Quantum Nonlocality under Latency Constraints

Dawei Ding*[1,2,3,4], Zhengfeng Ji†[5,6], Pierre Pocreau[7,8], Mingze Xu[9,10], and Xinyu Xu[1,3,4]

[1]Yau Mathematical Sciences Center, Tsinghua University, Beijing 100084, China
[2]Beijing Institute of Mathematical Sciences and Applications, Beijing, 101408, China
[3]Shanghai Institute for Mathematics and Interdisciplinary Sciences, Shanghai, 200433, China
[4]Center for Mathematics and Interdisciplinary Sciences, Fudan University, Shanghai, 200433, China
[5]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[6]Zhongguancun Laboratory, Beijing 100094, China
[7]Inria, Université Grenoble Alpes, 38000 Grenoble, France
[8]CNRS, Grenoble INP, LIG, Université Grenoble Alpes, 38000 Grenoble, France
[9]Zhili College, Tsinghua University, Beijing 100084, China
[10]Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Urbana, IL 61801, USA

Abstract

Bell inequality violation is the phenomenon where multiple non-communicating parties can exhibit correlations using quantum resources that are impossible if they can only use classical resources. One way to enforce non-communication is to apply a latency constraint: the parties must all produce outputs after they receive their inputs within a time window shorter than the speed of light delay between any pair of parties. If this latency constraint is relaxed so that a subset of the parties can communicate, we can obtain a new set of inequalities on correlations that extends Bell inequalities in a very natural way. Moreover, with this relaxed latency constraint, we can also have quantum communication between a subset of parties and thereby achieve possible quantum violations of these new inequalities. We ultimately wish to answer the fundamental question: "What are the physically realizable correlations between multiple parties under varying latency constraints?" To answer this question, we introduce latency-constrained games, a mathematical framework that extends nonlocal games to the setting where a subset of parties can communicate. The notion of latency-constrained games can have real-world applications, including high frequency trading, distributed computing, computer architecture, and distributed control systems.

Contents

*dding@simis.cn

†jizhengfeng@tsinghua.edu.cn

1 Introduction

One of the most famous and counterintuitive phenomena in quantum mechanics is the violation of Bell inequalities [1]. This phenomenon involves multiple parties that are not in communication but can exhibit correlations[1] that cannot be achieved using classical resources. One method to enforce non-communication between parties is to impose a *latency constraint*: the parties must complete the entire inequality-violating process from receiving their inputs to producing their outputs within a time window shorter than the speed of light delay between any pair of parties, making communication impossible. That is, because of the finite speed of light, by the theory of relativity *latency constraints translate into non-communication constraints.* This is also referred to as closing the locality loophole in experimental literature and has been demonstrated in early experiments [2]. Therefore, one interpretation of Bell inequalities is that they represent *fundamental bounds on correlations between classical parties under very low latency constraints.*

One natural follow-up question is the following: what if the latency constraint is relaxed? In the two-party case, if the latency constraint is loose enough to allow communication, the parties can exhibit any correlation by simply exchanging their inputs and choosing outputs accordingly. This entire process is classical, and so the latency constraint does not lead to any bounds on the correlations that classical parties can achieve. However, the situation becomes nontrivial when we go beyond two parties. For example, consider three parties A, B, C arranged collinearly such that A and B are close together, while C is farther away, as shown in Figure 1. For simplicity we will

[1] This is also known as a **behavior**, a conditional probability distribution of the parties' outputs conditioned on their inputs.

assume all three parties receive their inputs simultaneously and must all produce their outputs by time t, the latency constraint.

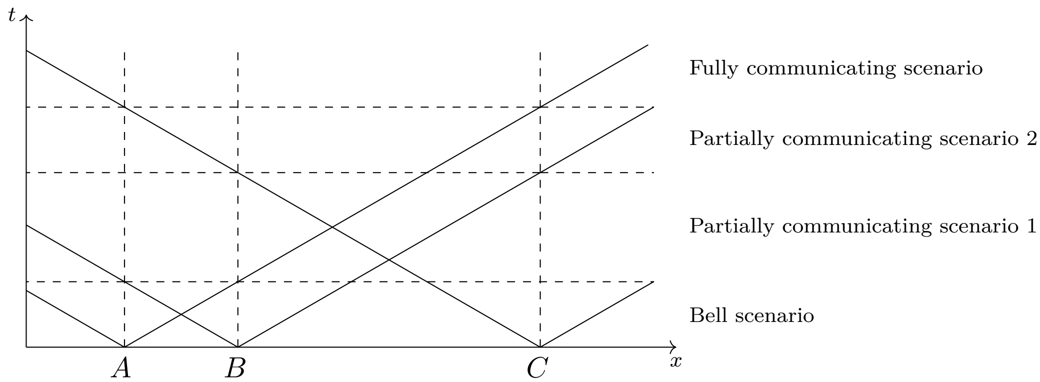


Figure 1: An illustration of the three parties A, B, C in the LC game that are collinear on the x-axis and their respective light cones. We assume the three parties receive their inputs simultaneously at time 0. They must produce their outputs within time t. Different latency constraints lead to inequivalent non-communication constraints.

Under the strictest latency constraints, all three parties cannot communicate and this becomes a regular Bell scenario with three parties. This is denoted by the lowest line in Figure 1. Now, consider slightly relaxing the latency constraint such that A and B can communicate with each other, but neither can communicate with C. This corresponds to the "Partially communicating scenario 1" in Figure 1. This is now a scenario that is *neither a Bell scenario where no parties can communicate nor the trivial, fully communicating scenario (no bounds on correlations)*. However, due to the latency constraint, there can still be bounds on the correlations exhibited by classical parties. For example, the CHSH inequalities between B and C and between A and C still need to be satisfied. However, bounds that holds in the no-communication Bell scenario, such as the CHSH inequality between A and B, can now be violated. By relaxing the latency constraint further, we can have the situation where A, B can communicate with each other and B, C can communicate with each other, but A, C cannot communicate. This is again an in-between scenario, corresponding to "Partially communicating scenario 2" in Figure 1. In this case, the CHSH inequality between B and C can now be violated, but the CHSH inequality between A and C still needs to be satisfied. Relaxing the latency constraint still further as in Figure 1 ends up with the trivial, fully communicating scenario.

This simple example reveals a new, natural area of research:

"What are the fundamental bounds on correlations between classical parties under *varying* latency constraints?"

In essence, we are *completing* the story of Bell inequalities if we are to interpret them as fundamental bounds on correlations under very low latency constraints. Furthermore, we ask:

"What is the maximum quantum violation of these bounds?"

As of this writing, the theory of quantum mechanics is the most general theory in describing physically realizable correlations between different parties. Hence, we are actually asking

"*What are the* <u>*physically realizable*</u> *correlations between multiple parties under varying latency constraints?*"

This one question underlies the central idea of our paper. That is, giving the parties all possible physical resources, what are the correlations they can exhibit under different latency constraints? This is a question of fundamental physical limits. Now, looser latency constraints imply that we can have strategies making use of quantum resources that are *more powerful than those of Bell scenarios.* Just as classical parties can send each other classical information and thereby violate Bell inequalities, quantum parties can send each other quantum information and possibly achieve stronger violations. We do not bound the amount of information exchanged during communication, but only impose latency constraints which dictate whether communication is possible or not. This necessitates a new mathematical framework that goes beyond that of the current theory of Bell nonlocality.

In this paper, we formulate an extension of the existing mathematical framework of **nonlocal games** from computer science that can be used to answer these questions. As our extension centers around the concept of latency constraints, we will use the name **latency-constrained (LC) games**. Our definition will mathematically formalize the concept of a latency constraint and the communication latency between different parties. We will define classical and quantum strategies as the most general operations that the parties can perform using classical and quantum resources, respectively, within the latency constraint. A crucial assumption we make throughout is that the parties' local operations consume zero time. In general, the local operation time is constrained by energy. By providing the parties with sufficient energy, the local operation time can be made arbitrarily short. This is analogous to the constraint that energy imposes on clock speed in digital computers. We will also consider the possibility of multiple rounds of communication and multiple inputs and outputs over time. We call the mathematical framework we introduce for this case **multi-step LC games**. This framework is more general and can be used to answer the question

"Given multiple parties with communication latency, what are the possible distributions over time-series of outputs they can produce, given a time-series of inputs?"

This is essentially an extension of the previous question to the multi-round, time-dependent setting. Furthermore, this framework uncovers a finer structure than what is drawn in Figure 1. That is, there is actually a regime that lies between the "Partially communicating scenario 1" and "Partially communicating scenario 2" where A, B can transmit to each other, receive that transmission, and then transmit to each other again. This we will call **back-and-forth communication**, as shown in Figure 2 for the same setup as Figure 1. This allows for quantum strategies where the two parties A, B can exchange inputs, perform a quantum measurement, and then exchange measurement results.

The concept of LC games also has real-world applications. As established in previous work [3, 4, 5], there exist real-world scenarios where we do wish to exhibit correlations between parties at timescales shorter than the speed of light delay. This equates to finding "real-world Bell scenarios". One notable example is in high frequency trading (HFT), where trading servers issue or cancel orders on timescales of microseconds and in the future even nanoseconds [6]. For stock exchanges separated by kilometers or more, this is already shorter than the speed of light delay [5]. Now, from the perspective of real-world scenarios, the concept of LC games is even more natural than Bell scenarios. For example, in HFT, the bound on correlations under a certain latency constraint can be identified as a bound on a figure of merit for an HFT strategy such as low risk or overall payoffs [3, 5] that

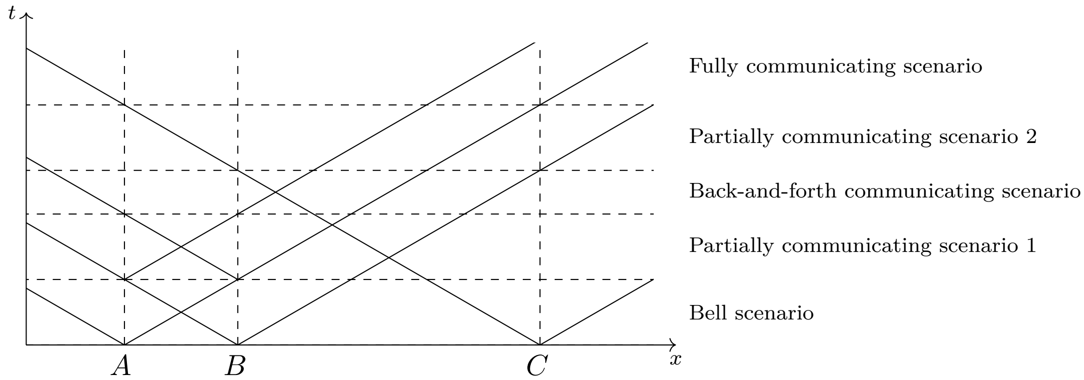


Figure 2: Parties A and B can have back-and-forth communication for a latency constraint that lies between those of the two partially communicating scenarios. Note that this is not an exhaustive list of all possible scenarios.

can be achieved within a certain period of time. Hence, our framework of LC games is, in this case, can be used to answer the question

"What is the lowest risk or highest payoff we can achieve within a certain latency constraint?"

Such a question is extremely relevant for HFT, where time is literally money. But in general, a separation between the answer to this question using classical or quantum resources can be interpreted as a mathematically provable *time advantage* that can be achieved by using quantum resources. That is, we can ask what is the *minimum time needed* to achieve a certain figure of merit, or utility [5], using classical or quantum resources. If the minimum time is shorter for quantum resources, this constitutes a provable quantum advantage in the time required. This applies to other real-world scenarios that can also be latency-sensitive to the point where speed of light delay is a concern, including (classical) distributed computing and computer architecture [5, 7]. Furthermore, as mentioned in [8, 4, 5], non-communication between parties can arise in real-world scenarios for reasons other than latency, such as a lack of a centralized control or the existence of privacy concerns, characteristics prevalent in distributed control systems [9, 10]. Our LC game framework naturally applies to these scenarios where there is centralized control or there are no privacy concerns for a subset of the parties. Lastly, our notion of multi-step LC games is very natural to real-world scenarios where parties are receiving a series of inputs and producing a series of outputs in real time.

Our paper is structured as follows. In Section 2 we first introduce the simplest type of LC game that is not just a conventional nonlocal game, where a subset of the parties can undergo a single round of communication. In Section 3 we give the definition of multi-step LC games, the most general latency-constrained game where parties can communicate, receive inputs, and give outputs over multiple rounds. In Section 4 we introduce numerical techniques that can be used to bound the maximum winning probability that a quantum strategy can achieve for this simplest type of LC game. This culminates in Figure 13 where we plot the classical and quantum values for an LC game *as a function of time*. We end in Section 5 with a discussion. In Section A, to be self-contained we give a short introduction to nonlocal games. In Section B, we conjecture a way to upper bound the quantum value for LC games. In Section C and Section D, we give numerical and analytic results on the quantum values of specific LC games.

Related work Overall, related works in the literature typically focus on generalizing local hidden-variable (LHV) models by allowing limited communication and seeing if quantum strategies (with no communication) can still rule out these stronger models. In particular, unlike our paper, *latency constraints is not a central concept in previous works.* Furthermore, we are not aware of any work that generalizes quantum strategies by allowing partial communication. There is a body of work concerning causal relationships and the types of quantum operations that can be performed [11, 12]. These works concentrate on what types of unitaries can be implemented, while our paper concerns what types of correlations can be exhibited.

Early work by Toner and Bacon [13] investigated the protocols which permit communication after the parties have chosen their measurements. This scenario was further investigated in Ref. [14]. Both works focused on the bipartite case with limits on the number of communicated bits. Subsequently, this scenario was extended to multipartite case and rigorously formalized using directed acyclic graphs (DAG) [15]. Refs. [16, 17, 18] analyzed one-way communication of inputs, outputs, or other classical messages in the bipartite setting using DAGs. It was shown that with three inputs and two outputs per party, one-way output communication alone is insufficient to reproduce nonlocal correlations. As for the multipartite case, the multistage game where the outcomes are announced one by one was discussed in Ref. [19].

Additionally, there are many works on Bell nonlocality in networks [20]. In these works, research typically focuses on setups with multiple independent quantum sources distributing entanglement across network nodes. Source-independence and no-signaling assumptions leads to constraints on classical and quantum strategies, which are used to derive network Bell inequalities [21] and certify full network nonlocality [22]. That is, this area of research is concerned with *weaker* classical and quantum strategies. In contrast, our work concerns *stronger* classical and quantum strategies justified by relaxing the latency constraint.

Communication between nodes is also allowed in genuine multipartite nonlocality (GMNL) [23] and the associated Svetlichny inequalities [24]. GMNL is defined as the set of correlations that cannot be expressed as a convex combination of correlations that are local with respect to any bipartition of the parties. For each bipartition, parties can communicate within their side of the partition but not with parties on the other side. However, these scenarios are very specific (communication allowed only within a group and everyone can communicate within the group) and are only analyzed up to deriving an inequality (a Svetlichny inequality) that cannot be violated with classical resources even if communication is allowed within the group. Quantum violations of this inequality are then analyzed with conventional quantum strategies for nonlocal games. In contrast, we consider dynamic quantum strategies involving quantum measurements and quantum communication between parties that naturally arises from the concept of a relaxed latency constraint. Such dynamic strategies do appear in [25]. However, our work makes different definitions and performs a comprehensive analysis based on our definitions:

- Latency constraints is main motivation for our new definitions, and so there is a clear physical interpretation of our mathematical framework. In particular, we allow for simultaneous two-way communication between a pair of parties.
- We consider the most general n-party scenario and mathematically formulate the concept of a latency constraint using a directed graph and an integer-valued latency function.
- We consider specific LC games and explicitly construct optimal quantum strategies.

- We introduce new numerical optimization algorithms for bounding winning probabilities achievable by quantum strategies and conduct numerical experiments on random games.

2 Latency-Constrained (LC) Games

We begin by introducing the simplest form of an LC game that extends beyond a conventional nonlocal game. This scenario corresponds to a physical setting in which the latency constraint permits a *single round of communication among a subset of the parties.* The pattern of allowed communication between parties can be represented by a directed graph G, where an edge (i, j) indicates that party i can send information to party j. For generality, we do not assume symmetry in communication—that is, if i can communicate with j, it does not imply that j can communicate with i. This is why we use directed graphs. We now formally define this generalization of a nonlocal game. For succinctness, we refer to this as a **latency-constrained game**, and use the term **multi-step latency-constrained game** to denote more complex settings where communication can go back and forth over multiple rounds. See Section 3 for more information.

We use n and $[n]$ to denote the number of parties and the set $\{1, 2, \ldots, n\}$ respectively.

Definition 1. *Let $n \geq 2$ be an integer, and let S_i and A_i be finite sets for each $i \in [n]$. An n-party **latency-constrained (LC) game** with input sets S_i and output sets A_i is defined by the tuple $(\mathcal{V}, \pi, G)$, where the probabilistic predicate $\mathcal{V}$ is a map*

$$\mathcal{V} : \prod_{i=1}^{n} A_i \times \prod_{i=1}^{n} S_i \to [0, 1],$$

*π is the input distribution over $\prod_{i=1}^{n} S_i$, and G is a directed graph with n vertices, referred to as the **connectivity graph** of the game.*

Note that a conventional nonlocal game is simply an LC game with G being the empty graph (no edges). Like nonlocal games, we can define **behavior** as a conditional probability distribution

$$p(a_1, a_2, \cdots, a_n | s_1, s_2, \cdots, s_n)$$

with corresponding **winning probability**

$$p_{\text{win}} \coloneqq \sum_{a_i \in A_i, s_i \in S_i} \pi(s_1, s_2, \cdots s_n) \cdot p(a_1, a_2, \cdots, a_n | s_1, s_2, \cdots, s_n) \mathcal{V}(a_1, a_2 \cdots a_n | s_1, s_2, \cdots s_n).$$

We will also use the term **maximum algebraic value** ω_a to mean the largest possible winning probability a behavior can attain:

$$\omega_a \coloneqq \sum_{s_i \in S_i} \pi(s_1, s_2, \cdots s_n) \cdot \max_{a_i \in A_i} \mathcal{V}(a_1, a_2 \cdots a_n | s_1, s_2, \cdots s_n).$$

We now proceed to define strategies for LC games.

2.1 Classical strategies

We want to understand what is the most general strategy using classical resources in the LC game setting. We first make some preliminary definitions. For a directed graph $G = (V, E)$, given $i \in V$, define

$$N_{\text{out}}(i) \coloneqq \{j \in V : (i, j) \in E\}$$

as the **out-neighborhood** of i and

$$N_{\text{in}}(i) \coloneqq \{j \in V : (j, i) \in E\}$$

as the **in-neighborhood** of i. We also define

$$N_{\text{out}}[i] \coloneqq N_{\text{out}}(i) \cup \{i\},\ N_{\text{in}}[i] \coloneqq N_{\text{in}}(i) \cup \{i\}$$

as the **closed out-neighborhood** and **closed in-neighborhood**, respectively.

Following the same logic as nonlocal games, we will define classical strategies by first considering the deterministic setting. When everything is deterministic, since classical information can be copied, *the only nontrivial information the parties can communicate are their inputs.* We thus define **deterministic strategies** as each party taking their local information and producing an output. But unlike nonlocal games, for LC games each party also has access to the inputs of his in-neighbors. We therefore have the following definition.

Definition 2. *Let $(\mathcal{V}, \pi, G)$ be an LC game with input and output sets S_i and A_i. Define the H_i,* ***past light cone of party i****, as*

$$H_i \coloneqq \prod_{j \in N_{\text{in}}[i]} S_j.$$

A ***deterministic strategy*** *of the game is given by functions $\{f_i\}_{i=1}^n$, where $f_i : H_i \to A_i$.*

The behavior realized by a deterministic strategy is simply the vector whose entry is indexed by $\mathbf{a}$ and $\mathbf{s}$ is

$$p(\mathbf{a}|\mathbf{s}) \coloneqq \prod_{i=1}^n \delta_{a_i, f_i(h_i)},$$

where $\mathbf{a} \in \prod_{i=1}^n A_i, \mathbf{s} \in \prod_{i=1}^n S_i$, and h_i is the element of H_i built from the corresponding elements in $\mathbf{s}$. We can compute the number of possible deterministic behaviors:

Proposition 3. *Given an LC game $(\mathcal{V}, \pi, G)$, the number of possible deterministic behaviors is*

$$\prod_{i=1}^n |A_i|^{|H_i|}. \tag{2.1}$$

Proof. Party i has $|H_i|$ possible input values, and for each input value, he can choose from $|A_i|$ possible output values, so the total number of possible deterministic behaviors is upper bounded by Equation (2.1). On the other hand, if the output differs for a certain input of a certain party, the resulting two behaviors will be different. Therefore, the number is exactly Equation (2.1). □

This result can be somewhat counter-intuitive since the parties can have overlapping inputs. However, what we are counting is the number of different possible deterministic strategies, not the number of different possible inputs.

Again following the same logic as nonlocal games, in the randomized setting, in addition to inputs, other nontrivial classical information are sources of randomness. However, this randomness could be shared with all parties before they start playing the LC game. Since we are assuming we can use all possible classical resources, we assume all sources of randomness are shared with all parties beforehand as an initial shared resource.[2] We thus define a **classical strategy** as a probabilistic mixture of deterministic strategies. That is, the set of all classical behaviors is by definition the convex hull of the set of deterministic behaviors. It is therefore a polytope with the number of vertices given in Equation (2.1). *The set of hyperplanes that defines this polytope via the Weyl-Minkowski theorem are then the generalization of Bell inequalities for varying latency constraints, in the case where a subset of parties can undergo a single round of communication.* This answers the first question posed in Section 1. We define the largest winning probability attainable by a classical strategy for an LC game as ω_c, which we call the **classical value**.

Also mentioned in Section 1 is that if every pair of parties can communicate with each other, then a classical strategy can realize any possible correlation. This is the "Fully communicating scenario" in Figure 1, which is defined by an LC game with connectivity graph being the complete directed graph (every pair of vertices are connected in both directions). We state this result explicitly in the general case:

Proposition 4. *Let $(\mathcal{V}, \pi, G)$ be an LC game where the connectivity graph G is a complete directed graph. Then, a classical strategy can realize all possible behaviors. In particular, it can achieve the maximum algebraic value ω_a.*

Proof. Let $p(\mathbf{a}|\mathbf{s})$ be a behavior. For every $\mathbf{s} \in \prod_{i=1}^{n} S_i$, the n parties have a corresponding shared randomness resource distributed according to the probability distribution $p(\mathbf{a}|\mathbf{s})$. Party i owns the a_i part of the shared randomness resource. Since G is a complete directed graph, every party has access to all of $\mathbf{s}$. Each party i then outputs a_i from their shared randomness resource corresponding to $\mathbf{s}$. The resulting behavior is exactly $p(\mathbf{a}|\mathbf{s})$.

In particular, the classical strategy can realize the behavior given by

$$p_0(\mathbf{a}|\mathbf{s}) := \delta_{\mathbf{a},\mathbf{a}_0(\mathbf{s})},$$

where

$$\mathbf{a}_0(\mathbf{s}) \in \operatorname{argmax}_{\mathbf{a}} \mathcal{V}(\mathbf{a}|\mathbf{s}).$$

This clearly achieves ω_a. □

We observe that the core mathematical object used to describe classical strategies for LC games *is exactly the same as that of conventional nonlocal games.* Both are simply a set of local discrete functions f_i. The only difference is that in LC games, the input distribution has some constraints, namely that overlapping inputs between different parties must be the same.

[2] This is the hidden variable λ in local hidden variable theories.

2.2 Quantum strategies

Quantum strategies for LC games, on the other hand, do *not* have the same core mathematical object as a quantum strategy for a nonlocal game. Like classical strategies, each party can transmit information to their out-neighbors with respect to the connectivity graph G. Unlike classical strategies, the parties can send *quantum information* to each other. We can describe such a strategy with the following definition.

