Abstracting Imperfect Information Away from Two-Player Zero-Sum Games

Anonymous authors
Paper under double-blind review

Abstract

In their seminal work, Nayyar et al. (2013) showed that imperfect information can be abstracted away from common-payoff games by having players publicly announce their policies as they play. This insight underpins sound solvers and decision-time planning algorithms for common-payoff games. Unfortunately, a naive application of the same insight to two-player zero-sum games fails because Nash equilibria of the game with public policy announcements may not correspond to Nash equilibria of the original game. As a consequence, existing sound decision-time planning algorithms require complicated additional mechanisms that have unappealing properties. The main contribution of this work is showing that certain regularized equilibria do not possess the aforementioned non-correspondence problem—thus, computing them can be treated as perfect information problems. This result yields a simplified framework for decision-time planning in two-player zero-sum games, void of the unappealing properties that plague existing decision-time planning algorithms.

1 Introduction

In single-agent settings, dynamic programming (Bertsekas, 2000) serves as a bedrock for reinforcement learning (Sutton & Barto, 2018), justifying approximating optimal policies by backward induction and facilitating a simple framework for decision-time planning. One might hope that dynamic programming could also provide similar grounding in multi-agent settings for well-defined notions of optimality, like optimal joint policies in common-payoff games, Nash equilibria in two-player zero-sum (2p0s) games, and team correlated equilibria in two-team zero-sum (2t0s) games. Unfortunately, this is not straightforwardly the case when there is imperfect information—i.e., one player has knowledge that another does not or two players act simultaneously. This difficulty arises from two causes, which we call the *backward dependence problem* and the *non-correspondence problem*.
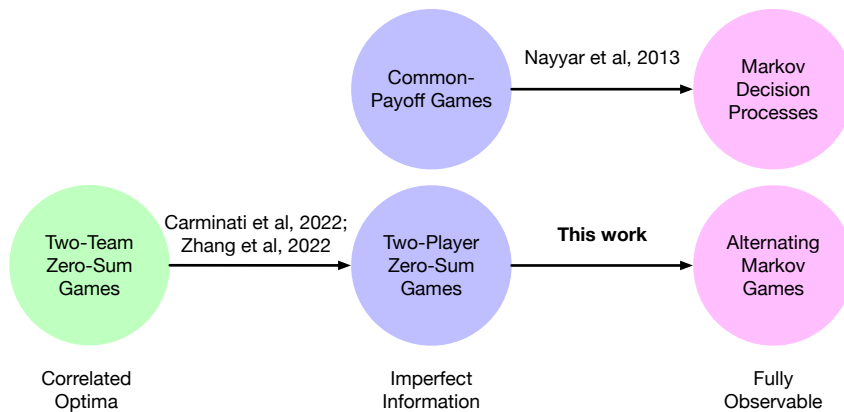

Figure 1: Our contribution in the context of related work, at an abstract level.

The backward dependence problem is that computing the expected return starting from a decision point generally requires knowledge about policies that were played up until now, in addition to the

policies that will be played going forward. This is in stark contrast to perfect information settings, in which the expected return starting from a decision point is independent of the policy played before arriving at the decision point. As a result of this bidirectional temporal dependence, backward induction arguments that work in perfect information settings fail in imperfect information settings.

In their seminal work, Nayyar et al. (2013) showed that the backward dependence problem can be resolved by having players publicly announce their policies as they play. Using this insight, a common-payoff game can be transformed into a public belief Markov decision processes (PuB-MDP). Importantly, deterministic optimal policies in the PuB-MDP can be mapped back to optimal joint policies of the original common-payoff game.

Having players publicly announce their policies can also be used to transform 2p0s games into alternating Markov games (AMGs) with public belief states, which we call public belief alternating Markov games (PuB-AMGs). AMGs are fully observable turn-based games (like Go and chess) and, therefore, are amenable to dynamic programming-based approaches (Littman, 1996). Unfortunately, computing Nash equilibria of PuB-AMGs carries little value because these Nash equilibria may not correspond with Nash equilibria in the original game (Ganzfried & Sandholm, 2015; Brown et al., 2020). Indeed, as we will show, they may correspond with arbitrarily exploitable policies. We call this problem *the non-correspondence problem*.

The main contribution of this work is showing that regularized minimax objectives that guarantee unique equilibria in subgames do not suffer from the non-correspondence problem. In other words, computing these uniqueness guaranteeing equilibria can be reduced to computing the associated equilibria in the PuB-AMG. Because guaranteeing uniqueness is as simple as adding entropy regularization (Perolat et al., 2021), our result is straightforward to apply in practice.

We view the primary importance of this reduction from imperfect information 2p0s games to perfect information games as three fold.

1. It is the first known reduction of its kind; specifically, it is the first regularized equilibrium preserving transformation from imperfect information 2p0s games to perfect information 2p0s games that is meaningful in the sense that regularized equilibria can be made arbitrarily close to Nash equilibria.

2. It yields a simple framework for decision-time planning in 2p0s games. The simplicity of this framework is in stark contrast to existing approaches (Brown & Sandholm, 2017a;b; Moravčík et al., 2017; Brown et al., 2020; Schmid et al., 2021), which are hampered by complications that involve undesirable facets, such as discontinuous functions.

3. It can be applied across the whole class of 2t0s games (as depicted in Figure 1) because of the recent results of Carminati et al. (2022a); Zhang et al. (2022); Carminati et al. (2022b), who showed that 2t0s games can be cast as 2p0s games.

2 NOTATION

2.1 FINITE-HORIZON SEQUENTIAL GAMES

We use the terminology finite-horizon sequential game to describe settings in which players act one-at-a-time and in which the game terminates after a fixed number of steps. This setting can express any (simultaneous-move) timeable perfect-recall finite-horizon game. See, for example, Kovařík et al. (2019) for more details.

Symbolically, we say a setting is a finite-horizon sequential game if it can be described by a tuple

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

where

- i ranges from 0 to $N-1$ and $\iota \in \{0, \ldots, N-1\}$ denotes the acting player.[1]
- $\mathbb{A}$ 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 (common knowledge to all players).

---

[1] Our usage of ι is informal but unambiguous in context.

- $\mathbb{H}_i = \cup_{t=0}^{T-1} (\mathbb{O}_{\text{pub}} \times \mathbb{O}_i)^t \times \mathbb{O}_{\text{pub}} \times \mathbb{O}_i$ is player i's action-observation histories (AOHs).
- $\mathbb{H}_{\text{pub}} = \cup_{t=0}^{T-1} \mathbb{O}_{\text{pub}}^t$ is the set of public histories.
- $\mathbb{H} \subset \mathbb{H}_0 \times \cdots \times \mathbb{H}_{N-1}$ is the set of histories.
- $\mu \in \Delta(\mathbb{H}_0(h_{\text{pub}}^0) \times \cdots \times \mathbb{H}_{N-1}(h_{\text{pub}}^0))$ is the initial history distribution.
- $\mathcal{O}_i \colon \mathbb{H} \to \mathbb{O}_i$ is player i's observation function.[2]
- $\mathcal{O}_{\text{pub}} \colon \mathbb{H} \to \mathbb{O}_{\text{pub}}$ is the public observation function.
- $\mathcal{R}_i \colon \mathbb{H} \times \mathbb{A} \to \mathbb{R}$ is the player i's reward function.
- $\mathcal{T} \colon \mathbb{H} \times \mathbb{A} \to \Delta(\mathbb{H})$ is the transition function.
- T is the time horizon at which the game terminates.

For a given $h_{\text{pub}} \in \mathbb{H}_{\text{pub}}$, we use $\mathbb{H}_i(h_{\text{pub}})$ to denote the set of AOHs for player i that are consistent with h_{pub} and $\mathbb{H}(h_{\text{pub}})$ to denote the set of histories that are consistent with h_{pub}. Also, for a history h, we use h_ι to denote the AOH for the acting player at history h. We use capitals of the same letters to denote random variables of the same types.

Special Cases This work will make use of the following special cases.

- **Two team zero sum**: Games in which $\{0, \ldots, N-1\}$ is a disjoint union of two blocks, where $\forall i, j, \mathcal{R}_i = \mathcal{R}_j$ if i, j belong to the same block and $\mathcal{R}_i = -\mathcal{R}_j$ if i, j belong to opposite blocks.
- **Common payoff**: Games in which $\forall i, j, \mathcal{R}_i = \mathcal{R}_j$.
- **Two player zero sum**: Games in which $N = 2$ and $\mathcal{R}_0 = -\mathcal{R}_1$.

In these special cases, the reward of all players is uniquely determined by the reward of any individual player. Thus, we will drop the player index i on the reward function and use $\mathcal{R} = \mathcal{R}_0$.

Subgames For a given finite-horizon sequential game, we use the term subgame to refer to a game that begins with initial history distribution $\mu \in \Delta(\mathbb{H}_0(h_{\text{pub}}) \times \cdots \times \mathbb{H}_{N-1}(h_{\text{pub}}))$ for some particular h_{pub} and is otherwise the same as the original game.

2.2 FINITE-HORIZON FULLY OBSERVABLE SEQUENTIAL GAMES

We also introduce specialized notation for fully observable settings. We use the terminology finite-horizon fully observable sequential game to describe tuples

$$\langle \mathbb{A}, \mathbb{S}, s^0, [\mathcal{R}_i], \mathcal{T}, T \rangle,$$

where

- $\mathbb{S}$ is the set of states.
- $s^0 \in \mathbb{S}$ is the initial state.
- $\mathcal{R}_i \colon \mathbb{S} \times \mathbb{A} \to \mathbb{R}$ is the player i's reward function.
- $\mathcal{T} \colon \mathbb{S} \times \mathbb{A} \to \Delta(\mathbb{S})$ is the transition function.
- $i, \iota, \mathbb{A}$, and T are defined as they were in the finite-horizon sequential game formalism.

Special Cases In the fully observable context, we are interested in the following settings.

- **Markov decision processes (MDPs)**: Games in which $N = 1$.
- **Alternating Markov games (AMGs)**: Games that are two player zero sum.

As before, because the reward of one player uniquely determines the others in these settings, we will use $\mathcal{R} = \mathcal{R}_0$.

