Computing Perfect Bayesian Equilibria, with Application to Empirical Game-Theoretic Analysis

Christine Konicki*
Michigan Tech Research Institute
Ann Arbor, USA
ckonicki@mtu.edu

Mithun Chakraborty
University of Michigan
Ann Arbor, USA
dcsmc@umich.edu

Michael P. Wellman
University of Michigan
Ann Arbor, USA
wellman@umich.edu

ABSTRACT

Perfect Bayesian Equilibrium (PBE) is a refinement of the Nash equilibrium for imperfect-information extensive-form games (EFGs) that enforces consistency between the two components of a solution: agents' strategy profile describing their decisions at information sets and the belief system quantifying their uncertainty over histories within an information set. We present a scalable approach for computing a PBE of an arbitrary two-player EFG. We adopt the definition of PBE enunciated by Bonanno in 2011 using a consistency concept based on the theory of belief revision due to Alchourrón, Gärdenfors, and Makinson. Our algorithm for finding a PBE is an adaptation of Counterfactual Regret Minimization (CFR) that minimizes the expected regret at each information set given a belief system, while maintaining the necessary consistency criteria. We prove that our algorithm is correct for two-player zero-sum games and has a reasonable slowdown in time-complexity relative to classical CFR given the additional computation needed for refinement. We also experimentally demonstrate the competent performance of PBE-CFR in terms of equilibrium quality and running time on medium-to-large non-zero-sum EFGs. Finally, we investigate the effectiveness of using PBE for strategy exploration in empirical game-theoretic analysis. Specifically, we compute PBE as a meta-strategy solver (MSS) in a tree-exploiting variant of Policy Space Response Oracles (TE-PSRO). Our experiments show that PBE as an MSS leads to higher-quality empirical EFG models with complex imperfect information structures compared to MSSs based on an unrefined Nash equilibrium.

KEYWORDS

Perfect Bayesian Equilibrium, Extensive-Form Empirical Game, Policy Space Response Oracles

ACM Reference Format:
Christine Konicki*, Mithun Chakraborty, and Michael P. Wellman. 2026. Computing Perfect Bayesian Equilibria, with Application to Empirical Game-Theoretic Analysis. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 9 pages. https://doi.org/10.65109/WQXT1004

*Konicki worked on this paper while a PhD student at the University of Michigan.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/WQXT1004

1 INTRODUCTION

Game theory offers a variety of approaches for formally representing strategic interactions among several autonomous agents and reasoning about their outcomes with the help of *solution concepts*. The preeminent game-theoretic solution concept is the Nash Equilibrium (NE), a strategy profile such that no agent can improve its payoff by unilaterally deviating from the profile. Ever since its introduction and the proof of its guaranteed existence for finite games [23], the NE has been the focus of several threads of theoretical and empirical research.

An important thread concerns *refinements* of the NE for *extensive-form games* (EFGs), tree-based representations of dynamic multi-agent interactions that explicitly capture the sequential nature of action-taking and conditioning on observations. In general, the NE of a game is non-unique, and a refinement is a set of criteria that selects plausible outcomes from among all Nash equilibria, given the characteristics of a class of games. The *subgame perfect equilibrium* (SPE) [26] is a natural refinement for an EFG with perfect information (i.e., when every agent knows the full game history leading up to each of its decision points). An SPE rules out non-credible threats by requiring the solution to induce a NE in each subgame. Under imperfect information, EFGs use the device of *information sets* to represent agents' inability to distinguish certain game histories; Kaminski [15] generalized SPEs to (potentially infinite) imperfect-information EFGs by refining the definition of subgames such that every information set is contained within a single subgame.

The most powerful NE refinements for imperfect-information EFGs augment the game solution space from that of strategy profiles to that of *assessments*. An assessment consists of a strategy profile and a *belief system*, a quantification of each agent's uncertainty over all decision points in each of its information sets via probability distributions. To be an equilibrium, an assessment must meet two conditions. First, it must satisfy *sequential rationality*, which stipulates that no unilateral deviation can improve expected utility at any information set. Second, all distributions induced by the assessment's strategies and beliefs must conform to Bayes' rule. However, game theorists have also given much thought to additional notions of *consistency* to be enforced *between* the two components of an assessment to address further plausibility issues, resulting in a few different refined solution concepts. Kreps and Wilson [18] proposed the *sequential equilibrium* (SE) that satisfies a topological consistency notion, called KW-consistency by Bonanno [5], which is unintuitive and hard to verify. A simplification of SE called the *weak sequential equilibrium* [22] imposes conformity to Bayes' rule only on information sets reached with positive probability under the strategy profile. Fudenberg and Tirole [12] introduced a solution concept of intermediate strength that they termed *perfect*

Bayesian equilibrium, but they demonstrated its construction only for a restricted class of games called multi-stage signaling games. A major issue with the practical implementation of these weaker equilibria is that the lack of concrete, general consistency restrictions on off-equilibrium paths makes them unsuitable as bases for general-purpose game-solving algorithms.

Bonanno [4, 5] introduced a new consistency notion that covers both on- and off-equilibrium paths, based on the theory of belief revision due to Alchourrón, Gärdenfors, and Makinson [1], which he termed *AGM-consistency* after the original authors. This notion requires the concept of *plausibility orders* over the nodes (representing histories or, equivalently, decision points) of the EFG tree based entirely on structural properties (edge incidence and information set membership). An assessment is AGM-consistent if it is possible to construct a plausibility order such that positive probabilities are assigned to nodes or edges of the EFG tree by the assessment if and only if they satisfy certain relationships under the order in question. Bonanno [5] reused the same term 'perfect Bayesian equilibrium' (PBE) for this refinement using AGM-consistency, which is still weaker than KW-consistency but easier to algorithmically verify in principle—we will use this definition of PBE in this paper. Although it is known that every finite EFG admits at least one PBE [5], the implementation and evaluation of robust, scalable algorithmic approaches towards the computation of a PBE for arbitrary dynamic games of imperfect information is an important practical question. This is the primary research question that motivates this work.

The framework of empirical game-theoretic analysis (EGTA) [31] is highly relevant to our work. For multiagent scenarios that are too complicated for an analytic representation but admit procedural descriptions that can be queried (e.g., a simulator), EGTA offers a toolkit for using data collected from such queries to estimate a coarser model, called an *empirical game*; a key idea is to make this model amenable to off-the-shelf game-solving algorithms so that approximate insights about the underlying scenario can be derived from it. A popular and powerful iterative approach to EGTA called *policy space response oracles* (PSRO) [3, 20] uses an arbitrary game-solving algorithm as a module called the *meta-strategy solver* (MSS), which provides a principled basis for exploring the underlying strategy space to augment the model. Whereas the empirical game in EGTA has commonly been maintained in the less expressive normal form, we developed a *tree-exploiting* variant of EGTA (TE-EGTA) in prior work [16, 17] that represents the empirical game as an EFG tree. This approach makes use of refined solution concepts feasible for empirical games and potentially conducive to higher-quality game models, and the MSS for a tree-based model provides a natural use case for our algorithmic contributions in this paper.

