Learning in two-player zero-sum partially observable Markov games with perfect recall

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the problem of learning a Nash equilibrium (NE) in an extensive game with imperfect information (EGII) through self-play. Precisely, we focus on two-player, zero-sum, episodic, tabular EGII under the *perfect-recall* assumption where the only feedback is realizations of the game (bandit feedback). In particular the *dynamics of the EGII is not known*—we can only access it by sampling or interacting with a game simulator. For this learning setting, we provide the Implicit Exploration Online Mirror Descent (IXOMD) algorithm. It is a model-free algorithm with a high-probability bound on convergence rate to the NE of order $1/\sqrt{T}$ where T is the number of played games. Moreover IXOMD is computationally efficient as it needs to perform the updates only along the sampled trajectory.

1 Introduction

We study the problem of *learning* a Nash equilibrium (NE, Nash Jr, 1950) in an *extensive game with imperfect information* (EGII, Osborne and Rubinstein, 1994). Precisely, we focus on two-player zero-sum EGII under the *perfect-recall* assumption (Kuhn, 1953). Perfect recall means that the players *do not forget* observations encountered or actions taken during the game. We model the game as a tabular, episodic (of horizon H) *partially observable Markov game* (POMG) with a state space of size S and action spaces of size A and B for the max- and min-player respectively. And where the observation spaces (or information set spaces) of size X for the max-player and Y for the min-player are partitions of the state space. In learning by *self play*, we control *both* the max *and* min-players. After T episodes of the game we are asked to return a profile that is close to a NE in terms of *exploitability* gap (Ponsen et al., 2011).

Full feedback In case when we have perfect knowledge of the game (i.e., the transition probabilities and rewards) there already exist several methods approximating the NE. The first line of work casts the problem through the sequence-form representation as a linear program which can be efficiently solved (Romanovsky, 1962; von Stengel, 1996; Koller et al., 1996) for games with moderate sizes of observation spaces X and Y. The sequence-from representation allows also to cast the problem as *finding a saddle point* (Hoda et al., 2010). It is then possible to adapt first-order methods such as Nesterov's smoothing (Nesterov, 2005) and `MirrorProx` (Nemirovski, 2004) to EGII, as done respectively by Hoda et al. (2010); Kroer et al. (2018) and Kroer et al. (2015, 2020). These methods have a rate of convergence of order $\tilde{\mathcal{O}}((X+Y)/T)$, where $\tilde{\mathcal{O}}$ hides poly-log terms in e^H, X, A, Y, B, T.[1] Note that *game-dependent* exponential rate could also be obtained with first-order methods, see Gilpin et al. (2012); Munos et al. (2020). Another important line of work relies on minimizing the *counterfactual regret* (Zinkevich et al., 2007). That is, they use an algorithm for adversarial bandits to locally minimize the regret of each player. A well-known example is the `CFR`

[1] Therefore, we hide *polynomial* dependence on the horizon H.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

algorithm by Zinkevich et al. (2007) based on the regret-matching algorithm (Hart and Mas-Colell, 2000; Gordon, 2007). There exist many other variants of it, such as CFR+ algorithm (Tammelin, 2014; Burch et al., 2019), see also Farina et al. (2019, 2021a). These algorithms however only enjoy a (known) guarantee of convergence of order $\widetilde{\mathcal{O}}((X\sqrt{A}+Y\sqrt{B})/\sqrt{T})$. Note that the two last approaches require to compute a full feedback: either some gradient for the first-order methods or the local regret for counterfactual regret minimization. Usually, this can be done by a complete traversal of the state space leading to a time-complexity of order $\mathcal{O}(S)$. It is possible to alleviate this time complexity to $\mathcal{O}(X+Y)$[2] by sampling (i.e., by sampling the transitions and the actions of the other player), see for example the external-sampling MCCFR algorithm (Lanctot et al., 2009; Farina et al., 2020).

Bandit feedback In this paper, we consider a more challenging setting where the agent *only observes realizations of the games (bandit feedback) and without any prior knowledge of the game*. Precisely, the rewards, the transition probabilities (sometimes modeled as the policy of a *chance player*), the observation/state space and its (tree) structure are unknown.

Bandit feedback, model-based To deal with this limited feedback, Zhou et al. (2020) consider model-based approach by using *posterior sampling* (PS, Strens, 2000) to learn a model and then use the CFR algorithm in games sampled from the posterior. They obtain a convergence rate of order $\widetilde{\mathcal{O}}(\max(XA+YB,\sqrt{S})/\sqrt{T})$ but only when the games are actually sampled according to the known prior. In addition, they still need to know the state space and its structure[3] in order to instantiate the prior. Instead, Zhang and Sandholm (2021) rely on the principle of optimism in presence of uncertainty to incrementally build a model of the game. Then, they feed optimistic local regrets to a counterfactual regret minimizer algorithm such as the CFR algorithm. They prove a high-probability bound on the exploitability gap of order $\widetilde{\mathcal{O}}((X\sqrt{A}+Y\sqrt{B})/\sqrt{T})$.

Bandit feedback, model-free This paper follows another line of work that consider *model-free* approach. A well known algorithm of this type is outcome-sampling MCCFR (Lanctot et al., 2009; Farina et al., 2020), which builds an importance-sampling estimate of the counterfactual regret given *exploration profile* (named balanced strategy by Farina et al., 2020). This exploration strategy should ensure that the players explore the information sets uniformly (i.e., such that all induced reach probabilities are lower-bounded by an absolute constant). Note that it is not clear how to find such an exploration profile without knowing the structure of the game.[3] In particular, following the uniform distribution over the actions at each information set is not necessarily a good choice, e.g., when the tree formed by the information set space is not balanced. This algorithm has a guarantee of order $\widetilde{\mathcal{O}}((X\sqrt{A}+Y\sqrt{B})/\sqrt{T})$ with high probability. Building on this idea, Farina and Sandholm (2021) propose to mix the exploration profile with one produced by a counterfactual regret minimizer such that CFR. They prove a high-probability bound on the exploitability gap of order $\widetilde{\mathcal{O}}(\mathrm{poly}(X,A,Y,B)/T^{1/4})$. Note that this bound is a consequence of a bound of the regret of both players (see Section 2) that holds even in the non-stochastic setting where an adversary picks a new game at each episode. Closer to our approach, Farina et al. (2021b) recast the problem as an adversarial bandit linear optimization problem (Flaxman et al., 2005; Abernethy et al., 2008, see also Section 3.1). Precisely they use the online mirror descent (OMD) algorithm with the dilated entropy distance-generating function (Hoda et al., 2010; Kroer et al., 2015) as regularizer. Then, the OMD algorithm is fed with an estimate of the losses of the reformulated adversarial bandit linear problem. The estimator is a generalization of the typical one-point linear regression (Dani et al., 2008). They obtain a rate of order $\widetilde{\mathcal{O}}((XA+YB)/\sqrt{T})$ which is similarly as the one of Farina and Sandholm (2021) also derived from a regret bound valid in the adversarial setting. However, this bound holds only in expectation and not in high probability.

