The Update Equivalence Framework for Decision-Time Planning

Samuel Sokota [1 †] **Gabriele Farina** [1 2] **David J. Wu** [2] **Hengyuan Hu** [3] **Kevin A. Wang** [2] **J. Zico Kolter** [1 4] **Noam Brown** [2]

Abstract

The process of revising (or constructing) a policy immediately prior to execution—known as decision-time planning—is key to achieving superhuman performance in perfect-information settings like chess and Go. A recent line of work has extended decision-time planning to more general *imperfect-information* settings, leading to superhuman performance in poker. However, this line of work involves subgames whose sizes scale exponentially in the number of bits of non-public information, making them unhelpful when the amount of non-public information is large. Motivated by this issue, we introduce an alternative perspective on decision-time planning: the *framework of update equivalence*. In this framework, decision-time planning algorithms are viewed as implementing updates of synchronous learning algorithms. This enables us to introduce a new family of principled decision-time planning algorithms that do *not* rely on public information, opening the door to sound and effective decision-time planning in settings with large amounts of non-public information. In experiments, members of this family produced comparable or superior results compared to state-of-the-art approaches in Hanabi and improved performance in 3x3 Abrupt Dark Hex and Phantom Tic-Tac-Toe.

1. Introduction

Decision-time planning (DTP) is the process of revising (or even constructing from scratch) a policy immediately prior to using that policy to make a decision. In settings involving strategic decision making, the benefits of DTP can be quite large. For example, while DTP approaches have achieved superhuman performance in chess (Campbell et al., 2002), Go (Silver et al., 2016), and poker (Moravčík et al., 2017; Brown & Sandholm, 2018; 2019), approaches without DTP remain non-competitive with top humans in these domains.

Currently, the dominant conceptual paradigm for DTP is based on the idea of solving (or improving the policy as much as possible for) subgames. In perfect-information games, a subgame is defined naturally as a game beginning from the current history that proceeds according to the same rules as the original game. In contrast, the definition of a subgame is significantly more nuanced in imperfect-information games, where counterfactual dependencies make naively considering subtrees in isolation unsound. While multiple definitions have been proposed, all those that facilitate sound guarantees (Nayyar et al., 2013; Burch et al., 2014; Moravcik et al., 2016; Brown & Sandholm, 2017; Moravčík et al., 2017; Brown et al., 2020; Sokota et al., 2023b) rely on a distribution known as a *public belief state (PBS)*—*i.e.*, the posterior over histories given public information and a historical joint policy.

Unfortunately, PBS-based planning has a fundamental limitation: It is not useful in settings with large amounts of non-public information. This shortcoming arises because the number of decision points in the support of the PBS distribution scales exponentially with the number of bits of non-public information. When the amount of non-public information is small, such as in poker, it is feasible to construct strong policies for all of these decision points. However, as the amount of non-public information grows, this becomes intractable under a fixed time budget. Indeed, in cases where there is no public information, PBS-based subgame solving requires solving the entire game.

In this work, motivated by the insufficiency of subgame-based conceptualizations, we advocate for an alternative perspective on DTP that we call the *framework of update equivalence*. Rather than viewing DTP algorithms as solving subgames, the framework of update equivalence views DTP algorithms as implementing updates of synchronous learning algorithms. Importantly, because these synchronous algorithms needn't involve PBSs, the framework of update equivalence has no inherent limitations arising from the amount of non-public information.

To facilitate the application of the update equivalence framework, we provide a general procedure for converting any

[†]Work done while part time at Meta AI [1]Carnegie Mellon University [2]Meta AI [3]Stanford University [4]Bosch Center for AI. Correspondence to: Samuel Sokota <ssokota@andrew.cmu.edu>.

synchronous learning algorithm using action-value feedback and local updates into a decision-time planner. The resulting approaches can be summarized as using search to estimate the current policy's action values and using the synchronous algorithm's update rule to update the policy. They are simple to implement, highly parallelizable, and can be scaled to settings with large amounts of private information using sequential generative modeling.

Using this procedure, we apply the update equivalence framework to justify two new DTP algorithms that we call *mirror descent update equivalent search (MD-UES)* and *magnetic mirror descent update equivalent search (MMD-UES)*. We test these algorithms in Hanabi (Bard et al., 2020), a fully cooperative game with a substantial amount of public information, and in two adversarial games with virtually no public information—3x3 Abrupt Dark Hex and Phantom Tic-Tac-Toe. In Hanabi, we find that MD-UES yields results that are competitive with or superior to modern PBS-based methods, which are considered state-of-the-art. In 3x3 Abrupt Dark Hex and Phantom Tic-Tac-Toe, we find that MMD-UES substantially lowers the approximate exploitability of a uniform random blueprint policy (i.e., the policy on top of which search is performed) and improves head-to-head performance against a variety of baselines.

Based on these results, we argue that the framework of update equivalence opens the door to sound and effective DTP and expert iteration (Anthony et al., 2017; Anthony, 2021) in settings with large amounts of non-public information.

2. Background and Notation

This section introduces the requisite information on finite-horizon partially observable stochastic games (POSGs), synchronous learning algorithms, and DTP algorithms necessary to discuss the update equivalence framework.

Partially Observable Stochastic Games The set of finite-horizon POSGs (Hansen et al., 2004) is a class of games that is equivalent to perfect recall timeable (Jakobsen et al., 2016) extensive form games (Kovarík et al., 2022). To describe POSGs, we use:

- g to notate the game itself,
- $s \in \mathbb{S}$ to notate Markov states,
- $a_i \in \mathbb{A}_i$ to notate player i's actions,
- $o_i \in \mathbb{O}_i$ to notate player i's observations,
- $h_i \in \mathbb{H}_i = \bigcup_t (\mathbb{O}_i \times \mathbb{A}_i)^t \times \mathbb{O}_i$ to notate i's decision points,
- $a \in \mathbb{A} = \times_i \mathbb{A}_i$ to notate joint actions,
- $h \in \mathbb{H} = \times_i \mathbb{H}_i$ to notate histories,
- $\mathcal{R}_i \colon \mathbb{S} \times \mathbb{A} \to \mathbb{R}$ to notate player i's reward function,
- $\mathcal{T} \colon \mathbb{S} \times \mathbb{A} \to \Delta(\mathbb{S})$ to notate the transition function,
- $\mathcal{O}_i \colon \mathbb{S} \times \mathbb{A} \to \mathbb{O}_i$ to notate player i's observation function.

Each player i interacts with the game via a policy π_i, which maps decision points to distributions over actions

$$\pi_i \colon \mathbb{H}_i \to \Delta(\mathbb{A}_i).$$

Given a joint policy $\pi = (\pi_i)_i$, player i's expected return is

$$\mathcal{J}_i(\pi) := \mathbb{E}\left[\sum_t \mathcal{R}_i(S^t, A^t) \,\middle|\, \pi\right]$$

where we use capital letters to denote random variables.

If the reward function for each player is the same, we say the game is *common payoff*. If there are two players and the reward functions of these players are the negations of one another, we say the game is *two player zero sum*.

We notate the expected value for an agent at a history h^t under joint policy π as

$$v_i^\pi(h^t) = \mathbb{E}\left[\sum_{t' \geq t} \mathcal{R}_i(S^{t'}, A^{t'}) \,\middle|\, \pi, h^t\right].$$

We notate the expected action value for an agent at a history h^t taking action a_i^t under joint policy π as

$$q_i^\pi(h^t, a_i^t) = \mathbb{E}\left[\mathcal{R}_i(S^t, A^t) + v_i^\pi(H^{t+1}) \,\middle|\, \pi, h^t, a_i^t\right].$$

The expected *value of a decision point* h_i^t is the weighted sum of history values, where each history is weighted by its probability, conditioned on the decision point and historical joint policy

$$v_i^\pi(h_i^t) = \mathbb{E}\left[v_i^\pi(H^t) \mid \pi, h_i^t\right].$$

Similarly, the expected *action value for action a_i^t at a decision point h_i^t* is defined as

$$q_i^\pi(h_i^t, a_i^t) = \mathbb{E}\left[q_i^\pi(H^t, a_i^t) \mid \pi, h_i^t\right].$$

Synchronous Learning Algorithms We say a learning algorithm is synchronous if its update function $\mathcal{U}^{\text{sync}} \colon (\pi_t, g) \mapsto \pi_{t+1}$ operates on the entire joint policy π simultaneously. We say a synchronous algorithm operates with local action value feedback if there exists $\mathcal{U}^{\text{sync}}_{\text{local}}$ such that, for each h_i,

$$\mathcal{U}^{\text{sync}}(\pi_t, g)(h_i) = \mathcal{U}^{\text{sync}}_{\text{local}}\left(\pi_t(h_i), q_i^{\pi_t}(h_i, \cdot)\right).$$

In other words, the update at a decision point depends only on the policy and action values at the decision point itself.

