A Regularization-Based Approach to Public Belief State Search for Adversarial Games

Sobhan Mohammadpour
MIT

Samuel Sokota
OpenAI

Brandon Kaplowitz
University of Oxford

Zico Kolter
Carnegie Mellon University

Noam Brown
OpenAI

Gabriele Farina
MIT

Abstract

The *public belief state* (PBS), which is the posterior over game histories conditioned on public information, is a fundamental abstraction for designing game-theoretically sound search algorithms for imperfect-information games. Existing sound PBS search techniques for adversarial games fall into two categories: gadget-game-based algorithms (including Libratus, DeepStack, Pluribus, Supremus, and Student of Games) and ReBeL-based algorithms. Both these categories possess disadvantages, including but not limited to a reliance on discontinuous functions, an inability to re-solve subgames, an implicit representation of policies, and high conceptual complexity. In this work, we propose Gravel, a new approach for PBS search—different from these previous two categories—that achieves game-theoretic soundness via regularization. Unlike the aforementioned approaches, Gravel relies on smooth functions, can initiate search at any point in the game, outputs explicit policies directly, and is underpinned by a simple conceptual framework. We demonstrate the effectiveness of Gravel on no-limit heads-up Texas hold'em—the most well-benchmarked problem in the literature—where it significantly outperforms ReBeL and Student of Games against Slumbot despite having only 1/10th of training samples and 1/32nd of the search budget.

1 Introduction

Search—also known as decision-time planning—is the process of revising or constructing a policy immediately before using that policy to make a decision. Search can yield vast improvements in playing strength by expending additional computation on the part of the game for which a decision is to be made. But designing search algorithms in imperfect-information games is complex, as the presence of information asymmetry violates the Markov assumption of dynamic programming.

One powerful abstraction for handling information asymmetry is the public belief state (PBS), which is a posterior over game histories conditioned on public information. PBSs capture all strategically relevant information, and thereby serve an analogous role to that of Markov states in fully observable settings [Nayyar et al., 2013]. Consequently, given a PBS, the search objective becomes conceptually straightforward: improve the policy for the subgame starting from that PBS [Dibangoye et al., 2013, Oliehoek, 2013]. This framework underpins many search algorithms for fully cooperative imperfect-information games [Lerer et al., 2020, Sokota et al., 2021, Fickinger et al., 2021, Sokota et al., 2022], as a single entity controls the policies of all agents, making the exact PBS deducible.

However, such an approach cannot be directly applied to adversarial imperfect-information games because opponent policies (and, as a result, PBSs) are unknown. The most straightforward workaround to this issue, sometimes referred to as *unsafe subgame solving*, is to assume that the opponent has been following the same policy as the search player thus far. However, as is suggested

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

by its name, unsafe subgame solving is unsound and can increase the exploitability of the search player [Ganzfried and Sandholm, 2015].

There are two main classes of approaches that improve upon the safety guarantees of unsafe subgame solving. The first is based on so-called gadget games, which are auxiliary games designed in such a way that their solutions are guaranteed not to increase exploitability [Burch et al., 2014, Moravcik et al., 2016, Moravčík et al., 2017, Brown and Sandholm, 2017, Brown et al., 2018]. Gadget games are the basis for most PBS search systems for adversarial games, including DeepStack [Moravčík et al., 2017], Libratus [Brown and Sandholm, 2018], Pluribus [Brown and Sandholm, 2019], Supremus [Zarick et al., 2020], and Student of Games [Schmid et al., 2023]. The second technique, which is the basis for the ReBeL algorithm [Brown et al., 2020], uses the idea of simulating the CFR-D algorithm [Burch et al., 2014].

Unfortunately, both gadget-game-based and ReBeL-based approaches have noteworthy downsides:

1. Learning a value network is difficult because information state-conditional equilibrium values are generally not unique, but cannot be mixed and matched with one another arbitrarily.
2. Learning a policy network is difficult because equilibrium policies can vary discontinuously as functions of subgame beliefs and gadget game alternative payoffs.
3. Gadget-game-based approaches require additional computation to estimate alternative payoff values.
4. ReBeL-based approaches cannot express their solution policy in closed form, which 1) makes it difficult for them to effectively warm start their searches from a policy network and 2) makes it difficult to use them as game solvers—a major use case for poker AI.
5. ReBeL-based approaches cannot safely perform search at every decision point—instead, they must wait until they reach the end of the depth-limited subgame that they previously searched over before they can safely perform search again.

Motivated by the deficiencies of gadget game- and ReBeL-based approaches, we propose a new, regularization-based class of approaches. This regularization-based approach is free of the downsides of the gadget game- and ReBeL-based approaches listed above, and on top of that, is conceptually simpler.

The basic idea is to use regularization to make *unsafe* subgame solving *safe*. We theoretically ground this technique by building on the results of Sokota et al. [2023b], who show that *unsafe* subgame solving is *safe* under exact solves with this kind of regularization, proving that safety is also recovered as the solve approaches exactness.

With this grounding in hand, we introduce the first practical instantiation of regularization-based PBS search, which we call Gravel. Gravel performs expert iteration [Anthony et al., 2017] using minimax entropy-regularized [Perolat et al., 2021] magnetic mirror descent (MMD) [Sokota et al., 2023a]. This configuration offers numerous advantages over those of past approaches, among the most exciting of which is the effectiveness with which its policy networks can be leveraged to warm start searches.

We empirically evaluate our method on the challenging established benchmark of no-limit Texas hold'em. We find that not only is our method substantially simpler than ReBeL, the closest algorithm to Gravel, but also that it achieves a 30% higher score while only requiring 1/10 of training samples and 1/32 of the search budget.

2 Background and Notation

We employ formalisms for two classes of games. The first, termed adversarial finite-horizon sequential games, involves two opposing players who take turns acting, with the game concluding after a predetermined number of steps; this formalism is equivalent to two-player zero-sum perfect-recall timeable [Jakobsen et al., 2016] extensive-form games, and encompasses the class of problems the paper regards. The second, alternating Markov games, is a special case of the first class of games in which there is a Markov state observable to all players; this formalism will allow us to describe the public belief alternating Markov game—a transformation of an adversarial finite-horizon sequential game that operates over public belief states. In turn, the public belief alternating Markov game

allows us to more rigorously articulate unsafe subgame solving, the reason that it is unsafe, and a result in the literature that suggests the existence of an approach to make it safe.

2.1 Adversarial Finite-Horizon Sequential Games

An adversarial finite-horizon sequential game is a tuple:

$$\langle \mathbb{A}, [\mathbb{O}_i], \mathbb{O}_{\text{pub}}, [\mathbb{H}_i], \mathbb{H}_{\text{pub}}, \mathbb{H}, \mu, [\mathcal{O}_i], \mathcal{O}_{\text{pub}}, \mathcal{R}, \mathcal{T}, t_{\max} \rangle, \text{ where}$$

- $i \in \{0, 1\}$ and $\iota = t \mod 2$ is the acting player at time step t.
- $\mathbb{A}$ is the set of actions.
- $\mathbb{O}_i$ is the set of private observations for player i.
- $\mathbb{O}_{\text{pub}}$ is the set of public observations (i.e., those that are immediate common knowledge).
- $\mathbb{H}_i = \cup_t (\mathbb{O}_{\text{pub}} \times \mathbb{O}_i \times \mathbb{A})^t \times \mathbb{O}_{\text{pub}} \times \mathbb{O}_i$ is the set of player i's information states.
- $\mathbb{H}_{\text{pub}} = \cup_t \mathbb{O}_{\text{pub}}^t$ is the set of public histories.
- $\mathbb{H} \subset \mathbb{H}_0 \times \mathbb{H}_1$ is the set of histories.
- $\mu \in \Delta(\mathbb{H}(h^0_{\text{pub}}))$ is the initial history distribution, where $\mathbb{H}(h_{\text{pub}})$ denotes the set of histories consistent with public history h_{pub}.
- $\mathcal{O}_i \colon \mathbb{H} \to \mathbb{O}_i$ is player i's observation function.
- $\mathcal{O}_{\text{pub}} \colon \mathbb{H} \to \mathbb{O}_{\text{pub}}$ is the public observation function.
- $\mathcal{R} \colon \mathbb{H} \times \mathbb{A} \to \mathbb{R}$ is player 0's reward function, which is the negation of player 1's reward function.
- $\mathcal{T} \colon \mathbb{H} \times \mathbb{A} \to \Delta(\mathbb{H})$ is the transition function.
- $t_{\max}$ is the time horizon at which the game terminates.

For an information state h_i, let $\mathbb{A}(h_i)$ be its legal actions; $\Delta(X)$ denotes the probability simplex on a finite set X.

2.2 Alternating Markov Games

An alternating Markov game (AMG) is a tuple

$$\langle \mathbb{A}, \mathbb{S}, s^0, \mathcal{R}, \mathcal{T}, t_{\max} \rangle, \text{ where}$$

- $\mathbb{S}$ is the set of Markov states.
- $s^0 \in \mathbb{S}$ is the initial state.
- $\mathcal{R} \colon \mathbb{S} \times \mathbb{A} \to \mathbb{R}$ is player 0's reward function, which is the negation of player 1's reward function.
- $\mathcal{T} \colon \mathbb{S} \times \mathbb{A} \to \Delta(\mathbb{S})$ is the transition function.
- i, ι, $\mathbb{A}$, and $t_{\max}$ are defined as they were in adversarial finite-horizon sequential games.

2.3 The Public Belief Alternating Markov Game

For every adversarial finite-horizon sequential game, there exists an associated public belief alternating Markov game (PuB-AMG) [Wiggers et al., 2016, Nayyar and Gupta, 2017, Brown et al., 2020]. The PuB-AMG differs from the underlying game from which it is derived in that the acting player must disclose its policy across all information states consistent with the public history at which it is making a decision. This allows both players to compute—and condition their policies on—the public posterior over histories of the underlying game, making the game perfect information.

Let

$$\langle \mathbb{A}, [\mathbb{O}_i], \mathbb{O}_{\text{pub}}, [\mathbb{H}_i], \mathbb{H}_{\text{pub}}, \mathbb{H}, \mu, [\mathcal{O}_i], \mathcal{O}_{\text{pub}}, \mathcal{R}, \mathcal{T}, t_{\max} \rangle,$$

be an adversarial finite-horizon sequential game. Then the associated PuB-AMG is

$$\langle \tilde{\mathbb{A}}, \tilde{\mathbb{S}}, \tilde{s}^0, \tilde{\mathcal{R}}, \tilde{\mathcal{T}}, \tilde{t}_{\max} \rangle, \text{ where}$$

- $i \in \{0, 1\}$ and $\iota = t \mod 2$ is the acting player at time step t.
- $\tilde{\mathbb{A}} = \bigcup_{h_{\text{pub}} \in \mathbb{H}_{\text{pub}}} \prod_{h_\iota \in \mathbb{H}_\iota(h_{\text{pub}})} \Delta(\mathbb{A}(h_\iota))$ is the set of *decision rules*, where $\mathbb{H}_\iota(h_{\text{pub}})$ denotes the set of information states for the acting player consistent with public history h_{pub}.
- $\tilde{\mathbb{S}} = \bigsqcup_{h_{\text{pub}} \in \mathbb{H}_{\text{pub}}} \Delta(\mathbb{H}(h_{\text{pub}}))$ is the set of public belief states (PBSs). The public-history label is part of the state, so this is a disjoint union of simplexes.
- $\tilde{\mathcal{R}}(\tilde{s}, \tilde{a}) = \mathbb{E}_{H \sim \tilde{s}} \mathbb{E}_{A \sim \tilde{a}(H_\iota)} \mathcal{R}(H, A)$, where we use capital letters to denote random variables.
- $\tilde{\mathcal{T}}$ is the PBS transition law obtained from $\mathcal{T}$ by the Bayesian update below.

- $\tilde{t}_{\max} = t_{\max}$.

The initial PBS is $\tilde{s}^0 = \mu$. A PuB-AMG policy deterministically selects a decision rule $\tilde{a} = \tilde{\pi}_\iota(\tilde{s})$; the underlying action is then sampled from $\tilde{a}(H_\iota)$. We also write $\tilde{\pi}(\tilde{s})$ when the acting component is understood. For a next public observation o and $h' \in \mathbb{H}(h_{\text{pub}}o)$, define

$$m_o(h') = \sum_{h,a} \tilde{s}(h)\tilde{a}(h_\iota, a)\mathcal{T}(h' \mid h, a), \qquad p_o = \sum_{h'} m_o(h'), \qquad \tilde{s}_o(h') = \frac{m_o(h')}{p_o} \quad (p_o > 0).$$

The sums range over consistent histories and legal actions, and observation o occurs with probability p_o. If $p_o = 0$, choose its posterior arbitrarily; all probability-weighted contributions from that branch are zero. Thus $\tilde{\mathcal{T}}$ assigns probability p_o to the labeled successor PBS $\tilde{s}_o$.

The Correspondence Mapping There is a natural correspondence between joint policies in the PuB-AMG and joint policies in the adversarial finite-horizon sequential game that generated that PuB-AMG, which maps each full-support PuB-AMG joint policy $\tilde{\pi}$ to the joint policy π that plays the same distribution over actions at each information state as $\tilde{\pi}$ under the PBS associated to $\tilde{\pi}$. This correspondence is called the correspondence mapping and is defined as follows.