Definition 5. *Let $(\mathcal{V}, \pi, G)$ be an LC game of n parties. Let B_i, $B_{i\to j}$ be Hilbert spaces for $i, j \in [n]$. A* ***quantum strategy*** *is a tuple*

$$(|\psi\rangle, \{W_i(s_i)\}_{i\in[n], s_i\in S_i}, \{M_i\}_{i=1}^n),$$

where $|\psi\rangle \in \bigotimes_{i=1}^n B_i$ is a quantum state,

$$W_i(s_i) : B_i \to B_i^{\text{out}} := \bigotimes_{j\in N_{\text{out}}[i]} B_{i\to j}.$$

is an isometry for all $s_i \in S_i$, and $M_i = \{\Pi_{i,a_i}\}_{a_i\in A_i}$ is a projective measurement on

$$B_i^{\text{in}} := \bigotimes_{j\in N_{\text{in}}[i]} B_{j\to i}.$$

In words, in a quantum strategy, the parties share a quantum state $|\psi\rangle$. Party i uses the isometry $W_i(s_i)$ which is dependent on his input s_i. The isometry maps his share of the quantum state in the Hilbert space B_i to a tensor product Hilbert space $B_i^{\text{out}} := \bigotimes_{j\in N_{\text{out}}[i]} B_{i\to j}$, where tensor factor $B_{i\to j}$ is the quantum system transmitted to party j. $B_{i\to i}$ is the quantum system that is left behind.[3] Hence, after the transmission, party i has an overall Hilbert space $B_i^{\text{in}} := \bigotimes_{j\in N_{\text{in}}[i]} B_{j\to i}$, on which he performs the measurement M_i. We can write the behavior realized by a quantum strategy for an LC game as

$$p(\mathbf{a}|\mathbf{s}) = \Big\langle \psi \,\Big|\, \bigotimes_{i=1}^n W_i(s_i)^\dagger \cdot \bigotimes_{i=1}^n \Pi_{i,a_i} \cdot \bigotimes_{i=1}^n W_i(s_i) \,\Big|\, \psi \Big\rangle. \tag{2.2}$$

Note that the notation is somewhat misleading as the tensor product structure of the measurements is *not* the same as that of the isometries. This is better explained by a figure: we draw a quantum strategy in Figure 3 for an LC game where the connectivity graph G is the graph

$$v_1 \leftrightarrows v_2 \leftrightarrows v_3.$$

We denote by ω_q the largest winning probability attainable by a quantum strategy for an LC game. We call this the **quantum value**.

Two remarks are worth making here about the quantum strategy of a graph game. First, we emphasize that *a quantum strategy for an LC game does not necessarily require the physical transmission of quantum systems in real time.* Instead, quantum teleportation—which only requires pre-established entanglement and real-time classical communication—is sufficient to turn quantum communications into classical ones.

[3] In general, the Hilbert spaces $B_{i\to j}$ can be input-dependent as well, but we can always enlarge the Hilbert space (by taking the direct sum over all inputs for example) so that in the end the isometries map to a fixed Hilbert space.

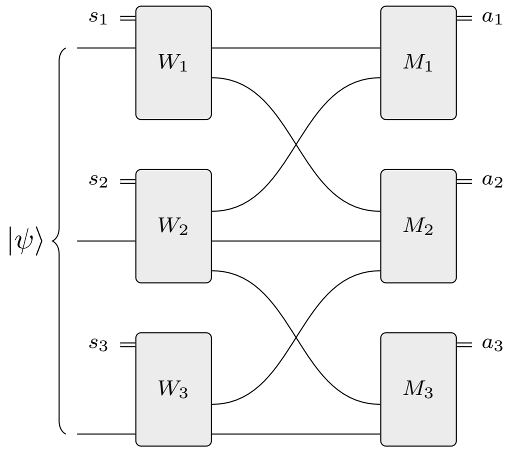


Figure 3: A quantum strategy for a graph game, where the connectivity graph G is the graph $v_1 \leftrightarrows v_2 \leftrightarrows v_3$.

Second, note that even though the measurement appears to occur at the end of the quantum strategy, this definition does not preclude the possibility that the parties perform a measurement first and then transmit the outcomes to one another. Moreover, this measurement can be made *input-dependent*, thereby recovering quantum strategies for conventional nonlocal games. Suppose party i wishes to perform a measurement described by the input-dependent projective measurement $\{\Pi_{i,a_i}(s_i)\}_{a_i \in A_i}$ on his share of the quantum state $|\psi\rangle$ and transmit the measurement result to party j. He can use the isometry $W_i(s_i)$ that satisfies

$$W_i(s_i) O_i W_i(s_i)^\dagger \coloneqq \sum_{a_i \in A_i} [\Pi_{i,a_i}(s_i) O_i \Pi_{i,a_i}(s_i) \otimes |a_i\rangle\langle a_i|]_{B_{i\to i}} \otimes |a_i\rangle\langle a_i|_{B_{i\to j}}. \tag{2.3}$$

where O_i is an arbitrary operator in the Hilbert space B_i and $|a_i\rangle$ denotes classical information encoded in a fixed basis. It is easy to see that $W_i(s_i)$ is an isometry for all $s_i \in S_i$. Indeed, for any quantum states $|\psi\rangle, |\phi\rangle \in B_i$,

$$\begin{aligned}
\langle\psi|W_i(s_i)^\dagger W_i(s_i)|\phi\rangle &= \mathrm{tr}[W_i(s_i)|\phi\rangle\langle\psi|W_i(s_i)^\dagger] \\
&= \sum_{a_i \in A_i} \mathrm{tr}[\Pi_{i,a_i}(s_i)|\phi\rangle\langle\psi|\Pi_{i,a_i}(s_i)] \\
&= \langle\psi|\phi\rangle\,.
\end{aligned}$$

After this isometry, party i and j receive system $B_{i\to i}$ and $B_{i\to j}$, respectively, and perform measurements in the fixed basis. In this way, party i effectively performs an input-dependent measurement and transmits the result to party j while keeping a copy for himself.

Now, let $\mathcal{Q}$ be the set of all quantum behaviors. Note that this set is only dependent on the input sets $\{S_i\}_{i=1}^n$, the output sets $\{A_i\}_{i=1}^n$, as well as the connectivity graph G. We can prove that this is convex via a proof similar to [26].

Proposition 6. *For any connectivity graph G, input sets S_i, and output sets A_i, the set $\mathcal{Q}$ is convex.*

Proof. The key idea of the proof is to use the direct sum of Hilbert spaces of the strategies that realize the original two behaviors. Note that in general, if we put a dimension constraint on the Hilbert spaces used, the set of quantum behaviors is not always convex [27].

Consider a convex combination of two quantum behaviors in $\mathcal{Q}$:

$$\begin{aligned}
\xi(\mathbf{a}|\mathbf{s}) &= \lambda p(\mathbf{a}|\mathbf{s}) + (1-\lambda)p'(\mathbf{a}|\mathbf{s}), \\
p(\mathbf{a}|\mathbf{s}) &= \langle\psi| \bigotimes_{i=1}^{n} W_i(s_i)^\dagger \bigotimes_{i=1}^{n} \Pi_{i,a_i} \bigotimes_{i=1}^{n} W_i(s_i)|\psi\rangle, \\
p'(\mathbf{a}|\mathbf{s}) &= \langle\psi'| \bigotimes_{i=1}^{n} W'_i(s_i)^\dagger \bigotimes_{i=1}^{n} \Pi'_{i,a_i} \bigotimes_{i=1}^{n} W'_i(s_i)|\psi'\rangle.
\end{aligned}$$

Denote the Hilbert spaces for $p(\mathbf{a}|\mathbf{s}), p'(\mathbf{a}|\mathbf{s})$ as $B_i, B_{i\to j}$ and $B'_i, B'_{i\to j}$, respectively. By replacing the underlying quantum system with the direct sum of the original two systems:

$$C_i \coloneqq B_i \oplus B'_i, C_{i\to j} \coloneqq B_{i\to j} \oplus B'_{i\to j}$$

and defining the corresponding $C_i^{\text{out}}, C_i^{\text{in}}$ systems as in Theorem 5, we can reconstruct $\xi(\mathbf{a}|\mathbf{s})$ as:

$$\xi(\mathbf{a}|\mathbf{s}) = \langle\phi| \bigotimes_{i=1}^{n} V_i(s_i)^\dagger \bigotimes_{i=1}^{n} \Xi_{i,a_i} \bigotimes_{i=1}^{n} V_i(s_i)|\phi\rangle,$$

where

$$\begin{aligned}
|\phi\rangle &\coloneqq \sqrt{\lambda}|\psi\rangle \oplus \sqrt{1-\lambda}|\psi'\rangle \in \bigotimes_{i=1}^{n} C_i, \\
V_i(s_i) &\coloneqq W_i(s_i) \oplus W'_i(s_i) : C_i \to C_i^{\text{out}}, \\
\Xi_{i,a_i} &\coloneqq \Pi_{i,a_i} \oplus \Pi'_{i,a_i} : C_i^{\text{in}} \to C_i^{\text{in}}.
\end{aligned}$$

The conclusion follows. □

We next prove a simple result regarding $\mathcal{Q}$ for different connectivity graphs.

Proposition 7. *Let $\{S_i\}_{i=1}^n$, $\{A_i\}_{i=1}^n$ be input and output sets, respectively, for n parties and $G = (V, E)$ a directed graph with n vertices. Let $\mathcal{Q}$ be the set of all possible quantum behaviors. Then, if $G' = (V, E')$ where $E' \subseteq E$, then the set of possible quantum behaviors $\mathcal{Q}'$ where G is replaced by G' is a subset of $\mathcal{Q}$.*

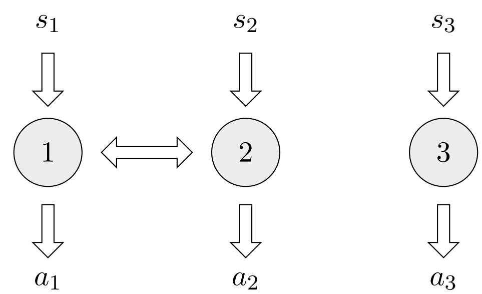
Proof. Consider a quantum strategy $(|\psi\rangle, \{W'_i(s_i)\}_{i\in[n], s_i\in S_i}, \{M'_i\}_{i=1}^n)$ where the connectivity graph is G'. Let $p'(\mathbf{a}|\mathbf{s})$ be the behavior realized. Define additional Hilbert spaces $B_{i\to j}$ where $(i,j) \in E\backslash E'$ to be trivial (one-dimensional) and add them to the output spaces of the isometries $W'_i(s_i)$:

$$W_i(s_i)|\phi\rangle_{B_i} \coloneqq W'_i(s_i)|\phi\rangle_{B_i} \otimes \bigotimes_{j:(i,j)\in E\backslash E'} |\xi_j\rangle_{B_{i\to j}},$$

where $|\xi_j\rangle_{B_{i\to j}}$ is a normalized state. We also add them to the input spaces of the measurements M'_i:

$$M_i\Big(|\phi\rangle_{B_i^{\text{in}}} \otimes \bigotimes_{j:(j,i)\in E\backslash E'} |\xi_j\rangle_{B_{j\to i}}\Big) \coloneqq (M'_i|\phi\rangle_{B_i^{\text{in}}}) \otimes \bigotimes_{j:(j,i)\in E\backslash E'} |\xi_j\rangle_{B_{j\to i}}.$$

Hence, $(|\psi\rangle, \{W_i(s_i)\}_{i\in[n], s_i\in S_i}, \{M_i\}_{i=1}^n)$ is a quantum strategy where the connectivity graph is G. Thus, $p'(\mathbf{a}|\mathbf{s}) \in \mathcal{Q}$. □



$a_1 = a_2$ and $a_1 \oplus a_3 = (s_1 \oplus s_2) \wedge s_3$.

Figure 4: The distributed CHSH game.

This corresponds to the fact that more correlations become possible as the latency constraint is relaxed. Hence, in particular, for a growing sequence of subgraphs

$$G_1 \subseteq G_2 \subseteq \cdots$$

that have the same vertex set, the maximum achievable winning probability by a quantum strategy ω_q is monotonically non-decreasing.

2.3 Examples: distributed games

In this section, we present examples of LC games where classical and quantum strategies have different winning probabilities. The examples we consider are "distributed" variants of well-known nonlocal games, specifically the CHSH game and the magic square game. While the original versions involve only two parties, the distributed variants feature three parties, with two of them collectively playing the role of a single party from the original game. Although this may appear to complicate the task, we allow communication according to a connectivity graph, enabling these two parties to communicate via a quantum strategy as defined in Theorem 5. We show that the three parties can achieve the quantum value of the original games. Our example not only separates classical and quantum strategies for LC games but also distinguishes quantum strategies with and without communication.

2.3.1 Distributed CHSH game

We first give a distributed version of the CHSH game. The three parties receive inputs from $S_1, S_2, S_3 = \{0, 1\}$ and give outputs in $A_1, A_2, A_3 = \{0, 1\}$. The predicate is given by

$$\mathcal{V}(a_1, a_2, a_3 | s_1, s_2, s_3) = \begin{cases} 1 & a_1 = a_2 \text{ and } (s_1 \oplus s_2) \wedge s_3 = a_1 \oplus a_3, \\ 0 & \text{otherwise.} \end{cases}$$

That is, each of the first two parties is playing the CHSH game with the third party. The first two parties must produce the same output. Moreover, one input to the CHSH game is distributed in the sense that it is the parity of the inputs of the first two parties, and neither party knows the actual input. Lastly, let π be the uniform distribution. The setup is shown in Figure 4.

We first claim that the distributed CHSH game has no quantum violation if the first two partiess cannot communicate with each other. That is, the latency constraint is very low. In this case, the connectivity graph is the empty graph and this becomes a conventional nonlocal game of three parties. We can write the winning probability as

$$\begin{aligned} p_{\text{win}} = \frac{1}{8}\Big[& p(000|000) + p(111|000) + p(000|010) + p(111|010) \\ & + p(000|100) + p(111|100) + p(000|110) + p(111|110) \\ & + p(000|001) + p(111|001) + p(110|011) + p(001|011) \\ & + p(110|101) + p(001|101) + p(000|111) + p(111|111)\Big]. \end{aligned} \tag{2.4}$$

It will be useful to express this in correlator form [28]. Given the input (s_1, s_2, s_3), the three parties measure binary observables $K_1^{s_1}, K_2^{s_2}, K_3^{s_3}$, respectively, with eigenvalues $+1, -1$. The measured eigenvalues $+1$ and -1 correspond to producing output bit values 0 and 1, respectively. We explicitly define correlators as

$$\begin{aligned} \langle K_i^{s_i} \rangle &:= \sum_{a_i=0}^{1} (2a_i - 1) \cdot p(a_i|s_i) \\ \langle K_i^{s_i} K_j^{s_j} \rangle &:= \sum_{a_i, a_j=0}^{1} (2a_i - 1)(2a_j - 1) \cdot p(a_i, a_j|s_i, s_j) \\ \langle K_i^{s_i} K_j^{s_j} K_k^{s_k} \rangle &:= \sum_{a_i, a_j, a_k=0}^{1} (2a_i - 1)(2a_j - 1)(2a_k - 1) \cdot p(a_i, a_j, a_k|s_i, s_j, s_k), \end{aligned}$$

where $p(a_i|s_i), p(a_i, a_j|s_i, s_j)$ are the marginal behaviors on party i and a pair of parties $i \neq j$, which exist due to the no-signaling condition. The equation

$$\begin{aligned} p(a_1, a_2, a_3|s_1, s_2, s_3) =& \frac{1}{8} \left[1 + (-1)^{a_1} \langle K_1^{s_1} \rangle + (-1)^{a_2} \langle K_2^{s_2} \rangle + (-1)^{a_3} \langle K_3^{s_3} \rangle + (-1)^{a_1 \oplus a_2} \langle K_1^{s_1} K_2^{s_2} \rangle \right. \\ & \left. + (-1)^{a_1 \oplus a_3} \langle K_1^{s_1} K_3^{s_3} \rangle + (-1)^{a_2 \oplus a_3} \langle K_2^{s_2} K_3^{s_3} \rangle + (-1)^{a_1 \oplus a_2 \oplus a_3} \langle K_1^{s_1} K_2^{s_2} K_3^{s_3} \rangle \right] \end{aligned}$$

holds for all $a_i, s_i \in \{0, 1\}$. For the outputs $(0,0,0), (1,1,1), (1,1,0), (0,0,1)$ that appear in p_{win}, we have

$$\begin{aligned} p(0,0,0|s_1,s_2,s_3) + p(1,1,1|s_1,s_2,s_3) &= \frac{1}{4}[1 + \langle K_1^{s_1} K_2^{s_2} \rangle + \langle K_1^{s_1} K_3^{s_3} \rangle + \langle K_2^{s_2} K_3^{s_3} \rangle] \\ p(1,1,0|s_1,s_2,s_3) + p(0,0,1|s_1,s_2,s_3) &= \frac{1}{4}[1 + \langle K_1^{s_1} K_2^{s_2} \rangle - \langle K_1^{s_1} K_3^{s_3} \rangle - \langle K_2^{s_2} K_3^{s_3} \rangle]. \end{aligned}$$

Thus, the winning probability p_{win} in Equation (2.4) can be expressed in a correlator form as

$$\begin{aligned} p_{\text{win}} =& \frac{1}{4} + \frac{1}{16}[\langle K_1^0 K_2^0 \rangle + \langle K_1^0 K_2^1 \rangle + \langle K_1^1 K_2^0 \rangle + \langle K_1^1 K_2^1 \rangle \\ & + \langle K_1^0 K_3^0 \rangle + \langle K_1^1 K_3^0 \rangle + \langle K_2^0 K_3^0 \rangle + \langle K_2^1 K_3^0 \rangle] \end{aligned} \tag{2.5}$$

Therefore, the RHS of Equation (2.5) is at most $\frac{3}{4}$. Note that since we used the no-signaling constraints in deriving Equation (2.5), $\frac{3}{4}$ is the maximum value for no-signaling behaviors rather than the maximum algebraic value ω_a, which is 1. The winning probability of $\frac{3}{4}$ can be attained for a classical strategy where all parties have a constant output of 0. Hence, the quantum value, which has to be between the classical and no-signaling value, also has to be $\frac{3}{4}$. The claim follows.

Now, we *slightly relax the latency constraint* so that the first two parties can communicate with each other, but no other communication is possible. That is, we have the connectivity graph G given by

$$v_1 \leftrightarrows v_2 \quad v_3.$$

This is possible if v_3 is slightly farther away, such as the setting of "Partial communicating scenario 1" shown in Figure 1. Here, we overload the letters v_1, v_2, v_3 to denote the parties as well as the vertices in G. We define the LC game, which we will call the **distributed CHSH game**, by the tuple $(\mathcal{V}, \pi, G)$. First of all, the classical value of this LC game is still $\frac{3}{4}$. This is because letting v_1, v_2 communicate effectively makes them one party as they can share inputs and locally compute outputs using their functions f_1, f_2 on their combined inputs.[4] This "aggregated" party now has to play the regular CHSH game with v_3. This can only have a winning probability of at most $\frac{3}{4}$ for a classical strategy. That is, even with the relaxed latency constraint, this is the maximum winning probability. *This is an example of a fundamental bound on classical parties in this new regime of relaxed latency constraints and is an analogue of Bell inequalities which hold in the regime of very low latency constraints.*[5]

However, we will find that in this relaxed latency constraint regime, *there is a quantum violation*. That is, although there was no quantum violation at first for very low latency constraints, by waiting a little longer a quantum violation becomes possible! Moreover, the quantum strategy we find wins the distributed CHSH game with probability $\cos^2 \frac{\pi}{8}$. This is the highest possible value for a quantum strategy. To see this, we aggregate v_1, v_2 and consider them as one party with input (s_1, s_2) and output (a_1, a_2). This can only increase the winning probability. Then, this aggregated party is effectively playing the usual CHSH game with v_3. The CHSH game has a quantum value of $\cos^2 \frac{\pi}{8}$ as a nonlocal game. Hence, the quantum value of the distributed CHSH game as an LC game is also $\cos^2 \frac{\pi}{8}$.

At first glance, it is not immediately clear how a quantum strategy for an LC game as defined in Theorem 5 can achieve this winning probability. In particular, for a quantum strategy, the first two parties can only communicate with each other once. If they can back-and-forth communicate, then the solution becomes trivial. In this case, v_1 and v_3 share a maximally entangled state

$$|\Phi^+\rangle := (|00\rangle + |11\rangle)/\sqrt{2}.$$

In the first round of communication, v_2 sends his input s_2 to v_1, who then computes $s_1 \oplus s_2$ and applies the corresponding measurement for an optimal quantum strategy of the original CHSH game on his share of $|\Phi^+\rangle$. v_3 applies his measurement according to s_3. Then, in the second round of communication, v_1 sends his output a_1 to v_2, who directly uses it as his own output. This clearly wins the distributed CHSH game with probability $\cos^2 \frac{\pi}{8}$, the quantum value of the original CHSH game.[6]

Surprisingly, we will find that a quantum strategy for an LC game, which only involves a single round of communication, also suffices to achieve this winning probability. From a physical perspective, this is interesting because it means that we can achieve the same winning probability

[4] This will be stated more explicitly when we prove Theorem 22.

[5] It is also an example of a Svetlichny inequality [24]. Now, previous works on violations of a Svetlichny inequality only consider using a quantum strategy for a conventional nonlocal game. By the above result, no such violation is possible. However, we will find that a violation is possible *if the quantum parties can communicate*, which is allowed by the relaxed latency constraint.

[6] Indeed, we will find that this is actually a quite general result when we prove Theorem 23.

while consuming only *half* the time. We consider a strategy where the first two parties' initial quantum system is in a maximally entangled logical qubit state with the third party:

$$|\psi\rangle_{B_1B_2B_3} := \frac{1}{\sqrt{2}}(|0_L0\rangle_{(B_1B_2)B_3} + |1_L1\rangle_{(B_1B_2)B_3}).$$

Here, the logical qubits $|i_L\rangle_{B_1B_2}$ are

$$|0_L\rangle_{B_1B_2} := \frac{1}{\sqrt{2}}\left(|0\rangle_{B_1}|1\rangle_{B_2} + |1\rangle_{B_1}|0\rangle_{B_2}\right),$$
$$|1_L\rangle_{B_1B_2} := \frac{1}{\sqrt{2}}\left(|0\rangle_{B_1}|0\rangle_{B_2} - |1\rangle_{B_1}|1\rangle_{B_2}\right),$$

which are two-qubit states residing in the $+1$ eigenspace of $Y_{B_1} \otimes Y_{B_2}$, with v_1 and v_2 each holding one qubit. We observe

$$\begin{aligned} X_{B_1} \otimes X_{B_2}|0_L\rangle_{B_1B_2} &= |0_L\rangle_{B_1B_2}, \quad X_{B_1} \otimes X_{B_2}|1_L\rangle_{B_1B_2} = -|1_L\rangle_{B_1B_2}, \\ Z_{B_1} \otimes Z_{B_2}|0_L\rangle_{B_1B_2} &= -|0_L\rangle_{B_1B_2}, \quad Z_{B_1} \otimes Z_{B_2}|1_L\rangle_{B_1B_2} = |1_L\rangle_{B_1B_2}. \end{aligned} \tag{2.6}$$

That is, both $X_{B_1} \otimes X_{B_2}$ and $-Z_{B_1} \otimes Z_{B_2}$ are logical Z operators on the logical qubits. Similarly, as

$$\begin{aligned} X_{B_1} \otimes Z_{B_2}|0_L\rangle_{B_1B_2} &= |1_L\rangle_{B_1B_2} \\ Z_{B_1} \otimes X_{B_2}|0_L\rangle_{B_1B_2} &= |1_L\rangle_{B_1B_2}, \end{aligned} \tag{2.7}$$

both $X_{B_1} \otimes Z_{B_2}$ and $Z_{B_1} \otimes X_{B_2}$ are logical X operators on the logical qubits. Based on the above observation, the following strategy allows both v_1 and v_2 to perform a CHSH game with v_3, while ensuring that v_1 and v_2 produce the same output:

- v_1 measures X_{B_1} on his qubit when $s_1 = 0$, and measures Z_{B_1} on his qubit when $s_1 = 1$. The measurement result is denoted by $m_1 \in \{1, -1\}$;
- v_2 measures X_{B_2} on his qubit when $s_2 = 0$, and measures Z_{B_2} on his qubit when $s_2 = 1$. The measurement result is denoted by $m_2 \in \{1, -1\}$;
- v_1 and v_2 share the measurement results m_1, m_2 and their inputs s_1, s_2 using a single-round of classical communication. They subsequently output

$$a_1 = a_2 = \frac{1}{2}[1 - (-1)^{s_1s_2}m_1m_2].$$

- v_3 measures $\frac{Z_{B_3}+X_{B_3}}{\sqrt{2}}$ on his qubit when $s_3 = 0$, and measures $\frac{Z_{B_3}-X_{B_3}}{\sqrt{2}}$ on his qubit when $s_3 = 1$. Then v_3 uses his measurement result $m_3 \in \{1, -1\}$ to give output $a_3 = \frac{1}{2}(1 - m_3)$;

With this strategy, both v_1 and v_2 effectively measure the observables $X_{B_1} \otimes X_{B_2}$, $X_{B_1} \otimes Z_{B_2}$, $Z_{B_1} \otimes X_{B_2}$, $-Z_{B_1} \otimes Z_{B_2}$ after communication for inputs $(s_1, s_2) \in \{(0,0), (0,1), (1,0), (1,1)\}$ respectively. These are exactly the logical Z observable for $(s_1, s_2) \in \{(0,0), (1,1)\}$ and logical X observable for $(s_1, s_2) \in \{(0,1), (1,0)\}$ on their logical qubit, as proved in Equations (2.6) and (2.7) above. Each of the first two parties v_1 and v_2, paired with the third party v_3, is therefore implementing an optimal quantum strategy for the usual CHSH game, and the outputs of the first two parties are guaranteed to be the same. According to Equation (2.4), the winning probability of this strategy in distributed CHSH game is

$$\begin{aligned}p_{win} =&\frac{1}{8}\left[p(m_1m_2 = m_3|000) + p(m_1m_2 = m_3|001) + p(m_1m_2 = m_3|010) + p(m_1m_2 = -m_3|011)\right.\\&\left.p(m_1m_2 = m_3|100) + p(m_1m_2 = -m_3|101) + p(-m_1m_2 = m_3|110) + p(-m_1m_2 = m_3|111)\right]\end{aligned}$$

Here, we have[7]

$$\begin{aligned}&p(m_1m_2 = m_3|000) + p(m_1m_2 = m_3|001)\\=&\frac{1 + \langle\psi|X \otimes X \otimes \frac{Z+X}{\sqrt{2}}|\psi\rangle}{2} + \frac{1 + \langle\psi|X \otimes X \otimes \frac{Z-X}{\sqrt{2}}|\psi\rangle}{2}\\=&1 + \frac{1}{\sqrt{2}}\langle\psi|X \otimes X \otimes Z|\psi\rangle\\=&1 + \frac{1}{\sqrt{2}},\end{aligned}$$

and similarly,

$$\begin{aligned}p(m_1m_2 = m_3|010) + p(m_1m_2 = -m_3|011) &= 1 + \frac{1}{\sqrt{2}}\langle\psi|X \otimes Z \otimes X|\psi\rangle = 1 + \frac{1}{\sqrt{2}},\\p(m_1m_2 = m_3|100) + p(m_1m_2 = -m_3|101) &= 1 + \frac{1}{\sqrt{2}}\langle\psi|Z \otimes X \otimes X|\psi\rangle = 1 + \frac{1}{\sqrt{2}},\\p(-m_1m_2 = m_3|110) + p(-m_1m_2 = m_3|111) &= 1 - \frac{1}{\sqrt{2}}\langle\psi|Z \otimes Z \otimes Z|\psi\rangle = 1 + \frac{1}{\sqrt{2}}.\end{aligned}$$

Therefore, this quantum strategy can achieve a winning probability of $\cos^2\frac{\pi}{8}$ as claimed.