1.1 Our Contributions

We propose a novel practical algorithm PBE-CFR (Algorithms 3.1, C.2, C.3) for computing a PBE [5] in arbitrary two-player EFGs. It is a non-trivial adaptation of the classic Counterfactual Regret Minimization (CFR) algorithm [33] that minimizes the expected regret at each information set given a belief system, while enforcing AGM-consistency.

- For two-player zero-sum games, we prove that the algorithm is correct by establishing a guarantee of convergence to an exactly sequentially rational solution and analyze its space and time complexity (Section 4).
- For two qualitatively different classes of two-player general-sum games, we experimentally demonstrate that PBE-CFR performs competently in practice both in approximation quality and running time (Sections 5.1 and 5.2).

We also report experiments that demonstrate the usefulness of PBE-CFR, vis-à-vis an unrefined NE obtained by classical CFR, for strategy exploration in TE-PSRO [16, 17]. In particular, we characterize how the speed of convergence to zero of the regret of the TE-PSRO empirical model with PBE as the MSS depends on the degree of coarsening of the information structure of the underlying game (Section 5.3). All appendices are available in the full version of the paper at http://arxiv.org/abs/2602.15233. The code for our implementation of PBE-CFR and all our experiments can be found at https://github.com/ckonicki-umich/AAMAS26/.

1.2 Further Related Work

We provide a more detailed review of consistency concepts for NE refinements for EFGs in Apps. H.1–H.3. Wellman et al. [31] provides an overview of EGTA techniques including PSRO. A body of work exists on algorithms for computing (approximate) NE refinements [2, 13, 24, 29, 30], but the scalability and practicality of these algorithms has not been adequately established, to the best of our knowledge; please see App. H.4 for further details. There is a rich literature on extensions of the CFR approach including warm-start CFR [7], CFR$^+$ [28], CFR-D [10], discounted CFR [9], linear CFR [8], deep CFR [6], Monte Carlo CFR [19], PCFR$^+$ [11], and dynamic discounted CFR [32]. In prior work, we developed a scalable, modular implementation [17] of the generalized backward induction algorithm [15] for computing an SPE of an imperfect-information EFG, and found in experiments that TE-PSRO with an SPE as MSS converges to a high-quality model faster than with an unrefined NE as MSS for diverse game classes. A similar treatment of PBE is a natural next step.

2 TECHNICAL PRELIMINARIES

A finite imperfect-information ***extensive-form game*** (EFG) is a tuple $G := \langle N, H, V, \{\mathcal{I}_j\}_{j=0}^{n}, \{A_j\}_{j=1}^{n}, X, P, u\rangle$, where

- $N = \{0, \ldots, n\}$ is the player set. Player 0 denotes ***Nature***, a non-strategic agent responsible for chance events that impact the course of play.
- H, the ***game tree***, is a finite tree rooted at node h_0 that captures players' dynamic interactions. Each node $h \in H$ represents a ***state*** or, equivalently, a ***history*** of the game beginning at h_0 (which corresponds to the empty history $\emptyset$). The ***terminal nodes*** $Z \subset H$ or leaves of the game tree represent possible end-states of the game. The remaining nodes $D = H \setminus Z$ are ***decision nodes***.
- $V : D \to N$ assigns a player to each decision node h. A node h where $V(h) = 0$ is called a ***chance node***.
- For each player $j \in N$, the set $\mathcal{I}_j$ is a partition of $V^{-1}(j)$ where each $I \in \mathcal{I}_j$ is an ***information set*** of j. All nodes $h \in I$ are indistinguishable to player j. $I(h)$ denotes the information set to which a node h belongs. We assume all information sets to be consistent with perfect recall [27, Definition 5.2.3].

- $A_j(I)$ denotes actions that player j can take at information set $I \in \mathcal{I}_j$.
- $X(h)$ is the set of possible outcomes of Nature's stochastic event at h.
- $P(\cdot|h)$ is the probability distribution over $X(h)$.
- $u : Z \to \mathbb{R}^n$ maps each terminal node z to a vector of players' ***utilities*** $\{u_j(z)\}_{j=1}^n$.

The directed edge connecting any $h \in I$ to its child represents a state transition resulting from $V(h)$'s move and is labeled with an outcome $x \in X(h)$ if $V(h) = 0$ or an action $a \in A_{V(h)}(I)$ otherwise, the child-node being denoted by hx or hs respectively. The set of nodes within H that succeed a given node h is denoted by $\text{Succ}(h)$. The function φ maps each node $h \in I \in \mathcal{I}_j$ to the ordered sequence of actions and chance outcomes observable to j from the root node leading up to I, according to the designated rules of the (imperfect-information) game. When the input is a terminal node $z \in Z$, which does not belong to any information set, φ returns a complete history from z to the root node, or the sequence of a specific player's actions given z and $j \in N$.

A ***pure strategy*** $\pi_j(\cdot)$ for player $j \in N \setminus \{0\}$ specifies the action $a \in A_j(I)$ that j selects at each information set $I \in \mathcal{I}_j$. More generally, a ***mixed strategy*** $\sigma_j(\cdot|I)$ defines a probability distribution over $A_j(I)$ at each information set of agent j where an action $a \in A_j(I)$ is selected with probability $\sigma_j(a|I)$. A ***strategy profile*** is a vector $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$, and $\boldsymbol{\sigma}_{-j}$ denotes the collection of strategies of all players other than j in $\boldsymbol{\sigma}$. Σ_j denotes the set of all strategies available to player j, and $\Sigma = \times_{j=1}^n \Sigma_j$ the space of strategy profiles.

The likelihood that node $h \in H$ is reached by strategy profile $\boldsymbol{\sigma}$ is given by its ***reach probability***

$$r(h, \boldsymbol{\sigma}) \coloneqq r_0(h) \prod_{j \in N \setminus \{0\}} r_j(h, \sigma_j),$$

where $r_j(h, \sigma_j)$ is the joint probability of player j choosing actions that lead to h according to σ_j at each of its decision nodes on the path to h; Nature's contribution $r_0(h)$ is the joint probability of each chance node along the path to h producing an outcome leading to h. The reach probability of information set I under $\boldsymbol{\sigma}$ is $r(I, \boldsymbol{\sigma}) = \sum_{h \in I} r(h, \boldsymbol{\sigma})$. A node or information set with a positive reach probability is said to be ***reachable*** under the given strategy profile. The ***expected utility*** of player j under a strategy profile $\boldsymbol{\sigma}$ is given by

$$U_j^E(\boldsymbol{\sigma}) \coloneqq \sum_{z \in Z} u_j(z) r(z, \boldsymbol{\sigma}).$$

The ***regret*** of player j at profile $\boldsymbol{\sigma}$ is given by

$$\text{Reg}_j(\boldsymbol{\sigma}) = \max_{\sigma \in \Sigma_j} U_j^E(\sigma, \boldsymbol{\sigma}_{-j}) - U_j^E(\boldsymbol{\sigma}).$$

We define the regret of a profile as the sum of player regrets, that is $\text{Reg}(\boldsymbol{\sigma}) = \sum_{j=1}^n \text{Reg}_j(\boldsymbol{\sigma})$. A strategy profile $\boldsymbol{\sigma}$ with $\text{Reg}(\boldsymbol{\sigma}) = 0$ is a ***Nash equilibrium*** (NE).

