathcal{A}_{[k]}.$$

Let us define the n-th level of the k-partite sequential NPA hierarchy for game $\mathcal{G}$. We consider the word sets at level n for each j:

$$\mathcal{W}^n_{[j]} := \{w \in \mathcal{A}_{[k]} \mid 0 \leq \deg(w) \leq n\} \subset \mathcal{A}_{[j]}, \tag{9}$$

with $\beta \in \mathrm{span}(\mathcal{W}^n_{[k]})$ and $T^k_{a_k|x_k} \cdots T^2_{a_2|x_2}(\mathcal{W}^n_{[1]}) \subset \mathcal{W}^n_{[k]}$. For the matrix $\Gamma^{(n)}$ indexed by $\mathcal{W}^n_{[k]}$, the k-partite sequential NPA hierarchy at level is defined as:

$$\begin{aligned}
\omega^n_{k\text{seqNPA}}(\mathcal{G}) \;:=\; \max \quad & \sum_{a_{[k]},x_{[k]}} \beta_{a_{[k]}x_{[k]}}\, \Gamma^{(n)}_{\mathbb{1},f_{a_{[k]}|x_{[k]}}} \\
\text{s.t.} \quad & \Gamma^{(n)} \succeq 0, \\
& \Gamma^{(n)}_{\mathbb{1},\mathbb{1}} = 1, \\
& \Gamma^{(n)}_{w,w'} = \Gamma^{(n)}_{v,v'} \quad \text{whenever } w^*w' = v^*v', \\
& \sum_{a_j} \Gamma^{(n)}\Big(T^k_{a_k|x_k} \cdots T^{j+1}_{a_{j+1}|x_{j+1}}\big(T^j_{a_j|x_j}(r^*s) - T^j_{a_j|x'_j}(r^*s)\big)\Big)_{w,v} = 0 \\
& \qquad\qquad \begin{array}{c} \forall j\in[2,k]\cap\mathbb{N}, \\ \forall a_{j+1},\ldots,a_k,x'_j,x_j,\ldots,x_k, \forall\, r,s\in\mathcal{W}^n_{[j-1]}, \\ \forall\, w,v\in T^k_{a_k|x_k}\cdots T^{j+1}_{a_{j+1}|x_{j+1}}(\mathcal{W}^n_{[j]}), \\ \deg w+\deg v+\deg r+\deg s\leq 2n \end{array} \quad \text{(party-}j\text{ operationally-non-signaling).}
\end{aligned} \tag{10}$$

Here, let L^{2n} be the normalized positive linear map associated with $\Gamma^{(n)}$, then party-j operationally-non-signaling constraint is equivalent to

$$\sum_{a_j} L^{2n} \circ T^k_{a_k|x_k} \cdots T^{j+1}_{a_{j+1}|x_{j+1}}\big(P^*T^j_{a_j|x_j}(S)Q\big) - \sum_{a_j} L^{2n} \circ T^k_{a_k|x_k} \cdots T^{j+1}_{a_{j+1}|x_{j+1}}\big(P^*T^j_{a_j|x'_j}(S)Q\big) = 0$$

for all $x'_j, x_j, \ldots, x_k, a_{j+1}, \ldots a_k$, $P, Q \in \mathrm{span}(\mathcal{W}^n_{[j]})$, and $S \in \mathrm{span}(\mathcal{W}^n_{[j-1]})$. This corresponds to the

degree n relaxation of [Bar+24, Eq. (38)].

With the notation introduced, we state all the k-partite results analogous to the bipartite and tripartite cases.

Theorem 3.6. *Let $p(a_{[k]}|x_{[k]})$ be a correlation for the k-partite nonlocal game $\mathcal{G}$. The following statements are equivalent:*

(i) The correlation $p(a_{[k]}|x_{[k]})$ arises from a k-partite sequential operationally-non-signaling strategy.

(ii) The correlation $p(a_{[k]}|x_{[k]})$ arises from a k-partite commuting operator strategy.

(iii) There exists a family of $\{\Gamma^{(n)}\}_n$ of feasible solutions to Eq. (10) such that $p(a_{[k]}|x_{[k]}) = \Gamma^{(n)}_{\mathbb{1},f_{a_{[k]}|x_{[k]}}}$ for all n.

Consequently, $\omega^n_{k\text{seqNPA}}(\mathcal{G}) \searrow \omega_{\text{qc}}(\mathcal{G})$ monotonically as $n \to \infty$.

Proof. This follows from a routine inductive generalization of the proof for Theorems 3.1 and 3.5. □

We now present the flatness condition and strict feasibility of the k-partite hierarchy achieved by a direct inductive generalization.

Definition 3.7. *For $n \in \mathbb{N}$, let $\Gamma^{(n)}$ be the solution for Eq. (10) at level n for some nonlocal game $\mathcal{G}$. Consider its block form*

$$\Gamma^{(n)} = \begin{pmatrix} \Gamma^{(n-1)} & M \\ M^* & N \end{pmatrix},$$

where $\Gamma^{(n-1)}$ is the principal block indexed by words in $\mathcal{W}^{n-1}_{[k]}$. We say that the solution $\Gamma^{(n)}$ is flat (or has a rank-loop) if

$$\operatorname{rank}(\Gamma^{(n)}) = \operatorname{rank}(\Gamma^{(n-1)}) < \infty.$$

Proposition 3.8. *The hierarchy of Eq. (10) for $\mathcal{G}$ admits a flat optimal solution at some finite level n if and only if the $\mathcal{G}$ admits a finite-dimensional optimal quantum strategy.*

Proof. This follows from the standard flatness argument, see e.g., [NPA08, Theorem 10]. □

Proposition 3.9. *For every level n, the SDP in Eq. (10) is strictly feasible. That is, there exists a moment matrix $\Gamma^{(n)}_{\text{strict}} \succ 0$ that satisfies all the linear constraints of Eq. (10).*

Proof. This is inductive from the proof of Theorem 3.4. □

4 Quantitative quantum soundness for multipartite compiled nonlocal games

In this section, we give our main result (Theorem 4.4): the quantitative quantum soundness statement for all multipartite compiled nonlocal games (as in Fig. 1(b)). We begin with introducing the necessary notations in Section 4.1. Then in Section 4.2, since the old technique of signaling decomposition of [Kle+25] cannot be extended to multipartite scenarios, we give novel proofs for multipartite signaling decomposition to show that the compiled strategy at any λ is negligibly

different from some solutions of Eq. (10) for any NPA level n. This new approach uses geometric arguments involved with the projection on the convex set of constraints of Eq. (10), and then fixing the positivity with strict feasibility followed by a renormalization, see Fig. 2 for illustration.