Subgames For a given finite-horizon fully observable sequential game, we use the term subgame to refer to a game that begins with initial state $s^0 = s$ for some particular s and otherwise proceeds by the same rules of the original game.

3 BACKGROUND

The Backward Dependence Problem To illustrate the presence of the backward dependence problem in even very simple settings, we show a cooperative matching pennies game in Figure 2. The goal of the game is for the blue player and the red player to select the same side of a coin. The

---

[2]We assume that, if i acts a time t, i's action is included in its private observation at time $t + 1$

blue player moves first; then, without observing the blue player's choice, the red player moves second. Because the red player does not observe the blue player's choice (as denoted by the dotted line between the two nodes), it must make the same decision at both nodes.

Now, let us consider the value for the red player. In perfect-information settings, because the red player is at a penultimate decision point, such a value would be equal to the expected return for the best action, independent of any prior events. However, here, with imperfect information, the expected return for the best action is equal to $\max(p, 1-p)$ where p is the probability that the blue player selects *heads*. If the blue player's policy is unknown, there is no way to compute this value, illustrating that the backward induction approach to learning in perfect information settings fails in imperfect information settings.

The Public Belief Markov Decision Process
We now describe Nayyar et al. (2013)'s transformation for turning common-payoff games into a public belief MDPs (called PuB-MDPs) in such a way that circumvents the backward dependence problem. Let

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

be a finite-horizon common-payoff game. Then we define the associated PuB-MDP as the following finite-horizon fully observable sequential game

$$\langle \tilde{\mathbb{A}}, \tilde{\mathbb{S}}, \tilde{s}^0, \tilde{\mathcal{R}}, \tilde{\mathcal{T}}, \tilde{T} \rangle,$$

where

Figure 2: The best action for the red player depends on the blue player's policy.

- $i = \iota = 0$.
- $\tilde{\mathbb{A}} = \{ \tilde{a} \mid \tilde{a} \colon \mathbb{H}_\iota(h_{\text{pub}}) \to \mathbb{A}, h_{\text{pub}} \in \mathbb{H}_{\text{pub}} \}$ is the set of *prescriptions*.
- $\tilde{\mathbb{S}} = \cup_{h_{\text{pub}}} \Delta(\mathbb{H}(h_{\text{pub}}))$ is the set of public belief states (PBSs).
- $\tilde{s}^0 = \mu$ is the initial PBS.
- $\tilde{\mathcal{R}} \colon \tilde{s}, \tilde{a} \mapsto r$, where $r = \mathbb{E}_{H \sim \tilde{s}} \mathcal{R}(H, \tilde{a}(H_\iota))$.
- $\tilde{\mathcal{T}}(\tilde{s}_{o_{\text{pub}}^{t+1}} \mid \tilde{s}^t, \tilde{a}) = \mathcal{P}(o_{\text{pub}}^{t+1} \mid \tilde{s}^t, \tilde{a}) = \mathbb{E}_{H^t \sim \tilde{s}^t} \mathbb{E}_{A^t \sim \tilde{a}(H_\iota)} \mathcal{P}(o_{\text{pub}}^{t+1} \mid H^t, A^t)$ where the PBS $\tilde{s}_{o_{\text{pub}}^{t+1}}$ is defined by $\tilde{s}_{o_{\text{pub}}^{t+1}}(h^{t+1}) = \mathcal{P}(h^{t+1} \mid \tilde{s}^t, \tilde{a}, o_{\text{pub}}^{t+1}) = \mathbb{E}_{H^t \sim \tilde{s}^t} \mathbb{E}_{A^t \sim \tilde{a}(H_\iota)} \mathcal{P}(h^{t+1} \mid H^t, A^t, o_{\text{pub}}^{t+1})$
- $\tilde{T} = T$.

Nayyar et al. (2013) showed that optimal deterministic policies in the PuB-MDP correspond with optimal joint policies for the common payoff game. Indeed, for the matching pennies game described in Figure 2, we can see that the PuB-MDP perspective resolves the backward dependence problem because the red player observes the blue player's decision rule. If the blue player plays *heads* with probability p, the red player can achieve a value of $\max(p, 1-p)$ by best responding. By backward induction, the blue player can then determine that it is optimal to use $p = 0$ or $p = 1$. Thus, the players can arrive at an optimal joint policy of the original game. For a more detailed discussion on the PuB-MDP, see, e.g., (Sokota, 2020; Sokota et al., 2021).
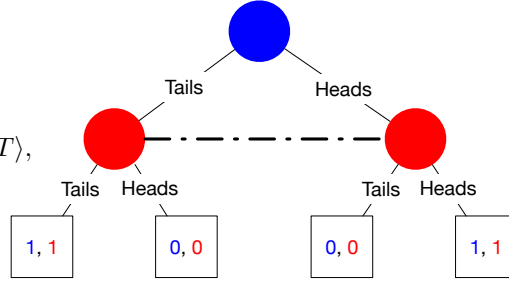
4 THE PUBLIC BELIEF ALTERNATING MARKOV GAME

Here, we introduce a construction analogous to the PuB-MDP that we call the PuB-AMG. Note that, while the name PuB-AMG is novel to this work, the idea already exists in literature (Brown et al., 2020; Buffet et al., 2020; Delage et al., 2021). Let

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

be a finite-horizon 2p0s sequential game. Then we define the associated PuB-AMG as the following finite-horizon fully observable sequential game

$$\langle \tilde{\mathbb{A}}, \tilde{\mathbb{S}}, \tilde{s}^0, \tilde{\mathcal{R}}, \tilde{\mathcal{T}}, \tilde{T} \rangle,$$

where

- i ranges from 0 to 1 and $\iota \in \{0, 1\}$ is the acting player.

- $\tilde{\mathbb{A}} = \{\tilde{a} \mid \tilde{a} \colon \mathbb{H}_\iota(h_{\text{pub}}) \to \Delta(\mathbb{A}), h_{\text{pub}} \in \mathbb{H}_{\text{pub}}\}$ is the set of *public decision rules* (or decision rules, for short). Note that public decision rules differ from prescriptions in that they map to a distribution over actions, rather than an action.

- $\tilde{\mathbb{S}}, \tilde{s}^0, \tilde{\mathcal{R}}, \tilde{\mathcal{T}}$, and $\tilde{T}$ are defined in the same fashion as in the PuB-MDP.

The Correspondence Mapping As alluded to earlier, a Nash equilibrium in the PuB-AMG may be undesirable because it does not necessarily correspond to a Nash equilibrium in the original game. Here, we make precise the notion of what we mean by correspond by defining a correspondence function $\Pi^\downarrow$ that maps public belief (joint) policies to joint policies of the original game. Given a PuB-AMG policy $\tilde{\pi}$, $\Pi^\downarrow(\tilde{\pi})$ is the joint policy that, for each AOH h_ι, plays actions with the probability that $\tilde{\pi}$ would at h_ι, assuming that h_ι was reached using $\tilde{\pi}$. (See Section A for a more rigorous definition.) *Importantly, $\Pi^\downarrow(\tilde{\pi})_i$ can be implemented in practice by running $\tilde{\pi}_i$ under the assumption that the opponent is playing according to $\tilde{\pi}_{-i}$.*

The Non-Correspondence Problem We can now make the non-correspondence problem precise. To illustrate, we show the perturbed variant of rock-paper-scissors described in (Brown et al., 2020) in Figure 3. In the game, the payouts are doubled if either player plays scissors. The unique Nash equilibrium of the game is $(R, P, S) \mapsto (0.4, 0.4, 0.2)$.
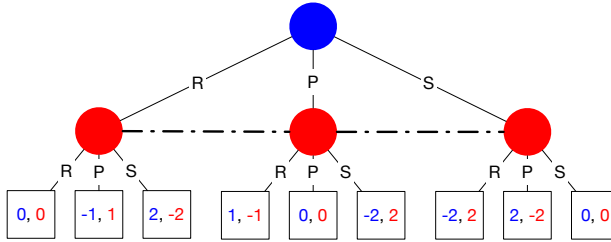

Figure 3: A perturbed variant of rock-paper-scissors.

Similarly to before, the red player can compute the associated value for each of the blue player's decision rules. Thus, the blue player can determine that the Nash equilibrium policy maximizes its value. It is at this point that the non-correspondence problem becomes apparent. Because the red player is conditioning on the blue player's decision rule, it achieves the the optimal value by playing any best response to the red player. In the perturbed rock-paper-scissors game, all policies are best responses to the Nash equilibrium. Thus, there is nothing constraining the red player to the Nash equilibrium policy of the original game.

A similar argument, detailed in Section B.1, leads to the following disappointing result.

Proposition 1. *A PuB-AMG Nash equilibrium $\tilde{\pi}$ may correspond with a joint policy $\Pi^\downarrow(\tilde{\pi})$ that is maximally exploitable.*

At an intuitive level, the non-correspondence problem arises because there is an important distinction between the public belief game and the original game. Specifically, in the public belief game, players acting earlier are forced to reveal their decision rules to players acting later. As a result, later acting players are able to "slack off" without losing any value because the earlier acting players cannot deviate to exploit them. In common-payoff games, this is a non-issue because the interests of every player are aligned. However, in 2p0s games, where there are adversarial interests, this distinction changes the strategic nature of the game in a more fundamental sense.

5 Correspondence of Uniqueness-Guaranteeing Equilibria

We will now show that the non-correspondence problem does not exist for minimax objectives that guarantee unique equilibria in subgames of the original game. The class of minimax objectives we consider generalizes the expected return minimax objective in that it includes objectives that may have dependence on policies beyond the actions they select.

Definition 5.1. *We use the term regularized minimax objective (or objective for short) to refer to mappings of the form*

$$\mathfrak{J} \colon \pi_0, \pi_1 \mapsto \mathbb{E}\left[\sum_{t=0}^{T-1} \mathfrak{R}(H^t, A^t, \pi(H_\iota^t)) \mid \pi_0, \pi_1\right]$$

where $\mathfrak{R}$ is a real-valued function.