For any h that precedes a terminal node z, we denote by $r(z|h, \boldsymbol{\sigma})$ the conditional reach probability of z according to $\boldsymbol{\sigma}$, given that h has already been reached. That is, $r(z|h, \boldsymbol{\sigma}) = r(z, \boldsymbol{\sigma})/r(h, \boldsymbol{\sigma})$ whenever h is reachable under $\boldsymbol{\sigma}$ and $z \in \text{Succ}(h)$ as well as the joint probability of all players choosing the right actions that lead to z starting from state h according to $\boldsymbol{\sigma}$. Moreover, the ***conditional expected utility*** of player j given that it is at a node h under a strategy profile $\boldsymbol{\sigma}$ is given by

$$U_j^E(\boldsymbol{\sigma}|h) \coloneqq \sum_{z \in Z} u_j(z) r(z|h, \boldsymbol{\sigma}) = \sum_{z \in Z \cap \text{Succ}(h)} u_j(z) r(z|h, \boldsymbol{\sigma}).$$

For an EFG with at least one non-singleton information set, players' uncertainty about game states is naturally captured by a ***system of beliefs*** denoted by μ and defined as a collection of probability distributions, one for each information set I. At an information set $I \in \mathcal{I}_j$ of player $j \in N \setminus \{0\}$, $\mu(\cdot|I)$ represents player j's belief about which tree node it is actually at; $\mu(h|I) \geq 0$ for every $h \in I$, and $\sum_{h \in I} \mu(h|I) = 1$. An ordered pair $(\boldsymbol{\sigma}, \mu)$ containing a strategy profile $\boldsymbol{\sigma}$ and a system of beliefs μ is called an ***assessment*** and serves as a solution candidate for an imperfect-information EFG. App. A.1 provides an example illustrating EFGs and assessments.

2.1 Perfect Bayesian Equilibrium

We now present the definition of the ***perfect Bayesian equilibrium*** (PBE) proposed by Bonanno [5]. The three defining properties of a PBE are sequential rationality, AGM-consistency, and compatibility of beliefs with Bayes' rule throughout the game tree.

Sequential rationality is the natural extension of subgame perfection from strategies to assessments. It stipulates that an assessment must induce an NE at each player's information set, conditioned on both the player's belief distribution at that information set and the assumption that the information set has been reached during gameplay. Let $U_j^B(\boldsymbol{\sigma}, \mu|I)$ denote the ***believed utility*** of player j at information set $I \in \mathcal{I}_j$ for playing strategy σ_j while the others play the profile $\boldsymbol{\sigma}_{-j}$, given its belief $\mu(\cdot|I)$; i.e.,

$$\begin{aligned} U_j^B(\boldsymbol{\sigma}, \mu|I) &\coloneqq \sum_{h \in I} \sum_{z \in Z} \mu(h|I)\, r(z|h, \boldsymbol{\sigma})\, u_j(z) \\ &= \sum_{h \in I} \mu(h|I)\, U_j^E(\boldsymbol{\sigma}|h) \\ &= \sum_{a \in A_j(I)} \sigma_j(a|I) \left(\sum_{h \in I} \mu(h|I)\, U_j^E(\boldsymbol{\sigma}|ha) \right) \end{aligned} \quad (1)$$

Definition 2.1 (Sequential Rationality). An assessment $(\boldsymbol{\sigma}, \mu)$ is ***sequentially rational*** if, at every information set $I \in \mathcal{I}_j$ of each player $j \in N \setminus \{0\}$,

$$U_j^B(\boldsymbol{\sigma}, \mu|I) \geq U_j^B\left(\sigma_j', \boldsymbol{\sigma}_{-j}, \mu|I\right), \ \forall \sigma_j' \in \Sigma_j.$$

It is sufficient to restrict σ_j' to pure strategy deviations at I [21].

The AGM-consistency criterion is based on the concept of a ***plausibility order*** over the nodes in H defined as follows [5].

Definition 2.2 (Plausibility Order). A plausibility order is a total preorder $\precsim$ on the set H that satisfies the following conditions:[1]

- For any node $h \in D$ and any action $a \in A_{V(h)}(I(h))$, it is impossible that $ha \prec h$.
- Every node $h \in D$ has at least one action $a \in A_{V(h)}(I(h))$ such that $ha \precsim h$; each a that satisfies $ha \precsim h$ also satisfies $h'a \precsim h'$ for all $h' \in I(h)$.

[1] We say that node a is *at least as plausible* as node b if $a \precsim b$; the symbols $\prec$ and $\sim$ have standard meanings given preorder $\precsim$.

- For every chance node h and every outcome $e \in X(h)$, $he \precsim h$.

Given a history h, we say that plausibility is *preserved* in another history $h' \in \text{Succ}(h)$ if $h' \precsim h$.

Definition 2.3 (AGM-consistency [5]). An assessment $(\boldsymbol{\sigma}, \mu)$ for game G is ***AGM-consistent*** if a plausibility order $\mathcal{P}$ can be constructed on H such that:

- For each node $h \in H$ and action $a \in A_{V(h)}(I(h))$, $\boldsymbol{\sigma}(a) > 0$ if and only if $h \sim ha$ in $\mathcal{P}$;
- For each chance node $h \in H$ and possible chance outcome $x \in X(h)$, $P(x|h) > 0$ if and only if $h \sim hx$ in $\mathcal{P}$;
- For each node $h \in H$, $\mu(h|I(h)) > 0$ if and only if $h \precsim h'$ in $\mathcal{P}$ for all $h' \in I(h)$.

A plausibility order $\mathcal{P}$ satisfying the three conditions in Definition 2.3 is said to ***rationalize*** the assessment $(\boldsymbol{\sigma}, \mu)$.

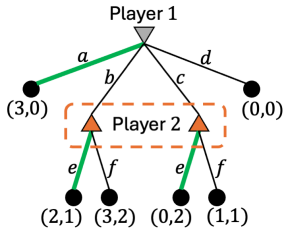


Figure 1: Example of an imperfect-information EFG from Bonanno [5], augmented with leaf utilities. There is one non-singleton information set (for Player 2) represented by the orange box. The equilibrium path induced by the AGM-consistent assessment (σ^*, μ^*) described in Example 2.5 is highlighted in green.

With this background in place and assuming familiarity with Bayes' rule, we now furnish the definition of PBE that we will use in the rest of the paper.

Definition 2.4 (Perfect Bayesian Equilibrium [5]). An assessment $(\boldsymbol{\sigma}, \mu)$ is a ***perfect Bayesian equilibrium*** for a given imperfect-information game G if it satisfies sequential rationality (Definition 2.1) and AGM-consistency (Definition 2.3), and every distribution in μ follows Bayes' rule given $\boldsymbol{\sigma}$; specifically, at every reachable information set $I \in \mathcal{I}_j$ for every player $j \in N \setminus \{0\}$ and every $h \in I$, $\mu(h) = \frac{r(h,\boldsymbol{\sigma})}{r(I,\boldsymbol{\sigma})} = \frac{r(h,\boldsymbol{\sigma})}{\sum_{h' \in I} r(h',\boldsymbol{\sigma})}$.

