The Partially Observable History Process

Dustin Morrill,[1] Amy R. Greenwald,[2] Michael Bowling[1]

[1]University of Alberta; Amii. Edmonton, Alberta, Canada
[2]Brown University, Providence, Rhode Island, United States
morrill@ualberta.ca, amy_greenwald@brown.edu, mbowling@ualberta.ca

Abstract

We introduce the partially observable history process (POHP) formalism for reinforcement learning. POHP centers around the actions and observations of a single agent and abstracts away the presence of other players without reducing them to stochastic processes. Our formalism provides a streamlined interface for designing algorithms that defy categorization as exclusively single or multi-agent, and for developing theory that applies across these domains. We show how the POHP formalism unifies traditional models including the Markov decision process, the Markov game, the extensive-form game, and their partially observable extensions, without introducing burdensome technical machinery or violating the philosophical underpinnings of reinforcement learning. We illustrate the utility of our formalism by concisely exploring observable sequential rationality, examining some theoretical properties of general immediate regret minimization, and generalizing the extensive-form regret minimization (EFR) algorithm.

1 Introduction

We develop the partially observable history process (POHP) that embodies the philosophical aspects of reinforcement learning. The formalism uses a few elementary mechanisms to analyze a single agent that makes observations and takes actions. The agent is responsible for managing their own representation of an environment that is, by default, massively more complicated than themselves. They are also responsible for and capable of evaluating themselves against goals set by their designer or themselves. To canonize the central role of the agent, our formalism abstracts away any other players without nullifying their agency.

The individual components of the POHP formalism are taken from two sequential decision-making frameworks, the extensive-form game (EFG) (Kuhn 1953) and the partially observable Markov decision process (POMDP) (Smallwood and Sondik 1973), along with a repeated game framework, the online decision process (see, *e.g.*, Greenwald, Li, and Marks (2006)). The result is a sequential decision-making formalism that is conceptually simpler than either of its two sequential decision-making progenitors. Other general formalisms such as the partially observable stochastic game (Hansen, Bernstein, and Zilberstein 2004), turn-
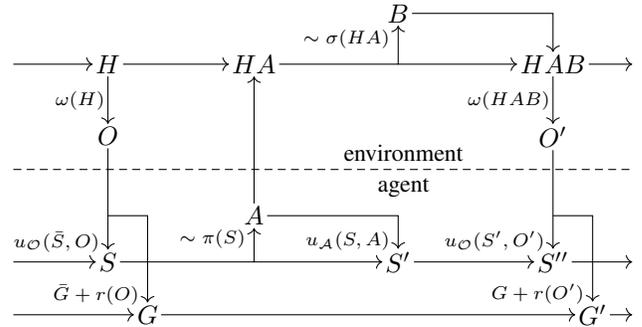
Figure 1: The evolution of a POHP environment and agent.

taking partially-observable Markov game (Greenwald, Li, and Sodomka 2017), and factored observation stochastic games (Kovařík et al. 2019) bring with them unnecessary complications for agent-centric reinforcement learning. The sequential decision-making setting presented by Farina, Kroer, and Sandholm (2019) shares spiritual similarities but it represents a less radical departure from the EFG model. Srinivasan et al. (2018)'s presentation of the EFG model using reinforcement learning and Markov decision process (MDP) terminology had a substantial influence on this work.

The POHP model is not meant replace any of these established formalisms, but rather fill a particular niche. We recommend our formalism in two cases: (i) when modeling a problem as an MDP would ignore imperfect information or the presence of other players, or (ii) when the extra structure of a more established model is unnecessary or burdensome.

2 Partially Observable History Process

The *partially observable history process* (*POHP*) model begins from the premise that an *agent* observes and influences an *environment*. We are principally concerned with the design of the agent and how well they navigate the environment. The environment may change without the agent's input and we attribute these changes to a *daimon*. Inspired by depictions in Greek mythology, our daimon is an inexplicable force that partially determines the evolution of the environment and shapes the agent's growth. The concept of a daimon is flexible enough that it can represent an adversary, a teammate, a teacher, chance, or any combination thereof.

2.1 The Environment and Daimon

The environment dynamics follow a simple continuing history model. History in this model refers to a simple ledger that permanently records actions. Given history h from the set of possible histories $\mathcal{H}$ and action a, the next history is always $ha \in \mathcal{H}$. The agent and daimon take turns choosing actions until the process terminates and the agent observes the daimon's actions only indirectly. Histories are partially ordered action strings so we use $h \sqsubset h'$ to denote that h is a predecessor of h', $|h|$ to denote the length of h, and use subscripts to reference substrings, *e.g.*, h_i is the ith action in h and $h_{\leq n}$ is the first n actions of h.

A POHP has four core objects: (i) a Boolean $\iota \in \{0, 1\}$ that indicates if the agent acts before the daimon, (ii) a function $\mathcal{A}$ that determines the set of legal actions after a given history, (iii) a function ω that generates the observation from a set of observations $\mathcal{O}$ at a given history immediately following a daimon action, and (iv) a function γ that determines the probability that the process continues after a daimon action. Two more objects are required to describe a POHP in full generality: a set of initial histories $\mathcal{H}_\varnothing$ and a probability distribution over these histories ξ. A POHP's history is initialized with one sampled from $\mathcal{H}_\varnothing$ according to ξ, but in most situations $\mathcal{H}_\varnothing$ consists of just the empty history, $\varnothing$.

The daimon behaves according to a *behavioral strategy* (also called a *policy*), σ, that assigns a probability distribution over legal actions to every history where it is the daimon's turn to act, $\mathcal{H}_\mathcal{O} = \{h \in \mathcal{H} \mid |h| \bmod 2 = \iota\}$. In history $h \in \mathcal{H}_\mathcal{O}$, the daimon chooses action $B \sim \sigma(h)$ and the history advances to hB, at which point the process continues if $\Gamma \sim \gamma(hB)$ and terminates otherwise. Since the agent waits for an observation in each $h \in \mathcal{H}_\mathcal{O}$, we call them *passive histories*, in contrast to the *active histories*, $\mathcal{H}_\mathcal{A} = \mathcal{H} \setminus \mathcal{H}_\mathcal{O}$, in which the agent acts. We do not ascribe any a priori motivation to the daimon.

2.2 An Abstract POHP Agent

The agent need only implement the observation and action interfaces provided by the POHP so there are many ways to construct the agent. The agent we work with has three conceptual modules: a state of mind, a behavior plan, and goals. Their state provides the context with which behaviors are chosen to advance them toward their goals. Formally, we define our POHP agent with a tuple, $(s_\varnothing, u_\mathcal{A}, u_\mathcal{O}, \pi, r)$.