Decision-Time Planning Algorithms As mentioned, a DTP algorithm is used to revise (or, in the extreme case, construct altogether) a policy immediately prior to using that policy to make a decision; we call a policy that is to be revised by a DTP algorithm a *blueprint* policy. We can think of a DTP algorithm as implementing an *update function* mapping each decision point to a revised policy,

$$\mathcal{U}^{\text{DTP}} \colon (h_i, *) \mapsto \delta \text{ where } \delta \in \Delta(\mathbb{A}_i)$$

where h_i is the decision point from which planning is being performed and $*$ acts as a stand in for any other information that a DTP algorithm may use, including (but not limited to) a blueprint policy or the transition model of the game.

3. The Framework of Update Equivalence

The framework of update equivalence is built on the observation that DTP algorithms and synchronous algorithms may induce the same updates, as is formalized below.

Definition 3.1 (Update Equivalence)**.** For a game g, a synchronous learning algorithm with updater $\mathcal{U}^{\text{sync}}$ and a DTP algorithm with updater $\mathcal{U}^{\text{DTP}}$ are *update equivalent* if, for any policy π_t and decision point h_i of any player, the synchronous learning algorithm and the decision-time planning algorithm induce the same updates:

$$\mathcal{U}^{\text{sync}}(\pi_t, g)(h_i) = \mathcal{U}^{\text{DTP}}(h_i, *(\pi_t, g)),$$

where $*(\pi_t, g)$ denotes the policy and game specific information used by the DTP algorithm.

There are at least two benefits to considering update equivalence relationships in the context of decision-time planning.

First, it begets an approach to analyzing DTP algorithms via their synchronous learning algorithm counterparts. (Note that, given any DTP algorithm, it is always possible to construct an update equivalent synchronous algorithm—simply define the synchronous algorithm's update to be that of Definition 3.1.) This avenue of analysis circumvents non-locality issues that have traditionally made the analysis of DTP algorithms painful by considering the global change induced by planning across all decision points.

Second, it enables the construction of new DTP algorithms. Specifically, it allows synchronous algorithms with desirable guarantees and locally structured updates to be used to generate new principled DTP algorithms. That these updates possess some local structure is important because DTP operates in a heavily time-constrained regime, making performing arbitrarily unstructured updates intractable.

We illustrate these benefits in the two coming subsections. In Section 3.1 we focus on synchronous learning algorithms that operate using local action-value feedback, providing a general procedure to generate DTP analogues of these algorithms and discussing three DTP algorithms generatable by this procedure, as well as their soundness. In Section 3.2 we also discuss synchronous algorithms that do not operate using local action-value feedback, demonstrating how framework of update equivalence can be used to generate or justify DTP algorithms with more complex structure.

Algorithm 1 Update Equivalent Search for $\mathcal{U}^{\text{sync}}_{\text{local}}$

```
Input: decision point h_i^t, joint policy π
Initialize q̂[a] as running mean tracker for a ∈ A_i
repeat
  Sample history H^t ~ P(H^t | h_i^t, π)
  for a ∈ A_i do
    Sample conditional return G ~ P(G | H^t, a, π)
    q̂[a].update_running_mean(G)
  end for
until search budget is exhausted
return U_local^sync(π(h_i^t), q̂)
```

3.1. Action-Value-Based Planners

The general procedure for converting synchronous algorithms whose update functions are continuous $\mathcal{U}^{\text{sync}}_{\text{local}}$ and operate on local action-value feedback into decision-time planners is given by Algorithm 1. In words, Algorithm 1 repeatedly samples histories conditioned upon the occupied decision point and the joint policy and acquires estimates of the sample returns for these histories via rollouts. When the computational budget available for DTP has been exhausted, Algorithm 1 returns the policy induced by the application of the local updater $\mathcal{U}^{\text{sync}}_{\text{local}}$ to the current joint policy and the estimated action values. As is formalized by the proposition below, decision-time-planning algorithms produced via Algorithm 1 approach update equivalence with the synchronous algorithms that induce them in the limit as computational budget grows large.

Proposition 3.2. *Let $\mathcal{U}^{sync}$ by a synchronous update induced by local action value updates $\mathcal{U}^{sync}_{local}$. Then, as the number of rollouts goes to infinity, the output of Algorithm 1, conditioned on $\mathcal{U}^{sync}_{local}$ and given inputs h_i^t, π, converges in probability to $\mathcal{U}^{sync}(\pi, g)(h_i^t)$.*

This proposition holds because $\hat{q}$ converges in probability to $q_i^\pi(h_i)$ by the law of large numbers and $\mathcal{U}^{\text{sync}}_{\text{local}}$ is continuous.

Policy Iteration Update Equivalent Search Algorithm 1 allows us to relate synchronous learning algorithms with DTP counterparts. A notable existing instance of such a relationship that has already been identified in literature is that between policy iteration and Monte Carlo search (Tesauro, 1995; Tesauro & Galperin, 1996), which is the name of the DTP algorithm resulting from plugging policy iteration's local update function into Algorithm 1. This local update can be written as

$$\mathcal{U}^{\text{sync}}_{\text{local}} \colon (\pi_t, q) \mapsto \delta \text{ where } \delta \in \Delta(\arg\max_a q(a)),$$

and where we abuse notation by using π_t and q to refer to elements of $\Delta(\mathbb{A}_i)$ and $\mathbb{R}^{|\mathbb{A}_i|}$, respectively, rather than the full joint policy and full action-value function. Because policy iteration possesses desirable properties in single-agent settings and two-player zero-sum (2p0s) games with perfect information (Littman, 1996), Monte Carlo search is also a sound approach to such settings.

Mirror Descent Equivalent Search Algorithm 1 also allows us to give novel justification to DTP approaches in settings in which acquiring guarantees has been historically difficult. One example of such as a setting is common-payoff games, where a naive application of Monte Carlo search can lead to bad performance as a result of the posterior over histories induced by the search policy diverging too far from the blueprint policy's posterior (Sokota et al., 2022). We show that this issue can be resolved using mirror descent (Beck & Teboulle, 2003; Nemirovsky & Yudin, 1983), which is a general approach to optimizing objectives that penalizes the distance between iterates. In the context of sequential decision making, a natural means of leveraging mirror descent is to instantiate simultaneous updaters $\mathcal{U}^{\text{sync}}_{\text{local}}$ at each decision point of the following form:

$$\mathcal{U}^{\text{sync}}_{\text{local}} \colon (\pi_t, q) \mapsto \arg\max_{\pi} \langle q, \pi \rangle - \frac{1}{\eta}\text{KL}(\pi, \pi_t),$$

where π_t is a local policy, η is a stepsize, and q is a local action value vector; in the case of discrete action spaces, this update reduces to the following closed form, which is sometimes called hedge:

$$\mathcal{U}^{\text{sync}}_{\text{local}}(\pi_t, q) \propto \pi_t e^{\eta q}.$$

We can show that the approach induced by conditioning Algorithm 1 on mirror descent's update function, which we call mirror descent update equivalent search (MD-UES), is principled by proving that its synchronous analogue (i.e., mirror descent) has a desirable improvement property. In the appendix, we provide such a proof, which leads to the theorem below.

Theorem 3.3. *Consider a common-payoff game. Let π_t be a joint policy having positive probability on every action at every decision point. Then, if we run mirror descent at every decision point with action-value feedback, for any sufficiently small stepsize $\eta > 0$,*

$$\mathcal{J}(\pi_{t+1}) \geq \mathcal{J}(\pi_t).$$

Magnetic Mirror Descent Equivalent Search Another example of a setting that has been historically difficult for DTP is 2p0s games. In such settings, neither policy iteration- nor mirror descent-based approaches yield useful policies—instead, these approaches tend to cycle in rock-paper-scissors-like ways without converging. In practice, this issue can be resolved using magnetic mirror descent (Sokota et al., 2023a), which is an extension of mirror descent with additional proximal regularization to a magnet that dampens these cycles. As with mirror descent, in the context of sequential decision making, a natural means of leveraging magnetic mirror descent is to instantiate simultaneous updaters $\mathcal{U}^{\text{sync}}_{\text{local}}$ at each decision point:

$$\mathcal{U}^{\text{sync}}_{\text{local}} \colon (\pi_t, q) \mapsto \arg\max_{\pi} \langle q, \pi \rangle - \frac{1}{\eta}\text{KL}(\pi, \pi_t) - \alpha \text{KL}(\pi, \rho),$$

where α is a regularization temperature and ρ is a local magnet; in the case of discrete action spaces, this update reduce to the following closed form:

$$\mathcal{U}^{\text{sync}}_{\text{local}}(\pi_t, q) \propto [\pi_t e^{\eta q} \rho^{\eta \alpha}]^{\frac{1}{1+\alpha\eta}}.$$

We observe that the approach induced by conditioning Algorithm 1 on magnetic mirror descent's update function, which we call magnetic mirror descent update equivalent search (MMD-UES), is principled if it is true that magnetic mirror descent with action-value feedback is principled in 2p0s games. Encouragingly, there is empirical evidence that this is true.

Remark 3.4. Sokota et al. (2023a) show that magnetic mirror descent with action-value feedback empirically exhibits reliable last-iterate convergence in 2p0s games.