The example below illustrates a PBE of an imperfect-information EFG with emphasis on its AGM-consistency; for an example of an assessment violating AGM-consistency, see App. A.2.

Example 2.5. Consider the 2-player, imperfect-information EFG depicted in Figure 1. Let $(\boldsymbol{\sigma}^*, \mu^*)$ be an assessment of this EFG where $\boldsymbol{\sigma}^*$ assigns a probability of 1 to each of actions a and e and 0 to every other edge, and $\mu^*(c) = 1$. To rationalize $(\boldsymbol{\sigma}^*, \mu^*)$, a plausibility order must require that $a \sim \emptyset$ and $a \precsim b, c, d$, since $\sigma_1^*(a) = 1$. Likewise, since $\sigma_2^*(e) = 1$, $b \sim be$ and $c \sim ce$. By extension, this means that $b \precsim bf$ and $c \precsim cf$. Since $\mu^*(c) = 1$ (and hence $\mu^*(b) = 0$), we must have that $c \precsim b$. Moreover, transitivity entails that $be \precsim bf$ and $ce \precsim cf$. No contradictions arise in this construction; in fact, there are multiple plausibility orders that rationalize $(\boldsymbol{\sigma}^*, \mu^*)$ depending on where nodes cf and d are placed in the order. Therefore, $(\boldsymbol{\sigma}^*, \mu^*)$ satisfies AGM-consistency; it trivially conforms to Bayes' rule, and it can be checked algebraically from definitions that it is also sequentially rational.

3 ALGORITHM FOR FINDING PBE

Before presenting our main algorithmic contribution, we will mention a collection of algorithms that we devised to verify whether a given assessment is a PBE of a given imperfect-information EFG, each focusing on one of the three conditions in Definition 2.4. We present pseudocode and written descriptions of verification methods IsSequentRational, SatisfiesBayes, and IsConsistent, respectively, in App. B; since PBE-CFR applies to two-player EFGs, we present the two-player versions of these procedures as Algorithms B.1, B.2, and B.3 respectively, but they can be naturally extended to an arbitrary number of players. In the rest of the paper, we will sometimes use $\boldsymbol{\sigma}(I)(a)$ to denote the probability assigned to action a at information set I by the strategy profile $\boldsymbol{\sigma}$ (i.e., $\sigma_j(a|I)$ where j is the player active at I).

We now present our central contribution PBE-CFR, an algorithm for computing a PBE of a given EFG; Algorithms 3.1, C.2, and C.3 provide the pseudocode for the main algorithm and its subroutines. PBE-CFR is an adaptation of CFR that minimizes what we call the ***believed regret*** of playing $\boldsymbol{\sigma}$ at each information set given a belief system μ while keeping μ consistent.

Let $U_j^B\left(\mu^t, \boldsymbol{\sigma}^t\big|_{I \to a} | I\right)$ denote the ***believed action utility*** of playing action a at I in iteration t of the algorithm. It can be computed in a way similar to $U_j^B(\boldsymbol{\sigma}, \mu | I)$ in Equation (1) except for marginalization over $A_j(I)$. In addition, we define

$$R_{j,imm}^T(I)(a) := \frac{1}{T}\sum_{t=1}^{T}\left[U_j^B\left(\mu^t, \boldsymbol{\sigma}^t\big|_{I\to a} | I\right) - U_j^B\left(\mu^t, \boldsymbol{\sigma}^t | I\right)\right]$$

Then, the ***immediate believed regret*** of playing $\boldsymbol{\sigma}$ at information set I at timestep T is given by

$$R_{j,imm}^T(I) := \max_{a \in A_j(I)} R_{j,imm}^T(I)(a)$$

An action $a^* \in \arg\max R_{j,imm}^T(I)$ is a local best response given I was reached. We now have all the notation we need for the pseudocode of PBE-CFR (Algorithm 3.1) and a sketch of all associated proofs (Section 4; see App. D for details).

We now describe the scheme of PBE-CFR in terms of two major but natural modifications to the original CFR algorithm [33]. First, in CFR, the counterfactual regrets of player j's strategy at information set I are weighted by the probability that I was reached by $\boldsymbol{\sigma}_{-j}$, given that player j played to reach I. Furthermore, when computing the average strategy for I at the end of CFR, every strategy $\boldsymbol{\sigma}_j^t(I)(a)$ is weighted by the likelihood $r_j(\boldsymbol{\sigma}^t, I)$ of that state being reached by I. Instead in PBE-CFR, we compute the believed utility $U^B(\boldsymbol{\sigma}, \mu | I)$ at every information set I given strategy profile $\boldsymbol{\sigma}$ and belief system μ (Definition 2.1), *given that it was reached* by $\boldsymbol{\sigma}$. Hence, we exclude the aforementioned probability of reaching I associated with $\boldsymbol{\sigma}_{-j}$ as part of $U^B(\boldsymbol{\sigma}, \mu | I)$ and also $r_j(\boldsymbol{\sigma}^t, I)$ at the end when computing the

Algorithm 3.1 PBE-CFR

Require: Input game G, number of timesteps T

for $I \in G$ **do**
  $j = V(I)$
  $\boldsymbol{\sigma}^1(I)(a) \leftarrow \frac{1}{|A_j(I)|}$ for all $a \in A_j(I)$
  $\mu(h \mid I) \leftarrow \frac{1}{|I(h)|}$ for all $h \in I$
  Initialize $R^T_{j,imm}(I)(a) \leftarrow 0$ for all $a \in A_j(I)$
  Initialize cumulative infoset strategy weights $S_I(a) \leftarrow 0$ for all $a \in A_j(I)$
  Initialize $U^E(\cdot \mid h) = 0$ for all $h \in I$
  Initialize $U^B(\cdot \mid I) = 0$
end for
for $t \in \{1, \ldots, T\}$ **do**
  $U^E(\boldsymbol{\sigma}^t \mid \emptyset) \leftarrow \textsc{TraverseWithBeliefs}\left(G, \emptyset, U^E, \mathbf{1}_3, \boldsymbol{\sigma}^t, \mu^t\right)$
  $\mu \leftarrow \textsc{UpdateBeliefs}(G, \boldsymbol{\sigma}^{t+1})$
end for
for $I \in G$ **do**
  $\boldsymbol{\sigma}^*(I) \leftarrow \textsc{Average}\left(\{\boldsymbol{\sigma}^t(I)\}_{t=1}^T\right)$
end for
$\mu^* \leftarrow \textsc{UpdateBeliefs}(G, \boldsymbol{\sigma}^*)$
return $\boldsymbol{\sigma}^*, \mu^*$

average strategy. Moreover, the immediate believed regret $R^T_{j,imm}(I)$ is computed cumulatively using the strategy $\boldsymbol{\sigma}^t$ at timestep t, the belief that node $h \in I$ has been reached $\mu^t(h \mid I)$ at timestep t, and the expected utility of taking each action $a \in A(I)$ at node h, $U^E_j\left(\boldsymbol{\sigma}^t\big|_{I \to a} \mid ha\right)$. U^E_j is computed separately during recursive calls to TraverseWithBeliefs (Algorithm C.2).