We summarize our results for the distributed CHSH game with in Table 1.

Connectivity graph	ω_c	ω_q	Quantum violation?
$v_1 \quad v_2 \quad v_3$	$\frac{3}{4}$	$\frac{3}{4}$	No
$v_1 \leftrightarrows v_2 \quad v_3$	$\frac{3}{4}$	$\cos^2\frac{\pi}{8}$	Yes

Table 1: Classical and quantum values for the distributed CHSH game with different connectivity graphs. The empty graph corresponds to a conventional nonlocal game, while the non-empty graph corresponds to an LC game.

We state again the miraculous conclusion: while a quantum violation is impossible under very low latency constraints where no parties can communicate, by slightly relaxing the latency constraint so that only v_1, v_2 can communicate with each other, a quantum violation becomes possible.

Additionally, the three-party distributed CHSH game can be naturally generalized to an $(n+m)$-party distributed CHSH game in the following way. Each party receives input $s_i \in \{0,1\}$, which is chosen uniformly at random, and produce outputs $a_i \in \{0,1\}$. The probabilistic predicate is given by

$$\mathcal{V}(\mathbf{a}|\mathbf{s}) = \begin{cases}1 & a_1 = a_2 = \cdots = a_n,\ a_{n+1} = a_{n+2} = \cdots = a_{n+m},\ (\bigoplus_{i=1}^n s_i) \wedge (\bigoplus_{j=n+1}^{n+m} s_j) = a_1 \oplus a_{n+1},\\0 & \text{otherwise.}\end{cases}$$

[7]The subscript B_1, B_2, B_3 for state and measurements are omitted for simplicity.

That is, there are two groups of parties, and each group collectively acts as a single party in the original CHSH game. The two inputs for this CHSH game correspond to the parity of the inputs from the parties within each group.

For such an $(n+m)$-party distributed CHSH game, we can consider a strategy where the parties share the $(n+m)$-qubit entangled state

$$|\psi^{(m+n)}\rangle = \frac{1}{\sqrt{2}}\left[|0_L^{(n)}\rangle\left(\cos\frac{\pi}{8}|0_L^{(m)}\rangle + \sin\frac{\pi}{8}|1_L^{(m)}\rangle\right) + |1_L^{(n)}\rangle\left(\sin\frac{\pi}{8}|0_L^{(m)}\rangle - \cos\frac{\pi}{8}|1_L^{(m)}\rangle\right)\right],$$

with each party holding one qubit. Here, $|0_L^{(n)}\rangle$ and $|1_L^{(n)}\rangle$ are the logical $|0\rangle$ and $|1\rangle$ of the code stabilized by $\{Y_iY_{i+1}\}_{i=1}^{n-1}$ with $Z_1Z_2\cdots Z_n$ serving as the logical Z operator. Then each party in the first group measures Z on his qubit when receives input 0 and measures X on his qubit when he receives input 1. It is straightforward to check that, for a Pauli string consisting of n X and Z operators, an odd number of X's corresponds to the logical X (or $-X$), while an even number corresponds to the logical Z (or $-Z$). Similarly, for the m parties in the second group, $|0_L^{(m)}\rangle$ and $|1_L^{(m)}\rangle$ are the logical $|0\rangle$ and $|1\rangle$ of the code stabilized by $\{Y_iY_{i+1}\}_{i=n+1}^{n+m-1}$, with $Z_{n+1}Z_{n+2}\cdots Z_{n+m}$ serving as the logical Z operator. Again, each of the m parties measures Z when receives input 0 and measures X when receives input 1, so as to implement two anticommuting logical operators $\pm X$ and $\pm Z$ depending on the parity $\bigoplus_{j=n+1}^{n+m} s_j$.

With the above shared entangled states and measurements, if the parties can have one round of communication within each group after their measurements in which they share the measurement results, they can win this $(n+m)$-party distributed CHSH game with probability $\cos^2\frac{\pi}{8}$, similar to the three-party case.

2.3.2 Distributed magic square game

We next consider a distributed version of the magic square game. In particular, we will use a variation of the magic square game [29] where one party is given a row or a column to fill in while the other party fills in a specific entry in the magic square. The first party's answer must fulfill the parity requirements of a magic square while the second party's entry must match the first party's. The input distribution is constrained so that the row or column given contains the specified entry. We draw the setup of the magic square game in Figure 5 and show the optimal quantum strategy.

$+1$	$+1$	$+1$	$+1$
-1	$+1$	-1	$+1$
-1	$+1$	?	$+1$
$+1$	$+1$	-1	

XI	IX	XX
IZ	ZI	ZZ
XZ	ZX	YY

Figure 5: The magic square game. An optimal quantum strategy that has unit winning probability is to measure the observables on the right-side figure.

In the distributed magic square game, v_1 and v_2 share the role of the party in the original game that has to fill in a specific entry. Instead of getting the full coordinates of that entry, the

x-coordinate is given to v_1 and the y-coordinate to v_2. More explicitly, the LC game is defined as follows.

1. The connectivity graph G is again given by
$$v_1 \leftrightarrows v_2 \qquad v_3.$$
2. v_1 receives input $s_1 \in [3]$, and v_2 receives input $s_2 \in [3]$. They each must output ± 1.
3. v_3 receives either s_1 or $s_2 + 3$ as his input s_3 with equal probability. If he receives $s_3 \in \{1, 2, 3\}$, he must fill in the s_3-th row with ± 1. While if he receives $s_3 \in \{4, 5, 6\}$, he must fill in the $(s_3 - 3)$-th column with ± 1.
4. The winning condition is:
   - v_3 fills the row or column with correct parity as according to the usual magic square game. The parity requirements are shown in Figure 5.
   - v_1 and v_2 have the same output, which should also be the same as what v_3 filled in for the entry (s_1, s_2).

There exists a quantum strategy with unit winning probability. In this strategy, two pairs of maximally entangled states
$$|\Phi^+\rangle^{\otimes 2} = \frac{1}{2}(|00\rangle + |11\rangle)^{\otimes 2}$$
are shared between v_1 and v_3. After receiving their inputs, v_1 perform measurements in the following bases depending on s_1:

$$\begin{aligned}
&s_1 = 1 : \{|{+}{+}\rangle, |{+}{-}\rangle, |{-}{+}\rangle, |{-}{-}\rangle\} \\
&s_1 = 2 : \{|00\rangle, |10\rangle, |01\rangle, |11\rangle\} \\
&s_1 = 3 : \{\frac{1}{\sqrt{2}}(|0+\rangle + |1-\rangle, \frac{1}{\sqrt{2}}(|0-\rangle + |1+\rangle, \frac{1}{\sqrt{2}}(|0+\rangle - |1-\rangle), \frac{1}{\sqrt{2}}(|0-\rangle - |1+\rangle\}.
\end{aligned}$$

The four measurement outcomes, as ordered above, are labeled via two binary digits $(m_1, m_1') \in \{+1, -1\}^2$. That is, the ordering is $(+1, +1), (+1, -1), (-1, +1), (-1, -1)$. Meanwhile, the measurement bases of v_3 are chosen in the same way as in the magic square game in [29]:

$$\begin{aligned}
&s_3 = 1 : \{|{+}{+}\rangle, |{+}{-}\rangle, |{-}{+}\rangle, |{-}{-}\rangle\} \\
&s_3 = 2 : \{|00\rangle, |10\rangle, |01\rangle, |11\rangle\} \\
&s_3 = 3 : \Big\{\frac{1}{\sqrt{2}}(|0+\rangle + |1-\rangle, \frac{1}{\sqrt{2}}(|0-\rangle + |1+\rangle, \frac{1}{\sqrt{2}}(|0+\rangle - |1-\rangle), \frac{1}{\sqrt{2}}(|0-\rangle - |1+\rangle)\Big\} \\
&s_3 = 4 : \{|{+}0\rangle, |{+}1\rangle, |{-}0\rangle, |{-}1\rangle\} \\
&s_3 = 5 : \{|0+\rangle, |1+\rangle, |0-\rangle, |1-\rangle\} \\
&s_3 = 6 : \Big\{\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle, \frac{1}{\sqrt{2}}(|01\rangle + |10\rangle, \frac{1}{\sqrt{2}}(|00\rangle - |11\rangle), \frac{1}{\sqrt{2}}(|01\rangle - |10\rangle)\Big\}.
\end{aligned}$$

and the four outcomes are also labeled via two binary digits $(m_3, m_3') \in \{+1, -1\}^{\times 2}$, ordered as $(+1, +1), (+1, -1), (-1, +1), (-1, -1)$.

After the measurements, v_1 sends his measurement outcome (m_1, m_1') to v_2, and v_2 sends his input s_2 to v_1. Finally, both v_1 and v_2 output the result of the function

$$f(s_2, m_1, m_1') \coloneqq \begin{cases} m_1 & \text{if } s_2 = 1, \\ m_1' & \text{if } s_2 = 2, \\ m_1 m_1' & \text{if } s_2 = 3. \end{cases}$$

Meanwhile, v_3 just fills the row or column specified by s_3 according to his measurement outcome exactly as in the quantum strategy for the variation of the magic square game. That is, if $s_3 \in \{1, 2, 3\}$, the s_3-th row is filled with $m_3, m_3', m_3 m_3'$, respectively. While if $s_3 \in \{4, 5, 6\}$, the $(s_3 - 3)$-th column is filled with $m_3, m_3', m_3 m_3'$, respectively.

We now show that this strategy wins the distributed magic square game with unit probability. We observe that v_1 and v_2 are measuring the same "effective observable" $M(s_1, s_2)$ consisting of v_1's projective measurement followed by classical post-processing $f(s_2, m, m')$. For example, for $s_1 = 1$, the effective measurements are

$$\begin{aligned}
M(1,1) &= |{+}{+}\rangle\langle{+}{+}| + |{+}{-}\rangle\langle{+}{-}| - |{-}{+}\rangle\langle{-}{+}| - |{-}{-}\rangle\langle{-}{-}| = X \otimes I, \\
M(1,2) &= |{+}{+}\rangle\langle{+}{+}| - |{+}{-}\rangle\langle{+}{-}| + |{-}{+}\rangle\langle{-}{+}| - |{-}{-}\rangle\langle{-}{-}| = I \otimes X, \\
M(1,3) &= |{+}{+}\rangle\langle{+}{+}| - |{+}{-}\rangle\langle{+}{-}| - |{-}{+}\rangle\langle{-}{+}| + |{-}{-}\rangle\langle{-}{-}| = X \otimes X.
\end{aligned}$$

Similarly, one can check that for other possible inputs (s_1, s_2), the effective measurements $M(s_1, s_2)$ are

$$\begin{aligned}
&M(2,1) = I \otimes Z, \ M(2,2) = Z \otimes I, \ M(2,3) = Z \otimes Z, \\
&M(3,1) = X \otimes Z, \ M(3,2) = Z \otimes X, \ M(3,3) = Y \otimes Y.
\end{aligned}$$

This is exactly the same as the quantum strategy shown on the right side of Figure 5. Therefore, v_1 and v_2 always produce identical outputs, and each of them can be effectively regarded as playing the variation of the magic square game [29] with v_3 using an optimal quantum strategy. Consequently, a quantum strategy with a single round of communication is sufficient to win the distributed magic square game with unit probability.

2.4 Forwarding strategies

We defined a quantum strategy for an LC game in Theorem 5. Ostensibly, the core mathematical object that describes it is not the same as that of quantum strategies for conventional nonlocal games. However, it is worthwhile to check this statement more carefully. In particular, we want to check that if we use the same mathematical object as a quantum strategy for a nonlocal game, but allow communication of inputs just like a classical strategy for an LC game, whether we obtain the same set of behaviors.

That is, we define a strategy where parties share inputs according the connectivity graph and then make a measurement on a shared quantum state given the received information. This strategy is indeed consistent with Theorem 5: if party i wants to share his input s_i to party j, he can simply use the isometry $W_i(s_i)$ which acts on an operator O_i on B_i by

$$W_i(s_i)(O_i)_{B_i} W_i(s_i)^\dagger \coloneqq (O_i)_{B_{i\to i}} \otimes |s_i\rangle\langle s_i|_{B_{i\to j}},$$

where $|s_i\rangle$ denotes classical information encoded in a fixed basis. Party j performs a measurement on his quantum system conditioned on the classical information encoded in the fixed basis. This

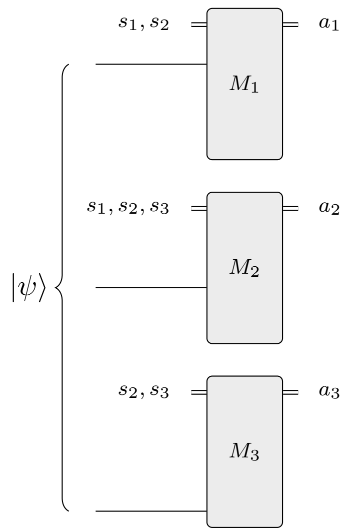


Figure 6: A forwarding strategy, where G is the graph $v_1 \leftrightarrows v_2 \leftrightarrows v_3$.

results in the strategy we described. In fact, we can give a new, simpler definition for this special class of quantum strategies.

Definition 8. *Let $(\mathcal{V}, \pi, G)$ be an LC game. Recall*

$$H_i := \prod_{j \in N_{\text{in}}[i]} S_j.$$

*is the past light cone of party i. Let B_i be Hilbert spaces for $i \in [n]$. A **forwarding strategy** is a tuple*

$$(|\psi\rangle, \{M_i(h_i)\}_{i \in [n], h_i \in H_i})$$

where $|\psi\rangle \in \bigotimes_{i=1}^{n} B_i$ is a quantum state and for fixed $i \in [n]$ and $h_i \in H_i$, $M_i(h_i) := \{\Pi_{i,a_i}(h_i)\}_{a_i \in A_i}$ is a projective measurement on B_i.

We choose the word "forwarding" because in such strategies, each party simply "forwards" his input to adjacent parties. Using the same notation as above, the realized behavior of a forwarding strategy, which we will call a **forwarding behavior**, is given by

$$p(\mathbf{a}|\mathbf{s}) := \Big\langle \psi \Big| \bigotimes_{i=1}^{n} \Pi_{i,a_i}(h_i) \Big| \psi \Big\rangle,$$

where

$$h_i := \prod_{j \in N_{\text{in}}[i]} s_j.$$

In Figure 6, we show a forwarding strategy for the same graph as Figure 3: $v_1 \leftrightarrows v_2 \leftrightarrows v_3$.

Now, it is clear that the mathematical object describing a forwarding strategy is exactly that of a quantum strategy for a conventional nonlocal game. This is exactly analogous to classical strategies for LC games. We give the following proposition to formalize this connection.

Proposition 9. *Let $(\mathcal{V}, \pi, G)$ be an LC game and let*

$$(|\psi\rangle, \{M_i(h_i)\}_{i\in[n],h_i\in H_i})$$

be a forwarding strategy. Then, the winning probability realized by this strategy is equal to that of a quantum strategy described by the same tuple for the nonlocal game $(\mathcal{V}', \pi')$, where

$$\mathcal{V}' : \prod_{i=1}^{n} A_i \times \prod_{i=1}^{n} H_i \to [0,1]$$
$$\mathcal{V}'(\mathbf{a}|h_1, h_2, \ldots, h_n) \coloneqq \mathcal{V}(\mathbf{a}|h_1(1), h_2(2), \ldots, h_n(n)),$$

$h_i(j)$ being the j-th element of

$$h_i \in H_i \coloneqq \prod_{k\in N_{\text{in}}[i]} S_k,$$

and

$$\pi' : \prod_{i=1}^{n} H_i \to [0,1]$$
$$\pi'(h_1, h_2, \ldots, h_n) \coloneqq \pi(h_1(1), h_2(2), \ldots, h_n(n)) \prod_{(i,j)\in E} \delta_{h_i(i),h_j(i)}.$$

In words, the equivalent nonlocal game takes as inputs for each party from the set H_i and the input distribution identifies all copies of S_i across different H_i's. We can see this pictorially in Figure 6 where we are now considering the multiple inputs for one party as one large input.

Proof. We can see this directly by writing down the two winning probabilities and finding them to be exactly the same. □

We now ask the question that motivated the definition of forwarding strategies. That is, given an LC game, let $\mathcal{Q}$ be the set of all quantum behaviors and $\mathcal{F}$ be the set of all forwarding behaviors. Is $\mathcal{F}$ equal to $\mathcal{Q}$? It's obvious that $\mathcal{F} \subseteq \mathcal{Q}$ since forwarding strategies are special cases of quantum strategies. Furthermore, each of these sets are convex ([26] and Theorem 6) and only depend on the input sets S_i, output sets A_i, and connectivity graph G. We find that these two sets are not always equal, *establishing that quantum strategies for LC games does indeed entail new mathematics.*

Consider the case of three parties and consider again the connectivity graph G:

$$v_1 \leftrightarrows v_2 \quad v_3. \tag{2.8}$$

At a high level, our counterexample is the following: Let v_1, v_3 play the CHSH game. Then, using a quantum strategy, v_2 can simply output the measurement results of v_1 using the isometry in Equation (2.3). A forwarding strategy cannot do this. This is because the behavior of v_1, v_3 self-tests the quantum state that v_1, v_3 are using, and therefore by monogamy of entanglement, v_2 cannot possibly know the measurement result of v_1. We state our result as a theorem and give a formal proof.

Theorem 10. *There exist input sets S_i, output sets A_i, and a connectivity graph G such that*

$$\mathcal{F} \subsetneq \mathcal{Q}.$$

Proof. Let G be the graph in Equation (2.8) with 3 vertices. Let

$$S_1, S_3 = \{0,1\}$$

be the possible inputs of v_1, v_3 respectively, while S_2 is a singleton. Since there's only one possible input, we will suppress s_2 in our notation for conciseness. Also, let

$$A_1, A_2, A_3 = \{0,1\}$$

be the possible outputs of v_1, v_2, v_3, respectively. Consider the behavior

$$q(a_1, a_2, a_3|s_1, s_3) := \begin{cases} p_{\text{CHSH}}(a_1, a_3|s_1, s_3) & a_2 = a_1 \\ 0 & \text{otherwise} \end{cases},$$

where $p_{\text{CHSH}}(a_1, a_3|s_1, s_3)$ is the behavior of v_1, v_3 implementing an optimal quantum strategy for the CHSH game.

Then, a quantum strategy for an LC game can achieve this behavior where v_1, v_3 implement an optimal quantum strategy for the CHSH game, and v_1 sends his measurement result a_1 to v_2. v_1, v_3 output their measurement results a_1, a_3, while v_2 outputs a_1. This clearly realizes the desired behavior.

We next show that no forwarding strategy can realize $q(a_1, a_2, a_3|s_1, s_3)$. We see that the marginal distribution $q(a_1, a_3|s_1, s_3)$ is simply $p_{\text{CHSH}}(a_1, a_3|s_1, s_3)$. Now, this behavior self-tests for the maximally entangled state of two qubits $|\Phi^+\rangle := \frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$ as well as the corresponding measurements P in an optimal quantum strategy for the CHSH game. That is, if q is a forwarding behavior:

$$q(a_1, a_2, a_3|s_1, s_3) = \langle\psi|_{B_1B_2B_3}\Pi_{1,a_1}(s_1) \otimes \Pi_{2,a_2}(s_1) \otimes \Pi_{3,a_3}(s_3)|\psi\rangle_{B_1B_2B_3},$$

there exists local isometries $V_1 : B_1 \to B_1' \otimes \bar{B}_1$, $V_3 : B_3 \to B_3' \otimes \bar{B}_3$ such that

$$(V_1 \otimes I_{B_2} \otimes V_3)(\Pi_{1,a_1}(s_1) \otimes I_{B_2} \otimes \Pi_{3,a_3}(s_3)|\psi\rangle_{B_1B_2B_3}) = (P_{1,a_1}(s_1) \otimes P_{3,a_3}(s_3)|\Phi^+\rangle_{B_1'B_3'}) \otimes |\xi\rangle_{\bar{B}_1B_2\bar{B}_3}, \tag{2.9}$$

where ξ is some state. Hence,

$$\begin{aligned}
&\langle\psi|\Pi_{1,a_1}(s_1) \otimes \Pi_{2,a_2}(s_1) \otimes \Pi_{3,a_3}(s_3)|\psi\rangle \\
&= \langle\psi|(V_1^\dagger \otimes I_{B_2} \otimes V_3^\dagger)(V_1 \otimes I_{B_2} \otimes V_3)(\Pi_{1,a_1}(s_1) \otimes \Pi_{2,a_2}(s_1) \otimes \Pi_{3,a_3}(s_3))|\psi\rangle \\
&= \langle\Phi^+|_{B_1'B_3'}\langle\xi|_{\bar{B}_1B_2\bar{B}_3}(P_{1,a_1}(s_1) \otimes P_{3,a_3}(s_3)|\Phi^+\rangle_{B_1'B_3'}) \otimes \Pi_{2,a_2}(s_1)|\xi\rangle_{\bar{B}_1B_2\bar{B}_3} \\
&= \langle\Phi^+|_{B_1'B_3'}P_{1,a_1}(s_1) \otimes P_{3,a_3}(s_3)|\Phi^+\rangle_{B_1'B_3'}\langle\xi|_{\bar{B}_1B_2\bar{B}_3}\Pi_{2,a_2}(s_1)|\xi\rangle_{\bar{B}_1B_2\bar{B}_3},
\end{aligned}$$

where $V^\dagger$ is the adjoint of V and the second equality follows from Equation (2.9) and from summing it over a_1, a_3. This implies that conditioned on s_1, s_3, the output of v_2 is independent of the output of v_1, a contradiction. □

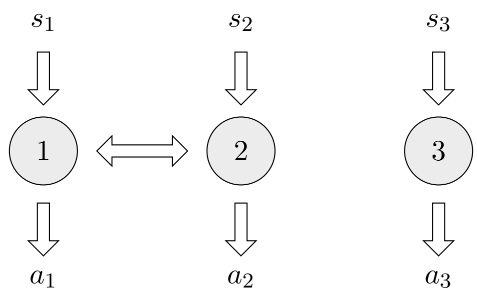
In fact, for any connectivity graph $G = (V, E)$ that is not weakly connected and not empty, $\mathcal{F} \subsetneq \mathcal{Q}$ for an appropriate choice of sets A_i, S_i. This is because we can always find a triple of vertices v_1, v_2, v_3 such that $(v_1, v_2) \in E$ but v_3 is not weakly connected to either. We can then define the

LC game where all parties that are not v_1, v_3 have singleton input sets, v_1, v_3 have 1-bit inputs, v_1, v_2, v_3 have 1-bit outputs, and all other parties have singleton output sets. That is, effectively all the parties other than v_1, v_2, v_3 are irrelevant. By the same argument as above, $\mathcal{F} \subsetneq \mathcal{Q}$.