3.2. Beyond Action-Value Based Planners

There are many DTP algorithms that cannot be derived from Algorithm 1 that nevertheless may be useful to view from the perspective of update equivalence. One class of such approaches are those that make updates at future decision points during search. Approaches of this form, such as Monte Carlo tree search (MCTS) (Coulom, 2006; Kocsis & Szepesvári, 2006; Browne et al., 2012), played an important role in successes in perfect information games (Silver et al., 2016; 2018; Schrittwieser et al., 2019; Antonoglou et al., 2022). Another class of such approaches are those than fine-tune the belief model (Sokota et al., 2022) to track the search policy.

To illustrate how the framework of update equivalence can help justify such approaches in the context of 2p0s games with imperfect information, we examine the synchronous counterparts to three variants of MMD-based searches. 1) With subgame updates: A variant in which the planning agent also performs MMD updates at its own future decision points. 2) With belief fine-tuning (BFT): A variant that implements an MMD update on top of a posterior fine-tuned to the search policy. 3) With opponent updates: A variant that implements an MMD update assuming a subgame joint policy in which the opponent has performed an MMD update at its next decision point.

Results are shown in Figure 1. The plot measures divergence to agent quantal response equilibrium (AQRE) (McKelvey & Palfrey, 1998) as a function of number of update iterations. We find that all three synchronous analogues exhibit empirical convergence, suggesting that these DTP algorithms may

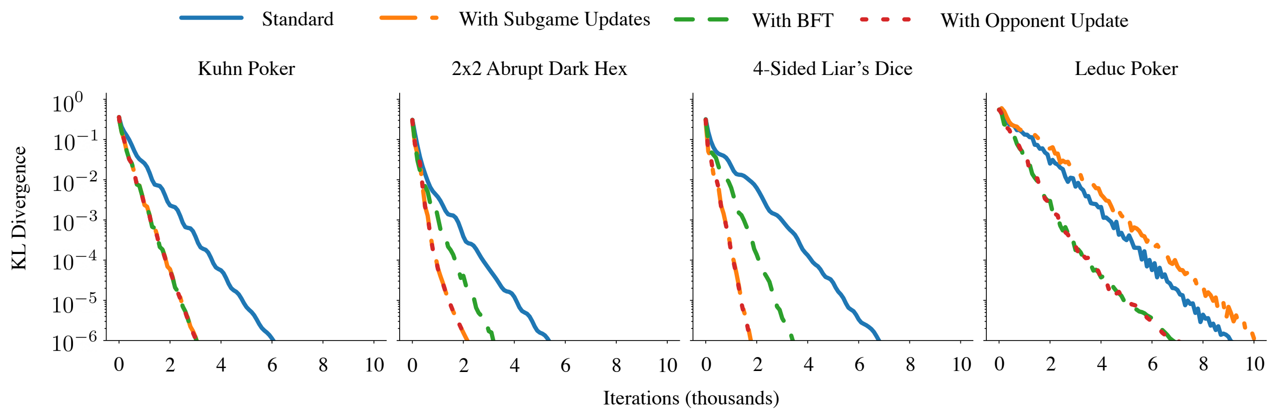

Figure 1. Divergence to AQRE as a function of iterations for synchronous analogues of variants of MMD-based search algorithms. Encouragingly, each variant exhibits empirical convergence.

be sound in 2p0s settings. More generally, these findings hint that, so long as MMD is used as a local updater, one has wide leeway in designing principled DTP algorithms for 2p0s games.

4. Experiments

In the previous section, we introduced the framework of update equivalence and showed how it enables us to both i) generate novel DTP algorithms, and ii) provide new analytical grounding for them. In this section, we add further evidence for the utility of this framework by showing that the novel DTP algorithms induced by it also perform well in practice. We focus on two settings with imperfect information: i) two variants of Hanabi (Bard et al., 2020), a fully cooperative card game in which PBS-based DTP approaches are considered state-of-the-art; and ii) 3x3 Abrupt Dark Hex and Phantom Tic-Tac-Toe, 2p0s games with virtually no public information.

4.1. Hanabi

We consider two versions of two-player Hanabi, a standard benchmark in the literature. The first variant, which is commonly played by humans, uses 5 cards and 8 hints. The second variant uses 7 cards and 4 hints and was first investigated by Hu et al. (2021) to illustrate the additional issues that arise when it becomes intractable to maintain a tabular posterior. To begin, we trained instances of independent PPO (Schulman et al., 2017) for each setting. We intentionally selected instances whose final performance roughly matched those of R2D2 (Kapturowski et al., 2019) instances that have been used to benchmark DTP algorithms (Fickinger et al., 2021; Sokota et al., 2022). Next, we trained a belief model for each PPO policy using Seq2Seq (Sutskever et al., 2014) with the same setup as Hu et al. (2021); in this setup, the belief model takes in the decision point of one player and predicts the private information of the other player, conditioned on the PPO policy having been played thus far.

We adapted our implementation of MD-UES from that of single-agent SPARTA with a learned belief model (Hu et al., 2021). Our adaptation involves three important changes. First, MD-UES performs search for all agents, rather than only one agent as was done in (Hu et al., 2021). Second, MD-UES plays the argmax[1] of a mirror descent update, rather than the argmax of the empirical Q-values from search, as SPARTA does. Third, MD-UES always plays its search policy. This contrasts both SPARTA (Lerer et al., 2020; Hu et al., 2021) and RLSearch (Fickinger et al., 2021), which, after doing search, perform a validation check to determine whether the search policy outperforms the blueprint policy; if the validation check fails, the blueprint policy is played instead of the search policy. This validation check is expensive and usually fails (SPARTA and RLSearch only play their search policies on a handful of turns). Thus, that MD-UES does not require it is a significant advantage.

We show the results of our experiments for 5-card 8-hint Hanabi in Table 1 with standard error over 2000 games. We compare MD-UES using a PPO blueprint policy against single- and multi-agent SPARTA and RLSearch, which are considered state-of-the-art, with exact belief models and an R2D2 blueprint policy; we also compare against single-agent SPARTA using the same PPO blueprint policy and Seq2Seq belief model. We find that, even with an approximate belief model, MD-UES matches the performance state-of-the-art approaches using exact beliefs. We show the results of our experiments for 7-card 4-hint Hanabi in Table 2 with standard error over 2000 games. In 7-card 4-hint Hanabi, the number of possible histories is too large to

[1]Note that this differs slightly from our description in the previous section in that we are playing the argmax, rather than sampling, from the updated policy.

Table 1. MD-UES (bold) compared to SPARTA and RLSearch in 5-card 8-hint Hanabi. Despite only using approximate beliefs, MD-UES roughly matches performance of prior methods using exact beliefs.

	(a) — **R2D2 Blueprint**					(b) — **PPO Blueprint**		
Search	-	SPARTA	RLSearch	SPARTA	RLSearch	-	SPARTA	**MD-UES**
Agents planned	-	Single	Single	Multi	Multi	-	Single	**Multi**
Belief model	-	Exact	Exact	Exact	Exact	-	Seq2Seq	**Seq2Seq**
Expected return	24.23	24.57	24.59	24.61	24.62	24.24	24.52	**24.62**
Standard error	± 0.04	± 0.03	± 0.02	± 0.02	± 0.03	± 0.02	± 0.02	**± 0.02**

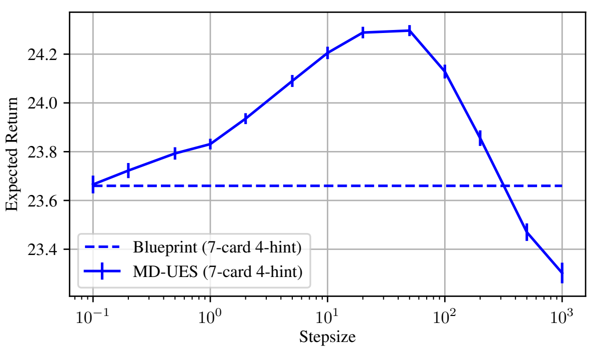

Figure 2. MD-UES Hanabi performance as a function of stepsize. Small step sizes provide less improvement over the blueprint; overly large step sizes cause the search policy to diverge too far from the blueprint, resulting in less improvement or even detriment.

enumerate. As a result, multi-agent SPARTA is inapplicable altogether, single-agent SPARTA and RLSearch require an approximate belief model, and multi-agent RLSearch requires both an approximate belief model and belief fine-tuning (Sokota et al., 2022). Our results suggest that MD-UES compares favorably against these approaches.

To offer further intuition for the behavior of MD-UES, we show the performance in 7-card 4-hint Hanabi as a function of mirror descent's stepsize in Figure 2. As might be expected, we find that improvement is a unimodal function of stepsize: If the stepsize is too small, opportunities to improve the blueprint policy are neglected; on the other hand, if the stepsize is too large, the search policy's posterior diverges too far from that of the blueprint for the search to provide useful feedback.

