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

Tadashi Kozuno*, Pierre Ménard*, Rémi Munos, Michal Valko (* equal contribution)

UNIVERSITY OF ALBERTA · OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG · DeepMind · Inria

Two-player Zero-sum IIG with Perfect Recall

$\mathcal{S}$: State space of size S, Horizon H

$\mathcal{X}$: Max-player's information set space of size X

$\mathcal{A}$: Max-player's action space of size A

$\mathcal{Y}$: Min-player's information set space of size Y

$\mathcal{B}$: Min-player's action space of size B

r_h, p_h: Reward/loss function and state-transition dynamics

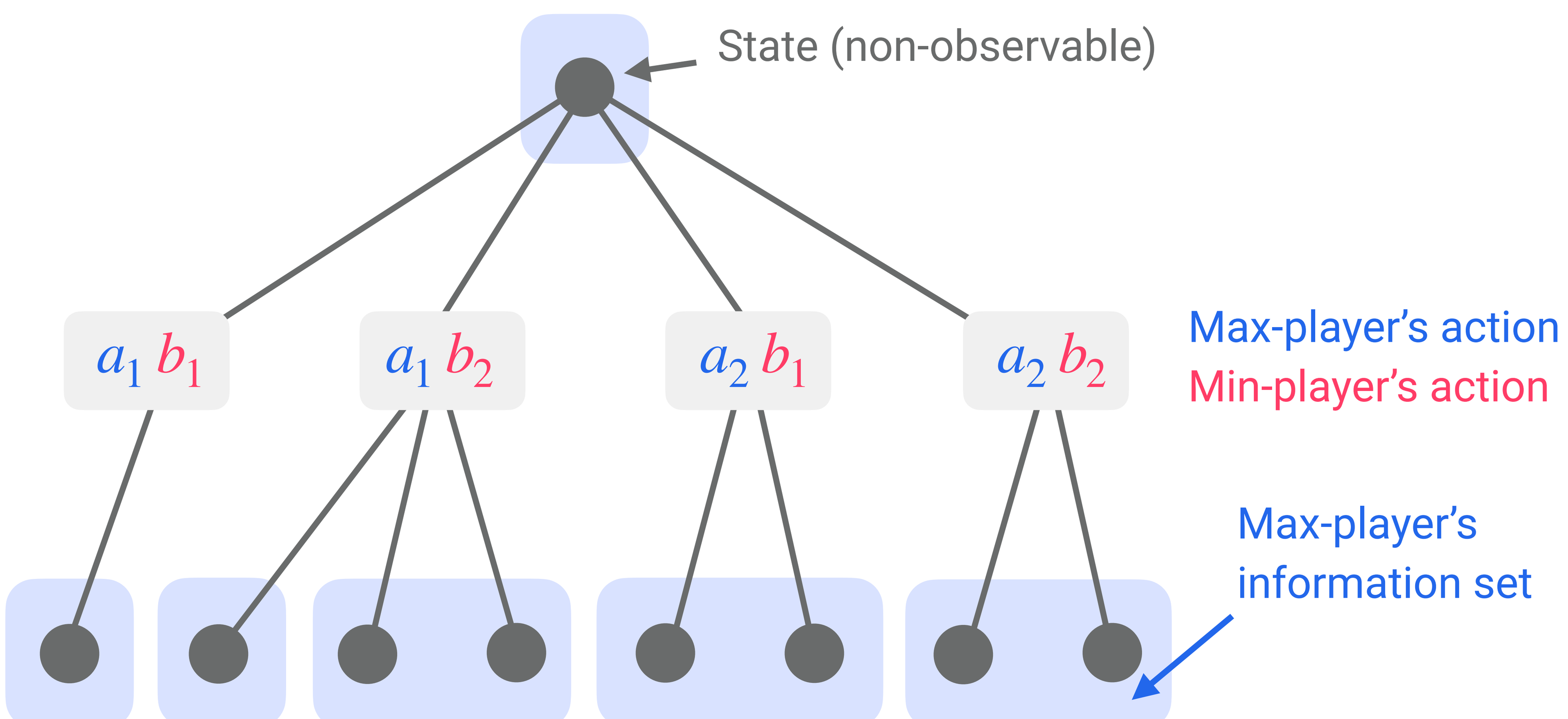
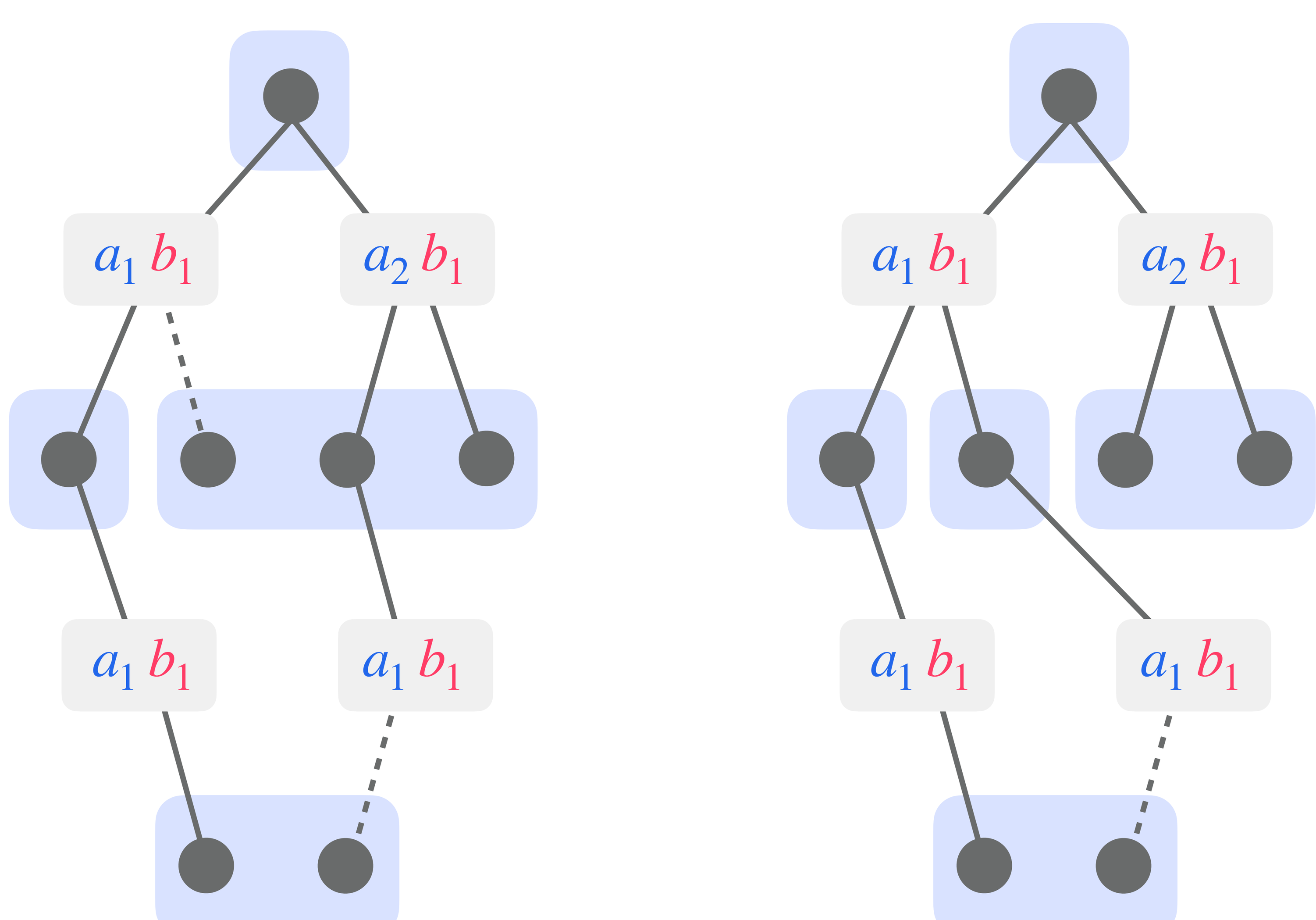