In order to obtain high-probability bounds we propose to use instead an importance sampling estimator of the losses with *implicit exploration* (Kocák et al., 2014; Neu, 2015). Indeed the implicit bias of this estimator allows to control effortlessly the variance of the estimate, see Lattimore and Szepesvári (2020, Chapter 12) for an in-depth discussion. Thus we propose the Implicit Exploration Online Mirror Descent (IXOMD) algorithm based on the OMD with the dilated entropy distance-generating function (using uniform weights) as a regularizer and add implicit exploration in the importance sampling

[2] Note that $\mathcal{O}(X+Y)$ is at most $\mathcal{O}(S)$.

[3] By structure we refer to the tree structure of the state space or observations spaces, see Section 2.

estimator of the losses. Using our new analysis of this particular combination, we prove a high-probability bound on the exploitability gap of the average profile of order $\widetilde{\mathcal{O}}((X\sqrt{A}+Y\sqrt{B})/\sqrt{T})$; cf. Table 1 to see how our result compares to the prior work mentioned above. Precisely, our bound is obtained by bounding the regret of each player if they both follow the policy prescribed by IXOMD. Note that the regret bound, e.g., of the max-player, of order $\widetilde{\mathcal{O}}(X\sqrt{AT})$, remains valid if the policy *and* the game are picked by an adversary at each episode. IXOMD shares some similarities with the one by Jin et al. (2020) designed for a different setting (see Remark 1). A notable difference is that we use the dilated entropy distance-generating function as a regularizer instead of the un-normalized Kullback-Leibler divergence (Rosenberg and Mansour, 2019). This choice of regularizer allows an efficient update of the current policy in $\mathcal{O}(H)$; see Farina et al. (2021b) and Section 3.3. Our result answers the open problem raised by Farina et al. (2021b) and Farina and Sandholm (2021) of providing an algorithm with high-probability regret bound scaling with $\sqrt{T}$ with a time-complexity of order $\mathcal{O}(HA)$ per episode. Moreover, IXOMD requires almost no prior knowledge of the game. In particular, we do not need to know the list of information sets in advance. We only require an oracle providing the possible actions at encountered information sets and a bound on A, B, and H to tune optimally the learning rate, see Remark 3.

We highlight our main contributions:

- The IXOMD algorithm to learn a NE of an EGII in self-play with limited feedback. It has a provably high-probability convergence rate of order $\widetilde{\mathcal{O}}((X\sqrt{A}+Y\sqrt{B})/\sqrt{T})$.
- If only one player follows IXOMD, e.g., the max-player, then its regret is at most w.h.p. $\widetilde{\mathcal{O}}(X\sqrt{AT})$. The important property of our result is that it remains valid even if the policy *and* the game are picked by an adversary at each episode. Furthermore, the time complexity of IXOMD per episode is of order $\mathcal{O}(HA)$. This answers an open problem by Farina et al. (2021b); Farina and Sandholm (2021).
- IXOMD only needs to know the possible actions at the encountered information sets and a bound on A, B, and H to tune the learning rate optimally. In particular, we do not need to know the list of information sets in advance.

Algorithm		**Adv. game**	**Rate**
Zhou et al. (2020)	model-based	no	$\widetilde{\mathcal{O}}(\max(X\sqrt{A}+Y\sqrt{B},\sqrt{S})/\sqrt{T})$ [1]
Zhang and Sandholm (2021)			$\widetilde{\mathcal{O}}((X\sqrt{A}+Y\sqrt{B})/\sqrt{T})$
Lanctot et al. (2009); Farina et al. (2020)	model-free		$\widetilde{\mathcal{O}}((X\sqrt{A}+Y\sqrt{B})/\sqrt{T})$
Farina and Sandholm (2021)		yes	$\widetilde{\mathcal{O}}(\mathrm{poly}(X,A,Y,B)/T^{1/4})$
Farina et al. (2021b)			$\widetilde{\mathcal{O}}((XA+YB)/\sqrt{T})$ [2]
IXOMD (this paper)			$\widetilde{\mathcal{O}}((X\sqrt{A}+Y\sqrt{B})/\sqrt{T})$

Table 1: Algorithms for computing a NE of an EGII with bandit feedback and their respective upper bound on the exploitability gap after T episode. With the adversarial game column we precise whether the algorithm could be used to obtain a $\sqrt{T}$-regret for one player when the other player and the game are chosen by an adversary at each episodes.

[1] Only in expectation according to a known prior on the game.
[2] Only in expectation.

2 Preliminaries

In this section, we introduce our notations and our setting—partially observable Markov game (POMG) with bandit feedback and perfect recall. For a positive integer i, we denote by $[i]$ the set $\{1,2,\dots,i\}$. For a finite set $\mathcal{A}$, we let $\Delta_{\mathcal{A}}$ (or $\Delta(\mathcal{A})$) denote the set of all probability distributions over $\mathcal{A}$.

Partially observable Markov game (POMG) In this paper we consider an episodic, tabular, two-player, zero-sum POMG $(\mathcal{S},\mathcal{X},\mathcal{Y},\mathcal{A},\mathcal{B},H,\{p_h\}_{h\in[H]},\{r_h\}_{h\in[H]})$, which consists of the following

components (Littman, 1994; Shapley, 1953): a finite state space $\mathcal{S}$ of size S, its information set spaces, i.e., partitions of $\mathcal{S}$, $\mathcal{X}$ (for the max-player) of size X and $\mathcal{Y}$ (for the min-player) of size Y, finite action spaces $\mathcal{A}$ (for the max-player) of size A and $\mathcal{B}$ (for the min-player) of size B, time-horizon $H \in \mathbb{N}$, initial state distribution $p_0 \in \Delta(\mathcal{S})$, a state-transition probability kernel $p_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to \Delta(\mathcal{S})$ for each $h \in [H]$, and a reward function $r_h : \mathcal{S} \times \mathcal{A} \times \mathcal{B} \to [0, 1]$ for each $h \in [H]$. For a state $s \in \mathcal{S}$ we denote by $x(s) \in \mathcal{X}$ and $y(s) \in \mathcal{Y}$ information sets such that $s \in x(s)$ and $y \in y(s)$.