Definition 5.2. *Given an objective $\mathfrak{J}$ induced by $\mathfrak{R}$ for some game, we can construct an objective $\tilde{\mathfrak{J}}$ for the associated public belief state game that is equivalent to $\mathfrak{J}$ in the sense that $\tilde{\mathfrak{J}}(\tilde{\pi}) = \mathfrak{J}(\Pi^{\downarrow}(\tilde{\pi}))$ by defining*

$$\tilde{\mathfrak{J}} \colon \tilde{\pi}_0, \tilde{\pi}_1 \mapsto \mathbb{E}\left[\sum_{t=0}^{T-1} \tilde{\mathfrak{R}}(\tilde{S}^t, \tilde{A}^t) \mid \tilde{\pi}_0, \tilde{\pi}_1\right],$$

where

$$\tilde{\mathfrak{R}} \colon (\tilde{s}, \tilde{a}) \mapsto \mathbb{E}_{H \sim \tilde{s}} \mathbb{E}_{A \sim \tilde{a}(H_\iota)} \mathfrak{R}(H, A, \tilde{a}(H_\iota)).$$

Next we formalize what we mean by uniqueness guaranteeing.

Definition 5.3. *For a particular game, we say a minimax objective $\mathfrak{J}$ is uniqueness guaranteeing if, for any subgame,*

$$\max_{\pi_0} \min_{\pi_1} \mathfrak{J}(\pi_0, \pi_1)$$

is guaranteed to have a unique solution (π_0^*, π_1^*).

We can now state our main result.

Theorem 1. *If $\tilde{\pi}$ is an equilibrium of a PuB-AMG under a uniqueness-guaranteeing objective, then its corresponding joint policy $\Pi^{\downarrow}(\tilde{\pi})$ is the equilibrium in the original game under the same uniqueness-guaranteeing objective.*

Proof. (Sketch) The first decision rule of any PuB-AMG equilibrium must correspond to the first decision rule of an equilibrium of the original game. Furthermore, if the objective is uniqueness-guaranteeing, subgame equilibria must be restrictions of the equilibrium of the whole game. Thus, by forward induction, PuB-AMG equilibria must correspond to the equilibrium of the original game. □

We detail the proof for Theorem 1 in Section B.2. Due to recent work, Theorem 1 can be generalized to the entire class of 2t0s games.

Corollary 1. *Computing team-correlated equilibria of 2t0s games under a uniqueness guaranteeing objectives can be reduced to computing an equilibrium of a PuB-AMG under the same uniqueness guaranteeing objectives.*

Proof. This follows immediately from chaining the results of Carminati et al. (2022a); Zhang et al. (2022); Carminati et al. (2022b), who provide a reduction from 2t0s games to 2p0s games via intra-team public policy announcements, with Theorem 1. □

6 Sufficient Conditions for Uniqueness Guaranteeing

Theorem 1 tells us that uniqueness guaranteeing is a sufficient condition to ameliorate the non-correspondence problem. We now discuss objectives that are uniqueness guaranteeing and general sufficient conditions for guaranteeing uniqueness.

Definition 6.1. *We call the objective $\mathfrak{J}$ induced by*

$$\mathfrak{R} \colon (h, a, \delta) \mapsto \begin{cases} \mathcal{R}(h, a) - \alpha KL(\delta, \rho(h_\iota)) & \iota = 0 \\ \mathcal{R}(h, a) + \alpha KL(\delta, \rho(h_\iota)) & \iota = 1, \end{cases}$$

for some reference policy ρ, a MiniMaxKL objective.

Definition 6.2. *We call the objective $\mathfrak{J}$ induced by*

$$\mathfrak{R} \colon (h, a, \delta) \mapsto \begin{cases} \mathcal{R}(h, a) + \alpha \mathcal{H}(\delta) & \iota = 0 \\ \mathcal{R}(h, a) - \alpha \mathcal{H}(\delta) & \iota = 1, \end{cases}$$

for some reference policy ρ, a MiniMaxEnt objective.

Remark 1. *MiniMaxEnt is the special case of MiniMaxKL in which ρ is uniform.*

To our knowledge, MiniMaxKL objectives were first introduced by Perolat et al. (2021), who showed the following result using Lyapunov techniques.

Theorem 2. *(Perolat et al., 2021) MiniMaxKL objectives are uniqueness guaranteeing for interior ρ.*

While the MiniMaxKL equilibrium is likely the most useful uniqueness-guaranteeing equilibrium concept, it is conceivable that other uniqueness-guaranteeing concepts may be useful. Thus, we provide a generalization below.

Theorem 3. *Consider an objective $\mathfrak{J}$ induced by*

$$\mathfrak{R}\colon (h, a, \delta) \mapsto \begin{cases} \mathcal{R}(h, a) - \psi(\delta, h_\iota) & \iota = 0 \\ \mathcal{R}(h, a) + \psi(\delta, h_\iota) & \iota = 1 \end{cases}$$

and define a policy greedification function

$$g\colon [-\mathfrak{M}, \mathfrak{M}]^{|\mathbb{A}|} \times \mathbb{H}_\iota \to \Delta(\mathbb{A})$$

where $\mathfrak{M} \in \mathbb{R}$ is the maximum of the absolute values of the expected returns of $\mathfrak{J}$ and where

$$g\colon q, h_\iota \mapsto arg\,max_{\delta \in \Delta(\mathcal{A})} \langle \delta, q \rangle - \psi(\delta, h_\iota).$$

For each AOH at which a player acts h_ι, g maps possible regularized action values q to the policy that is greedy with respect to the regularized objective under those regularized action values.

If, for all h_ι, $\psi(\cdot, h_\iota)$ is continuous and $g(\cdot, h_\iota)$ is i) well defined, ii) continuous, and iii) has an interior image, then the objective $\mathfrak{J}$ is uniqueness guaranteeing.

Proof. (Sketch) Consider the following facts:

- Because of our assumptions on ψ and g, an equilibrium must exist.
- Because g has an interior image, the best response to an equilibrium must be unique.
- Because the game is two player zero sum, there cannot exist incompatible equilibria.

These facts can only be simultaneously satisfied if there is a unique equilibrium. $\square$

We detail the proof for Theorem 3 in Section B.3.

Remark 2. *The premises of Theorem 3 are satisfied if $\psi(\cdot, h_\iota)$ is bounded and is strictly convex and differentiable on its interior with $\lim_{\delta \to \delta'} ||\nabla_\delta \psi(\delta, h_\iota)|| = +\infty$ for δ' on the boundary of $\Delta(\mathbb{A})$.*

Remark 3. *One example of an objective covered by Theorem 3, but not by Theorem 2, is that which is induced by setting $\psi(\cdot, h_\iota)$ to a sum of a reverse KL divergence to an interior point and a bounded differentiable convex function.*

7 DISCUSSION

In the previous section, we showed that computing certain regularized equilibria in 2p0s games can be reduced to perfect information problems. This may give the impression that it is now possible to apply perfect information algorithms out-of-the-box to solve imperfect information problems. While this is technically true, it is not a good way to view the value of our reduction. Indeed, because the PuB-AMG has an action space of dimensionality on the order of $|\mathbb{A}|^{|\mathbb{H}_\iota(h_{\text{pub}})|}$, naively pursuing such an approach would quickly become intractable as the size of the game increases.[3]

Instead, we see two main ways to approach solving regularized PuB-AMGs. The first is to adapt heuristic search value iteration (Smith & Simmons, 2004) into a tabular regularized PuB-AMG solver. Encouragingly, this has already been done for PuB-MDPs (Dibangoye et al., 2013b) and for unregularized Pub-AMGs (Horák & Bošanský, 2019; Buffet et al., 2020; Delage et al., 2021). The second is to use the regularized PuB-AMG as a building block for expert iteration with function approximation (Anthony et al., 2017). This approach would look almost identical to ReBeL (Brown et al., 2020) but have a few key differences: i) It would use a regularized objective, rather than an unregularized one as ReBeL does; ii) It would use the beliefs induced by its own policy at test time, rather than the fictitious beliefs that ReBeL uses; iii) It would (optionally) be able to perform re-planning (e.g., wherein a multi-ply search is only used to make the immediate decision), whereas ReBeL must play its search policy until the end of the subgame that was searched over.

---

[3]Although an attempt has been made to scale up such approaches for the PuB-MDP (Foerster et al., 2019).

8 EXPERIMENTS

We perform two experiments in which we naively tabularly solve small PuB-AMGs using magnetic mirror descent (Sokota et al., 2022a) to offer further evidence for our results. We show the results for perturbed rock-paper-scissors Figure 4 and include results for Kuhn poker, as well as the details of our solving procedures, in Section C.
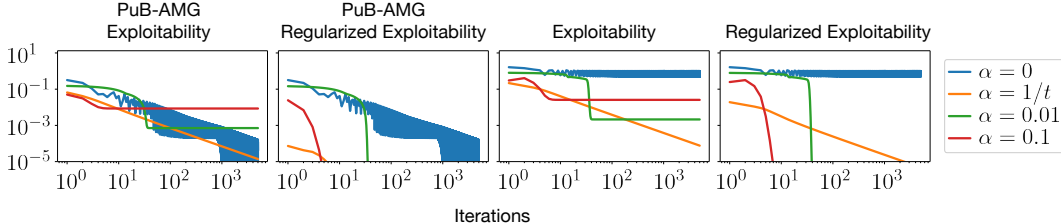

Figure 4: Results for perturbed rock-paper-scissors.

On the far left, we show exploitability in the PuB-AMG. The iterates from the unregularized objective (blue) trend and the objective with annealed regularization (orange) trend toward zero. The iterates from the objective with constant regularization converge to a constant positive exploitability.

On the middle left, we show the regularized exploitability (i.e., exploitability under the regularized objective) in the PuB-AMG of the objective associated for the iterate. We observe that all objectives induce iterates that converge to zero.

On the middle right, we show the exploitability in the original game. Because the non-correspondence exists for the second moving player, the exploitabilities of the iterates from the unregularized objective (blue) remain high, despite that exploitability is going to zero in the PuB-AMG. The objectives with fixed regularization (green, red) induce iterates that converge to lower, but non-zero, exploitability values. The objective with annealed regularization (orange) induces iterates that converge to zero in exploitability.

On the far right, we show the regularized exploitability in the original game. We observe that, in accordance Theorem 1, the approaches with non-zero regularization that converge to zero regularized exploitability in the PuB-AMG also converge to zero regularized exploitability in the original game. In contrast, the unregularized approach does not converge, despite converging in the PuB-AMG.