The gap in realizable behaviors proved in Theorem 10 also translates to a gap in achievable winning probabilities for the corresponding LC game. We consider the LC game $\mathcal{G} = (\mathcal{V}, \pi, G)$ where

$$\mathcal{V}(a_1, a_2, a_3|s_1, s_3) := \mathcal{V}_{\text{CHSH}}(a_1, a_3|s_1, s_3) \cdot [a_2 = a_1],$$

where $\mathcal{V}_{\text{CHSH}}$ is the winning condition for the CHSH game, and G is given by Equation (2.8). We also assume a uniform input distribution $\pi(s_1, s_3) := \frac{1}{4}$. We will call this the **extended CHSH game** as per [30], but as an LC game with partial communication allowed. We show the extended CHSH game schematically in Figure 7. We denote the maximum achievable winning probabilities



$a_1 = a_2$ and $a_1 \oplus a_3 = s_1 \wedge s_3$.

Figure 7: The extended CHSH game.

of classical, quantum, and forwarding strategies as $\omega_c, \omega_q, \omega_f$, respectively. Then, it is clear that $\omega_q \geq \cos^2(\pi/8)$ by simply letting v_1, v_3 implement an optimal quantum strategy for the CHSH game and then v_1 send his output to v_2, who simply outputs exactly that. This actually achieves the quantum value by the same argument as that of the distributed CHSH game. That is,

$$\omega_q = \cos^2(\pi/8).$$

However, it is easy to see that $\omega_f < \cos^2(\pi/8)$ again via self-testing. In fact, we can actually prove a stronger result:

$$\omega_f = \frac{3}{4},$$

the classical value of the CHSH game. To see this, let

$$p(a_1, a_2, a_3|s_1, s_3) = \langle \psi | \Pi_{1,a_1}(s_1) \otimes \Pi_{2,a_2}(s_1) \otimes \Pi_{3,a_3}(s_3) | \psi \rangle$$

be a forwarding behavior. By Theorem 9,

$$\omega_f(\mathcal{G}) = \omega_q(\mathcal{G}_1),$$

where $\mathcal{G}_1$ is a nonlocal game, namely the 1-extension [30] of the CHSH game. By Theorem 2.1 in [30], the classical and no-signaling values $\omega_c(\mathcal{G}_1), \omega_{\text{ns}}(\mathcal{G}_1)$ is equal to the classical value of the CHSH game, $\frac{3}{4}$. Thus, $\omega_q(\mathcal{G}_1) = \frac{3}{4}$. We conclude $\omega_f(\mathcal{G}) = \frac{3}{4}$ as desired. We can summarize this result with a proposition.

Proposition 11. *Let the LC game $\mathcal{G}$ be defined as above. Then,*

$$\omega_q(\mathcal{G}) = \cos^2(\pi/8) > \omega_f(\mathcal{G}) = \frac{3}{4}.$$

3 Multi-Step Latency-Constrained (LC) Games

Latency-constrained games as defined in Theorem 1 is the simplest extension of nonlocal games that arises from latency constraints. Effectively, it allows for only a single round of communication between a subset of the parties. To disambiguate, we will refer to this type of LC game as a **simple latency-constrained game**. Now, in real-world scenarios, the parties can have multiple rounds of communication, and may even receive inputs and create outputs over different rounds. An example is in HFT where different trading servers at different stock exchanges around the globe are communicating and making trades over an extended period of time. We would therefore like to define a multi-round extension of LC games. Now, communication between different parties may incur different latencies, for example due to different geographical distances. Hence, we need to introduce a notion of time. That is, we want an extension of LC games involving multiple *time steps*. We first define for a non-negative integer τ, $[0:\tau] \coloneqq \{0\} \cup [\tau]$.

Definition 12. *Let $n \geq 2$, $\tau \geq 0$ be integers. Let $S_i^{(t)}, A_i^{(t)}$ be finite sets for $i \in [n], t \in [0:\tau]$, and*

$$S^{(t)} \coloneqq \prod_{i=1}^{n} S_i^{(t)}, A^{(t)} \coloneqq \prod_{i=1}^{n} A_i^{(t)}.$$

We define a probabilistic predicate $\mathcal{V}$ which is a map

$$\mathcal{V} : \prod_{t=0}^{\tau} A^{(t)} \times \prod_{t=0}^{\tau} S^{(t)} \to [0,1].$$

Furthermore, we define the input distribution π which is a probability distribution on $\prod_{t=0}^{\tau} S^{(t)}$. Finally, let

$$\ell : [n] \times [n] \to \mathbb{Z}_+$$

be a function such that

$$\forall i \in [n],\ \ell(i,i) = 1.$$

*This will be called the **latency function**. Then, a **τ-step latency-constrained (LC) game** is the tuple $(\mathcal{V}, \pi, \ell)$.*

Here, $t \in [0:\tau]$ specifies the time step. Each party is to receive an input and produce an output at each time step from $t = 0$ to $t = \tau$. $S_i^{(t)}$ is the input set for party i at time step t and $A_i^{(t)}$ is the corresponding output set. The probabilistic predicate evaluates the total "score" of the combined outputs from all time steps of the parties given the combined inputs from all time steps. π describes the probability distribution of the different inputs. Finally, the latency function ℓ encodes the latency information. That is, $\ell(i,j)$ is the number of time steps required for party i to transmit information to party j. We will assume this is always positive.[8] Moreover, we define for all $i \in [n]$

$$\ell(i,i) = 1$$

[8] If $\ell(i,j) = \ell(j,i) = 0$ then party i and j are the same party. If one direction is positive and the other zero, defining a strategy becomes messy. We can also justify the positivity of the latency by the finite speed of light for parties that are spatially separated.

to more conveniently define strategies for τ-step LC games below. For the sake of generality we make no additional assumptions on the latency function. For example, it could be asymmetric or violate the triangle inequality and therefore is not a distance function. *When ℓ is the speed of light delay, then the corresponding realizable behaviors reflect fundamental physical limits.* Note here we are assuming that local operations consume zero time. The only operation that consumes time is communication between the parties, which is given by the latency function ℓ. As mentioned in Section 1, this is justified by giving the parties more energy, thereby allowing them to arbitrarily shorten their local operation times. Our assumption is also a good approximation to real-world scenarios where latencies are much longer than local operation times. An example would be where the parties are physically far from each other. We can define a **behavior** $p(\mathbf{a}|\mathbf{s})$ and corresponding **winning probability** p_{win} for a τ-step LC game analogously to the LC game case in Section 2.

In Theorem 12 we seem to be assuming that time is discrete and that receiving inputs, producing outputs, and communication are done in a synchronous manner. However, the definition is more general than it may seem. The time values we want to include in our definition are the communication latencies and the times at which the parties need to receive inputs and produce outputs. Without loss of generality, we can assume all of these time values are rational numbers since the rationals are a dense subset of the real numbers and all physical clocks must have a finite precision. Let the set of all time values be

$$V_T \subseteq \mathbb{Q}.$$

This set is finite. We can then set one time step as a length of time

$$t_0 \coloneqq \frac{1}{\operatorname{lcm}\{q : p/q \in V_T, \gcd(p,q) = 1\}}, \tag{3.1}$$

where $\operatorname{lcm}, \gcd$ are short for least common multiple and greatest common denominator, respectively. It is clear that all time values in V_T are integer multiples of t_0. For cases where parties are receiving inputs, producing outputs, and communicating asynchronously, we can simply let $S_i^{(t)}$ or $A_i^{(t)}$ to be singletons when there is no input or output, respectively, at time step t for party i. The parties can also choose to communicate nothing at a particular time step. This will depend on the specific strategy they choose to implement, which is what we will next discuss. Note that a 0-step LC game is just a nonlocal game. A 1-step LC game where $S_i^{(1)}, A_i^{(0)}$ are singletons is a simple LC game.

Our τ-step LC games framework can be used in real-world scenarios by specifying the points of time at which inputs are received and outputs are produced for a set of spatially separated parties with corresponding communication latencies. This determines the input sets $S_i^{(t)}$ and output sets $A_i^{(t)}$, as well as the latency function ℓ. The probabilistic predicate $\mathcal{V}$ then specifies what is the desired time-series of outputs given the time-series of inputs, while the π is the prior on the inputs. By optimizing the winning probability using classical strategies versus quantum strategies, we can determine whether a quantum advantage is possible. This framework is a lot more natural than [5] because real-world scenarios often involve a sequence of inputs and outputs and can involve communication (with latency). Hence, we expect our framework to be useful for identifying real-world applications.

We also mention that our τ-step LC games framework goes beyond conventional nonlocal games *even in the two-party setting.* This is because the two parties receive multiple inputs and produce multiple outputs, and they can also communicate, albeit with latency. For example, this can be used to analyze real-world scenarios where outputs in the past can affect the game being played in the present. We say such games has **memory**. This can be applicable to scenarios such as [7], where previous routing of customer requests can have effects later in time.

For general τ-step LC games we will mostly make definitions and prove some simple properties. Future work will be to establish a more comprehensive theory of τ-step LC games. Such a theory could possibly make use of mathematical tools from works such as [25].

3.1 Classical strategies

As we argued in Section 2.1, we can always copy classical information, so in the deterministic setting the only nontrivial information that the parties can transmit to each other are their inputs. Hence, we can define a deterministic strategy for a τ-step LC game as follows.

Definition 13. *Let $(\mathcal{V}, \pi, \ell)$ be a τ-step LC game. Let*

$$H_i^{(t)} := \prod_{t' \leq t} S_i^{(t')} \times \prod_{\substack{j \neq i \\ \ell(i,j) \leq t}} \prod_{t'=0}^{t-\ell(i,j)} S_j^{(t')}$$

*be the **past light cone of party i at time step t**. A **deterministic strategy** is given by functions $\{f_i^{(t)}\}_{i \in [n], t \in [0:\tau]}$, where*

$$f_i^{(t)} : H_i^{(t)} \to A_i^{(t)}.$$

That is, in a deterministic strategy, each party produces an output at time step t given all his past inputs and the inputs of all parties that can be sent to him by time step t.

The behavior of a deterministic strategy is simply

$$p(\mathbf{a}|\mathbf{s}) := \prod_{t=0}^{\tau} \prod_{i=1}^{n} \delta_{a_i^{(t)}, f_i^{(t)}(h_i^{(t)})},$$

where $h_i^{(t)}$ is the element of $H_i^{(t)}$ made up of corresponding elements in $\mathbf{s}$. Using the same proof as Theorem 3, we can establish the following.

Proposition 14. *Given a τ-step LC game $(\mathcal{V}, \pi, l)$, the number of possible deterministic behaviors is*

$$\prod_{t=0}^{\tau} \prod_{i=1}^{n} |A_i^{(t)}|^{|H_i^{(t)}|}. \tag{3.2}$$

A **classical strategy** is a probabilistic mixture of deterministic strategies where shared randomness can be used. Since the set of all classical behaviors is by definition the convex hull of the set of deterministic behaviors, it is a polytope with the number of vertices given in Equation (3.2). Again, via the Weyl-Minkowski theorem, *we can obtain the fundamental bounds on classical correlations on parties receiving multiple inputs and producing multiple outputs over time and each party having different communication latencies.*

3.2 Quantum strategies

The core mathematical object describing a quantum strategy for a τ-step LC game involves a set of **interaction tensors** W. For a τ-step LC game with n parties, party i at time step t uses the interaction tensor $W_i^{(t)}$ which has $n+1$ input legs and $n+1$ output legs. This is shown diagrammatically in Figure 8. The first input leg will take in the input from $S_i^{(t)}$, having dimension

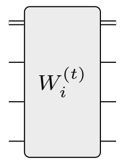


Figure 8: The interaction tensor $W_i^{(t)}$ for party i at time step t.

$|S_i^{(t)}|$. The first output leg will yield a probability distribution on the outputs in $A_i^{(t)}$, having dimension $|A_i^{(t)}|$. These legs correspond to classical systems. The other n output legs are quantum systems sent to other parties (including party i himself). That is, these connect with input legs of other interaction tensors in the future (higher t). Specifically, the j-th quantum output leg of $W_i^{(t)}$ will connect with the i-th quantum input leg of $W_j^{(t+\ell(i,j))}$. Also, since we assumed $\ell(i,i) = 1$ for all $i \in [n]$, the quantum system that party i sends to himself can just be regarded as a local quantum register that is left behind and is sent to the interaction tensor for the next time step: $W_i^{(t+1)}$.

This specifies the connections of all the legs except for interaction tensors near the initial and final time steps. An "early-time" interaction tensor of party i at time step t may have an input leg corresponding to party j that does not connect with an interaction tensor where $t - \ell(j,i) < 0$. That is, no quantum system was transmitted to the interaction tensor via that leg because not enough time elapsed for an initial transmission from j to i. We will assume these legs are trivial (dimension 1). Similarly, the "late-time" interaction tensors of party i at time step t may have an output leg corresponding to party j that does not connect with an interaction tensor where $t + \ell(i,j) > \tau$. That is, no interaction tensor receives their transmission because the τ-step LC game already ended by the time the transmission is completed. We will also assume these legs are trivial.

To keep everything physical, we require that for any interaction tensor $W_i^{(t)}$, for a fixed classical input leg index $s_i^{(t)}$, the map from the remaining (quantum) input legs to the remaining output legs (including the classical leg) is an isometry. With these interaction tensors and the initial shared quantum state, we can define a quantum strategy:

Definition 15. *Let $(\mathcal{V}, \pi, \ell)$ be a τ-step LC game. A **quantum strategy** is a tuple*

$$(|\psi\rangle, \{W_i^{(t)}\}_{i\in[n],t\in[0:\tau]}).$$

Hence, the behavior of a quantum strategy is given by

$$p(\mathbf{a}|\mathbf{s}) = \||\psi(\mathbf{a}|\mathbf{s})\rangle\|^2$$

where this time

$$\mathbf{a} \coloneqq (a_i^{(t)})_{i\in[n],t\in[0:\tau]},\ \mathbf{s} \coloneqq (s_i^{(t)})_{i\in[n],t\in[0:\tau]}$$

and

$$|\psi(\mathbf{a}|\mathbf{s})\rangle \coloneqq \prod_{t=0}^{\tau}\left[\bigotimes_{i=1}^{n}\langle a_i^{(t)}|W_i^{(t)}|s_i^{(t)}\rangle\right]|\psi\rangle, \tag{3.3}$$

the matrix product being taken from right to left. The sandwiching of the $W_i^{(t)}$ interaction tensor in Equation (3.3) sets the classical input leg index to be $s_i^{(t)} \in S_i^{(t)}$ and the classical output leg index to be $a_i^{(t)} \in A_i^{(t)}$. Note that $|\psi(\mathbf{a}|\mathbf{s})\rangle$ is not a normalized vector. *A quantum strategy captures the most general quantum protocol that the parties can implement in τ time steps, where $\ell(i,j)$ is how long it takes for i to signal to j.* In Figure 9, we draw a diagram of a quantum strategy for three parties where $\ell(i,j) = \max\{1, |i-j|\}$. Note again that *it is not strictly necessary for the parties to send quantum systems in a hardware implementation*: they could simply perform quantum teleportation by using shared entanglement and communicating classical information.

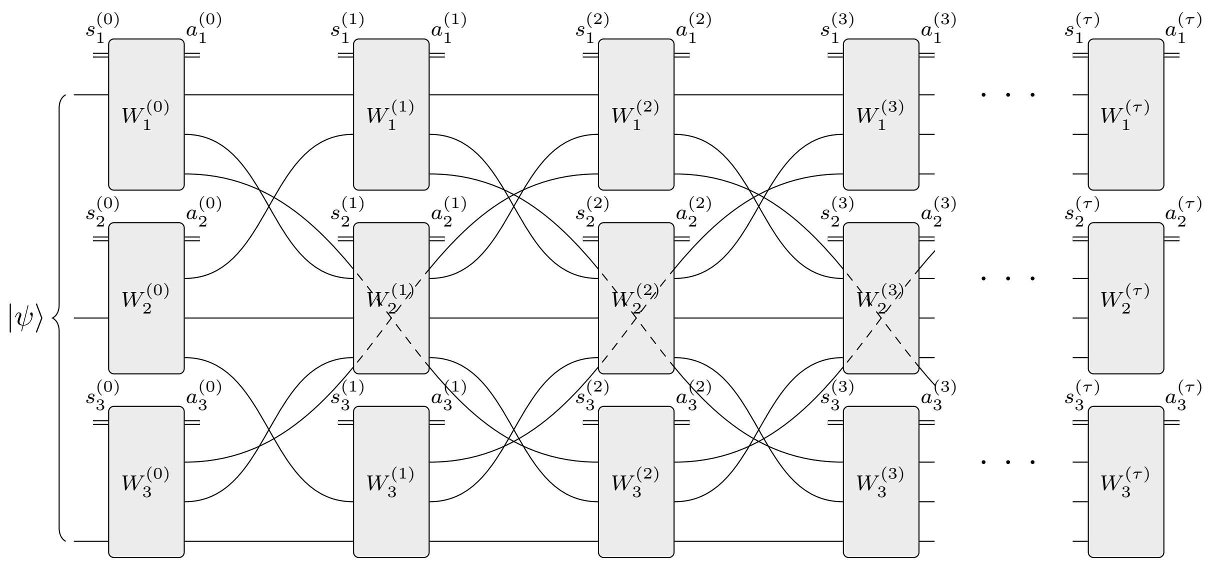


Figure 9: A quantum strategy for a three-party, τ-step latency-constrained game, where the latency function is given by $\ell(i,j) = \max\{1, |i-j|\}$. The wires represent communication channels, with time flowing to the right. For instance, the first output wire of $W_1^{(0)}$ corresponds to the local register of party v_1, the second wire takes one time step to reach party v_2, and the third wire takes two time steps to reach v_3. Trivial legs are omitted.

Let $\mathcal{Q}$ be the set of all quantum behaviors in a τ-step LC game $(\mathcal{V}, \pi, \ell)$. Similar to Theorem 6, we can prove the convexity of this set.

Proposition 16. *For any τ-step LC game $(\mathcal{V}, \pi, \ell)$, the set $\mathcal{Q}$ is convex.*

Proof. Consider a convex combination of two quantum behaviors in $\mathcal{Q}$:

$$\xi(\mathbf{a}|\mathbf{s}) = \lambda p(\mathbf{a}|\mathbf{s}) + (1-\lambda)p'(\mathbf{a}|\mathbf{s}),$$

where

$$p(\mathbf{a}|\mathbf{s}) = \||\psi(\mathbf{a}|\mathbf{s})\rangle\|^2, \quad |\psi(\mathbf{a}|\mathbf{s})\rangle = \prod_{t=0}^{\tau}\left[\bigotimes_{i=1}^{n}\langle a_i^{(t)}|W_i^{(t)}|s_i^{(t)}\rangle\right]|\psi\rangle,$$

$$p'(\mathbf{a}|\mathbf{s}) = \||\psi'(\mathbf{a}|\mathbf{s})\rangle\|^2, \quad |\psi'(\mathbf{a}|\mathbf{s})\rangle = \prod_{t=0}^{\tau}\left[\bigotimes_{i=1}^{n}\langle a_i^{(t)}|W_i'^{(t)}|s_i^{(t)}\rangle\right]|\psi'\rangle.$$

By replacing the underlying quantum system with the direct sum of the original two systems, we can reconstruct $\xi(\mathbf{a}|\mathbf{s})$:

$$\xi(\mathbf{a}|\mathbf{s}) = \||\phi(\mathbf{a}|\mathbf{s})\rangle\|^2,$$

where

$$|\phi(\mathbf{a}|\mathbf{s})\rangle = \prod_{t=0}^{\tau}\left[\bigotimes_{i=1}^{n}\langle a_i^{(t)}|\left(W_i^{(t)} \oplus W_i'^{(t)}\right)|s_i^{(t)}\rangle\right]\left(\sqrt{\lambda}|\psi\rangle \oplus \sqrt{1-\lambda}|\psi'\rangle\right).$$

The direct sums can be defined similarly to Theorem 6.

□

3.3 Single-input single-output (SISO) LC games

Although simple LC games extend Bell inequalities by relaxing the latency constraint, it only assumes a single round of communication between the parties. To include multiple rounds of communication, we need to define a special class of τ-step LC games:

Definition 17. *A τ-step LC game $(\mathcal{V}, \pi, \ell)$ is **single-input single-output (SISO)** if $\forall t > 0$, $S_i^{(t)}$ are singletons and $\forall t < \tau$, $A_i^{(t)}$ are singletons.*

That is, the parties only receive inputs in the very beginning at time step 0, and they only give outputs at the very end at time step τ. In between they are able to transmit information to each other, but they do not receive inputs or give outputs. Therefore this is almost equivalent to conventional nonlocal games, the only difference being the addition of a latency function. For conciseness, we will define

$$S_i := S_i^{(0)}, A_i := A_i^{(\tau)}.$$

Note that a 1-step SISO game[9] is a simple LC game. In Figure 10 we show a 2-step SISO game with 3 parties where $\ell(1,2) = \ell(2,1) = 1$, $\ell(2,3) = \ell(3,2) = 2$, and $\ell(1,3), \ell(3,1) > 2$.

A SISO game is naturally interpreted as multiple parties that receive inputs and need to produce their outputs *within latency constraint τ, given their pairwise communication latencies.* If τ is too small, no communication is possible, and this becomes a conventional nonlocal game. On the other hand, if the τ is sufficiently large for all the parties to receive all the inputs, then the game is trivial. That is, we can prove:

Proposition 18. *Let $(\mathcal{V}, \pi, \ell)$ be a SISO, τ-step LC game. Suppose $\tau \geq \ell_{\max}$. Then, a classical strategy can realize all possible behaviors. In particular, it can achieve the maximum algebraic value ω_a.*

[9]For conciseness we will sometimes omit "LC".

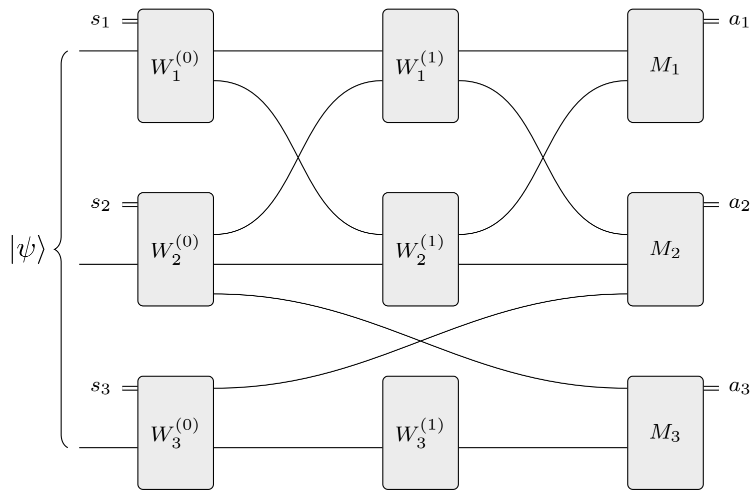


Figure 10: A 2-step SISO game with 3 parties where $\ell(1,2) = \ell(2,1) = 1$, $\ell(2,3) = \ell(3,2) = 2$, and $\ell(1,3), \ell(3,1) > 2$. For clarity we denote the last isometries as M since they are essentially measurements.