Definition 2.1 (Correspondence Mapping)**.** *Let $\tilde{\pi}$ be a joint PuB-AMG policy. Then the correspondence mapping is defined as $\Pi^{\downarrow}\colon \{\tilde{\pi} \mid \tilde{\pi}$ induces full support over information states.$\} \to \{\pi\}$ where, for any h_ι^t, the output policy π maps $h_\iota^t \mapsto \tilde{\pi}(\tilde{s})(h_\iota^t)$, and where $\tilde{s}$ is the PBS induced by reaching the public history h_{pub} associated with h_ι^t using $\tilde{\pi}$.*

2.4 Unsafe Subgame Solving and the Non-Correspondence Problem

A desirable property of the PuB-AMG is that it is fully observable, which makes performing search conceptually straightforward. Thus, it would be convenient to be able to perform search in the original game through the PuB-AMG in a manner analogous to PBS search approaches for fully cooperative games. However, agents in adversarial settings, unlike those in cooperative settings, have no reason to disclose their policy computation procedures—as a result, they cannot generally compute the actual PBS.

A naive workaround to the inaccessibility of the opponent's policy computation procedure is to substitute the agent's own policy computation procedure for that of the opponent for the purpose of computing the PBS. Performing search starting from a PBS computed in this way is known as *unsafe subgame solving*. For unsafe subgame solving to be sound, it would need to be the case that good PuB-AMG joint policies correspond to good policies in the underlying game. As the name of this approach suggests, such a correspondence does not generally hold, even at Nash equilibria.

Proposition 2.1 (Sokota et al. [2023b])**.** *A PuB-AMG Nash equilibrium $\tilde{\pi}$ may correspond with a joint policy $\Pi^{\downarrow}(\tilde{\pi})$ that is maximally exploitable.*

This issue is known in the literature as the non-correspondence problem.

2.5 Recovery of Equilibrium Correspondence

An important result is that equilibrium correspondence is recovered under a certain class of objectives that guarantee unique equilibria.

Definition 2.2 (Uniqueness-Guaranteeing Objective)**.** *Consider an objective of the following form*

$$\mathfrak{J}\colon \pi_0, \pi_1 \mapsto \mathbb{E}_{\pi_0,\pi_1}\left[\sum_t \mathfrak{R}(H^t, A^t, \pi(H_\iota^t))\right]$$

where $\mathfrak{R}$ is a real-valued function. Every such objective $\mathfrak{J}$ possesses a PuB-AMG analog $\tilde{\mathfrak{J}}$ that is equivalent to $\mathfrak{J}$ in the sense that $\tilde{\mathfrak{J}}(\tilde{\pi}) = \mathfrak{J}(\Pi^{\downarrow}(\tilde{\pi}))$ that is defined by

$$\tilde{\mathfrak{J}}\colon \tilde{\pi}_0, \tilde{\pi}_1 \mapsto \mathbb{E}_{\tilde{\pi}_0,\tilde{\pi}_1}\left[\sum_t \tilde{\mathfrak{R}}(\tilde{S}^t, \tilde{A}^t)\right], \text{ where } \tilde{\mathfrak{R}}\colon (\tilde{s}, \tilde{a}) \mapsto \mathbb{E}_{H\sim\tilde{s}}\mathbb{E}_{A\sim\tilde{a}(H_\iota)}\mathfrak{R}(H, A, \tilde{a}(H_\iota)).$$

An objective $\mathfrak{J}$ is uniqueness guaranteeing *(UG) if every original-game subgame has an equilibrium that is unique on its reachable information states. A subgame begins at a distribution in $\Delta(\mathbb{H}(h_{pub}))$ over histories consistent with a public state.*

Theorem 2.3 (Sokota et al. [2023b])**.** *If $\tilde{\pi}$ is an equilibrium of a PuB-AMG under a UG objective, then its corresponding joint policy $\Pi^{\downarrow}(\tilde{\pi})$ is the equilibrium of the original game under the same UG objective.*

Theorem 2.3 offers hope that an unsafe subgame solving-like approach could be made effective under a uniqueness-guaranteeing objective.

3 Convergence of Exploitability

The results in existing literature on the relationship between PuB-AMG exploitability and exploitability in the underlying game pertain only to equilibria. As a grounding for practical algorithms, such results have limited applicability, as it is not generally feasible to compute equilibria exactly. To put practical regularization-based PBS search approaches on more solid footing, we show that solutions that approach zero exploitability in the PuB-AMG under UG objectives correspond to solutions that approach zero exploitability in the underlying game under the same objectives.

Toward doing so, we derive the following local continuity property of the correspondence mapping $\Pi^{\downarrow}$ in Appendix B.

Proposition 3.1 (Local Continuity)**.** *Let $\{\tilde{\pi}_n\}_n$ be continuous PuB-AMG joint policies converging uniformly to a continuous PuB-AMG joint policy $\tilde{\pi}^*$. If $\tilde{\pi}^*$ induces full support over information states, then $\Pi^{\downarrow}(\tilde{\pi}_n)$ is defined for all sufficiently large n and*

$$\Pi^{\downarrow}(\tilde{\pi}_n) \to \Pi^{\downarrow}(\tilde{\pi}^*).$$

Proof sketch. The proof tracks public-history reach probabilities and conditional PBSs by induction over public histories. Uniform convergence of the PuB-AMG policies and continuity of $\tilde{\pi}^*$ imply convergence of the unnormalized next-history probabilities. The full-support condition implies positive reach for every feasible decision public history, which keeps the relevant Bayes denominators bounded away from zero, so the induced PBSs converge. The same reach-probability convergence shows that the correspondence mapping is defined for all sufficiently large n. Evaluating the uniformly convergent policies at these convergent PBSs gives convergence of the corresponding original-game policies. □

Proposition 3.1, together with Theorem 2.3, shows that approaching a full-support PuB-AMG equilibrium (in the uniform policy topology) induces joint policies that approach the original-game equilibrium. We now state the corresponding exploitability guarantee, which is proved in Appendix C.

Theorem 3.1 (Exploitability Convergence)**.** *Let the UG objective be induced by a continuous $\mathfrak{R}$. Let $\{\tilde{\pi}_n\}_n$ be a sequence of continuous PuB-AMG joint policies converging uniformly to a continuous PuB-AMG joint policy $\tilde{\pi}^*$. Assume that $\tilde{\pi}^*$ induces full support over information states. Then $\Pi^{\downarrow}(\tilde{\pi}_n)$ is defined for all sufficiently large n, and*

$$\{\mathit{expl}_{\mathrm{pub}}(\tilde{\pi}_n)\}_n \to 0$$

implies that

$$\{\mathit{expl}_{\mathrm{orig}}(\Pi^{\downarrow}(\tilde{\pi}_n))\}_n \to 0.$$

Proof sketch. PuB-AMG exploitability is continuous at continuous policies, so the limiting policy $\tilde{\pi}^*$ has zero PuB-AMG exploitability and is therefore a PuB-AMG equilibrium. By Theorem 2.3, $\Pi^{\downarrow}(\tilde{\pi}^*)$ is the original-game equilibrium. The full-support assumption makes the correspondence mapping continuous at $\tilde{\pi}^*$, so $\Pi^{\downarrow}(\tilde{\pi}_n) \to \Pi^{\downarrow}(\tilde{\pi}^*)$. Continuity of original-game exploitability gives the claim. □

Additionally, under the curvature and regularity assumptions described in Appendix C.1, we establish a square-root exploitability-transfer bound from the PuB-AMG to the original game.

Theorem 3.2 (Quantitative Exploitability Transfer)**.** *Under the assumptions on $\mathfrak{R}$, $\tilde{\pi}$, and $\tilde{\pi}^*$ in Appendix C.1, there exists a constant C independent of $\tilde{\pi}$ such that*

$$\mathit{expl}_{\mathrm{orig}}\big(\Pi^{\downarrow}(\tilde{\pi})\big) \le C\sqrt{\mathit{expl}_{\mathrm{pub}}(\tilde{\pi})}.$$

Proof sketch. The curvature assumption first relates PuB-AMG policy error to the square root of PuB-AMG exploitability. Controlling changes in the induced beliefs transfers this bound to the corresponding original-game policy. Finally, payoff regularity converts the resulting policy error into a bound on original-game exploitability. □

Theorem 3.2 also yields a finite-error bound for an idealized model of search with an approximate value function. Partition the horizon into $N = \lceil t_{\max}/d \rceil$ blocks of d steps. At the PBS $\tilde{s}$ starting each block, let $\tilde{\pi}$ play a solution of the depth-d PuB-AMG subgame from $\tilde{s}$ whose truncated leaves are scored by a value function $\hat{V}$.

Corollary 3.3 (Search with approximate values)**.** *Suppose $|\hat{V}(\tilde{s}) - \tilde{V}^*(\tilde{s})| \leq \epsilon_V$ at every block-boundary PBS, where $\tilde{V}^*$ is the equilibrium value, and that each subgame solution lets neither player gain more than ϵ_K over the value of its subgame by deviating. Then $\mathit{expl}_{\mathrm{pub}}(\tilde{\pi}) \leq N(\epsilon_K + 2\epsilon_V)$, and if $\tilde{\pi}$ satisfies the assumptions of Theorem 3.2,*

$$\mathit{expl}_{\mathrm{orig}}\big(\Pi^{\downarrow}(\tilde{\pi})\big) \leq C\sqrt{N(\epsilon_K + 2\epsilon_V)}.$$

Proof sketch. The PuB-AMG is a perfect-information game, so an induction over blocks applies: changing the leaf values of a subgame by at most ϵ_V changes its value by at most ϵ_V, and each block adds at most $\epsilon_K + 2\epsilon_V$ to either player's best-response gain. The full proof is in Appendix C.2. □

Regularization bias Both theorems measure exploitability under the regularized objective. For Gravel's minimax-KL objective with regularization strength α, removing the regularizer changes every payoff by at most $b = \alpha\, t_{\max} \log A_{\max}$, where $A_{\max}$ is the largest number of legal actions at an information state, so

$$\mathrm{expl}^{\mathrm{unreg}}_{\mathrm{orig}}(\pi) \leq \mathrm{expl}^{\mathrm{reg}}_{\mathrm{orig}}(\pi) + \alpha\, t_{\max} \log A_{\max}.$$

This bias vanishes as $\alpha \to 0$, but solving the regularized game takes more search iterations at smaller α (Figure 1), which is why we anneal α during training.

4 Gravel

Inspired by Theorem 3.1, we propose Gravel, a regularization-based expert-iteration algorithm for imperfect-information games. Gravel performs decision-time search in the PuB-AMG but replaces the unregularized game objective with a uniqueness-guaranteeing minimax-KL objective [Perolat et al., 2021]. For a history h, action a, and local decision rule δ, define

$$\mathfrak{R}(h, a, \delta) = \mathcal{R}(h, a) - (-1)^{\mathbb{I}[\iota(h)=1]}\alpha\, \mathrm{KL}(\delta \,\|\, u_{h_\iota}),$$

where u_{h_ι} is the uniform distribution over legal actions at information state h_ι. The corresponding objective is

$$\mathfrak{J}(\pi_0, \pi_1) = \mathbb{E}_{\pi_0,\pi_1}\left[\sum_t \mathfrak{R}\left(H^t, A^t, \pi_{\iota(H^t)}(\cdot \mid H^t_\iota)\right)\right].$$

For player 0 decision states the KL term is subtracted from player 0's payoff; for player 1 decision states it is added to player 0's payoff, thereby always acting as a penalty for the optimizing player. This objective leaves the game zero-sum while ensuring each player's local optimization problem is strongly regularized. As a result, the PuB-AMG equilibrium is unique in the sense needed by Definition 2.2, and the approximate-search guarantee in Theorem 3.1 applies as search approaches the regularized PuB-AMG equilibrium.

At a decision point with public history h_{pub}, Gravel constructs the PBS $\tilde{s}$ induced by the observed public trajectory under the current Gravel policy for both players. This is the same belief used by unsafe subgame solving; the difference is that Gravel solves the resulting PBS subgame under $\mathfrak{J}$ rather than under the original unregularized payoff. The root of the searched subgame is a chance node that samples an underlying history from $\tilde{s}$. The subgame then follows the original transition and reward dynamics for d steps, or until a true terminal state is reached. At depth-d truncated states, Gravel substitutes a learned PBS value function $v(\tilde{s}', h')$ for the remaining continuation value. Thus,

a search call is a PBS-rooted, value-augmented search problem: its interior is the original extensive-form dynamics with rewards regularized by $\mathfrak{R}$, while its truncated leaves are backed up from learned PBS continuation values rather than from payoffs defined inside the finite tree itself.

Gravel solves each depth-limited subgame with magnetic mirror descent (MMD) [Sokota et al., 2023a]. The search policy is initialized from the policy network. Given the current search policy π_k, Gravel computes action values for every decision point in the subgame by backward induction, using v at truncated leaves and including the regularization terms from descendant decision points. For a player 0 information state h_0, the next iterate is

$$\pi_{k+1}(\cdot \mid h_0) = \underset{\pi \in \Delta(\mathbb{A}(h_0))}{\arg\max} \left\{ \mathbb{E}_{A \sim \pi} q_k(h_0, A) - \alpha \,\mathrm{KL}(\pi \,\|\, u_{h_0}) - \beta \,\mathrm{KL}(\pi \,\|\, \pi_k(\cdot \mid h_0)) \right\},$$

where q_k measures player 0's reward and excludes the current decision's KL term. Let t_r be the depth of the root PBS $\tilde{s}$ and $T = \min(t_r + d,\, t_{\max})$ the leaf depth of the searched subgame; then

$$q_k(h_\iota^{t_0}, a) = \mathbb{E}_{\pi_k}\Big[\mathcal{R}(H^{t_0}, a) + \sum_{t=t_0+1}^{T-1} \mathfrak{R}\left(H^t, A^t, \pi_k(\cdot \mid H_\iota^t)\right) + v(\tilde{S}^T, H^T) \;\Big|\; H_\iota^{t_0} = h_\iota^{t_0}, A^{t_0} = a\Big],$$

with $v \equiv 0$ at terminal histories, so a subgame that ends before depth T contributes only its rewards. The update has the closed form

$$\pi_{k+1}(a \mid h_0) \propto u_{h_0}(a)^{\alpha/\tau} \pi_k(a \mid h_0)^{\beta/\tau} \exp(q_k(h_0, a)/\tau), \quad \tau = \alpha + \beta,$$

which is smooth in both the action values and the previous iterate. Player 1 instead minimizes the expected q_k plus both KL terms; its closed form uses $\exp(-q_k(h_1, a)/\tau)$. After K iterations, Gravel outputs the behavior policy induced by the running sequence-form average [Von Stengel, 1996] of the iterates, denoted $\bar{\pi}$. We average in sequence form rather than independently averaging local action probabilities so that the output corresponds to a single coherent extensive-form policy.

