A Minimal Decision Capacity Threshold Prevents Catastrophic Exploitation in Self-Play RL

Arahan Kujur KUJURARAHAN@GMAIL.COM
Independent Researcher

Abstract

We show that a minimal threshold in decision capacity determines whether self-play reinforcement learning agents collapse under asymmetric rule perturbations. In Kuhn and Leduc Poker, we remove Player 0's ability to bet or raise—either at all decision nodes or only at the opening move. Across five seeds with paired card deals: (i) removing the bet/raise action at all decision nodes (capacity 0 in Kuhn; residual capacity >0 in Leduc, where fold/check-call remain) causes adaptive Q-learning to collapse toward exploitation (Kuhn: -0.93; Leduc: -0.31), while a frozen Q-learning baseline stays near -0.14, confirming the collapse is co-adaptation-driven; (ii) preserving a single decision point stabilises Q-learning near Nash equilibrium (Kuhn: -0.07; Leduc: -0.10); (iii) the pattern is timing-invariant: early, mid, and late perturbation produce identical collapse severity; (iv) collapse is fast, occurring within four episodes on average in Kuhn. These results reveal a structural instability in learning dynamics: equilibrium behavior becomes unsustainable once agents lose all contingent responses. We provide empirical evidence for a sharp threshold effect between zero and minimal decision capacity, observed consistently across two small imperfect-information poker games (Kuhn and Leduc) rather than being specific to Kuhn. We frame the collapse as a learning-dynamics instability measured against the original-game Nash value, not as a claim that the exploiting opponent behaves irrationally.

1. Introduction

Multi-agent reinforcement learning (MARL) agents trained through self-play have achieved superhuman performance in complex games [4], yet their robustness to structural changes in the environment remains poorly understood. Prior work has focused primarily on stochastic perturbations to observations or rewards, or on opponent modelling under distribution shift, leaving structural changes to the action space relatively unexplored. When the rules of a game change asymmetrically—for instance, when one player loses access to certain actions—how do self-play dynamics respond?

We study this question in two imperfect-information poker games: Kuhn Poker and Leduc Poker. After a training phase of normal self-play, we remove Player 0's ability to bet (Kuhn) or raise (Leduc) at a subset of decision nodes. This creates a controlled perturbation to the action space that reduces the agent's *decision capacity*—the number of information sets at which the agent retains more than one legal action.

Our central finding is a sharp threshold effect. When decision capacity drops to zero (every decision is forced), adaptive Q-learning collapses to near-maximal exploitation through a co-adaptation spiral. When even a single decision point is preserved, Q-learning stabilises near Nash equilibrium. A frozen Q-learning baseline—which stops updating at the perturbation point—does not collapse, isolating continued self-play adaptation as the mechanism rather than the constraint itself. We refer to this phenomenon as a decision-capacity collapse in self-play learning dynamics.

We further show that the collapse is timing-invariant (identical outcomes whether the perturbation is applied early, mid, or late in training) and replicates across both games, ruling out Kuhn-specific artifacts. These results indicate that collapse is not a function of training quality or convergence level, but a structural property of the constrained game. More broadly, this suggests that classical equilibrium concepts may be insufficient to characterise stability in learning-driven multi-agent systems under structural constraints. Our contributions are:

- We identify a sharp, discontinuous threshold in decision capacity that governs whether learning dynamics converge to equilibrium or collapse to deterministic exploitation.
- We isolate co-adaptation under constraint as the mechanism via a frozen baseline comparison.
- We demonstrate timing-invariant collapse across perturbation schedules, showing the effect is structural rather than training-dependent.
- We replicate the phenomenon across two structurally distinct poker games (Kuhn and Leduc).

2. Related Work

Regret minimisation in games. Counterfactual regret minimisation (CFR) [6] is the standard algorithm for computing Nash equilibria in imperfect-information games. Its convergence guarantees in two-player zero-sum settings are well-established, and it has been scaled to large poker variants [1]. We use CFR as a planning baseline whose frozen strategy represents the best-case response to perturbation without adaptation.

Self-play reinforcement learning. Self-play has driven major advances in game-playing agents, from TD-Gammon [5] through AlphaZero [4]. However, self-play dynamics can be unstable: agents may cycle, overfit to their own weaknesses, or fail to generalise [3]. Our work isolates a specific failure mode—co-adaptation-driven collapse under asymmetric constraint changes—that is distinct from the cyclic instabilities studied in general-sum games.