9 RELATED WORK

Public Belief States in Common-Payoff Games In the sense of providing reductions for multi-agent problems using PBSs, our work is similar to those of Nayyar et al. (2013), Dibangoye et al. (2013b), and Oliehoek (2013). As discussed in the background, Nayyar et al. (2013) provided a reduction from solving common-payoff games to solving belief MDPs; independently, Dibangoye et al. (2013b) and Oliehoek (2013) discovered similar reductions. These ideas have been leveraged in a large body of work in decentralized control literature (Lessard & Nayyar, 2013; Nayyar et al., 2014; Arabneydi & Mahajan, 2014; Ouyang et al., 2015; Vasconcelos & Martins, 2016; Tavafoghi et al., 2016; Afshari & Mahajan, 2018; Gagrani & Nayyar, 2018; Tavafoghi et al., 2018; Zhang et al., 2019; Gupta, 2021) and machine learning literature (Dibangoye et al., 2013a; MacDermed & Isbell, 2013; Dibangoye et al., 2014a;b; Dibangoye & Buffet, 2018; Foerster et al., 2019; Lerer et al., 2020; Sokota et al., 2021; Fickinger et al., 2021; Sokota et al., 2022b; Kao et al., 2022). Use cases include game solving (Dibangoye et al., 2013b), expert iteration (Sokota et al., 2021), and decision-time planning (Lerer et al., 2020; Fickinger et al., 2021; Sokota et al., 2022b).

Public Belief States in Two-Player Zero-Sum Games PBSs have also been used in many works in the context of 2p0s games. For our purposes, we taxonomize these into those concerned with solving the PuB-AMG (Wiggers et al., 2016; Horák & Bošanský, 2019; Buffet et al., 2020; Delage et al., 2021) and those concerned with sound decision-time planning and expert iteration (Burch

et al., 2014; Moravcik et al., 2016; Brown & Sandholm, 2017a;b; Moravčík et al., 2017; Brown et al., 2018; Serrino et al., 2019; Brown et al., 2020; Schmid et al., 2021).

The former group is concerned with analyzing the structure of the PuB-AMG and using HSVI (Smith & Simmons, 2004) to solve the PuB-AMG. Our work is complementary in the sense that it shows that using an approach such as HSVI to solve a regularized PuB-AMG would yield a regularized equilibrium in the original game.

The latter group can be broken down into two subgroups, those that use opt-out values to circumvent the non-correspondence problem (Brown & Sandholm, 2017a;b; Moravčík et al., 2017; Schmid et al., 2021) and those that use no-regret learning to circumvent the non-correspondence problem (Brown et al., 2020). Both possess substantial downsides. For the opt-out value approach: i) the policy and value are discontinuous functions of the opt-out values[4], and ii) the opt-values must be approximated separately from self play. For the no-regret learning approach: i) the search policy must be played for the entire subgame that was searched over (i.e., re-planning is not allowed), ii) the search algorithm must be no regret, and iii) the policy is a discontinuous function of the PBS. In contrast, decision-time planning using a regularized objective in the PuB-AMG involves no opt-out values, involves no value discontinuities (as is shown in Section D), allows for re-planning, and is search algorithm agnostic.

MiniMaxKL Objectives in Two-Player Zero-Sum Games As discussed, prior to this work, Perolat et al. (2021) made use of MiniMaxEnt and MiniMaxKL objectives; they have also been used in more recent works (Cen et al., 2021; Zeng et al., 2022; Sokota et al., 2022a; Perolat et al., 2022). Our use case for MiniMaxKL objectives (eliminating the non-correspondence problem) differs substantially from that of these works (inducing the last iterate convergence of model-free algorithms).

Public Belief States in Two-Team Zero-Sum Games In the sense of providing reductions for multi-agent problems using PBSs, our work is similar to a recent body of literature (Carminati et al., 2022a; Zhang et al., 2022; Carminati et al., 2022b) showing that solving 2t0s games can be reduced to solving a 2p0s game by using intra-team policy announcements. As articulated by Corollary 1, combining our results with these yields a reduction from computing regularized team correlated equilibria in 2t0s games to computing regularized equilibria in PuB-AMGs.

Stackelberg Games Public belief games with two-time steps are closely related to Stackelberg games (von Stackelberg, 1934; Schelling, 1960; Gibbons, 1992). A Stackelberg game is one in which a distinguished leader publicly commits to a strategy and a follower best responds to it, resulting in a bilevel optimization problem. As with a public belief game, when a Stackelberg game is two player zero sum, Stackelberg equilibrium coincides with Nash equilibrium *for the leader*, but the follower's best response is generally highly exploitable in the game without public commitments. While there exist tie-breaking procedures in Stackelberg literature (e.g., strong or weak Stackelberg equilibrium), they *do not* resolve the non-correspondence issue.

10 CONCLUSION AND FUTURE WORK

In this work, we provided a reduction from computing regularized equilibria of 2p0s games to computing regularized equilibria of PuB-AMGs. We see this contribution as resolving an important gap in literature between common-payoff games and 2p0s games.

We see three impactful directions for future work. The first involves comparing a high performance implementation of a regularized-objective-in-the-PuB-AMG approach to expert iteration (Anthony et al., 2017) to those of existing approaches (Brown et al., 2020; Schmid et al., 2021); while we have shown here that a a regularized-objective-in-the-PuB-AMG approach possesses favorable properties in comparison to ReBeL (Brown et al., 2020) and Player of Games (Schmid et al., 2021), verifying that these advantages manifest in practice would be a valuable contribution. Second, by providing a simpler approach to working with PBSs in 2p0s games, our work provides further motivation for developing new approaches for approximating PBSs at scale; while Sokota et al. (2022b) recently made progress in this direction by showing that fine-tuning can effectively approximate PBSs, amortized approximation of PBSs remains an open problem. Finally, third, we believe it may be possible to extend some weaker form of the results from our work to general sum settings.

---

[4]Though Schmid et al. (2021) show that certain approximate value functions can be made continuous.

REFERENCES

Mohammad Afshari and Aditya Mahajan. Team optimal decentralized state estimation. In *2018 IEEE Conference on Decision and Control (CDC)*, pp. 5044–5050, 2018. doi: 10.1109/CDC.2018.8619493.

Thomas Anthony, Zheng Tian, and David Barber. Thinking fast and slow with deep learning and tree search. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 5366–5376, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Jalal Arabneydi and Aditya Mahajan. Team optimal control of coupled subsystems with mean-field sharing. In *53rd IEEE Conference on Decision and Control*, pp. 1669–1674, 2014. doi: 10.1109/CDC.2014.7039639.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2nd edition, 2000. ISBN 1886529094.

Noam Brown and Tuomas Sandholm. Libratus: The superhuman ai for no-limit poker. In Carles Sierra (ed.), *IJCAI*, pp. 5226–5228. ijcai.org, 2017a. ISBN 978-0-9992411-0-3. URL http://dblp.uni-trier.de/db/conf/ijcai/ijcai2017.html#BrownS17.

Noam Brown and Tuomas Sandholm. Safe and nested subgame solving for imperfect-information games. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 689–699, Red Hook, NY, USA, 2017b. Curran Associates Inc. ISBN 9781510860964.

Noam Brown, Tuomas Sandholm, and Brandon Amos. Depth-limited solving for imperfect-information games. In *NeurIPS*, 2018.

Noam Brown, Anton Bakhtin, Adam Lerer, and Qucheng Gong. Combining deep reinforcement learning and search for imperfect-information games. *Advances in Neural Information Processing Systems*, 33, 2020.

Olivier Buffet, Jilles Dibangoye, Aurélien Delage, Abdallah Saffidine, and Vincent Thomas. On Bellman's Optimality Principle for zs-POSGs. working paper or preprint, December 2020. URL https://hal.inria.fr/hal-03080287.

Neil Burch, Michael Johanson, and Michael Bowling. Solving imperfect information games using decomposition. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, pp. 602–608. AAAI Press, 2014.

Luca Carminati, Federico Cacciamani, Marco Ciccone, and Nicola Gatti. Public information representation for adversarial team games, 2022a. URL https://arxiv.org/abs/2201.10377.

Luca Carminati, Federico Cacciamani, Marco Ciccone, and Nicola Gatti. A marriage between adversarial team games and 2-player games: Enabling abstractions, no-regret learning, and subgame solving. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2638–2657. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/carminati22a.html.

Shicong Cen, Yuting Wei, and Yuejie Chi. Fast policy extragradient methods for competitive games with entropy regularization. In *NeurIPS*, 2021.

Aurélien Delage, Olivier Buffet, and Jilles Dibangoye. HSVI fo zs-POSGs using Concavity, Convexity and Lipschitz Properties. 37 pages, 4 figures, 4 tables, 3 algorithms, October 2021. URL https://hal.inria.fr/hal-03523399.

Jilles Dibangoye and Olivier Buffet. Learning to act in decentralized partially observable MDPs. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1233–1242. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/dibangoye18a.html.

Jilles S. Dibangoye, Christopher Amato, Olivier Buffet, and Francçois Charpillet. Exploiting separability in multiagent planning with continuous-state mdps. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, AAMAS '14, pp. 1281–1288, Richland, SC, 2014a. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450327381.

Jilles S. Dibangoye, Olivier Buffet, and François Charpillet. Error-bounded approximations for infinite-horizon discounted decentralized pomdps. In *Machine Learning and Knowledge Discovery in Databases*, pp. 338–353, Berlin, Heidelberg, 2014b. Springer-Verlag. ISBN 978-3-662-44847-2. doi: 10.1007/978-3-662-44848-9_22. URL https://doi.org/10.1007/978-3-662-44848-9_22.

Jilles Steeve Dibangoye, Chris Amato, Arnaud Doniec, and François Charpillet. Producing efficient error-bounded solutions for transition independent decentralized mdps. In *AAMAS*, 2013a.

Jilles Steeve Dibangoye, Christopher Amato, Olivier Buffet, and François Charpillet. Optimally solving dec-pomdps as continuous-state mdps. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, pp. 90–96. AAAI Press, 2013b. ISBN 9781577356332.

Arnaud Fickinger, Hengyuan Hu, Brandon Amos, Stuart Russell, and Noam Brown. Scalable online planning via reinforcement learning fine-tuning. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=D0xGh031I9m.