The second change is that after updating $\boldsymbol{\sigma}$ for timestep $t+1$ and returning from the original call to TraverseWithBeliefs, μ is also updated for the next timestep using UpdateBeliefs (Algorithm C.3). UpdateBeliefs first constructs a plausibility order $\mathcal{P}$ given $\boldsymbol{\sigma}^{t+1}$ and then computes μ^{t+1} for each information set I as follows. If I is off of the equilibrium path, the nodes of I are divided into two tiers according to their relative plausibilities in $\mathcal{P}$, with the most plausible nodes being added to set V. $\mu^{t+1}(\cdot \mid I)$ is set to the uniform distribution over all nodes in V and 0 for all nodes excluded from V. If I is on the equilibrium path, meaning $r(I, \boldsymbol{\sigma}^{t+1}) > 0$, then μ^{t+1} is updated using Bayes' rule and the reach probabilities of each node in I given $\boldsymbol{\sigma}^{t+1}$.

4 THEORETICAL RESULTS

Our first result establishes that the space and time complexity of PBE-CFR is polynomial as a function of the input game size and the number of timesteps T. The proof is relegated to App. D.1.

Theorem 4.1. *The worst-case space and time complexities of PBE-CFR are $O(|H| \cdot |A_{max}|^2)$ and $O(T \cdot |H| \cdot |A_{max}|^2)$ respectively, where A_{max} is the largest action set across all players' information sets.*

Next, we prove that PBE-CFR is guaranteed to converge to a PBE for two-player zero-sum EFGs. We will use the concept of ***local sequential rationality*** which means that the property of sequential rationality at information set I holds for all strategies that differ from $\boldsymbol{\sigma}$ only at I. Hendon et al. [14] state that if beliefs μ are consistent, we need only consider these local deviations at each information set I in order to verify sequential rationality for I. The one-shot deviation principle follows:

Definition 4.2 (One-shot deviation). Let $(\boldsymbol{\sigma}, \mu)$ be an assessment that satisfies ***local sequential rationality*** at every information set, meaning that the property of sequential rationality at information set I holds for all strategies that differ from $\boldsymbol{\sigma}$ only at I. If $(\boldsymbol{\sigma}, \mu)$ is also consistent, then $(\boldsymbol{\sigma}, \mu)$ is sequentially rational and therefore a sequential equilibrium.

We prove that the assessment $(\boldsymbol{\sigma}^*, \mu^*)$ returned by PBE-CFR satisfies sequential rationality at every player information set.

Theorem 4.3. *In a two-player zero-sum game, for any information set $I \in \mathcal{I}_j$, $j \neq 0$, a consistent assessment $(\boldsymbol{\sigma}^*, \mu^*)$, and any strategy profile $\sigma'_j \in \Sigma_j$,*

$$U^B_j\left(\sigma'_j \boldsymbol{\sigma}^*_{-j}, \mu^* \mid I\right) \leq U^B_j\left(\boldsymbol{\sigma}^*, \mu^* \mid I\right).$$

We break the proof down into lemmas and provide all omitted proofs of these lemmas in Appendices D.2 through D.5. We first demonstrate that the immediate believed regret at any information set I after running PBE-CFR for T timesteps, given by $R^T_{j,imm}(I)$, is equal to the immediate believed regret of the average strategy $\boldsymbol{\sigma}^*$ given a consistent belief μ^*.

Lemma 4.4. *$(\boldsymbol{\sigma}^*, \mu^*)$ is an AGM-consistent assessment rationalized by plausibility order $\mathcal{P}$, and μ is Bayesian relative to $\mathcal{P}$.*

Absent the algorithm, the immediate believed regret of the returned assessment $(\boldsymbol{\sigma}^*, \mu^*)$ at information set I is given by

$$R^*_{j,imm}(I) = \max_{a \in A_j(I)} U^B_j\left(\mu^*, \boldsymbol{\sigma}^*\big|_{I \to a} \mid I\right) - U^B_j\left(\mu^*, \boldsymbol{\sigma}^* \mid I\right).$$

If the immediate believed regret after T timesteps at information set I given the strategy $\boldsymbol{\sigma}^t$ and belief μ^t at each timestep t can be written in accordance with the domain of regret-matching, Blackwell's approachability theorem applies, and convergence is guaranteed for two-player zero-sum games. In a zero-sum game, the range of utilities to player j is $\Delta_{u,j} = \max_{z \in Z} u_j(z) - \min_{z \in Z} u_j(z)$; given this range, we have the following lemma for convergence:

Lemma 4.5. *For any information set $I \in \mathcal{I}_j$ in a two-player zero-sum game, where $R^*_{j,imm}(I)$ denotes the immediate believed regret of the average strategy $\boldsymbol{\sigma}^*$ given belief μ^* at I and $R^T_{j,imm}(I)$ denotes the cumulative immediate believed regret at I after T timesteps,*

$$R^*_{j,imm}(I) \leq R^T_{j,imm}(I) \leq \varepsilon,$$

satisfying local sequential rationality for large enough T where

$$T \leq \left(\frac{\Delta_{u,j}|A_j(I)|}{\varepsilon}\right)^2.$$

We now show that the one-shot deviation principle is satisfied, completing the proof of Theorem 4.3.

Lemma 4.6. *For a given finite EFG G, any player j, and a consistent assessment $(\mu^*, \boldsymbol{\sigma}^*)$ learned through PBE-CFR, if $\pi'_j = \{a \in \arg\max_{a \in A(I)} R^*_{j,imm}(I)\}_{I \in \mathcal{I}_j}$, then π'_j is a sequential best response to $(\mu^*, \boldsymbol{\sigma}^*)$ $\iff$ $\pi'_j(I)$ is a local best response to $(\mu^*, \pi'_j, \boldsymbol{\sigma}^*_{-j})$ for all $I \in \mathcal{I}_j$.*

Lemma 4.7. *If local sequential rationality is satisfied at every information set by strategy π'_j, then the consistent assessment $(\boldsymbol{\sigma}^*, \mu^*)$ is also sequentially rational, with $R^*_{j,imm}(I) \leq \frac{\Delta_{u,j}|A_j(I)|}{\sqrt{T}}$ at every information set.*

5 EXPERIMENTS

5.1 Experimental Setup

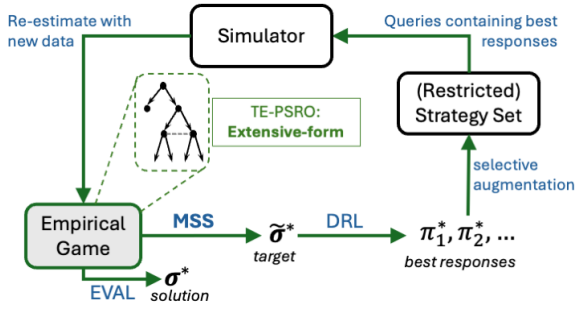


Figure 2: TE-PSRO Schematic: Empirical game is extensive-form, so PBE may be used as MSS and/or EVAL.