Robustness in multi-agent RL. Prior work on robustness in MARL has focused on adversarial perturbations to observations or rewards [2], domain randomisation, or opponent modelling under distribution shift. Our setting differs: we perturb the *action space* of one player asymmetrically, eliminating decision points rather than adding noise. This structural perturbation reveals a qualitative threshold effect that continuous perturbations would not produce.

3. Background

Kuhn Poker. A two-player zero-sum game with three cards ($J < Q < K$), two actions (pass, bet), and an ante of 1. The Nash equilibrium value for Player 0 is $-1/18 \approx -0.056$.

Leduc Poker. A two-player zero-sum game with six cards (J, Q, K in two suits), three actions (fold, check/call, raise), and two rounds of fixed-limit betting (round 1: raise size 2; round 2: raise size 4; maximum two raises per round). The Nash equilibrium value for Player 0 is approximately -0.087. Our from-scratch CFR implementation converges to -0.0866 after 5,000 iterations, matching the known analytical value.

Agents. We evaluate three agent types:

- **CFR**: Counterfactual regret minimisation, trained to Nash equilibrium offline (deterministic full-tree enumeration). The frozen strategy is shared across seeds.
- **Q-Learning**: Tabular, ε-greedy ($\varepsilon = 0.15$), with Monte Carlo terminal updates. Continues learning after perturbation.
- **QL-Frozen**: Identical to Q-Learning during training. At the perturbation point, the Q-table and ε are frozen (greedy execution only).

4. Methodology

Each experiment runs five seeds, each consisting of 20,000 self-play episodes. A perturbation is applied to Player 0 at episode 10,000 unless otherwise noted.

Decision capacity. We define *decision capacity* as the number of reachable information sets in which the affected agent retains more than one legal action after perturbation. Capacity 0 means every decision is forced; capacity 1 means exactly one information set retains a genuine choice. In Kuhn, removing the bet action at all of Player 0's nodes yields capacity 0: every Player 0 information set becomes forced (forced check or forced fold). In Leduc, removing the *raise* action does *not* reach capacity 0, because fold and check/call remain legal at multiple Player 0 information sets; Player 0 therefore retains genuine decisions (residual capacity >0). To avoid ambiguity we refer to the Leduc condition as *raise removal* (reduced capacity) rather than capacity 0 throughout, and we report the residual decision points explicitly in Section 5.7.

Perturbation mechanism. Perturbations are implemented as *action masking*: at the targeted information sets, the bet action (Kuhn) or raise action (Leduc) is removed from Player 0's legal action set, and the policy selects among the remaining legal actions only. *Full removal* masks the action at all of Player 0's information sets; *root-only removal* masks it only at the opening information set, leaving the downstream call/fold decision (the "pb" node in Kuhn) intact. No other agent, reward, or game parameter is altered at the perturbation point, so any change in outcome is attributable to the masked actions and to continued self-play.

Reference value. All rewards are reported from Player 0's perspective and compared against the *original-game* Nash value (Kuhn: $-1/18 \approx -0.056$; Leduc: ≈ -0.087). We interpret a post-perturbation collapse as a *learning-dynamics instability*—self-play fails to settle at the safe value of the constrained game—rather than as evidence that the exploiting opponent is irrational. Exploiting a fully forced opponent is itself a rational best response; our claim is that *continued* self-play drives the dynamics to that exploitative fixed point, which a frozen policy (QL-Frozen) avoids under the same constraint.

RNG design. Four independent random number generator streams are derived per seed: (1) card deals, shared between all agents for paired comparison; (2) CFR action selection; (3) Q-Learning action selection; (4) QL-Frozen action selection. CFR training uses deterministic full-tree enumeration and runs once outside the seed loop.

Statistical analysis. We report paired t-tests across seeds, bootstrap 95% confidence intervals (10,000 resamples), and Cohen's d effect sizes. An adaptive burn-in (first 25% of each phase, capped at 5,000/2,000 episodes) is excluded for stable estimates. Due to low across-seed variance under paired evaluation, effect sizes are large and should be interpreted qualitatively as indicating direction and reliability rather than as conventional benchmarks.

Time-to-collapse. The first episode at which the 200-episode moving average drops below -0.5.

5. Experiments

5.1. Kuhn Poker: Full Removal (Capacity 0)

Bet is removed from Player 0 at all decision nodes. Player 0 has zero remaining decisions: forced check, forced fold.