4.1 Notations preliminaries

According to [Bar+25], at the security parameter λ, an efficient (quantum polynomial-time) strategy $S = (S_\lambda)$ of a k-partite compiled nonlocal game gives rise to states $\sigma^\lambda : \mathcal{A}_{[k]} \to \mathbb{C}$ such that

$$p^\lambda(a_{[k]}|x_{[k]}) = \sigma^\lambda(T^k_{a_k|x_k} \cdots T^2_{a_2|x_2}(f_{a_{[1]}|x_{[1]}})).$$

Instead of the operationally-non-signaling condition as in Eq. (10), we only have a weakly-non-signaling condition. That is, for all $j \in [2,k] \cap \mathbb{N}$ and for all $x'_j, x_j, \ldots, x_k, a_{j+1}, \ldots a_k$ and for every operator $P, Q \in \mathcal{A}_{[j]}$ and $R \in \mathcal{A}_{[j-1]}$, there exists a negligible function $\mathrm{negl}_{S,P,R,Q}(\lambda)$, depending on the strategy S and operators P, Q, R, such that

$$\begin{aligned}\Big|\sum_{a_j} \sigma^\lambda \circ T^k_{a_k|x_k} \cdots T^{j+1}_{a_{j+1}|x_{j+1}}\big(P^* T^j_{a_j|x_j}(R)Q\big) \\ - \sum_{a_j} \sigma^\lambda \circ T^k_{a_k|x_k} \cdots T^{j+1}_{a_{j+1}|x_{j+1}}\big(P^* T^j_{a_j|x'_j}(R)Q\big)\Big| \le \mathrm{negl}_{S,P,R,Q}(\lambda).\end{aligned} \tag{11}$$

Next, fix the NPA level $n \in \mathbb{N}$ with word sets for all $j \in [k]$. Consider the associated compiled moment matrix

$$(\Gamma^{(n),\lambda}_{\mathrm{comp}})_{w,v} := \sigma^\lambda(w^* v), \quad \Gamma^{(n),\lambda}_{\mathrm{comp}} \succeq 0,$$

for all $w, v \in \mathcal{W}^n_{[k]}$ that is "almost" a feasible solution of Eq. (10). Observe that $\left\|\Gamma^{(n),\lambda}_{\mathrm{comp}}\right\|_{\mathrm{op}} \le \left|\mathcal{W}^n_{[k]}\right|$ regardless of λ. Indeed, since σ^λ is contractive and the generators of C^*-algebra $\mathcal{A}_{[k]}$ are projective, all diagonals satisfy $(\Gamma^{(n),\lambda}_{\mathrm{comp}})_{w,w} \le 1$, thus

$$\left\|\Gamma^{(n),\lambda}_{\mathrm{comp}}\right\|_{\mathrm{op}} = \max \text{eigenvalue of } \Gamma^{(n),\lambda}_{\mathrm{comp}} \le \mathrm{Tr}\left(\Gamma^{(n),\lambda}_{\mathrm{comp}}\right) \le \left|\mathcal{W}^n_{[k]}\right|.$$

(In fact, the same argument shows that any level-n normalized moment matrix on $\mathcal{W}^n_{[k]}$ admits the same operator norm upper bound.)

To formalize and quantify the notion of "almost feasibility", we define a linear map $\mathcal{E}^n_k$. This map consolidates all linear constraints of Eq. (10) by mapping a Hermitian matrix H to a single vector, and H satisfies these constraints if and only if it lies in the kernel of $\mathcal{E}^n_k$, $\ker(\mathcal{E}^n_k)$. Specifically, let $\mathrm{Herm}(\mathcal{W}^n_{[k]})$ be the space of Hermitian matrices indexed by words in $\mathcal{W}^n_{[k]}$. Let

$$I^n_{\mathrm{Han}} = \{(w, v, w', v') \mid w^* v = w'^* v'\}$$

be the index pairs for the Hankel condition (symmetry of moment matrices) and, for each $j \in [2,k] \cap \mathbb{N}$,

let

$$\begin{aligned} I^n_{j-\text{ons}} = \{&(a_{j+1},\dots,a_k,x'_j,x_j,\dots,x_k,w,v,r,s) \mid \\ &\forall a_{j+1},\dots,a_k,x'_j,x_j,\dots,x_k,\ w,v\in\mathcal{W}^n_{[j]}, \\ &r,s\in\mathcal{W}^n_{[j-1]},\ \deg w+\deg v+\deg r+\deg s\le 2n\} \end{aligned}$$

be the set of tuples defining the operationally-non-signaling constraints of party j. Define

$$\mathcal{E}^n_k : \mathrm{Herm}(\mathcal{W}^n_{[k]}) \to \mathbb{C}^{|I^n_{\text{Han}}|} \oplus (\oplus_j \mathbb{C}^{|I^n_{j-\text{ons}}|})$$

$$H \mapsto \begin{pmatrix} (H_{w,v} - H_{w',v'})_{(w,v,w',v')\in I^n_{\text{Han}}} \\ \Big(\sum_{a_j} H\Big(T^k_{a_k|x_k}\cdots T^{j+1}_{a_{j+1}|x_{j+1}}\big(T^j_{a_j|x_j}(r^*s) - T^j_{a_j|x'_j}(r^*s)\big)\Big)_{\substack{T^k_{a_k|x_k}\cdots T^{j+1}_{a_{j+1}|x_{j+1}}(w),\\ T^k_{a_k|x_k}\cdots T^{j+1}_{a_{j+1}|x_{j+1}}(v)}}\Big)_{\substack{(a_{j+1},\dots,a_k,x'_j,x_j,\dots,x_k,\\ w,v,r,s)\in I^n_{j-\text{ons}}}} \end{pmatrix}. \tag{12}$$

Here, the top I^n_{Han} block of the output vector corresponds to the Hankel condition and the bottom $I^n_{j-\text{ons}}$ block corresponds to the operationally-non-signaling constraints for every party j. The norm of the vector $\mathcal{E}^n_k(H)$ can then serve as a measure of the solution's "almost-ness" as formalized in Theorem 4.2.

