RESEARCH UNIVERSITY
EXCELLENCE INITIATIVE

AGH

Akademia Górniczo-Hutnicza
im. Stanisława Staszica w Krakowie

AGH University of Krakow

Multi-Agent Reinforcement Learning for Nonlocal Games via Shared VQCs

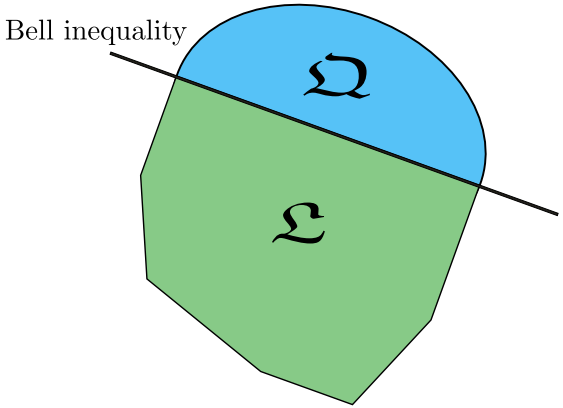
Dawid Mazur[1] Piotr Gawron[2,3]

[1]Faculty of Computer Science, AGH University of Krakow,
al. Mickiewicza 30, 30-059 Cracow, Poland

[2]Nicolaus Copernicus Astronomical Center, Polish Academy of Sciences,
ul. Bartycka 18, 00-716 Warsaw, Poland

[3]Center of Excellence in Artificial Intelligence, AGH University of Krakow,
al. Mickiewicza 30, 30-059 Cracow, Poland

Motivation



CHSH game

- questions and answers are bits: $x, y, a, b \in \{0, 1\}$
- (x, y) drawn uniformly
- agents win iff $a \oplus b = x \wedge y$

(x, y)	win	lose
$(0, 0)$	$a = b$	$a \neq b$
$(0, 1)$	$a = b$	$a \neq b$
$(1, 0)$	$a = b$	$a \neq b$
$(1, 1)$	$a \neq b$	$a = b$

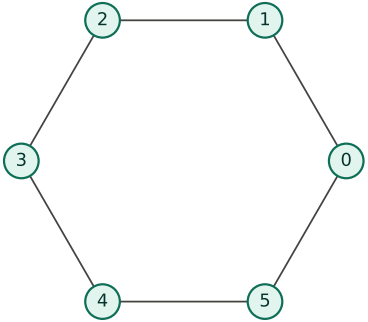
- local (LHV) value: $w_{\mathfrak{L}} = \frac{3}{4} = 0.75$
- quantum value: $w_{\mathfrak{Q}} = \cos^2(\pi/8) \approx 0.854$

Magic Square game

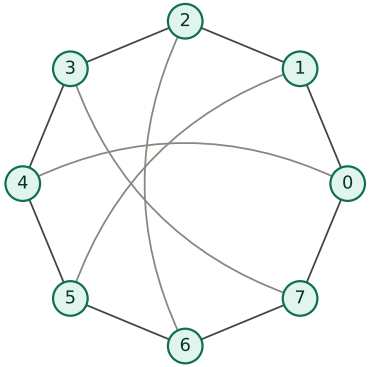
	$y=0$	$y=1$	$y=2$
$x=0$		b_1	
$x=1$	a_1	$a_2=b_2$	a_3
$x=2$		b_3	

- Alice fills her row x with ± 1 entries: $a_1 a_2 a_3 = +1$
- Bob fills his column y with ± 1 entries: $b_1 b_2 b_3 = -1$
- they win iff they agree on the shared cell: $a_2 = b_2$
- local (LHV) value: $w_{\mathfrak{L}} = \frac{8}{9} \approx 0.889$
- quantum value: $w_{\mathfrak{Q}} = 1$ (perfect strategy)

Rendezvous problem



(a) Cycle graph, $|V| = 6$



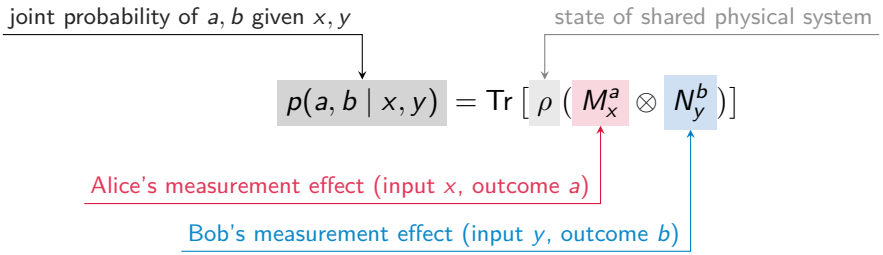
(b) Cubic graph, $|V| = 8$

Figure: The two graph topologies used in the Rendezvous game.