Training follows the expert-iteration template [Anthony et al., 2017], alternating between search and supervised learning. During self-play, Gravel samples root PBSs from the current policy, runs the search procedure above, and stores two kinds of examples in replay buffers. For the policy network, the input is a pair $(\tilde{s}, h_\iota)$ and the target is the policy found by search $\bar{\pi}(\cdot \mid h_\iota)$ at that decision point. We train the policy network by minimizing the forward KL from the search target to the network prediction. For the value network, the input is again $(\tilde{s}, h_\iota)$ and the target is the regularized player 0 value of the averaged searched policy. At player 0 states this is

$$\mathbb{E}_{A \sim \bar{\pi}(\cdot \mid h_0)}\left[q_{\bar{\pi}}(h_0, A)\right] - \alpha \,\mathrm{KL}(\bar{\pi}(\cdot \mid h_0) \,\|\, u_{h_0}),$$

and at player 1 states the KL term is added. We fit v with a mean-squared error loss.

The training sampler must expose the networks to the PBSs that will later be queried by search. Accordingly, after solving a subgame, Gravel samples successor PBSs from both the final averaged policy $\bar{\pi}$ and randomly selected intermediate MMD iterates. At test time, Gravel samples actions from $\bar{\pi}$. Because safety comes from the regularized objective rather than from a gadget-game boundary condition, Gravel can re-solve at any later decision point; when the next decision already lies inside an untruncated searched tree, it can simply continue using the policy computed by the previous search. The fixed-temperature controlled comparisons instead use a mean-squared-error policy loss.

5 Experiments

Architecture and optimization Both of our policy and value networks are MLPs with GELU [Hendrycks and Gimpel, 2016] activation functions. We use concatenated activations [Shang et al., 2016, Abbas et al., 2023] as we found residual connections harmful. We avoid layer normalization [Ba et al., 2016], as it can increase the Lipschitz constant of the network. Given a PBS, the value network predicts the value conditioned on the hand of each player in a single forward pass. At any information state, the policy network takes as input the root PBS, the public history from the root PBS, and a description of the valid actions and outputs a distribution over all possible actions for all compatible information-states. We use the Polar Express variant of Muon [Amsel et al., 2026] for updating the intermediate weight matrices and Adam [Kingma and Ba, 2015] for other parameters of the MLPs. We train the value network with the MSE loss and the policy network with the forward KL loss.

Table 1: Unregularized exploitability (lower is better), in native payoff units for LD2D3F and mbb/h for TEH. Values are means with 95% t-confidence intervals over three seeds. K denotes search iterations. ReBeL is our own implementation; "Muon" trains ReBeL with Muon instead of Adam for intermediate weight matrices.

Solver		LD2D3F		TEH	
		K	Exploitability	K	Exploitability
Gravel (MD)		48	0.015 ± 0.006	80	66 ± 60
FP	Learned initial policy	48	0.098 ± 0.013	80	889 ± 89
	Uniform initial policy	48	0.072 ± 0.018	80	552 ± 27
ReBeL	Random iterate, own setting	512	0.105 ± 0.027	512	240 ± 105
	Average iterate, own setting	512	0.389 ± 0.138	512	391 ± 162
	Random iterate, Gravel's K	48	0.150 ± 0.173	80	705 ± 136
	Muon, own K	512	0.153 ± 0.176	512	180 ± 89
	Muon, Gravel's K	48	0.187 ± 0.162	80	880 ± 673

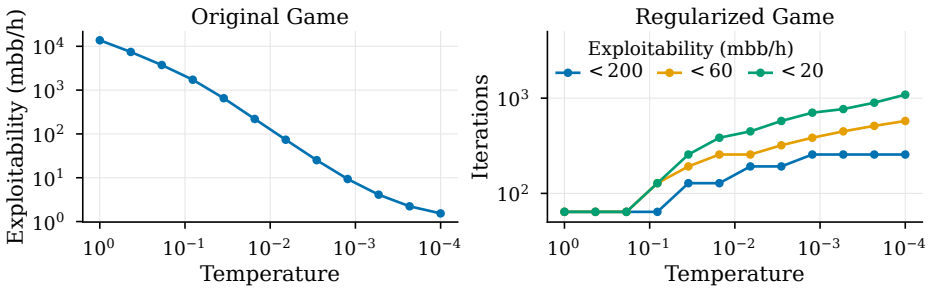


Figure 1: Left: unregularized exploitability of the regularized TEH equilibrium as a function of temperature. Right: search iterations until the regularized exploitability first falls below 200, 60 and 20 mbb/h at each temperature; lower temperatures need more iterations.

Benchmarks We evaluate Gravel on the two-dice, three-face Liar's Dice (LD2D3F) and heads-up no-limit Texas hold'em poker (HUNL) and its turn-endgame hold'em (TEH) variant where both players automatically check or call on the first two streets. We handle "off-tree" actions by including a random raise size in addition to actions for fold, check, call, all-in, and other raises: raise by half pot and pot as the first raise of the street, and raise by 0.75 pot otherwise. Whenever we observe an off-tree action, we modify the random raise and resolve the subgame.

Controlled comparisons We compare Gravel (MD), Gravel with fictitious play (FP), and ReBeL on LD2D3F and TEH as they are small enough to calculate exact exploitability. Table 1 reports unregularized exploitability over three seeds, with ReBeL both at its own best setting and at Gravel's iteration count. There are a few things worth noting: first, the correspondence problem is visible as the exploitability of ReBeL goes up when it plays average iterate. Second, Gravel needs drastically fewer iterations than ReBeL as using the iteration count of Gravel with ReBeL drastically increases the exploitability. Lastly, the FP variant of Gravel is also not able to leverage the policy network as initializing the search procedure with a uniform policy boosts performance.

The cost of regularization As shown in Figure 1, finding a high-quality solution in the unregularized TEH game requires that the amount of regularization not be too large, but solving the game at a lower regularization level requires more search iterations. Since we train with few iterations, we are not able to learn a strong policy when starting from an initially low temperature. As a consequence, we anneal the temperature while increasing the number of search iterations.

Searching more at test time Increasing the number of search iteration at test time is a distribution shift and is considered harmful, but Figure 3 shows a more nuanced picture: from epoch 200

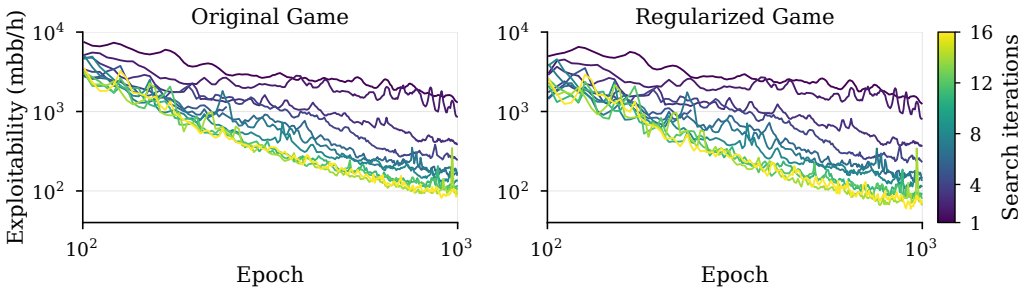


Figure 2: Exploitability, in mbb/h, for different numbers of search iterations. Left original game, right regularized game.

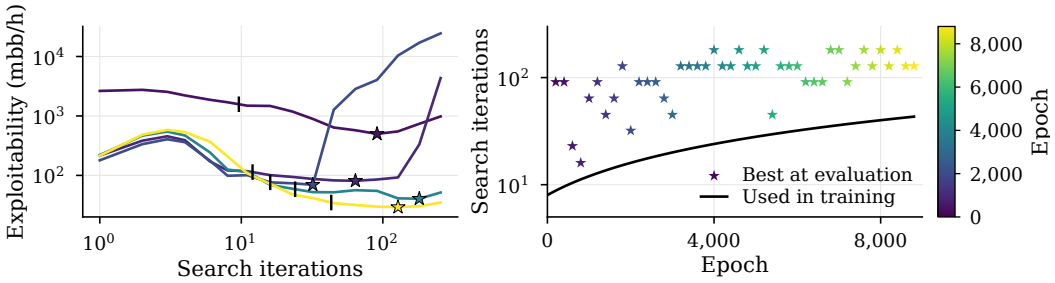


Figure 3: TEH search budget. Left: exploitability at selected checkpoints, with stars marking the best number of search iterations and vertical ticks the number used in training. Right: the best number of search iterations at each evaluated checkpoint from epoch 200 onward, compared with the number used in training.

onward, the best number of search iterations exceeds the number used at training time, but can start deteriorating the solution.

For HUNL, we scale up the training by taking more gradient steps per epoch. We compare Gravel against Slumbot [Jackson, 2013]. We play 1.5M hands against Slumbot and report our variance-reduced AIVAT [Burch et al., 2018] score with a 95% confidence interval in Table 2, alongside other scores found in the literature. We acknowledge that Table 2 is an uncontrolled, historical comparison of head-to-head playing strength against a common opponent. Gravel was trained for nine days on one node with eight H200 GPUs, whereas ReBeL reports 90 machines with eight V100 GPUs each for data generation.

Table 2: Performance against Slumbot, measured in mbb/h, compared to other self-play PBS search methods. This is an uncontrolled historical comparison of playing strength, not of exploitability; ReBeL and Student of Games values are as reported in their papers, under different compute budgets, abstractions, and error-bar conventions. For Gravel, values are AIVAT-adjusted [Burch et al., 2018] average performance over 1.5 million hands. We estimate our 95% confidence intervals from asymptotic normality under the CLT.

Gravel	ReBeL [Brown et al., 2020]	Student of Games [Schmid et al., 2023]
60 ± 2	45 ± 5	7 ± 3

6 Related Work

The technology for decomposing imperfect-information games into independently analyzable subgames was first conceived by Burch et al. [2014]. Burch et al. [2014] proposed and proved the soundness of the CFR-D algorithm, which leverages a construct that would later be coined a *gadget*

game. Gadget games were the basis for DeepStack [Moravčík et al., 2017], Libratus [Brown and Sandholm, 2018], Supremus [Zarick et al., 2020] and Pluribus [Brown and Sandholm, 2019].

However, the manner in which these approaches handle their value functions (or lack thereof) leaves things to be desired. Libratus does not have a value function at all; as a result, it can only perform search when the game is sufficiently close to the end. DeepStack and Supremus have value functions, but they are trained offline using randomly sampled PBSs, which can result in large distribution shifts at test time. Finally, Pluribus trains a tabular value function using poker-specific abstractions, making it difficult to generalize to other settings.

These shortcomings were addressed by both ReBeL [Brown et al., 2020] and Student of Games [Schmid et al., 2023], which learn their value functions via self-play. Like the previously discussed related works, Schmid et al. [2023] achieve safety using gadget games. In contrast, Brown et al. [2020] used a new approach based on simulating the CFR-D algorithm.

In the sense of performing PBS search with a self-play-learned value function, ReBeL and Student of Games are the most closely related works to our own. However, our approach differs substantially at a conceptual level. Rather than relying on gadget games or CFR-D simulation for safety, our approach has the straightforward interpretation of solving a regularized perfect-information game. This results in different properties, such as continuous and unique equilibria [Sokota et al., 2023b], compared to gadget games and CFR-D simulation.

7 Conclusion and Future Work

We proposed an approach for imperfect-information search that overcomes the non-correspondence problem by using regularization. In doing so, our approach overcomes the typical challenges encountered in gadget-game-based and ReBeL-based approaches, and leads to a significantly simpler and theoretically grounded method. The price of this conceptual simplification is that, in principle, regularization might slightly shift the equilibrium of the game. We empirically demonstrate that this concern does not translate into degraded performance. In fact, we find that the opposite is true: not only is our method substantially simpler, but also it scores 30% higher against Slumbot than the reported score of its closest comparison point, ReBeL, all while requiring only 1/10th the number of training samples and 1/32nd of the search budget.

Broader Impacts

This work provides a significant methodological advance in solving large partially observable stochastic games and can immediately be applied to better understand optimal strategies in games like poker, as well as potentially extended to other impactful real-world scenarios well-modeled by these games. Like any advance, these methods could potentially be misused (such as for cheating) and require responsible usage, though less advanced methods exist for solving similar zero-sum settings.

Acknowledgment

SM is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) PSG-D Fellowship 599271-2025. GF is supported in part by NSF Award CCF-2443068 and IIS-2552046, ONR grant N00014-25-1-2296, and an AI2050 Early Career Fellowship.

References

Zaheer Abbas, Rosie Zhao, Joseph Modayil, Adam White, and Marlos C Machado. Loss of plasticity in continual deep reinforcement learning. In *Conference on lifelong learning agents*, pages 620–636. PMLR, 2023.