Learning procedure The players play this game for T episodes, following so-called policies. A policy μ of the max-player is a sequence $(\mu_h)_{h\in[H]}$ of mappings from $\mathcal{X}_h$ to $\Delta_{\mathcal{A}}$. ($\mathcal{X}_h \subset \mathcal{X}$ is define later.) A policy ν of the min-player is similarly defined. We let $\Pi_{\max}$ and $\Pi_{\min}$ denote the sets of max- and min-players' policies, respectively. The t-th episode of the game proceeds as follows: an initial state s_1^t is sampled from p_0. At the step h, the max- and min-players (resp.) observe their information sets $x_h^t := x(s_h^t)$ and $y_h^t := y(s_h^t)$. Given the information, the max- and min-players (resp.) choose and execute actions $a_h^t \sim \mu_h^t(\cdot|x_h)$ and $b_h^t \sim \nu_h^t(\cdot|y_h)$. As a result, the current state transitions to a next state $s_{h+1}^t \sim p_h(\cdot|s_h^t, a_h^t, b_h^t)$, and the max- and min-players receive rewards $r_h^t := r_h(s_h^t, a_h^t, b_h^t)$ and $-r_h^t$, respectively. This is repeated until a time step H, at which the episode finishes.

Tree-like game structure and perfect recall assumption We assume that the game has a tree-like structure: for any state $s \in \mathcal{S}$ there is a unique step h and history $(s_1, a_1, b_1, \ldots, s_h = s)$ to reach s. Precisely, for any policy of the players, for any realization of the game (i.e., trajectory) $(s'_k, a'_k, b'_k)_{k\in[H]}$, conditionally to $s'_i = s$ it almost surely holds that $i = h$ and $(s'_1, \ldots, s'_h) = (s_1, \ldots, s_h)$. We also assume perfect recall, which means that each player remembers their past observations and actions. For example in case of the max-player, for each information set $x \in \mathcal{X}$ there is a unique history $(x_1, a_1, \ldots, x_h = x)$ up to x. These assumptions require that $\mathcal{X}$ can be partitioned to H subsets $(\mathcal{X}_h)_{h\in[H]}$ such that $x_h \in \mathcal{X}_h$ is reachable only at time step h (otherwise there would be two different histories up to x_h). $\mathcal{S}$ and $\mathcal{Y}$ can be also partitioned into H subsets $(\mathcal{S}_h)_{h\in[H]}$, and $(\mathcal{Y}_h)_{h\in[H]}$, respectively.

Given the assumptions above, there exists a unique history $(s_1, a_1, b_1, \ldots, s_h = s, a_h = a, b_h = b)$ ending with $(s_h = s, a_h = a, b_h = b)$ for any state $s \in \mathcal{S}_h$, the max-player's action $a \in \mathcal{A}$, and the min-player's action $b \in \mathcal{B}$. Accordingly the probability of $s_h = s, a_h = a, b_h = b$ can be computed by $p_h^{\mu,\nu}(s, a, b) = p_{1:h}(s)\mu_{1:h}(s, a)\nu_{1:h}(s, b)$, where

$$
\begin{aligned}
p_{1:h}(s) &:= p_0(s_1) \textstyle\prod_{h'=1}^{h-1} p_{h'}(s_{h'+1}|s_{h'}, a_{h'}, b_{h'}),\\
\mu_{1:h}(s, a) &:= \mu_{1:h}(x(s), a) := \textstyle\prod_{h'=1}^{h} \mu_{h'}(a_{h'}|x(s_{h'})),\\
\nu_{1:h}(s, b) &:= \nu_{1:h}(y(s), b) := \textstyle\prod_{h'=1}^{h} \nu_{h'}(b_{h'}|y(s_{h'})).
\end{aligned}
$$

With abuse of notation we let $\mu_{1:h-1}(s) := \mu_{1:h-1}(x(s)) := \mu_{1:h-1}(s_{h-1}, a_{h-1})$, $p_h^{\mu,\nu}(s) := p_{1:h}(s)\mu_{1:h-1}(s, a)\nu_{1:h-1}(s, b)$ and $p_h^{\mu,\nu}(x) := \sum_{s\in x(s)} p_h^{\mu,\nu}(s)$ for any information set $x \in \mathcal{X}_h$. We use $\nu_{1:h-1}$ similarly.

Bandit feedback We assume that the value of $r_h(s, a, b)$ is revealed to the players only when actions $a \in \mathcal{A}$ and $b \in \mathcal{B}$ are taken in a state $s \in \mathcal{S}$ at time step h. (That being said, the players are not aware of the underlying state s.) Furthermore we assume that the players know neither the state transition dynamics nor even the set of states $\mathcal{S}$. These assumptions impose a significant difficulty since the players need to carefully play the game trying different actions to obtain the information of the game.

Remark 1. *Jin et al. (2020) consider a similar setting (from the view point of the max-player) of learning adversarial MDPs with bandit feedback wherein the reward function is chosen by an adversary. Our setting is different in that the players have only* imperfect information*, and that the* state transition dynamics is changing due to the learning opponents. *Nonetheless the tree structure and perfect recall assumptions allow a simple and efficient model-free algorithm that we provide.*

Regret and Nash Equilibrium (NE) For policies μ and ν we define the expected return (of the max-player) by $V^{\mu,\nu} := \mathbb{E}^{\mu,\nu}[\sum_{h=1}^{H} r_h(s_h, a_h, b_h)]$, where $\mathbb{E}^{\mu,\nu}$ means that actions are selected as described above. For sequences of policies $(\mu^t)_{t\in[T]} \in \Pi_{\max}^T$ and $(\nu^t)_{t\in[T]} \in \Pi_{\min}^T$, the regret of the

Algorithm 1: IXOMD for the Max-Player

Input: IX hyper-parameter $\gamma \in (0, \infty)$ and OMD's learning rate $\eta \in (0, \infty)$.
Output: A near-NE policy for the max-player.