State of mind. A snapshot of the agent's state of mind is given concrete form in their *information state*, which is set to a given initial information state $s_\varnothing$ at the beginning of the POHP. Information state evolves according to update functions $u_\mathcal{A}$ and $u_\mathcal{O}$, which describe how actions and observations are processed, respectively.[1] After the agent chooses

---

[1] Splitting information state updates into action and observation specific updates has a few benefits: (i) it avoids treating the initial observation update in POHPs where the daimon acts first as a special case, (ii) it allows us to refer to the agent's state of mind immediately after choosing an action, which we make use of in our reduction to Markov decision processes and in intermediate proof steps, and (iii) it allows the agent to forget the action they just chose without waiting for an observation, which could be useful in some

action a in information state s, their information state is updated to $s' = u_\mathcal{A}(s, a)$ while they wait for an observation. After receiving observation o, their information state is updated to $u_\mathcal{O}(s', o)$ and the agent chooses another action a' to begin the cycle again. Ultimately, each history $h \in \mathcal{H}$ yields an information state, so we recursively define a unified update function,

$$u : h \mapsto \begin{cases} s_\varnothing & \text{if } h \in \mathcal{H}_\varnothing \\ u_\mathcal{A}\big(u(h_{<|h|}), h_{|h|}\big) & \text{if } h_{<|h|} \in \mathcal{H}_\mathcal{A} \\ u_\mathcal{O}\big(u(h_{<|h|}), \omega(h)\big) & \text{o.w.} \end{cases}$$

Each information state corresponds to an *information set*, $I(s) = \{h \mid u(h) = s\}$, which is the set of histories the environment could be in given the agent's information state is s. We denote the set of information states that could ever be generated as $\mathcal{S}$, and we partition them into the passive information states where the agent awaits an observation, $\mathcal{S}_\mathcal{O}$, and the active information states where the agent acts, $\mathcal{S}_\mathcal{A}$. We overload $\mathcal{S}_\mathcal{A}(s, a) = \{u_\mathcal{O}(u_\mathcal{A}(s, a), \omega(hab))\}_{h \in I(s), b \in \mathcal{A}(ha)}$ as the set of child active information states following s and action a.

Behavior. The agent acts by sampling actions from a behavioral strategy, $\pi \in \Pi$, where probability distributions over legal actions are assigned to information states. At history h in active information state s, the agent chooses an action by sampling from their *immediate strategy* at s, $\pi(s) \in \Delta(\mathcal{A}(h))$, where $\Delta(\mathcal{A}(h))$ is the probability simplex over $\mathcal{A}(h)$. We assume that the agent can always determine the legal actions from their information state so we overload $\mathcal{A}(s) = \mathcal{A}(h)$ for all $s \in \mathcal{S}$ and $h \in s$.

Goals. A bounded reward function, $r : \mathcal{O} \to [-U, U]$, provides quantitative feedback to the agent about their progress toward their goals. The *return* (cumulative reward) that the agent acquires from active history $h \in \mathcal{H}_\mathcal{A}$ is $G_h(\pi; \sigma) = \sum_{i=1}^{\infty} Y_i r\big(\omega(H_i)\big)$, where the initial history in the trajectory is $H_1 = h$, the agent's action on each step is $A_i \sim \pi\big(u(H_i)\big)$, the daimon's action on each step is $B_i \sim \sigma(H_i A_i)$, the history is updated as the concatenation $H_{i+1} = H_i A_i B_i$, and the continuation indicator is the product $Y_{i+1} = Y_i \Gamma_i \in \{0, 1\}$ with $Y_1 = 1$ and $\Gamma_i \sim \gamma(H_i)$.

Generally, the agent's goal is to maximize their return. The fact that the daimon's strategy is unknown and their actions are only partially observed prevents us from immediately formulating this goal as an optimization problem. Neither can an equilibrium concept be proposed as a solution concept without presupposing incentives and a level of rationality for the daimon. Hindsight rationality (Morrill et al. 2021b), in contrast, is well suited as a solution concept for POHPs as it focuses on self-improvement grounded in experience and requires no assumptions about the daimon.

However, since no history may ever repeat and repetition is a key requirement of hindsight rationality, we will only consider hindsight rationality in the context of a repeated POHP. Before each round t begins, the agent chooses strategy π^t and the daimon chooses strategy σ^t. The POHP plays out according to these strategies, after which the agent receives reward information. The agent can then compare the

---

applications involving asynchronous processing.

returns they achieved with π^t with those they could have achieved with alternative behavior.[2] This online decision process is well defined as long as the POHP terminates almost surely so that the agent is unlikely to be stuck in a single round forever.

2.3 Reach Probabilities

A derived property of POHPs that we will make use of later is the reach probability. Consider random history H generated according to agent strategy π, daimon strategy σ, and continuation function γ. The probability that a given history $h_\varnothing \sqsubseteq h \sqsubseteq H$ follows from the chain rule of probability, $\mathbb{P}_{\pi,\sigma}[\bigcup_{\bar{h}\in\mathcal{H}_\varnothing} \bar{h} \sqsubseteq h \sqsubseteq H] = \prod_{i=|h_\varnothing|+1}^{|h|} \mathbb{P}_{\pi,\sigma}[h_i \mid h_{<i}]$ where

$$\mathbb{P}_{\pi,\sigma}[h_i \mid h_{<i}] = \begin{cases} \gamma(h_{<i})\pi(h_i \mid u(h_{<i})) & \text{if } h_{<i} \in \mathcal{H}_\mathcal{A} \\ \sigma(h_i \mid h_{<i}) & \text{o.w.} \end{cases}$$

We denote $\mathbb{P}_{\pi,\sigma}[h] = \mathbb{P}_{\pi,\sigma}[\bigcup_{\bar{h}\in\mathcal{H}_\varnothing} \bar{h} \sqsubseteq h \sqsubseteq H]$ and refer to this quantity as h's *reach probability*. We can decompose

$$\mathbb{P}_{\pi,\sigma}[h] = \underbrace{\prod_{i=|h_\varnothing|+1,\, i \bmod 2=\iota}^{|h|} \gamma(h_{<i-1})\sigma(h_i \mid h_{<i})}_{\mathbb{P}_\sigma[h]}$$

$$\underbrace{\prod_{i=|h_\varnothing|+1,\, i \bmod 2=1-\iota}^{|h|} \pi(h_i \mid u(h_{<i}))}_{\mathbb{P}_\pi[h]}$$

according to the probability that the daimon and agent play their parts of h, where we have grouped the continuation probabilities with the daimon's and set $\gamma(h_{<|h_\varnothing|}) = 1$.