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained policy optimization. In *International Conference on Machine Learning (ICML)*, pages 22–31, 2017. URL `https://proceedings.mlr.press/v70/achiam17a.html`.

Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. In *International Conference on Learning Representations (ICLR)*, 2026.

Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. In *Neural Information Processing Systems (NIPS)*, 2017.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Noam Brown and Tuomas Sandholm. Safe and nested subgame solving for imperfect-information games. In *Neural Information Processing Systems (NIPS)*, 2017.

Noam Brown and Tuomas Sandholm. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 2018.

Noam Brown and Tuomas Sandholm. Superhuman ai for multiplayer poker. *Science*, 2019.

Noam Brown, Tuomas Sandholm, and Brandon Amos. Depth-limited solving for imperfect-information games. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong. Combining deep reinforcement learning and search for imperfect-information games. *Neural Information Processing Systems (NeurIPS)*, 2020.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3–4):231–357, 2015. doi: 10.1561/2200000050.

Neil Burch, Michael Johanson, and Michael Bowling. Solving imperfect information games using decomposition. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2014.

Neil Burch, Martin Schmid, Matej Moravcik, Dustin Morrill, and Michael Bowling. Aivat: A new variance reduction technique for agent evaluation in imperfect information games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Jilles Steeve Dibangoye, Christopher Amato, Olivier Buffet, and François Charpillet. Optimally solving dec-pomdps as continuous-state mdps. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

Arnaud Fickinger, Hengyuan Hu, Brandon Amos, Stuart Russell, and Noam Brown. Scalable online planning via reinforcement learning fine-tuning. In *Neural Information Processing Systems (NeurIPS)*, 2021.

Sam Ganzfried and Tuomas Sandholm. Endgame solving in large imperfect-information games. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2015.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Samid Hoda, Andrew Gilpin, Javier Peña, and Tuomas Sandholm. Smoothing techniques for computing nash equilibria of sequential games. *Mathematics of Operations Research*, 35(2):494–512, 2010.

Eric Jackson. Slumbot nl: Solving large games with counterfactual regret minimization using sampling and distributed processing. In *AAAI Workshop on Computer Poker and Incomplete Information*, 2013.

Sune K. Jakobsen, Troels B. Sørensen, and Vincent Conitzer. Timeability of extensive-form games. In *ACM Conference on Innovations in Theoretical Computer Science*, 2016.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 795–811, 2016. doi: 10.1007/978-3-319-46128-1_50. URL https://arxiv.org/abs/1608.04636.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Adam Lerer, Hengyuan Hu, Jakob N. Foerster, and Noam Brown. Improving policies via search in cooperative partially observable games. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

Matej Moravcik, Martin Schmid, Karel Ha, Milan Hladik, and Stephen Gaukrodger. Refining subgames in large imperfect information games. *AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 2017.

Ashutosh Nayyar and Abhishek Gupta. Information structures and values in zero-sum stochastic games. In *American Control Conference (ACC)*, 2017.

Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 2013.

Frans A. Oliehoek. Sufficient plan-time statistics for decentralized pomdps. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2013.

Julien Perolat, Remi Munos, Jean-Baptiste Lespiau, Shayegan Omidshafiei, Mark Rowland, Pedro Ortega, Neil Burch, Thomas Anthony, David Balduzzi, Bart De Vylder, Georgios Piliouras, Marc Lanctot, and Karl Tuyls. From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In *International Conference on Machine Learning (ICML)*, 2021.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 627–635, 2011. URL `https://proceedings.mlr.press/v15/ross11a.html`.

Martin Schmid, Matej Moravčík, Neil Burch, Rudolf Kadlec, Josh Davidson, Kevin Waugh, Nolan Bard, Finbarr Timbers, Marc Lanctot, G. Zacharias Holland, Elnaz Davoodi, Alden Christianson, and Michael Bowling. Student of games: A unified learning algorithm for both perfect and imperfect information games. *Science Advances*, 2023.

Wenling Shang, Kihyuk Sohn, Diogo Almeida, and Honglak Lee. Understanding and improving convolutional neural networks via concatenated rectified linear units. In *international conference on machine learning*, pages 2217–2225. PMLR, 2016.

Samuel Sokota, Edward Lockhart, Finbarr Timbers, Elnaz Davoodi, Ryan D'Orazio, Neil Burch, Martin Schmid, Michael Bowling, and Marc Lanctot. Solving common-payoff games with approximate policy iteration. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Samuel Sokota, Hengyuan Hu, David J Wu, J Zico Kolter, Jakob Nicolaus Foerster, and Noam Brown. A fine-tuning approach to belief state modeling. In *International Conference on Learning Representations (ICLR)*, 2022.

Samuel Sokota, Ryan D'Orazio, J Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *International Conference on Learning Representations (ICLR)*, 2023a.

Samuel Sokota, Ryan D'Orazio, Chun Kai Ling, David J Wu, J Zico Kolter, and Noam Brown. Abstracting imperfect information away from two-player zero-sum games. In *International Conference on Machine Learning (ICML)*, 2023b.

Bernhard Von Stengel. Efficient computation of behavior strategies. *Games and Economic Behavior*, 14(2):220–246, 1996.

Auke J. Wiggers, Frans A. Oliehoek, and Diederik M. Roijers. Structure in the value function of two-player zero-sum games of incomplete information. In *European Conference on Artificial Intelligence (ECAI)*, 2016.

Ryan Zarick, Bryan Pellegrino, Noam Brown, and Caleb Banister. Unlocking the potential of deep counterfactual value networks. *arXiv preprint arXiv: 2007.10442*, 2020.

A Conventions and Standard Results

Time, legal actions, and policies. Write $-i = 1 - i$. At time $0 \le t < t_{\max}$, the history H^t records the actions $A^0, \dots, A^{t-1}$; the action A^t and the next transition produce H^{t+1}, and $H^{t_{\max}}$ is terminal. We write $\mathrm{depth}(h_{\text{pub}}) = t$ for the number of completed transitions, so the root has depth zero even when there is an initial public observation. Each transition adds a public observation, which may be null, and $\iota(h) = \mathrm{depth}(h_{\text{pub}}) \bmod 2$. Early termination can be padded with deterministic zero-reward steps; these steps have zero policy error and are excluded from support and reach-ratio conditions below. Only histories feasible from μ under some policy are included. For an information state h_i, let $\mathbb{A}(h_i)$ be its legal actions and put $A_{\max} = \max_{i,h_i} |\mathbb{A}(h_i)|$. Here $\Delta(X)$ denotes the probability simplex on a finite set X. A behavioral policy $\pi_i(h_i) = \pi_i(\cdot \mid h_i)$ belongs to $\Delta(\mathbb{A}(h_i))$, and $\pi = (\pi_0, \pi_1)$ is a joint policy. We write $\pi_i(h_i, a) = \pi_i(a \mid h_i)$ for an action probability, and use the same coordinate notation for decision rules. Mixed profiles such as (π_i, π'_{-i}) always place components in player order.

Reach probabilities. For a joint policy x in either game, let $z_x(h_{\text{pub}})$ be its public-history reach probability and $\tilde{s}_{x,h_{\text{pub}}}$ its conditional PBS. For h_i at public history h_{pub}, write $\mathbb{H}(h_{\text{pub}}, h_i)$ for the histories consistent with both and define

$$z_x(h_i) = z_x(h_{\text{pub}}) \sum_{h \in \mathbb{H}(h_{\text{pub}}, h_i)} \tilde{s}_{x,h_{\text{pub}}}(h).$$

Full information-state support means $z_x(h_i) > 0$ for every feasible decision information state. A posterior at zero reach may be arbitrary. For $\pi = \Pi^{\downarrow}(\tilde{\pi})$, we have $z_\pi = z_{\tilde{\pi}}$ and the same conditional PBSs wherever they are reached. We use z_n and $\tilde{s}_{n,h_{\text{pub}}}$ for policies indexed by n, and a subscript $*$ for the reference policy. The marginal of a PBS on player i's information state is $\tilde{s}_i(h_i) = \sum_{h \in \mathbb{H}(h_{\text{pub}}, h_i)} \tilde{s}(h)$.

Exploitability. For the objective $\mathfrak{J}$ and its public analog $\tilde{\mathfrak{J}}$, define

$$\mathrm{expl}_{\text{orig}}(\pi) = \frac{1}{2}\left[\sup_{\pi'_0} \mathfrak{J}(\pi'_0, \pi_1) - \inf_{\pi'_1} \mathfrak{J}(\pi_0, \pi'_1)\right],$$

$$\mathrm{expl}_{\text{pub}}(\tilde{\pi}) = \frac{1}{2}\left[\sup_{\tilde{\pi}'_0} \tilde{\mathfrak{J}}(\tilde{\pi}'_0, \tilde{\pi}_1) - \inf_{\tilde{\pi}'_1} \tilde{\mathfrak{J}}(\tilde{\pi}_0, \tilde{\pi}'_1)\right].$$

Deviations range over all behavioral policies in the original game and rule-selecting policies in the public game. Zero exploitability is equivalent to being a saddle point. Unless marked unregularized, both exploitabilities use the same objective $\mathfrak{R}$. The notation $\mathfrak{J}_{\tilde{s}}$ and $\tilde{\mathfrak{J}}_{\tilde{s}}$ denotes the corresponding continuation objectives starting at PBS $\tilde{s}$.

Policy distances. For original-game joint policies π, π', let

$$\|\pi - \pi'\|_{\text{loc}} = \max_{i,h_i} \|\pi_i(h_i) - \pi'_i(h_i)\|_1.$$

For two decision rules at the same public history, $\|\cdot\|_{\text{loc}}$ instead takes the maximum over the acting player's information states there. An empty maximum is zero. For a local action distribution p, write

$$\mathfrak{R}(h, p) := \mathbb{E}_{A \sim p}[\mathfrak{R}(h, A, p)] = \sum_{a \in \mathbb{A}(h_\iota)} p(a)\mathfrak{R}(h, a, p).$$

Induction over time gives the same action and transition laws under $\tilde{\pi}$ and $\pi = \Pi^{\downarrow}(\tilde{\pi})$, so their payoffs agree. Without full support, assigning arbitrary policies at unreached information states still preserves the payoff of that joint profile; this does not assert that $\Pi^{\downarrow}$ acts separately on the two players' components.

We use the uniform policy topology to define uniform convergence of PuB-AMG policies.

Definition A.1 (Uniform policy topology)**.** *For PuB-AMG policies $\tilde{\pi}$ and $\tilde{\pi}'$, define*

$$d_\infty(\tilde{\pi}, \tilde{\pi}') = \sup_{\tilde{s}, h_\iota, a} \left|\tilde{\pi}(\tilde{s})(h_\iota, a) - \tilde{\pi}'(\tilde{s})(h_\iota, a)\right|,$$

where h_ι ranges over the information states in the decision rule available at $\tilde{s}$. The uniform policy topology *is the topology induced by d_∞. Thus $\tilde{\pi}_n$ converges uniformly to $\tilde{\pi}^*$ exactly when*

$$\sup_{\tilde{s},h_\iota,a} |\tilde{\pi}_n(\tilde{s})(h_\iota, a) - \tilde{\pi}^*(\tilde{s})(h_\iota, a)| \to 0.$$

Lemma A.2 (Maximum theorem)**.** *Let $\mathbb{X}$ be a metric space, let $\mathbb{Y}$ be a compact metric space, and let $f\colon \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ be continuous. Then the function*

$$g(x) = \max_{y \in \mathbb{Y}} f(x, y)$$

is continuous. The same conclusion holds with $\min$ *in place of* $\max$*.*

B Continuity Results

Proposition 3.1 (Local Continuity)**.** *Let $\{\tilde{\pi}_n\}_n$ be continuous PuB-AMG joint policies converging uniformly to a continuous PuB-AMG joint policy $\tilde{\pi}^*$. If $\tilde{\pi}^*$ induces full support over information states, then $\Pi^{\downarrow}(\tilde{\pi}_n)$ is defined for all sufficiently large n and*

$$\Pi^{\downarrow}(\tilde{\pi}_n) \to \Pi^{\downarrow}(\tilde{\pi}^*).$$

Proof. For a public history h_{pub}, let $z_n(h_{\text{pub}}) = \Pr_{\tilde{\pi}_n}(h_{\text{pub}})$ be its reach probability under $\tilde{\pi}_n$, and let $\tilde{s}_{n,h_{\text{pub}}}$ be the PBS conditional on h_{pub} whenever $z_n(h_{\text{pub}}) > 0$. We prove by induction on the depth of h_{pub} that

$$z_n(h_{\text{pub}}) \to z_*(h_{\text{pub}}) \quad \text{and} \quad \tilde{s}_{n,h_{\text{pub}}} \to \tilde{s}_{*,h_{\text{pub}}}$$

for every decision public history h_{pub} with $\mathbb{H}(h_{\text{pub}}) \neq \emptyset$, with $\tilde{s}_{n,h_{\text{pub}}}$ defined for all sufficiently large n. The claim is immediate at the root public history: its reach probability is one and its PBS is the fixed initial distribution μ.