4.2. 3x3 Abrupt Dark Hex and Phantom Tic-Tac-Toe

Compared to benchmarking performance in common-payoff games, benchmarking performance in 2p0s is more difficult because the usual metric of interest—exploitability (*i.e.*, the expected return of a best responder)—is not cheaply estimable. Indeed, providing a reasonable upper bound requires training an approximate best response (Timbers et al., 2022), which may involve many episodes of training. To facilitate this approximate best response training under a modest computation budget, we focus our experiments on a blueprint policy that selects actions uniformly at random at every decision point so that rollouts may be performed quickly; furthermore, rather than using a learned belief model, we track an approximate posterior using particle filtering (Doucet & Johansen, 2009) with 10 particles. If there are particles remaining, our implementation of MMD-UES performs a rollout for each particle for each legal action to the end of the game and performs an MMD update on top of the empirical means of the returns. If these particles have run out (i.e., no particle is consistent with the planning agent's decision point), we set the agent to execute the (uniformly random) blueprint policy.

We empirically investigate MMD-UES in 3x3 Abrupt Dark Hex and Phantom Tic-Tac-Toe, two standard benchmarks available in OpenSpiel (Lanctot et al., 2019). We show approximate exploitability results in Table 3. We computed these results using OpenSpiel's (Lanctot et al., 2019) DQN (Mnih et al., 2015) best response code, trained for 10 million time steps. We show standard error over 5 DQN best response training seeds and 2000 final evaluation games for the fully trained model. We compare against a bot that plays the first legal action, the uniform random blueprint, Liang et al. (2018)'s implementation of independent PPO (Schulman et al., 2017), Lanctot et al. (2019)'s implementation of NFSP (Heinrich & Silver, 2016), and MMD (Sokota et al., 2023a). For the learning agents, we ran 5 seeds and include checkpoints trained for both 1 million time steps and 10 million time steps.

We find that MMD-UES reduces the approximate exploitability of a uniformly random blueprint by more than a third, despite that it uses a meager 10 particle approximate posterior. Furthermore, compared to the baselines, MMD-UES achieves lower approximate exploitability than all of the non-MMD based approaches.

In addition to approximate exploitability results, we also investigate the performance of MMD-UES in head-to-head matchups. We show results using the uniform random

Table 2. MD-UES (bold) compared to SPARTA and RLSearch in 7-card 4-hint Hanabi; MD-UES compares favorably to these approaches.

	(a) — **R2D2 Blueprint**				(b) — **PPO Blueprint**		
Search	-	RLSearch	RLSearch	RLSearch	-	SPARTA	**MD-UES**
Agents planned	-	Single	Single	Multi	-	Single	**Multi**
Belief model	-	Seq2Seq	BFT	BFT	-	Seq2Seq	**Seq2Seq**
Expected return	23.67	24.14	24.18	24.18	23.66	24.17	**24.28**
Standard error	$\pm$ 0.02	$\pm$ 0.04	$\pm$ 0.03	$\pm$ 0.03	$\pm$ 0.03	$\pm$ 0.03	$\pm$ **0.02**

Table 3. Approximate exploitability in 3x3 Abrupt Dark Hex and Phantom Tic-Tac-Toe, on a 0 to 100 scale. MMD-UES (bold) substantially reduces Random's exploitability.

Agent	3x3 Ab. DH	Phantom TTT
1st Legal Action	100 ± 0	100 ± 0
Random	74 ± 1	78 ± 0
PPO(1M steps)	85 ± 6	89 ± 6
PPO(10M steps)	100 ± 0	90 ± 4
NFSP(1M steps)	91 ± 4	95 ± 1
NFSP(10M steps)	59 ± 1	78 ± 5
Random+MMD-UES	$\mathbf{50 \pm 1}$	$\mathbf{50 \pm 1}$
MMD(1M steps)	34 ± 2	37 ± 1
MMD(10M steps)	20 ± 1	15 ± 1

blueprint with 10 particles and also a MMD(1M) blueprint with 100 particles in Figure 3. The values shown are averages over 10,000 games with bootstrap estimates of 95% confidence intervals. We find that MMD-UES tends to improve the performance of the blueprint policies in both cases.

5. Related Work

We subdivide related work into three groups: approaches to DTP that rely on PBSs, approaches to DTP that are not handicapped by the amount of non-public information, and structurally similar approaches to DTP as those that naturally arise out of the framework of update equivalence.

PBS-Based Approaches to Imperfect Information The first sound approaches to DTP arose in the context of common-payoff games from the work of Nayyar et al. (2013). Nayyar et al. (2013) showed that common-payoff games can be transformed into Markov decision processes (MDPs) in which the Markov states are the PBSs of the original game, thereby facilitating planning via PBSs.

Sound DTP algorithms for 2p0s games involve additional subtlety because solutions of the analogous public belief game do not generally correspond to solutions of the original game. The approaches to resolving this non-correspondence problem involve opt-out values (Brown & Sandholm, 2017; Moravčík et al., 2017), no-regret learning (Brown et al., 2020), and regularization (Sokota et al., 2023b).

Sound DTP algorithms for two-team zero-sum games involve using Nayyar et al. (2013)-like transformations (Carminati et al., 2022a; Zhang et al., 2022a; Carminati et al., 2022b), which reduce two-team zero-sum games to 2p0s games, resolving the non-correspondence problem.

These techniques have led to successes in common-payoff games (Lerer et al., 2020; Sokota et al., 2021), 2p0s games (Moravčík et al., 2017; Brown & Sandholm, 2018; Zarick et al., 2020; Brown et al., 2020; Schmid et al., 2021), and two-team zero-sum games (Zhang et al., 2022b), in settings in which the amount of non-public information is small.

The presence of PBSs in these approaches motivated work improving their scalability to settings with larger amounts of non-public information. There have been two notable advancements on this front. First, Fickinger et al. (2021) show how Lerer et al. (2020)'s approach can be extended to settings in which the support of the PBS is too large to separately perform search for each decision point by fine-tuning the policy in the style of (Anthony et al., 2019). Second, Sokota et al. (2022) show how Fickinger et al. (2021)'s approach can be extended to settings in which the PBS is too large to track tabularly by inductively fine-tuning the belief model. However, these advancements do not address the fundamental limitations of PBSs.

Other Approaches to Imperfect Information Motivated by the deficiencies of PBS-based DTP described above, a small group of existing works has advocated for alternative approaches to DTP for common-payoff games (Tian et al., 2020) and 2p0s games (Zhang & Sandholm, 2021).

Tian et al. (2020)'s approach relies on a result that shows that it is possible to decompose the change in expected return for two different joint policies across decision points. By leveraging this result, Tian et al. (2020) introduce a search procedure called joint policy search (JPS) that is guaranteed to not decrease the expected return.

Zhang & Sandholm (2021)'s approach is based around the

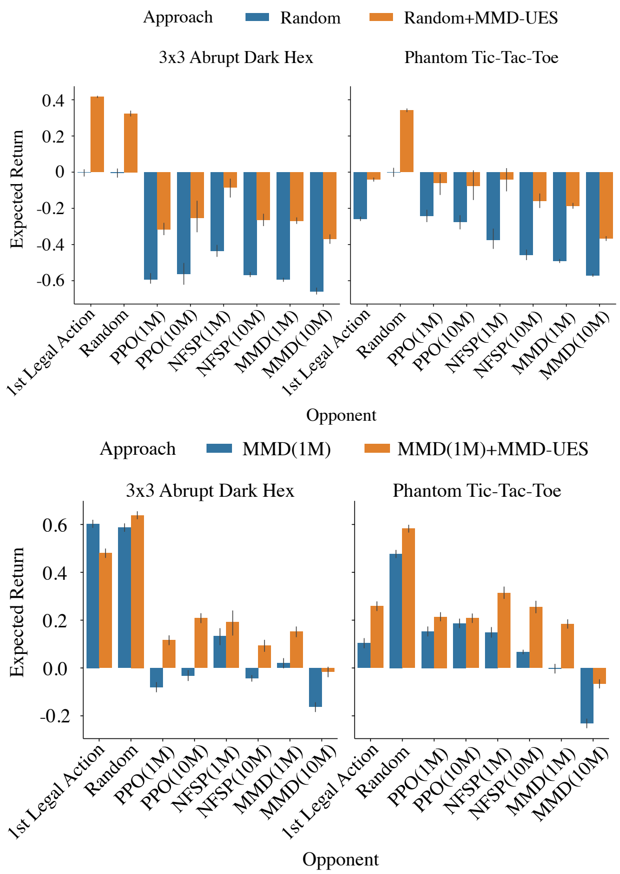

Figure 3. Expected return of MMD-UES compared to uniform random and MMD(1M) blueprint policies in head-to-head matches in 3x3 Abrupt Dark Hex and Phantom Tic-Tac-Toe. MMD-UES tends to improve head-to-head expected return.

insight that, in practice, it is effective to consider a subgame that excludes most of the decision points supported by the PBS. Specifically, Zhang & Sandholm (2021) advocate in favor of solving a maxmargin subgame (Moravcik et al., 2016) that includes the planning agent's true decision point, as well as any opponent decision points that are possible from the perspective of the planning agent. Despite proving the existence of games in which this approach, which they call 1-KLSS, increases the exploitability of the policy, Zhang & Sandholm (2021) find experimentally that, on small and medium-sized benchmark games, 1-KLSS reliably decreases exploitability.