Table 1: Kuhn Poker, full removal (capacity 0). Mean Player 0 reward with 95% bootstrap CIs across five seeds.

Agent	Pre (95% CI)	Post (95% CI)	p	d
CFR	-0.053 $[-0.066, -0.041]$	-0.231 $[-0.235, -0.228]$	<0.0001	-11.9
Q-Learning	-0.035 $[-0.046, -0.025]$	$\mathbf{-0.927}$ $[-0.929, -0.925]$	<0.0001	-66.1

As shown in Table 1, Q-Learning collapses to near -1.0: the self-play opponent learns that Player 0 always folds and converges to unconditional betting. The ε-floor (0.15) prevents the average from reaching exactly -1.0. CFR drops to -0.23, bounded because its opponent also plays Nash.

Collapse speed. Q-Learning collapsed in all five seeds within a mean of four episodes after perturbation.

Figure 1: Kuhn Poker, full removal: moving-average Player 0 reward. Q-Learning collapses within a mean of four episodes after perturbation at episode 10,000, illustrating the transition to a deterministic exploitation regime.

5.2. Kuhn Poker: Root-Only Removal (Capacity 1)

Bet is removed from Player 0 only at the root. Player 0 retains the call/fold decision at the "pb" node.

Table 2: Kuhn Poker, root-only removal (capacity 1).

Agent	Pre (95% CI)	Post (95% CI)	p	d
CFR	-0.053 $[-0.066, -0.041]$	-0.061 $[-0.065, -0.055]$	0.27	-0.6
Q-Learning	-0.035 $[-0.046, -0.025]$	-0.073 $[-0.081, -0.064]$	0.001	-3.6

As shown in Table 2, CFR barely shifts ($p = 0.27$, not significant)—removing an action at an indifference point does not change the equilibrium value. Q-Learning drops modestly ($\Delta = -0.037$) then stabilises: the single remaining decision point forces the opponent to bet honestly, bounding exploitation.

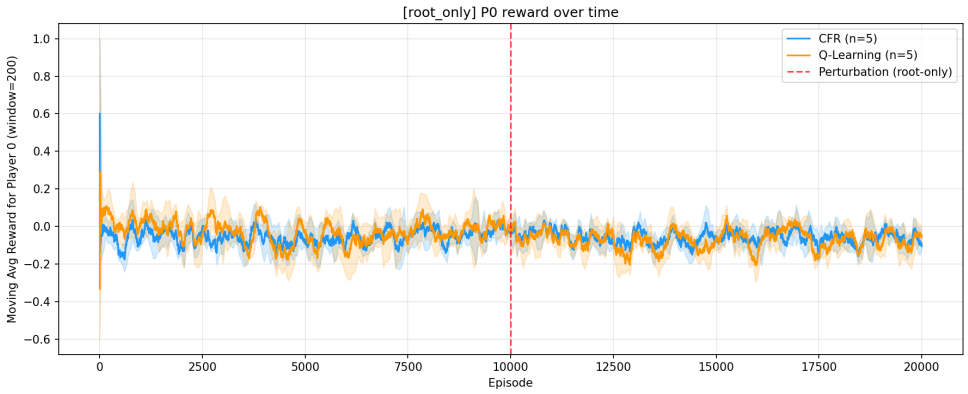

Figure 2: Kuhn Poker, root-only removal: a single retained decision point is sufficient to prevent collapse. Q-Learning stabilises near Nash equilibrium after a brief transient dip.

5.3. Decision Capacity Sweep

Table 3: Kuhn Poker capacity sweep: post-perturbation mean Player 0 reward.

Capacity	Description	CFR Post	QL Post
0	All actions removed	-0.231	$\mathbf{-0.927}$
1	Root bet removed, call/fold kept	-0.061	-0.073
2	No perturbation (control)	-0.062	-0.034

As shown in Table 3, the jump from capacity 0 to capacity 1 is catastrophic ($\Delta = -0.85$ for Q-Learning). The jump from 1 to 2 is marginal ($\Delta = +0.04$). The relationship between decision