Jakob Foerster, Francis Song, Edward Hughes, Neil Burch, Iain Dunning, Shimon Whiteson, Matthew Botvinick, and Michael Bowling. Bayesian action decoder for deep multi-agent reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1942–1951. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/foerster19a.html.

Mukul Gagrani and Ashutosh Nayyar. Thompson sampling for some decentralized control problems. In *2018 IEEE Conference on Decision and Control (CDC)*, pp. 1053–1058, 2018. doi: 10.1109/CDC.2018.8619423.

Sam Ganzfried and Tuomas Sandholm. Endgame solving in large imperfect-information games. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '15, pp. 37–45, Richland, SC, 2015. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450334136.

R. Gibbons. *A Primer in Game Theory*. Pearson Education, 1992.

Abhishek Gupta. Existence of team-optimal strategies in teams with countable observation spaces. *IEEE Transactions on Automatic Control*, 66(10):4792–4798, 2021. doi: 10.1109/TAC.2020.3037047.

Karel Horák and Branislav Bošanský. Solving partially observable stochastic games with public observations. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.33012029. URL https://doi.org/10.1609/aaai.v33i01.33012029.

Hsu Kao, Subramanian, and Vijay. Common information based approximate state representations in multi-agent reinforcement learning. *25th International Conference on Artifi- cial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR*, 151, 2022. URL https://par.nsf.gov/biblio/10332971.

Vojtech Kovarík, Martin Schmid, Neil Burch, Michael Bowling, and Viliam Lisý. Rethinking formal models of partially observable multiagent decision making. *CoRR*, abs/1906.11110, 2019. URL http://arxiv.org/abs/1906.11110.

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

Laurent Lessard and Ashutosh Nayyar. Structural results and explicit solution for two-player lqg systems on a finite time horizon. In *52nd IEEE Conference on Decision and Control*, pp. 6542–6549, 2013. doi: 10.1109/CDC.2013.6760924.

Michael Lederman Littman. Algorithms for sequential decision making, 1996.

Liam MacDermed and Charles L. Isbell. Point based value iteration with optimal belief compression for dec-pomdps. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, pp. 100–108, Red Hook, NY, USA, 2013. Curran Associates Inc.

Matej Moravcik, Martin Schmid, Karel Ha, Milan Hladik, and Stephen Gaukrodger. Refining subgames in large imperfect information games. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), Feb. 2016. doi: 10.1609/aaai.v30i1.10033. URL `https://ojs.aaai.org/index.php/AAAI/article/view/10033`.

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*, 356(6337):508–513, 2017. doi: 10.1126/science.aam6960. URL `https://www.science.org/doi/abs/10.1126/science.aam6960`.

Ashutosh Nayyar, Aditya Mahajan, and Demosthenis Teneketzis. Decentralized stochastic control with partial history sharing: A common information approach. *IEEE Transactions on Automatic Control*, 58(7):1644–1658, 2013. doi: 10.1109/TAC.2013.2239000.

Ashutosh Nayyar, Abhishek Gupta, Cédric Langbort, and Tamer Başar. Common information based markov perfect equilibria for stochastic games with asymmetric information: Finite games. *IEEE Transactions on Automatic Control*, 59(3):555–570, 2014. doi: 10.1109/TAC.2013.2283743.

Frans A. Oliehoek. Sufficient plan-time statistics for decentralized pomdps. In *IJCAI*, 2013.

Yi Ouyang, Hamidreza Tavafoghi, and Demosthenis Teneketzis. Dynamic oligopoly games with private markovian dynamics. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pp. 5851–5858, 2015. doi: 10.1109/CDC.2015.7403139.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8525–8535. PMLR, 18–24 Jul 2021. URL `https://proceedings.mlr.press/v139/perolat21a.html`.

Julien Perolat, Bart de Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T. Connor, Neil Burch, Thomas Anthony, Stephen McAleer, Romuald Elie, Sarah H. Cen, Zhe Wang, Audrunas Gruslys, Aleksandra Malysheva, Mina Khan, Sherjil Ozair, Finbarr Timbers, Toby Pohlen, Tom Eccles, Mark Rowland, Marc Lanctot, Jean-Baptiste Lespiau, Bilal Piot, Shayegan Omidshafiei, Edward Lockhart, Laurent Sifre, Nathalie Beauguerlange, Remi Munos, David Silver, Satinder Singh, Demis Hassabis, and Karl Tuyls. Mastering the game of stratego with model-free multiagent reinforcement learning, 2022. URL `https://arxiv.org/abs/2206.15378`.

T.C. Schelling. *The Strategy of Conflict*. Harvard University Press, 1960.

Martin Schmid, Matej Moravcik, Neil Burch, Rudolf Kadlec, Joshua Davidson, Kevin Waugh, Nolan Bard, Finbarr Timbers, Marc Lanctot, Zach Holland, Elnaz Davoodi, Alden Christianson, and Michael Bowling. Player of games. *CoRR*, abs/2112.03178, 2021. URL `https://arxiv.org/abs/2112.03178`.

Jack Serrino, Max Kleiman-Weiner, David C. Parkes, and Joshua B. Tenenbaum. Finding friend and foe in multi-agent games. In *NeurIPS*, 2019.

Trey Smith and Reid Simmons. Heuristic search value iteration for pomdps. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pp. 520–527, Arlington, Virginia, USA, 2004. AUAI Press. ISBN 0974903906.

Samuel Sokota. Solving common-payoff games with approximate policy iteration. Master's thesis, University of Alberta, 2020.

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. 2021.

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, 2022a. URL https://arxiv.org/abs/2206.05825.

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*, 2022b. URL https://openreview.net/forum?id=ckZY7DGa7FQ.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Hamidreza Tavafoghi, Yi Ouyang, and Demosthenis Teneketzis. On stochastic dynamic games with delayed sharing information structure. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 7002–7009, 2016. doi: 10.1109/CDC.2016.7799348.

Hamidreza Tavafoghi, Yi Ouyang, and Demosthenis Teneketzis. A sufficient information approach to decentralized decision making. In *2018 IEEE Conference on Decision and Control (CDC)*, pp. 5069–5076. IEEE Press, 2018. doi: 10.1109/CDC.2018.8619040. URL https://doi.org/10.1109/CDC.2018.8619040.

Marcos M. Vasconcelos and Nuno C. Martins. The structure of optimal communication policies for remote estimation over the collision channel with private and common observations. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pp. 320–326, 2016. doi: 10.1109/CDC.2016.7798289.

H. von Stackelberg. *Marktform und Gleichgewicht*. Springer, Vienna, 1934.

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 *ECAI*, 2016.

Sihan Zeng, Thinh T. Doan, and Justin Romberg. Regularized gradient descent ascent for two-player zero-sum markov games, 2022. URL https://arxiv.org/abs/2205.13746.

B. Zhang, Gabriele Farina, and Tuomas Sandholm. Team belief dag form: A concise representation for team-correlated game-theoretic decision making. *ArXiv*, abs/2202.00789, 2022.

K. Zhang, Erik Miehling, and Tamer Başar. Online planning for decentralized stochastic control with partial history sharing. *2019 American Control Conference (ACC)*, pp. 3544–3550, 2019.

A DEFINITIONS

First, we formalize our definition of the correspondence mapping $\Pi^{\downarrow}$ in Algorithm 1 below.

Algorithm 1 Correspondence Mapping
---
procedure $\Pi^{\downarrow}(\tilde{\pi})$
    queue $\leftarrow [\tilde{s}^0]$
    $\pi \leftarrow \{\}$
    while len(queue) > 0 **do**
        $\tilde{s} \leftarrow$ queue.pop()
        $\tilde{a} \leftarrow \tilde{\pi}(\tilde{s})$ ▷ Assume $\tilde{\pi}$ is deterministic for reachable $\tilde{s}$.[5]
        for $h \in \mathrm{supp}(\tilde{s})$ **do**
            $\pi(h_\iota) = \tilde{a}(h_\iota)$
        end for
        for $\tilde{s}' \in \mathrm{supp}(\tilde{\mathcal{T}}(\tilde{s}, \tilde{a}))$ **do**   ▷ Continue along states reachable by $\tilde{\pi}$.
            queue.append($\tilde{s}'$)
        end for
    end while
    for untouched h_ι **do**
        Set $\pi(h_\iota)$ arbitrarily.   ▷ These information states are unreachable.
    end for
    return π
end procedure

Next, we define a canonical choice function $\Pi^{\uparrow}$, which maps each joint policy to a corresponding PuB-AMG policy.

Algorithm 2 Canonical Choice Mapping

1: **procedure** $\Pi^{\uparrow}(\pi)$
2:     $\tilde{\pi} \leftarrow \{\}$
3:     **for** all $\tilde{s}$ **do**
4:         $\tilde{a} \leftarrow \{\}$
5:         **for** $h \in \mathrm{supp}(\tilde{s})$ **do**
6:             $\tilde{a}(h_\iota) = \pi(h_\iota)$ ▷ Do what π does.
7:         **end for**
8:         $\tilde{\pi}(\tilde{s}) = \tilde{a}$
9:     **end for**
10:     **return** $\tilde{\pi}$
11: **end procedure**

In short, Algorithm 2 yields a PuB-AMG policy in which the agents play according to π irrespective of the public belief. Therefore, we have that $\mathfrak{R}(\pi) = \tilde{\mathfrak{R}}(\Pi^{\uparrow}(\pi))$. Note that, in contrast to the correspondence mapping, the canonical choice mapping is invariant to opponent policy. Thus, we also allow $\Pi^{\uparrow}$ to be applied directly to individual player policies.

We also introduce some additional definitions.

Definition A.1. *For a minimax objective $\mathfrak{J}$, the value of the game under $\mathfrak{J}$ is $\max_{\pi'_0} \min_{\pi'_1} \mathfrak{J}(\pi'_0, \pi'_1) = \min_{\pi'_1} \max_{\pi'_0} \mathfrak{J}(\pi'_0, \pi'_1)$.*

Remark 4. *Uniqueness-guaranteeing objectives guarantee a well-defined value. This follows immediately from the fact that both players can guarantee the unique equilibrium value.*

Definition A.2. *For a minimax objective $\mathfrak{J}$, the best response value to π_0 under $\mathfrak{J}$ is $\min_{\pi'_1} \mathfrak{J}(\pi_0, \pi'_1)$; analogously, the best response value to π_1 under $\mathfrak{J}$ is $\max_{\pi'_0} \mathfrak{J}(\pi'_0, \pi_1)$. We denote the best response*

---

[5]This assumption is not required, but makes for cleaner presentation.

to π_i as $BR(\pi_i)$. A policy is part of a Nash equilibrium if the best response value to it is equal to the value of the game.