Comparing JPS and 1-KLSS to the update equivalence framework, the update equivalence framework may offer the advantage of being easier to understand, especially for researchers not coming from imperfect-information games backgrounds. As far as theoretical grounding, both JPS and MD-UES possess provable improvement guarantees; on the other hand, both 1-KLSS and MMD-UES lack formal theoretical grounding (though there is hope that formal grounding for MMD-UES is possible). Regarding performance in common-payoff games, we would describe MD-UES performing comparably with or superior to PBS-based methods in Hanabi as a strong result; however, it is difficult to compare this result with those of JPS, as it is has neither been benchmarked on Hanabi nor against PBS-based search methods. It is also difficult to make any definitive statements regarding relative performance compared to 1-KLSS. We provide a very limited comparison to some of the numbers that Zhang & Sandholm (2021) reported for small games in the appendix; however, it is unlikely that these small games qualitatively resemble the larger regimes for which the methods were intended.

Structurally Similar Approaches While the motivation for the update equivalence framework most closely resembles of the works of Tian et al. (2020) and Zhang & Sandholm (2021), natural instances of the update equivalence framework are structurally more similar to search algorithms unrelated to resolving issues with PBS-based planning. As discussed previously, arguably the most fundamental instance of the framework of update equivalence is Monte Carlo search (MCS) (Tesauro & Galperin, 1996), which was used prominently in TD-Gammon (Tesauro, 1995). MCS's synchronous equivalent is policy iteration, as was articulated by Tesauro & Galperin (1996) themselves: "[MCS] basically implements a single step of policy iteration."

Anthony et al. (2019); Anthony (2021); Hamrick et al. (2021); Lerer et al. (2020); Hu et al. (2021) recently investigated MCS in a variety of settings and report positive results. Anthony et al. (2019); Anthony (2021) investigate MCS in the context of Hex, finding that it yields perhaps surprisingly good performance relative to MCTS. Hamrick et al. (2021) also report positive results for MCS, finding that it achieves comparable performance with MCTS across a variety of settings, with the exception of Go. Lerer et al. (2020) and Hu et al. (2021) investigate a variant of MCS in the context of Hanabi (under the name single-agent SPARTA and learned belief search) and also show strong results.

In their work, Anthony (2021) also investigate a search algorithm called policy gradient search that performs policy gradient updates at future decision points. They find that a variant of this approach that involves persistent regularization toward this blueprint policy tends to outperform not regularizing. This approach is very similar to some of the variants of MMD-UES investigated in Section 3.2. We feel that the combination of policy gradient search and the framework of update equivalence may be a fruitful direction in the pursuit of algorithms that perform well in both perfect and imperfect information games.

Separately from Anthony (2021), Jacob et al. (2022) also investigate approaches to DTP involving regularization toward a blueprint. Jacob et al. (2022) show empirically that, by performing regularized search, it is possible to increase expected return of an imitation learned policy without a loss of prediction accuracy (for the policy being imitated). The most immediately related experiments to this work are those concerning Hanabi, in which Jacob et al. (2022) investigated MD-UES (under the name piKL SPARTA) applied on top of imitation learned blueprint policies. Jacob et al. (2022) note that this approach reliably increases the performance of weak policies, but neither recognize that it possesses an improvement guarantee nor that it is simply performing a hedge update, as we do in this work. Jacob et al. (2022)'s experiments on Diplomacy (Paquette et al., 2019) are also related in that their approach is similar to a follow-the-regularized-leader analogue of MMD-UES. This approach played an important role in recent empirical successes for Diplomacy (Bakhtin et al., 2022; Meta Fundamental AI Research Diplomacy Team et al., 2022), suggesting that the framework of update equivalence is a natural approach to general-sum settings.

6. Conclusion and Future Work

In this work, motivated by the deficiencies of subgame-based search, we advocated for a new paradigm to DTP that we call the framework of update equivalence. We showed how the framework of update equivalence can be used to generate and ground new DTP algorithms for imperfect-information games. Furthermore, we showed that these algorithms can achieve competitive (or superior) performance compared with state-of-the-art subgame-based methods in Hanabi and can reduce approximate exploitability in 3x3 Abrupt Dark Hex and Phantom Tic-Tac-Toe.

We believe the framework of update equivalence opens the door to many exciting possibilities for DTP and expert iteration in settings with large amounts of imperfect information. In our opinion, the most exciting of these possibilities is the prospect of extending algorithms closely resembling AlphaZero (Silver et al., 2018) and Stochastic MuZero (Antonoglou et al., 2022) to imperfect-information and general-sum games. We hope to pursue these directions in future work.

7. Acknowledgements

We thank Brian Hu Zhang for helpful discussions.

References

Anthony, T., Tian, Z., and Barber, D. 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.

Anthony, T., Nishihara, R., Moritz, P., Salimans, T., and Schulman, J. Policy gradient search: Online planning and expert iteration without search trees. *CoRR*, abs/1904.03646, 2019. URL http://arxiv.org/abs/1904.03646.

Anthony, T. W. *Expert iteration*. PhD thesis, UCL (University College London), 2021.

Antonoglou, I., Schrittwieser, J., Ozair, S., Hubert, T. K., and Silver, D. Planning in stochastic environments with a learned model. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=X6D9bAHhBQ1.

Bakhtin, A., Wu, D. J., Lerer, A., Gray, J., Jacob, A. P., Farina, G., Miller, A. H., and Brown, N. Mastering the game of no-press diplomacy via human-regularized reinforcement learning and planning, 2022. URL https://arxiv.org/abs/2210.05492.

Bard, N., Foerster, J. N., Chandar, S., Burch, N., Lanctot, M., Song, H. F., Parisotto, E., Dumoulin, V., Moitra, S., Hughes, E., Dunning, I., Mourad, S., Larochelle, H., Bellemare, M. G., and Bowling, M. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence*, 280:103216, 2020. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2019.103216. URL https://www.sciencedirect.com/science/article/pii/S0004370219300116.

Beck, A. and Teboulle, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003. ISSN 0167-6377.

Brown, N. and Sandholm, T. 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, 2017. Curran Associates Inc. ISBN 9781510860964.

Brown, N. and Sandholm, T. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359:418 – 424, 2018.

Brown, N. and Sandholm, T. Superhuman ai for multiplayer poker. *Science*, 365(6456):885–890, 2019. doi: 10.1126/science.aay2400. URL https://www.science.org/doi/abs/10.1126/science.aay2400.

Brown, N., Bakhtin, A., Lerer, A., and Gong, Q. Combining deep reinforcement learning and search for imperfect-information games. *Advances in Neural Information Processing Systems*, 33, 2020.

Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S., and Colton, S. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43, 2012. doi: 10.1109/TCIAIG.2012.2186810.

Burch, N., Johanson, M., and Bowling, M. 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.

Campbell, M., Hoane, A., and hsiung Hsu, F. Deep blue. *Artificial Intelligence*, 134 (1):57–83, 2002. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(01)00129-1. URL https://www.sciencedirect.com/science/article/pii/S0004370201001291.

Carminati, L., Cacciamani, F., Ciccone, M., and Gatti, N. Public information representation for adversarial team games, 2022a. URL https://arxiv.org/abs/2201.10377.

Carminati, L., Cacciamani, F., Ciccone, M., and Gatti, N. A marriage between adversarial team games and 2-player games: Enabling abstractions, no-regret learning, and subgame solving. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (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.

Coulom, R. Efficient selectivity and backup operators in monte-carlo tree search. In van den Herik, H. J., Ciancarini, P., and Donkers, H. H. L. M. (eds.), *Computers and Games*, volume 4630 of *Lecture Notes in Computer Science*, pp. 72–83. Springer, 2006. ISBN 978-3-540-75537-1. URL http://dblp.uni-trier.de/db/conf/cg/cg2006.html#Coulom06.

Doucet, A. and Johansen, A. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, 12, 01 2009.

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

Hamrick, J. B., Friesen, A. L., Behbahani, F., Guez, A., Viola, F., Witherspoon, S., Anthony, T., Buesing, L. H., Veličković, P., and Weber, T. On the role of planning in model-based deep reinforcement learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=IrM64DGB21.

Hansen, E. A., Bernstein, D. S., and Zilberstein, S. Dynamic programming for partially observable stochastic games. In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, pp. 709–715. AAAI Press, 2004. ISBN 0262511835.

Heinrich, J. and Silver, D. Deep reinforcement learning from self-play in imperfect-information games, 2016. URL https://arxiv.org/abs/1603.01121.

Hu, H., Lerer, A., Brown, N., and Foerster, J. N. Learned belief search: Efficiently improving policies in partially observable settings. *CoRR*, abs/2106.09086, 2021. URL https://arxiv.org/abs/2106.09086.

Jacob, A. P., Wu, D. J., Farina, G., Lerer, A., Hu, H., Bakhtin, A., Andreas, J., and Brown, N. Modeling strong and human-like gameplay with KL-regularized search. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9695–9728. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/jacob22a.html.

Jakobsen, S. K., Sørensen, T. B., and Conitzer, V. Timeability of extensive-form games. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16, pp. 191–199, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450340571. doi: 10.1145/2840728.2840737. URL https://doi.org/10.1145/2840728.2840737.