Example 4.1. *We give an example for the tripartite case with $k=3$ with an efficient strategy S_3. In the notation of Section 3.2, the weakly non-signaling conditions for Alice and Bob are, respectively,*

$$\begin{aligned} |\sum_a \sigma^\lambda(P^*T_{a|x}(U)Q) - \sum_a \sigma^\lambda(P^*T_{a|x'}(U)Q)| &\le \mathrm{negl}_{S_3,P,U,Q}(\lambda), \\ |\sum_b \sigma^\lambda\circ T_{a|x}(R^*T_{b|y}(O)U) - \sum_b \sigma^\lambda\circ T_{a|x}(R^*T_{b|y'}(O)U)| &\le \mathrm{negl}_{S_3,R,O,U}(\lambda), \end{aligned}$$

for all operators $P,Q\in\mathcal{A}_{ABC}$, $R,U\in\mathcal{A}_{BC}$, and $O\in\mathcal{A}_C$. The constraint-testing map $\mathcal{E}^n_3$ is

$$\mathcal{E}^n_3 : \mathrm{Herm}(\mathcal{W}^n_{ABC}) \to \mathbb{C}^{|I^n_{\text{Han}}|} \oplus \mathbb{C}^{|I^n_{A-\text{ons}}|} \oplus \mathbb{C}^{|I^n_{B-\text{ons}}|}$$

$$H \mapsto \begin{pmatrix} (H_{w,v} - H_{w',v'})_{(w,v,w',v')\in I^n_{\text{Han}}} \\ \Big(H\Big(\sum_a \big(T_{a|x}(r^*s) - T_{a|x'}(r^*s)\big)\Big)_{w,v}\Big)_{(x,x',w,v,r,s)\in I^n_{A-\text{ons}},} \\ \Big(H\Big(T_{a|x}\big(\sum_b T_{b|y}(t^*u) - T_{b|y'}(t^*u)\big)\Big)_{T_{a|x}(r),T_{a|x}(s)}\Big)_{(y,y',a,x,r,s,t,u)\in I^n_{B-\text{ons}}} \end{pmatrix},$$

where

$$I^n_{A-\text{ons}} = \{(x,x',w,v,r,s) \mid \forall x,x',\ w,v\in\mathcal{W}^n_{ABC},\ r,s\in\mathcal{W}^n_{BC},\ \deg w+\deg v+\deg r+\deg s\le 2n\}$$

is the set of tuples defining Alice's operationally-non-signaling constraints and

$$I^n_{B-\text{ons}} = \{(y,y',a,x,r,s,t,u) \mid \forall y,y',a,x,\ r,s\in\mathcal{W}^n_{BC},\ t,u\in\mathcal{W}^n_C,\ \deg r+\deg s+\deg t+\deg u\le 2n\}$$

is the corresponding set for Bob's.

4.2 Geometric proofs for quantitative quantum soundness

We first formalize the notation of "almost feasibility" of the compiled moment matrix $\Gamma_{\mathrm{comp}}^{(n),\lambda}$ as follows.

Lemma 4.2. *For the compiled moment matrix $\Gamma_{\mathrm{comp}}^{(n),\lambda}$ extracted from an efficient (quantum polynomial-time) strategy $S = (S_\lambda)$ for a k-partite compiled nonlocal game, there exists a negligible function $\mathrm{negl}_{S,n}^{\mathrm{lem}}(\lambda)$, dependent on S, n, such that*

$$\left\| \mathcal{E}_k^n(\Gamma_{\mathrm{comp}}^{(n),\lambda}) \right\|_2 \leq \mathrm{negl}_{S,n}^{\mathrm{lem}}(\lambda),$$

where $\|\cdot\|_2$ denotes the usual Euclidean norm.

Proof. Clearly $(\Gamma_{\mathrm{comp}}^{(n),\lambda})_{w,v} - (\Gamma_{\mathrm{comp}}^{(n),\lambda})_{w',v'} = 0$ if $w^*v = w'^*v'$ by its definition. On the other hand, by Eq. (11),

$$\begin{aligned}
&\Big| \sum_{a_j} \Gamma_{\mathrm{comp}}^{(n),\lambda} \Big(T^k_{a_k|x_k} \cdots T^{j+1}_{a_{j+1}|x_{j+1}} \big(T^j_{a_j|x_j}(r^*s) - T^j_{a_j|x'_j}(r^*s) \big) \Big)_{\substack{T^k_{a_k|x_k} \cdots T^{j+1}_{a_{j+1}|x_{j+1}}(w), \\ T^k_{a_k|x_k} \cdots T^{j+1}_{a_{j+1}|x_{j+1}}(v)}} \Big| \\
&= \left| \sum_{a_j} \sigma^\lambda \circ T^k_{a_k|x_k} \cdots T^{j+1}_{a_{j+1}|x_{j+1}} \big(w^* T^j_{a_j|x_j}(r^*s) v \big) - \sum_{a_j} \sigma^\lambda \circ T^k_{a_k|x_k} \cdots T^{j+1}_{a_{j+1}|x_{j+1}} \big(w^* T^j_{a_j|x'_j}(r^*s) v \big) \right| \\
&\leq \mathrm{negl}_{S,w,r,s,v}(\lambda) \leq \max_{w,r,s,v} \mathrm{negl}_{S,w,r,s,v}(\lambda) := \mathrm{negl}'_{S,n}(\lambda),
\end{aligned}$$

where the maximum makes sense because the degree n word sets $\mathcal{W}^n_{[j-1]}, \mathcal{W}^n_{[j]}$ are finite. Then, as the I^n_{Han}-block is already 0 for $\mathcal{E}_k^n(\Gamma_{\mathrm{comp}}^{(n),\lambda})$, one has

$$\left\| \mathcal{E}_k^n(\Gamma_{\mathrm{comp}}^{(n),\lambda}) \right\|_2 \leq \sqrt{\sum_j \left| I^n_{j-\mathrm{ons}} \right|} \cdot \mathrm{negl}'_{S,n}(\lambda) := \mathrm{negl}_{S,n}^{\mathrm{lem}}(\lambda).$$

□

Now, we are ready for the proof of our main technical result as illustrated by Fig. 2.

Theorem 4.3. *Let $\mathcal{G}$ be a k-partite nonlocal game with $\mathcal{G}_{\mathrm{comp}}$ as its compiled version. Let $S = (S_\lambda)_\lambda$ be an arbitrary efficient (quantum polynomial-time) strategy employed by the prover, and consider the corresponding algebraic compiled strategy state $\sigma^\lambda : \mathcal{A}_{[k]} \to \mathbb{C}$ due to [Bar+25]. For every $n \in \mathbb{N}$, let $(\Gamma_{\mathrm{comp}}^{(n),\lambda})_{w,v} := \sigma^\lambda(w^*v)$ be the associated level-n moment matrix.*

Then, there exists a (S,n)-dependent negligible function $\mathrm{negl}_{S,n}(\lambda)$ (goes to zero faster than the reciprocal of any polynomial in λ) and a feasible solution $\Gamma^{(n),\lambda}$ of Eq. (10) such that

$$\left\| \Gamma_{\mathrm{comp}}^{(n),\lambda} - \Gamma^{(n),\lambda} \right\|_{\mathrm{op}} \leq \mathrm{negl}_{S,n}(\lambda),$$

where $\|\cdot\|_{\mathrm{op}}$ denotes the operator spectral norm.