Rendezvous Game

- players on vertices of a graph $G = (V, E)$ — V vertices, E edges; Alice's question $x \in V$ and Bob's question $y \in V$ are their own starting vertices
- answer: a move — to a neighboring vertex (or "stay") — chosen without knowledge of the other's start
- win condition: both agents on the *same* vertex after the move
- achievable values of the game depend on graph topology

Correlations



- all joint probabilities $p(a, b \mid x, y)$ are the *correlations*
- quantum set $\mathfrak{Q}$ — correlations from any shared state ρ
- local set $\mathfrak{L}$ — correlations from *separable* ρ
- $\mathfrak{L} \subsetneq \mathfrak{Q}$ — *entangled* ρ can produce correlations no separable state reproduces

Nonlocal Games

Definition (Nonlocal Game)

Question sets X and Y, answer sets A and B, a probability distribution $q : X \times Y \to [0, 1]$ over questions, and a predicate function $W : X \times Y \times A \times B \to \{0, 1\}$ together define a nonlocal game $\mathcal{G} = (W, q)$.

How a round is played:

1. a strategy is selected before the start of a game
2. questions $(x, y) \sim q$ are drawn — Alice gets x and Bob gets y
3. Alice answers $a \in A$, Bob answers $b \in B$
4. they **win** iff $W(a, b \mid x, y) = 1$

Value of a game

$w_{\mathfrak{L}}$, $w_{\mathfrak{Q}}$ — best winning probability achievable with strategies from $\mathfrak{L}$, $\mathfrak{Q}$.

Dec-POMDP

agents and their joint action space

environment state and how it evolves

$$\mathcal{M} = \langle \mathcal{D}, \mathcal{A}, \mathcal{S}, T, \mathcal{O}, O, R, h, b_0 \rangle$$

what each agent actually sees — own observation only

shared reward, episode horizon, initial state

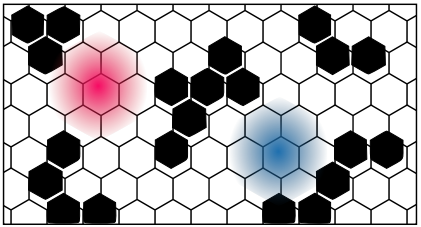
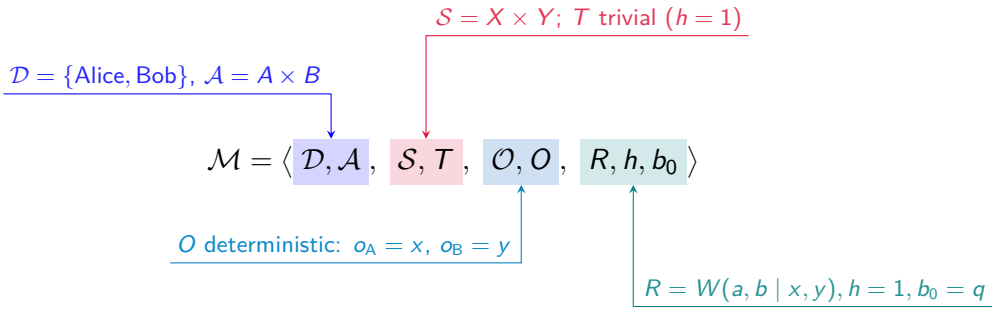


Figure: Example environment with two agents — Alice (red) and Bob (blue).