Kapturowski, S., Ostrovski, G., Dabney, W., Quan, J., and Munos, R. Recurrent experience replay in distributed reinforcement learning. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=r1lyTjAqYX.

Kocsis, L. and Szepesvári, C. Bandit based monte-carlo planning. In *Proceedings of the 17th European Conference on Machine Learning*, ECML'06, pp. 282–293, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 354045375X. doi: 10.1007/11871842_29. URL https://doi.org/10.1007/11871842_29.

Kovarík, V., Schmid, M., Burch, N., Bowling, M., and Lisý, V. Rethinking formal models of partially observable multiagent decision making. *Artif. Intell.*, 303:103645, 2022.

doi: 10.1016/j.artint.2021.103645. URL https://doi.org/10.1016/j.artint.2021.103645.

Lanctot, M., Lockhart, E., Lespiau, J.-B., Zambaldi, V., Upadhyay, S., Pérolat, J., Srinivasan, S., Timbers, F., Tuyls, K., Omidshafiei, S., Hennes, D., Morrill, D., Muller, P., Ewalds, T., Faulkner, R., Kramár, J., Vylder, B. D., Saeta, B., Bradbury, J., Ding, D., Borgeaud, S., Lai, M., Schrittwieser, J., Anthony, T., Hughes, E., Danihelka, I., and Ryan-Davis, J. OpenSpiel: A framework for reinforcement learning in games. *CoRR*, abs/1908.09453, 2019. URL http://arxiv.org/abs/1908.09453.

Lerer, A., Hu, H., Foerster, J., and Brown, N. Improving policies via search in cooperative partially observable games. In *AAAI Conference on Artificial Intelligence*, 2020.

Liang, E., Liaw, R., Nishihara, R., Moritz, P., Fox, R., Goldberg, K., Gonzalez, J., Jordan, M., and Stoica, I. RLlib: Abstractions for distributed reinforcement learning. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3053–3062. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/liang18b.html.

Littman, M. L. Algorithms for sequential decision making, 1996.

McKelvey, R. D. and Palfrey, T. R. Quantal response equilibria for extensive form games. *Experimental Economics*, 1:9–41, 1998.

Meta Fundamental AI Research Diplomacy Team, F., Bakhtin, A., Brown, N., Dinan, E., Farina, G., Flaherty, C., Fried, D., Goff, A., Gray, J., Hu, H., et al. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378 (6624):1067–1074, 2022.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D. Human-level control through deep reinforcement learning. *Nature*, 518 (7540):529–533, February 2015. ISSN 00280836. URL http://dx.doi.org/10.1038/nature14236.

Moravcik, M., Schmid, M., Ha, K., Hladik, M., and Gaukrodger, S. J. Refining subgames in large imperfect information games. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pp. 572–578. AAAI Press, 2016.

Moravčík, M., Schmid, M., Burch, N., Lisý, V., Morrill, D., Bard, N., Davis, T., Waugh, K., Johanson, M., and Bowling, M. 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.

Nayyar, A., Mahajan, A., and Teneketzis, D. 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.

Nemirovsky, A. and Yudin, D. *Problem Complexity and Method Efficiency in Optimization*. A Wiley-Interscience publication. Wiley, 1983. ISBN 9780471103455.

Paquette, P., Lu, Y., Bocco, S. S., Smith, M., O-G, S., Kummerfeld, J. K., Pineau, J., Singh, S., and Courville, A. C. No-press diplomacy: Modeling multi-agent gameplay. *Advances in Neural Information Processing Systems*, 32, 2019.

Perolat, J., Munos, R., Lespiau, J.-B., Omidshafiei, S., Rowland, M., Ortega, P., Burch, N., Anthony, T., Balduzzi, D., De Vylder, B., Piliouras, G., Lanctot, M., and Tuyls, K. From poincaré recurrence to convergence in imperfect information games: Finding equilibrium via regularization. In Meila, M. and Zhang, T. (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.

Schmid, M., Moravcik, M., Burch, N., Kadlec, R., Davidson, J., Waugh, K., Bard, N., Timbers, F., Lanctot, M., Holland, Z., Davoodi, E., Christianson, A., and Bowling, M. Player of games. *CoRR*, abs/2112.03178, 2021. URL https://arxiv.org/abs/2112.03178.

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., and Silver, D. Mastering atari, go, chess and shogi by planning with a learned model, 2019. URL http://arxiv.org/abs/1911.08265. cite arxiv:1911.08265.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, jan 2016. ISSN 0028-0836. doi: 10.1038/nature16961.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018. doi: 10.1126/science.aar6404. URL https://www.science.org/doi/abs/10.1126/science.aar6404.

Sokota, S., Lockhart, E., Timbers, F., Davoodi, E., D'Orazio, R., Burch, N., Schmid, M., Bowling, M., and Lanctot, M. Solving common-payoff games with approximate policy iteration. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 9695–9703. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17166.

Sokota, S., Hu, H., Wu, D. J., Kolter, J. Z., Foerster, J. N., and Brown, N. A fine-tuning approach to belief state modeling. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=ckZY7DGa7FQ.

Sokota, S., D'Orazio, R., Kolter, J. Z., Loizou, N., Lanctot, M., Mitliagkas, I., Brown, N., and Kroer, C. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=DpE5UYUQzZH.

Sokota, S., D'Orazio, R., Ling, C. K., Wu, D. J., Kolter, J. Z., and Brown, N. Abstracting imperfect information away from two-player zero-sum games, 2023b. URL https://arxiv.org/abs/2301.09159.

Sutskever, I., Vinyals, O., and Le, Q. V. Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

Tesauro, G. Temporal difference learning and td-gammon. *J. Int. Comput. Games Assoc.*, 18(2):88, 1995. URL http://dblp.uni-trier.de/db/journals/icga/icga18.html#Tesauro95.

Tesauro, G. and Galperin, G. On-line policy improvement using monte-carlo search. In Mozer, M., Jordan, M., and Petsche, T. (eds.), *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. URL https://proceedings.neurips.cc/paper/1996/file/996009f2374006606f4c0b0fda878af1-Paper.pdf.

Tian, Y., Gong, Q., and Jiang, T. Joint policy search for multi-agent collaboration with imperfect information. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Timbers, F., Bard, N., Lockhart, E., Lanctot, M., Schmid, M., Burch, N., Schrittwieser, J., Hubert, T., and Bowling, M. Approximate exploitability: Learning a best response. In Raedt, L. D. (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 3487–3493. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/484. URL https://doi.org/10.24963/ijcai.2022/484. Main Track.

Zarick, R., Pellegrino, B., Brown, N., and Banister, C. Unlocking the potential of deep counterfactual value networks. *ArXiv*, abs/2007.10442, 2020.

Zhang, B. and Sandholm, T. Subgame solving without common knowledge. In *Neural Information Processing Systems*, 2021.

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

Zhang, B. H., Carminati, L., Cacciamani, F., Farina, G., Olivieri, P., Gatti, N., and Sandholm, T. Subgame solving in adversarial team games. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=Roiw2Trm-qP.

Zinkevich, M., Johanson, M., Bowling, M., and Piccione, C. Regret minimization in games with incomplete information. In Platt, J., Koller, D., Singer, Y., and Roweis, S. (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper/2007/file/08d98638c6fcd194a4b1e6992063e944-Paper.pdf.

A. Theory

In this section, we prove Theorem 3.3. To start, consider the following more general lemma.

Lemma A.1 (Folklore). *Let $f \in \mathcal{C}^1(\mathcal{X})$, where $\mathcal{X} \subseteq \mathbb{R}^d$ is convex and compact, $x_0 \in \mathcal{X}$, and x^+ be the solution to the mirror descent step*

$$x^+ := \arg\min_{x \in \mathcal{X}} \left\{ \eta \langle \nabla f(x_0), x \rangle + D_\varphi(x, x_0) \right\}.$$

where φ is differentiable and 1-strongly convex with respect to a norm $\|\cdot\|$. Then, for η small enough, if $x^+ \neq x_0$,

$$f(x^+) < f(x_0),$$

Proof. From the first-order necessary optimality conditions for the mirror descent step, we have

$$\langle \eta \nabla f(x_0) + \nabla \varphi(x^+) - \nabla \varphi(x_0), \hat{x} - x^+ \rangle \geq 0 \quad \forall x \in \mathcal{X}.$$

Hence, letting $\hat{x} := x_0$, and using the strong convexity of φ and the assumption that $x^+ \neq x_0$, we obtain

$$\langle \nabla f(x_0), x^+ - x_0 \rangle \leq -\frac{1}{\eta} \|x^+ - x_0\|^2 < 0,$$

and with a further application of the Cauchy-Schwarz inequality,

$$\|x^+ - x_0\| \leq \eta \|\nabla f(x_0)\|.$$

By continuity of the gradient of f, there must exist $\epsilon > 0$ such that

$$\langle \nabla f(x), x^+ - x_0 \rangle < 0 \quad \forall x \in B(x_0, \epsilon).$$

Hence, the mean value theorem guarantees that

$$f(x^+) - f(x^0) = \langle \nabla f(\xi), x^+ - x_0 \rangle$$

for some ξ on the line connecting x_0 to x^+. So, as long as $\eta \|\nabla f(x_0)\| < \epsilon$, we have

$$f(x^+) < f(x_0),$$

as we wanted to show. □

Theorem 3.3 Consider a common-payoff game. Let π_t be a joint policy having positive probability on every action at decision point. Then, if we run mirror descent at every decision point with action-value feedback, for any sufficiently small stepsize η,

$$\mathcal{J}(\pi_{t+1}) \geq \mathcal{J}(\pi_t).$$

Proof. This follows immediately from Lemma A.1. □

B. Experiments

In this section, we provide further details about some of our empirical results and also show some additional results.