Definition A.3. *For a minimax objective $\mathfrak{J}$, the exploitability π under $\mathfrak{J}$ is:*

$$\frac{-\min_{\pi'_1} \mathfrak{J}(\pi_0, \pi'_1) + \max_{\pi'_0} \mathfrak{J}(\pi'_0, \pi_1)}{2}.$$

A joint policy is a Nash equilibrium if it has exploitability zero.

Definition A.4. *For a minimax objective $\mathfrak{J}$ induced by*

$$\mathfrak{R} \colon (h, a, \delta) \mapsto \begin{cases} \mathcal{R}(h, a) - \psi(\delta, h_\iota) & \iota = 0 \\ \mathcal{R}(h, a) + \psi(\delta, h_\iota) & \iota = 1, \end{cases}$$

the action value for action a at AOH h_ι^t under joint policy π is

$$Q(h_\iota, a) = (-1)^{\mathbb{I}[\iota=1]} \mathbb{E}\left[\mathfrak{R}(H^t, A^t, a \mapsto \mathbb{I}[a^t = a]) + \sum_{t' > t}^{T} \mathfrak{R}(H^{t'}, A^{t'}, \pi(A^{t'})) \mid \pi, h_\iota^t, a^t \right].$$

In words, it is the expected future value to the acting player for taking a at $h_\iota^{t'}$ assuming that both players have played according to π up until now and will continue to play according to π hereinafter.

B THEORY

We now detail the proofs of our theoretical results.
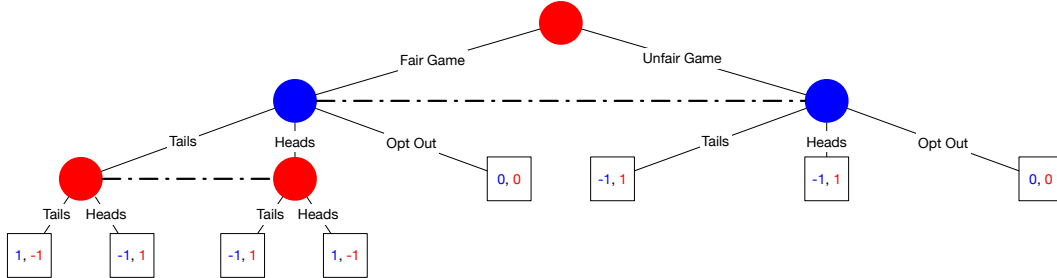
B.1 PROPOSITION 1

Figure 5: A rigged and adversarial variant of matching pennies.

Proposition 1 A PuB-AMG Nash equilibrium $\tilde{\pi}$ may correspond with a joint policy $\Pi^\downarrow(\tilde{\pi})$ that is maximally exploitable.

Proof. We show that this worst case can be realized in the 2p0s rigged adversarial matching pennies game depicted in Figure 5. The game starts with two options for the red player: it can either decide to make the game fair or to rig the game. Then, without having observed the red player's decision, the blue player decides whether to opt out of the game altogether, in which case both players receive a payout of 0, or to play adversarial matching pennies. If the blue player opts in and the game is rigged, the red player receives a payout of 1 independent of the blue player's selection. If the blue player opts in and the game is not rigged, the blue player receives a payout of 1 if the players select the same side of the coin; otherwise, if the players selected opposite sides of the coin, the red player receives a payout of 1.

In the game, the blue player's only Nash equilibrium strategy is to opt out with probability one. The red player's Nash equilibria strategies require at least one of i) rigging the game with probability one and ii) mixing 50-50 between tails and heads.

Now, consider the following PuB-AMG policy, where superscript denotes time step within the game:

- $\tilde{\pi}^0(\emptyset) = (\text{Fair}, \text{Unfair}) \mapsto (1, 0)$.

- $\tilde{\pi}^1(\tilde{\pi}^0) = \begin{cases} (\text{Tails}, \text{Heads}, \text{Out}) \mapsto (1/2, 1/2, 0) & \tilde{\pi}^0(\text{Fair}) = 1 \\ (\text{Tails}, \text{Heads}, \text{Out}) \mapsto (0, 0, 1) & \text{otherwise.} \end{cases}$

- $\tilde{\pi}^2(\tilde{\pi}^0, \tilde{\pi}^1) = \begin{cases} (\text{Tails}, \text{Heads}) \mapsto (0, 1) & \tilde{\pi}^1(\text{Tails}) \geq 1/2 \\ (\text{Tails}, \text{Heads}) \mapsto (1, 0) & \text{otherwise.} \end{cases}$

We claim that $\tilde{\pi}$ is a PuB-AMG Nash equilibrium. To see this, first consider that the expected return is 0: the red player always opts in, the blue player mixes evenly between heads and tails, and the red player always selects tails. Next consider that the red player has no incentive to deviate to an unfair game, because the blue player will opt out, yielding an expected return of zero. Also consider the blue player has no incentive to place additional mass on opting out, as it yields an expected return of zero. Furthermore, the blue player has no incentive to select a different mixture of heads and tails, as doing so will decrease its expected return since the red player best responds at the final time step. Lastly, consider that the red player is best responding at the final time step and, therefore, has no incentive to deviate.

Then, consider that the corresponding policy $\pi = \Pi^\downarrow(\tilde{\pi})$ is as follows:

- π^0: $(\text{Fair}, \text{Unfair}) \mapsto (1, 0)$.

- π^1: $(\text{Tails}, \text{Heads}, \text{Out}) \mapsto (1/2, 1/2, 0)$.

- π^2: $(\text{Tails}, \text{Heads}) \mapsto (0, 1)$.

We claim that this policy is maximally exploitable. To see this, consider that a red player that always rigs the game achieves an expected return of one against the blue player's policy, and consider that a blue player that always selects heads achieves an expected return of one against the red player's policy. $\qquad\square$

B.2 THEOREM 1: CORRESPONDENCE OF UNIQUENESS-GUARANTEEING EQUILIBRIA

To prove Threom 1, we first require some lemmas.

Lemma 1. *The best response value to π_i in the original game is equal to the best response value of $\Pi^\uparrow(\pi_i)$ in PuB-AMG.*

Proof. This follows because player $-i$ has no mechanism to exploit $\Pi^\uparrow(\pi_i)$ beyond that of the original game, since $\Pi^\uparrow(\pi_i)$ ignores belief information.

More formally, consider

$$
\begin{aligned}
\min_{\tilde{\pi}'_1} \tilde{\mathfrak{J}}(\Pi^\uparrow(\pi_0), \tilde{\pi}'_1) &= \tilde{\mathfrak{J}}(\Pi^\uparrow(\pi_0), \text{BR}(\Pi^\uparrow(\pi_0))) \\
&= \mathfrak{J}(\Pi^\downarrow(\Pi^\uparrow(\pi_0), \text{BR}(\Pi^\uparrow(\pi_0)))) \\
&= \mathfrak{J}(\Pi^\downarrow(\Pi^\uparrow(\pi_0), \text{BR}(\Pi^\uparrow(\pi_0)))_0, \Pi^\downarrow(\Pi^\uparrow(\pi_0), \text{BR}(\Pi^\uparrow(\pi_0)))_1) \\
&= \mathfrak{J}(\pi_0, \Pi^\downarrow(\Pi^\uparrow(\pi_0), \text{BR}(\Pi^\uparrow(\pi_0)))_1) \\
&\geq \mathfrak{J}(\pi_0, \text{BR}(\pi_0)) \\
&= \min_{\pi'_1} \mathfrak{J}(\pi_0, \pi'_1).
\end{aligned}
$$

The first equality follows by definition of the best response function BR. The second equality because $\Pi^\downarrow$ preserves expected return. The third equality is notational expansion. The fourth equality follows because π_0 and $\Pi^\downarrow(\Pi^\uparrow(\pi_0), \text{BR}(\Pi^\uparrow(\pi_0)))_0$ can only differ at AOHs that are not reached when playing against $\Pi^\downarrow(\Pi^\uparrow(\pi_0), \text{BR}(\Pi^\uparrow(\pi_0)))_1$. The inequality and final equality follow by definition of best response.

Also, consider

$$
\begin{aligned}
\min_{\pi_1'} \mathfrak{J}(\pi_0, \pi_1') &= \mathfrak{J}(\pi_0, \mathrm{BR}(\pi_0)) \\
&= \tilde{\mathfrak{J}}(\Pi^\uparrow(\pi_0), \Pi^\uparrow(\mathrm{BR}(\pi_0))) \\
&\geq \tilde{\mathfrak{J}}(\Pi^\uparrow(\pi_0), \mathrm{BR}(\Pi^\uparrow(\pi_0))) \\
&= \min_{\tilde{\pi}_1'} \tilde{\mathfrak{J}}(\Pi^\uparrow(\pi_0), \tilde{\pi}_1).
\end{aligned}
$$

The first equality follows by definition of the best response function BR. The second equality follows because $\Pi^\uparrow$ preserves expected return. The inequality and final equality follows by definition of best response.

These two inequalities can only be true if $\min_{\pi_1'} \tilde{\mathfrak{J}}(\Pi^\uparrow(\pi_0), \pi_1') = \min_{\pi_1'} \mathfrak{J}(\pi_0, \pi_1')$. An analogous argument shows the same result for π_1. $\qquad\square$

Corollary 2. *The exploitability of π in the original game is equal to the exploitability of $\Pi^\uparrow(\pi)$ in PuB-AMG.*

Proof. This follows immediately from Lemma 1 and the fact that exploitability is defined in terms of best response values. $\qquad\square$

Corollary 3. *The value of the PuB-AMG under a uniqueness-guaranteeing objective is well-defined and equal to that of the original game.*

Proof. Note that it suffices to show that PuB-AMGs are guaranteed to have an equilibrium with the same expected return as the equilibrium of the original game. Then consider $\tilde{\pi} = \Pi^\uparrow(\pi)$, where π is an equilibrium. Then, since π is an equilibrium and, per Corollary 2, $\Pi^\uparrow$ preserves exploitability, $\tilde{\pi}$ is an equilibrium. Additionally, since $\Pi^\uparrow$ preserves expected return, the original game and the PuB-AMG possess equilibria π and $\tilde{\pi}$, respectively, that yield the same expected return. $\qquad\square$

We are now ready to prove our two main lemmas.