Proof. Since, $\tau \geq \ell_{\max}$, all parties have access to all inputs. The rest follows the proof of Theorem 4. □

This is similar to the result in Theorem 4. However, if the time is not sufficiently long, then only limited communication is possible. In fact, by looking at the classical value $\omega_c(\tau)$ or quantum value $\omega_q(\tau)$ for the τ-step SISO game $(\mathcal{V}, \pi, \ell)$ for different τ, ceteris paribus, we are actually computing the highest possible winning probability achievable with classical or quantum resources *under varying latency constraints*, thereby answering the questions posed in Section 1. This can also be interpreted as the achievable winning probability with classical or quantum resources *as a function of time*. It is clear that this function is monotonically non-decreasing. We can then take the difference $\omega_q(\tau) - \omega_c(\tau)$ to evaluate the quantum advantage as a function of time.

We can also define a quantum advantage from the other direction. We can define $\tau_c(\alpha)$ for $\alpha \in [0, \omega_a]$ as the **$\boldsymbol{\alpha}$-classical threshold time**: the minimum number of time steps for a classical strategy to achieve a winning probability of α or higher. We can similarly define $\tau_q(\alpha)$ as the **$\boldsymbol{\alpha}$-quantum threshold time**. If

$$\tau_q(\alpha) < \tau_c(\alpha),$$

for some α, then we have a *mathematically provable time advantage using quantum resources* as described in Section 1 for this SISO game and specific α. Note that Theorem 18 implies that for all $\alpha \in [0, \omega_a]$,

$$\tau_c(\alpha), \tau_q(\alpha) \leq \ell_{\max}.$$

For a SISO game, we observe that both classical and quantum strategies lead to behaviors which satisfy a relaxation of the no-signaling condition. Namely, parties that are within each others' light cones can communicate, but the behavior of parties outside of each others' light cones must be no-signaling with respect to each other. Just like simple LC games, information regarding which parties are in each others' light cones can be captured by a directed[10] graph:

[10]This corresponds to the fact that we did not assume $\ell(i,j)$ is symmetric.

Definition 19. *Let $(\mathcal{V}, \pi, \ell)$ be a SISO, τ-step LC game. Its* ***connectivity graph*** $G = (V, E)$ *is defined as:*

$$V \coloneqq [n]$$

and

$$E \coloneqq \{(i, j) : i, j \in V, i \neq j, \ell(i, j) \leq \tau\}.$$

That is, the connectivity graph's vertices are the n parties, and two vertices are connected if one can communicate to the other within τ time steps. Hence, the connectivity graph describes which parties can communicate with which other parties. We then define the generalization of no-signaling behaviors as follows:

Definition 20. *Let $G = (V, E)$ be a directed graph. We say a behavior $p(\mathbf{a}|\mathbf{s})$, is* ***G-signaling*** *if for fixed $i \in [n]$, $s_1, s_2, \cdots, s_{i-1}, s_{i+1}, \cdots, s_n$,*

$$\sum_{\substack{a_j \in A_j \\ j \in N_{\text{out}}[i]}} p(a_1, a_2, \cdots, a_n | s_1, s_2, \cdots, s_{i-1}, s_i, s_{i+1}, \cdots, s_n)$$

is independent of $s_i \in S_i$.

In words, this is saying that the marginal distribution of the outputs for parties that are not out-neighbors of a party is independent of the input of that party. It is clear that quantum and classical strategies for a SISO game realize behaviors that are G-signaling, where G is the connectivity graph. Note that $(V, \emptyset)$-signaling is equivalent to no-signaling. Furthermore, any marginal of a G-signaling behavior that only involve parties that are not adjacent (in either direction) of each other is no-signaling for every possible set of inputs of the parties that were traced out. That is, let $V' \subseteq V$ such that

$$\forall i, j \in V', (i, j) \notin E.$$

Then, for fixed $\{s_i\}_{v_i \notin V'}$

$$\sum_{\substack{a_i \in A_i \\ v_i \notin V'}} p(a_1, a_2, \cdots, a_n | s_1, s_2, \cdots, s_n)$$

is a no-signaling behavior.

We can prove an upper bound on the dimension of the set of G-signaling behaviors $\mathcal{B}_G$. Based on the notations introduced in Section 2.1, for any $X \subseteq V$, define

$$N_{\text{in}}[X] \coloneqq \bigcup_{i \in X} N_{\text{in}}[i].$$

Proposition 21.

$$\dim \mathcal{B}_G \leq \sum_{X \in \mathcal{P}^+(V)} \left(\prod_{w \in N_{\text{in}}[X]} |S_w| \right) \left(\prod_{t \in X} (|A_t| - 1) \right),$$

where $\mathcal{P}^+(V)$ is the set of all nonempty subsets of V.

Proof. For simplicity, in this proof, we assume $A_i = \{0, 1, ..., |A_i| - 1\}$. The upper bound in this proposition can be proven by an explicit parametrization in [31] from an observation that all conditional probabilities involving the outcome 0 can be reconstructed using the normalization conditions and signaling conditions. For example, from the signaling condition

$$\sum_{a_j=0}^{|A_j|-1} p(a_i, a_j|h_i, h_j) = p(a_i|h_i),$$

one may express

$$p(a_i, 0|h_i, h_j) = p(a_i|h_i) - \sum_{a_j=1}^{|A_j|-1} p(a_i, a_j|h_i, h_j).$$

In this parametrization scheme, for each subset $X \subseteq V$, there are $\prod_{w\in N_{\text{in}}[X]} |S_w|$ input possibilities and for each input scenario, there are $\prod_{t\in X}(|A_t|-1)$ specifiable output probabilities. The summation over all the subsets then gives an upper bound on the dimension.

□

3.3.1 Aggregating parties

In latency-constrained settings, we often encounter situations where a group of parties can communicate, while different groups cannot communicate with each other. This is similar to the setup in Svetlichny scenarios [24]. To motivate this from a latency standpoint, we can imagine parties are located in various cities. Parties in the same city can communicate at low latencies due to geographical proximity, while parties located in different cities communicate with much higher latencies. In this scenario, it is natural to ask when it is possible to "aggregate" multiple parties. This is a concept that appeared in some of our examples above, including the distributed CHSH game in Section 2.3.1. That is, in what situations can we replace a group of parties by one party that aggregates their inputs and outputs? We give sufficient conditions for the specific case of SISO games.

Define a **bidirected clique** of a directed graph as a subgraph that is itself a complete directed graph (every pair of vertices is connected in both directions). Now, it is clear that for classical strategies for a SISO game, the connectivity graph G actually encodes all the information we need, since the past light cone of a party is fully determined by G. Since SISO games only have a single input and output set just like conventional nonlocal games, to more succinctly state our results below, we will sometimes refer to a classical strategy for a SISO game with connectivity graph G as a ***G*-classical strategy** for the corresponding nonlocal game. Then, we can show that parties in an isolated, bidirected clique of G can be collapsed to one party.

Proposition 22. *Let $(\mathcal{V}, \pi, \ell)$ be a SISO, τ-step LC game and let $G = (V, E)$ be the connectivity graph. Let $C = (V_C, E_C) \subseteq G$ be a bidirected clique that is isolated, that is, vertices in C are not connected to vertices outside of C. We define G' where we replace C in G by a single isolated vertex v_C. We define*

$$S_{v_C} \coloneqq \prod_{v'\in V_C} S_{v'}$$

and

$$A_{v_C} \coloneqq \prod_{v' \in V_C} A_{v'}.$$

Then, $p(\mathbf{a}|\mathbf{s})$ is a G-classical behavior for the nonlocal game with the original input and ouput sets iff it is also a G'-classical behavior for the nonlocal game with the input and output sets aggregated as above.

Proof. Consider the case of a deterministic strategy. Every party $v' \in V_C$ receives the inputs of all the other parties in V_C. Hence, each output $a_{v'}$ is a function of S_{v_C}. The collective output $\prod_{v' \in V_C} a'_v$ is in A_{v_C}. Thus the statement is true for deterministic strategies. We can then take convex combinations of deterministic strategies to obtain our conclusion. □

For example, we can start with a graph

$$G = v_1 \leftrightarrows v_2 \quad v_3$$

and after aggregation we obtain the graph

$$G' = v_C \quad v_3,$$

where v_C has input set $S_1 \times S_2$ and output set $A_1 \times A_2$.

In the quantum case, a similar statement would need to involve the latency function ℓ. Specifically, if there is a group of parties that cannot communicate with other parties, and there is a party within the group who has enough time to receive inputs from and send back outputs to all other parties within the group, then we can aggregate that group into one party. We state this as a proposition.

Proposition 23. *Let $(\mathcal{V}, \pi, \ell)$ be a SISO, τ-step LC game and $G = (V, E)$ its connectivity graph. Suppose for some $V_C \subseteq V$, for all $i \in V_C$, $j \in V \setminus V_C$,*

$$(i, j), (j, i) \notin E.$$

Further suppose there exists $v_0 \in V_C$ such that for all $i \in V_C$,

$$\ell(i, v_0) + \ell(v_0, i) \leq \tau \tag{3.4}$$

Now, replace all the parties in V_C with one party v_C with input and output sets

$$S_{v_C} \coloneqq \prod_{i \in V_C} S_i$$

and

$$A_{v_C} \coloneqq \prod_{i \in V_C} A_i,$$

respectively. We define the latency function

$$\ell'(j, k) \coloneqq \ell(j, k)$$

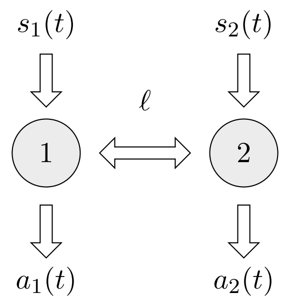


Figure 11: A continuous-time LC game with two parties.

for $j, k \in V \setminus V_C$ *and*

$$\ell'(v_C, j), \ell'(j, v_C) > \tau$$

for all $j \neq v_C$.[11] *Then,* $p(\mathbf{a}|\mathbf{s})$ *is a quantum behavior for* $(\mathcal{V}, \pi, \ell)$ *iff it is also a quantum behavior, aggregating the inputs and outputs as above, for the* τ*-step LC game* $(\mathcal{V}', \pi', \ell')$ *where* $\mathcal{V}'$ *and* π' *are obtained from* $\mathcal{V}, \pi$ *by replacing the parties in* V_C *with one party* v_C *and using the above question and answer sets.*

Proof. The forward direction is trivial because anything the parties in V_C can do, since they cannot transmit anything to other parties, can be captured by an S_{v_C}-dependent measurement on their overall share of the quantum state with outcomes in A_{v_C}.

As for the backwards direction, we observe that party v_C in the aggregated setting is completely isolated and simply can only do an S_{v_C}-dependent measurement on his share of the quantum state with results in A_{v_C}. Going to the non-aggregated setting, we can give $v_0 \in V_C$ the entire share of the quantum state owned by v_C. At time step 0, each party in V_C sends his input to v_0. v_0 then performs the S_{v_C}-dependent measurement with results in A_{v_C} that v_C uses in the aggregated setting. The measurement results are then sent to all the parties in V_C. By Equation (3.4), they all receive the measurement results before time step τ and produce their respective outputs at time step τ. The conclusion follows. $\square$

3.4 Continuous-time LC games

Although our definition of τ-step LC games can approximate physical latency-constrained scenarios arbitrarily well, it is still worthwhile to consider how an LC game should be formulated in the continuous-time setting where the parties are continuously receiving inputs and creating outputs. This would be more natural from a physics perspective. In the same spirit, we will also assume inputs and outputs are continuous, in particular real-valued [32]. We will describe in words how continuous-time LC games can be formulated. We leave for future work a more formal mathematical formulation and subsequent analyses.

For simplicity, we consider the two-party case as drawn in Figure 11. Let $\tau \geq 0$ be a real number and let $\mathcal{S}_1(t), \mathcal{S}_2(t)$, where $t \in [0, \tau]$, be continuous-time stochastic processes that take values in $\mathbb{R}$. We interpret these as inputs to v_1 and v_2, respectively. Likewise, we let $\mathcal{A}_1(t), \mathcal{A}_2(t)$ be continuous-time stochastic processes that take values in $\mathbb{R}$. These will be the outputs of v_1 and v_2,

[11] Any values larger than τ suffices, enforcing the fact that v_C cannot communicate with other parties.

respectively. Let $s_1(t), s_2(t)$ and $a_1(t), a_2(t)$ be real-valued functions on $[0, \tau]$. We are interested in the **continuous-time behavior**

$$p(\mathcal{A}_1(t) = a_1(t), \mathcal{A}_2(t) = a_2(t) \text{ for } t \leq \tau | \mathcal{S}_1(t) = s_1(t), \mathcal{S}_2(t) = s_2(t) \text{ for } t \leq \tau).$$

That is, we are interested in the probability of a *trajectory* of the outputs $a_1(t), a_2(t)$ for $t \in [0, \tau]$ given the trajectory of the inputs $s_1(t), s_2(t)$. Thus, unlike τ-step LC games where the behavior $p(\mathbf{a}|\mathbf{s})$ is a function whose domain is a Cartesian product of discrete sets S_i, A_i, in the continuous-time setting, the behavior is a function whose domain is a Cartesian product of the space of functions

$$\{f : [0, \tau] \to \mathbb{R}\}.$$

Then, we can define the latency function ℓ in the continuous-time setting as real-valued:

$$\ell(v_1, v_2), \ell(v_2, v_1) \in \mathbb{R}_+,$$

where $\ell(v_1, v_2)$ is the latency from v_1 to v_2 and vice versa.

We can then define strategies as involving continuous-time classical or quantum operations. Again we will assume local operations consume zero time. This is somewhat extreme in the continuous-time setting but as we discussed in Section 1, this can be justified by simply increasing the energy. In the end, we are concerned with fundamental physical limits:

- Given all possible classical physical resources (including arbitrarily large energies), what correlations can be exhibited?
- Given all possible quantum physical resources (including arbitrarily large energies), what correlations can be exhibited?

Our assumption therefore is consistent with the premise of these questions.

A classical strategy can include actions such as v_1 at time t sending classical information to v_2, such as his input $s_1(t)$ or his output $a_1(t)$. v_2 will receive the information at time $t + \ell(v_1, v_2)$. v_2 can apply a function on this information and other information that he has to produce his output $s_2(t + \ell(v_1, v_2))$. Similarly, quantum strategies can include actions such as v_2 locally preparing and sending a quantum system $|\psi\rangle_{B_2}$ at time t to v_1. v_1 would receive this quantum system at time $t + \ell(v_2, v_1)$. He can locally perform a measurement $\{\Pi_{1,a_1}\}_{a_1}$ on $|\psi\rangle_{B_2}$ together with some quantum system on his local quantum register to obtain his output $a_1(t + \ell(v_2, v_1))$. In essence, these strategies are the continuous-time versions of classical and quantum strategies for τ-step LC games.

The above two questions about fundamental physical limits then translate to the mathematical questions:

"What is the set of continuous-time behaviors realized by classical strategies?"
"What is the set of continuous-time behaviors realized by quantum strategies?"

We leave the rigorous formulation and answering of these mathematical questions to future work.

4 Numerical Techniques and Examples

We want to answer the second, and equivalently third, question in Section 1 of how much quantum resources, which are the most general physical resources, can violate the bounds on classical correlations in the latency-constrained games setting. More generally, such a question naturally extends to SISO games. That is, we want to numerically compute the quantum value ω_q for arbitrary simple LC games and SISO games and compare it to the classical value ω_c.

In general, optimizing conventional nonlocal games is already a computationally hard problem [33, 34, 35]. Hence, we can only hope to provide some heuristic algorithms. Standard methods for nonlocal games, like the see-saw algorithm [36] or the Navascues-Pironio-Acín (NPA) algorithm [37, 38], may also apply to specific LC and SISO games, depending on their graph G. For example, in the case where G is the empty graph or we can aggregate parties in a SISO game such that it reduces to a conventional nonlocal game. In general, however, such methods cannot be used directly on games that cannot be reduced to conventional nonlocal games.

For this reason, we provide in Section 4.1 a generalization of the see-saw algorithm to compute lower bounds on the quantum values of arbitrary simple LC games. Such a method can naturally extend to SISO games. Upper bounds on the quantum value of LC games could be obtained by a generalization of the NPA hierarchy along the lines of [39], for which we conjecture a possible approach in Section B.

4.1 Lower bounds via a see-saw algorithm

We can compute lower bounds on the quantum value ω_q of any LC game via a generalization of the see-saw algorithm [36] used to compute lower bounds for conventional nonlocal games. We first re-write the behavior Equation (2.2) in the mixed quantum formalism:

$$p(\mathbf{a}|\mathbf{s}) = \operatorname{tr}\left[\bigotimes_{i=1}^{n}\Pi_{i,a_i}\left(\bigotimes_{i=1}^{n}\mathcal{W}_i(s_i)\right)(\rho)\right],$$

where, writing L as the space of linear operators on a Hilbert space,

$$\rho \in L\Big(\bigotimes_{i=1}^{n} B_i\Big)$$

is a quantum state and where, for each party i,

$$\mathcal{W}_i(s_i) : L(B_i) \to L(B_i^{\text{out}})$$

is a quantum channel parametrized by s_i, and $\Pi_i = \{\Pi_{i,a_i}\}_{a_i \in A_i}$ is a projective measurement with projectors $\Pi_{i,a_i} \in L\left(B_i^{\text{in}}\right)$.

Together, a parametrized quantum channel and a projective measurement define the "local" strategy of a party i. A key observation is that such a strategy is formally equivalent to a one-slot quantum comb [40, 41, 25], which are higher-order quantum operations [42] corresponding to the combination of two quantum operations connected by memory link, which in this case is the space $B_{i\to i}$.

To represent the strategy of a party as a single higher-order operation we use the "link product" [40, 41] denoted '$*$', which represents the composition of operators at the level of their Choi matrices [43, 44]. It is defined for any matrices $X_{CD} \in L(CD)$ and $Y_{DE} \in L(DE)$ by

$$X_{CD} * Y_{DE} \coloneqq \operatorname{tr}_D \left[(X_{CD} \otimes I_E)^{\top[D]} \cdot (I_C \otimes Y_{DE}) \right] \in L(CE),$$

where $\cdot^{\top[D]}$ and tr_D are respectively the partial transpose and the partial trace on the space D, and where I_E is the identity on space E. We start by writing $\mathsf{W}_i(s_i)$ as the non-normalized Choi operator of $\mathcal{W}_i(s_i)$ defined as

$$\mathsf{W}_i(s_i) \coloneqq \sum_{j,k=1}^{\dim B_i} |j\rangle\langle k| \otimes \mathcal{W}_i(s_i)(|j\rangle\langle k|) \in L\left(B_i \otimes B_i^{\text{out}}\right).$$

The local strategy of party i can then be written as a collection of operators $\{\mathsf{K}_{i,a_i}(s_i)\}_{a_i,s_i}$ with $\mathsf{K}_{i,a_i}(s_i) \in L\left(B_i \otimes B_{(i)}^{\text{out}} \otimes B_{(i)}^{\text{in}}\right)$ where

$$B_{(i)}^{\text{in}} := \bigotimes_{j\in N_{\text{in}}(i)} B_{j\to i},\; B_{(i)}^{\text{out}} := \bigotimes_{j\in N_{\text{out}}(i)} B_{i\to j},$$

effectively removing the memory link $B_{i\to i}$ (see the definitions of $N_{\text{in}}(i)$ and $N_{\text{out}}(i)$ in Section 2.1). These operators are defined as

$$\begin{aligned}\mathsf{K}_{i,a_i}(s_i) &\coloneqq \mathsf{W}_i(s_i) * \Pi_{i,a_i},\\ &= \operatorname{tr}_{B_{i\to i}}\left[\left(\mathsf{W}_i(s_i)^{\top[B_{i\to i}]} \otimes I_{B_{(i)}^{\text{in}}}\right) \cdot \left(I_{B_i} \otimes I_{B_{(i)}^{\text{out}}} \otimes \Pi_{i,a_i}\right)\right],\end{aligned}$$

By construction, because the link product of two positive matrices is positive [40, 41], the $\mathsf{K}_{i,a_i}(s_i)$ are positive matrices. Furthermore, the property that they can be decomposed in two quantum operations $\mathsf{W}_i(s_i)$ and Π_i can be characterized by a set of normalization and projective constraints [40, 41, 25, 42]. More precisely, writing $\mathsf{K}_i(s_i) = \sum_{a_i\in A_i} \mathsf{K}_{i,a_i}(s_i)$, for any $s_i \in S_i$, the set $\{\mathsf{K}_{i,a_i}(s_i)\}_{a_i}$ defines a local strategy for party i if and only if it satisfies

$$\operatorname{tr} \mathsf{K}_i(s_i) = \dim B_i \dim B_{(i)}^{\text{in}} \tag{4.1}$$

$$\operatorname{tr}_{B_{(i)}^{\text{in}}}[\mathsf{K}_i(s_i)] \otimes I_{B_{(i)}^{\text{in}}} = \dim B_{(i)}^{\text{in}} \mathsf{K}_i(s_i) \tag{4.2}$$

$$\operatorname{tr}_{B_{(i)}^{\text{in}} \otimes B_{(i)}^{\text{out}}} \mathsf{K}_i(s_i) = \dim B_{(i)}^{\text{in}} I_{B_i} \tag{4.3}$$

$$\text{and } \mathsf{K}_{i,a_i}(s_i) \geq 0, \text{ for all } a_i \in A_i. \tag{4.4}$$

These are all permissible constraints for a semidefinite program (SDP). We can now rewrite behavior Equation (2.2) in the Choi's picture as[12]

$$p(\mathbf{a}|\mathbf{s}) = \mathsf{K}_{1,a_1}(s_1) * \cdots * \mathsf{K}_{N,a_N}(s_N) * \rho.$$

The see-saw algorithm is then based on the bound

$$\omega_q \geq \sum_{s\in S} \pi(s) \sum_{a\in A} \mathcal{V}(a|s)\big(\mathsf{K}_{1,a_1}(s_1) * \cdots * \mathsf{K}_{N,a_N}(s_N) * \rho\big), \tag{4.5}$$

[12]Note the link product is associative.

and the outline of the algorithm is as follows. First, we set the dimensions of all the quantum systems $B_i, B_{i\to j}$. Then, we generate a random density matrix ρ by taking a Haar random pure state of dimension $(\dim B_i)^2$ of which we trace out a $\dim B_i$-dimensional system.

We also generate random positive semidefinite matrices $\{K_{i,a_i}(s_i)\}_{a_i,s_i}$ for each party i by generating a random operator with an ancillary system $B_{|A_i|}$ of dimension $|A_i|$, that we trace out to generate $|A_i|$ random elements. This can be thought of as generating operators where each party's measurements had been purified. More precisely, we generate Haar random pure states of dimension $(\dim B_i \dim B^{\text{out}}_{(i)} \dim B^{\text{in}}_{(i)}\ |A_i|)^2$ of which we trace out a $(\dim B_i \dim B^{\text{out}}_{(i)} \dim B^{\text{in}}_{(i)}\ |A_i|)$-dimensional system, defining operators $\overline{K}_i(s_i) \in L(B_i \otimes B^{\text{out}}_{(i)} \otimes B^{\text{in}}_{(i)} \otimes B_{|A_i|})$. We then normalize these so as to satisfy Equation (4.1) and project them to obtain operators $\widetilde{K}_i(s_i)$ such that $\operatorname{tr}_{B_{|A_i|}} \widetilde{K}_i(s_i)$ satisfies both Equation (4.2) and Equation (4.3) and furthermore such that

$$\operatorname{tr}_{B_{|A_i|}} \widetilde{K}_i(s_i) = \operatorname{tr}_{B^{\text{in}}_{(i)} B_{|A_i|}}[\widetilde{K}_i(s_i)] \otimes \frac{1}{\dim B^{\text{in}}_{(i)}} I_{B^{\text{in}}_{(i)}}.$$

This last equation ensures that the system contained in $B_{|A_i|}$ has no influence on previous operations, and correspond to a purification of parties' i measurement. Finally, if these operators are not positive semidefinite, we mix them with the identity to obtain positive eigenvalues and satisfy Equation (4.4):

$$\tilde{\mathsf{K}}_i(s_i) := (1-\lambda)\cdot \widetilde{K}_i(s_i) + \lambda \cdot \frac{1}{\dim B^{\text{out}}_{(i)}|A_i|} I_{B_i B^{\text{out}}_{(i)} B^{\text{in}}_{(i)} B_{|A_i|}},$$

where λ is the minimum eigenvalue of $\widetilde{K}_i(s_i)$.