Initialize $\mu_h^1(a_h|x_h) \leftarrow 1/A$ for each $(x_h, a_h, h) \in \mathcal{X}_h \times \mathcal{A} \times [H]$.
for $t = 1, \ldots, T$ **do**
  for $h = 1, \ldots, H$ **do**
    Observe x_h^t, execute $a_h^t \sim \mu_h^t(\cdot|x_h^t)$, and receive r_h^t.
  end
  Set $Z_{H+1}^t \leftarrow 1$.
  for $h = H, \ldots, 1$ **do**
    Construct the IX loss estimate $\widetilde{\ell}_h^t$ by

$$\widetilde{\ell}_h^t \leftarrow \frac{1 - r_h^t}{\mu_{1:h}^t(x_h^t, a_h^t) + \gamma}.$$

    Compute for each $h \in [H]$ with $Z_{H+1}^t \leftarrow 1$

$$Z_h^t \leftarrow 1 - \mu_h^t(a_h^t|x_h^t) + \mu_h^t(a_h^t|x_h^t) \exp\Big(-\eta \widetilde{\ell}_h^t + \log Z_{h+1}^t\Big).$$

    Update μ^t to μ^{t+1} at x_h^t by

$$\mu_h^{t+1}(a_h|x_h^t) \leftarrow \begin{cases} \mu_h^t(a_h|x_h^t) \exp\Big(-\eta \widetilde{\ell}_h^t + \log Z_{h+1}^t - \log Z_h^t\Big) & \text{if } a_h = a_h^t \\ \mu_h^t(a_h|x_h^t) \exp(-\log Z_h^t) & \text{otherwise} \end{cases}$$

    and $\mu^{t+1}(\cdot|x_h) \leftarrow \mu^t(\cdot|x_h)$ at other information sets $x_h \in \mathcal{X}_h$.
  end
end
return *Policy $\overline{\mu}$ which is the average of $\mu_1, \ldots, \mu_T$ defined in Theorem 1.*

max-player relative to some policy $\mu^\dagger \in \Pi_{\max}$ is defined as

$$\mathfrak{R}_{\max}^T(\mu^\dagger) := \sum_{t=1}^T \Big(V^{\mu^\dagger, \nu^t} - V^{\mu^t, \nu^t}\Big). \tag{1}$$

Similarly $\sum_{t=1}^T (V^{\mu^t, \nu^t} - V^{\mu^t, \nu^\dagger})$ is the min-player's regret relative to some $\nu^\dagger \in \Pi_{\min}$.

Our aim is to compute a NE. The following well-known folklore theorem[4] states that this problem can be converted to the regret minimization problem.

Theorem 1. *For each $(x_h, a_h, h) \in \mathcal{X}_h \times \mathcal{A} \times [H]$, define the average profile $(\overline{\mu}, \overline{\nu})$ by*

$$\overline{\mu}_h(a_h|x_h) := \frac{\sum_{t=1}^T \mu_{1:h}^t(x_h, a_h)}{\sum_{t=1}^T \mu_{1:h-1}^t(x_h)} \quad \textit{and} \quad \overline{\nu}_h(b_h|y_h) := \frac{\sum_{t=1}^T \nu_{1:h}^t(y_h, b_h)}{\sum_{t=1}^T \nu_{1:h-1}^t(y_h)},$$

if the sum of the denominator is non-zero, otherwise as the uniform distribution over actions. If for some non-negative real value ε, we have that $(\mathfrak{R}_{\max}^T(\mu^\dagger) + \mathfrak{R}_{\min}^T(\nu^\dagger))/T \le \varepsilon$ for any profile $(\mu^\dagger, \nu^\dagger)$, then $(\overline{\mu}, \overline{\nu})$ are ε-NE, i.e., $\max_{\mu \in \Pi_{\max}} V^{\mu, \overline{\nu}} - \min_{\nu \in \Pi_{\min}} V^{\overline{\mu}, \nu} \le \varepsilon$.

Given Theorem 1, we consider how to minimize the regret for the max- and min-players; or how to control the regret such that it grows sublinearly. The subsequent section presents an algorithm, which we call implicit exploration online mirror descent (IXOMD), that accomplishes this goal.

3 Implicit Exploration Online Mirror Descent (IXOMD)

Due to the symmetry of the players, it suffices to consider only the learning of the max-player. Therefore we mainly focus on it and denote the max-player's regret (1) by $\mathfrak{R}^T(\mu^\dagger)$. We first convert

[4] For example see Farina et al. (2019) or Lanctot et al. (2009).

the original regret minimization problem into a adversarial linear bandits one. Then we give an explanation behind the use of implicit exploration and introduce our algorithm, IXOMD, whose pseudocode is given in Algorithm 1. For simplicity, we gives a simple to read but memory-inefficient version as it keeps action probabilities for *each* information set. However, in Appendix B we provide a memory-efficient version.

Additional notation For a policy $\mu \in \Pi_{\max}$ and a sequence of functions $f := (f_h)_{h\in[H]}$, where $f_h : \mathcal{X}_h \times \mathcal{A} \to \mathbb{R}$, we denote the scalar product $\sum_{h\in[H]} \sum_{x_h\in\mathcal{X}_h, a\in\mathcal{A}} \mu_{1:h}(x_h, a_h) f_h(x_h, a_h)$ by $\langle \mu, f\rangle$. We let $\mathcal{F}^{t-1}$ be the sigma-algebra generated by variables up to the beginning of the t-th episode, i.e., $\{s_h^\tau, a_h^\tau, b_h^\tau\}_{h\in[H],\tau\in[t-1]}$. We let $\mathbb{E}^{t-1}[\cdot] := \mathbb{E}[\cdot|\mathcal{F}^{t-1}]$.

3.1 Conversion to online linear regret minimization

Note that for any profile (μ, ν), we have

$$\begin{aligned} V^{\mu,\nu} &= \sum_{h=1}^{H} \sum_{s_h\in\mathcal{S}_h, a_h\in\mathcal{A}, b_h\in\mathcal{B}} p_{1:h}(s_h)\mu_{1:h}(s_h, a_h)\nu_{1:h}(s_h, b_h) r_h(s_h, a_h, b_h) \\ &= \sum_{h=1}^{H} \sum_{x_h\in\mathcal{X}_h, a_h\in\mathcal{A}} \mu_{1:h}(x_h, a_h) \sum_{s_h\in x_h, b_h\in\mathcal{B}} p_{1:h}(s_h)\nu_{1:h}(s_h, b_h) r_h(s_h, a_h, b_h), \end{aligned}$$

where the second line follows from the facts that $\mu_{1:h}$ is dependent on (x_h, a_h) rather than (s_h, a_h), and $\sum_{s_h\in\mathcal{S}_h} f(s_h) = \sum_{x_h\in\mathcal{X}_h} \sum_{s_h\in x_h} f(s_h)$ for any function $f : \mathcal{S} \to \mathbb{R}$. Therefore defining a loss by

$$\ell_h^t(x_h, a_h) := \sum_{s_h\in x_h, b_h\in\mathcal{B}} p_{1:h}(s_h)\nu_{1:h}^t(s_h, b_h)(1 - r_h(s_h, a_h, b_h)),$$

we can rewrite the regret (1) as[5]

$$\mathfrak{R}^T(\mu^\dagger) = \sum_{t=1}^{T} \langle \mu^t - \mu^\dagger, \ell^t\rangle . \tag{2}$$

This result tells us that we may convert the original regret minimization problem to a linear one in which we choose μ^t such that $\mathfrak{R}^T(\mu^\dagger)$ grows sublinearly.