We begin with an overview the TE-PSRO framework [16, 17] and the two parameterized classes of general-sum imperfect-information games, GenGoof and Bargain, which we use in our experiments.

Policy space response oracles (PSRO) [20] is a powerful implementation of the empirical game-theoretic analysis (EGTA) approach [31]. Given access to a simulator that encapsulates the full procedural description of a prohibitively complex game (called the ***true game***), EGTA uses accumulated simulation data to induce a coarser but more tractable model of the game called the ***empirical game***. The empirical game generally covers a restricted space of the original strategy profile space. In iterative approaches to EGTA, analysis of the empirical game drives further refinement of the model through extension of the profile space.

In PSRO, the model is updated in the following steps, illustrated in Figure 2). First, an arbitrary game-solving algorithm, called the meta-strategy solver (MSS) in this context, is applied to the current empirical game to obtain a solution called the *target*. Then, each agent's best response (BR) to this target is approximated by a single-agent deep reinforcement learning (DRL) approach in an environment consistent with the simulator. Finally, agents' strategy sets are augmented with the respective BRs, and the simulator is queried to obtain further data (payoff information, in particular) to complete the latest empirical game iterate. Moreover, a game-solver termed EVAL, not necessarily the same as the MSS, gauges model quality and decides whether further refinement is necessary.

In prior work [16], we introduced the Tree-Exploiting PSRO (TE-PSRO) variant where the empirical game is in extensive form, though still a coarser version of a full description of the true game. We followed up with methodological advances to improve the tradeoff between tractability and fidelity of the induced empirical game, and hence the scalability of TE-PSRO for imperfect-information games [17]. We devised an *abstraction framework* in which each edge of the extensive-form empirical game represents a DRL-derived implicit *policy* executable in the simulator, allowing much of the underlying state and observation spaces to remain implicit in the model. We also employed a parameterized heuristic to control the growth of the empirical game tree by adding edges induced by the latest BR policies at select information sets only. For a fixed integer M, we first estimated the gains of playing BR policies rather than target policies at candidate information sets of the current model, constructed a softmax distribution over these information sets using those gains, and then sampled (up to) M information sets from this distribution for adding edges to. We adopt this framework in this paper too and call M the ***growth parameter***.

GenGoof$_K$ [17], parameterized by a positive integer $K > 1$, generalizes the 2-player version of the widely studied symmetric zero-sum card game Goofspiel [25] to $K-1$ rounds and arbitrary real-valued utilities. We start with a support of K possible stochastic outcomes and a categorical distribution over them sampled uniformly at random. At the start of each round, Nature uniformly samples one outcome without replacement, re-normalizing the distribution over the residual support for the next round; then, players 1 and 2 sequentially choose one of K respective actions each, observing the full history of all previous rounds and the latest revealed stochastic outcome. For each triplet of stochastic outcome and players' actions, a uniformly random finite reward is sampled for each player and publicly revealed; the utility of each player on termination is the sum of rewards over all rounds.

Additionally, we introduce a novel modified version of this game class called **PrivateGenGoof$_K$** which differs from GenGoof$_K$ in the following way only: in each round, player 2 observes player 1's action before moving but neither player observes the revealed stochastic outcome, the history of past rounds still being public. If the true game is PrivateGenGoof, we tend to have more non-singleton information sets in empirical games in TE-PSRO iterations than those for GenGoof.

Bargain [17] is a finite-horizon negotiation game where two players engage in an alternating-offer bargaining protocol to decide how to split a public pool of indivisible items of multiple types between themselves. Each player has a vector of private valuations over item types, satisfying mild assumptions, as well as an *outside offer* in the form of a private set of items of the same types. At the start of the game, Nature picks valuation vectors and outside offers from public probability distributions. Each player is also allowed to communicate to the other a binary signal (high/low) indicating whether the value of their private offer exceeds a fixed threshold; we encode the decision of whether or not to disclose this coarsened information by another binary signal (true/false) called *revelation*. The game proceeds in rounds, with players 1 and 2 sequentially taking one action each from the following options in each round: accept the other player's latest offer (if any), walk away (ending the game), or produce an offer-revelation combination. An offer takes the form of a proposed partition of the pool between the agents. If an offer is accepted by the other player in any round, the game stops, the pool is split accordingly, and each agent's realized utility is the total value of their share in the split; otherwise, negotiation fails and each agent receives their outside offer. Each agent's utility is geometrically discounted over rounds.

Detailed descriptions of these game classes along with additional references and respective parameter configurations used in

our experiments are available in Appendices E.1 (GenGoof), E.2 (PrivateGenGoof) and F (Bargain).

5.2 PBE-CFR Performance Evaluation

In our first set of experiments, we estimated the effectiveness of PBE-CFR in approximating a PBE of a general-sum imperfect-information game as well as the memory and wall time needed for convergence. We generated test games of varying complexity by running multiple iterations of TE-PSRO (which we call epochs to distinguish them from CFR/PBE-CFR iterations) on several parameterized instances of PrivateGenGoof$_4$ and PrivateGenGoof$_5$; in each epoch, we used deep Q-networks for best-response approximation, using the same methodology as Konicki et al. [17]. This resulted in approximately 1200 empirical games for PrivateGenGoof$_4$ and approximately 800 for PrivateGenGoof$_5$ across all epochs. 2 and 3 GB of memory were sufficient for completing every full TE-PSRO run for PrivateGenGoof$_4$ and PrivateGenGoof$_5$ respectively on our local computing cluster using a single core.

We gauged the approximation quality of PBE-CFR by measuring how close it gets to achieving local sequential rationality, which implies sequential rationality by the one-shot deviation principle (Section 4). Note that the solution generated by PBE-CFR satisfies the other two defining criteria of PBE by construction. We applied PBE-CFR to each of our PrivateGenGoof empirical games with different values of the total number of PBE-CFR iterations T. For each solution, we computed the regret at each information set of not choosing another action $a \in A_j(I)$, given the assessment (σ^*, μ^*), and recorded the maximum of all these regrets, termed the ***worst-case local regret***. Table 1 shows the resulting worst-case local regrets, averaged over all empirical games for each PrivateGenGoof variant: for all T, we obtain regret values of the order of 10^{-3} to 10^{-2} for leaf utilities of the order of 10^1, with a slight reduction as T increases. This suggests that PBE-CFR closely approximates a PBE of these general-sum games.

T	PrivateGenGoof$_4$	PrivateGenGoof$_5$
500	0.0104	0.0113
1000	0.0080	0.0099
2000	0.0078	0.0097
5000	0.0073	0.0096

Table 1: Worst-case local regret of PBE in PrivateGenGoof$_4$ and PrivateGenGoof$_5$ for various values of T.

To assess speed, we applied traditional CFR with the same values of T to each empirical game in parallel to PBE-CFR, and recorded the respective running times. Figure 3 provides scatter plots of these running times against the sizes of the corresponding games measured in terms of the total number of information sets of both players for a representative value of T. PBE-CFR running times are typically larger than but of the same order of magnitude as those for CFR. The slowdown is reasonable given the additional modules that PBE-CFR needs to execute to ensure equilibrium refinement. Plots for other values of T, being qualitatively similar, are omitted.

