Adapting to game trees in zero-sum imperfect information games

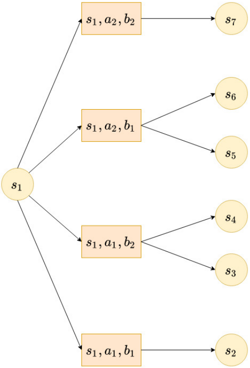
Côme Fiegel[1], Pierre Ménard[2], Tadashi Kozuno[3] ,
Rémi Munos[4], Vianney Perchet[1,5], Michal Valko[4]

[1]ENSAE Paris [2]ENS Lyon [3]Omron Sinic X [4]DeepMind [5]Criteo AI Lab

International Conference on Machine Learning 2023

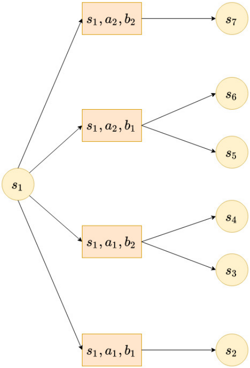
Zero-sum (two players) imperfect information games

- State space $\mathcal{S}$, initial state $s_1 \in \mathcal{S}$ and horizon $H > 0$
- At timestep $h \in [1..H]$, the two players take actions $a \in \mathcal{A}$ and $b \in \mathcal{B}$
- Reward $r_h(s, a, b) \in [0, 1]$ and transition to the next state $p_h(\cdot|s, a, b)$



Zero-sum (two players) imperfect information games

- State space $\mathcal{S}$, initial state $s_1 \in \mathcal{S}$ and horizon $H > 0$
- At timestep $h \in [1..H]$, the two players take actions $a \in \mathcal{A}$ and $b \in \mathcal{B}$
- Reward $r_h(s, a, b) \in [0, 1]$ and transition to the next state $p_h(\cdot|s, a, b)$

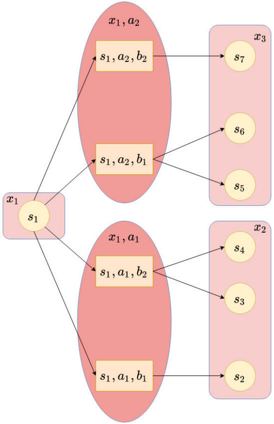


Assumptions:

Zero-sum: max-player receives r_h, min-player receives $-r_h$

Zero-sum (two players) imperfect information games

- State space $\mathcal{S}$, initial state $s_1 \in \mathcal{S}$ and horizon $H > 0$
- At timestep $h \in [1..H]$, the two players take actions $a \in \mathcal{A}$ and $b \in \mathcal{B}$
- Reward $r_h(s, a, b) \in [0, 1]$ and transition to the next state $p_h(\cdot|s, a, b)$



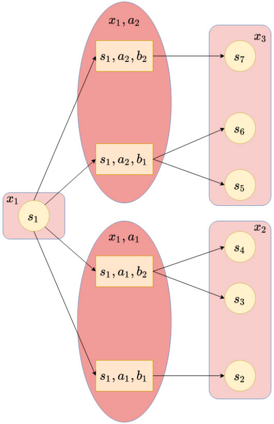
Assumptions:

Zero-sum: max-player receives r_h, min-player receives $-r_h$

Imperfect information: Players only observe information sets $x(s) \in \mathcal{X}$ and $y(s) \in \mathcal{Y}$

Zero-sum (two players) imperfect information games

- State space $\mathcal{S}$, initial state $s_1 \in \mathcal{S}$ and horizon $H > 0$
- At timestep $h \in [1..H]$, the two players take actions $a \in \mathcal{A}$ and $b \in \mathcal{B}$
- Reward $r_h(s, a, b) \in [0, 1]$ and transition to the next state $p_h(\cdot|s, a, b)$



Assumptions:

Zero-sum: max-player receives r_h, min-player receives $-r_h$

Imperfect information: Players only observe information sets $x(s) \in \mathcal{X}$ and $y(s) \in \mathcal{Y}$

Perfect recall: Players do not forget past observations and actions

Approximate Nash equilibrium

Policies: Non-deterministic $\mu = (\mu(\cdot|x))_{x \in \mathcal{X}}$ and $\nu = (\nu(\cdot|y))_{y \in \mathcal{Y}}$

Approximate Nash equilibrium

Policies: Non-deterministic $\mu = (\mu(\cdot|x))_{x \in \mathcal{X}}$ and $\nu = (\nu(\cdot|y))_{y \in \mathcal{Y}}$