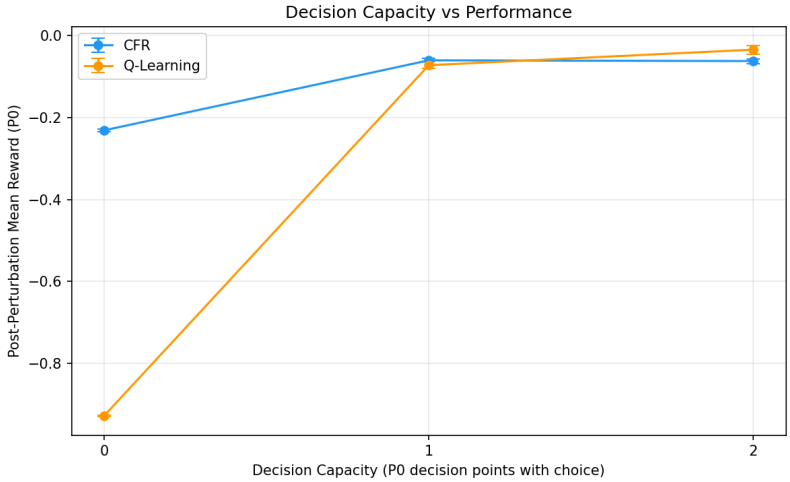

Figure 3: Decision capacity versus post-perturbation reward with 95% bootstrap CIs. The discontinuity between capacity 0 and capacity 1 confirms a sharp threshold rather than a gradual decline.

capacity and exploitability is discontinuous at the zero-to-one boundary—a sharp threshold effect, not a smooth gradient (Figure 3).

5.4. Frozen vs. Adaptive Q-Learning

Table 4: Frozen versus adaptive Q-Learning under full removal (Kuhn).

Agent	Post (95% CI)	Collapsed
CFR	-0.231 $[-0.235, -0.228]$	4/5 transient
Q-Learning	$\mathbf{-0.927}$ $[-0.929, -0.925]$	5/5 (delay: 4 eps)
QL-Frozen	-0.141 $[-0.406, -0.007]$	2/5

As shown in Table 4, the paired comparison between Q-Learning and QL-Frozen yields $\Delta = -0.787$, $p = 0.004$, $d = -2.6$. The frozen agent avoids collapse because its policy does not shift toward always-fold. Without the co-adaptation spiral—in which the opponent learns to exploit Player 0's increasing passivity—the perturbation alone causes only moderate degradation.

This isolates continued adaptation under constraint, rather than the quality of the pre-perturbation policy, as the mechanism of collapse. Because the only difference between the two arms is whether learning continues after the perturbation—the constraint, pre-perturbation policy, and card deals are identical—the contrast is a clean causal identification rather than a correlational observation. The perturbation creates the vulnerability; self-play learning converts it into exploitation.

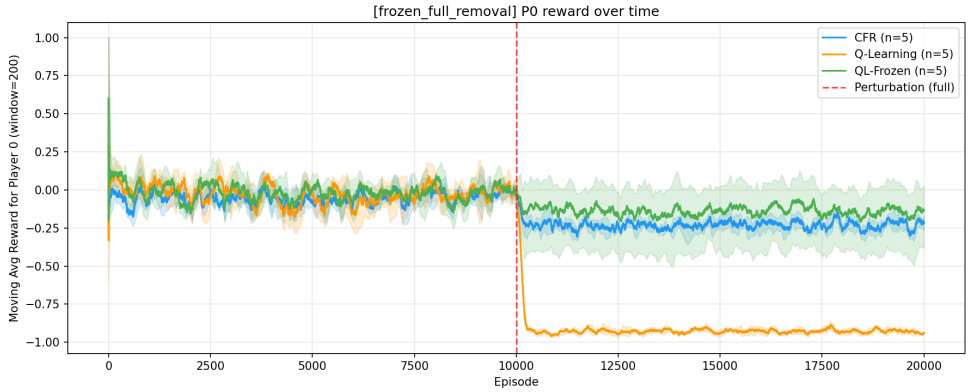

Figure 4: Frozen versus adaptive Q-Learning under full removal. QL-Frozen avoids catastrophic collapse, isolating continued co-adaptation—not the perturbation itself—as the mechanism driving exploitation.

Under root-only removal (capacity 1), no significant difference exists between adaptive and frozen Q-Learning ($p = 0.32$). When capacity ≥ 1, neither agent collapses regardless of whether learning continues.

5.5. Perturbation Severity Sweep

We vary both *when* the perturbation is applied (episode 3,000, 10,000, or 17,000) and *how strongly* (severe: all nodes; mild: root only).