Lemma 2. *Let π be the equilibrium of a uniqueness-guaranteeing objective. Let $\tilde{s}$ define a subgame of the original game induced by playing π for some number of steps. Then the unique equilibrium of the subgame $\pi^{\tilde{s}}$, considered as an independent game, is the restriction $\pi^{|\tilde{s}}$ of π to the subgame.*

Proof. If $\pi^{|\tilde{s}}$ is not an equilibrium of the subgame, then it must be exploitable in the subgame. This means that either

$$
\min_{\pi_1'} \mathfrak{J}^{\tilde{s}}(\pi_0^{|\tilde{s}}, \pi_1') < \mathfrak{J}^{\tilde{s}}(\pi_0^{|\tilde{s}}, \pi_1^{|\tilde{s}}) \quad \text{or} \quad \mathfrak{J}^{\tilde{s}}(\pi_0^{|\tilde{s}}, \pi_1^{|\tilde{s}}) < \max_{\pi_0'} \mathfrak{J}^{\tilde{s}}(\pi_0', \pi_1^{|\tilde{s}})
$$

Without loss of generality, assume the former. Let $\pi_1^{\mathrm{br}} = \mathrm{argmin}_{\pi_1'} \mathfrak{J}^{\tilde{s}}(\pi^{|\tilde{s}}, \pi_1')$; let $\mathcal{P}^\pi(\tilde{s})$ represent the probability of reaching $\tilde{s}$ using policy π; let t be the time step corresponding to $\tilde{s}$ and let $\mathcal{J}^{<t}$ denote the expected return prior to time t. Further, let $\tilde{s}' \neq \tilde{s}$ range over the possible subgames entered at time t if $\tilde{s}$ is not entered. Then

$$
\begin{aligned}
\mathfrak{J}(\pi_0, \pi_1) &= \mathfrak{J}^{<t}(\pi_0, \pi_1) + \mathcal{P}^\pi(\tilde{s}) \mathfrak{J}^{\tilde{s}}(\pi_0^{|\tilde{s}}, \pi_1^{|\tilde{s}}) + \sum_{\tilde{s}' \neq \tilde{s}} \mathcal{P}^\pi(\tilde{s}') \mathfrak{J}^{\tilde{s}'}(\pi_0^{|\tilde{s}'}, \pi_1^{|\tilde{s}'}) \\
&> \mathfrak{J}^{<t}(\pi_0, \pi_1) + \mathcal{P}^\pi(\tilde{s}) \mathfrak{J}^{\tilde{s}}(\pi_0^{|\tilde{s}}, \pi_1^{\mathrm{br}}) + \sum_{\tilde{s}' \neq \tilde{s}} \mathcal{P}^\pi(\tilde{s}') \mathfrak{J}^{\tilde{s}'}(\pi_0^{|\tilde{s}'}, \pi_1^{|\tilde{s}'}) \\
&= \mathfrak{J}(\pi_0, [\pi_1^{\mathrm{br}}, \pi_1^{|-\tilde{s}}]).
\end{aligned}
$$

Here, the first line decomposes the expected return into 1) that which is accrued prior to time t, 2) that which is accrued in subgame $\tilde{s}$, and 3) that which is accrued after time t outside of subgame $\tilde{s}$. The second line invokes our assumption that $\pi_1^{|\tilde{s}}$ does not achieve the best response value against $\pi_0^{|\tilde{s}}$ and $\mathcal{P}^\pi(\tilde{s}) > 0$. The third line re-assembles the expected return, where we use $[\pi_1^{\mathrm{br}}, \pi_1^{|-\tilde{s}}]$ to denote a policy that plays π_1 outside $\tilde{s}$ and π_1^{br} inside $\tilde{s}$.

In total, we have shown that if $\pi^{|\tilde{s}}$ is not an equilibrium in the subgame induced by $\tilde{s}$, then π is not an equilibrium because π_1 is not a best response. Thus, $\pi^{|\tilde{s}}$ must be an equilibrium of the subgame. Therefore, because $\mathfrak{J}$ is uniqueness guaranteeing, we must have $\pi^{\tilde{s}} = \pi^{|\tilde{s}}$. $\qquad\square$

Lemma 3. *If $\tilde{\pi}$ is an equilibrium of the PuB-AMG, then the decision rule for the first time step $\Pi^{\downarrow}(\tilde{\pi})^0$ must be part of an equilibrium policy in the original game.*

Proof. Without loss of generality, assume that $\iota = 0$ at the first time step. Also, use π_0^{-0} to denote the part of π_0 that is relevant after the first time step. Also, let $[\pi'^0, \pi_0'^{-0}]$ denote a policy for $i = 0$ that plays according to π'^0 at the first time step and $\pi_0'^{-0}$ otherwise. Let $\tilde{\pi}$ be an equilibrium of the PuB-AMG. Then observe

$$\max_{\pi_0'} \min_{\pi_1'} \mathfrak{J}(\pi_0', \pi_1') = \max_{\tilde{\pi}_0'} \min_{\tilde{\pi}_1'} \tilde{\mathfrak{J}}(\tilde{\pi}_0', \tilde{\pi}_1') \tag{1}$$

$$= \max_{\tilde{\pi}_0'^{-0}} \min_{\tilde{\pi}_1'} \tilde{\mathfrak{J}}([\tilde{\pi}^0, \tilde{\pi}_0'^{-0}], \tilde{\pi}_1) \tag{2}$$

$$= \max_{\tilde{\pi}_0'^{-0}} \min_{\tilde{\pi}_1'} \mathfrak{J}(\Pi^{\downarrow}([\tilde{\pi}^0, \tilde{\pi}_0'^{-0}], \tilde{\pi}_1')) \tag{3}$$

$$= \max_{\tilde{\pi}_0'^{-0}} \min_{\tilde{\pi}_1'} \mathfrak{J}([\Pi^{\downarrow}(\tilde{\pi})^0, \Pi^{\downarrow}([\tilde{\pi}^0, \tilde{\pi}_0'^{-0}], \tilde{\pi}_1')_0^{-0}], \Pi^{\downarrow}([\tilde{\pi}^0, \tilde{\pi}_0'^{-0}], \tilde{\pi}_1')_1) \tag{4}$$

$$= \max_{\pi_0'^{-0}} \min_{\pi_1'} \mathfrak{J}([\Pi^{\downarrow}(\tilde{\pi})^0, \pi_0'^{-0}], \pi_1') \tag{5}$$

$$= \min_{\pi_1'} \mathfrak{J}([\Pi^{\downarrow}(\tilde{\pi})^0, \arg\max_{\pi_0'^{-0}} \min_{\pi_1''} \mathfrak{J}([\Pi^{\downarrow}(\tilde{\pi})^0, \pi_0'^{-0}], \pi_1'')], \pi_1'). \tag{6}$$

Here, the first equality follows by Corollary 3; the second equality follows because $\tilde{\pi}^0$ is part of an equilibrium; the third equality follows because $\tilde{\mathfrak{J}}(\tilde{\pi}') = \mathfrak{J}(\Pi^{\downarrow}(\tilde{\pi}'))$; the fourth line equality follows because the image of the correspondence mapping for the first time step is invariant to the PuB-AMG policy at later time steps; the fifth line follows because each player can express any policy in the original game through $\Pi^{\downarrow}$, up to reachability, and because changes over unreachable AOHs do not change the expected return; the sixth line follows because the evaluation of an argmax is equal to the max.

This chain of equalities shows that the best response value to

$$[\Pi^{\downarrow}(\tilde{\pi})^0, \arg\max_{\pi_0'^{-0}} \min_{\pi_1''} \mathfrak{J}([\Pi^{\downarrow}(\tilde{\pi})^0, \pi_0'^{-0}], \pi_1'')]$$

is equal to the value of the game. Thus, $\Pi^{\downarrow}(\tilde{\pi})^0$ is part of an equilibrium. $\qquad\square$

Theorem 1 *If $\tilde{\pi}$ is an equilibrium of the PuB-AMG induced by a uniqueness-guaranteeing objective, then its projection $\Pi^{\downarrow}(\tilde{\pi})$ is an equilibrium in the original game.*

Proof. Lemma 3 shows this to be true for the first time step. Now assume this is true up to time step t and consider time step $t + 1$. Then, for a particular reachable $\tilde{s}^{t+1}$, the PuB-AMG subgame starting at this point is the PuB-AMG of the subgame of the original game starting from $\tilde{s}^{t+1}$. Thus, the PuB-AMG strategy for $\tilde{s}^{t+1}$ must correspond to an equilibrium of the subgame of the original game, as per Lemma 3. Furthermore, because the minimax objective is uniqueness guaranteeing, the equilibrium strategy of the subgame of the original game must the unique restriction of the equilibrium of the original game to that subgame, as per Lemma 2. $\qquad\square$

B.3 Theorem 3: Sufficient Conditions for Uniqueness Guaranteeing

Theorem 3 Consider an objective induced by

$$\mathfrak{R}\colon h, a, \delta \mapsto \begin{cases} \mathcal{R}(h, a) - \psi(\delta, h_\iota) & \iota = 0 \\ \mathcal{R}(h, a) + \psi(\delta, h_\iota) & \iota = 1 \end{cases}$$

and define a policy greedification function

$$g\colon q, h_\iota \mapsto \arg\max_{\delta \in \Delta(\mathcal{A})} \langle \delta, q \rangle - \psi(\delta, h_\iota) \text{ where } g\colon [-\mathfrak{M}, \mathfrak{M}]^{|\mathbb{A}|} \times \mathbb{H}_\iota \to \Delta(\mathbb{A})$$

and where $\mathfrak{M}$ is the maximum of the absolute values of the returns of the objective. Then, if, for all h_ι, $\psi(\cdot, h_\iota)$ is continuous and $g(\cdot, h_\iota)$ is i) well defined, ii) continuous, and iii) has an interior image, the objective $\mathfrak{J}$ is uniqueness guaranteeing.