Assume the claim holds for a public history h_{pub} and consider a one-step extension h^+_{pub} that is also a decision public history, with $\mathbb{H}(h^+_{\text{pub}}) \neq \emptyset$. For all sufficiently large n, $\tilde{s}_{n,h_{\text{pub}}}$ is defined by the induction hypothesis. For such n and any $h' \in \mathbb{H}(h^+_{\text{pub}})$, define the unnormalized probability

$$m_n(h^+_{\text{pub}}, h') = z_n(h_{\text{pub}}) \mathbb{E}_{H \sim \tilde{s}_{n,h_{\text{pub}}}} \mathbb{E}_{A \sim \tilde{\pi}_n(\tilde{s}_{n,h_{\text{pub}}})(h_\iota(H))} \left[\mathcal{T}(h' \mid H, A)\right].$$

Here H is the current underlying history conditional on h_{pub}, A is the action drawn from the decision rule chosen by $\tilde{\pi}_n$ at the current PBS, and $h_\iota(H)$ is the acting player's information state in history H. The definition is only needed for $h' \in \mathbb{H}(h^+_{\text{pub}})$; histories outside this set have zero posterior mass after observing h^+_{pub} and are not included in the normalization. Thus $m_n(h^+_{\text{pub}}, h')$ is the probability mass of first reaching h_{pub} and then transitioning in one step to the particular underlying history h'; it is unnormalized because it has not yet been divided by the total probability of the new public history h^+_{pub}.

The prefactor $z_n(h_{\text{pub}})$ and the distribution of H converge by the induction hypothesis. Moreover, for every history h and action a,

$$\tilde{\pi}_n(\tilde{s}_{n,h_{\text{pub}}})(h_\iota(h), a) \to \tilde{\pi}^*(\tilde{s}_{*,h_{\text{pub}}})(h_\iota(h), a),$$

because $\tilde{\pi}_n \to \tilde{\pi}^*$ uniformly and $\tilde{\pi}^*$ is continuous. Since the expectation is over finite sets and $\mathcal{T}$ is fixed, $m_n(h^+_{\text{pub}}, h') \to m_*(h^+_{\text{pub}}, h')$ for every such h'. Summing over histories consistent with h^+_{pub} gives $z_n(h^+_{\text{pub}}) \to z_*(h^+_{\text{pub}})$.

Since $\tilde{\pi}^*$ induces full support over decision information states, every such decision public history has positive reach probability, so $z_*(h^+_{\text{pub}}) > 0$. Thus, for all sufficiently large n, $z_n(h^+_{\text{pub}}) > 0$, and Bayes' rule gives

$$\tilde{s}_{n,h^+_{\text{pub}}}(h') = \frac{m_n(h^+_{\text{pub}}, h')}{z_n(h^+_{\text{pub}})} \to \frac{m_*(h^+_{\text{pub}}, h')}{z_*(h^+_{\text{pub}})} = \tilde{s}_{*,h^+_{\text{pub}}}(h').$$

This completes the induction.

For any original-game information state h_ι with public history h_{pub}, let $\mathbb{H}(h_{\text{pub}}, h_\iota)$ be the set of histories consistent with both h_{pub} and h_ι. The reach probability of h_ι under $\tilde{\pi}_n$ is

$$z_n(h_{\text{pub}}) \sum_{h \in \mathbb{H}(h_{\text{pub}}, h_\iota)} \tilde{s}_{n, h_{\text{pub}}}(h).$$

By the preceding induction, this converges to the corresponding reach probability under $\tilde{\pi}^*$, which is positive by full support. Since there are finitely many information states, $\Pi^{\downarrow}(\tilde{\pi}_n)$ is defined for all sufficiently large n. Discard finitely many initial terms so that it is defined for every remaining n. Now fix an original-game information state h_ι and let h_{pub} be its public history. By the definition of $\Pi^{\downarrow}$,

$$\Pi^{\downarrow}(\tilde{\pi}_n)(h_\iota) = \tilde{\pi}_n(\tilde{s}_{n, h_{\text{pub}}})(h_\iota) \quad \text{and} \quad \Pi^{\downarrow}(\tilde{\pi}^*)(h_\iota) = \tilde{\pi}^*(\tilde{s}_{*, h_{\text{pub}}})(h_\iota).$$

The first action distribution converges to the second by uniform convergence of $\tilde{\pi}_n$ and continuity of $\tilde{\pi}^*$. Since there are finitely many information states and actions, this convergence holds uniformly over original-game information states and actions. □

C Exploitability Results

Lemma C.1 (Exploitability continuity)**.** *For a UG objective induced by a continuous $\mathfrak{R}$, PuB-AMG exploitability is continuous at every continuous PuB-AMG joint policy, and original-game exploitability is continuous at every original-game joint policy.*

Proof. For the original game, the behavioral policy spaces are finite products of simplexes and hence compact. The objective $\mathfrak{J}$ is continuous in the joint policy, because it is a finite sum of products of action probabilities, transition probabilities, and the continuous reward term $\mathfrak{R}$; continuity of $\mathfrak{R}$ on its compact domain also implies boundedness. Therefore

$$\pi_1 \mapsto \max_{\rho_0} \mathfrak{J}(\rho_0, \pi_1) \quad \text{and} \quad \pi_0 \mapsto \min_{\rho_1} \mathfrak{J}(\pi_0, \rho_1)$$

are continuous by Lemma A.2. Since original-game exploitability is

$$\text{expl}_{\text{orig}}(\pi) = \frac{1}{2} \left(\max_{\rho_0} \mathfrak{J}(\rho_0, \pi_1) - \min_{\rho_1} \mathfrak{J}(\pi_0, \rho_1) \right),$$

it is continuous.

For the PuB-AMG, the Markov state space is the finite union of compact PBS simplexes. A behavioral policy is therefore a function from this PBS state space to finite-dimensional compact decision-rule sets, each a product of action simplexes. The argument below uses the compactness of the PBS simplexes and of the decision-rule set available at each PBS; it does not require the behavioral-policy function space itself to be compact. Let

$$\mathcal{B}_0(\tilde{\sigma}_1) = \max_{\tilde{\rho}_0} \tilde{\mathfrak{J}}(\tilde{\rho}_0, \tilde{\sigma}_1)$$

be player 0's best-response value against a continuous player-1 policy $\tilde{\sigma}_1$. We show that $\mathcal{B}_0$ is continuous under the uniform policy topology; the argument for player 1's best-response value is symmetric. The value $\mathcal{B}_0(\tilde{\sigma}_1)$ is obtained by finite-horizon backward induction. For fixed $\tilde{\sigma}_1$, let $b_{\tilde{\sigma}_1}(\tilde{s})$ denote player 0's optimal continuation value from PBS $\tilde{s}$ when player 1 follows $\tilde{\sigma}_1$. Thus $\mathcal{B}_0(\tilde{\sigma}_1) = b_{\tilde{\sigma}_1}(\tilde{s}^0)$. At a terminal PBS $\tilde{s}$, $b_{\tilde{\sigma}_1}(\tilde{s}) = 0$. Assume inductively that the map $(\tilde{\sigma}_1, \tilde{s}) \mapsto b_{\tilde{\sigma}_1}(\tilde{s})$ is continuous. For a nonterminal PBS $\tilde{s}$ and an available decision rule $\tilde{a}$, define the Bellman objective

$$\Phi_{\tilde{\sigma}_1}(\tilde{s}, \tilde{a}) = \tilde{\mathfrak{R}}(\tilde{s}, \tilde{a}) + \sum_{o_{\text{pub}}} p(o_{\text{pub}} \mid \tilde{s}, \tilde{a}) b_{\tilde{\sigma}_1}(\tilde{s}_{o_{\text{pub}}}(\tilde{s}, \tilde{a}))$$

where o_{pub} ranges over the finite set of next public observations.

We first check that the map $(\tilde{\sigma}_1, \tilde{s}, \tilde{a}) \mapsto \Phi_{\tilde{\sigma}_1}(\tilde{s}, \tilde{a})$ is continuous. The only point requiring care is a public observation whose probability is zero. For such an observation, the product

$$p(o_{\text{pub}} \mid \tilde{s}, \tilde{a}) b_{\tilde{\sigma}_1}(\tilde{s}_{o_{\text{pub}}}(\tilde{s}, \tilde{a}))$$

is interpreted as zero, so the value is independent of how the unreachable posterior is selected. Continuity of $\tilde{\mathfrak{R}}$ on its compact domain makes one-step rewards bounded, and the finite horizon gives a policy-independent bound on all continuation values. Therefore, when the observation probability tends to zero, this product tends to zero. When the limiting observation probability is positive, the posterior $\tilde{s}_{o_{\text{pub}}}$ is obtained by normalizing continuous unnormalized probabilities, and continuity follows from the induction hypothesis for $b_{\tilde{\sigma}_1}$. Thus each summand is continuous, and hence so is $\Phi_{\tilde{\sigma}_1}$.

At a player-0 state, the Bellman update is

$$\max_{\tilde{a}} \Phi_{\tilde{\sigma}_1}(\tilde{s}, \tilde{a}),$$

where the maximum is over the compact decision-rule set available at $\tilde{s}$. On each public-history component, this decision-rule set is fixed, so the update is continuous by Lemma A.2. At a player-1 state, the Bellman update is $\Phi_{\tilde{\sigma}_1}(\tilde{s}, \tilde{\sigma}_1(\tilde{s}))$. The evaluation map $(\tilde{\sigma}_1, \tilde{s}) \mapsto \tilde{\sigma}_1(\tilde{s})$ is continuous at continuous policies under the uniform topology: if $d_\infty(\tilde{\sigma}_1, \tilde{\sigma}_1')$ is small, then $\tilde{\sigma}_1(\tilde{s})$ and $\tilde{\sigma}_1'(\tilde{s})$ are close at every PBS $\tilde{s}$, while continuity of $\tilde{\sigma}_1$ handles small changes in $\tilde{s}$. Thus the player-1 Bellman update is continuous by composition with the joint continuity of Φ. Thus the backward-induction values are continuous by induction, and $\mathcal{B}_0$ is continuous. The best-response value for player 1 is continuous by the symmetric argument. Hence

$$\text{expl}_{\text{pub}}(\tilde{\pi}) = \frac{1}{2}\left(\max_{\tilde{\rho}_0} \tilde{\mathfrak{J}}(\tilde{\rho}_0, \tilde{\pi}_1) - \min_{\tilde{\rho}_1} \tilde{\mathfrak{J}}(\tilde{\pi}_0, \tilde{\rho}_1)\right)$$

is continuous at continuous PuB-AMG joint policies. □

Theorem 3.1 (Exploitability Convergence)**.** *Let the UG objective be induced by a continuous $\mathfrak{R}$. Let $\{\tilde{\pi}_n\}_n$ be a sequence of continuous PuB-AMG joint policies converging uniformly to a continuous PuB-AMG joint policy $\tilde{\pi}^*$. Assume that $\tilde{\pi}^*$ induces full support over information states. Then $\Pi^{\downarrow}(\tilde{\pi}_n)$ is defined for all sufficiently large n, and*

$$\{\textit{expl}_{\text{pub}}(\tilde{\pi}_n)\}_n \to 0$$

implies that

$$\{\textit{expl}_{\text{orig}}(\Pi^{\downarrow}(\tilde{\pi}_n))\}_n \to 0.$$

Proof. Write expl_{pub} and $\text{expl}_{\text{orig}}$ for exploitability in the PuB-AMG and in the original game, respectively. By Lemma C.1 and the assumption $\text{expl}_{\text{pub}}(\tilde{\pi}_n) \to 0$,

$$\text{expl}_{\text{pub}}(\tilde{\pi}^*) = \lim_{n\to\infty} \text{expl}_{\text{pub}}(\tilde{\pi}_n) = 0.$$

Thus $\tilde{\pi}^*$ is a PuB-AMG equilibrium under the UG objective. By Theorem 2.3, $\Pi^{\downarrow}(\tilde{\pi}^*)$ is the equilibrium of the original game under the same objective, so

$$\text{expl}_{\text{orig}}(\Pi^{\downarrow}(\tilde{\pi}^*)) = 0.$$

By Proposition 3.1, $\Pi^{\downarrow}(\tilde{\pi}_n)$ is defined for all sufficiently large n and, along this tail,

$$\Pi^{\downarrow}(\tilde{\pi}_n) \to \Pi^{\downarrow}(\tilde{\pi}^*).$$

Applying original-game exploitability continuity from Lemma C.1 gives

$$\lim_{n\to\infty} \text{expl}_{\text{orig}}(\Pi^{\downarrow}(\tilde{\pi}_n)) = \text{expl}_{\text{orig}}(\Pi^{\downarrow}(\tilde{\pi}^*)) = 0,$$

which is the desired conclusion. □

C.1 Quantitative Exploitability Transfer

The proof has three steps: public exploitability controls public-policy error; controlling changes in beliefs transfers this error to the original-game policy; and payoff regularity converts the resulting policy error into original-game exploitability. We give the complete argument first, then prove its supporting estimates in the same order.

Policies and notation. Let $\tilde{\pi}$ be the evaluated PuB-AMG joint policy and $\tilde{\pi}^*$ the reference equilibrium. Their corresponding original-game policies are

$$\pi = \Pi^{\downarrow}(\tilde{\pi}), \qquad \pi^* = \Pi^{\downarrow}(\tilde{\pi}^*), \qquad \epsilon = \mathrm{expl}_{\mathrm{pub}}(\tilde{\pi}).$$

Using Appendix A, write $z = z_\pi = z_{\tilde{\pi}}$ for the evaluated reach probabilities, $\tilde{s}_{h_{\mathrm{pub}}}$ for the evaluated PBSs, and $\tilde{s}_{*,h_{\mathrm{pub}}}$ for the reference PBSs. All values below use player 0's payoff: player 0 maximizes and player 1 minimizes. Sums over (i, h_i) range over decision information states, and h_{pub} denotes the public history associated with h_i.