The conditional probability

$$\mathbb{P}_{\pi,\sigma}[h' \mid h] = \mathbb{P}_{\pi,\sigma}[h \sqsubseteq H, h' \sqsubseteq H]/\mathbb{P}_{\pi,\sigma}[h]$$

is the probability that history $h' \sqsubseteq H$ given $h \sqsubseteq H$. If h' and h are unrelated in that $h' \not\sqsubseteq h \not\sqsubseteq h'$, then it is not possible for H to realize both, so the joint probability $\mathbb{P}_{\pi,\sigma}[h,h'] = 0$, and consequently $\mathbb{P}_{\pi,\sigma}[h' \mid h] = 0$. If $h' \sqsubseteq h$ then H always realizes h' when h is realized, therefore, $\mathbb{P}_{\pi,\sigma}[h,h'] = \mathbb{P}_{\pi,\sigma}[h]$ and $\mathbb{P}_{\pi,\sigma}[h' \mid h] = 1$. The last case is $h \sqsubseteq h'$, where

$$\mathbb{P}_{\pi,\sigma}[h' \mid h] = \mathbb{P}_{\pi,\sigma}[h,h']/\mathbb{P}_{\pi,\sigma}[h] = \mathbb{P}_{\pi,\sigma}[h']/\mathbb{P}_{\pi,\sigma}[h].$$

3 Representing Traditional Models

3.1 Games

A *game* is an N player interaction where each player simultaneously chooses a strategy and immediately receives a payoff from a bounded utility function (von Neumann and Morgenstern 1947). There may also be an extra "chance player", denoted c, who "decides" chance events like die rolls with strategy π_c. A game described in this way is called a *normal-form game* (NFG).

---

[2]The agent may estimate the returns for alternative behavior using importance corrections if this information is not provided explicitly at the end of each round, similarly to how reward functions are estimated in adversarial bandit contexts (see, *e.g.*, Lattimore and Szepesvári (2020)).

Algorithm 1: The procedure for playing an N player game in POHP-form.

1: **Input:** turn function $p : \mathcal{H} \to \{c\} \cup \{i\}_{i=1}^N$,
2:     legal actions function $\mathcal{A}$,
3:     terminal histories $\mathcal{Z} \subseteq \mathcal{H}$
4:     or continuation function $\gamma : \mathcal{H} \to \Delta\{0,1\}$,
5:     information partitions $\{\mathcal{I}_i\}_{i\in\{c\}\cup\{j\}_{j=1}^N}$
6:     or observation functions $\{\omega_i : \mathcal{H} \to \mathcal{S}_i\}_{i\in\{c\}\cup\{j\}_{j=1}^N}$,
7:     and utility functions $\{\upsilon_i : \mathcal{Z} \to [-U, U]\}_{i=1}^N$.
8: **for** $i \in \{c\} \cup \{j\}_{j=1}^N$ **do**
9:     $\omega_i(h) \leftarrow I$ **for** $h \in I \in \mathcal{I}_i$ **if** ω_i undefined
10: $\gamma \leftarrow h \mapsto \mathbb{1}\{h \notin \mathcal{Z}\}$ **if** γ undefined
11: $H \leftarrow \varnothing$
12: $\Gamma \leftarrow 1$
13: **while** Γ **do**
14:     **send** $\omega_i(H)$ to player $p(H)$
15:     **receive** $A \in \mathcal{A}(H)$ **from** player $p(H)$
16:     $H \leftarrow HA$
17:     **sample** $\Gamma \sim \gamma(H)$
18: **for** $i = 1, 2, \ldots, N$ **do**
19:     **send** $\omega_i(H) = (H, \upsilon_i(H))$ **to** player i

For any given player, i, we can represent i's view of the game with a POHP, $\mathcal{G}_i$, where the agent represents i and the daimon represents the other $N - 1$ players and chance in aggregate. We can also represent chance's view of the game with a POHP where the agent's strategy is fixed to π_c. The histories, action sets, and continuation function across all $N + 1$ of these POHPs are shared but the first turn indicator and observation functions are specific to each player. The reward functions for each player must also reflect the game's payoffs. After each player chooses an agent strategy for their POHP, all the POHPs are evaluated together, sharing the same history, and each player receives a return in their POHP that equals their payoff in the game. Together, the set of POHPs, $\{\mathcal{G}_i\}_{i\in\{c\}\cup\{j\}_{j=1}^N}$, represents what we could call a *POHP-form game*. See Algorithm 1 for a programmatic description of how a game can be played out in POHP form.

In each $\mathcal{G}_i$, the daimon's strategy, σ_i, must reflect those of the other players. If we have a *turn function* $p : \mathcal{H} \to \{c\} \cup \{j\}_{j=1}^N$ that determines which player acts after a given history h, we can constrain σ_i to conform to the agent strategies from the other POHPs as $\sigma_i(h) = \pi_{p(h)}(u_{p(h)}(h))$.

A game described with histories and turns is called an *extensive-form game* (EFG) (Kuhn 1953). Any NFG can be converted into extensive form by serializing each decision. Of course, players who act later are not allowed to observe previous actions, and this is traditionally specified through information partitions. Each player, i, is assigned a set of information sets as their *information partition*, denoted $\mathcal{I}_i$. Typically, EFGs also define a set of *terminal histories*, $\mathcal{Z} \subseteq \mathcal{H}$, which is constructed so that every history eventually terminates. As in a NFG, payoffs are given to players upon termination.

Since the POHP and EFG share the same history-based

progression, representing an EFG in POHP-form simply requires that information partitions, terminal histories, and utility functions are faithfully reconstructed in the POHP. If player i's observation function returns its given history's information set in $\mathcal{I}_i$ and player i's observation update function replaces the current information state with its given observation, then the set of information sets on active information states reproduces $\mathcal{I}_i$. Formally, $\{I(s)\}_{s \in \mathcal{S}_{i,\mathcal{A}}} = \mathcal{I}_i$. Constructing the information states for each player in this way ensures that all information partitions are respected. We can add terminal histories to a POHP by setting $\gamma(h) = \mathbb{1}\{h \notin \mathcal{Z}\}$. To respect the EFG's utility function for each player i, $v_i : \mathcal{Z} \to [-U, U]$, we set player i's rewards for all observations to zero except those following terminal histories, z, at which point $r_i(\omega_i(z)) = v_i(z)$.

