Adapting to game trees in zero-sum imperfect information games

Côme Fiegel, Pierre Ménard, Tadashi Kozuno, Rémi Munos, Vianney Perchet and Michal Valko

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

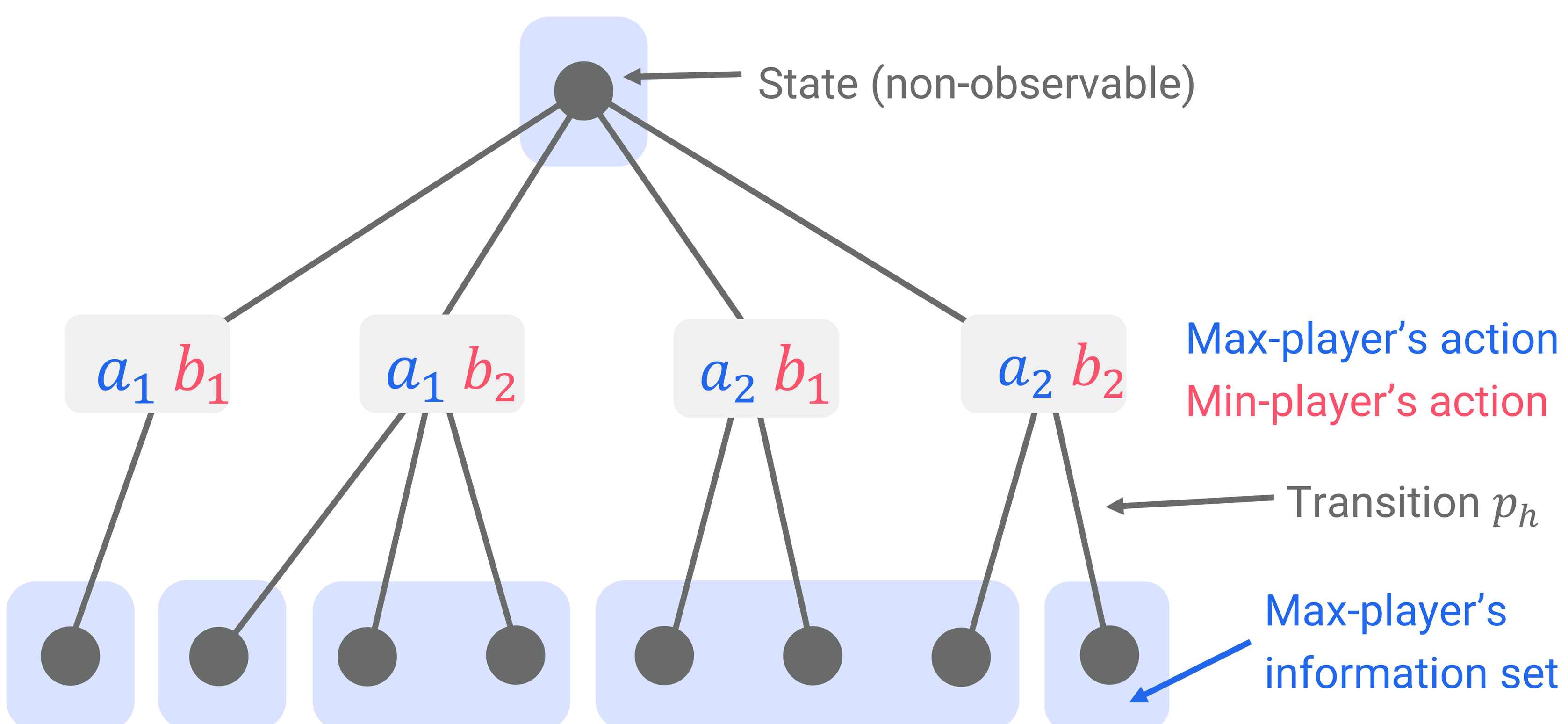


Fig 1. An IIG with $H = 2$, $\mathcal{A} = \{a_1, a_2\}$, and $\mathcal{B} = \{b_1, b_2\}$. Only max-player's information sets are shown.

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:

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

For a sequence of profiles (μ^t, ν^t), the regret of the max-player is

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

The time-averaged profile $(\bar{\mu}, \bar{\nu})$ is an ε-NE with:

$$\varepsilon = \frac{\mathfrak{R}^T_{max} + \mathfrak{R}^T_{min}}{T}\,.$$

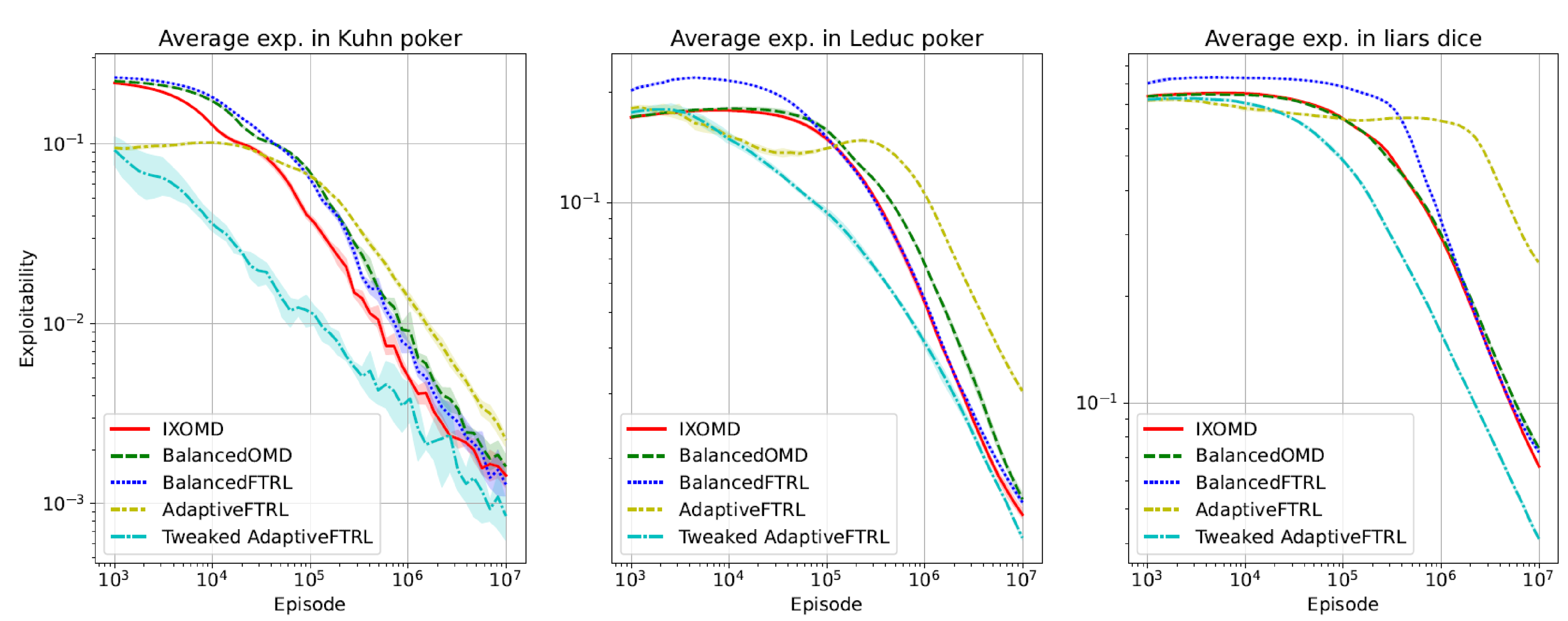
The Problem

Find an approximation through self-play of an optimal strategy for a zero-sum imperfect information game only using trajectory feedback.

.

Our Contributions

- *Propose two computationally efficient algorithms, by combining implicit exploration and follow the regularized leader.*
- *If applied by both players, the first has an optimal high-probability sample complexity of order $H\,(AX + BY)/\varepsilon^2$ requiring the knowledge of the structure.*
- *The second has a sample complexity of order $H^2(AX + BY)/\varepsilon^2$ without this knowledge, using an adaptive regularization.*



Performances of various algorithms with respect to the number of episodes. The vertical axis corresponds to the smallest ε such that the output is an ε-NE.

Algorithm	Sample complexity	Structure-free
MCCFR (Farina et al., 2020; Bai et al., 2022)	$\widetilde{\mathcal{O}}(H^4(AX+BY)/\varepsilon^2)$	✗
IXOMD (Kozuno et al., 2021)	$\widetilde{\mathcal{O}}(H^2(AX^2+BY^2)/\varepsilon^2)$	✓
Balanced OMD (Bai et al., 2022)	$\widetilde{\mathcal{O}}(H^3(AX+BY)/\varepsilon^2)$	✗
Balanced FTRL (this paper)	$\widetilde{\mathcal{O}}(H(AX+BY)/\varepsilon^2)$	✗
Adaptive FTRL (this paper)	$\widetilde{\mathcal{O}}(H^2(AX+BY)/\varepsilon^2)$	✓
Lower bound (this paper)	$\widetilde{\mathcal{O}}(H(A_{\mathcal{X}}+B_{\mathcal{Y}})/\varepsilon^2)$	

Sample complexity for various algorithms. Structure-free means that the algorithm does not need to know the structure of the information set spaces in advance.

Algorithm 1 Adaptive FTRL for the max-player

1: **Input:**
  Base learning rate η and IX bias γ
  Uniform policy μ^0
  $\mu_{1:h}(x, a)$ denotes the combined probability for the max-player of choosing actions that lead to (x, a)
2: **For** $t = 1$ to T :
  For all h and $x_h \in \mathcal{X}_h$, compute learning rate:
    $\eta^t_h(x_h) \leftarrow \min_{x'_{h'} \geq x_h} \eta/(1 + \widetilde{P}^{t-1}_h(x'_{h'}))$
    For all $a_h \in \mathcal{A}(x_h)$, compute bias rate:
      $\gamma^t_h(x_h, a_h) \leftarrow \gamma/(1 + \widetilde{P}^{t-1}_h(x_h, a_h))$
  Compute update:
    $\mu^t \leftarrow \operatorname{argmin}_{\mu} \sum_h \left\langle \mu_{1:h}, \widetilde{L}^{t-1}_h \right\rangle + \mathcal{D}_{\eta^t}(\mu, \mu^0)$
  with $\mathcal{D}_{\eta}(\mu, \mu^0) = \sum_x \mu_{1:h}(x)\mathrm{KL}(\mu(\cdot|x), \mu^1(\cdot|x))/\eta(x)$
  For $h = 1$ to H:
    Observe information set x^t_h
    Execute $a^t_h \sim \mu^t(.|x^t_h)$ and **receive** reward r^t_h
    $\widetilde{L}^t_h \leftarrow \widetilde{L}^{t-1}_h + \mathbb{I}_{\{x^t_h, a^t_h\}}(1 - r^t_h)/(\mu^t_{1:h} + \gamma^t_h)$
    $\widetilde{P}^t_h \leftarrow \widetilde{P}^{t-1}_h + \mathbb{I}_{\{x^t_h, a^t_h\}}/(\mu^t_{1:h} + \gamma^t_h)$
  Return average policy $\overline{\mu}$