The two policy errors. The public policies can be compared at the *same* evaluated belief:

$$e(h_i) = \left\| \tilde{\pi}_i(\tilde{s}_{h_{\mathrm{pub}}})(h_i) - \tilde{\pi}^*_i(\tilde{s}_{h_{\mathrm{pub}}})(h_i) \right\|_1 . \tag{1}$$

The corresponding original-game policies instead use their *own* induced beliefs. We distinguish their two accumulated errors by

$$\Delta = \sum_{i,h_i} z(h_i) e(h_i), \qquad \Delta_{\mathrm{orig}} = \sum_{i,h_i} z(h_i) \|\pi_i(h_i) - \pi^*_i(h_i)\|_1. \tag{2}$$

Weighting by reach probability lets the argument account for how often each information state is visited. The proof follows the chain

$$\epsilon \implies \Delta = O(\sqrt{\epsilon}) \implies \Delta_{\mathrm{orig}} = O(\Delta) \implies \mathrm{expl}_{\mathrm{orig}}(\pi) = O(\Delta_{\mathrm{orig}}).$$

C.1.1 Assumptions and equilibrium values

Objective and reference equilibrium. Assume a UG objective with $\mathfrak{R}(h, a, \cdot)$ finite and continuous on the closed legal-action simplex. Both $\tilde{\pi}$ and $\tilde{\pi}^*$ have full information-state support. The reference is *subgame-perfect*: at every PBS $\tilde{s}$, reached or unreached, all unilateral deviations $\tilde{\pi}'_0$ and $\tilde{\pi}'_1$ satisfy

$$\tilde{\mathfrak{J}}_{\tilde{s}}(\tilde{\pi}'_0, \tilde{\pi}^*_1) \le \tilde{\mathfrak{J}}_{\tilde{s}}(\tilde{\pi}^*) =: \tilde{V}^*(\tilde{s}) \le \tilde{\mathfrak{J}}_{\tilde{s}}(\tilde{\pi}^*_0, \tilde{\pi}'_1).$$

Payoff regularity. Assume

$$\begin{aligned} |\mathfrak{R}(h, a, p)| &\le R_{\max}, \\ |\mathfrak{R}(h, a, p) - \mathfrak{R}(h, a, q)| &\le L_R \|p - q\|_1. \end{aligned} \tag{3}$$

The first bound covers all legal policies, including deviations. The second is needed only between the compared original policies, with no interior restriction on their fixed opponent.

Reference-policy regularity. At every decision public history, assume

$$\left\| \tilde{\pi}^*(\tilde{s}_{h_{\mathrm{pub}}}) - \tilde{\pi}^*(\tilde{s}_{*,h_{\mathrm{pub}}}) \right\|_{\mathrm{loc}} \le L \left\| \tilde{s}_{h_{\mathrm{pub}}} - \tilde{s}_{*,h_{\mathrm{pub}}} \right\|_1 . \tag{4}$$

Only these pairs of PBSs require a common finite L; neither regularity of $\tilde{\pi}$ nor reference regularity at deviation-induced PBSs or zero-marginal boundaries is assumed.

Reach under opponent deviations. Assume a common bound $\overline{\mathcal{C}} < \infty$ on

$$\mathcal{C} = \max_{i,h_i} \sup_{\pi'_{-i}} \frac{z_{(\pi_i, \pi'_{-i})}(h_i)}{z(h_i)} \le \overline{\mathcal{C}}. \tag{5}$$

This compares how often an information state can be reached under an opponent deviation with its reach under the evaluated joint policy. A uniform reach margin $z(h_i) \ge \eta > 0$ is sufficient, since every numerator is at most one and hence $\mathcal{C} \le 1/\eta$.

Reward curvature. For Theorem 3.2, assume that the acting player's expected reward is uniformly strongly concave in their local policy. In the player-0 payoff convention, this means

$$p \longmapsto \mathfrak{R}(h, p) \quad \text{is} \quad \begin{cases} \mu_{\mathfrak{R}}\text{-strongly concave}, & \iota(h) = 0, \\ \mu_{\mathfrak{R}}\text{-strongly convex}, & \iota(h) = 1, \end{cases} \tag{6}$$

on the legal-action simplex in ℓ_1, with a common $\mu_{\mathfrak{R}} > 0$ and the usual quadratic factor $1/2$.

The value of one decision rule. To use curvature, compare a decision rule with the equilibrium rule at the same PBS. Define

$$Q^*(\tilde{s}, \tilde{a}) = \tilde{\mathfrak{R}}(\tilde{s}, \tilde{a}) + \sum_o p_o \tilde{V}^*(\tilde{s}_o),$$

where $p_o, \tilde{s}_o$ are induced by $\tilde{a}$. Thus Q^* is the value of choosing $\tilde{a}$ now and using equilibrium continuation values thereafter. Lemma C.6, proved in Appendix C.1.6, shows that strong reward curvature implies the following uniform *quadratic growth* condition: at every evaluated decision PBS $\tilde{s}$, with acting player i, every legal rule $\tilde{a}$ satisfies

$$\left|\tilde{V}^*(\tilde{s}) - Q^*(\tilde{s}, \tilde{a})\right| \geq \frac{\mu_Q}{2} \sum_{h_i} \tilde{s}_i(h_i) \|\tilde{a}(h_i) - \tilde{\pi}_i^*(\tilde{s})(h_i)\|_1^2 \tag{7}$$

with $\mu_Q = \mu_{\mathfrak{R}}$. The transfer argument only needs this weaker condition with a common $\mu_Q > 0$. It is imposed directly on the constrained decision-rule value in the displayed PBS-weighted norm, without a differentiability assumption. It is implied by the Polyak–Łojasiewicz condition and by strong convexity [Karimi et al., 2016].

All bounds are uniform over evaluated policies sharing these parameter bounds, including zero error. The transfer constant depends only on $t_{\max}, L, R_{\max}, L_R, \overline{\mathcal{C}}, \mu_Q$, with $\mu_Q = \mu_{\mathfrak{R}}$ under the stated reward-curvature assumption.

C.1.2 The proof in three steps

Proof of Theorem 3.2. Step 1: public exploitability controls public-policy error. Quadratic growth bounds the squared error of a decision rule by its loss in equilibrium value. Subgame best-response bounds control the sum of these losses along play. Lemma C.2 makes this precise and uses Cauchy–Schwarz to obtain

$$\Delta \leq 2t_{\max}\sqrt{\epsilon/\mu_Q}.$$

Step 2: control the effect of changing beliefs. The error Δ compares both public policies at the evaluated beliefs, whereas Δ_{orig} compares the policies at their respective induced beliefs. Controlling belief drift and applying the reference-policy bound Equation (4) gives, by Proposition C.1,

$$\Delta_{\text{orig}} \leq (1 + 2L)^{t_{\max}-1}\Delta.$$

Step 3: policy error controls original-game exploitability. Correspondence (Theorem 2.3) makes π^* unexploitable. The payoff comparison and the uniform reach-ratio bound control the change in each best-response value when replacing π^* by π. As shown in Theorem C.5,

$$\text{expl}_{\text{orig}}(\pi) \leq \frac{(L_R + t_{\max}R_{\max})\overline{\mathcal{C}}}{2}\Delta_{\text{orig}}.$$

Combining the three estimates proves the claim, with

$$C = \frac{(L_R + t_{\max}R_{\max})t_{\max}(1 + 2L)^{t_{\max}-1}\overline{\mathcal{C}}}{\sqrt{\mu_Q}}.$$

□

A stronger policy conclusion under a reach margin. If additionally $z(h_i) \geq \eta > 0$ uniformly, the maximum-error bound in Proposition C.1, together with Step 1, gives

$$\|\Pi^{\downarrow}(\tilde{\pi}) - \pi^*\|_{\text{loc}} = O\left(\sqrt{\text{expl}_{\text{pub}}(\tilde{\pi})}\right). \tag{8}$$

At zero error, the induced behavioral policies coincide even without this additional margin: full support and $\Delta_{\text{orig}} = 0$ suffice. The public policies may still differ away from their induced PBSs.

C.1.3 Step 1: from public exploitability to policy error

The following lemma supplies the square-root estimate used in the main proof. It uses subgame perfection to compare continuation values and quadratic growth to turn value losses into squared policy errors.

Lemma C.2 (Public exploitability controls policy error)**.** *Under the reference-equilibrium and quadratic growth assumptions above, the errors in Equations* (1) *and* (2) *satisfy*

$$\sum_{i,h_i} z(h_i)e(h_i)^2 \le \frac{4t_{max}\epsilon}{\mu_Q}, \qquad \Delta \le 2t_{max}\sqrt{\epsilon/\mu_Q}.$$

Proof. Continuation best-response gaps. Define

$$U(\tilde{s}) = \sup_{\tilde{\pi}'_0} \tilde{\mathfrak{J}}_{\tilde{s}}(\tilde{\pi}'_0, \tilde{\pi}_1), \qquad D(\tilde{s}) = \inf_{\tilde{\pi}'_1} \tilde{\mathfrak{J}}_{\tilde{s}}(\tilde{\pi}_0, \tilde{\pi}'_1), \qquad \gamma = U - D.$$

The gap $\gamma(\tilde{s})$ measures the two players' best-response opportunities from $\tilde{s}$; at the root, $\gamma(\mu) = 2\epsilon$. Let P_t be the evaluated expected payoff before depth t. Each player can follow $\tilde{\pi}$ to depth t, then use independent approximate best responses at distinct public-history labels. Earlier rewards and reaches stay fixed, so letting approximation errors vanish gives

$$U(\mu) \ge P_t + \sum_{h_{\text{pub}}:\text{depth}(h_{\text{pub}})=t} z(h_{\text{pub}})U(\tilde{s}_{h_{\text{pub}}}),$$
$$D(\mu) \le P_t + \sum_{h_{\text{pub}}:\text{depth}(h_{\text{pub}})=t} z(h_{\text{pub}})D(\tilde{s}_{h_{\text{pub}}}).$$

Subtracting cancels P_t:

$$\sum_{h_{\text{pub}}:\text{depth}(h_{\text{pub}})=t} z(h_{\text{pub}})\gamma(\tilde{s}_{h_{\text{pub}}}) \le U(\mu) - D(\mu) = 2\epsilon.$$

Policy error at one PBS. By subgame perfection, $Q^*(\tilde{s}, \tilde{a})$ is the public security value with only the first rule fixed to $\tilde{a}$: successor subgames have disjoint public-history labels and equilibrium values $\tilde{V}^*(\tilde{s}_o)$. For the played rule $\tilde{a} = \tilde{\pi}_i(\tilde{s})$, constraining the entire policy instead of only its first rule and applying the saddle inequalities gives

$$i = 0: \quad D(\tilde{s}) \le Q^*(\tilde{s}, \tilde{a}) \le \tilde{V}^*(\tilde{s}) \le U(\tilde{s}),$$
$$i = 1: \quad D(\tilde{s}) \le \tilde{V}^*(\tilde{s}) \le Q^*(\tilde{s}, \tilde{a}) \le U(\tilde{s}).$$

Thus $|\tilde{V}^*(\tilde{s}) - Q^*(\tilde{s}, \tilde{a})| \le \gamma(\tilde{s})$. At a fixed public history h_{pub}, let i be the acting player and set $\tilde{s} = \tilde{s}_{h_{\text{pub}}}$. Applying Equation (7) to the played rule $\tilde{a} = \tilde{\pi}_i(\tilde{s})$ gives

$$\frac{\mu_Q}{2} \sum_{h_i} \tilde{s}_i(h_i)e(h_i)^2 \le \left|\tilde{V}^*(\tilde{s}) - Q^*(\tilde{s}, \tilde{a})\right| \le \gamma(\tilde{s}).$$

Multiplying by $z(h_{\text{pub}})$ and using $z(h_i) = z(h_{\text{pub}})\tilde{s}_i(h_i)$ yields

$$\frac{\mu_Q}{2} \sum_{h_i \in \mathbb{H}_i(h_{\text{pub}})} z(h_i)e(h_i)^2 \le z(h_{\text{pub}})\gamma(\tilde{s}_{h_{\text{pub}}}).$$

Accumulating error over the horizon. Summing over public histories at depth t collects all decision information states at that depth. The preceding bound on their weighted gaps therefore gives

$$\frac{\mu_Q}{2} \sum_{i,h_i:\text{depth}(h_{\text{pub}})=t} z(h_i)e(h_i)^2 \le \sum_{h_{\text{pub}}:\text{depth}(h_{\text{pub}})=t} z(h_{\text{pub}})\gamma(\tilde{s}_{h_{\text{pub}}})$$
$$\le 2\epsilon.$$

Dividing by $\mu_Q/2$ and summing over $0 \le t < t_{\text{max}}$ gives

$$\sum_{i,h_i} z(h_i)e(h_i)^2 \le \frac{2}{\mu_Q} \sum_{t<t_{\text{max}}} 2\epsilon = \frac{4t_{\text{max}}\epsilon}{\mu_Q}.$$

At each depth, a trajectory visits at most one decision information state of the acting player. The reach probabilities of these mutually exclusive states therefore sum to at most one, giving

$$\sum_{i,h_i} z(h_i) = \sum_{t<t_{\max}} \sum_{i,h_i:\mathrm{depth}(h_{\mathrm{pub}})=t} z(h_i) \le \sum_{t<t_{\max}} 1 = t_{\max}.$$

Cauchy–Schwarz therefore gives

$$\Delta^2 \le \left(\sum_{i,h_i} z(h_i)\right) \sum_{i,h_i} z(h_i) e(h_i)^2 \le \frac{4t_{\max}^2 \epsilon}{\mu_Q}.$$

□

C.1.4 Step 2: from changes in beliefs to original-game policy error

This step accounts for the distinction between the two errors in Equation (2). We first control how changing an original-game policy changes trajectory probabilities and conditional beliefs. We then apply this estimate recursively to compare the evaluated policy with the reference policy.

The trajectory bound is a finite-horizon analogue of the visitation-distribution bound in Achiam et al. [2017, Lemma 3, supplementary material], combined with an elementary normalization bound for conditional PBSs.