3.2 Markov Models

The POHP model allows agents to construct their own complicated notions of state but forces conceptual simplicity on environment state, *i.e.*, its history. However, a popular class of models are Markov models where the environment has a more complicated notion of state and this state evolves according to Markovian dynamics. That is, there may be many histories that lead to the same environment state and the state on the next step is determined, up to stochasticity, by the current environment state and action. The transition probabilities must be constant across each history in an environment state, otherwise transitions would depend on past environment states, violating the Markov property. Agent information states in a POHP may lack the Markov property because the daimon's strategy may depend on the history.

A straightforward way to represent environment states is with the passive information states of the chance player in a POHP-form game. At each of chance's passive histories h, each non-chance player plays an action in turn, which advances the history to $h' = ha_1 \ldots a_N$ where chance updates their information state to $s_{h'} = u_c(h') \in \mathcal{S}_{c,\mathcal{A}}$. Chance then chooses which of their passive information states is next by sampling A_c from $\pi_c(s_{h'})$, resulting in a transition to $s_{h'A_c} = u_c(h'A_c) \in \mathcal{S}_{c,\mathcal{O}}$. A Markovian transition between s_h and $s_{h'A_c}$ can be enforced by restricting chance's observation function so that $\omega_c(ha_1 \ldots a_N) = \omega_c(\bar{h}a_1 \ldots a_N)$ for all joint player actions $a_1 \ldots a_N$ and histories $\bar{h} \in I(s_h)$. Enforcing this constraint for each history h ensures that if $u_c(\bar{h}) = s_h$, then, given joint player actions $a_1 \ldots a_N$,

$$
\begin{aligned}
u_c(\bar{h}a_1 \ldots a_N A_c) &= u_{c,\mathcal{A}}(u_{c,\mathcal{O}}(u_c(\bar{h}), \omega_c(\bar{h}a_1 \ldots a_N)), A_c) \\
&= u_{c,\mathcal{A}}(u_{c,\mathcal{O}}(s_h, \omega_c(ha_1 \ldots a_N)), A_c) \\
&= s_{h'A_c}
\end{aligned}
$$

with transition probability $\pi_c(A_c \mid s_{h'})$ for all $\bar{h}$.

A general Markovian model is the *partially observable Markov game* (*POMG*) (Hansen, Bernstein, and Zilberstein 2004).[3] In this model, each environment state represents its own NFG, so all players simultaneously choose an action

and receive a reward that depends on the state and joint action selection. The next state (and thus the next NFG) is determined by the current state and the joint action selection. Just as we can serialize any NFG by introducing a turn function and selectively hiding actions, we can serialize a POMG into a *turn-taking POMG* (*TT-POMG*) (Greenwald, Li, and Sodomka 2017) by serializing each of its component NFGs.

The TT-POMG formalism was developed to convert EFGs into Markov models (Greenwald, Li, and Sodomka 2017), so naturally the POHP-form of a POMG is similar to that of an EFG, except that chance's observations are constrained so that chance's passive information states can play the role of the POMG environment states. The actions that each player plays are only revealed to the other players after chance's current passive information state transitions to the next one, corresponding to the POMG's next NFG. Players also receive a reward at this time based on the utilities of the previous NFG. This model is typically presented as a continuing process with discounting, and we can replicate the same setup by setting the continuation probability $\gamma(h)$ to the discount factor for each of chance's passive histories h and $\gamma(h') = 1$ for all other histories h'.

Providing full observability to player i in a POHP-form POMG is simply a matter of adding the constraint that chance's passive information states are isomorphic to player i's active information states. That is, there must be a bijection where chance's information state is $s' \in \mathcal{S}_{c,\mathcal{O}}$ whenever player i's information state is $s \in \mathcal{S}_{i,\mathcal{A}}$, and *vice-versa*. Effectively, player i's active information state is always chance's passive information state and thus also the POMG environment state. One trivial way to enforce this constraint is for player i's observation function to return chance's passive information state, *i.e.*, $\omega_i(ha_1 \ldots a_{i-1}) = u_c(h)$, and for player i's observation update function to replace their current state with the given observation. If all players are granted full observability, then a POMG becomes, naturally, a *Markov game* (Shapley 1953). Furthermore, a single-player Markov game or POMG reduces to a *Markov decision process* (*MDP*) or *partially observable MDP* (*POMDP*) (Smallwood and Sondik 1973), respectively, and this is true when models are represented either in their canonical or POHP-forms.

4 The Sub-POHP and Learning

We now describe how sub-POHPs can be constructed in finite-horizon POHPs with timed updates, and show how observable sequential rationality (Morrill et al. 2021b) is naturally defined in terms of sub-POHPs.

A POHP has a *finite horizon* if every history eventually terminates deterministically. We enforce this by selecting a subset of histories, $\mathcal{Z} \subseteq \mathcal{H}$ where $\gamma(z) = 0$ for all $z \in \mathcal{Z}$. The agent's updates are *timed* as long as the agent's action update function records the number of actions the agent has taken. A finite horizon and timed updates ensure that the number of histories in each information set is finite and the same information state is never encountered twice before termination. Thus, the information states are partially ordered and we can write $s \prec s'$ to denote that information state s is a predecessor of s'.

---

[3]A Markov game is also often called a "stochastic game", but a core feature of this model is Markovian transitions, not stochasticity. This leads us to prefer the term "Markov game".

4.1 Beliefs and Realization Weights

Given that the agent's information state is s, how likely is it that the agent is in a particular history $h \in I(s)$? Traditionally, this is called the agent's *belief* (about which history they are in) at s. According to Bayes' rule, $\mathbb{P}_{\pi,\sigma}[h \mid s] = \mathbb{P}_{\pi,\sigma}[s \mid h]\mathbb{P}_{\pi,\sigma}[h]/\mathbb{P}_{\pi,\sigma}[s]$. Since $h \in I(s)$, $\mathbb{P}_{\pi,\sigma}[s \mid h] = 1$. The agent's information state is s only if the random history H lands in $I(s)$, so we can describe the event of realizing s as the union of history realization events. Since we assume the agent's updates are timed, there is at most one prefix of H in $I(s)$, which means that each $h \sqsubseteq H$ event for $h \in I(s)$ is disjoint. The probability of their union is thus the sum

$$\mathbb{P}_{\pi,\sigma}[s] = \mathbb{P}_{\pi,\sigma}\left[\bigcup_{\substack{h \in I(s), \\ \bar{h} \in \mathcal{H}_{\varnothing}}} \bar{h} \sqsubseteq h \sqsubseteq H\right] = \sum_{h \in I(s)} \mathbb{P}_{\pi,\sigma}[h].$$

The agent's belief at s is $\xi_s^{\pi,\sigma} : h \mapsto \mathbb{P}_{\pi,\sigma}[h]/\mathbb{P}_{\pi,\sigma}[s]$.