Proof. We aim to construct a feasible solution $\Gamma^{(n),\lambda}$ that is negligibly close to the compiled moment matrix $\Gamma_{\mathrm{comp}}^{(n),\lambda}$, see Fig. 2 for a pictorial illustration. To this end, consider the projection onto $\ker(\mathcal{E}_k^n)$

$$\Pi := \mathbb{1} - (\mathcal{E}_k^n)^\dagger \mathcal{E}_k^n,$$

where $(\mathcal{E}_k^n)^\dagger$ denotes the Moore-Penrose pseudo-inverse of $\mathcal{E}_k^n$ (see e.g., [GV13]). Thus, the projection of the compiled moment matrix

$$\Gamma_1^{(n),\lambda} := \Pi(\Gamma_{\text{comp}}^{(n),\lambda})$$

satisfies all linear constraints of Eq. (10) and

$$\begin{aligned}\left\|\Gamma_1^{(n),\lambda} - \Gamma_{\text{comp}}^{(n),\lambda}\right\|_{\text{op}} &= \left\|(\mathcal{E}_k^n)^\dagger \mathcal{E}_k^n(\Gamma_{\text{comp}}^{(n),\lambda})\right\|_{\text{op}} \leq \left\|(\mathcal{E}_k^n)^\dagger\right\|_{\text{op}} \left\|\mathcal{E}_k^n(\Gamma_{\text{comp}}^{(n),\lambda})\right\|_2 \\ &\leq \left\|(\mathcal{E}_k^n)^\dagger\right\|_{\text{op}} \text{negl}_{S,n}^{\text{lem}}(\lambda) := \text{negl}'_{S,n}(\lambda),\end{aligned} \tag{13}$$

where we use Theorem 4.2 and absorb the bounded constant $\left\|(\mathcal{E}_k^n)^\dagger\right\|_{\text{op}}$ into the negligible function.

While $\Gamma_1^{(n),\lambda}$ satisfies almost all constraints of Eq. (10), it is no longer normalized nor necessarily positive semidefinite. Nonetheless, by the fact that $\Gamma_{\text{comp}}^{(n),\lambda} \succeq 0$ and Eq. (13), Weyl's inequality [Wey12] implies that the minimum eigenvalue $\mu_0(\Gamma_1^{(n),\lambda})$ of $\Gamma_1^{(n),\lambda}$ satisfies

$$\mu_0(\Gamma_1^{(n),\lambda}) \geq -\,\text{negl}'_{S,n}(\lambda), \tag{14}$$

i.e., $\Gamma_1^{(n),\lambda}$ is almost positive semidefinite. Furthermore, normalization is almost fulfilled in the sense that

$$\begin{aligned}\left|(\Gamma_1^{(n),\lambda})_{\mathbb{1},\mathbb{1}} - 1\right| &= \left|(\Gamma_1^{(n),\lambda})_{\mathbb{1},\mathbb{1}} - (\Gamma_{\text{comp}}^{(n),\lambda})_{\mathbb{1},\mathbb{1}}\right| \leq \left\|\Gamma_1^{(n),\lambda} - \Gamma_{\text{comp}}^{(n),\lambda}\right\|_{\max} \\ &\leq \left\|\Gamma_1^{(n),\lambda} - \Gamma_{\text{comp}}^{(n),\lambda}\right\|_{\text{op}} \leq \text{negl}'_{S,n}(\lambda),\end{aligned} \tag{15}$$

where $\|\cdot\|_{\max}$ denotes the max norm and $\|\cdot\|_{\max} \leq \|\cdot\|_{\text{op}}$ is immediate. The last observation is that $\left\|\Gamma_1^{(n),\lambda}\right\|_{\text{op}} \leq \left\|\Gamma_{\text{comp}}^{(n),\lambda}\right\|_{\text{op}} \leq \left|\mathcal{W}_{[k]}^n\right|$, independent of λ, due to the contractivity of Π.

Next, consider the strictly feasible solution $\Gamma_{\text{strict}}^{(n)}$ of Theorem 3.9 and denote its minimal eigenvalue by $\mu_n > 0$. With a convex combination of $\Gamma_{\text{strict}}^{(n)}$ and $\Gamma_1^{(n),\lambda}$, define

$$\Gamma_2^{(n),\lambda} := \frac{\mu_n}{\mu_n + \text{negl}'_{S,n}(\lambda)} \Gamma_1^{(n),\lambda} + \frac{\text{negl}'_{S,n}(\lambda)}{\mu_n + \text{negl}'_{S,n}(\lambda)} \Gamma_{\text{strict}}^{(n)}.$$

It follows from Eq. (14) and Weyl's inequality [Wey12] that the minimal eigenvalue $\mu_0(\Gamma_2^{(n),\lambda})$ of $\Gamma_2^{(n),\lambda}$ satisfies

$$\mu_0(\Gamma_2^{(n),\lambda}) \geq \frac{\mu_n}{\mu_n + \text{negl}'_{S,n}(\lambda)} \underbrace{\mu_0(\Gamma_1^{(n),\lambda})}_{\geq -\,\text{negl}'_{S,n}(\lambda)} + \frac{\text{negl}'_{S,n}(\lambda)}{\mu_n + \text{negl}'_{S,n}(\lambda)} \underbrace{\mu_0(\Gamma_{\text{strict}}^{(n)})}_{=\mu_n} \geq 0,$$

consequently $\Gamma_2^{(n),\lambda} \succeq 0$. Thus, the matrix $\Gamma_2^{(n),\lambda}$ satisfies all constraints of Eq. (10) except for the normalization, because of the convexity of the constraint set. Moreover,

$$\left\|\Gamma_1^{(n),\lambda} - \Gamma_2^{(n),\lambda}\right\|_{\text{op}} = \frac{\text{negl}'_{S,n}(\lambda)}{\mu_n + \text{negl}'_{S,n}(\lambda)} \left\|\Gamma_1^{(n),\lambda} - \Gamma_{\text{strict}}^{(n)}\right\|_{\text{op}} \leq \text{negl}'_{S,n}(\lambda) \left\|\Gamma_1^{(n),\lambda} - \Gamma_{\text{strict}}^{(n)}\right\|_{\text{op}} := \text{negl}''_{S,n}(\lambda), \tag{16}$$