3.2 Loss estimation and implicit exploration

To solve the regret minimization problem (2) with bandit feedback, we need to estimate ℓ^t. Its importance sampling estimator is $\widehat{\ell}_h^t(x_h, a_h) := \mathbb{I}_{\{x_h=x_h^t, a_h=a_h^t\}}(1 - r_h^t)/\mu_{1:h}^t(x_h, a_h)$. However we estimate the loss by

$$\widetilde{\ell}_h^t(x_h, a_h) := \frac{\mathbb{I}_{\{x_h=x_h^t, a_h=a_h^t\}}}{\mu_{1:h}^t(x_h, a_h) + \gamma}(1 - r_h^t), \tag{3}$$

where γ is a positive real value and a hyper-parameter. This estimator is used by implicit exploration in bandits (IX, Kocák et al., 2014; Neu, 2015; Lattimore and Szepesvári, 2020, Chapter 12), and we therefore refer to it as the IX estimator. Note that IX uses a biased estimate, but it prevents the variance of the IX estimator from becoming too large.

[5] As introduced at **Additional notation**, $\langle \mu^t, \widetilde{\ell}^t\rangle = \sum_{h=1}^{H} \sum_{x_h\in\mathcal{X}_h, a\in\mathcal{A}} \mu_{1:h}^t(x_h, a_h)\widetilde{\ell}_h^t(x_h, a_h)$. Hence the meaning of μ^t here is abused, and we are viewing it as a sequence $(\mu_{1:h}^t)_{h\in[H]}$ of functions. In this case, μ^t must satisfy the following two conditions: (non-negativity) $\mu_{1:h}^t(x_h, a_h) \geq 0$ for any $x_h \in \mathcal{X}_h$ and $h \in [H]$; (consistency) $\sum_{a_h\in\mathcal{A}} \mu_{1:h}^t(x_h, a_h) = \mu_{1:h-1}^t(x_{h-1}, a_{h-1})$ for any $x_h \in \mathcal{X}_h$ and $h \in \{2, \dots, H\}$, where (x_{h-1}, a_{h-1}) is a unique predecessor of x_h, and $\sum_{a_1\in\mathcal{A}} \mu_{1:1}^t(x_1, a_1) = 1$ for any $x_1 \in \mathcal{X}_1$. Nonetheless there is a bijective mapping between $\Pi_{\max}$ and the set of μ^t satisfying these two conditions. Therefore we do not discern these two sets.

3.3 Efficient implementation

Given a loss estimate, we find μ^{t+1} by solving

$$\mu^{t+1} := \operatorname*{arg\,min}_{\mu \in \Pi_{\max}} \eta \left\langle \mu, \widetilde{\ell}^t \right\rangle + \mathrm{D}\big(\mu \| \mu^t\big)\,, \tag{4}$$

where D is the *dilated* entropy distance-generating function (Kroer et al., 2015) defined by

$$\mathrm{D}(\mu \| \mu') := \sum_{h=1}^{H} \sum_{x_h \in \mathcal{X}_h, a_h \in \mathcal{A}} \mu_{1:h}(x_h, a_h) \log \frac{\mu_h(a_h|x_h)}{\mu'_h(a_h|x_h)}\,.$$

Note that D is a Bregman divergence. The update in (4) has an easy implementation, as explained next. For more details of its derivation, please refer to Appendix A. To compute a new policy, we first need to compute for each $h \in [H]$,

$$\begin{aligned} Z_h^t &:= \sum_{a_h \in \mathcal{A}} \mu_h^t(a_h|x_h^t) \exp\Big(\mathbb{I}_{a_h = a_h^t}\Big(-\eta \widetilde{\ell}_H^t(x_h^t, a_h) + \log Z_{h+1}^t\Big)\Big) \\ &= 1 - \mu_h^t(a_h^t|x_h^t) + \mu_h^t(a_h^t|x_h^t) \exp\Big(-\eta \widetilde{\ell}_H^t(x_h^t, a_h) + \log Z_{h+1}^t\Big), \end{aligned}$$

with $Z_{H+1}^t := 1$. Then, we can compute a new policy by

$$\mu_h^{t+1}(a_h|x_h^t) = \mu_h^t(a_h|x_h^t) \exp\Big(\mathbb{I}_{a_h = a_h^t}\Big(-\eta \widetilde{\ell}_h^t(x_h^t, a_h) + \log Z_{h+1}^t\Big) - \log Z_h^t\Big)\,. \tag{5}$$

Note that this policy is updated *only* at the information sets visited along the t-th trajectory. This implies that the update requires $\mathcal{O}(HA)$ time-complexity per episode. The update of the average policy $\overline{\mu}$ is a bit more involved and costly, leading to a total time-complexity (for T episodes) of order $\mathcal{O}\big(TH(A\max(TH, X) + B\max(TH, Y)\big)$, see Appendix B.

4 Theoretical Analysis of IXOMD

We now analyze IXOMD. It has the following guarantee, which we shall prove in the present section.