An assignment of beliefs to each information state is called a *system of beliefs*. A problem that arises in defining a complete system of beliefs from a given π–σ pair is that some information states may be unrealizable ($\mathbb{P}_{\pi,\sigma}[s] = 0$). Different rationality assumptions lead to different ways of constructing complete belief systems and corresponding notions of equilibria (see, *e.g.*, Kreps and Wilson (1982); Breitmoser, Tan, and Zizzo (2010); Dekel and Siniscalchi (2015)). However, from a hindsight perspective, only realizable information states could have been observed by the agent, and only behavior in realizable states could have impacted the agent's return. Thus, beliefs at unreachable information states are naturally left undefined.

As a consequence, information state realization probabilities hold special significance in hindsight analysis, as they determine whether or not a state is observable. More generally, they provide a measure of importance to each information state. Let J be the random step in the trajectory $\{H_i\}_{i=1}^{\infty}$ where $u(H_J) = s$ or $J = \infty$ if information state s is never realized. The return from H_1 can be split as

$$G_{H_1}(\pi;\sigma) = \mathbb{1}\{J < \infty\}\sum_{i=1}^{J-1} Y_i r\big(\omega(H_i)\big)$$

$$+ \mathbb{1}\{J = \infty\}\sum_{i=1}^{\infty} Y_i r\big(\omega(H_i)\big)$$

$$+ \underbrace{\mathbb{1}\{J < \infty\}\sum_{i=J}^{\infty} Y_i r\big(\omega(H_i)\big)}_{s\text{'s contribution.}}$$

Since $H_J \sim \xi_s^{\pi,\sigma}$, the expectation of s's contribution is the *realization-weighted expected return* from s,

$$v_s(\pi;\sigma) = \mathbb{P}_{\pi,\sigma}[s]\mathbb{E}_{H \sim \xi_s^{\pi,\sigma}}[G_H(\pi;\sigma)], \quad (1)$$

where $\xi_s^{\pi,\sigma}$ is defined arbitrarily if $\mathbb{P}_{\pi,\sigma}[s] = 0$.

4.2 Observable Sequential Rationality

Here we capitalize on the generality of our POHP definition. An agent belief can be used as a distribution over initial histories to define a POHP, which in this context we call a *sub-POHP*. Thus, every realizable information state s admits a sub-POHP where $\xi_s^{\pi,\sigma}$ is the probability distribution over the histories in $I(s)$ and the turn indicator is $\mathbb{1}\{s \in \mathcal{S}_{\mathcal{A}}\}$.

Sequential rationality can then be defined as optimal behavior within every sub-POHP with respect to an assignment of beliefs to unrealizable information states. This definition is equivalent to sequential rationality in a single-player EFG (Kreps and Wilson 1982). *Observable sequential rationality* (Morrill et al. 2021b) merely drops the requirement that play must be rational at unrealizable information states. The key value determining observable sequential rationality is in fact Eq. (1), the realization-weighted expected return.

As with rationality in normal and extensive-form games, we can generalize the idea of observable sequential rationality to samples from a joint distribution of agent strategy–daimon strategy pairs (traditionally called a *recommendation distribution*) and deviations. A *deviation* is a transformation that generates alternative agent behavior, *i.e.*, a function $\phi : \mathcal{X} \to \mathcal{X}$ where $\mathcal{X}$ is the set of *pure strategies* for the agent that play a single action deterministically in every active information state. We denote the complete set of such transformations, known as the set of *swap deviations*, as $\Phi_{\mathcal{X}}^{\mathrm{SW}}$. We can now give a generalized definition of observable sequential rationality in a POHP.

Definition 1. *A recommendation distribution, $\mu \in \Delta(\mathcal{X} \times D)$, where $\mathcal{X}$ and D are the sets of pure strategies for the agent and daimon, respectively, is observably sequentially rational for the agent with respect to a set of deviations, $\Phi \subseteq \Phi_{\mathcal{X}}^{\mathrm{SW}}$, if the maximum benefit for every deviation, $\phi \in \Phi$, according to the realization-weighted expected return from every information state, $s \in \mathcal{S}$, is non-positive,*

$$\mathbb{E}_{(x,d) \sim \mu}[v_s(\phi(x); d) - v_s(\phi_{\prec s}(x); d)] \leq 0,$$

where $\phi_{\prec s}$ is the deviation that applies ϕ only before s, i.e., $[\phi_{\prec s}x](\bar{s}) = [\phi x](\bar{s})$ if $\bar{s} \prec s$ and $x(\bar{s})$ otherwise.

If ϕ always deterministically plays to reach s, then this definition becomes equivalent to Morrill et al. (2021b)'s. The hindsight analogue to Definition 1 follows.

Definition 2. *Define the* full regret *from information state s as $\rho_s(\phi, \pi; \sigma) = v_s(\phi(\pi); \sigma) - v_s(\pi; \sigma)$. An agent is observably sequentially hindsight rational if they are a no-full-regret learner in every realizable information state within a given POHP with respect to $\Phi \subseteq \Phi_{\mathcal{X}}^{\mathrm{SW}}$. That is, the agent generates for any $T > 0$ a sequence of strategies, $(\pi^t)_{t=1}^{T}$, where $\lim_{T \to \infty} \frac{1}{T}\sum_{t=1}^{T} \rho_s(\phi, \pi^t; \sigma^t) \leq 0$ at each s for each $\phi \in \Phi$ under any sequence of daimon strategies $(\sigma^t)_{t=1}^{T}$.*

4.3 Local Learning

Consider a local learning problem in a repeated finite-horizon POHP with timed updates based on the realization-weighted expected return at each active information state s. Given a set of deviations, $\Phi \subseteq \Phi_{\mathcal{X}}^{\mathrm{SW}}$, we can construct a set of truncated deviations, $\Phi_{\prec s} = \{\phi_{\prec s}\}_{\phi \in \Phi}$, where each deviation in $\Phi_{\prec s}$ applies a deviation from Φ until after an action has been taken in s, at which point the rest of the strategy is left unmodified. Each truncated deviation represents a way

that the agent could play to and in s so a natural local learning problem is for the agent to choose their actions at s so that there is no beneficial truncated deviation.