Table 5: Perturbation timing × severity (Kuhn). Post-perturbation Q-Learning reward.

Timing	Pert. Episode	Severe (cap. 0)	Mild (cap. 1)
Early	3,000	$\mathbf{-0.926}$	-0.063
Mid	10,000	$\mathbf{-0.927}$	-0.073
Late	17,000	$\mathbf{-0.925}$	-0.061

As shown in Table 5, under severe perturbation collapse is **timing-invariant**: whether Q-values have barely begun learning (early) or are well-converged (late), the outcome is identical. Under mild perturbation, no collapse occurs at any timing. Collapse is structural, not dependent on training stage or convergence level. What determines the outcome is whether the perturbation eliminates all contingent responses, not when it occurs.

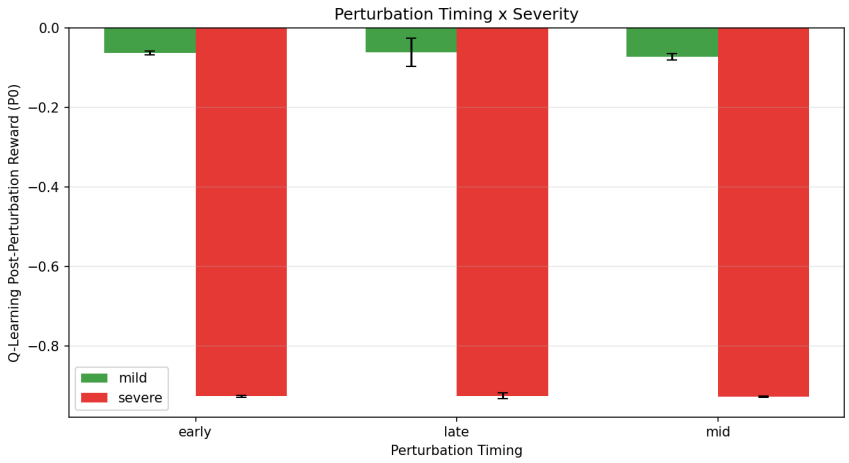

Figure 5: Perturbation timing × severity. Collapse severity is determined entirely by capacity, not by training stage: severe perturbation produces identical outcomes at all timings.

5.6. Stochastic Masking

When the root-only perturbation is applied with 50% probability per episode (stochastic masking), neither agent collapses: CFR post $= -0.066$ ($p = 0.14$), Q-Learning post $= -0.049$ ($p = 0.06$). Intermittent perturbation does not trigger the exploitation spiral.

5.7. Leduc Poker: Cross-Game Replication

Both agents degrade under raise removal. Q-Learning's post-perturbation drop is not statistically significant at conventional thresholds ($p = 0.07$), but the direction of the effect is consistent with Kuhn and aligns with the capacity-based explanation. We emphasise that, under our own definition, this Leduc condition is *not* capacity 0: removing the raise action leaves fold and check/call legal at multiple Player 0 information sets (Player 0 retains a genuine decision at every reachable informa-

Table 6: Leduc Poker, raise removal (raise removed at all Player 0 nodes). Note that this condition is *not* capacity 0: fold and check/call remain legal, so Player 0 retains residual decisions.

Agent	Pre (95% CI)	Post (95% CI)	p	d
CFR	-0.094 $[-0.111, -0.078]$	-0.281 $[-0.310, -0.247]$	0.002	-3.4
Q-Learning	-0.122 $[-0.146, -0.098]$	-0.307 $[-0.429, -0.219]$	0.07	-1.1

tion set where it can still fold or call), so residual decision capacity is strictly positive. The collapse is correspondingly less extreme than in Kuhn (-0.31 versus -0.93): Leduc tests the threshold from the *reduced-capacity* side rather than the strict zero-capacity side, and the milder collapse is exactly what the capacity-monotone reading of our results predicts. Collapse speed is also slower (mean delay: 104 episodes for Q-Learning) due to the larger state space.