B.1. Beyond Action-Value-Based Planners

First, we describe in greater detail the DTP algorithms that we investigated in Section 3.2.

With Subgame Updates With subgame updates differs from MMD-UES in the feedback it uses. In particular, rather than using the action values for the current policy, with subgame updates uses the action-values for the joint policy induced by performing MMD updates at its own future decision points (but leaving the opponent's policy fixed). Because the opponent is fixed, in tabular settings, we can compute the feedback for the synchronous analogue of this approach using one backward induction pass for each player.

With Belief Fine-Tuning With BFT differs from MMD-UES in the distribution it samples histories from. In particular, rather than sampling from the distribution induced by the current policy, it samples from the distribution induced by the search policy for each player.

With Opponent Update With opponent update differs from MMD-UES in the feedback it uses. In particular, rather than using the action values for the current policy, with opponent update uses the action values for the joint policy induced by performing an MMD update for the opponent at the next time step. Note that, as a DTP algorithm, this approach would involve two belief model sampling steps: one to sample an opponent decision point for updating and one to sample an unbiased history for that opponent information state.

Agent Quantal Response Equilibria Solving In Figure 4, we show the convergence results from the main body, along with the exploitabilities of the corresponding iterates. For these experiments, we used $\alpha = 0.1$ and $\eta = \alpha/10$, except for with opponent updates on Leduc, where we used $\eta = \alpha/20$, and with subgame updates on Leduc, where we used $\eta = \alpha/50$.

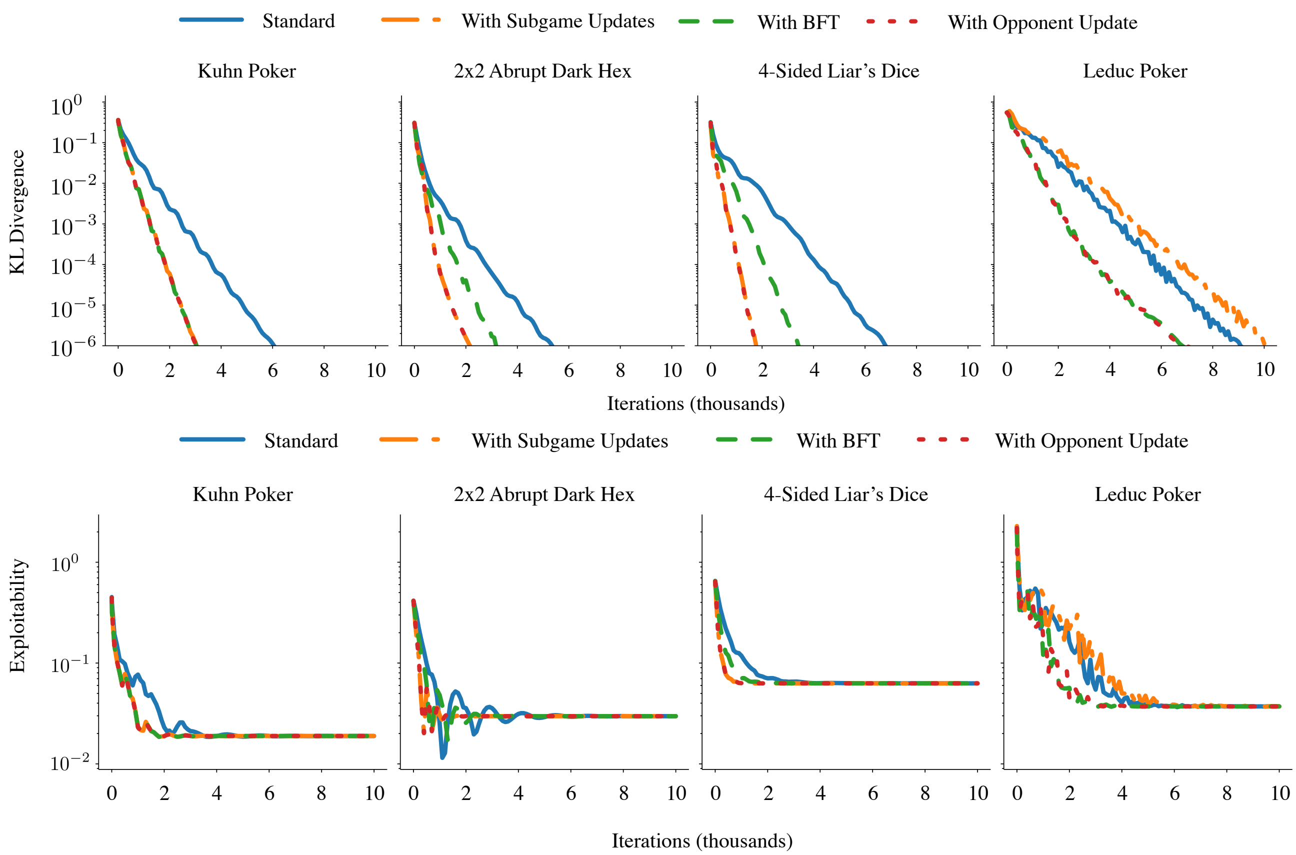

Figure 4. Solving for agent quantal response equilibria using synchronous analogues of variants of MMD-UES.

MiniMaxEnt Equilibrium Solving In Figure 5, we show convergence results for solving for MiniMaxEnt equilibria, which are the solutions of MiniMaxEnt objectives (Perolat et al., 2021). A MiniMaxEnt objective is an objective of the form

$$\mathcal{J}_i \colon \pi \mapsto \mathbb{E}\left[\sum_t \mathcal{R}_i(S^t, A^t) + \alpha\mathcal{H}(\pi_i(H_i^t)) - \alpha\mathcal{H}(\pi_{-i}(H_{-i}^t)) \mid \pi\right].$$

We used $\alpha = 0.1$ and $\eta = \alpha/10$, except for with opponent update on Leduc, where we used $\eta = \alpha/20$, and with subgame updates on Leduc, which used $\eta = \alpha/50$. We again observe empirical convergence.

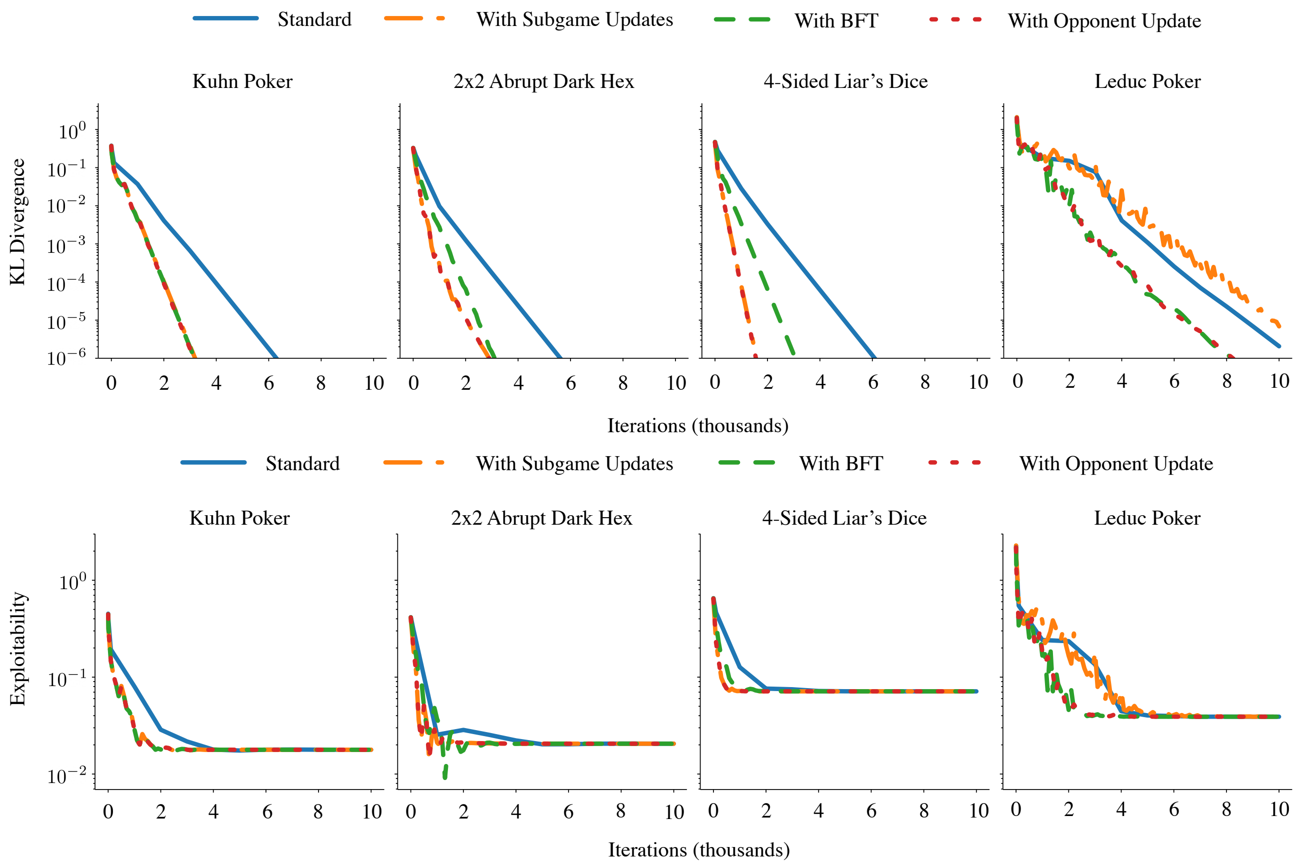

Figure 5. Solving for MiniMaxEnt equilibria using synchronous analogues of variants of MMD-UES.