Theorem 2 (regret bound of IXOMD)**.** *For any $\mu^\dagger \in \Pi_{\max}$ the regret (1) satisfies the following guarantee with probability at least $1-\delta$,*

$$\mathfrak{R}^T(\mu^\dagger) \le H\sqrt{2T\iota} + \gamma TXA + \frac{\iota}{2\gamma} + \frac{X \log A}{\eta} + \eta(1+H)TXA + \frac{\eta(1+H)\iota}{2\gamma}\,,$$

where $\iota := \log(3/\delta)$. In particularly, setting $\eta = \sqrt{\frac{\log A}{T(1+H)A}}$ and $\gamma = \sqrt{\frac{\iota}{2TXA}}$ results in

$$\mathfrak{R}^T(\mu^\dagger) \le H\sqrt{2T\iota} + \sqrt{2TXA\iota} + X\sqrt{T(1+H)A\log A} + \sqrt{\frac{2(1+H)XA\log A}{2\iota}}\,.$$

Remark 2. *We emphasize that this result is agnostic of the min-player. In particular, the same result holds for learning in a partially observable MDP with adversarial state-transition dynamics and reward function, as long as assumptions similar to the tree-like structure and perfect recall hold.*

Remark 3. *In Theorem 2, we adjusted η and γ using T, H, X, and A. Even when we know T only, setting $\eta = 1/\sqrt{T}$ and $\eta = \sqrt{\iota/T}$ guarantees $\mathfrak{R}^T(\mu^\dagger) \le \mathcal{O}(\mathrm{poly}(H, X, A, \log(1/\delta))\sqrt{T})$. If we additionally know H and A (which is likely to be the case), but* do not *know X, setting $\eta = \sqrt{\log A/(T(1+H)A)}$ and $\gamma = \sqrt{\iota/(2TA)}$ results in $X\sqrt{2TA\iota}$ in lieu of $\sqrt{2TXA\iota}$.*

A similar result holds for the min-player thanks to the symmetry. From Theorem 1 and 2, it follows that the average policies $(\overline{\mu}^t, \overline{\nu}^t)$ are close to a Nash equilibrium with high probability.

Corollary 2.1. *Suppose that both max- and min-players learn their policies by IXOMD with the setting[6] of η and γ in Theorem 2. Then, with probability at least $1-2\delta$ the average profile $(\overline{\mu}^t, \overline{\nu}^t)$ defined in Theorem 1 are ε-Nash equilibrium, with*

$$\varepsilon := \widetilde{\mathcal{O}}\left(\frac{1}{\sqrt{T}}\left(X\sqrt{HA} + Y\sqrt{HB} + H\right)\right),$$

where $\widetilde{\mathcal{O}}$ hides terms that are poly-log in X, A, Y, B, H, and $1/\delta$.

[6] Note that A and X must be replaced with B and Y (resp.) for the min-player's η and γ.

4.1 Proof of Theorem 2

In the first step, we decompose the regret (2) to three terms as

$$\mathfrak{R}^T(\mu^\dagger) = \underbrace{\sum_{t=1}^T \left\langle \mu^t, \ell^t - \widetilde{\ell}^t \right\rangle}_{\text{BIAS 1}} - \underbrace{\sum_{t=1}^T \left\langle \mu^\dagger, \ell^t - \widetilde{\ell}^t \right\rangle}_{\text{BIAS 2}} + \underbrace{\sum_{t=1}^T \left\langle \mu^t - \mu^\dagger, \widetilde{\ell}^t \right\rangle}_{\text{REGRET}} . \tag{6}$$

Then, we prove a high-probability upper-bound for each term. After deriving each upper-bound, Theorem 2 follows simply by taking the union bound over the three terms.

BIAS 1 It holds that BIAS 1 $\leq H\sqrt{2T\iota} + \gamma TXA$ with probability at least $1 - \delta/3$. To see that this is true, we first deduce that

$$\begin{aligned}\left\langle \mu^t, \widetilde{\ell}^t \right\rangle &= \sum_{h=1}^H \sum_{x_h \in \mathcal{X}_h, a \in \mathcal{A}} \mu^t_{1:h}(x_h, a_h) \frac{\mathbb{I}_{\{x_h = x^t_h, a_h = a^t_h\}}}{\mu^t_{1:h}(x_h, a_h) + \gamma} (1 - r^t_h) \\ &\leq \sum_{h=1}^H \sum_{x_h \in \mathcal{X}_h, a \in \mathcal{A}} \mathbb{I}_{\{x_h = x^t_h, a_h = a^t_h\}} = H ,\end{aligned}$$

where the inequality follows from $\mu^t_{1:h}(x_h, a_h)/(\mu^t_{1:h}(x_h, a_h) + \gamma) \leq 1$, and $0 \leq 1 - r^t_h \leq 1$. By Hoeffding-Azuma inequality, we deduce that $\sum_{t=1}^T \langle \mu^t, \widetilde{\ell}^t - \mathbb{E}^{t-1}[\widetilde{\ell}^t] \rangle \geq -H\sqrt{2T\iota}$ with probability at least $1 - \delta/3$. Next, we deduce that

$$\begin{aligned}\left\langle \mu^t, \ell^t - \mathbb{E}^{t-1}\left[\widetilde{\ell}^t\right] \right\rangle &= \sum_{h=1}^H \sum_{x_h \in \mathcal{X}_h, a \in \mathcal{A}} \mu^t_{1:h}(x_h, a_h) \left(1 - \frac{\mu^t_{1:h}(x_h, a_h)}{\mu^t_{1:h}(x_h, a_h) + \gamma}\right) \ell^t_h(x_h, a_h) \\ &= \sum_{h=1}^H \sum_{x_h \in \mathcal{X}_h, a \in \mathcal{A}} \mu^t_{1:h}(x_h, a_h) \frac{\gamma \ell^t_h(x_h, a_h)}{\mu^t_{1:h}(x_h, a_h) + \gamma} \\ &\leq \gamma \sum_{h=1}^H \sum_{x_h \in \mathcal{X}_h, a \in \mathcal{A}} \ell^t_h(x_h, a_h) \leq \gamma XA ,\end{aligned}$$

where again, the first inequality follows from $\mu^t(x_h, a_h)/(\mu^t_{1:h}(x_h, a_h) + \gamma) \leq 1$. Combining both bounds, we obtain the claimed result.

BIAS 2 It holds that BIAS 2 $\leq \iota/(2\gamma)$ with probability at least $1 - \delta/3$. This bound immediately follows from Lemma 1 of Neu (2015).

REGRET It holds that

$$\text{REGRET} \leq \frac{X \log A}{\eta} + \eta(1 + H)TXA + \frac{\eta(1 + H)\iota}{2\gamma},$$

with probability at least $1 - \delta/3$. The proof of this result is lengthy and technical. Therefore, we provide a proof sketch and defer the full proof and the proof of the lemmas to Appendix C. For brevity, we use the following notation: $\widetilde{\ell}^t_h := \widetilde{\ell}^t_h(x^t_h, a^t_h)$ and $\mu^t_h := \mu^t_h(x^t_h, a^t_h)$.