Value: $V^{\mu,\nu} = \mathbb{E}^{\mu,\nu}\left[\sum_{h=1}^{H} r_h\right]$

Approximate Nash equilibrium

Policies: Non-deterministic $\mu = (\mu(\cdot|x))_{x \in \mathcal{X}}$ and $\nu = (\nu(\cdot|y))_{y \in \mathcal{Y}}$

Value: $V^{\mu,\nu} = \mathbb{E}^{\mu,\nu}\left[\sum_{h=1}^{H} r_h\right]$

(μ, ν) is a **Nash equilibrium** if $\mu \in \operatorname{argmax} V^{\cdot,\nu}$ and $\nu \in \operatorname{argmin} V^{\mu,\cdot}$

Approximate Nash equilibrium

Policies: Non-deterministic $\mu = (\mu(\cdot|x))_{x \in \mathcal{X}}$ and $\nu = (\nu(\cdot|y))_{y \in \mathcal{Y}}$

Value: $V^{\mu,\nu} = \mathbb{E}^{\mu,\nu}\left[\sum_{h=1}^{H} r_h\right]$

(μ, ν) is a **Nash equilibrium** if $\mu \in \operatorname{argmax} V^{\cdot,\nu}$ and $\nu \in \operatorname{argmin} V^{\mu,\cdot}$

Objective $\rightarrow$ Approximate a **Nash equilibrium**

Sequential learning

Interaction with the game: T episodes played using freely chosen profiles $(\mu^t, \nu^t)_{t \in [1..T]}$

Sequential learning

Interaction with the game: T episodes played using freely chosen profiles $(\mu^t, \nu^t)_{t \in [1..T]}$

Regrets: $\mathfrak{R}^{T}_{\max} = \max_{\mu} \sum_t \left[V^{\mu, \nu^t} - V^{\mu^t, \nu^t} \right]$, $\mathfrak{R}^{T}_{\min} = \max_{\nu} \sum_t \left[V^{\mu^t, \nu^t} - V^{\mu^t, \nu} \right]$

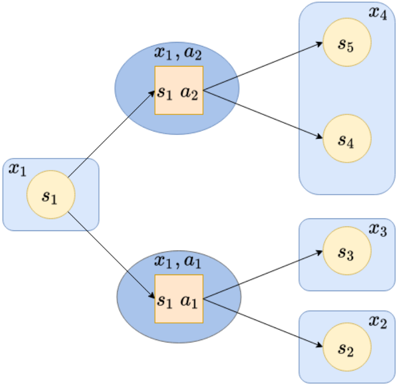
Sequential learning

Interaction with the game: T episodes played using freely chosen profiles $(\mu^t, \nu^t)_{t\in[1..T]}$

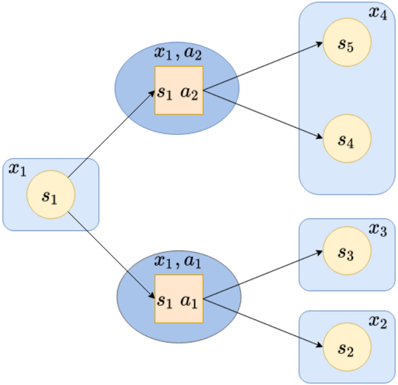
Regrets: $\mathfrak{R}^T_{\max} = \max_\mu \sum_t \left[V^{\mu,\nu^t} - V^{\mu^t,\nu^t}\right]$, $\mathfrak{R}^T_{\min} = \max_\nu \sum_t \left[V^{\mu^t,\nu^t} - V^{\mu^t,\nu}\right]$

Small **regrets** $\iff$ average profile $(\overline{\mu}, \overline{\nu})$ approximates a **Nash equilibrium**

Max player's point of view



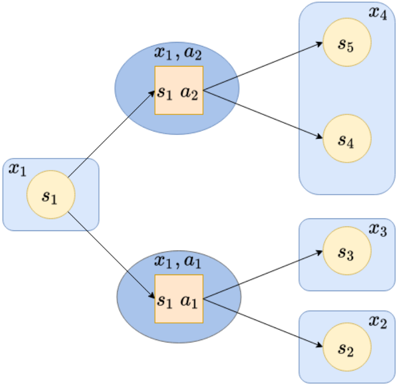
Max player's point of view



Advantage

Episodic MDP with a tree-structure

Max player's point of view



Advantage

Episodic MDP with a tree-structure

Difficulty

Adversarial transitions p^t that change between episodes

Back to regret minimization

Objective: minimize $\mathfrak{R}^T_{\max} = \max_\mu \sum_{t=1}^T \left[V^{\mu,\nu^t} - V^{\mu^t,\nu^t}\right]$