Solving for Nash Equilibria Next we show that these synchronous analogues can be made to converge to Nash equilibria by annealing the amount of regularization used. We show that these results in Figure 6 compared against CFR (Zinkevich et al., 2007). The hyperparameters for these results are shown in Table 4.

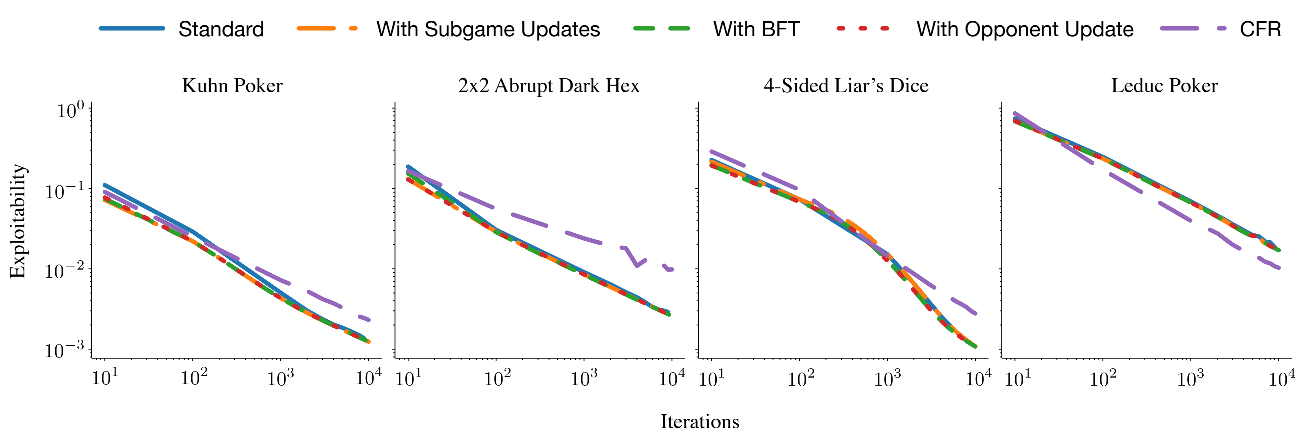

Figure 6. Exploitability of different MMD-UES analogues with annealed regularization.

Method\Game	Kuhn Poker	2x2 Abrupt Dark Hex	4-Sided Liar's Dice	Leduc Poker
Standard	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{2}{\sqrt{t}}$	$\alpha_t = \frac{5}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$
With Subgame Updates	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{2\sqrt{t}}$	$\alpha_t = \frac{5}{\sqrt{t}}, \eta_t = \frac{1}{5\sqrt{t}}$
With BFT	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{2}{\sqrt{t}}$	$\alpha_t = \frac{5}{\sqrt{t}}, \eta_t = \frac{1}{2\sqrt{t}}$
With Opponent Update	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{5}{\sqrt{t}}, \eta_t = \frac{1}{2\sqrt{t}}$

Table 4. Schedules for Figure 6.

Solving for Nash Equilibria with MiniMaxEntRL objectives We also show analogous results for MiniMaxEnt objectives in Figure 7. The hyperparameters for these experiments are shown in Table 5.

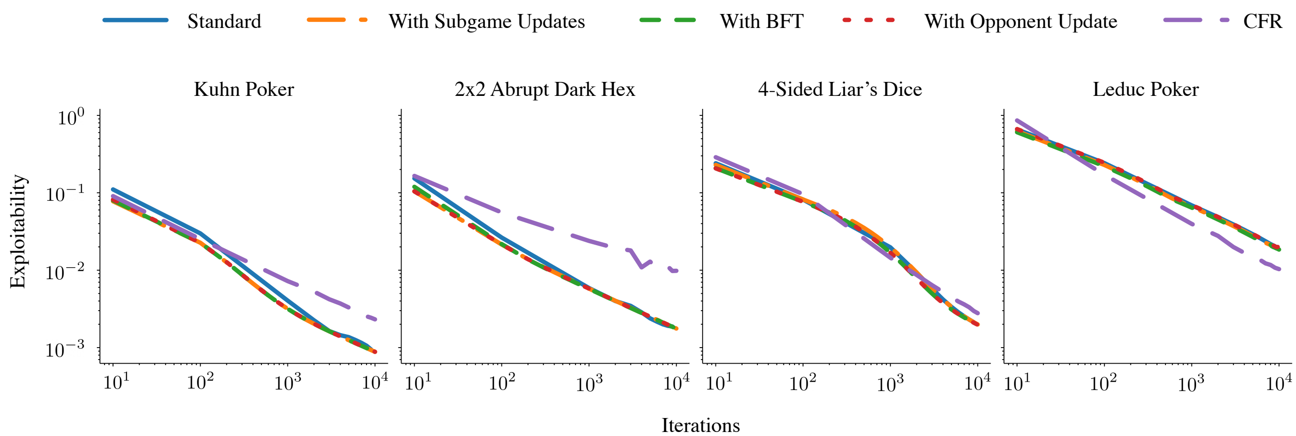

Figure 7. Exploitability of different MMD-UES analogues under MiniMaxEnt objectives with annealed regularization.

Method\Game	Kuhn Poker	2x2 Abrupt Dark Hex	4-Sided Liar's Dice	Leduc Poker
One-Step Search	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{2}{\sqrt{t}}$	$\alpha_t = \frac{5}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$
Multi-Step Search	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{2\sqrt{t}}$	$\alpha_t = \frac{5}{\sqrt{t}}, \eta_t = \frac{1}{5\sqrt{t}}$
BFT Search	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{2}{\sqrt{t}}$	$\alpha_t = \frac{5}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$
Opponent Search	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{1}{\sqrt{t}}, \eta_t = \frac{1}{\sqrt{t}}$	$\alpha_t = \frac{5}{\sqrt{t}}, \eta_t = \frac{1}{2\sqrt{t}}$

Table 5. Schedules for Figure 7.

B.2. Hanabi

For our Hanabi experiments, we used $\eta = 20$ for the MD-UES results in Tables 1 and 2. We performed search with 10,000 samples. On average, search took about 1.86 seconds per move using two GPUs.

B.3. 3x3 Abrupt Dark Hex and Phantom-Tic-Tac-Toe

For our 3x3 Abrupt Dark Hex and Phantom-Tic-Tac-Toe experiments with a uniform blueprint, we used $\eta = 50$, $\alpha = 0.01$, and set ρ to be uniform. For the MMD(1M) blueprint, used $\eta = 10$, $\alpha = 0.05$, and set ρ to be uniform. For particle filtering, we sampled 10 particles for the uniform blueprint and 100 particles for the MMD(1M) blueprint from the start of the game independently at every decision point to reduce bias. For the baselines, we used the same setup as Sokota et al. (2023a).

B.4. Small Comparison to 1-KLSS

We provide a very limited comparison between MMD-UES and 1-KLSS (Zhang & Sandholm, 2021) in two of the small games in which Zhang & Sandholm (2021) provided results: Kuhn poker and 2x2 Abrupt Dark Hex. We computed blueprints of similar exploitability values using MMD. We show results in Table 6. In Kuhn Poker, we find that 1-KLSS yields a larger improvement while, in 2x2 Abrupt Dark Hex, we find that MMD-UES yields a larger improvement. We emphasize that these results should not be read into too much, as this comparison is very limited and in a very different qualitative regime from the large games for which 1-KLSS and MMD-UES are intended.

	1-KLSS		MMD-UES	
	BP	Search	BP	Search
Kuhn Poker	.0124	.0015	.0123	.0114
2x2 Abrupt Dark Hex	.0683	.0625	.0664	.0282

Table 6. Small comparison against 1-KLSS. For MMD we used $\eta = 8, \alpha = 0.2$; for 2x2 Abrupt Dark Hex we used $\eta = 0.2, \alpha = 0.05$