Lemma 4 states that $\eta\left\langle \mu^t - \mu^\dagger, \widetilde{\ell}^t \right\rangle = \mathrm{D}\big(\mu^\dagger \| \mu^t\big) - \mathrm{D}\big(\mu^\dagger \| \mu^{t+1}\big) + \mathrm{D}\big(\mu^t \| \mu^{t+1}\big)$, which implies

$$\begin{aligned}\sum_{t=1}^T \eta\left\langle \mu^t - \mu^\dagger, \widetilde{\ell}^t \right\rangle &= \mathrm{D}\big(\mu^\dagger \| \mu^1\big) - \mathrm{D}\big(\mu^\dagger \| \mu^{T+1}\big) + \sum_{t=1}^T \mathrm{D}\big(\mu^t \| \mu^{t+1}\big) \\ &\leq \mathrm{D}\big(\mu^\dagger \| \mu^1\big) + \sum_{t=1}^T \mathrm{D}\big(\mu^t \| \mu^{t+1}\big) .\end{aligned}$$

It is easy to upper-bound the first term by $X \log A$. The second term is trickier.

It turns out that $\mathrm{D}(\mu^t \| \mu^{t+1}) = \eta \langle \mu^t, \widetilde{\ell}^t \rangle + \log Z_1^t$ holds for any $t \in [T]$. Therefore, we need an upper bound on $\log Z_1^t$ and need to show it is close to $-\eta \langle \mu^t, \widetilde{\ell}^t \rangle$. This indeed holds. We have that

$$Z_1^t = 1 + \sum_{h=1}^{H} \mu_{1:h}^t \exp\left(-\eta \sum_{h'=1}^{h-1} \widetilde{\ell}_{h'}^t\right) \left(\exp\left(-\eta \widetilde{\ell}_h^t\right) - 1\right).$$

Using $\log(1+x) \le x$ for any $x \in (-1, \infty)$ and $\exp(-x) \le 1 - x + x^2$ for any $x \in (0, \infty)$, we deduce that

$$\log Z_1^t \le \sum_{h=1}^{H} \mu_{1:h}^t \exp\left(-\eta \sum_{h'=1}^{h-1} \widetilde{\ell}_{h'}^t\right) \left(-\eta \widetilde{\ell}_h^t + \eta^2 \left(\widetilde{\ell}_h^t\right)^2\right).$$

Using $1 - \exp(-x) \le x$ for $x \in \mathbb{R}$ yields

$$\mu_{1:h}^t \left(1 - \exp\left(-\eta \sum_{h'=1}^{h-1} \widetilde{\ell}_{h'}^t\right)\right) \le \eta \mu_{1:h}^t \sum_{h'=1}^{h-1} \widetilde{\ell}_{h'}^t \le \eta H\,,$$

where the last inequality follows from $\mu_{1:h}^t \widetilde{\ell}_{h'}^t = \mu_{1:h}^t (1 - r_h^t)/(\mu_{1:h'}^t + \gamma) \le 1$ for any $h' \le h$. Accordingly,

$$\mathrm{D}\big(\mu^t \| \mu^{t+1}\big) \le \eta^2 H \sum_{h=1}^{H} \widetilde{\ell}_h^t + \eta^2 \sum_{h=1}^{H} \mu_{1:h}^t \exp\left(-\eta \sum_{h'=1}^{h-1} \widetilde{\ell}_{h'}^t\right) \left(\widetilde{\ell}_h^t\right)^2 \le \eta^2 (1+H) \sum_{h=1}^{H} \widetilde{\ell}_h^t\,,$$

where again, we used $\mu_{1:h}^t \widetilde{\ell}_{h'}^t \le 1$. Recalling that $\widetilde{\ell}_h^t$ is non-zero only at (x_h^t, a_h^t), we have that $\sum_{h=1}^{H} \widetilde{\ell}_h^t = \sum_{h=1}^{H} \sum_{x_h \in \mathcal{X}_h, a_h \in \mathcal{A}} \widetilde{\ell}_h^t(x_h, a_h)$. Thus we can use Lemma 1 by Neu (2015) and obtain

$$\begin{aligned}
\eta^2 (1+H) \sum_{t=1}^{T} \sum_{h=1}^{H} \widetilde{\ell}_h^t &\le \eta^2 (1+H) \sum_{t=1}^{T} \sum_{h=1}^{H} \sum_{x_h \in \mathcal{X}_h, a_h \in \mathcal{A}} \ell_h^t(x_h, a_h) + \frac{\eta^2 (1+H) \iota}{2\gamma} \\
&\le \eta^2 (1+H) T X A + \frac{\eta^2 (1+H) \iota}{2\gamma}
\end{aligned}$$

with probability at least $1 - \delta/3$. Consequently, we obtain the claimed bound of REGRET.

5 Conclusion

We theoretically studied the problem of learning a NE of a EGII under a perfect-recall assumption. We provided the IXOMD algorithm based on OMD with the dilated entropy distance-generating function as a regularizer and implicit exploration for estimation of the losses. We proved a high-probability bound on the convergence rate to the NE of order $\widetilde{\mathcal{O}}(X\sqrt{A} + Y\sqrt{B})/\sqrt{T})$ derived from a regret bound of order $\widetilde{\mathcal{O}}((X\sqrt{AT})$ (for the max-player). Notably, the regret bound remains valid in the adversarial setting (where the other player and the game are picked by an adversary). Furthermore, due to our choice of the regularizer, the updates of the policy (e.g., of the max-player) could be implemented with a time-complexity of $\mathcal{O}(HA)$ per episode, which makes IXOMD also computationally efficient. However, the computation of the average profile is more costly, leading to a total time complexity (after T episodes) of order $\mathcal{O}\big(TH(A \max(TH, X) + B \max(TH, Y)\big)$.

An interesting next direction would be to characterize the problem-independent optimal regret, e.g., for the max-player, in our setting. We conjecture that it is of order $\widetilde{\mathcal{O}}(\sqrt{XAT})$ even in the adversarial setting (where the min-player and the game are picked by an adversary). This would make our current bound to be loose by a factor $\sqrt{X}$. Finally, we also hypothesize that the time-complexity for the computation of the average policy could be potentially improved.

References

Jacob Abernethy, U C Berkeley, and Alexander Rakhlin. Competing in the Dark : An Efficient Algorithm for Bandit Linear Optimization. In *Conference on Learning Theory*, 2008.