To apply deviations to the agent's behavioral strategies, notice that sampling an action for each information state under timed updates yields a pure strategy. Thus, a behavioral strategy defines a probability distribution over the set of pure strategies, $\mathcal{X}$. We overload $\pi : \mathcal{X} \to \Delta(\mathcal{X})$ to return the probability of a given pure strategy under behavioral strategy $\pi \in \Pi$. From this perspective, π may be called a *mixed strategy*. The transformation of π by deviation ϕ is the pushforward measure $\phi(\pi)$ defined pointwise by $[\phi\pi](x') = \sum_{x \in \phi^{-1}(x')} \pi(x)$ for all $x' \in \mathcal{X}$, where $\phi^{-1} : x' \mapsto \{x \mid \phi(x) = x'\}$ is the pre-image of ϕ.

The *immediate regret* at information state s for not employing truncated deviation $\phi_{\preceq s}$ is a difference in realization-weighted expected return under $\xi_s^{\phi_{\prec s}(\pi),\sigma}$:

$$\rho_s(\phi_{\preceq s}, \pi; \sigma) = v_s(\phi_{\preceq s}(\pi); \sigma) - v_s(\phi_{\prec s}(\pi); \sigma)$$
$$= \mathbb{P}_{\phi_{\prec s}(\pi),\sigma}[s]\mathbb{E}[G_H(\phi_{\preceq s}(\pi); \sigma) - G_H(\pi; \sigma)].$$

Intuitively, it is the advantage that $\phi_{\preceq s}(\pi)$ has over π in s assuming that the agent plays to s according to $\phi_{\prec s}(\pi)$.[4] Sadly, it can be impossible to prevent information state s's immediate regret with respect to $\Phi_{\preceq s}$ from growing linearly in a repeated POHP.

Theorem 1. *An agent with timed updates cannot generally prevent immediate regret from growing linearly in a finite-horizon repeated POHP.*

Proof. Consider a two action, two information state POHP where information state s transitions to s' where the reward is $+1$ if the agent chooses the same action in both s and s', and -1 otherwise. The two *external* (constant) deviations, $\phi^{\to 1}$ and $\phi^{\to 2}$, that choose the same actions in both information states always achieve a value of $+1$. At s', the agent has to choose between achieving value with respect to the play of $\phi^{\to 1}$ or $\phi^{\to 2}$ in s. If the agent chooses action #1, then $v_s(\phi_{\prec s}^{\to 1}(\pi); \sigma) = +1$ but $v_s(\phi_{\prec s}^{\to 2}(\pi); \sigma) = -1$, and *vice-versa* if they choose action #2. Therefore, the agent minimizes their maximum regret by always playing uniform random and suffering a regret of $+1$ on every round. $\square$

What if we assume a stronger property, *perfect recall*? Perfect recall requires that every bit of information from every action and observation is encoded in the information state, *e.g.*, update functions that concatenate the previous information state with the given action or observation. Agents with perfect recall remember each of their actions and observations. This ensures that each information state s' is either the initial information state or has a single parent information state s, *i.e.*, $u(h_{<|h|}) = s$ for each history $h \in I(s')$.

A perfect-recall agent can only play to reach each history in a given information state, s, equally, *i.e.*, $\mathbb{P}_\pi[h] = \mathbb{P}_\pi[h']$ for all $h, h' \in I(s)$. If we define $\mathbb{P}_\pi[s] \propto \sum_{h \in I(s)} \mathbb{P}_\pi[h]$,

then perfect recall implies that $\mathbb{P}_\pi[s] = \mathbb{P}_\pi[h']$ for any history $h' \in I(s)$. The probability of realizing s simplifies to

$$\mathbb{P}_{\pi,\sigma}[s] = \sum_{h \in I(s)} \mathbb{P}_\pi[h]\mathbb{P}_\sigma[h] = \mathbb{P}_\pi[s] \sum_{h \in I(s)} \mathbb{P}_\sigma[h].$$

The belief about any history $h \in I(s)$ then simplifies to

$$\xi_s^{\pi,\sigma}(h) = \frac{\mathbb{P}_\pi[h]\mathbb{P}_\sigma[h]}{\mathbb{P}_\pi[s] \sum_{h \in I(s)} \mathbb{P}_\sigma[h]} = \frac{\mathbb{P}_\sigma[h]}{\sum_{h \in I(s)} \mathbb{P}_\sigma[h]}.$$

The realization-weighted expected return simplifies to

$$v_s(\pi; \sigma) = \mathbb{P}_\pi[s] \underbrace{\sum_{h \in I(s)} \mathbb{P}_\sigma[h]\mathbb{E}[G_h(\pi; \sigma)]}_{v_s^{\mathrm{CF}}(\pi;\sigma)}.$$

The sum denoted $v_s^{\mathrm{CF}}(\pi; \sigma)$ is recognizable as the *counterfactual value* (Zinkevich et al. 2007) of s, which does not depend on π's play at s's predecessors. Immediate regret becomes weighted immediate *counterfactual regret*,

$$\rho_s(\phi_{\preceq s}, \pi; \sigma) = \mathbb{P}_{\phi_{\prec s}(\pi)}[s](v_s^{\mathrm{CF}}(\phi_{\preceq s}(\pi); \sigma) - v_s^{\mathrm{CF}}(\pi; \sigma)).$$

Since the counterfactual value function does not depend on π's play at s's predecessors, perfect recall avoids the difficulty that leads to Theorem 1 and allows a reduction from minimizing immediate regret to minimizing *time selection regret* in the *prediction with expert advice* setting (Blum and Mansour 2007). In a repeated POHP where the agent and daimon choose $\pi^t \in \Pi$ and $\sigma^t \in \Sigma$ on each round t the (time dependent) counterfactual value function $t \mapsto v_s^{\mathrm{CF}}(\cdot; \sigma^t)$ fills the role of the prediction-with-expert-advice reward function and the (time dependent) reach probability function $w_{s,\phi} : t \mapsto \mathbb{P}_{\phi_{\prec s}(\pi^t)}[s]$ fills the role of a time selection function. The growth of cumulative immediate regret can therefore be controlled, in principle, to a sublinear rate by deploying Blum and Mansour (2007)'s algorithm or time selection regret matching (Morrill et al. 2021a).