Lemma C.3 (Trajectory comparison)**.** *Let π, π' be original-game joint policies with the same initial law μ. Write $p_\pi^t(h) = \Pr_\pi(H^t = h)$ and $q_\pi^t(h,a) = p_\pi^t(h)\pi_\iota(h_\iota, a)$, and similarly for π'. Set $D_t = \mathbb{E}_\pi \|\pi_\iota(H_\iota^t) - \pi'_\iota(H_\iota^t)\|_1$ for $0 \le t < t_{max}$. Then $\|q_\pi^t - q_{\pi'}^t\|_1 \le \sum_{k\le t} D_k$ for $0 \le t < t_{max}$, and, for $0 \le t \le t_{max}$,*

$$\sum_{h_{pub}:\mathrm{depth}(h_{pub})=t} z_\pi(h_{pub}) \left\|\tilde{s}_{\pi,h_{pub}} - \tilde{s}_{\pi',h_{pub}}\right\|_1 \le 2\sum_{k<t} D_k.$$

Proof. History distributions. Set $M_t = \|p_\pi^t - p_{\pi'}^t\|_1$. Adding and subtracting $p_\pi^t(h)\pi'_\iota(h_\iota, a)$ gives

$$\begin{aligned} q_\pi^t(h,a) - q_{\pi'}^t(h,a) &= \big[p_\pi^t(h)\pi_\iota(h_\iota,a) - p_\pi^t(h)\pi'_\iota(h_\iota,a)\big] \\ &\quad + \big[p_\pi^t(h)\pi'_\iota(h_\iota,a) - p_{\pi'}^t(h)\pi'_\iota(h_\iota,a)\big] \\ &= p_\pi^t(h)\big[\pi_\iota(h_\iota,a) - \pi'_\iota(h_\iota,a)\big] \\ &\quad + \big[p_\pi^t(h) - p_{\pi'}^t(h)\big]\pi'_\iota(h_\iota,a). \end{aligned}$$

Taking absolute values and summing, using $\sum_a \pi'_\iota(h_\iota,a) = 1$, yields

$$\begin{aligned} \|q_\pi^t - q_{\pi'}^t\|_1 &\le \sum_h p_\pi^t(h)\|\pi_\iota(h_\iota) - \pi'_\iota(h_\iota)\|_1 \\ &\quad + \sum_h |p_\pi^t(h) - p_{\pi'}^t(h)| = D_t + M_t. \end{aligned}$$

The game's transition law gives $p_\pi^{t+1}(h') = \sum_{h,a} q_\pi^t(h,a)\mathcal{T}(h' \mid h,a)$. Its nonnegative probabilities sum to one over h', so the triangle inequality gives $M_{t+1} \le \|q_\pi^t - q_{\pi'}^t\|_1$. Since $M_{t+1} \le M_t + D_t$ and $M_0 = 0$, induction gives $M_t \le \sum_{k<t} D_k$. Substituting into the earlier bound yields $\|q_\pi^t - q_{\pi'}^t\|_1 \le D_t + M_t \le \sum_{k\le t} D_k$.

Conditional beliefs. At a fixed public history, abbreviate the PBSs by $\tilde{s}_\pi, \tilde{s}_{\pi'}$ and write $m_\pi = z_\pi(h_{\mathrm{pub}})\tilde{s}_\pi$, $m_{\pi'} = z_{\pi'}(h_{\mathrm{pub}})\tilde{s}_{\pi'}$. Since each PBS sums to one, $z_\pi(h_{\mathrm{pub}}) = \sum_h m_\pi(h)$, and similarly for π'. Thus

$$\begin{aligned} |z_{\pi'}(h_{\mathrm{pub}}) - z_\pi(h_{\mathrm{pub}})| &= \left|\sum_h \big(m_{\pi'}(h) - m_\pi(h)\big)\right| \\ &\le \sum_h |m_{\pi'}(h) - m_\pi(h)| = \|m_\pi - m_{\pi'}\|_1. \end{aligned}$$

Combining this bound with the identity

$$z_\pi(h_{\text{pub}})(\tilde{s}_\pi - \tilde{s}_{\pi'}) = m_\pi - m_{\pi'} + [z_{\pi'}(h_{\text{pub}}) - z_\pi(h_{\text{pub}})]\tilde{s}_{\pi'}$$

gives

$$z_\pi(h_{\text{pub}})\|\tilde{s}_\pi - \tilde{s}_{\pi'}\|_1 \le 2\|m_\pi - m_{\pi'}\|_1.$$

This also holds at zero reach. For $h \in \mathbb{H}(h_{\text{pub}})$ at depth t, we have $m_\pi(h) = p^t_\pi(h)$, and similarly for π'. The sets $\mathbb{H}(h_{\text{pub}})$ partition the depth-t histories, so summing gives

$$\begin{aligned}
&\sum_{h_{\text{pub}}:\text{depth}(h_{\text{pub}})=t} z_\pi(h_{\text{pub}}) \left\|\tilde{s}_{\pi,h_{\text{pub}}} - \tilde{s}_{\pi',h_{\text{pub}}}\right\|_1 \\
&\quad \le 2 \sum_{h_{\text{pub}}:\text{depth}(h_{\text{pub}})=t} \sum_{h\in\mathbb{H}(h_{\text{pub}})} |p^t_\pi(h) - p^t_{\pi'}(h)| \\
&\quad = 2M_t \le 2\sum_{k<t} D_k.
\end{aligned}$$

□

The next proposition combines this control of belief drift with the Lipschitz bound on the reference policy. Its recurrence separates direct policy error from the error caused by evaluating the reference at different beliefs.

Proposition C.1 (Reach-weighted policy comparison)**.** *Under Equation* (4)*, let* $e(h_i)$ *and* Δ *be as in Equations* (1) *and* (2)*. Then*

$$\sum_{i,h_i} z(h_i)\|\pi_i(h_i) - \pi^*_i(h_i)\|_1 \le (1+2L)^{t_{max}-1}\Delta.$$

If additionally $z(h_i) \ge \eta > 0$ *for every decision information state, then the maximum local original-game policy error satisfies*

$$\|\pi - \pi^*\|_{\text{loc}} = \max_{i,h_i} \|\pi_i(h_i) - \pi^*_i(h_i)\|_1 \le \frac{(1+2L)^{t_{max}-1}}{\eta}\Delta.$$

Proof. For $0 \le t < t_{\text{max}}$, define

$$\begin{aligned}
D_t &= \sum_{i,h_i:\text{depth}(h_{\text{pub}})=t} z(h_i)\|\pi_i(h_i) - \pi^*_i(h_i)\|_1, \\
E_t &= \sum_{i,h_i:\text{depth}(h_{\text{pub}})=t} z(h_i)e(h_i).
\end{aligned}$$

Here h_{pub} is the public history associated with h_i, and $\Delta = \sum_{t<t_{\text{max}}} E_t$. *Comparing policies at one information state.* By the correspondence mapping, $\pi_i(h_i) = \tilde{\pi}_i(\tilde{s}_{h_{\text{pub}}})(h_i)$ and $\pi^*_i(h_i) = \tilde{\pi}^*_i(\tilde{s}_{*,h_{\text{pub}}})(h_i)$. Adding and subtracting $\tilde{\pi}^*_i(\tilde{s}_{h_{\text{pub}}})(h_i)$, the triangle inequality and Equation (4) give

$$\begin{aligned}
\|\pi_i(h_i) - \pi^*_i(h_i)\|_1 &\le e(h_i) + \left\|\tilde{\pi}^*_i(\tilde{s}_{h_{\text{pub}}})(h_i) - \tilde{\pi}^*_i(\tilde{s}_{*,h_{\text{pub}}})(h_i)\right\|_1 \\
&\le e(h_i) + L\|\tilde{s}_{h_{\text{pub}}} - \tilde{s}_{*,h_{\text{pub}}}\|_1.
\end{aligned}$$

Accumulating error over the horizon. Multiplying by $z(h_i)$ and summing over depth-t decision information states, then grouping by public history and applying Lemma C.3, gives

$$D_t \le E_t + L \sum_{h_{\text{pub}}:\text{depth}(h_{\text{pub}})=t} z(h_{\text{pub}})\|\tilde{s}_{h_{\text{pub}}} - \tilde{s}_{*,h_{\text{pub}}}\|_1 \le E_t + 2L\sum_{k<t} D_k.$$

Thus $B_t = \sum_{k<t} D_k$ satisfies $B_0 = 0$ and $B_{t+1} \le (1+2L)B_t + E_t$, so $B_{t_{\text{max}}} \le (1+2L)^{t_{\text{max}}-1}\Delta$.

The stronger maximum-error bound. For the maximum local error, assume $z(h_i) \ge \eta > 0$ at every decision information state. For each (i, h_i), the weighted bound just proved gives

$$\begin{aligned}
\eta\|\pi_i(h_i) - \pi^*_i(h_i)\|_1 &\le z(h_i)\|\pi_i(h_i) - \pi^*_i(h_i)\|_1 \\
&\le \sum_{j,h_j} z(h_j)\|\pi_j(h_j) - \pi^*_j(h_j)\|_1 \\
&\le (1+2L)^{t_{\text{max}}-1}\Delta.
\end{aligned}$$

Dividing by η and taking the maximum over (i, h_i) proves the local-error bound. □

C.1.5 Step 3: from policy error to original-game exploitability

We first compare two policies against a fixed opponent, allowing the opponent to be arbitrary. We then use the reach-ratio assumption to make that comparison uniform over best responses.

The next lemma is a simulation-lemma-style bound, related to the visitation-weighted performance comparison of Ross et al. [2011, Theorem 2.2], with an additional term for policy-dependent rewards.

Lemma C.4 (Payoff comparison)**.** *Under Equation* (3)*, let π_i, π_i' be policies for player i. Against any fixed opponent β_{-i},*

$$|\mathfrak{J}(\pi_i, \beta_{-i}) - \mathfrak{J}(\pi_i', \beta_{-i})| \le (L_R + t_{max} R_{\max}) \sum_{h_i} z_{(\pi_i, \beta_{-i})}(h_i) \|\pi_i(h_i) - \pi_i'(h_i)\|_1.$$

Proof. For $i = 0$, apply Lemma C.3 to $\pi = (\pi_0, \beta_1)$ and $\pi' = (\pi_0', \beta_1)$, and put $F = \sum_{t<t_{\max}} D_t$. Only player 0's decisions contribute, so F is the weighted policy error in the statement. With $r_\pi^t(h, a) = \mathfrak{R}(h, a, \pi_\iota(h_\iota))$ and similarly for π',

$$\mathfrak{J}(\pi) - \mathfrak{J}(\pi') = \sum_{t<t_{\max}} \left[\langle q_\pi^t, r_\pi^t - r_{\pi'}^t \rangle + \langle q_\pi^t - q_{\pi'}^t, r_{\pi'}^t \rangle \right],$$

$$|\mathfrak{J}(\pi) - \mathfrak{J}(\pi')| \le L_R F + R_{\max} \sum_{t<t_{\max}} \sum_{k \le t} D_k \le (L_R + t_{\max} R_{\max}) F.$$

Reward differences vanish at player 1's decisions because β_1 is fixed. Interchanging the player indices gives the other case; neither needs a margin on the fixed opponent. □

To compare best-response values, we use $|\sup f - \sup g| \le \sup |f - g|$ and its infimum analogue. The reach ratio converts the weights under an opponent deviation into the evaluated weights used by Δ.

Theorem C.5 (Reach-weighted exploitability transfer)**.** *Under Equations* (3) *and* (4)*, suppose the objective is UG and $\tilde{\pi}^*$ is a PuB-AMG Nash equilibrium from μ. Let $\mathcal{C}$ be the opponent-deviation reach ratio defined in Equation* (5)*. Then $\mathrm{expl}_{\mathrm{orig}}(\pi) = O(\mathcal{C}\Delta)$. If $z(h_i) \ge \eta > 0$, then $\mathcal{C} \le 1/\eta$ and $\mathrm{expl}_{\mathrm{orig}}(\pi) = O(\Delta)$, uniformly under common parameter bounds.*

Proof. Apply Lemma C.4 to each best-response value, using $z_{(\pi_i, \pi'_{-i})}(h_i) \le \mathcal{C} z(h_i)$ and $|\sup f - \sup g| \le \sup |f - g|$ (and its infimum analogue). Since exploitability is half the best-response gap,

$$|\mathrm{expl}_{\mathrm{orig}}(\pi) - \mathrm{expl}_{\mathrm{orig}}(\pi^*)| \le \frac{(L_R + t_{\max} R_{\max})\mathcal{C}}{2} \sum_{i, h_i} z(h_i) \|\pi_i(h_i) - \pi_i^*(h_i)\|_1.$$

Correspondence (Theorem 2.3) gives zero reference exploitability; Proposition C.1 finishes the bound. Each numerator reach is at most one, giving $\mathcal{C} \le 1/\eta$. □

C.1.6 Why strong reward curvature gives quadratic growth

The main argument above uses quadratic growth of Q^*. This final supporting lemma proves it from the strong reward-curvature assumption Equation (6). Its proof mixes continuation policies in sequence form, retaining the strong curvature at the first decision.

This uses realization-plan mixing [Von Stengel, 1996] and the perspective construction underlying sequence-form regularization [Hoda et al., 2010, Section 3]; the final quadratic-gap bound is the standard consequence of strong concavity [Bubeck, 2015, Section 3.4].