Fig 1. An IIG with $H = 2, \mathcal{A} = \{a_1, a_2\}$, and $\mathcal{B} = \{b_1, b_2\}$. Only max-player's information sets are shown.



Actions forgotten

Information set forgotten

Fig 2. Examples where the perfect-recall assumption is not met at transitions indicated by dashed lines.

The Problem

Find a Nash equilibrium (NE) of an imperfect information game (IIG) with perfect recall, with high probability, only using bandit feedbacks.

Our Contributions

- *Propose a computationally efficient model-free algorithm called* IXOMD*, by combining implicit exploration (IX) and online mirror descent (OMD).*
- *A high-prob exploitability gap bound of order $1/\sqrt{T}$.*
- *A high-prob regret bound of order $\sqrt{T}$.*

T: Number of game plays

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 $t = 1, \dots, T$ **do**
  for $h = 1, \dots, H$ **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 $h = H, \dots, 1$ **do**
    Construct the IX loss estimate $\widetilde{\ell}_h^t$ by

$$\widetilde{\ell}_h^t \leftarrow \frac{1 - r_h^t}{\prod_{i=1}^h \mu_i^t(a_i^t|x_i^t) + \gamma}.$$

    For each $h \in [H]$ (with $Z_{H+1}^t \leftarrow 1$)

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

    Update μ^t to μ^{t+1} at x_h^t by

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

    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 *Average policy* $\overline{\mu}$

Regret, Average Profile, and Nash Equilibrium

For a profile (μ, ν), the expected return (of the max-player) is defined by

$$V^{\mu,\nu} := \mathbb{E}^{\mu,\nu}\left[\sum_{h=1}^H r_h(s_h, a_h, b_h)\right].$$

When a profile (μ, ν) satisfies the following, it is said to be an ε-NE:

$$\max_{\mu'} V^{\mu',\nu} - \min_{\nu} V^{\mu,\nu'} \le \varepsilon$$

The LHS is an exploitability gap. For a sequence of profiles (μ^t, ν^t), the regret of the max-player, relative to some policy μ, is defined as

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

An average profile $(\bar{\mu}, \bar{\nu})$ is a profile such that

$$V^{\bar{\mu},\nu} = \sum_{t=1}^T V^{\mu^t,\nu}/T \text{ and } V^{\mu,\bar{\nu}} = \sum_{t=1}^T V^{\mu,\nu^t}/T$$

for any profile (μ, ν). It is guaranteed to exist and computable.

Main Theorem

Let $\delta \in (0,1)$. If the max-player is trained by IXOMD with appropriate learning rate and IX parameter, then with probability at least $1-\delta$, its regret $\mathfrak{R}_{\max}^T$ is bounded by $\tilde{O}(X\sqrt{AT})$. If the min-player is trained similarly, then with probability at least $1-\delta$, the average profile $(\bar{\mu}, \bar{\nu})$ is ε-Nash equilibrium, where

$$\varepsilon := \tilde{O}\left((X\sqrt{A} + Y\sqrt{B})/\sqrt{T}\right).$$

Comparison to Previous Results

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 IIG with bandit feedback and their respective upper bound on the exploitability gap after T episodes. In 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.