4.4 General Immediate Regret Minimization

The algorithm that applies a no-time-selection-regret algorithm to minimize immediate regret in every active information state generalizes the *extensive-form regret minimization* (*EFR*) algorithm (Morrill et al. 2021a) in that $\Phi \subseteq \Phi_{\mathcal{X}}^{\mathrm{SW}}$ may be any set of deviations rather than a set of behavioral deviations. Just as Morrill et al. (2021a) shows that EFR is hindsight rational in EFGs, we can prove that general immediate regret minimization achieves the same in POHPs, though now we can easily present this result as a consequence of observable sequential hindsight rationality. In principle, our generalized algorithm could compete with the set of swap deviations when given this set (or the set of internal deviations) as a parameter argument, however, circular dependencies between immediate strategies at different information states prevents our algorithm from being efficiently implemented with such a deviation set.

Observable sequential hindsight rationality depends on full regret so we relate immediate regret to full regret with two lemmas and conclude with the algorithm's regret bound.

---

[4] The term "advantage" here is chosen deliberately as immediate regret is analogous to advantage in MDPs (Baird 1994), with respect to a given rather than optimal policy (see, *e.g.*, Kakade (2003)).

Lemma 1. *In a finite-horizon POHP, the realization-weighted expected return of active information state s under perfect recall recursively decomposes as*

$$v_s(\pi; \sigma) = \mathbb{P}_\pi[s] r_s(\pi; \sigma) + \sum_{s' \in \bigcup_{a \in \mathcal{A}(s)} \mathcal{S}_\mathcal{A}(s,a)} v_{s'}(\pi; \sigma)$$

where $r_s(\pi; \sigma) = \sum_{h \in I(s)} \mathbb{P}_\sigma[h] \mathbb{E}[r(\omega(hAB))]$, and expectations are taken over $A \sim \pi(s)$ and $B \sim \sigma(hA)$.

Proof. The counterfactual value decomposes as

$$v_s^{\mathrm{CF}}(\pi; \sigma) = \sum_{h \in I(s)} \mathbb{P}_\sigma[h] \mathbb{E}[r(\omega(hAB)) + \Gamma G_{hAB}(\pi; \sigma)]$$
$$= r_s(\pi; \sigma) + \mathbb{E}\Big[\sum_{h \in I(s), b \in \mathcal{A}(hA)} \mathbb{P}_\sigma[hAb] G_{hAb}(\pi; \sigma)\Big]$$
$$= r_s(\pi; \sigma) + \mathbb{E}\Big[v_{u_\mathcal{A}(s,A)}^{\mathrm{CF}}(\pi; \sigma)\Big],$$

where $\Gamma \sim \gamma(h)$. Multiplying by the reach weight,

$$v_s(\pi; \sigma)$$
$$= \mathbb{P}_\pi[s] r_s(\pi; \sigma) + \sum_{a \in \mathcal{A}(s)} \mathbb{P}_\pi[s] \pi(a \mid s) v_{u_\mathcal{A}(s,a)}^{\mathrm{CF}}(\pi; \sigma)$$
$$= \mathbb{P}_\pi[s] r_s(\pi; \sigma) + \sum_{a \in \mathcal{A}(s)} v_{u_\mathcal{A}(s,a)}(\pi; \sigma). \qquad (2)$$

Furthermore,

$$v_{u_\mathcal{A}(s,a)}(\pi; \sigma)$$
$$= \sum_{s' \in \mathcal{S}_\mathcal{A}(s,a)} \sum_{h \in u_\mathcal{A}(s,a)} \mathbb{1}\{u(h) = s'\} \mathbb{P}_{\pi,\sigma}[h] \mathbb{E}[G_h(\pi; \sigma)]$$
$$= \sum_{s' \in \mathcal{S}_\mathcal{A}(s,a)} \underbrace{\sum_{h' \in I(s')} \mathbb{P}_{\pi,\sigma}[h'] \mathbb{E}[G_{h'}(\pi; \sigma)]}_{v_{s'}(\pi; \sigma)}. \qquad (3)$$

Substituting Eq. (3) into Eq. (2) completes the proof. $\qquad \square$

Lemma 2. *In a finite-horizon POHP, the full regret with respect to $\phi \in \Phi_\mathcal{X}^{\mathrm{SW}}$ under perfect recall at active information state s recursively decomposes as*

$$\rho_s(\phi, \pi; \sigma) = \rho_s(\phi_{\preceq s}, \pi; \sigma) + \sum_{s' \in \bigcup_{a \in \mathcal{A}(s)} \mathcal{S}_\mathcal{A}(s,a)} \rho_{s'}(\phi, \pi; \sigma).$$

Proof.

$$\rho_s(\phi, \pi; \sigma)$$
$$= v_s(\phi(\pi); \sigma) \overbrace{- v_s(\phi_{\preceq s}(\pi); \sigma) + v_s(\phi_{\preceq s}(\pi); \sigma)}^{0} - v_s(\pi; \sigma)$$

where $\underbrace{}_{\rho_s(\phi_{\preceq s}, \pi; \sigma)}$

$$= \rho_s(\phi_{\preceq s}, \pi; \sigma)$$
$$+ \underbrace{\mathbb{P}_{\phi(\pi)}[s] r_s(\pi; \sigma) - \mathbb{P}_{\phi_{\preceq s}(\pi)}[s] r_s(\pi; \sigma)}_{0}$$
$$+ \sum_{a \in \mathcal{A}(s)} \underbrace{v_{u_\mathcal{A}(s,a)}(\phi(\pi); \sigma) - v_{u_\mathcal{A}(s,a)}(\phi_{\preceq s}(\pi); \sigma)}_{\rho_{u_\mathcal{A}(s,a)}(\phi, \pi; \sigma)}.$$

Applying Eq. (3) to sum over active information states,

$$= \rho_s(\phi_{\preceq s}, \pi; \sigma) + \sum_{s' \in \bigcup_{a \in \mathcal{A}(s)} \mathcal{S}_\mathcal{A}(s,a)} \underbrace{v_{s'}(\phi(\pi); \sigma) - v_{s'}(\pi; \sigma)}_{\rho_{s'}(\phi, \pi; \sigma)}. \quad \square$$

Theorem 2. *If a perfect recall agent's cumulative immediate regret with respect to $\Phi \subseteq \Phi_\mathcal{X}^{\mathrm{SW}}$ at each information state s in a repeated finite-horizon POHP is upper bounded by $f(T) \geq 0$, $f(T) \in \mathrm{o}(T)$ after T rounds, then the agent's cumulative full regret at each s is sublinear, upper bounded according to $\sum_{t=1}^{T} \rho_s(\phi, \pi^t; \sigma^t) \leq |\mathcal{S}_{s,\mathcal{A}}| f(T)$, where $\mathcal{S}_{s,\mathcal{A}} = \{s' \in \mathcal{S}_\mathcal{A} \mid s \preceq s'\}$ is the active information states in the sub-POHP rooted at s. Such an agent is therefore observably sequentially hindsight rational with respect to Φ.*