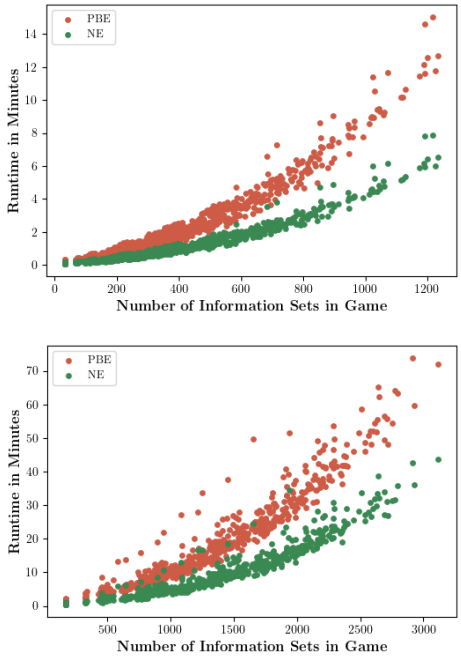
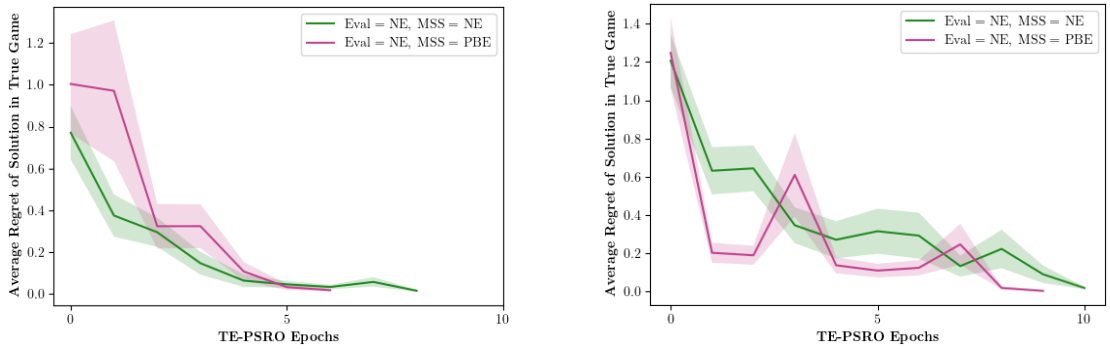


Figure 3: Time required by CFR and PBE-CFR for games generated from PrivateGenGoof$_4$ (top) and PrivateGenGoof$_5$ (bottom) with $T = 1000$.

5.3 Application to TE-PSRO as MSS

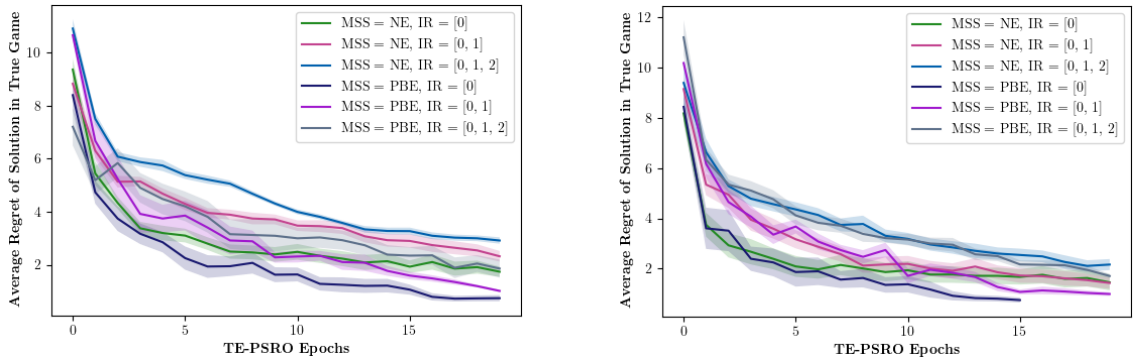
We conducted another set of experiments to evaluate the advantage that may be gained by using PBE computed using PBE-CFR as the MSS in TE-PSRO. We used traditional CFR, which approximates NE with no guarantee of refinement, as a baseline MSS for the same true game(s) and parameter configurations. We drew multiple true game instances from the Bargain and GenGoof$_4$ classes and used several values of the growth parameter M from $\{1, 2, 4, 8, 16\}$. We set $T = 500$ for both CFR and PBE-CFR. Additionally for GenGoof$_4$, we experimented with different degrees of coarseness of the empirical games by specifying which rounds' stochastic event could be included in the empirical game tree, which we denote by IR to refer to ***included rounds***, zero-indexed with respect to the root; e.g., $IR = [0, 1]$ means that the third and last stochastic event in the true game is necessarily abstracted away from every TE-PSRO-induced empirical game by construction. For each empirical game, we used the NE returned by CFR as the EVAL regardless of the MSS and used the regret of this EVAL, computed with respect to the true game, as the metric of model quality. This is the regret that we plot on the vertical axis in Figures 4 and 5.

Figure 4 shows how regret varies over TE-PSRO epochs for Bargain, averaged over 25 trials. Figure 5 shows the same for GenGoof$_4$, averaged over 5 trials, for each of the 3 IR treatments. Error bars correspond to a 95% confidence interval. These figures correspond to two representative values of M for each true game class; plots for all the values of M we considered are available in App. G. Figure 4 does not support a clear winner for Bargain: the regret curves stay close to each other while converging to approximately zero regret, with NE and PBE slightly outperforming



(a) $M = 4$ **(b)** $M = 8$

Figure 4: Average regret of σ^* evaluated in Bargain over the course of TE-PSRO's runtime, using NE or PBE as the MSS.



(a) $M = 2$ **(b)** $M = 4$

Figure 5: Average regret of σ^* evaluated in GenGoof$_4$ over the course of TE-PSRO's runtime, using NE or PBE as the MSS.

the other as an MSS for $M = 4$ and $M = 8$ respectively. By contrast, in Figure 5, PBE mostly appears to outperform NE more clearly as an MSS for GenGoof$_4$ under the same IR treatment.

We offer a potential intuitive explanation for this difference in PBE performance between the two game classes, based on our observations of the structural evolution of the respective empirical game sequences over TE-PSRO epochs. For GenGoof$_4$, player 1's action in the current round is hidden from player 2. As M increases, more information sets for both players have their action spaces augmented with new best response policies, leading to more non-singleton information sets belonging to player 2; this might be the reason why an MSS that incorporates beliefs for player 2 (PBE) is beneficial. For Bargain empirical games, imperfect information manifests only at the beginning of the game due to the opponent's outside offer signal being hidden, but this only persists as long as at least one agent keeps it signal hidden; thus, substantial portions of the empirical games for Bargain ended up containing primarily singleton information sets, rendering a refined MSS less useful.

6 DISCUSSION

We proposed the first algorithm that efficiently and effectively approximates a general PBE concept for arbitrary two-player EFGs of imperfect information. Our algorithm specifically addresses the PBE concept defined by Bonanno [5]. It is based on two non-trivial modifications to the classic CFR algorithm, which approximates unrefined NE.