Problem Formulation

We express a nonlocal game $\mathcal{G} = (W, q)$ as a Dec-POMDP:



Model

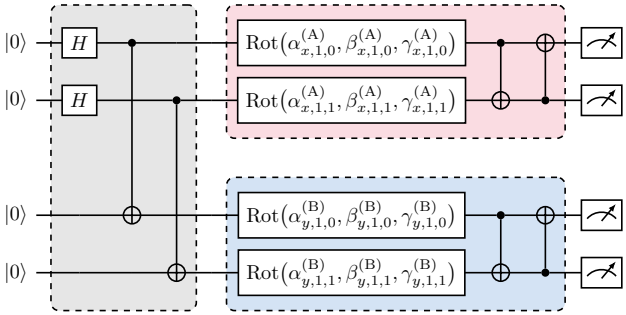


Figure: Variational quantum circuit for Alice ($x \in X$) and Bob ($y \in Y$), each with $n = 2$ qubits and $L = 1$ layer. Gray: entangled-state preparation. Red/blue: Alice's/Bob's variational unitary, weights indexed by their question.

Model

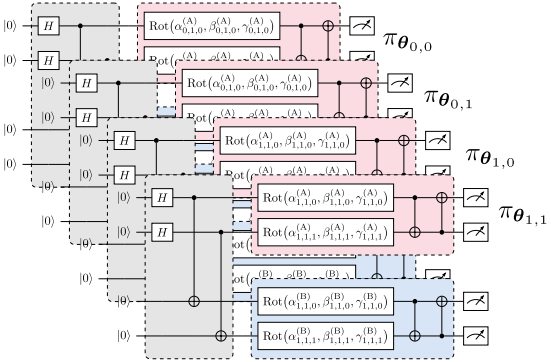


Figure: Constructed circuits for a game with question sets $X = Y = \{0, 1\}$, entangled initial state, $n = 2$ qubits per agent, $L = 1$ layer. Each question pair (x, y) gets its own circuit, and thus its own policy $\pi_{\boldsymbol{\theta}}(a, b \mid x, y)$.

From Measurement Probabilities to Policy

- computational-basis projectors $\Pi_i = |i\rangle\langle i|$, $i \in \{0, \ldots, 2^n - 1\}$

$$\tilde{p}(k, m \mid x, y) = \mathrm{Tr}\left[\, |\psi(\boldsymbol{\theta}_{x,y})\rangle \langle\psi(\boldsymbol{\theta}_{x,y})| \left(\Pi_k \otimes \Pi_m\right)\right]$$

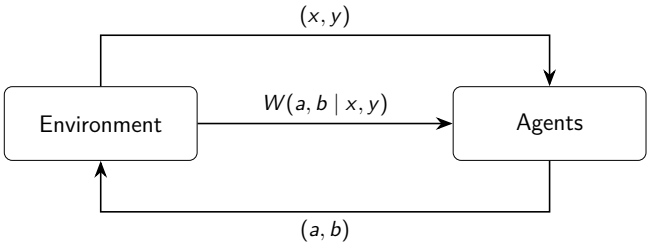
- raw outcomes $k, m \in \{0, \ldots, 2^n - 1\}$ — not yet the actions in A, B
- partition into families $\{\mathcal{P}_a^{(\mathrm{A})}\}_{a \in A}$, $\{\mathcal{P}_b^{(\mathrm{B})}\}_{b \in B}$ — one subset per action

$$p(a, b \mid x, y) = \sum_{k \in \mathcal{P}_a^{(\mathrm{A})}} \sum_{m \in \mathcal{P}_b^{(\mathrm{B})}} \tilde{p}(k, m \mid x, y)$$

$$\pi_{\boldsymbol{\theta}}(a, b \mid x, y) := p(a, b \mid x, y)$$

- joint policy $\pi_{\boldsymbol{\theta}}$: parametrized by the shared VQC — factorizes into independent per-agent policies or not, depending on the shared resource
- **in training**: exact probability vector, no sampling
- **at execution**: each agent samples one outcome from their own subsystem, mapped through the same partition

REINFORCE Training



Monte Carlo REINFORCE gradient (batch of N trajectories)

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=0}^{N-1} W(a_i, b_i \mid x_i, y_i) \nabla_{\boldsymbol{\theta}} \ln \pi_{\boldsymbol{\theta}}(a_i, b_i \mid x_i, y_i)$$

- weight update ($z = 0, \ldots, Z-1$ batches): $\boldsymbol{\theta}_{z+1} = \boldsymbol{\theta}_z + \eta \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_z)$