where the constant $\left\|\Gamma_1^{(n),\lambda} - \Gamma_{\text{strict}}^{(n)}\right\|_{\text{op}}$ is absorbed into the negligible function, since both $\left\|\Gamma_1^{(n),\lambda}\right\|_{\text{op}}$ and $\left\|\Gamma_{\text{strict}}^{(n)}\right\|_{\text{op}}$ are upper bounded by $\left|\mathcal{W}_{[k]}^n\right|$ that is independent of λ. Note that by the definition of $\Gamma_2^{(n),\lambda}$, $\mu_n > 0$, $\text{negl}'_{S,n}(\lambda) \geq 0$, and Eq. (15),

$$\left|(\Gamma_2^{(n),\lambda})_{\mathbb{1},\mathbb{1}} - 1\right| = \frac{\mu_n}{\mu_n + \text{negl}'_{S,n}(\lambda)} \left|(\Gamma_1^{(n),\lambda})_{\mathbb{1},\mathbb{1}} - 1\right| \leq \frac{\mu_n \, \text{negl}'_{S,n}(\lambda)}{\mu_n + \text{negl}'_{S,n}(\lambda)} \leq \text{negl}'_{S,n}(\lambda). \tag{17}$$

Finally, we normalize $\Gamma_2^{(n),\lambda}$ and obtain

$$\Gamma^{(n),\lambda} := \Gamma_2^{(n),\lambda} / (\Gamma_2^{(n),\lambda})_{\mathbb{1},\mathbb{1}},$$

which is a proper feasible solution of Eq. (10) since all constraints (that are not normalization) are homogeneous and thus preserved by this renormalization. It follows from Eqs. (13), (16) and (17) that

$$\begin{aligned}
\left\|\Gamma_{\text{comp}}^{(n),\lambda} - \Gamma^{(n),\lambda}\right\|_{\text{op}} &\leq \left\|\Gamma_{\text{comp}}^{(n),\lambda} - \Gamma_1^{(n),\lambda}\right\|_{\text{op}} + \left\|\Gamma_1^{(n),\lambda} - \Gamma_2^{(n),\lambda}\right\|_{\text{op}} + \left\|\Gamma_2^{(n),\lambda} - \Gamma^{(n),\lambda}\right\|_{\text{op}} \\
&\leq \text{negl}'_{S,n}(\lambda) + \text{negl}''_{S,n}(\lambda) + \left|1 - \frac{1}{(\Gamma_2^{(n),\lambda})_{\mathbb{1},\mathbb{1}}}\right| \left\|\Gamma_2^{(n),\lambda}\right\|_{\text{op}} \\
&\leq \text{negl}'_{S,n}(\lambda) + \text{negl}''_{S,n}(\lambda) + \left|\frac{\left\|\Gamma_2^{(n),\lambda}\right\|_{\text{op}}}{(\Gamma_2^{(n),\lambda})_{\mathbb{1},\mathbb{1}}}\right| \text{negl}'_{S,n}(\lambda) := \text{negl}_{S,n}(\lambda),
\end{aligned}$$

where $\text{negl}_{S,n}(\lambda)$ is a negligible function by the fact that $\left\|\Gamma_2^{(n),\lambda}\right\|_{\text{op}} / \left|(\Gamma_2^{(n),\lambda})_{\mathbb{1},\mathbb{1}}\right|$ is a bounded constant. □

A direct consequence of the existence and closeness statement from Theorem 4.3 is the following quantitative quantum soundness statement for all multipartite compiled nonlocal games.

Corollary 4.4. *Let $\mathcal{G}$ be a k-partite nonlocal game with $\mathcal{G}_{\text{comp}}$ as its compiled version. Let $S = (S_\lambda)_\lambda$ be an arbitrary efficient (quantum polynomial-time) strategy employed by the prover. Then for every $n \geq 1$, there exists a negligible function $\text{negl}_{S,n}(\lambda)$ (dependent on the QHE scheme, the strategy S, and the NPA level n) such that*

$$\omega_\lambda(\mathcal{G}_{\text{comp}}, S) \leq \omega_{k\text{seqNPA}}^n(\mathcal{G}) + \text{negl}_{S,n}(\lambda), \tag{18}$$

where $\omega_\lambda(\mathcal{G}_{\text{comp}}, S)$ is the prover's Bell score using S and $\omega_{k\text{seqNPA}}^n(\mathcal{G})$ is the optimal value of the hierarchy Eq. (10) at level n.

Furthermore, if the game $\mathcal{G}$ admits a finite-dimensional optimal strategy, then there exists a negligible function $\text{negl}_S(\lambda)$ (dependent on the QHE scheme, the strategy S) such that

$$\omega_\lambda(\mathcal{G}_{\text{comp}}, S) \leq \omega_{\text{q}}(\mathcal{G}) + \text{negl}_S(\lambda), \tag{19}$$

where $\omega_{\text{q}}(\mathcal{G})$ is the optimal tensor product quantum score.

Proof. Let $\sigma^\lambda : \mathcal{A}_{[k]} \to \mathbb{C}$ be the corresponding algebraic compiled strategy state due to [Bar+25].

For each $n \geq 1$, consider the associated compiled moment matrix

$$(\Gamma^{(n),\lambda}_{\text{comp}})_{w,v} := \sigma^\lambda(w^* v) \succeq 0$$

for all $w, v \in \mathcal{W}^n_{[k]}$.

By Theorem 4.3, there exists an n-dependent negligible function $\text{negl}'_{S,n}(\lambda)$ and a feasible solution $\Gamma^{(n),\lambda}$ for every λ of the k-partite sequential NPA hierarchy of Eq. (10) such that

$$\left\| \Gamma^{(n),\lambda}_{\text{comp}} - \Gamma^{(n),\lambda} \right\|_{\text{op}} \leq \text{negl}'_{S,n}(\lambda).$$