FTRL approach : $\mu^{t+1} = \operatorname{argmax}_\mu \sum_{k=1}^t \tilde{V}^{\mu,\nu^k} - \Psi(\mu)$ with

- $\tilde{V}^{\cdot,\nu^t}$ estimated value at episode t as a function of μ
- Ψ the regularizer

Back to regret minimization

Objective: minimize $\mathfrak{R}^T_{\max} = \max_\mu \sum_{t=1}^T \left[V^{\mu,\nu^t} - V^{\mu^t,\nu^t}\right]$

FTRL approach : $\mu^{t+1} = \operatorname{argmax}_\mu \sum_{k=1}^t \tilde{V}^{\mu,\nu^k} - \Psi(\mu)$ with

- $\tilde{V}^{\cdot,\nu^t}$ estimated value at episode t as a function of μ
- Ψ the regularizer

How to choose Ψ?

Regularizer choice

First choice: **BalancedFTRL**

- $\Psi_p(\mu) = \textit{negentropy}\,(\text{information set} \mid \mu, p)/\eta$
- Compute balanced transitions $p^\star$ and use Ψ_{p^*}
- $\mathfrak{R}^T_{\mathbf{max}} = \widetilde{\mathcal{O}}\left(\sqrt{H\,|\mathcal{X}|\,|\mathcal{A}|\,T}\right)$

Regularizer choice

First choice: **BalancedFTRL**

- $\Psi_p(\mu) = \textit{negentropy}\,(\text{information set} \mid \mu, p)/\eta$
- Compute balanced transitions $p^\star$ and use Ψ_{p^*}
- $\mathfrak{R}^T_{\mathbf{max}} = \widetilde{\mathcal{O}}\left(\sqrt{H\,|\mathcal{X}|\,|\mathcal{A}|\,T}\right)$

Impossible with initially unknown tree structure

Regularizer choice

First choice: **BalancedFTRL**

- $\Psi_p(\mu) = \textit{negentropy}\,(\text{information set} \mid \mu, p)/\eta$
- Compute balanced transitions $p^\star$ and use Ψ_{p^*}
- $\mathfrak{R}^T_{\mathbf{max}} = \widetilde{\mathcal{O}}\left(\sqrt{H\,|\mathcal{X}|\,|\mathcal{A}|\,T}\right)$

Impossible with initially unknown tree structure

Second choice: **AdaptiveFTRL**

- Estimate cumulative transitions $\tilde{P}^t$
- "Replace" $p^\star$ with $\tilde{P}^t$
- $\mathfrak{R}^T_{\mathbf{max}} = \widetilde{\mathcal{O}}\left(H\sqrt{|\mathcal{X}|\,|\mathcal{A}|\,T}\right)$

Conclusion

Algorithm	Sample complexity	Structure-free
IXOMD	$\widetilde{\mathcal{O}}(H^2(\lvert\mathcal{X}\rvert^2\lvert\mathcal{A}\rvert + \lvert\mathcal{Y}\rvert^2\lvert\mathcal{B}\rvert)/\epsilon^2)$	✓
BalancedOMD	$\widetilde{\mathcal{O}}(H^3(\lvert\mathcal{X}\rvert\lvert\mathcal{A}\rvert + \lvert\mathcal{Y}\rvert\lvert\mathcal{B}\rvert/\epsilon^2))$	✗
BalancedFTRL	$\widetilde{\mathcal{O}}(H(\lvert\mathcal{X}\rvert\lvert\mathcal{A}\rvert + \lvert\mathcal{Y}\rvert\lvert\mathcal{B}\rvert)/\epsilon^2)$	✗
AdaptiveFTRL	$\widetilde{\mathcal{O}}(H^2(\lvert\mathcal{X}\rvert\lvert\mathcal{A}\rvert + \lvert\mathcal{Y}\rvert\lvert\mathcal{B}\rvert)/\epsilon^2)$	✓
Lower bound	$\widetilde{\mathcal{O}}(H(\lvert\mathcal{X}\rvert\lvert\mathcal{A}\rvert + \lvert\mathcal{Y}\rvert\lvert\mathcal{B}\rvert)/\epsilon^2)$	-

Contact: come.fiegel@gmail.com