The $\{\mathsf{K}_{i,a_i}(s_i)\}_{a_i\in A_i, s_i\in S_i}$ are then obtained by measuring system $B_{|A_i|}$ in the computational basis, which can be expressed in the Choi picture as:

$$\mathsf{K}_{i,a_i}(s_i) = \tilde{\mathsf{K}}_i(s_i) * |a_i\rangle\langle a_i|_{B_{|A_i|}}. \tag{4.6}$$

Once a random solution is initialized, we iteratively optimize the RHS of Equation (4.5), optimizing each of ρ and $\{\mathsf{K}_{i,a_i}(s_i)\}_{a_i\in A_i, s_i\in S_i}$ satisfying the relevant constraints, ceteris paribus. Each iteration is clearly an SDP. This leads to a monotonically non-decreasing lower bound on ω_q. We stop the algorithm when a round of optimization (of the state and each operator K_i) does not change the lower bound more than a given tolerance, typically, 10^{-6}.

4.2 LC games

Using this generalized see-saw, as well as more standard tools used for nonlocal games, we can now look at different LC and SISO games and compare their classical and quantum values for different latency constraints and connectivity graphs. We consider in this section the case of LC games, while SISO games will be considered in Section 4.3.

4.2.1 Random XOR games

We apply the generalized see-saw to three-party XOR games. As we discuss below, in this case it is actually sufficient to use a see-saw algorithm for conventional nonlocal games, but we perform this calculation to demonstrate the use of the generalized see-saw algorithm and to finally motivate the example in Section 4.3.2. We consider random three-party XOR games defined by a uniform distribution π on their input and with probabilistic predicates of the form

$$\mathcal{V}(a_1, a_2, a_3, s_1, s_2, s_3) = \beta_{s_1,s_2,s_3}(a_1 \oplus a_2 \oplus a_3),$$

where for all $(s_1, s_2, s_3) \in \{0,1\}^3$ we have

$$\beta_{s_1,s_2,s_3}(a_1 \oplus a_2 \oplus a_3) = \max[\hat{\beta}_{s_1,s_2,s_3} \cdot (-1)^{a_1 \oplus a_2 \oplus a_3}, 0] \tag{4.7}$$

for some coefficients $-1 \leq \hat{\beta}_{s_1,s_2,s_3} \leq 1$.

Consider for instance the game defined by the coefficients

$$\hat{\beta}_{0,0,0} = 0.438 \qquad \hat{\beta}_{1,0,0} = 0.580 \tag{4.8}$$
$$\hat{\beta}_{0,0,1} = 0.610 \qquad \hat{\beta}_{1,0,1} = -0.502 \tag{4.9}$$
$$\hat{\beta}_{0,1,0} = 0.520 \qquad \hat{\beta}_{1,1,0} = -0.724 \tag{4.10}$$
$$\hat{\beta}_{0,1,1} = -0.466 \qquad \hat{\beta}_{1,1,1} = -0.220, \tag{4.11}$$

which was randomly generated. We choose to use it because it exhibits some interesting properties as summarized in Table 2.

Connectivity graph	ω_c	Lower bound ω_q	Upper bound ω_q	Quantum violation?
(1) $v_1 \quad v_2 \quad v_3$	0.34300	0.37157	0.37157 (NPA)	Yes
$(2a)$ $v_1 \leftrightarrows v_2 \quad v_3$	0.34300	0.37440	0.37440 (NPA aggreg)	Yes
$(2b)$ $v_1 \quad v_2 \leftrightarrows v_3$	0.34300	0.37219	0.37219 (NPA aggreg)	Yes
$(2c)$ $v_1 \leftrightarrows v_3 \quad v_2$	0.34300	0.37347	0.37347 (NPA aggreg)	Yes
(3) $v_1 \leftrightarrows v_3 \leftrightarrows v_2$	0.50750	0.5075	0.5075 (Algebraic)	No

Table 2: Computed classical and quantum values of the game defined by the coefficients given in Equations (4.8) to (4.11) with respect to different connectivity graphs. For the empty graph (1), the upper bound on ω_q is obtained via the NPA hierarchy.[13] For the graphs $(2a)$, $(2b)$ and $(2c)$ the upper bounds on ω_q are obtained via the NPA hierarchy on the non-local game obtained by aggregating respectively (v_1, v_2), (v_2, v_3), and (v_1, v_3) into a single party, which can only make the strategy more powerful. Meanwhile the upper bound for the graph (3) is given by $\omega_a = \sum_{s_1,s_2,s_3} \hat{\beta}_{s_1,s_2,s_3}$. The classical values are obtained by enumerating all the possible classical strategies as given by Theorem 2. The lower bounds on ω_q are obtained by running the generalized see-saw with where the quantum systems $B_i, B_{i\to j}$ are qubits.

We first observe that in Table 2 all connectivity graphs except (3) exhibits a classical-quantum separation. This shows whether or not the an LC game has a quantum violation can depend on the connectivity graph, as in the case of the the distributed CHSH game in Section 2.3.1. The absence of a separation for (3) is something generic to XOR games, as for such games, classical strategies can always achieve the algebraic bound on the graph (3), by having parties v_1 and v_3 producing a

[13]We here use the second level of the hierarchy.

fixed output 0, and v_2 producing its output depending on the full input (s_1, s_2, s_3) which as been communicated to him by v_1 and v_2.

The particularity of XOR games also appears in the results for the graphs $(2a), (2b)$ and $(2c)$ where the lower and upper bounds on ω_q are matching, meaning that the see-saw converged for each graph to the optimal quantum strategy when the parties are aggregated. This is because v_1 and v_2 can communicate their inputs, so their joint behavior can be restricted without loss of generality to one of them outputting a fixed value 0, and the other one producing its output depending on their joint input (s_1, s_2). Hence, for XOR games, it is sufficient to the see-saw algorithm for conventional nonlocal games.

Finally we observe in Table 2 that the classical values are the same for all connectivity graphs except (3), where it reaches the algebraic value. This means that for any physical arrangement of the parties (which of the two parties are physically closer), relaxing the latency constraints such that two physically close classical parties exchange their inputs does not help for this particular game. This in contrast to what happens in the quantum case, where the quantum value can be different depending on the physical arrangement. In fact, we can order the different graphs based on their quantum values

$$\omega_q^{(1)} < \omega_q^{(2b)} < \omega_q^{(2c)} < \omega_q^{(2a)} < \omega_q^{(3)}.$$

Here, we stress that because the solutions of the different SDP's are obtained with a very small duality gap (typically on the order of 10^{-8}), these results give genuine separations between these different graphs.

4.2.2 Extended XOR game

We next examine whether the separation of Theorem 10 between forwarding strategies and quantum strategies obtained for the extended CHSH games can occur for other LC games. In particular, can we obtain such a separation by extending [30] other nonlocal games other than CHSH game?

For any $n, m \in \mathbb{N}$, an extended XOR game is defined by a Boolean function $f : S_2 \times S_3 \to \{0, 1\}$ with $S_2 = \{0, \dots, n-1\}$ and $S_3 = \{0, \dots, m-1\}$ and a uniform prior π on the inputs. The first party receives a fixed input $s_1 = 0$, while the two others receive an input from S_2 and S_3 respectively. They all produce an output in $A_1, A_2, A_3 = \{0, 1\}$. The probabilistic predicate is given by

$$\mathcal{V}(a_1, a_2, a_3, s_1, s_2, s_3) = \begin{cases} 1 & a_1 = a_2,\ f(s_2, s_3) = a_1 \oplus a_3 \\ 0 & \text{otherwise.} \end{cases}$$

We consider such games on the connectivity graph

$$v_1 \leftrightarrows v_2 \quad v_3, \tag{4.12}$$

To find a separation between forwarding and quantum strategies, we first compute upper-bounds on the quantum value obtained via forwarding strategies using Theorem 9. Indeed, for any game $(\mathcal{V}, \pi, G)$ as we can obtain upper bounds on ω_f directly via the NPA hierarchy [37, 38], which gives upper bounds on the quantum value of the nonlocal game $(\mathcal{V}', \pi')$. The lower bounds on ω_q are computed via the generalized see-saw of Section 4.1 where $B_{1\to 2}, B_{2\to 1}$ are qubits and v_2, v_3 share a two-qubit entangled quantum state.

By considering 50 extended XOR games for Boolean functions selected uniformly at random with $n = m = 3$ we find a separation for 35 games. This confirms that the separation we observed in the extended CHSH game applies naturally to a wide range of games. We also observe that for these 50 random extended XOR games, the upper-bound on forward strategies obtained via the NPA are all reached by classical solutions, just as in the case of the extended CHSH game in Section 2.4 (see Appendix C for more details on these simulations[14]). This may be because XOR games with a quantum-classical separation and a unique optimal quantum strategy (up to isometry) [45] may be sufficient for this to hold via a proof similar to that of Theorem 11.

4.3 SISO games

We now apply our numerical techniques to SISO games. In fact, we now have the numerical tools necessary to analyze a specific *physical layout* where three parties v_1, v_2, v_3 are arranged in an isosceles triangle as shown in Figure 12. Parties v_1 and v_2 are spatially separated by distance d,

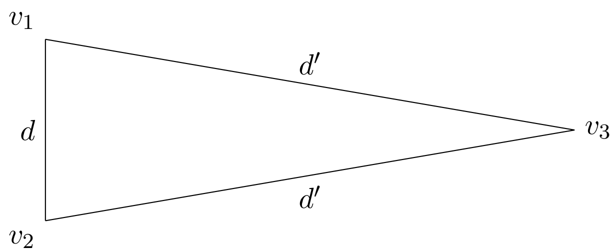


Figure 12: Three parties arranged in an isosceles triangle where $d' > 2d$.

whereas v_3 is at a larger distance $d' > 2d$ from both v_1 and v_2. We wish to derive fundamental physical limits on their correlations by setting the communication latency to be the distance divided by the speed of light c. We then map this physical layout to a SISO game via the procedure given in Section 3:[15]

- We assume all parties receive their inputs at time 0 and must produce their outputs at time t, the latency constraint.
- The latency functions take values d/c or d'/c.
- Assume $V_T = \{t, d/c, d'/c\} \subseteq \mathbb{Q}$. Define t_0 as in Equation (3.1).
- We will thereby consider a SISO, τ-step LC game where $\tau = t/t_0$ and the latency function ℓ takes values in $\{d/(ct_0), d'/(ct_0)\}$.

Now, we observe that depending on the value of the latency constraint t, we end up with different special cases of SISO games:

1. For $t \in [0, d/c)$,[16] no parties can communicate and this is a conventional nonlocal game.
2. For $t \in [d/c, 2d/c)$, v_1 and v_2 can only communicate once and so this is a simple LC game.

[14]All the code and numerical solutions are accessible on the online repository `https://github.com/pierrepocreau/Latency-constrained-games`.

[15]Note this would be more natural if we use a continuous-time definition of LC games as in Section 3.4.

[16]This is an interval of real numbers, not rational numbers, but we will ignore this for now.

3. For $t \in [2d/c, d'/c)$, this is an SISO game where v_1 and v_2 can back-and-forth communicate and thus can be aggregated into one party.

4. For $t \geq d'/c$, all parties have access to all inputs and thus can realize all possible correlations.

For simple LC games, we can use the generalized see-saw algorithm in Section 4.1 to compute a lower bound on the quantum value. When we can aggregate v_1 and v_2, this becomes a conventional two-party nonlocal game, for which the standard numerical methods of nonlocal games can be used. Hence, we can compute bounds on the classical and quantum values for all these different ranges of latency constraints. This will culminate in the graph shown in Figure 13 where we plot the classical and quantum values for a particular LC game with this physical layout as a function of time!

4.3.1 Three-party XOR games

Given the physical layout and the induced τ and ℓ, we need to next choose the probabilistic predicate $\mathcal{V}$ and input distribution π. We first consider the three-party XOR games introduced in Section 4.2.1.

Now, consider the case where $t < d/c$ such that no can parties communicate. This is then a conventional nonlocal game. We can write the winning probability of a three-party XOR game with uniformly distributed inputs as

$$p_{\text{win}}(\tau = 0) = \frac{1}{8} \sum_{s_1,s_2,s_3} \sum_{a_1,a_2,a_3} \beta_{s_1,s_2,s_3}(a_1 \oplus a_2 \oplus a_3) p(a_1, a_2, a_3 | s_1, s_2, s_3).$$

The quantum value ω_q can be upper bounded using the NPA hierarchy [37, 38], which we denote as ω_q^{upper}.[17] And the classical value ω_c can be obtained via linear programming [46, 47].

When the latency constraint is relaxed to $t \in [d/c, d'/c)$, v_1 and v_2 can communicate while v_3 cannot communicate with either. In this regime, v_1 and v_2 can be regarded as one aggregated party, even when $t < 2d/c$. As mentioned in Section 4.2.1, this is because they can adopt the following strategy:

1. v_1 gives a constant output $a_1 = 0$, and send his input s_1 to v_2.

2. v_2 measures binary observables[18] $K_{1,2}^{s_1,s_2}$ dependent on the aggregated inputs (s_1, s_2) to produce his output a_2.

The winning probability becomes

$$p_{\text{win}}(\tau = 4) = \frac{1}{8} \sum_{s_1,s_2,s_3} \sum_{a_2,a_3} \beta_{s_1,s_2,s_3}(a_2 \oplus a_3) p(a_2, a_3 | (s_1, s_2), s_3).$$

Thus, the quantum value of the aggregated game ω_q^{agg} can be computed exactly using an SDP [48], and the classical value ω_c^{agg} is again obtained via linear programming.

Building on the above observation, we now investigate the behavior of three-party XOR games before and after aggregation, corresponding to $t \in [0, d/c)$ and $t \in [d/c, d'/c)$, respectively. For

[17] We here use the third level of the NPA hierarchy.

[18] See Section 2.3.1.

simplicity we assume $\beta_{s_1,s_2,s_3}(a_1 \oplus a_2 \oplus a_3)$ are Boolean functions (takes values in $\{0,1\}$). We investigate the following examples:[19]

$$\begin{aligned}
\mathrm{XOR}_1 &: \hat{\beta}_{s_1,s_2,s_3} = (-1)^{s_1 s_2 s_3}, \\
\mathrm{XOR}_2 &: \hat{\beta}_{s_1,s_2,s_3} = (-1)^{s_1 s_2}, \\
\mathrm{XOR}_3 &: \hat{\beta}_{s_1,s_2,s_3} = (-1)^{s_1 s_3}, \\
\mathrm{XOR}_4 &: \hat{\beta}_{s_1,s_2,s_3} = (-1)^{s_1 (s_2 + s_3)}.
\end{aligned}$$

The quantum and classical values of these three-party XOR games before and after aggregating v_1 and v_2 are listed in Table 3. Note for the quantum value ω_q, we can give explicit quantum strategies that match the upper bounds ω_q^{upper} from the NPA algorithm within numerical precision. Furthermore, for these special three-party XOR games, the quantum values ω_q^{agg} can be verified analytically by transforming XOR_i into a generalized CHSH game [49]. See Section D for details. As

Game	ω_c	ω_q	ω_c^{agg}	ω_q^{agg}
XOR_1	$\frac{7}{8}$	0.8750	$\frac{7}{8}$	$\frac{4+\sqrt{10}}{8}$
XOR_2	$\frac{3}{4}$	0.8536	1	1
XOR_3	$\frac{3}{4}$	0.8536	$\frac{3}{4}$	$\cos^2 \frac{\pi}{8}$
XOR_4	$\frac{3}{4}$	0.7500	$\frac{3}{4}$	$\cos^2 \frac{\pi}{8}$

Table 3: Quantum and classical values of three-party XOR games XOR_i under different latency constraints. For ω_q, we compute an upper bound via the NPA hierarchy and find matching lower bounds up to numerical precision in Section D.

shown in Table 3, for XOR_1 and XOR_4, a quantum advantage does not exist when $t \in [0, d/c)$, but appears when communication is allowed when $t \in [d/c, d'/c)$ due to the relaxed latency constraint. Meanwhile for XOR_2, we see the opposite phenomenon: while where was a quantum advantage for $t \in [0, d/c)$, it disappears for $t \in [d/c, d'/c)$. Finally, for XOR_3, communication has no effect on the classical or quantum values. This shows that even for three-party XOR games where β is a Boolean function, the classical and quantum values can exhibit diverse behaviors as a function of time.

4.3.2 Perturbed XOR game

For the three parties XOR games studied in the previous section, one round of communication is equivalent to back-and-forth communication, for both classical and quantum strategies. While for classical strategies such an equivalence holds for all games (parties sharing all their input information in the first round of communication do not gain anything from a second round), we might expect a separation for the quantum case.

Here we aim to find such separation by considering an XOR game defined in Section 4.2.1, but with the addition of a random perturbation. This perturbation is introduced to break the structural properties of XOR games, which prevent all separations between one-round and back-and-forth communication. Perturbed XOR games are defined with a uniform distribution π on their inputs and with a probabilistic predicate

$$\mathcal{V}(a_1, a_2, a_3, s_1, s_2, s_3) = (1-\lambda) \cdot \beta_{s_1,s_2,s_3}(a_1 \oplus a_2 \oplus a_3) + \lambda \cdot \eta(a_1, a_2, a_3, s_1, s_2, s_3),$$

[19] Here we use the coefficients $\hat{\beta}_{s_1,s_2,s_3}$ in Equation (4.7) to define the Boolean functions.

for some coefficient $\lambda \in [0,1]$, and where for all $(s_1, s_2, s_3) \in \{0,1\}^3$ we have

$$\beta_{s_1,s_2,s_3}(a_1 \oplus a_2 \oplus a_3) = \max[\hat{\beta}_{s_1,s_2,s_3} \cdot (-1)^{a_1 \oplus a_2 \oplus a_3}, 0]$$

for some coefficients $-1 \leq \hat{\beta}_{s_1,s_2,s_3} \leq 1$, and $\eta(a_1, a_2, a_3, s_1, s_2, s_3) \in [0,1]$.

The specific example we consider is defined with $\lambda = 0.25$ and with the same coefficients $\hat{\beta}_{s_1,s_2,s_3}$ as that of Section 4.2.1,

$$\hat{\beta}_{0,0,0} = 0.438 \qquad \hat{\beta}_{1,0,0} = 0.580 \tag{4.13}$$

$$\hat{\beta}_{0,0,1} = 0.610 \qquad \hat{\beta}_{1,0,1} = -0.502 \tag{4.14}$$

$$\hat{\beta}_{0,1,0} = 0.520 \qquad \hat{\beta}_{1,1,0} = -0.724 \tag{4.15}$$

$$\hat{\beta}_{0,1,1} = -0.466 \qquad \hat{\beta}_{1,1,1} = -0.220, \tag{4.16}$$

while the perturbation η is given as a matrix in Table 4.

s\a	000	001	010	011	100	101	110	111
000	0.805	0.808	0.515	0.286	0.054	0.383	0.408	0.045
001	0.049	0.999	0.652	0.235	0.435	0.974	0.898	0.844
010	0.392	0.493	0.677	0.061	0.556	0.271	0.880	0.064
011	0.679	0.870	0.227	0.895	0.872	0.019	0.707	0.001
100	0.503	0.437	0.203	0.325	0.806	0.316	0.149	0.699
101	0.449	0.799	0.236	0.320	0.800	0.507	0.506	0.236
110	0.015	0.933	0.086	0.845	0.368	0.951	0.399	0.936
111	0.556	0.240	0.741	0.674	0.684	0.464	0.222	0.641

Table 4: Perturbation values $\eta(a_1, a_2, a_3, s_1, s_2, s_3)$ for each output (a_1, a_2, a_3) and input (s_1, s_2, s_3).

We run numerical simulations in order to compute the classical value of this game, as well as lower and upper bounds on ω_f and ω_q for different latency constraints (note that by definition $\omega_f \leq \omega_q$). We further assume that the parties are arranged in the isosceles triangle of Figure 12. These numerical results are summarized in Table 5.

The first observation we can make in Table 5 is that, except for the latency constraint $t \in [d/c, 2d/c)$, where v_1 and v_2 have time for one round of communication, all the lower and upper bounds on ω_q match, meaning that the see-saw algorithm converges to the optimal quantum solution. We also see that there exists a separation between classical and quantum values for all latency constraints except $t \geq d'/c$ for which both classical and quantum solutions can reach the maximum algebraic value since they have sufficient time to share all of the inputs.

By relaxing the latency constraints, we also see an increase in the quantum value of the game, as shown in Figure 13. Several things can be said here. As mentioned previously, both one-round and back-and-forth communication are equivalent for classical strategies and lead to the same classical value. This may not be the case for the quantum value in our perturbed XOR game. Here we observe a gap between the lower bound obtained on latency constraints $t \in [d/c, 2d/c)$ and the quantum value obtainable for latencies $t \in [2d/c, d'/c)$. However, we must stress that this gap is not proof of a separation, as we lack an upper bound on the quantum value achievable for latencies $t \in [d/c, 2d/c)$. In particular, the lower bound obtained numerically on ω_q for general quantum strategies (0.41770) is lower than the quantum value for forwarding strategies $\omega_f = 0.42419$. This implies that our

Latency constraint t	Value ω_c	Value ω_f	Lower bound ω_q	Upper bound ω_q	Quantum violation?
$t \in [0, d/c)$	0.38650	0.40052	0.40052	0.40052 (NPA)	Yes
$t \in [d/c, 2d/c)$	0.40384	0.42419	0.41770 (*)	0.42896 (NPA aggreg)	Yes
$t \in [2d/c, d'/c)$	0.40384	N/A	0.42896	0.42896 (NPA aggreg)	Yes
$t \geq d'/c$	0.59563	N/A	0.59563	0.59563 (Algebraic)	No

Table 5: Computed classical and quantum values of the perturbed XOR game with coefficients given in Equations (4.13) to (4.16) and perturbation matrix in Table 4, defined with $\lambda = 0.25$. For the latency $t \in [0, d/c)$, the upper bound on ω_q is obtained via the NPA hierarchy. For the latencies $t \in [d/c, 2d/c)$ and $t \in [2d/c, d'/c)$, the upper bounds on ω_q are obtained via the NPA hierarchy on the nonlocal game obtained by aggregating (v_1, v_2) into a single party as explained in Section 3.3.1. Meanwhile classical and quantum values for $t \geq d'/c$ is given by the maximum algebraic value $(1-\lambda) \cdot \sum_{s_1,s_2,s_3} \hat{\beta}_{s_1,s_2,s_3} + \lambda \cdot \sum_{s_1,s_2,s_3} \max_{a_1,a_2,a_3}[\eta(a_1, a_2, a_3, s_1, s_2, s_3)]$. The quantum strategies are optimized with quantum systems $B_i, B_{i\to j}$ being qubits. The values for forwarding strategies are obtained with matching lower and upper bounds obtained respectively via see-saw and the NPA hierarchy. The lower bound on ω_q for $t \in [d/c, 2d/c)$ is obtained with the generalized see-saw algorithm and correspond to the best results obtained on 1000 random initial solutions. (*) Note that, surprisingly, the generalized see-saw converges to quantum strategies that achieve lower winning probabilities than forwarding strategies. The lower bounds on ω_q for latency $t \in [0, d/c)$ and $t \in [2d/c, d'/c)$ are obtained with a see-saw algorithm on conventional nonlocal games with parties each holding a qubit. The nonlocal game for $t \in [2d/c, d'/c)$ is obtained by aggregating v_1 and v_2. Note that for $t \in [0, d/c)$, ω_f and ω_q coincide by definition.

see-saw algorithm is not finding the optimal quantum strategy. However, for $\lambda = 0$, the see-saw does converge to the optimal quantum solution, as shown in Table 2, which means it is still very capable. For $\lambda = 0.25$, even with 1000 random trials, the see-saw converges on local maxima that are far from the optimal value. That is, there might exist quantum solutions that achieve a quantum value of 0.42896, which would close the gap between one-round and back-and-forth communication for this LC game.