Then,

$$\begin{aligned}
\omega_\lambda(\mathcal{G}_{\text{comp}}, S) = \sigma^\lambda(\beta) &= \sum_{a_{[k]}, x_{[k]}} \beta_{a_{[k]} x_{[k]}} (\Gamma^{(n),\lambda}_{\text{comp}})_{\mathbb{1}, f_{a_{[k]}|x_{[k]}}} \\
&= \underbrace{\sum_{a_{[k]}, x_{[k]}} \beta_{a_{[k]} x_{[k]}} \Gamma^{(n),\lambda}_{\mathbb{1}, f_{a_{[k]}|x_{[k]}}}}_{\text{score of a possibly non-optimal feasible solution}} + \sum_{a_{[k]}, x_{[k]}} \beta_{a_{[k]} x_{[k]}} \left(\underbrace{(\Gamma^{(n),\lambda}_{\text{comp}})_{\mathbb{1}, f_{a_{[k]}|x_{[k]}}} - \Gamma^{(n),\lambda}_{\mathbb{1}, f_{a_{[k]}|x_{[k]}}}}_{\text{bounded by the max norm}} \right) \\
&\leq \omega^n_{k\text{seqNPA}}(\mathcal{G}) + \underbrace{\left| \sum_{a_{[k]}, x_{[k]}} \beta_{a_{[k]} x_{[k]}} \right|}_{\text{game-related constant}} \left\| \Gamma^{(n),\lambda}_{\text{comp}} - \Gamma^{(n),\lambda} \right\|_{\max} \\
&\leq \omega^n_{k\text{seqNPA}}(\mathcal{G}) + \text{const} \cdot \left\| \Gamma^{(n),\lambda}_{\text{comp}} - \Gamma^{(n),\lambda} \right\|_{\text{op}} \\
&\leq \omega^n_{k\text{seqNPA}}(\mathcal{G}) + \text{const} \cdot \text{negl}'_{S,n}(\lambda) = \omega^n_{k\text{seqNPA}}(\mathcal{G}) + \text{negl}_{S,n}(\lambda),
\end{aligned}$$

where $\text{negl}_{S,n}(\lambda) := \text{const} \cdot \text{negl}'_{S,n}(\lambda)$ is again negligible.

Finally, if $\mathcal{G}$ admits a finite-dimensional optimal strategy, Theorem 3.8 implies the existence of some fixed level n' with some flat optimal solution such that $\omega^{n'}_{k\text{seqNPA}}(\mathcal{G}) = \omega_{\text{q}}(\mathcal{G})$. We are done by letting $\text{negl}_S(\lambda) := \text{negl}_{S,n'}(\lambda)$. □

Note that we recover [Kle+25, Theorem A] the quantitative quantum soundness statement for all bipartite compiled nonlocal games by setting $k = 2$, and the asymptotic results of [Bar+25] by letting $n \to \infty$ thanks to Theorem 3.6.

Acknowledgments

MB acknowledges funding from QuantEdu France, a state aid managed by the French National Research Agency for France 2030 with the reference ANR-22-CMAS-0001. IK was supported by the Slovenian Research and Innovation Agency program P1-0222 and grants J1-50002, N1-0217, J1-3004, J1-50001, J1-60011, J1-60025. Partially supported by the Fondation de l'École polytechnique as part of the Gaspard Monge Visiting Professor Program. IK thanks École Polytechnique and Inria for hospitality during the preparation of this manuscript. DL acknowledges support from the Quantum Advantage Pathfinder (QAP) research program within the UK's National Quantum Computing Center (NQCC). MOR, LT, and XX acknowledge funding by the ANR for the JCJC grant LINKS (ANR-23-CE47-0003) and T-ERC QNET (ANR-24-ERCS-0008), by INRIA and CIEDS in the

Action Exploratoire project DEPARTURE. MOR, IK, LT, and XX acknowledge support by the European Union's Horizon 2020 Research and Innovation Programme under QuantERA Grant Agreement no. 731473 and 101017733.

References

[Arv69] William B Arveson. "Subalgebras of C^*-algebras". In: *Acta Math.* 123 (1969), pp. 141–224. ISSN: 0001-5962. DOI: 10.1007/BF02392388.

[Bar+25] Matilde Baroni, Dominik Leichtle, Siniša Janković, and Ivan Šupić. *Bounding the asymptotic quantum value of all multipartite compiled non-local games*. 2025. arXiv: 2507.12408 [quant-ph]. URL: https://arxiv.org/abs/2507.12408.

[Bar+24] Matilde Baroni, Quoc-Huy Vu, Boris Bourdoncle, Eleni Diamanti, Damian Markham, and Ivan Šupić. "Quantum bounds for compiled XOR games and d-outcome CHSH games". In: *arXiv preprint arXiv:2403.05502* (2024). URL: https://arxiv.org/abs/2403.05502.

[Bel64] John S Bell. "On the Einstein Podolsky Rosen paradox". In: *Physics Physique Fizika* 1.3 (1964), p. 195.

[Bra18] Zvika Brakerski. "Quantum FHE (almost) as secure as classical". In: *Annual International Cryptology Conference*. Springer. 2018, pp. 67–95.

[Bru+14] Nicolas Brunner, Daniel Cavalcanti, Stefano Pironio, Valerio Scarani, and Stephanie Wehner. "Bell nonlocality". In: *Reviews of modern physics* 86.2 (2014), pp. 419–478.

[Cla+69] John F Clauser, Michael A Horne, Abner Shimony, and Richard A Holt. "Proposed experiment to test local hidden-variable theories". In: *Physical review letters* 23.15 (1969), p. 880.

[Cui+25] David Cui, Chirag Falor, Anand Natarajan, and Tina Zhang. *A convergent sum-of-squares hierarchy for compiled nonlocal games*. 2025. arXiv: 2507.17581 [quant-ph]. URL: https://arxiv.org/abs/2507.17581.

[Cui+24] David Cui, Giulio Malavolta, Arthur Mehta, Anand Natarajan, Connor Paddock, Simon Schmidt, Michael Walter, and Tina Zhang. "A Computational Tsirelson's Theorem for the Value of Compiled XOR Games". In: *arXiv preprint arXiv:2402.17301* (2024). URL: https://arxiv.org/abs/2402.17301.

[Giu+15] Marissa Giustina, Marijn AM Versteegh, Sören Wengerowsky, Johannes Handsteiner, Armin Hochrainer, Kevin Phelan, Fabian Steinlechner, Johannes Kofler, Jan-Åke Larsson, Carlos Abellán, et al. "Significant-loophole-free test of Bell's theorem with entangled photons". In: *Physical review letters* 115.25 (2015), p. 250401.

[GV13] Gene H. Golub and Charles F. Van Loan. *Matrix Computations - 4th Edition*. Philadelphia, PA: Johns Hopkins University Press, 2013. DOI: 10.1137/1.9781421407944. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781421407944. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781421407944.

[Hen+15] Bas Hensen, Hannes Bernien, Anaïs E Dréau, Andreas Reiserer, Norbert Kalb, Machiel S Blok, Just Ruitenberg, Raymond FL Vermeulen, Raymond N Schouten, Carlos Abellán, et al. "Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres". In: *Nature* 526.7575 (2015), pp. 682–686.