Proof. Working from each terminal information state where the full and immediate regret are equal toward s at the root of any given sub-POHP, we recursively bound the cumulative full regret at every information state according to Lemma 2. Every active information state adds at most $f(T)$ to the cumulative full regret at s and there are $|\mathcal{S}_{s,\mathcal{A}}|$ active information states in s's sub-POHP so the cumulative full regret at s is no more than $|\mathcal{S}_{s,\mathcal{A}}| f(T)$. $\qquad \square$

5 Conclusion

The POHP formalism may be useful in modeling continual learning problems where environments are expansive, unpredictable, and dynamic. Good performance here demands that the agent continually learns, adapts, and re-evaluates their assumptions. We suspect that hindsight rationality could serve as the learning objective for such problems if it could be formulated for a single agent lifetime rather than over a repeated POHP.

Our analysis of general immediate regret minimization for POHPs and the impossibility result of Theorem 1 brings up questions about how far this procedure can be generalized. Regret decomposition is based on a perfect-recall, realization-weighted variant of Kakade (2003)'s performance difference lemma (Lemma 5.2.1). Mohammedalamen et al. (2021) use this to show how the counterfactual regret minimization (CFR) (Zinkevich et al. 2007) (EFR with counterfactual deviations (Morrill et al. 2021a)) can be applied to continuing, discounted MDPs with reward uncertainty. The POHP formalism can perhaps allow us to better understand when immediate and full regret can be minimized without perfect recall by considering Lanctot et al. (2012)'s well-formed-game conditions together with Mohammedalamen et al. (2021)'s analysis.

The POHP formalism allows agents to determine their own representation of the environment. This opens the way to direct discussions and comparisons of representations and updating schemes. One particular direction that is made natural by the POHP model's action–observation interface is predictive state representations (PSRs) (Singh et al. 2003; Singh, James, and Rudary 2012). While PSRs were developed to model Markovian dynamical systems with at most one controller, the POHP model could facilitate an extension to multi-agent settings.

Acknowledgments

References

Baird, L. C. 1994. Reinforcement learning in continuous time: Advantage updating. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, volume 4, 2448–2453. IEEE.

Blum, A.; and Mansour, Y. 2007. From external to internal regret. *Journal of Machine Learning Research*, 8(Jun): 1307–1324.

Breitmoser, Y.; Tan, J. H.; and Zizzo, D. J. 2010. On the beliefs off the path: Equilibrium refinement due to quantal response and level-k. *Nottingham University Business School Research Paper*, (2010-07).

Dekel, E.; and Siniscalchi, M. M. 2015. Epistemic Game Theory. *Handbook of Game Theory With Economic Applications*, 4: 619–702.

Farina, G.; Kroer, C.; and Sandholm, T. 2019. Online convex optimization for sequential decision processes and extensive-form games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 1917–1925.

Greenwald, A.; Li, J.; and Sodomka, E. 2017. *Solving for Best Responses and Equilibria in Extensive-Form Games with Reinforcement Learning Methods*, 185–226. Cham: Springer International Publishing. ISBN 978-3-319-47843-2.

Greenwald, A.; Li, Z.; and Marks, C. 2006. Bounds for Regret-Matching Algorithms. In *ISAIM*.

Hansen, E. A.; Bernstein, D. S.; and Zilberstein, S. 2004. Dynamic programming for partially observable stochastic games. In *AAAI*, volume 4, 709–715.

Kakade, S. M. 2003. *On the sample complexity of reinforcement learning*. Ph.D. thesis, UCL (University College London).

Kovařík, V.; Schmid, M.; Burch, N.; Bowling, M.; and Lisỳ, V. 2019. Rethinking formal models of partially observable multiagent decision making. *arXiv preprint arXiv:1906.11110*.

Kreps, D. M.; and Wilson, R. 1982. Sequential equilibria. *Econometrica*, 50(4): 863–894.

Kuhn, H. W. 1953. Extensive Games and the Problem of Information. *Contributions to the Theory of Games*, 2: 193–216.

Lanctot, M.; Burch, N.; Zinkevich, M.; Bowling, M.; and Gibson, R. G. 2012. No-Regret Learning in Extensive-Form Games with Imperfect Recall. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 65–72.

Lattimore, T.; and Szepesvári, C. 2020. *Bandit algorithms*. Cambridge University Press.

Mohammedalamen, M.; Morrill, D.; Sieusahai, A.; Satsangi, Y.; and Bowling, M. 2021. Learning to Be Cautious.

Morrill, D.; D'Orazio, R.; Lanctot, M.; Wright, J. R.; Bowling, M.; and Greenwald, A. R. 2021a. Efficient Deviation Types and Learning for Hindsight Rationality in Extensive-Form Games. In *Thirty-Eighth International Conference on Machine Learning, July 18-24, 2021, virtual*.

Morrill, D.; D'Orazio, R.; Sarfati, R.; Lanctot, M.; Wright, J. R.; Greenwald, A. R.; and Bowling, M. 2021b. Hindsight and Sequential Rationality of Correlated Play. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, February 2-9, 2021, virtual*.

Shapley, L. S. 1953. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100.

Singh, S.; James, M.; and Rudary, M. 2012. Predictive state representations: A new theory for modeling dynamical systems. *arXiv preprint arXiv:1207.4167*.

Singh, S. P.; Littman, M. L.; Jong, N. K.; Pardoe, D.; and Stone, P. 2003. Learning predictive state representations. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, 712–719.

Smallwood, R. D.; and Sondik, E. J. 1973. The optimal control of partially observable Markov processes over a finite horizon. *Operations research*, 21(5): 1071–1088.

Srinivasan, S.; Lanctot, M.; Zambaldi, V.; Pérolat, J.; Tuyls, K.; Munos, R.; and Bowling, M. 2018. Actor-Critic Policy Optimization in Partially Observable Multiagent Environments. In *Advances in Neural Information Processing Systems*.

von Neumann, J.; and Morgenstern, O. 1947. *The Theory of Games and Economic Behavior*. Princeton University Press, 2nd edition.

Zinkevich, M.; Johanson, M.; Bowling, M. H.; and Piccione, C. 2007. Regret Minimization in Games with Incomplete Information. In *Advances in Neural Information Processing Systems*, 1729–1736.