Given the ability to compute PBE, we investigate the opportunity to employ PBE for strategy exploration, as the MSS in a tree-exploiting variant of PSRO. We conduct experiments on two parameterized game classes, a general-sum variant on the card game Goofspiel, and a bargaining game with signaling options. We assess effectiveness in terms of the rate of convergence to low-regret empirical games, compared to unrefined NE as MSS. Our results suggest that the benefit of PBE-as-MSS can depend significantly on structural properties of the game concerned. In particular, we found the performance of PBE-as-MSS to be better for Goofspiel than for our bargaining game, as empirical game trees for the former tended to have more downstream non-singleton information sets.

Natural future research directions include assessing PBE as an MSS for other game classes (e.g., poker) and improvements to PBE-CFR by invoking variants of CFR (Section 1.2).

ACKNOWLEDGMENTS

This work was supported in part by the US National Science Foundation under CRII Award 2153184.

REFERENCES

[1] Carlos E. Alchourrón, Peter Gärdenfors, and David C. Makinson. 1985. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic* 50, 2 (1985), 510–530.

[2] Salman Azhar, Andrew McLennan, and John H. Reif. 2005. Computation of equilibria in noncooperative games. *Computers and Mathematics with Applications* 50, 5 (2005), 823–854.

[3] Ariyan Bighashdel, Yongzhao Wang, Stephen McAleer, Rahul Savani, and Frans A. Oliehoek. 2024. Policy space response oracles: A survey. In *33rd International Joint Conference on Artificial Intelligence (IJCAI)*. 7951–7961.

[4] Giacomo Bonanno. 2011. AGM belief revision in dynamic games. In *13th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*. 37–45.

[5] Giacomo Bonanno. 2011. AGM-consistency and perfect Bayesian equilibrium. Part I: Definition and properties. *International Journal of Game Theory* 42 (2011), 562–592.

[6] Noam Brown, Adam Lerer, Sam Gross, and Tuomas Sandholm. 2019. Deep counterfactual regret minimization. In *36th International Conference on Machine Learning (ICML)*. 793–802.

[7] Noam Brown and Tuomas Sandholm. 2014. Regret transfer and parameter optimization. In *28th AAAI Conference on Artificial Intelligence (AAAI)*. 594–601.

[8] Noam Brown and Tuomas Sandholm. 2019. Solving imperfect-information games via discounted regret minimization. In *33rd AAAI Conference on Artificial Intelligence (AAAI)*. 1829–1836.

[9] Noam Brown and Tuomas Sandholm. 2019. Superhuman AI for multiplayer poker. *Science* 365, 6456 (2019), 885–890.

[10] Neil Burch, Michael Johanson, and Michael Bowling. 2014. Solving imperfect information games using decomposition. In *28th AAAI Conference on Artificial Intelligence (AAAI)*. 602–608.

[11] Gabriele Farina, Christian Kroer, and Tuomas Sandholm. 2021. Faster game solving via predictive Blackwell approachability: Connecting regret matching and mirror descent. In *35th AAAI Conference on Artificial Intelligence (AAAI)*. 5363–5371.

[12] Drew Fudenberg and Jean Tirole. 1991. Perfect Bayesian equilibrium and sequential equilibrium. *Journal of Economic Theory* 53 (1991), 236–260.

[13] Moritz Graf, Thorsten Engesser, and Bernhard Nebel. 2024. Symbolic computation of sequential equilibria. In *23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 715–723.

[14] Ebbe Hendon, Hans Jørgen Jacobsen, and Birgitte Sloth. 1996. The one-shot deviation principle for sequential rationality. *Games and Economic Behavior* 12 (1996), 274–282. Issue 2.

[15] Marek Mikolaj Kaminski. 2019. Generalized backward induction: Justification for a folk algorithm. *Games* 10 (2019). Issue 34.

[16] Christine Konicki, Mithun Chakraborty, and Michael P. Wellman. 2022. Exploiting extensive-form structure in empirical game-theoretic analysis. In *18th International Conference on Web and Internet Economics (WINE)*. 132–149.

[17] Christine Konicki, Mithun Chakraborty, and Michael P. Wellman. 2025. Policy abstraction and Nash refinement in tree-exploiting PSRO. In *24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*. 1163–1171.

[18] David Kreps and Robert Wilson. 1982. Sequential equilibrium. *Econometrica* 50 (1982), 863–894.

[19] Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael Bowling. 2009. Monte Carlo sampling for regret minimization in extensive games. In *23rd Annual Conference on Neural Information Processing Systems*.

[20] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. 2017. A unified game-theoretic approach to multiagent reinforcement learning. In *31st Annual Conference on Neural Information Processing Systems (NeurIPS)*. 4190–4203.

[21] Dustin Morrill, Ryan D'Orazio, Reca Sarfati, Marc Lanctot, James R. Wright, Amy R. Greenwald, and Michael Bowling. 2021. Hindsight and sequential rationality of correlated play. In *35th AAAI Conference on Artificial Intelligence (AAAI)*.

[22] Roger Myerson. 1991. *Game Theory*. Harvard University Press.

[23] John Nash. 1951. Non-cooperative games. *Annals of Mathematics* 54, 2 (1951), 286–295.

[24] Fabio Panozzo. 2014. *Algorithms for the verification, computation and learning of equilibria in extensive-form games*. Ph.D. Dissertation. Polytechnic University of Milan.

[25] Glenn C. Rhoads and Laurent Bartholdi. 2012. Computer solution to the game of pure strategy. *Games* 3, 4 (2012), 150–156.

[26] Richard Selten. 1965. Spieltheoretische behandlung eines oligopolmodells mit nachfrageträgheit–Teil I: Bestimmung des dynamischen preisgleichgewichts. *Zeitschrift für die gesamte Staatswissenschaft* 121 (1965), 301–324.

[27] Yoav Shoham and Kevin Leyton-Brown. 2008. *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*. Cambridge University Press.

[28] Oskari Tammelin. 2014. Solving large imperfect information games using CFR+. arXiv:1407.5042 [cs.GT] arXiv: 1407.5042.

[29] Vinzenz Thoma, Vitor Bosshard, and Sven Seuken. 2023. Computing perfect Bayesian equilibria in sequential auctions with verification. 14158–14166.

[30] Theodore L Turocy. 2010. Computing sequential equilibria using agent quantal response equilibria. *Economic Theory* 42, 1 (2010), 255–269.

[31] Michael P. Wellman, Karl Tuyls, and Amy Greenwald. 2025. Empirical game theoretic analysis: A survey. *Journal of Artificial Intelligence Research* 82 (2025), 1017–1076.

[32] Hang Xu, Kai Li, Haobo Fu, Qiang Fu, Junliang Xing, and Jian Cheng. 2024. Dynamic discounted counterfactual regret minimization. In *12th International Conference on Learning Representations (ICLR)*.

[33] Martin Zinkevich, Michael Johanson, Michael H. Bowling, and Carmelo Piccione. 2007. Regret minimization in games with incomplete information. In *21st Annual Conference on Neural Information Processing Systems*.