[HR19] Paweł Horodecki and Ravishankar Ramanathan. "The relativistic causality versus no-signaling paradigm for multi-party correlations". In: *Nature Communications* 10.1 (Apr. 2019). ISSN: 2041-1723. DOI: 10.1038/s41467-019-09505-2. URL: http://dx.doi.org/10.1038/s41467-019-09505-2.

[Ji+22] Zhengfeng Ji, Anand Natarajan, Thomas Vidick, John Wright, and Henry Yuen. *MIP*=RE*. 2022. arXiv: 2001.04383 [quant-ph]. URL: https://arxiv.org/abs/2001.04383.

[Kal+23] Yael Kalai, Alex Lombardi, Vinod Vaikuntanathan, and Lisa Yang. "Quantum Advantage from Any Non-Local Game". In: *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*. ACM. 2023, pp. 1617–1628. DOI: 10.1145/3564246.3585164. URL: https://dl.acm.org/doi/10.1145/3564246.3585164.

[KMP22] Igor Klep, Victor Magron, and Janez Povh. "Sparse noncommutative polynomial optimization". In: *Math. Program.* 193.2 (B) (2022), pp. 789–829. ISSN: 0025-5610. DOI: 10.1007/s10107-020-01610-1.

[Kle+25] Igor Klep, Connor Paddock, Marc-Olivier Renou, Simon Schmidt, Lucas Tendick, Xiangling Xu, and Yuming Zhao. *Quantitative Quantum Soundness for Bipartite Compiled Bell Games via the Sequential NPA Hierarchy*. 2025. arXiv: 2507.17006 [quant-ph]. URL: https://arxiv.org/abs/2507.17006.

[Kul+25] Alexander Kulpe, Giulio Malavolta, Connor Paddock, Simon Schmidt, and Michael Walter. "A bound on the quantum value of all compiled nonlocal games". In: *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*. 2025, pp. 222–233.

[Las01] Jean B Lasserre. "Global optimization with polynomials and the problem of moments". In: *SIAM Journal on optimization* 11.3 (2001), pp. 796–817.

[MW23] Victor Magron and Jie Wang. *Sparse polynomial optimization: theory and practice*. World Scientific, 2023.

[Mah20] Urmila Mahadev. "Classical homomorphic encryption for quantum circuits". In: *SIAM Journal on Computing* 52.6 (2020), FOCS 18–189.

[MPW24] Arthur Mehta, Connor Paddock, and Lewis Wooltorton. "Self-testing in the compiled setting via tilted-CHSH inequalities". In: *arXiv preprint arXiv:2406.04986* (2024). URL: https://arxiv.org/abs/2406.04986.

[Mey+24] Uta Isabella Meyer, Ivan Šupić, Frédéric Grosshans, and Damian Markham. "Self-Testing Graph States Permitting Bounded Classical Communication". In: *arXiv preprint arXiv:2404.03496* (2024).

[MB23] Gláucia Murta and Flavio Baccari. "Self-testing with dishonest parties and device-independent entanglement certification in quantum communication networks". In: *Physical Review Letters* 131.14 (2023), p. 140201.

[NZ23] Anand Natarajan and Tina Zhang. "Bounding the quantum value of compiled nonlocal games: from CHSH to BQP verification". In: *2023 IEEE 64th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2023, pp. 1342–1348.

[NPA08] Miguel Navascués, Stefano Pironio, and Antonio Acín. "A convergent hierarchy of semidefinite programs characterizing the set of quantum correlations". In: *New Journal of Physics* 10.7 (2008), p. 073013.

[Oza13] Narutaka Ozawa. "Tsirelson's problem and asymptotically commuting unitary matrices". In: *Journal of Mathematical Physics* 54.3 (Mar. 2013), p. 032202. ISSN: 0022-2488. DOI: 10.1063/1.4795391. eprint: https://pubs.aip.org/aip/jmp/article-pdf/doi/10.1063/1.4795391/16053470/032202_1_online.pdf. URL: https://doi.org/10.1063/1.4795391.

[Par03] Pablo A Parrilo. "Semidefinite programming relaxations for semialgebraic problems". In: *Mathematical programming* 96 (2003), pp. 293–320.

[Pau02] Vern Paulsen. *Completely Bounded Maps and Operator Algebras*. Vol. 78. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2002. ISBN: 9780511546631. DOI: 10.1007/978-0-08-092496-0.

[PNA10] Stefano Pironio, Miguel Navascués, and Antonio Acin. "Convergent relaxations of polynomial optimization problems with noncommuting variables". In: *SIAM Journal on Optimization* 20.5 (2010), pp. 2157–2180.

[Rag03] Maxim Raginsky. "Radon–Nikodym derivatives of quantum operations". In: *Journal of Mathematical Physics* 44.11 (Nov. 2003), pp. 5003–5020. ISSN: 1089-7658. DOI: 10.1063/1.1615697. URL: http://dx.doi.org/10.1063/1.1615697.

[Sai+15] Ana Belén Sainz, Nicolas Brunner, Daniel Cavalcanti, Paul Skrzypczyk, and Tamás Vértesi. "Postquantum steering". In: *Physical review letters* 115.19 (2015), p. 190403.

[Sca12] Valerio Scarani. "The device-independent outlook on quantum physics". In: *Acta Physica Slovaca* 62.4 (2012), pp. 347–409.

[SW08] Volkher B Scholz and Reinhard F Werner. "Tsirelson's problem". In: *arXiv preprint arXiv:0812.4305* (2008). URL: https://arxiv.org/abs/0812.4305.

[Sve87] George Svetlichny. "Distinguishing three-body from two-body nonseparability by a Bell-type inequality". In: *Phys. Rev. D* 35 (10 May 1987), pp. 3066–3069. DOI: 10.1103/PhysRevD.35.3066. URL: https://link.aps.org/doi/10.1103/PhysRevD.35.3066.

[Tav+24] Armin Tavakoli, Alejandro Pozas-Kerstjens, Peter Brown, and Mateus Araújo. "Semidefinite programming relaxations for quantum correlations". In: *Reviews of Modern Physics* 96.4 (Dec. 2024). ISSN: 1539-0756. DOI: 10.1103/revmodphys.96.045006. URL: http://dx.doi.org/10.1103/RevModPhys.96.045006.

[Wey12] Hermann Weyl. "Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung)". In: *Mathematische Annalen* 71.4 (1912), pp. 441–479.

[XRK25] Xiangling Xu, Marc-Olivier Renou, and Igor Klep. "Quantitative Tsirelson's Theorems via Approximate Schur's Lemma and Probabilistic Stampfli's Theorems". In: *arXiv preprint arXiv:2505.22309* (2025). URL: https://arxiv.org/abs/2505.22309.