Lemma C.6 (Curvature of a public decision rule)**.** *Under the equilibrium and strong reward curvature assumptions above, if player* 0 *acts at $\tilde{s}$, then every legal rule $\tilde{a}$ satisfies*

$$\tilde{V}^*(\tilde{s}) - Q^*(\tilde{s}, \tilde{a}) \ge \frac{\mu_{\mathfrak{R}}}{2} \sum_{h_0} \tilde{s}_0(h_0) \|\tilde{a}(h_0) - \tilde{\pi}_0^*(\tilde{s})(h_0)\|_1^2.$$

For player 1*, the same bound holds with $Q^* - \tilde{V}^*$ on the left and player-*1 *components on the right. Zero-marginal information states contribute zero. Thus Equation* (7) *holds with $\mu_Q = \mu_{\mathfrak{R}}$.*

Proof. Fixing the first decision rule. Fix $\tilde{s}$ where player 0 acts. The first-rule-constrained security value is

$$F(\tilde{a}) = \sup_{\pi_0 :\, \pi_0|_{\text{first}}=\tilde{a}} \inf_{\beta_1} \mathfrak{J}_{\tilde{s}}(\pi_0, \beta_1) = Q^*(\tilde{s}, \tilde{a}).$$

Indeed, fixing the first rule fixes the immediate reward and successor law. Distinct public observations have disjoint continuation information states, so their security values optimize independently. Original/public value equivalence [Sokota et al., 2023b, Corollary B.3] applies in each successor subgame; zero-probability branches vanish.

Mixing continuation policies. Compactness and continuity give optimal constrained plans $\pi_0^{\tilde{a}}, \pi_0^{\tilde{b}}$. Let $x_{\tilde{a}}(h_0), x_{\tilde{b}}(h_0)$ be their sequence-form realization weights for the preceding own-action sequences: the products of player 0's preceding action probabilities within the subgame. For $0 < \lambda < 1$, form

$$\pi_0^{\lambda}(h_0) = \frac{\lambda x_{\tilde{a}}(h_0)\pi_0^{\tilde{a}}(h_0) + (1-\lambda)x_{\tilde{b}}(h_0)\pi_0^{\tilde{b}}(h_0)}{\lambda x_{\tilde{a}}(h_0) + (1-\lambda)x_{\tilde{b}}(h_0)},$$

with any legal distribution at a zero denominator. Induction over own actions gives $x_\lambda = \lambda x_{\tilde{a}} + (1-\lambda)x_{\tilde{b}}$. Against a fixed β_1, let $z_j(h)$ be full-history reach under (π_0^j, β_1). Perfect recall factors $z_j(h) = c_\beta(h)x_j(h_0)$ into fixed root, chance, and opponent factors and player 0's own reach. Consequently,

$$z_\lambda(h) = \lambda z_{\tilde{a}}(h) + (1-\lambda)z_{\tilde{b}}(h),$$
$$z_\lambda(h)\pi_0^{\lambda}(h_0) = \lambda z_{\tilde{a}}(h)\pi_0^{\tilde{a}}(h_0) + (1-\lambda)z_{\tilde{b}}(h)\pi_0^{\tilde{b}}(h_0).$$

Using reward curvature. Rewards at player 1's decisions therefore mix affinely. At player 0's decisions with $z_\lambda(h) > 0$, apply concavity in Equation (6) with mixing weight $\lambda z_{\tilde{a}}(h)/z_\lambda(h)$ and multiply by $z_\lambda(h)$. This bounds the expected reward below by the mixture of the two plans' rewards; zero-reach terms vanish. At the first decision $x_{\tilde{a}} = x_{\tilde{b}} = 1$, $z_j(h) = \tilde{s}(h)$, and the mixed rule is $\lambda\tilde{a} + (1-\lambda)\tilde{b}$. Keeping the strong concavity term there and summing over histories gives

$$\mathfrak{J}_{\tilde{s}}(\pi_0^{\lambda}, \beta_1) \geq \lambda\mathfrak{J}_{\tilde{s}}(\pi_0^{\tilde{a}}, \beta_1) + (1-\lambda)\mathfrak{J}_{\tilde{s}}(\pi_0^{\tilde{b}}, \beta_1) + \frac{\mu_{\mathfrak{R}}}{2}\lambda(1-\lambda)d^2(\tilde{a}, \tilde{b}),$$
$$d^2(\tilde{a}, \tilde{b}) := \sum_{h_0} \tilde{s}_0(h_0)\|\tilde{a}(h_0) - \tilde{b}(h_0)\|_1^2.$$

Taking the opponent infimum and using $\inf(f+g) \geq \inf f + \inf g$ proves the same strong concavity inequality for F. *Comparing with the equilibrium rule.* Subgame perfection makes $\tilde{a}^* = \tilde{\pi}_0^*(\tilde{s})$ optimal, so

$$F(\tilde{a}^*) \geq F((1-\lambda)\tilde{a}^* + \lambda\tilde{a}) \geq (1-\lambda)F(\tilde{a}^*) + \lambda F(\tilde{a}) + \frac{\mu_{\mathfrak{R}}}{2}\lambda(1-\lambda)d^2(\tilde{a}, \tilde{a}^*).$$

Subtract $(1-\lambda)F(\tilde{a}^*)$, divide by λ, and let $\lambda \downarrow 0$. Since $F(\tilde{a}^*) = \tilde{V}^*(\tilde{s})$, this proves the player-0 bound. For player 1, use $\inf_{\pi_1}\sup_{\beta_0}\mathfrak{J}_{\tilde{s}}(\beta_0, \pi_1)$ with the first rule fixed. Convexity reverses the inequalities and gives the bound on $Q^* - \tilde{V}^*$. □

C.1.7 KL specialization

The entropy curvature used below is the standard ℓ_1 strong convexity of negative entropy, equivalently Pinsker's inequality [Bubeck, 2015, Section 4.3].

Corollary C.7 (KL specialization)**.** *For Gravel's minimax-KL objective, assume the reference conditions above and fix $\alpha > 0$. Write*

$$\mathrm{KL}(p\|u_{h_i}) = \sum_a p_a \log\frac{p_a}{u_{h_i,a}},$$

where u_{h_i} is uniform on legal actions, logarithms are natural, and $0\log 0 = 0$. Then strong reward curvature holds with $\mu_{\mathfrak{R}} = \alpha$, and Equation (7) *holds with $\mu_Q = \alpha$. A common legal-action margin $\kappa > 0$ for π, π^* supplies the payoff Lipschitz and reach bounds; Equation* (4) *remains a separate assumption. For fixed α and $\epsilon = \mathit{expl}^{\mathrm{reg}}_{\mathrm{pub}}(\tilde{\pi})$,*

$$\mathit{expl}^{\mathrm{unreg}}_{\mathrm{orig}}(\pi) \leq O(\sqrt{\epsilon}) + \alpha t_{max}\log A_{\max}.$$

Proof. On the simplex interior,

$$v^\top \nabla_p^2 \operatorname{KL}(p\|u)v = \sum_a \frac{v_a^2}{p_a} \geq \|v\|_1^2.$$

Cauchy–Schwarz gives the inequality; integration along segments and continuity extend modulus-one strong convexity to the boundary. At player 0's decisions, $\mathfrak{R}(h,p)$ is affine minus $\alpha \operatorname{KL}$, hence α-strongly concave. At player 1's decisions it is affine plus $\alpha \operatorname{KL}$, hence α-strongly convex. Lemma C.6 therefore gives quadratic growth with $\mu_Q = \alpha$. Also $0 \leq \operatorname{KL}(p\|u_{h_i}) = \log|\mathbb{A}(h_i)| + \sum_a p_a \log p_a \leq \log A_{\max}$, bounding rewards even for boundary-valued deviations. On $p_a \geq \kappa$, the gradient coordinates $1 + \log(p_a/u_a)$ are bounded. The action margin and finitely many feasible paths with positive initial and chance probabilities give a common positive reach margin, hence a uniform bound on $\mathcal{C}$. Finally, removing regularization changes every original-game payoff by at most $b = \alpha t_{\max} \log A_{\max}$. Each best-response extremum, and hence half their difference, changes by at most b; apply Theorem 3.2. □

All bounds keep the stated parameters fixed; no solver iteration rate or annealing guarantee is asserted.

C.2 Search with approximate values

Proof of Corollary 3.3. Number the blocks $k = 0, \ldots, N-1$; block k starts at depth kd. For a block-boundary PBS $\tilde{s}$, write $G(\tilde{s}, W)$ for the depth-d PuB-AMG subgame rooted at $\tilde{s}$ whose truncated leaves $\tilde{s}'$ pay $W(\tilde{s}')$, and $\operatorname{val} G(\tilde{s}, W)$ for its value. All values use player 0's payoff.

Values are stable under leaf perturbations. Since $\tilde{\pi}^*$ is subgame-perfect, its restriction to the block is an equilibrium of $G(\tilde{s}, \tilde{V}^*)$, so $\operatorname{val} G(\tilde{s}, \tilde{V}^*) = \tilde{V}^*(\tilde{s})$. Every joint policy's payoff in $G(\tilde{s}, W)$ is the expected interior reward plus the expected leaf payoff, so shifting the leaf payoffs by at most ϵ_V in sup norm shifts every payoff, and hence the max-min value, by at most ϵ_V:

$$\left|\operatorname{val} G(\tilde{s}, \hat{V}) - \tilde{V}^*(\tilde{s})\right| \leq \epsilon_V.$$

Induction. Let $B_k(\tilde{s})$ be player 0's best-response value against $\tilde{\pi}_1$ from a block-k boundary PBS $\tilde{s}$. We show $B_k(\tilde{s}) \leq \tilde{V}^*(\tilde{s}) + (N-k)(\epsilon_K + 2\epsilon_V)$. At $k = N$ the game is over and both sides are zero. Suppose the claim holds for $k+1$. In block k, player 1 plays its component σ_1 of the solution of $G(\tilde{s}, \hat{V})$, which depends only on $\tilde{s}$. For any deviation σ_0 of player 0 within the block, with interior reward r and next boundary PBS $\tilde{S}'$,

$$\begin{aligned}
\mathbb{E}\big[r + B_{k+1}(\tilde{S}')\big] &\leq \mathbb{E}\big[r + \tilde{V}^*(\tilde{S}')\big] + (N-k-1)(\epsilon_K + 2\epsilon_V) \\
&\leq \mathbb{E}\big[r + \hat{V}(\tilde{S}')\big] + \epsilon_V + (N-k-1)(\epsilon_K + 2\epsilon_V) \\
&\leq \operatorname{val} G(\tilde{s}, \hat{V}) + \epsilon_K + \epsilon_V + (N-k-1)(\epsilon_K + 2\epsilon_V) \\
&\leq \tilde{V}^*(\tilde{s}) + (N-k)(\epsilon_K + 2\epsilon_V).
\end{aligned}$$

The third line is the solver guarantee and the last is the stability bound. Taking the supremum over σ_0 proves the claim; the case $k+1 = N$, where $\tilde{S}'$ is terminal, is the same with zero leaf values. The symmetric argument gives $\inf_{\tilde{\pi}_1'} \tilde{\mathfrak{J}}(\tilde{\pi}_0, \tilde{\pi}_1') \geq \tilde{V}^*(\mu) - N(\epsilon_K + 2\epsilon_V)$, so $\operatorname{expl}_{\mathrm{pub}}(\tilde{\pi}) \leq N(\epsilon_K + 2\epsilon_V)$. Theorem 3.2 then gives the second bound, and Corollary C.7 adds at most $\alpha\, t_{\max} \log A_{\max}$ for the unregularized exploitability. □

D Hyperparameters

All hyperparameters and configurations are available at `https://github.com/sobhanmp/gravel`.

E Controlled-game experimental details

Games and metric LD2D3F gives each player two independently and uniformly rolled private dice with three faces. Players alternate between increasing a bid and calling the preceding bid. A

bid specifies a quantity from one to the number of dice present in the game (4) and a face (1 to 3); bids are ordered first by quantity and then by face. The highest face is wild and counts toward bids on either of the other faces. A call ends the game: the bidder wins if the total matching count is at least the bid quantity, and the caller wins otherwise, with payoffs $+1$ and -1. TEH starts on board $3\mathrm{s}\,7\mathrm{h}\,\mathrm{Td}\,\mathrm{Ks}$ with a 2,000-chip pot and 19,000 chips remaining per player. Each player holds two private cards, with the initial joint deal uniform over disjoint hands consistent with the board. Play consists of turn betting, a uniformly dealt legal river card, and river betting; folds award the pot to the other player, and showdowns compare the best five-card hands, splitting ties. Payoffs are net chip gains, including the 1,000 chips each player has already contributed to the initial pot. Its betting abstraction includes fold, check/call, all-in, half-pot and pot-sized first raises, and three-quarter-pot subsequent raises. These fixed-game comparisons do not evaluate off-tree HUNL play. The reported hard exploitability is unregularized, even when the search procedure optimizes a regularized objective. It is half the sum of the two players' best-response improvements against the extracted strategy. TEH values are in milli-big-blinds per hand (mbb/h), with a 100-chip big blind. LD2D3F values are in the game's native payoff units and are not numerically comparable with TEH values.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The theoretical claims are reflected in the theory and the performance claim is reflected in the experiment.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The main text discusses the gap between the theoretical guarantees and the practical algorithm, the bias introduced by regularization, and the scope of the empirical evaluation; the paragraphs after Theorem 3.2 state the scope of the theorems.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Yes, we provide the proof and use standard assumptions.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, we provide the hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Yes, we provide code that is easy to run.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, we provide all of the aforementioned details if applicable; there are no data splits.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We provide error bars for the experiment comparing against Slumbot. We instead opt for running more experiments for some of the other plots.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See github repo.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, we were careful to meet the Ethics Guidelines in conducting this research.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    See the Broader Impacts section.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: cf previous question.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: All code is original.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: There are some comments and examples of usage in the provided code.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or research with human subjects were used.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No humans were studied.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: No LLMs were a component of the core methodology of this research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.