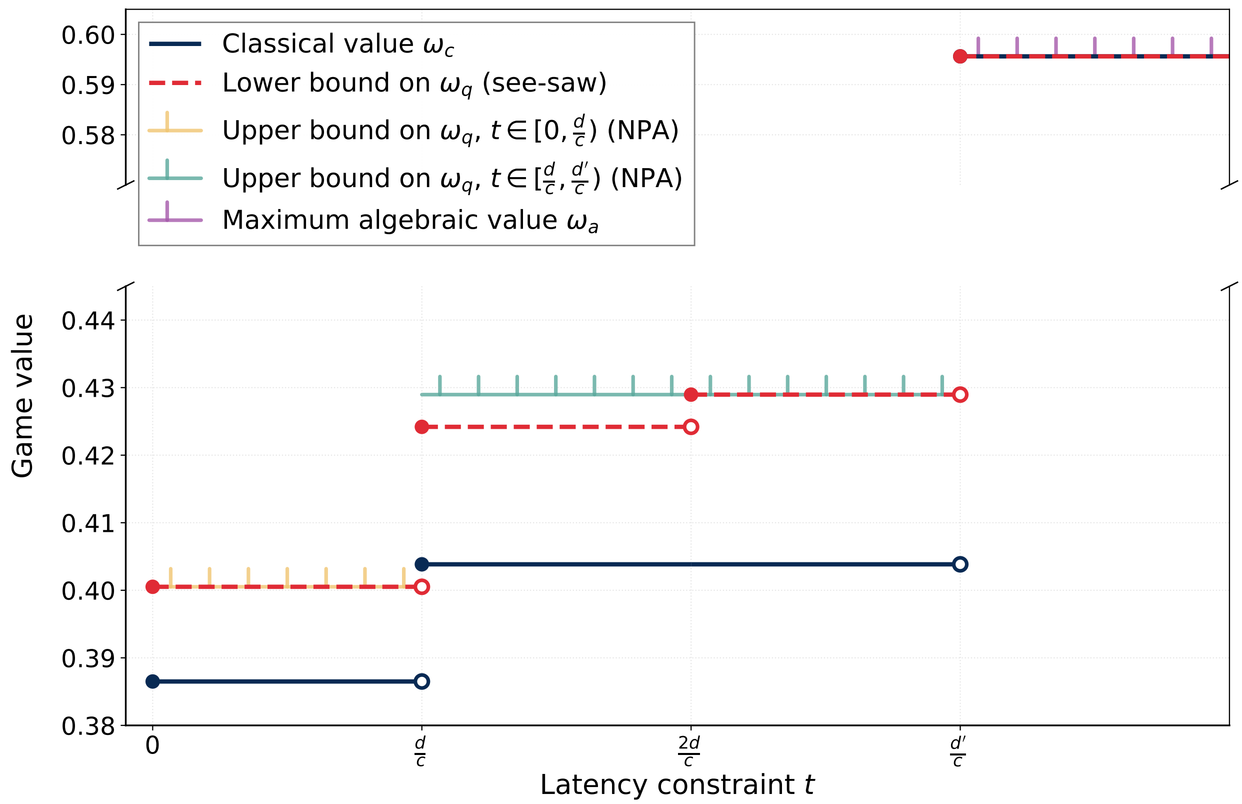


Figure 13: Graphic summarizing the results of Table 5 where we take $d' = 3d$ for convenience. Parties arranged in an isosceles triangle as in Figure 12 can achieve progressively higher classical and quantum values by relaxing the latency constraint.

5 Discussion

Although the violation of Bell inequalities is an extremely well-known phenomenon in physics, taught in undergraduate physics classes and mentioned in every modern textbook on quantum mechanics, we provide an interpretation of Bell inequalities that is not as well appreciated in the literature. The central concept that we introduce is a *latency constraint*. By setting the latency constraint lower than the speed of light delay between all pairs of parties, the set of inequalities that the parties' correlations must satisfy are exactly Bell inequalities. In the multipartite (more than 2) setting,[20] by setting the latency constraint a bit higher so that a subset of the parties can

[20]Note as mentioned in Section 3, we can stay in the two-party setting for multi-step LC games.

communicate, we obtain a new set of inequalities that generalizes Bell inequalities in a very natural way. And just like Bell inequalities, these new inequalities are fundamental bounds on local hidden variable theories. In other words, *the correlation of classical parties must satisfy these bounds within the specific latency constraint.* We saw examples of these bounds in Section 2.3 and Section 3.3, as well as quantum violations of these bounds.

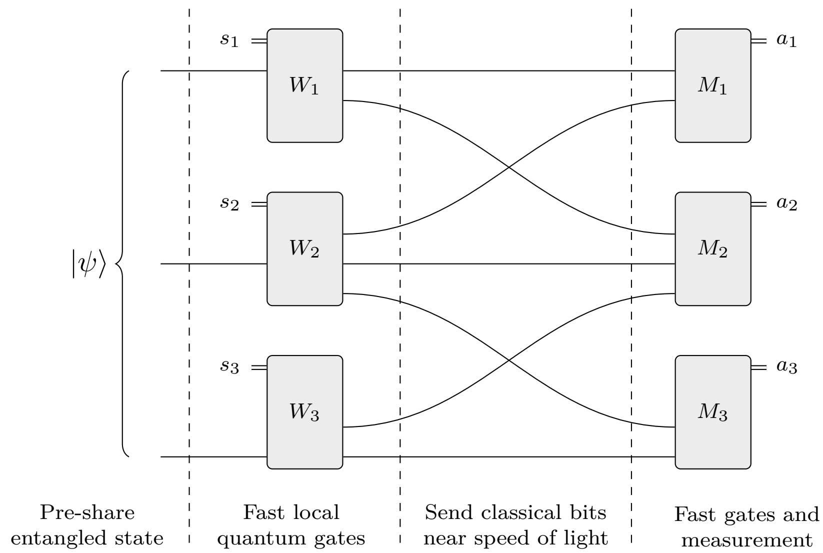


Figure 14: Experimental operations corresponding to different parts of the quantum strategy in Figure 3 (connectivity graph $v_1 \leftrightarrows v_2 \leftrightarrows v_3$). These operations include what is necessary to implement the quantum teleportation protocol.

Experimental realization The next natural step is to conduct experiments that can achieve quantum violations in this new, relaxed latency constraint regime. In general, this will require, in addition to quantum measurements on a shared entangled state for a Bell inequality violation, *real-time quantum communication.* The parties will need to perform quantum operations on their quantum register depending on their received input and send a part of their quantum registers to each other as specified by the quantum strategy. As we mentioned in Section 2 and Section 3, quantum teleportation is sufficient for the communication step. The parties can pre-share maximally entangled states and use them for quantum teleportation in real time. In particular, they only need to send each other classical bits as specified in the quantum teleportation protocol. This is summarized in Figure 14, where we label the experimental operations corresponding to different parts of the quantum strategy drawn in Figure 3.

Now, the main technical challenge[21] is the following: *the entire quantum strategy needs to be completed within the latency constraint.* Fast quantum measurements in the computational basis at or below microsecond timescales are already possible [50, 51, 52]. Quantum gates that the parties need to perform are usually even faster than the measurements for most physical platforms, assuming they can be efficiently decomposed into hardware-native gates. For the communication

[21] Other than the already recognized challenges in quantum network experiments such as achieving long-distance and long-lasting quantum entanglement.

step, to save time the transmitted classical bits ideally should be moving near the speed of light in vacuum. This can be approximately achieved via free-space optical communication, including radio transmission, microwave towers, or laser links. This is one reason quantum teleportation is preferred over physically sending quantum systems, as the latter, if we require high probability of success, usually requires a fiber connection between the parties. Light travels much slower in fiber than in free space. Finally, we ideally should use quantum memories that can serve as local quantum registers and can store entanglement, both for measurement and for quantum teleportation purposes. Current physical platforms such as trapped ions, quantum dots, nitrogen-vacancy (NV) centers, and possibly neutral atoms used in emerging quantum networks can serve as such quantum memories [53].

Real-world applications From a fundamental physics perspective, such experiments constitute a refutation of local hidden variable theory in a new regime. From an applied quantum technologies perspective, such experiments physically realize the mathematically provable time advantage achievable by quantum resources as mentioned in Section 1. Again, this is extremely important in latency-sensitive settings such as high frequency trading. It would be interesting to find specific examples of this, or at least toy models such as those of [8, 5].

One immediate possibility is to consider the same HFT setting in [5] with instrument X being traded in NYSE (New York Stock Exchange), but this time the same instrument X is also being traded in NASDAQ. Meanwhile, instrument Y is traded in CME far away in Chicago. Instruments X and Y are sometimes positively correlated, sometimes negatively correlated. The trading servers decide to issue a buy or sell order first in order to minimize the overall risk of their trading strategy.[22] If the instruments are positively correlated, the servers trading X first issue the order opposite of what the server trading Y first issues. If negative, the servers first issue the same order. Now, NYSE and NASDAQ's data centers are both in New Jersey, only 56.3 km apart. Meanwhile, the server in the CME data center in Aurora, IL is about 1100 km away. Thus, for latency constraints between 0.188 and 3.9 ms, the trading servers at NYSE and NASDAQ can communicate with each other with one round of communication but not with CME. The connectivity graph for the corresponding LC game is thus

$$\text{CME} \quad \text{NYSE} \leftrightarrows \text{NASDAQ}.$$

Since the HFT scenario in [5] maps to the CHSH game, assuming we want to exactly copy the NYSE decision at NASDAQ (otherwise there is no net trade of instrument X), *this is exactly the extended CHSH game with communication between two parties considered in Section 2.4*. That is, the task from the HFT perspective is two-fold:

1. The 3 trading servers at the 3 exchanges make the correct decisions for hedging purposes.

2. The NYSE and NASDAQ servers make the same decision.

By Theorem 11, quantum resources cannot achieve a higher probability of success in this task over classical resources ($\omega_f = \frac{3}{4}$ is an upper bound) for latency constraints below 0.188 ms, but *can for latency constraints between 0.188 and 3.9 ms*. Apart from HFT, it would also be interesting to

[22]Note that in HFT the ultra-fast trade decisions are actually order cancellations. We can consider the same scenario but instead of issuing orders we are canceling orders. A similar logic applies.

consider other real-world scenarios such as ad hoc network routing [9], rendezvous on graphs [10], and distributed systems [7] where a subset of the parties can communicate.

With the framework of τ-step LC games, we can also consider real-world scenarios where the parties are receiving multiple inputs and producing multiple outputs over the course of time. This should be a common characteristic in many settings. For example, in HFT, this framework allows for the exciting possibility to analyze a "full trading scenario" where trading servers are receiving information from their local exchanges, communicating with each other, and making trades *in real time*. We can define the probabilistic predicate $\mathcal{V}$, input distribution π, and latency function ℓ for a specific scenario, and then compute the classical and quantum values for the τ-step LC game defined thereby. A separation between the two would entail a mathematically provable quantum advantage for this trading scenario. Other latency-sensitive scenarios can be analyzed in a like manner.

Generalizations We can further generalize LC games to more complicated settings. The most interesting generalization would be to involve *parties in motion*. In this case, the distance between parties and therefore the latency function ℓ_L, defined as the speed of light delay, itself will be *time-dependent*. Taking relativistic effects into account, for parties moving near the speed of light, ℓ_L will also be dependent on the parties' velocities. However, in general, we would need to specify a frame of reference by which we define time. Given this complication, it may be more desirable to define a *coordinate-free notion of communication latency*. That is, we define moving parties as world lines in spacetime. Then, the transmitting parties emit photons, which are light-like curves. The receiving parties then receive the transmission when the light-like curves intersect their world lines, as shown in Figure 15 for a particular frame of reference. Going further, we can include

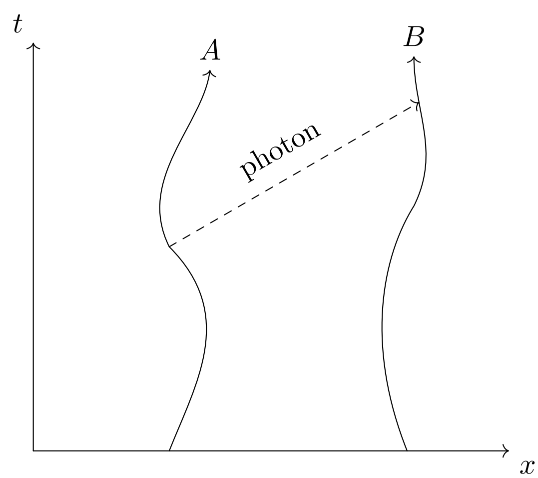


Figure 15: Spacetime diagram of two moving parties where party A emits a photon to party B.

gravitational effects by considering an LC game with parties in curved spacetime. In particular, it would be extremely interesting to consider a specific physical scenario that requires a general relativistic treatment and analyze the fundamental limits on correlations the parties can exhibit using classical versus quantum resources. One exciting example is to consider two parties, where one party is orbiting a black hole outside the event horizon while the other is falling into the black hole. The analysis of this scenario from an LC game perspective could possibly have bearing on the black hole information paradox [54].

Other possible generalizations include

- Nonlinear objective functions. Bell inequalities are linear combinations of terms in the behavior. We can consider nonlinear functions of the behavior [55, 39] and analyze corresponding fundamental limits for an LC game.

- Non-cooperative games. There is a body of literature on the usage of quantum entanglement to obtain more and better (in terms of social welfare) Nash equilibria in non-cooperative games [56, 57, 58, 59, 60]. In the LC games setting, we can analyze what Nash equilibria can be reached if some parties can communicate.

- Photon loss. In many Bell experiments, entangled photon pairs are used as the shared quantum system between the different parties. These photons are readily absorbed by physical media, and the consequent effects on the possible strategies and the behaviors realized can be described mathematically [61, 5, 62]. Compared to conventional nonlocal games, in the LC games setting, a new phenomenon appears: parties that can communicate can tell each other whether or not photon loss has occurred.

Open problems There is a plethora of open problems regarding LC games and their generalizations. One important open question is whether back-and-forth communication between two parties can achieve higher winning probabilities compared to a single round of communication. Figure 13 seems to give evidence of this, but is not sufficient. In general, we would need to prove a nontrivial upper bound on ω_q for single-round quantum strategies as defined in Theorem 5. One possible approach is to define an NPA-like hierarchy of upper bounds that converge to the quantum value as conjectured in Section B. Such a result will also better characterize the set of quantum behaviors $\mathcal{Q}$ for simple LC games. Another important direction is to mathematically analyze the threshold times $\tau_c(\alpha), \tau_q(\alpha)$ defined in Section 3.3. This would shed more light on the time advantage that quantum resources can achieve in SISO games. Lastly, it will be important to find more efficient numerical optimization algorithms for LC games. This is especially pressing since the numerical optimizers for conventional nonlocal games can already face scaling problems (consider state-of-the-art optimizers such as that of [63]). An efficient optimizer is crucial for designing experiments and for evaluating real-world applications.

Author contributions. D.D. conceived the idea of extending Bell inequalities by relaxing latency constraints, was in charge of the overall direction and planning, and was primarily responsible for the writing of the manuscript. Z.J. contributed to the definition of simple and multi-step LC games, led the effort in elucidating the structure and separations among classes of LC game strategies, and improved the presentation of the manuscript. P.P. devised the generalization of the see-saw method to LC games, was responsible for its implementation, and analyzed the simulations for random, extended and perturbed XOR games. M.X. contributed to the proofs of the basic properties of LC games. X.X. contributed to analyzing distributed games and analyzed specific three-party XOR games. All authors discussed the results and commented on the manuscript.

Acknowledgments. We would like to thank Nafiz Ishtiaque, Ziwen Liu, Guorui Ma, Veronica Pasquarella, Richard Schoen, and Jie Wang for helpful discussions. We thank Yuqing Li for writing one of the SDP algorithms that we used in our paper. DD would like to thank God for all of His provisions.

References

[1] John S Bell. On the Einstein Podolsky Rosen paradox. *Physics Physique Fizika*, 1(3):195, 1964.

[2] Gregor Weihs, Thomas Jennewein, Christoph Simon, Harald Weinfurter, and Anton Zeilinger. Violation of Bell's inequality under strict Einstein locality conditions. *Physical Review Letters*, 81(23):5039, 1998.

[3] Adam Brandenburger and Pierfrancesco La Mura. Team decision problems with classical and quantum signals. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2058):20150096, 2016.

[4] Mario Szegedy, Dawei Ding, and Yaoyun Shi. Systems and methods for coordinating decisions between non-communicating parties, December 17 2020. US Patent 11676104B2.

[5] Dawei Ding and Liang Jiang. Coordinating decisions via quantum telepathy. *arXiv preprint arXiv:2407.21723*, 2024.

[6] Andrew Haldane. Patience and finance. `https://www.bis.org/review/r100909e.pdf`, 2010. Online; accessed 16 February 2024.

[7] Francisco Ferreira da Silva and Stephanie Wehner. Entanglement improves coordination in distributed systems. In *Proceedings of the 2nd Workshop on Quantum Networks and Distributed Quantum Computing*, pages 14–20, 2025.

[8] Adam Brandenburger and Pierfrancesco La Mura. Quantum-assisted loan balancing in communication-constrained wide-area physical networks, November 24 2016. US Patent 10056983B2.

[9] Masoud Hasanpour, Shahin Shariat, Payam Barnaghi, Seyed Amir Hoseinitabatabaei, Seiamak Vahid, and Rahim Tafazolli. Quantum load balancing in ad hoc networks. *Quantum Information Processing*, 16(6):148, 2017.

[10] Giuseppe Viola and Piotr Mironowicz. Quantum strategies for rendezvous and domination tasks on graphs with mobile agents. *Physical Review A*, 109(4):042201, 2024.

[11] Robin Lorenz and Jonathan Barrett. Causal and compositional structure of unitary transformations. *Quantum*, 5:511, 2021.

[12] David Beckman, Daniel Gottesman, Michael A Nielsen, and John Preskill. Causal and localizable quantum operations. *Physical Review A*, 64(5):052309, 2001.

[13] D. Bacon and B. F. Toner. Bell inequalities with auxiliary communication. *Phys. Rev. Lett.*, 90:157904, Apr 2003.

[14] Katherine Maxwell and Eric Chitambar. Bell inequalities with communication assistance. *Phys. Rev. A*, 89:042108, Apr 2014.

[15] M. G. M. Moreno, Samuraí Brito, Ranieri V. Nery, and Rafael Chaves. Device-independent secret sharing and a stronger form of Bell nonlocality. *Phys. Rev. A*, 101:052339, May 2020.

[16] R. Chaves, R. Kueng, J. B. Brask, and D. Gross. Unifying framework for relaxations of the causal assumptions inBell 's theorem. *Phys. Rev. Lett.*, 114:140403, Apr 2015.

[17] Jonatan Bohr Brask and Rafael Chaves. Bell scenarios with communication. *Journal of Physics A: Mathematical and Theoretical*, 50(9):094001, 2017.

[18] Carlos Vieira, Carlos de Gois, Pedro Lauand, Lucas E. A. Porto, Sébastien Designolle, and Marco Túlio Quintino. Can outcome communication explain Bell nonlocality?, 2025.

[19] George Moreno, Ranieri Nery, Alberto Palhares, and Rafael Chaves. Multistage games and Bell scenarios with communication. *Phys. Rev. A*, 102:042412, Oct 2020.

[20] Armin Tavakoli, Alejandro Pozas-Kerstjens, Ming-Xing Luo, and Marc-Olivier Renou. Bell nonlocality in networks. *Reports on Progress in Physics*, 85(5):056001, 2022.

[21] Denis Rosset, Cyril Branciard, Tomer Jack Barnea, Gilles Pütz, Nicolas Brunner, and Nicolas Gisin. Nonlinear Bell inequalities tailored for quantum networks. *Phys. Rev. Lett.*, 116:010403, Jan 2016.

[22] Alejandro Pozas-Kerstjens, Nicolas Gisin, and Armin Tavakoli. Full network nonlocality. *Phys. Rev. Lett.*, 128:010403, Jan 2022.

[23] Daniel Collins, Nicolas Gisin, Sandu Popescu, David Roberts, and Valerio Scarani. Bell-type inequalities to detect true n-body nonseparability. *Phys. Rev. Lett.*, 88:170405, Apr 2002.

[24] George Svetlichny. Distinguishing three-body from two-body nonseparability by a Bell-type inequality. *Phys. Rev. D*, 35:3066–3069, May 1987.

[25] Gus Gutoski and John Watrous. Toward a general theory of quantum games. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 565–574, 2007.

[26] Itamar Pitowsky. The range of quantum probability. *Journal of Mathematical Physics*, 27(6):1556–1565, 1986.

[27] John Matthew Donohue and Elie Wolfe. Identifying nonconvexity in the sets of limited-dimension quantum correlations. *Phys. Rev. A*, 92:062120, Dec 2015.

[28] Nicolas Brunner, Daniel Cavalcanti, Stefano Pironio, Valerio Scarani, and Stephanie Wehner. Bell nonlocality. *Reviews of modern physics*, 86(2):419–478, 2014.

[29] Alex Arkhipov. Extending and characterizing quantum magic games. *arXiv:1209.3819*, 2012.

[30] Ben Toner. Monogamy of non-local quantum correlations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 465(2101):59–69, 2009.

[31] Stefano Pironio. Lifting Bell inequalities. *Journal of Mathematical Physics*, 46(6):062112, 06 2005.

[32] Netanel Aharon, Shai Machnes, Benni Reznik, Jonathan Silman, and Lev Vaidman. Continuous input nonlocal games. *Natural computing*, 12(1):5–8, 2013.

[33] Julia Kempe, Hirotada Kobayashi, Keiji Matsumoto, Ben Toner, and Thomas Vidick. Entangled games are hard to approximate. *SIAM Journal on Computing*, 40(3):848–877, 2011.

[34] Zhengfeng Ji. Classical verification of quantum proofs. In *Proceedings of the Forty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '16, page 885–898, New York, NY, USA, 2016. Association for Computing Machinery.

[35] Zhengfeng Ji, Anand Natarajan, Thomas Vidick, John Wright, and Henry Yuen. MIP*= RE. arXiv:2001.04383, 2020.

[36] Reinhard F Werner and Michael M Wolf. Bell inequalities and entanglement. *arXiv preprint quant-ph/0107093*, 2001.

[37] Miguel Navascués, Stefano Pironio, and Antonio Acín. Bounding the set of quantum correlations. *Physical Review Letters*, 98(1):010401, 2007.

[38] Miguel Navascués, Stefano Pironio, and Antonio Acín. A convergent hierarchy of semidefinite programs characterizing the set of quantum correlations. *New J. Phys.*, 10(7):073013, July 2008.

[39] Igor Klep, Victor Magron, Jurij Volčič, and Jie Wang. State polynomials: positivity, optimization and nonlinear Bell inequalities. *Mathematical Programming*, 207(1):645–691, 2024.

[40] Giulio Chiribella, Giacomo Mauro D'Ariano, and Paolo Perinotti. Quantum circuit architecture. *Phys. Rev. Lett.*, 101(6):060401, 2008.

[41] Giulio Chiribella, Giacomo Mauro D'Ariano, and Paolo Perinotti. Theoretical framework for quantum networks. *Phys. Rev. A*, 80(2):022339, 2009.

[42] Philip Taranto, Simon Milz, Mio Murao, Marco Túlio Quintino, and Kavan Modi. Higher-order quantum operations, 2025.

[43] Man-Duen Choi. Completely positive linear maps on complex matrices. *Lin. Algebra Appl.*, 10(3):285–290, 1975.

[44] A. Jamiołkowski. Linear transformations which preserve trace and positive semidefiniteness of operators. *Rep. Math. Phys.*, 3(4):275–278, 1972.

[45] Carl A. Miller and Yaoyun Shi. Optimal robust self-testing by binary nonlocal XOR games. In Simone Severini and Fernando Brandao, editors, *8th Conference on the Theory of Quantum Computation, Communication and Cryptography (TQC 2013)*, volume 22 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 254–262, Dagstuhl, Germany, 2013. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[46] Marek Zukowski, Dagomir Kaszlikowski, Adam Baturo, and Jan-Åke Larsson. Strengthening the bell theorem: conditions to falsify local realism in an experiment. *arXiv preprint quant-ph/9910058*, 1999.

[47] Dagomir Kaszlikowski, Piotr Gnaciński, Marek Żukowski, Wieslaw Miklaszewski, and Anton Zeilinger. Violations of local realism by two entangled n-dimensional systems are stronger than for two qubits. *Physical Review Letters*, 85(21):4418, 2000.

[48] Richard Cleve, Peter Hoyer, Benjamin Toner, and John Watrous. Consequences and limits of nonlocal strategies. In *Proceedings. 19th IEEE Annual Conference on Computational Complexity, 2004.*, pages 236–249. IEEE, 2004.

[49] Antonio Acín, Serge Massar, and Stefano Pironio. Randomness versus nonlocality and entanglement. *Phys. Rev. Lett.*, 108:100402, Mar 2012.

[50] 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. *Nature*, 526(7575):682–686, 2015.

[51] Lynden K Shalm, Evan Meyer-Scott, Bradley G Christensen, Peter Bierhorst, Michael A Wayne, Martin J Stevens, Thomas Gerrits, Scott Glancy, Deny R Hamel, Michael S Allman, et al. Strong loophole-free test of local realism. *Physical review letters*, 115(25):250402, 2015.