Proof. First, we show that such an equilibrium is guaranteed to exist. Let $\mathcal{F}\colon [-\mathfrak{M}, \mathfrak{M}]^{|\mathcal{H}_\iota||\mathcal{A}|} \to [-\mathfrak{M}, \mathfrak{M}]^{|\mathcal{H}_\iota||\mathcal{A}|}$ be a function that maps each vector $[q_{h_\iota}]_{h_\iota}$ to the action-value vector for the joint policy dictated by the application of g to $[(q_{h_\iota}, h_\iota)]_{h_\iota}$. Note that $\mathcal{F}$ is well defined—i.e., the ensuant action values are always well defined—because g maps to the interior, so every history is reached with positive probability. Also note that this function is continuous, by the continuity of g and ψ, and single valued because g is single valued. Thus, because $[-\mathfrak{M}, \mathfrak{M}]^{|\mathcal{H}_\iota||\mathcal{A}|}$ is compact and convex, by Brouwer's fixed point theorem, a fixed point must exist. The policy corresponding to these fixed-point action values is an equilibrium. This follows because, by backward induction, each player is optimally responding to the other, holding the other fixed.

Now we show that there is a unique equilibrium. Note that, for any fixed opponent, the optimal policy at any decision point reached with positive probability must be full support because g's image is within the interior. By forward induction, this means that every equilibrium must be full support at every decision point. Now, note that, by backward induction, the best responses to full support policies are unique because g is single valued with an interior image. In aggregate, these two things show that any equilibrium is strict—i.e., the only best response to one part of the equilibrium is the other part of the equilibrium. Now, assume there exist two distinct equilibria π and π'. Without loss of generality, assume that π_0 performs at least as well as π'_0 against π_1. If π_0 performs equally well, there is a contradiction because π'_0 is not the unique best response. If π_0 outperforms π', there is a contradiction because π' is not at equilibrium. Thus, the equilibrium must be unique.

The result follows because this proof also holds for every subgame of the original game. $\square$

C EXPERIMENTS

C.1 MAGNETIC MIRROR DESCENT

In our experiments, we use magnetic mirror descent (MMD) (Sokota et al., 2022a) as our game solver. In the instance of MMD we use, updates are of the form

$$\pi_{t+1} = \text{argmax}_\pi \mathbb{E}_{A\sim\pi} q_{\pi_t}(A) + \alpha\mathcal{H}(\pi) - \frac{1}{\eta}\text{KL}(\pi, \pi_t) \tag{7}$$

where π_t is the current policy and q_t is the MiniMaxEnt Q-value vector for time t. This update possesses the closed form

$$\pi_{t+1} \propto \left[\pi_t e^{\eta q_{\pi_t}}\right]^{\frac{1}{1+\alpha\eta}}. \tag{8}$$

The fixed point of equation (8) is a policy satisfying

$$\pi_* = \arg\max_\pi \mathbb{E}_{A\sim\pi} q(A) + \alpha\mathcal{H}(\pi) \propto e^{q/\alpha}. \tag{9}$$

C.2 TABULAR PUB-AMG POLICIES

In PuB-AMGs, the state space is continuous. Thus, it may not be possible to express a fully specified PuB-AMG policy in tabular form. We describe how we handle this issue for perturbed rock-paper-scissors and Kuhn poker, respectively, in subsequent subsections.

In both settings, we solve the games using full feedback, meaning that we compute exact Q-values and update the policy for every AOH.

C.2.1 PERTURBED ROCK-PAPER-SCISSORS

In perturbed rock-paper-scissors, the first moving player's state space is trivial; thus, its policy can be expressed exactly. Also, the (regularized) best response of the second player can be computed in closed form using equation. Thus, we set the second moving player's (regularized) PuB-AMG
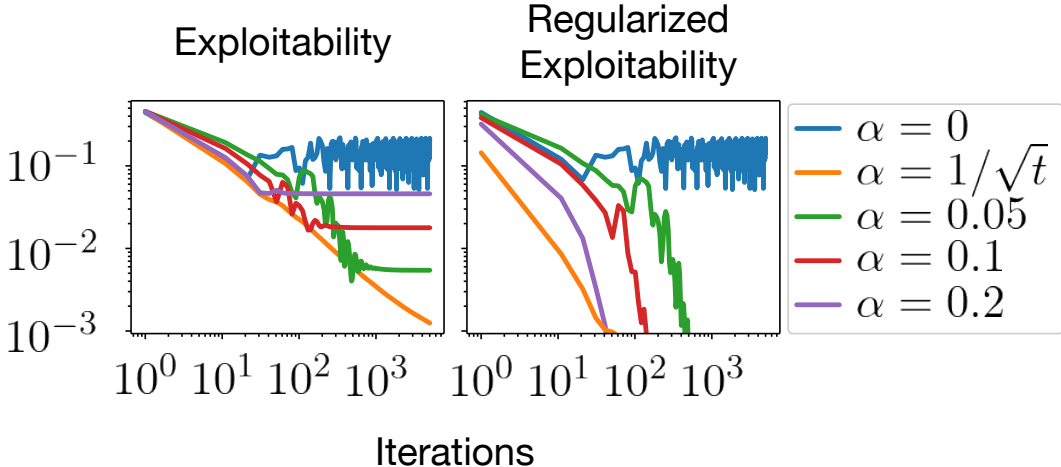
Figure 6: Results for Kuhn poker.

Nash equilibrium policy to induced by equation (9). We update the first moving player's policy using equation (8) where q_t is the feedback induced by the second moving player's (regularized) PuB-AMG Nash equilibium policy.

C.2.2 KUHN POKER

We also investigate MiniMaxEnt objectives in an extensive-form game—Kuhn poker. In Kuhn poker, there are up to three time steps. For the third time step, we use the (regularized) Pub-AMG Nash equilibrium policy induced by equation (9). For the first time step, at each iteration, we update the policy at each information state using MMD on the feedback from the previous time step. For the second time step, at iteration t, holding fixed the iteration t decision rule for the first time step, we use the policy induced by performing $\sqrt{t}$ iterations of MMD against the (regularized) PuB-AMG Nash equilibrium policy of the third time step. As $\sqrt{t}$ grows large, we expect the decision rules for the second time step to approximate a PuB-AMG best response.

We show the results for the original game in Figure 6.[6] Qualitatively, they are analogous to those from the perturbed rock-paper-scissors game. The unregularized objective induces high exploitability iterates (blue) that do not converge in the original game; the objectives with fixed regularization (purple, red, greed) converge to constant exploitability and zero regularized exploitability in the original game; the objective with annealed regularization converges to zero exploitability and zero regularized exploitability.

D CONTINUITY IN THE PUB-AMG

Lemma 4. *Let f_1, f_2 be real-valued continuous functions with shared compact domain $\mathcal{X} \times \mathcal{Y}$. Furthermore, assume their max-min values are attained and the following inequality holds for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$:*

$$|f_1(x, y) - f_2(x, y)| < \epsilon.$$

Then it follows that

$$|[\max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} f_1(x, y)] - [\max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} f_2(x, y)]| < \epsilon.$$

---

[6]We omit PuB-AMG (regularized) exploitability, as it is difficult to compute exactly in this case.

Proof. Note that by assumption we have for all (x, y) within the domains of f_1, f_2 it holds

$$f_1(x, y) \leq f_2(x, y) + \epsilon, \tag{10}$$
$$f_2(x, y) \leq f_1(x, y) + \epsilon. \tag{11}$$

Therefore,

$$\max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} f_1(x, y) = \min_{y \in \mathcal{Y}} f_1(x_*, y) \text{ for } x_* \in \arg\max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} f_1(x, y)$$
$$\leq \min_{y \in \mathcal{Y}} f_2(x_*, y) + \epsilon$$
$$\leq \max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} f_2(x, y) + \epsilon.$$

Where the first inequality is due to taking the min of both sides of equation 10. Following the same steps starting with f_2 and using equation 11 gives

$$\max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} f_2(x, y) \leq \max_{x \in \mathcal{X}} \min_{y \in \mathcal{Y}} f_1(x, y) + \epsilon.$$

These two inequalities together yield the result. □

Theorem 4. *The minimax PuB-AMG value function $\tilde{v}_*$ is continuous for any objective $\mathfrak{J}$ for which policy evaluation is continuous.*

Proof. Fix $\epsilon > 0$. Let b and b' differ in total variation distance by less than

$$\delta = \frac{\epsilon}{2\mathfrak{M}}$$

. Then observe that, for a joint policy π, we have that

$$|v_\pi(b) - v_\pi(b')| = |\sum_h b(h) v_\pi(h) - b'(h) v_\pi(h)|$$
$$\leq \sum_h |b(h) v_\pi(h) - b'(h) v_\pi(h)|$$
$$= \mathfrak{M} \sum_h |b(h) - b'(h)|$$
$$< 2\mathfrak{M}\delta$$
$$= \epsilon,$$

where $v_\pi(b)$ is the expected return under $\mathfrak{J}$ to playing π starting from the subgame defined by b.

Then

$$|\tilde{v}_*(b) - \tilde{v}_*(b')| = |[\max_{\tilde{\pi}_0} \min_{\tilde{\pi}_1} \tilde{v}_{\tilde{\pi}}(b)] - [\max_{\tilde{\pi}_0} \min_{\tilde{\pi}_1} \tilde{v}_{\tilde{\pi}}(b')]|$$
$$= |[\max_{\pi_0} \min_{\pi_1} v_\pi(b)] - [\max_{\pi_0} \min_{\pi_1} v_\pi(b')]|$$
$$< \epsilon.$$

The first equality follows by definition of $\tilde{v}_*$. The second equality follows from Corollary 3. The third equality follows from Lemma 4.

□

Corollary 4. *The minimax PuB-AMG value function $\tilde{v}_*$ is continuous for the standard expected return objective.*

Corollary 5. *The minimax PuB-AMG value function $\tilde{v}_*$ is continuous for any objective satisfying premises of Theorem 3.*