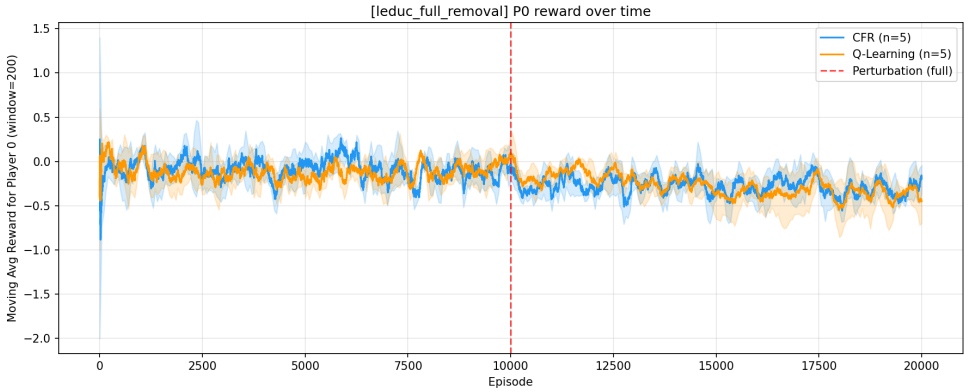

Figure 6: Leduc Poker, full removal: both agents degrade after perturbation. Q-Learning shows larger cross-seed variance than in Kuhn, reflecting the greater residual action space in Leduc.

Under root-only removal in Leduc, neither agent degrades significantly (CFR: $p = 0.75$; Q-Learning: $p = 0.40$). The capacity-1 stabilisation observed in Kuhn replicates in Leduc.

Cross-game comparison. The qualitative pattern is consistent across both games: severe degradation under the most-restrictive removal, and stability once a contingent response is preserved. The quantitative severity tracks residual decision capacity. This is clearest at the extremes: Kuhn full removal reaches true capacity 0 and collapses to -0.93, whereas Leduc raise removal retains fold/check-call (residual capacity >0) and degrades only to -0.31. The two games thus probe different points on the same capacity–exploitability relationship rather than contradicting it.

Table 7: Cross-game comparison of post-perturbation Q-Learning reward. "Full/most-restrictive removal" denotes capacity 0 in Kuhn (all Player 0 actions forced) but only *raise removal* in Leduc, where fold/check-call remain (residual capacity >0); the two are not directly equivalent in capacity terms.

	Kuhn	Leduc
Most-restrictive removal (QL Post)	$\mathbf{-0.927}$	$\mathbf{-0.307}$
capacity of this condition	0	>0 (raise removed)
collapse delay	4 eps	104 eps
Root-only removal (QL Post)	-0.073	-0.102
collapse?	No	No

5.8. Variance Decomposition

We decompose post-perturbation reward variance into environment (card deals) and policy (action selection) components by fixing one RNG source across seeds (Table 8).

Table 8: Variance decomposition under full removal (Kuhn).

Agent	V_{total}	V_{env}	V_{policy}	$V_{\text{interaction}}$
CFR	2.1×10^{-5}	1.5×10^{-5}	2.8×10^{-5}	-2.2×10^{-5}
Q-Learning	5×10^{-6}	7×10^{-6}	6×10^{-6}	-8×10^{-6}

As shown in Table 8, post-perturbation Q-Learning variance is extremely low across all sources. The near-zero total variance confirms that collapse corresponds to a deterministic fixed point of the self-play dynamics: the opponent converges to the same exploitation strategy regardless of card randomness or exploration noise. This indicates that collapse is not merely instability but convergence to a deterministic attractor under self-play dynamics—a qualitative distinction from the high-variance oscillations typically associated with training failure.

6. Why a Single Decision Point Bounds Exploitation

We give an informal argument for why retaining even one genuine decision changes the qualitative outcome, sketching the connection between decision capacity and exploitability that our measurements reflect. We do not claim a general theorem; the goal is to make the empirical threshold intelligible rather than to prove it.

Setup. After perturbation, fix Player 0's reachable information sets and let C denote the decision capacity (the number of information sets at which Player 0 retains more than one legal action). Self-play lets Player 1 adapt toward a best response to Player 0's induced policy π_0. Write $V_{\max}$ for the maximum per-hand loss the game can inflict on Player 0.

Capacity 0. If $C = 0$, every Player 0 action is forced, so π_0 is a fixed map independent of Player 1. Player 1 then faces no contingent threat: it can condition freely on its private card and on Player 0's deterministic behaviour, and the value to Player 0 is the minimum over Player 1 strategies of a

payoff function that Player 0 cannot influence. Nothing bounds Player 1's exploitation except $V_{\max}$. In Kuhn this is precisely the always-fold/always-bet attractor with value near -1 (Table 1).