[52] 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. *Physical review letters*, 115(25):250401, 2015.

[53] Khabat Heshami, Duncan G England, Peter C Humphreys, Philip J Bustard, Victor M Acosta, Joshua Nunn, and Benjamin J Sussman. Quantum memories: emerging applications and recent advances. *Journal of modern optics*, 63(20):2005–2028, 2016.

[54] Joseph Polchinski. The black hole information problem. In *New Frontiers in Fields and Strings: TASI 2015 Proceedings of the 2015 Theoretical Advanced Study Institute in Elementary Particle Physics*, pages 353–397. World Scientific, 2017.

[55] Denis Rosset, Cyril Branciard, Tomer Jack Barnea, Gilles Pütz, Nicolas Brunner, and Nicolas Gisin. Nonlinear Bell inequalities tailored for quantum networks. *Physical review letters*, 116(1):010403, 2016.

[56] Vincenzo Auletta, Diodato Ferraioli, Ashutosh Rai, Giannicola Scarpa, and Andreas Winter. Belief-invariant and quantum equilibria in games of incomplete information. *Theoretical Computer Science*, 895:151–177, 2021.

[57] Anna Pappa, Niraj Kumar, Thomas Lawson, Miklos Santha, Shengyu Zhang, Eleni Diamanti, and Iordanis Kerenidis. Nonlocality and conflicting interest games. *Physical review letters*, 114(2):020401, 2015.

[58] Katarzyna Bolonek-Lasoń. Three-player conflicting interest games and nonlocality. *Quantum Information Processing*, 16(8):186, 2017.

[59] Faisal Shah Khan, Neal Solmeyer, Radhakrishnan Balu, and Travis Humble. Quantum games: a review of the history, current state, and interpretation. *arXiv preprint arXiv:1803.07919*, 2018.

[60] Alastair A Abbott, Mehdi Mhalla, and Pierre Pocreau. Improving social welfare in non-cooperative games with different types of quantum resources. *Quantum*, 8:1376, 2024.

[61] Nicolas Gigena, Ekta Panwar, Giovanni Scala, Mateus Araújo, Máté Farkas, and Anubhav Chaturvedi. Robust self-testing of Bell inequalities tilted for maximal loophole-free nonlocality, 2024.

[62] Dawei Ding, Yuqing Li, Mingze Xu, and Xinyu Xu. Algorithms for optimizing Bell inequalities with photon loss. In preparation.

[63] Luke Mortimer. Bounding large-scale Bell inequalities. *Physical Review A*, 111(5):052442, 2025.

[64] Oded Regev and Thomas Vidick. Quantum XOR games. *ACM Trans. Comput. Theory*, 7(4), August 2015.

[65] Jens Eisert, Martin Wilkens, and Maciej Lewenstein. Quantum games and quantum strategies. *Phys. Rev. Lett.*, 83:3077–3080, Oct 1999.

[66] Anna Pappa, Niraj Kumar, Thomas Lawson, Miklos Santha, Shengyu Zhang, Eleni Diamanti, and Iordanis Kerenidis. Nonlocality and conflicting interest games. *Phys. Rev. Lett.*, 114:020401, Jan 2015.

[67] John F. Clauser, Michael A. Horne, Abner Shimony, and Richard A. Holt. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.*, 23:880–884, Oct 1969.

[68] MOSEK ApS. The MOSEK optimization toolbox for MATLAB manual. Version 10.1. `https://www.mosek.com/`.

A Nonlocal Games

In this section, we give a brief introduction to nonlocal games. A nonlocal game typically involves multiple parties that cannot communicate with each other. This is shown schematically in Figure 16. The formal definition is as follows.

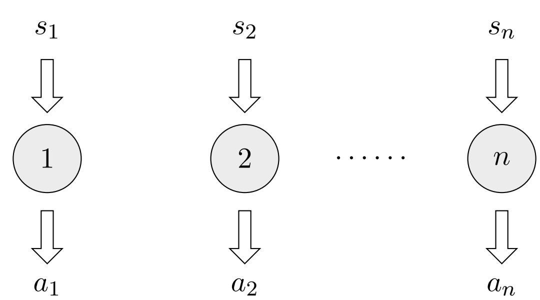


Figure 16: A nonlocal game with n parties. Each party i receives input s_i and outputs a_i.

Definition 24. *Let $n \geq 2$ be an integer. Let S_i, A_i be finite sets for $i \in [n]$. We define a probabilistic predicate $\mathcal{V}$ which is a map*

$$\mathcal{V} : \prod_{i=1}^{n} A_i \times \prod_{i=1}^{n} S_i \to [0, 1].$$

Furthermore, we define the input distribution π which is a probability distribution on $\prod_{i=1}^n S_i$. A ***nonlocal game*** *is described by the tuple $(\mathcal{V}, \pi)$.*

Informally speaking, in a nonlocal game, each party i receives input s_i and produces an output a_i. All parties have full knowledge of the input distribution π and the winning predicate $\mathcal{V}$. The goal is to maximize the winning probability. Parties may agree on a joint strategy in advance based on their knowledge, but once the game starts, communication is forbidden. When playing the game, implementation of their strategy realizes a **behavior**, defined as a conditional probability distribution

$$p(a_1, a_2, \cdots, a_n | s_1, s_2, \cdots, s_n).$$

A behavior is interpreted as the distribution over the different outputs the parties produce, conditioned on their inputs. By plugging the behavior into the predicate, we can compute the **winning probability**:

$$p_{\text{win}} \coloneqq \sum_{a_i \in A_i, s_i \in S_i} \pi(s_1, s_2, \cdots s_n) \cdot p(a_1, a_2, \cdots, a_n | s_1, s_2, \cdots, s_n) \mathcal{V}(a_1, a_2 \cdots a_n | s_1, s_2, \cdots s_n).$$

We proceed to define what strategies the parties can use with classical resources. We first introduce **deterministic strategies**.

Definition 25. *Let $(\mathcal{V}, \pi)$ be a nonlocal game. A* ***deterministic strategy*** *is given by functions $\{f_i\}_{i=1}^n$, where*

$$f_i : S_i \to A_i.$$

The behavior realized by a deterministic strategy is simply

$$p(\mathbf{a}|\mathbf{s}) = \prod_{i=1}^n \delta_{a_i, f_i(s_i)},$$

where $\mathbf{a} \in \prod_{i=1}^n A_i, \mathbf{s} \in \prod_{i=1}^n S_i$.

More generally, for a **classical strategy**, parties can use shared randomness. In this case, a classical strategy can be expressed as a probabilistic mixture of deterministic strategies, and the behavior realized is therefore a convex combination of behaviors realized by deterministic behaviors. A bound on the largest winning probability attainable by a classical strategy is known as a **Bell inequality**.

For a quantum strategy, parties can have access to a shared entangled state and perform local quantum measurements according to the inputs they receive. We therefore have the following definition.

Definition 26. *Let $(\mathcal{V}, \pi)$ be a nonlocal game. Let B_i be a Hilbert space for $i \in [n]$. A* ***quantum strategy*** *is a tuple $(|\psi\rangle, \{M_i(s_i)\}_{i \in [n], s_i \in S_i})$, where*

$$|\psi\rangle \in \bigotimes_{i=1}^n B_i$$

is a quantum state and

$$M_i(s_i) = \{\Pi_{i,a_i}(s_i)\}_{a_i \in A_i}$$

is a projective measurement on B_i for each s_i. The behavior realized by a quantum strategy is given by

$$p(\mathbf{a}|\mathbf{s}) = \langle \psi | \bigotimes_{i=1}^n \Pi_{i,a_i}(s_i) | \psi \rangle.$$

Classical behaviors form a subset of quantum behaviors. However, with quantum resources, parties have the potential to achieve a higher winning probability. This phenomenon is called **Bell inequality violation**. Several nonlocal games are known for which this occurs [64, 65, 66], specific examples being the CHSH game [1, 67] and the magic square game [29].

B Towards upper bounds via a generalized NPA algorithm

As for upper bounds on ω_q for LC games, consider again the expression for the behavior of a quantum strategy for an LC game:

$$p(\mathbf{a}|\mathbf{s}) = \Big\langle \psi \Big| \bigotimes_{i=1}^{n} W_i(s_i)^\dagger \bigotimes_{i=1}^{n} \Pi_{i,a_i} \bigotimes_{i=1}^{n} W_i(s_i) \Big| \psi \Big\rangle.$$

As in Section 4.1, we can take the Choi state representation:

$$|W_i(s_i)\rangle := \frac{1}{\sqrt{\dim B_i}} \sum_{j=1}^{\dim B_i} |j\rangle \otimes W_i(s_i)|j\rangle$$

which is a state on $B_i \otimes B_i^{\text{out}}$. We can then write the behavior as

$$p(\mathbf{a}|\mathbf{s}) = \dim B \operatorname{tr}\Big[\Big(\bigotimes_{i=1}^{n} |W_i(s_i)\rangle\langle W_i(s_i)|\Big) \cdot \Big(|\psi\rangle\langle\psi|^T \otimes \bigotimes_{i=1}^{n} \Pi_{i,a_i}\Big)\Big]$$

With this form, we can use the formalism in [39] to express this in terms of a noncommutative polynomial optimization with multiple state symbols and a tracial state symbol. First, write the objective function

$$\begin{aligned}
&\sum_{\mathbf{s}\in\prod_{i=1}^n S_i} \pi(\mathbf{s}) \sum_{\mathbf{a}\in\prod_{i=1}^n A_i} \mathcal{V}(\mathbf{a}|\mathbf{s}) p(\mathbf{a}|\mathbf{s}) \\
&= \sum_{\mathbf{s}\in\prod_{i=1}^n S_i} \pi(\mathbf{s}) \sum_{\mathbf{a}\in\prod_{i=1}^n A_i} \mathcal{V}(\mathbf{a}|\mathbf{s}) \operatorname{tr}\Big[\Big(\bigotimes_{i=1}^{n} |W_i(s_i)\rangle\langle W_i(s_i)|\Big) \cdot \Big(\dim B|\psi\rangle\langle\psi|^T \otimes \bigotimes_{i=1}^{n} \Pi_{i,a_i}\Big)\Big]. \qquad \text{(B.1)}
\end{aligned}$$

Let z be a noncommutative variable. For each $a_i \in A_i, i \in [n]$, let x_{i,a_i} also be noncommutative variables and for each $\mathbf{s} \in \prod_{i=1}^n S_i$, let $\varsigma_{\mathbf{s}}$ be state symbols. Lastly, let Tr be the tracial state symbol (the normalized trace for finite dimensional Hilbert spaces). We can re-write the optimization of Equation (B.1) as

$$\sum_{\mathbf{s}\in\prod_{i=1}^n S_i} \pi(\mathbf{s}) \sum_{\mathbf{a}\in\prod_{i=1}^n A_i} \mathcal{V}(\mathbf{a}|\mathbf{s}) \varsigma_{\mathbf{s}}(z \prod_{i=1}^{n} x_{i,a_i}),$$

where the product is taken from right to left, such that

$$
\begin{gathered}
x_{i,a_i}^2 = x_{i,a_i}, \sum_{a_i \in A_i} x_{i,a_i} = 1, \\
x_{i,a_i} x_{i,a_i'} = 0,\ a_i \neq a_i', \\
[x_{i,a_i}, x_{i',a_{i'}}] = 0,\ i \neq i', \\
z \succeq 0, \\
\mathrm{Tr}[z] = 1, \\
[x_{i,a_i}, z] = 0, \\
\varsigma_{\mathbf{s}}(w_1 \cdots w_\ell) = \varsigma_{\mathbf{s}}(w_1) \cdots \varsigma_{\mathbf{s}}(w_\ell) \text{ for } w_i \in \langle x_{i,a_i} : a_i \in A_i \rangle \\
\text{where } i \text{ are distinct parties that do not share an edge,} \\
\varsigma_{\mathbf{s}}(w) = \mathrm{Tr}[w] \text{ for } w \in \langle z \rangle.
\end{gathered}
$$

Note here that we don't impose the purity of the quantum states corresponding to $z, \varsigma_{\mathbf{s}}$, but this is not a problem because mixed states will simply lead to a convex combination of the objective function values corresponding to pure states.

We leave to future to prove that there is an SDP hierarchy that converges to the optimal value. We would need to prove a series of results similar to that of [39] but where we have both tracial and state constraints. We also need multiple different state symbols $\varsigma_{\mathbf{s}}$.

C Details on Extended XOR games

In this Appendix are detailed the upper and lower bounds obtained on ω_q and ω_f respectively for fifty random extended XOR games as defined in Section 4.2.2 with the connectivity graph G

$$v_1 \leftrightarrows v_2 \quad v_3.$$

These results are summarized in Table 6, where the ID corresponds to the decimal number associated to the truth table of a Boolean function $f : S_2 \times S_3 \to \{0,1\}$ defining an extended XOR game. More explicitly, writing $S_2 = \{0, \ldots, n-1\}$ and $S_3 = \{0, \ldots, m-1\}$, a Boolean function is defined by a length $n \cdot m$ truth table, such that for any $(i,j) \in S_2 \times S_3$, $f(i,j)$ is evaluated to the bit at position $m \cdot i + j$. We consider in particular the case $n = m = 3$, in that case, the truth table $[0,0,0,0,0,0,1,0,0]$ corresponds for instance to the ID 4.

The upper bounds on ω_f are obtained using the NPA hierarchy on the transformed nonlocal game as given in Theorem 9. The lower bounds on w_q are obtained for each game with the see-saw algorithm described in Section 4.1, with a two dimensional quantum channel between v_1 and v_2, and a four dimensional quantum system shared between v_2 and v_3, while v_1 initially does not hold any quantum system. Each value is taken as the best among that obtained for 20 random initial strategies and the semidefinite programs are solved with MOSEK [68]. We find a separation between forwarding and quantum strategies for 35 of the 50 random games[23]. We also include the classical values ω_c and $\omega_c(\emptyset)$ of each game, with $\omega_c(\emptyset)$ being the classical value obtained on the empty graph

$$v_1 \quad v_2 \quad v_3.$$

[23]The code and numerical quantum strategies are accessible on the online repository `https://github.com/pierrepocreau/Latency-constrained-games`

$\omega_c, \omega_c(\emptyset)$ are obtained by enumerating all possible classical strategies. We can observe in Table 6 that the best classical strategies on the graph G always match the upper-bound ω_f, and thus that the best forwarding strategies for such games are essentially classical. Each semidefinite program converging with a small duality gap (well below 1e−6), any separation of the order of 1e−3 translate to a meaningful separation between the different type of strategies.

Table 6: Classical values $\omega_c(\emptyset)$ and ω_c as well as lower and upper bounds on ω_q and ω_f for 50 random XOR games, with $n = m = 3$. "Gap" refers to the difference between the lower bound on ω_q and the upper bound on ω_f.

ID	$\omega_c(\emptyset)$	ω_c	Upper ω_f	Lower ω_q	Gap
5	7.778e-01	8.889e-01	8.889e-01	8.928e-01	3.948e-03
11	7.778e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
17	7.778e-01	7.778e-01	7.778e-01	7.778e-01	-9.197e-10
26	7.778e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
37	7.778e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
69	7.778e-01	7.778e-01	7.778e-01	7.778e-01	-3.765e-10
73	1.000e+00	1.000e+00	1.000e+00	1.000e+00	1.025e-08
85	6.667e-01	8.889e-01	8.889e-01	8.928e-01	3.948e-03
88	7.778e-01	7.778e-01	7.778e-01	7.778e-01	-5.602e-10
105	8.889e-01	8.889e-01	8.889e-01	8.889e-01	-1.566e-09
152	7.778e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
161	6.667e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
184	6.667e-01	8.889e-01	8.889e-01	8.928e-01	3.948e-03
205	7.778e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
206	6.667e-01	8.889e-01	8.889e-01	8.889e-01	-4.864e-09
252	6.667e-01	8.889e-01	8.889e-01	8.928e-01	3.948e-03
257	7.778e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
259	6.667e-01	8.889e-01	8.889e-01	8.928e-01	3.948e-03
275	6.667e-01	8.889e-01	8.889e-01	8.928e-01	3.948e-03
289	7.778e-01	8.889e-01	8.889e-01	8.928e-01	3.948e-03
289	7.778e-01	8.889e-01	8.889e-01	8.889e-01	-3.510e-09
314	6.667e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
316	7.778e-01	8.889e-01	8.889e-01	8.928e-01	3.948e-03
335	7.778e-01	7.778e-01	7.778e-01	7.778e-01	1.051e-08
363	7.778e-01	8.889e-01	8.889e-01	8.928e-01	3.948e-03
386	7.778e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
404	7.778e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
487	7.778e-01	8.889e-01	8.889e-01	8.889e-01	-2.048e-09
488	6.667e-01	8.889e-01	8.889e-01	8.928e-01	3.948e-03
494	7.778e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
514	8.889e-01	8.889e-01	8.889e-01	8.928e-01	3.948e-03
552	7.778e-01	8.889e-01	8.889e-01	8.889e-01	-1.754e-08
568	6.667e-01	1.000e+00	1.000e+00	1.000e+00	3.335e-08
571	6.667e-01	8.889e-01	8.889e-01	8.928e-01	3.948e-03

Continued on next page

ID	$\omega_c(\emptyset)$	ω_c	Upper ω_f	Lower ω_q	Gap
583	6.667e-01	8.889e-01	8.889e-01	8.928e-01	3.947e-03
590	6.667e-01	1.000e+00	1.000e+00	1.000e+00	-2.155e-10
628	6.667e-01	8.889e-01	8.889e-01	8.928e-01	3.948e-03
679	6.667e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
681	6.667e-01	8.889e-01	8.889e-01	8.928e-01	3.948e-03
745	7.778e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
773	7.778e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
776	7.778e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
791	6.667e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
807	7.778e-01	8.889e-01	8.889e-01	8.928e-01	3.947e-03
809	7.778e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
824	6.667e-01	8.889e-01	8.889e-01	8.928e-01	3.948e-03
844	7.778e-01	7.778e-01	7.778e-01	8.333e-01	5.556e-02
885	7.778e-01	8.889e-01	8.889e-01	8.889e-01	-3.362e-09
935	7.778e-01	7.778e-01	7.778e-01	7.778e-01	3.024e-10
987	8.889e-01	8.889e-01	8.889e-01	8.889e-01	7.772e-09

D Three-party XOR games with aggregated parties

For the three-party XOR games $\{\mathrm{XOR}_i\}_{i=1}^4$ in Section 4.3.1, the quantum upper bound ω_q^{upper} is sufficiently tight, in the sense that we can find quantum strategies that achieve a winning probability that matches this upper bound within numerical precision. For XOR_1 and XOR_4, the strategy is that every party always outputs 0, and the quantum value $\frac{7}{8}$ and $\frac{3}{4}$ are trivially achieved. For XOR_2, let v_3 always output 0, then winning probability of XOR_2 becomes

$$\begin{aligned} p_{\text{win}} =& \frac{1}{8}\sum_{s_1,s_2,s_3}\sum_{a_1,a_2}\beta_{s_1,s_2,s_3}(a_1\oplus a_2)p(a_1,a_2|s_1,s_2)p(0|s_3) \\ =& \frac{1}{4}\sum_{s_1,s_2}\sum_{a_1,a_2}\beta'_{s_1,s_2}(a_1\oplus a_2)p(a_1,a_2|s_1,s_2), \end{aligned}$$

with

$$\beta'_{s_1,s_2}(a_1\oplus a_2) \coloneqq \max[(-1)^{s_1s_2}\cdot(-1)^{a_1\oplus a_2},0].$$

This is nothing but the winning probability of a CHSH game between the parties v_1 and v_2. Hence there exists quantum strategy achieves the value $\cos^2\frac{\pi}{8}$ for XOR_2, which matches the upper bound 0.8536 computed via the third-level of the NPA hierarchy. Similarly, for XOR_3, let v_2 always output 0. The three-party XOR game then reduces to a CHSH game between v_1 and v_3. Then they can win this game with probability $\cos^2\frac{\pi}{8}$.

When the latency constraint is relaxed such that $t\in[d/c,d'/c)$, we can aggregate v_1,v_2 and the winning probability is given by

$$p_{\text{win}}^{\text{agg}} = \frac{1}{8}\sum_{s_1,s_2,s_3}\sum_{a_2,a_3}\beta_{s_1,s_2,s_3}(a_2\oplus a_3)p(a_2,a_3|(s_1,s_2),s_3).$$

For the game XOR_1, its probabilistic predicate is

$$\beta_{s_1,s_2,s_3}(a_2\oplus a_3) = \max[(-1)^{s_1s_2s_3}\cdot(-1)^{a_2\oplus a_3},0]$$

We can let $\tilde{s}_2 \coloneqq s_1 \cdot s_2$ and define a binary-input, binary-output 2-party nonlocal game, whose probabilistic predicate is

$$\beta'_{\tilde{s}_2,s_3}(a_2 \oplus a_3) = \max[(-1)^{\tilde{s}_2 s_3} \cdot (-1)^{a_2 \oplus a_3}, 0]$$

and the four inputs $\{(\tilde{0},0),(\tilde{0},1),(\tilde{1},0),(\tilde{1},1)\}$ are chosen with probability $\{\frac{3}{8},\frac{3}{8},\frac{1}{8},\frac{1}{8}\}$, respectively. Its winning probability is

$$p'_{\text{win}} = \frac{1}{8}\sum_{\tilde{s}_2,s_3}\sum_{a_2,a_3}(2+(-1)^{\tilde{s}_2})\beta'_{\tilde{s}_2,s_3}(a_2 \oplus a_3)p(a_2,a_3|\tilde{s}_2,s_3)$$

For the game XOR_1 after aggregation, the parties can achieve the behavior

$$p(a_2,a_3|(0,0),s_3) = p(a_2,a_3|(0,1),s_3) = p(a_2,a_3|(1,0),s_3) = p(a_2,a_3|\tilde{0},s_3)$$

by producing the same output for $(s_1,s_2) \in \{(0,0),(0,1),(1,0)\}$, so they win with the same probability as p'_{win}. Meanwhile, in the binary-input, binary-output 2-party nonlocal game, the parties can realize the behavior

$$p(a_2,a_3|\tilde{0},s_3) = \frac{1}{3}\left[p(a_2,a_3|(0,0),s_3) + p(a_2,a_3|(0,1),s_3) + p(a_2,a_3|(1,0),s_3)\right]$$

to win with the same probability as p_{win} in XOR_1, by randomly choosing a strategy from three candidates when receives input $\tilde{s}_2 = \tilde{0}$. That is, the three-party XOR game XOR_1 is equivalent to the binary-input, binary-output 2-party nonlocal game above, when v_1 and v_2 can communicate. For the nonlocal game, if we map the outputs $\{0,1\}$ respectively to outputs $\{+1,-1\}$ of some binary observables $K_i^{s_i}$ respectively for each party, we can express the winning probability p'_{win} of a quantum strategy in operator form as

$$p'_{\text{win}} = \frac{1}{2} + \frac{1}{16}\langle B^{\alpha=3}_{CHSH}\rangle,$$

where $\langle\cdot\rangle$ denotes the expectation with some quantum state $|\psi\rangle$, with

$$B^{\alpha=3}_{CHSH} = 3K_2^{\tilde{0}}K_3^0 + 3K_2^{\tilde{0}}K_3^1 + K_2^{\tilde{1}}K_3^0 - K_2^{\tilde{1}}K_3^1.$$

Here, $B^{\alpha=3}_{CHSH}$ is a generalized CHSH operator

$$B^{\alpha}_{\text{CHSH}} \coloneqq \alpha\langle A^0B^0\rangle + \alpha\langle A^0B^1\rangle + \langle A^1B^0\rangle - \langle A^1B^1\rangle,$$

whose quantum value is $2\sqrt{1+\alpha^2}$, derived analytically in [49]. Consequently, the quantum value ω_q^{agg} of the aggregated three-party XOR game XOR_1 is $\frac{4+\sqrt{10}}{8}$.

Similarly, for XOR_3 and XOR_4, we can take the map $\{(0,0),(0,1)\} \mapsto \tilde{0}$, $\{(1,0),(1,1)\} \mapsto \tilde{1}$ on (s_1,s_2), which shows that both XOR_3 and XOR_4 is equivalent to the CHSH game when v_1 and v_2 are aggregated. So their quantum value is $\cos^2\frac{\pi}{8}$.

Finally, for XOR_2, let v_3 always output 0, then the game reduces to a CHSH game between v_1 and v_2 as discussed in Section 4.3.1. Since v_1 and v_2 can win with certainty if they can communicate. So the quantum value ω_q^{agg} of the aggregated XOR_2 game is 1.