A Appendix

We give a brief comparison of the two related previous works [Bar+25; Kle+25].

A.1 The bipartite sequential NPA hierarchy of [Kle+25]

As the first attempt to bridge the gap between the standard NPA hierarchy and the sequential protocols central to this work (Fig. 1), the authors of [Kle+25] introduced a sequential variant of

the NPA hierarchy for *bipartite* games. Different from our Heisenberg picture inspired hierarchy in Section 3, they model the scenario from the *Schrödinger picture*: after Alice receives question x and produces answer a, the shared state σ collapses into a new, subnormalized state $\sigma_{a|x}$ for Bob, that can now perform a measurement y and obtain an output b. Instead of a single moment matrix representing the global pre-measurement state, the hierarchy is defined via a collection of moment matrices labeled by the actions of Alice $\Theta^{(n)}(a|x)$, which are indexed by words built from Bob's operators $f_{b|y}$.

The constraint of operationally-non-signaling translates to this family of moment matrices as follows.

Definition A.1 (Bipartite sequential NPA hierarchy [Kle+25])**.** *For a bipartite game $\mathcal{G}$, the level-n sequential NPA relaxation is the solution to the following SDP, defined in terms of a collection of subnormalized moment matrices $\Theta^{(n)}(a|x)$ indexed by words of length $\leq n$ in letters $\{f_{b|y}\}$:*

$$\begin{aligned}
\omega^n_{\text{seqNPA}}(\mathcal{G}) \quad &= \max_{\Theta^{(n)}(a|x)} \sum_{a,b,x,y} \beta_{abxy} \Theta^{(n)}(a|x)_{1,f_{b|y}} \\
s.t. \quad & \Theta^{(n)}(a|x) \succeq 0, \quad \forall a, x, \\
& \sum_a \Theta^{(n)}(a|x) = \sum_a \Theta^{(n)}(a|x') := \Theta^{(n)} \quad \forall x, x' \quad \textit{(operationally-non-signaling condition)} \\
& 1 = \Theta^{(n)}_{\mathbb{1},\mathbb{1}} \qquad \textit{(normalization).}
\end{aligned} \tag{20}$$

This hierarchy effectively characterizes the set of bipartite sequential quantum correlations and the commuting observable quantum correlations as $n \to \infty$, and possesses several key properties:

(a) *Soundness and monotone convergence.* Like the standard NPA hierarchy from Theorem 2.5, it is sound,

$$\omega_{\text{qc}}(\mathcal{G}) \leq \omega^n_{\text{seqNPA}}(\mathcal{G}),$$

and converges to the commuting-operator value monotonically,

$$\omega^1_{\text{seqNPA}}(\mathcal{G}) \geq \omega^2_{\text{seqNPA}}(\mathcal{G}) \geq \cdots \searrow \omega_{\text{qc}}(\mathcal{G}).$$

(b) *Relation to standard NPA.* At any finite level n, it is equivalent to a relaxed version of the standard NPA hierarchy where Alice's operators $f_{a|x}$ do not satisfy the POVM completeness condition: $\sum_a f_{a|x} \neq \mathbb{1}$. It only need to appear to be complete when testing with length $\leq n$ words of Bob's operators $f_{b|y}$.

(c) *Duality.* Its conic dual problem is equivalent to a sparse sum-of-squares (SOS) hierarchy [KMP22; MW23]. Similar to its dual counterpart, the bipartite sequential NPA hierarchy can be less computationally demanding, as the moment matrices $\Theta^{(n)}(a|x)$ indexed by only $f_{b|y}$ are much smaller than the standard bipartite NPA moment matrices.

However, as discussed in Section 1, this Schrödinger-picture construction based on post-measurement states is *inherently bipartite*. It does not naturally extend to the multipartite setting, as tracking the sequence of post-measurement states loses too much of the necessary algebraic structure.

A.2 The multipartite algebraic framework of [Bar+25]

In order to extend the operator-algebraic techniques of [Kul+25] from the bipartite to the multipartite setting, the authors of [Bar+25] propose a composable algebraic structure that is very well-suited to characterize sequential players. For the notation convenience, let us focus on the tripartite scenarios. Their key idea is the new concept of universal C*-algebras of sequential projective measurements, which generalizes the classical universal algebra of static measurements, i.e., what before we were referring to as the "symbols" for Bob $f_{b|y}$. More concretely, let us consider the non-trivial case of three players in a sequence. The action of Alice is modeled as an algebraic state $\varphi_{a|x}$, Bob performs a transformation and in the end Charlie measures; or equivalently in the Heisenberg picture, Charlie performs a measurement and Bob implements a transformation that is pulling back Bob's measurement to the space in which Alice's state is defined. In this picture, Charlie and Bob jointly perform a measurement, whose only constraint is that Bob acts before Charlie, hence the latter cannot signal to the first, while the contrary is allowed.

In more detail, [Bar+25] define the universal C*-algebra $\mathcal{A}_{B\to C}$ of Bob's and Charlie's sequential measurements using the following relations on its generators $\{f_{bc|yz}\}$:

$$f^*_{bc|yz} = f_{bc|yz}, \quad f_{bc|yz} f_{b'c'|yz} = \delta_{b,b'}\delta_{c,c'} f_{bc|yz}, \quad \sum_{b,c} f_{bc|yz} = \mathbb{1}, \quad \sum_{c} f_{bc|yz} = \sum_{c} f_{bc|yz'}, \ \forall z, z'.$$

Note that $\mathcal{A}_{B\to C}$ is precisely $\mathcal{A}_{BC}$ that we introduce in Section 3.2. Letting $\mathcal{A}_C$ be the universal C*-algebra of Charlie's measurements, there exists then a family of *-homomorphisms $T_{b|y} : \mathcal{A}_C \to \mathcal{A}_{B\to C}$, taking generators to generators naturally in the following way

$$T_{b|y}(f_{c|z}) = f_{bc|yz}.$$

Note, that by construction, summing over b yields unital *-homomorphisms $T_y = \sum_b T_{b|y}$. In this way, the maps $T_{b|y}$ can be thought of as embeddings of $\mathcal{A}_C$ into subalgebras of $\mathcal{A}_{B\to C}$ labeled by b and y.

The power of this approach lies in two important features:

1. *Composability*, it is very immediate to understand how to increase the number of sequential players: it is sufficient to add the labels and the additional "causal" constraint;

2. It provides a very *compact framework*, because all of the players strategies are defined on a single big algebra, with a rich internal structure that retrieves the action of the single players, which is captured by *-homomorphisms.