Neil Burch, Matej Moravcik, and Martin Schmid. Revisiting CFR+ and Alternating Updates. *Journal of Artificial Intelligence Research*, 2019.

Varsha Dani, Thomas Hayes, and Sham Kakade. The Price of Bandit Information for Online Optimization. In J C Platt, D Koller, Y Singer, and S Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20, 2008.

Gabriele Farina and Tuomas Sandholm. Model-Free Online Learning in Unknown Sequential Decision Making Problems and Games. In *AAAI Conference on Artificial Intelligence*, 2021.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Optimistic Regret Minimization for Extensive-Form Games via Dilated Distance-Generating Functions. In *Advances in Neural Information Processing Systems*, 2019.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Stochastic Regret Minimization in Extensive-Form Games. In *International Conference on Machine Learning*, 2020.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Faster Game Solving via Predictive Blackwell Approachability: Connecting Regret Matching and Mirror Descent. In *AAAI Conference on Artificial Intelligence*, 2021a.

Gabriele Farina, Christian Kroer, and Tuomas Sandholm. Bandit Linear Optimization for Sequential Decision Making and Extensive-Form Games. In *AAAI Conference on Artificial Intelligence*, 2021b.

Abraham D. Flaxman, Adam Tauman Kalai, and H. Brendan McMahan. Online Convex Optimization in the Bandit Setting: Gradient Descent without a Gradient. In *Annual ACM-SIAM Symposium on Discrete Algorithms*, 2005.

Andrew Gilpin, Javier Peña, and Tuomas Sandholm. First-order algorithm with $O(ln(1/\epsilon)))$ convergence for ϵ-equilibrium in two-person zero-sum games. *Mathematical Programming*, 2012.

Geoffrey J Gordon. No-regret Algorithms for Online Convex Programs. In *Advances in Neural Information Processing Systems*, 2007.

Sergiu Hart and Andreu Mas-Colell. A Simple Adaptive Procedure Leading to Correlated Equilibrium. *Econometrica*, 2000.

Samid Hoda, Andrew Gilpin, Javier Peña, and Tuomas Sandholm. Smoothing Techniques for Computing Nash Equilibria of Sequential Games. *Mathematics of Operations Research*, 2010.

Chi Jin, Tiancheng Jin, Haipeng Luo, Suvrit Sra, and Tiancheng Yu. Learning Adversarial Markov Decision Processes with Bandit Feedback and Unknown Transition. In *International Conference on Machine Learning*, 2020.

Tomáš Kocák, Gergely Neu, Michal Valko, and Rémi Munos. Efficient learning by implicit exploration in bandit problems with side observations. In *Advances in Neural Information Processing Systems*, 2014.

Daphne Koller, Nimrod Megiddo, and Bernhard Von Stengel. Efficient Computation of Equilibria for Extensive Two-Person Games. *Games and Economic Behavior*, 1996.

Christian Kroer, Kevin Waugh, Fatma Kilinç-Karzan, and Tuomas Sandholm. Faster First-Order Methods for Extensive-Form Game Solving . In *ACM Conference on Economics and Computation*, 2015.

Christian Kroer, Gabriele Farina, and Tuomas Sandholm. Solving Large Sequential Games with the Excessive Gap Technique. In *Advances in Neural Information Processing Systems*, 2018.

Christian Kroer, Kevin Waugh, Fatma Kılınç-Karzan, and Tuomas Sandholm. Faster algorithms for extensive-form game solving via improved smoothing functions. *Mathematical Programming*, 2020.

Harold W Kuhn. Extensive Games and the Problem of Information. *Annals of Mathematics Studies*, 1953.

Marc Lanctot, Kevin Waugh, Martin Zinkevich, and Michael Bowling. Monte Carlo Sampling for Regret Minimization in Extensive Games. In *Advances in Neural Information Processing Systems*, 2009.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Michael L. Littman. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *International Conference on Machine Learning*, 1994.

Remi Munos, Julien Perolat, Jean-Baptiste Lespiau, Mark Rowland, Bart De Vylder, Marc Lanctot, Finbarr Timbers, Daniel Hennes, Shayegan Omidshafiei, Audrunas Gruslys, et al. Fast Computation of Nash Equilibria in Imperfect Information Games. In *International Conference on Machine Learning*, 2020.

John F Nash Jr. Equilibrium points in N-person games. *Proceedings of the National Academy of Sciences of the United States of America*, 1950.

Arkadi Nemirovski. Prox-Method with Rate of Convergence $O(1/t)$ for Variational Inequalities with Lipschitz Continuous Monotone Operators and Smooth Convex-Concave Saddle Point Problems. *SIAM Journal on Optimization*, 2004.

Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 2005.

Gergely Neu. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. In *Advances in Neural Information Processing Systems*, 2015.

Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. The MIT Press, 1994.

Marc Ponsen, Steven De Jong, and Marc Lanctot. Computing Approximate Nash Equilibria and Robust Best-Responses Using Sampling. *Journal of Artificial Intelligence Research*, 2011.

J. V. Romanovsky. Reduction of a game with complete memory to a matricial game. *Dokl. Akad. Nauk SSSR*, 1962.

Aviv Rosenberg and Yishay Mansour. Online Convex Optimization in Adversarial Markov Decision Processes. In *International Conference on Machine Learning*, 2019.

Lloyd S Shapley. Stochastic Games. *Proceedings of the National Academy of Sciences of the United States*, 1953.

Malcolm Strens. A Bayesian Framework for Reinforcement Learning. In *International Conference on Machine Learning*, 2000.

Oskari Tammelin. Solving Large Imperfect Information Games Using CFR+. *arXiv preprint arXiv:1407.5042*, 2014.

Bernhard von Stengel. Efficient Computation of Behavior Strategies. *Games and Economic Behavior*, 1996.

Brian Hu Zhang and Tuomas Sandholm. Finding and Certifying (Near-) Optimal Strategies in Black-Box Extensive-Form Games. In *AAAI Conference on Artificial Intelligence*, 2021.

Yichi Zhou, J. Li, and J. Zhu. Posterior sampling for multi-agent reinforcement learning: solving extensive games with imperfect information. In *International Conference on Learning Representations*, 2020.

Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. Regret Minimization in Games with Incomplete Information. *Advances in neural information processing systems*, 2007.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] The computation of the average profile in Thereom 1 needs relatively large number of computations.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We did not, because the present paper is purely theoretical work, and it is difficult to predict any negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]