CHSH Results

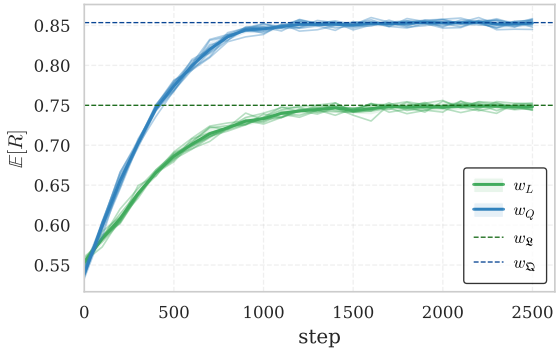


Figure: Training curves for the CHSH game, averaged over ten runs. Here, w_Q and w_L denote agents with entangled and product-state initialization, with solid lines showing the mean and shading one standard deviation. Horizontal lines mark the quantum ($w_{\mathfrak{Q}}$) and LHV ($w_{\mathfrak{L}}$) upper bounds.

Magic Square Results

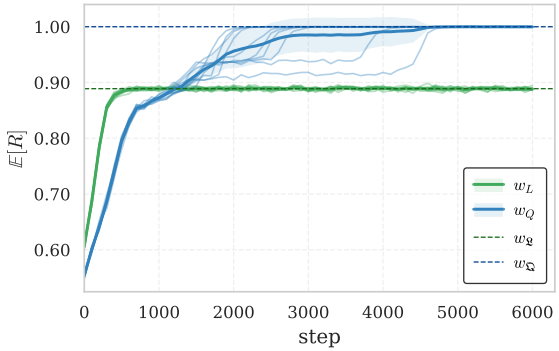


Figure: Training curves for the Magic Square game, averaged over ten runs. Here, w_Q and w_L denote agents with entangled and product-state initialization, with solid lines showing the mean and shading one standard deviation. Horizontal lines mark the quantum ($w_{\mathfrak{Q}}$) and LHV ($w_{\mathfrak{L}}$) upper bounds.

Rendezvous Results — Cycle Graph

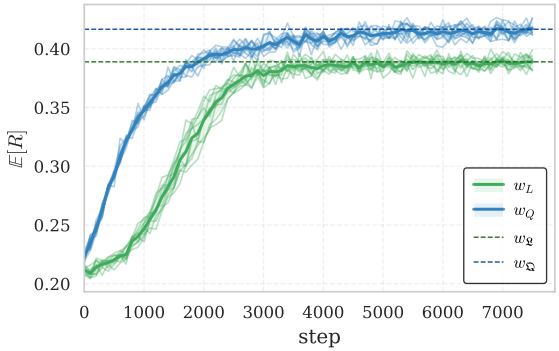


Figure: Training curves for the Rendezvous game for cycle graph, averaged over ten runs. Here, w_Q and w_L denote agents with entangled and product-state initialization, with solid lines showing the mean and shading one standard deviation. Horizontal lines mark the quantum ($w_{\mathfrak{Q}}$) and LHV ($w_{\mathfrak{L}}$) upper bounds.

Rendezvous Results — Cubic Graph

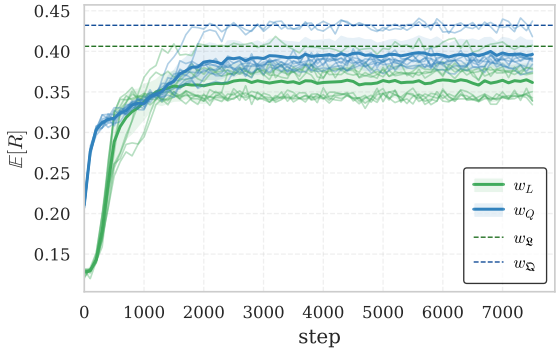


Figure: Training curves for the Rendezvous game for cubic graph, averaged over ten runs. Here, w_Q and w_L denote agents with entangled and product-state initialization, with solid lines showing the mean and shading one standard deviation. Horizontal lines mark the quantum ($w_{\mathfrak{Q}}$) and LHV ($w_{\mathfrak{L}}$) upper bounds.

Thank you for your attention!