Capacity ≥ 1. If Player 0 retains a genuine choice at some information set I, it can mix between actions at I and thereby punish any over-committed opponent. A purely exploitative pure strategy for Player 1 leaves it exploitable at I: e.g. the single retained call/fold in Kuhn means that if Player 1 bets too aggressively, Player 0 calls with strong hands and Player 1 loses the surplus. To avoid being punished, Player 1 must keep its behaviour honest in expectation, which caps Player 0's loss at the equilibrium value of the *reduced* game—close to the original Nash value (Table 2). The decisive factor is the *existence* of a punishing response, not the number of such responses, which is why the jump from $C = 0$ to $C = 1$ is discontinuous while the jump from $C = 1$ to $C = 2$ is marginal.

Implication. Reading these two regimes together, the attainable exploitation is bounded above by a quantity that drops sharply between $C = 0$ (loss $\approx V_{\max}$) and $C \geq 1$ (loss $\approx$ reduced-game Nash). The discontinuity we measure (Table 3, Figure 3) is the empirical signature of this bound, and the milder Leduc degradation is consistent with its residual capacity remaining strictly positive. A formal treatment would state this as an exploitability bound monotone in C for a given extensive-form game; we leave the general statement to future work and present the above as intuition aligned with our data.

7. Conclusion

We have shown that a minimal threshold in decision capacity determines whether self-play RL agents catastrophically collapse under asymmetric action-space perturbations. A single remaining contingent response—call/fold in Kuhn, any non-forced action in Leduc—compels the opponent to maintain strategic diversity, bounding losses near Nash equilibrium. When no contingent responses remain, the opponent can enforce a dominant strategy through co-adaptation: the constrained agent's forced passivity removes any incentive for honest play, collapsing the game to deterministic exploitation within episodes.

Within the two games we study, this threshold behaves as a structural property: across Kuhn and Leduc it does not depend on how well-trained the agent is or when the constraint is imposed—only on whether the agent retains at least one genuine decision. We deliberately stop short of asserting a general law; establishing whether the same threshold holds beyond small tabular poker is left to future work.

Limitations. We do not claim generality beyond the regime we study, and several scope limitations should be read alongside our results. *(i) Algorithm.* We use tabular, ε-greedy Q-learning; whether the same threshold appears under SARSA, policy-gradient self-play, NFSP-style learning, or deep self-play with function approximation is untested. *(ii) Scale.* Our games are small (Kuhn: 12 information sets; Leduc: 288) with exact CFR solutions; larger extensive-form games, larger state/action spaces, and continuous action spaces remain open. *(iii) Game class.* We study two-player zero-sum poker, so relevance to non-zero-sum or cooperative multi-agent settings is not established. *(iv) Exploration and seeds.* We fix $\varepsilon = 0.15$ and use five seeds; we do not sweep the exploration rate, and although five seeds suffice for the large Kuhn effects, statistical power is limited for subtler comparisons (e.g., Leduc Q-learning raise removal, $p = 0.07$). *(v) Metric.* We report reward against the self-play opponent relative to the original-game Nash value rather than exploitability or best-response/distance-to-constrained-Nash metrics; the latter would more directly

substantiate the game-theoretic interpretation. Accordingly we frame our findings as empirical evidence in small imperfect-information poker games rather than a general principle.

Future work. Natural extensions include: (i) scaling to larger games using neural network function approximation; (ii) evaluating additional self-play learning algorithms (e.g., SARSA, policy-gradient, NFSP) to test whether the threshold is specific to tabular Q-learning; (iii) reporting exploitability and best-response metrics against the constrained-game equilibrium to sharpen the game-theoretic claim; (iv) studying whether the threshold generalises to non-zero-sum or cooperative settings; (v) formalising the connection between decision capacity and exploitability bounds in extensive-form games (Section 6); and (vi) investigating whether graduated capacity reduction produces smooth degradation in richer games.

References

[1] Michael Bowling, Neil Burch, Michael Johanson, and Oskari Tammelin. Heads-up limit hold'em poker is solved. *Science*, 347(6218):145–149, 2015.

[2] Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. In *International Conference on Learning Representations*, 2020.

[3] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, et al. OpenSpiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

[4] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362 (6419):1140–1144, 2018.

[5] Gerald Tesauro. TD-Gammon, a self-teaching backgammon program, achieves master-level play. *Neural Computation*, 6(2):215–219, 1994.

[6] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret minimization in games with incomplete information. In *Advances in Neural Information Processing Systems*, volume 20, 2007.