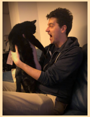
DeepMind

Model-free learning for two-player partially observable zero-sum Markov games

Tadashi Kozuno[1], Pierre Ménard[2], Rémi Munos[3,4], **Michal Valko**[3,4]

[1]University of Alberta [2]Otto von Guericke Universität [3]DeepMind [4]Inria



ELLIS@Milan Artificial Intelligence workshop 9/9/2022

Think of playing poker: What do we bring?

- Model-free algorithm for two-player zero-sum games

Think of playing poker: What do we bring?

- Model-free algorithm for two-player zero-sum games
  - **Only updates the policy along the trajectory**
  - Works under perfect-recall assumption

Think of playing poker: What do we bring?

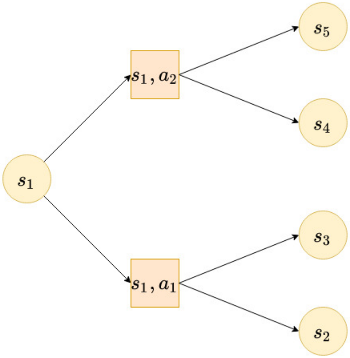
- Model-free algorithm for two-player zero-sum games
  - **Only updates the policy along the trajectory**
  - Works under perfect-recall assumption
- Algorithm only needs bandit (trajectory) feedback
  - Dynamics of the game does not need to be known
  - Converges to Nash with the rate of $1/\sqrt{T}$ w.h.p.

Part I: Tree Markov Decision Process (MDP)

- state space $\mathcal{S}$ of size S
- action space $\mathcal{A}$ of size A
- horizon H
- reward functions $r_h(s, a)$ and transition probabilities $p_h(\cdot|s, a)$

Part I: Tree Markov Decision Process (MDP)

- state space $\mathcal{S}$ of size S
- action space $\mathcal{A}$ of size A
- horizon H
- reward functions $r_h(s,a)$ and transition probabilities $p_h(\cdot|s,a)$



Assumptions:
Tree structure: all the history in the state, for any $s \in S$ there is a unique step h and sub-trajectory leading to $s_1, a_1, \ldots, s_h = s$

Value of policy $\mu_h(a|s)$

$$V^{\mu} = \sum_{s,a,h} p^{\mu}_{1:h}(s,a) r_h(s,a)$$

Value of policy $\mu_h(a|s)$

$$V^\mu = \sum_{s,a,h} p^\mu_{1:h}(s,a) r_h(s,a)$$

Reach probability (because of tree structure)

$$p^\mu_{1:h}(s,a) = \underbrace{\mu_{1:h}(s,a)}_{\text{agent}} \underbrace{p^\emptyset_h(s)}_{\text{nature}}$$

where along trajectory $s_1, a_1, \ldots, s_h = s, a_h = a$, realization plan of agent/nature

$$\mu_{1:h}(s,a) = \prod_{h'=1}^{h} \mu_{h'}(a_{h'}|s_{h'})$$

$$p^\emptyset_h(s) = p_0(s_1) \prod_{h'=1}^{h-1} p_{h'}(s_{h'+1}|s_{h'}, a_{h'})$$

Value of policy $\mu_h(a|s)$

$$V^\mu = \sum_{s,a,h} p^\mu_{1:h}(s,a) r_h(s,a)$$
$$= \sum_{s,a,h} \mu_{1:h}(s,a) p^\emptyset_h(s) r_h(s,a)$$

Reach probability (because of tree structure)

$$p^\mu_{1:h}(s,a) = \underbrace{\mu_{1:h}(s,a)}_{\text{agent}} \underbrace{p^\emptyset_h(s)}_{\text{nature}}$$

where along trajectory $s_1, a_1, \ldots, s_h = s, a_h = a$, realization plan of agent/nature

$$\mu_{1:h}(s,a) = \prod_{h'=1}^{h} \mu_{h'}(a_{h'}|s_{h'})$$
$$p^\emptyset_h(s) = p_0(s_1) \prod_{h'=1}^{h-1} p_{h'}(s_{h'+1}|s_{h'}, a_{h'})$$

Realization plan of policy $\mu_h(a|s)$ with trajectory $s_1, a_1, \ldots, s_h = s, a_h = a$,

$$\mu_{1:h}(s, a) = \prod_{h'=1}^{h} \mu_{h'}(a_{h'} | s_{h'})$$

From realization plan to policy $\mu_{1:h}(s, a)$

$$\mu_h(a|s) = \frac{\mu_{1:h}(s, a)}{\sum_b \mu_{1:h}(s, b)}$$

Learning in adversarial MDP

Learning procedure: For episode $t \in [T]$:

- Nature chooses transitions and rewards (r_h^t, p_h^t)
- Agent chooses a policy $(\mu_h^t)_{h \in [H]}$
- For $h \in [H]$:
  - observe the current state s_h^t
  - take action $a_h^t \sim \mu_h^t(\cdot|s_h^t)$
  - get reward $r_h^t = r_h^t(s_h^t, a_h^t)$
  - next state $s_{h+1}^t \sim p_h^t(\cdot|s_h^t, a_h^t)$

$\rightarrow$ trajectory (bandit) feedback: $(s_1^t, a_1^t, r_1^t, \dots, s_h^t, a_H^t, r_H^t)$

Regret

$$\mathfrak{R}^T = \max_{\mu} \sum_{t=1}^{T} V^{t,\mu} - V^{t,\mu^t}$$

Conversion to online linear regret minimization

Value (tree structure)

$$V^{t,\mu} = \sum_{s,a,h} \mu_{1:h}(s,a) \underbrace{p^{\emptyset}_{1:h}(s) r_h(s,a)}_{:= g^t_h(s,a)} = \langle \mu, g^t \rangle$$

Loss instead of gain Why?→convenient for analysis

$$\ell^t_h(s,a) = p^{\emptyset,t}_{1:h}(s)\big(1 - r^t_h(s,a)\big)$$

Online "bandit" linear optimization $\mathfrak{R}^T = \max_\mu \sum_{t=1}^T \langle \mu^t - \mu, \ell^t \rangle$

→**Online Mirror Descent (OMD)**

$$\mu^{t+1} = \operatorname*{argmin}_\mu \eta \langle \mu, \widehat{\ell}^{\,t} \rangle + D(\mu, \mu^t)$$

→→ Which estimate $\widehat{\ell}^{\,t}$?
→→ Which regularizer $D(\cdot,\cdot)$?

Putting $+\gamma$ into denominator…

Regularised graph cuts (2008)

$$\ell_u = (L_{uu} + \gamma_g I)^{-1} W_{ul} \ell_l$$

Kernel bandits (2013)

$$\hat{\mu}_{a,t} = k^{\mathsf{T}}_{x_{a,t},t}(K_t + \gamma I)^{-1} y_t$$

Implicit exploration (2014)

$$\hat{\ell}_{t,i} = \frac{\ell_{t,i}}{o_{t,i} + \gamma_t} \mathbb{1}_{\{(I_t \to i) \in G_t\}}$$

Spectral bandits (2014)

$$\log \frac{|\mathbf{V}_T|}{|\mathbf{\Lambda}|} \leq \max \sum_{i=1}^{N} \log\left(1 + \frac{t_i}{\lambda_i}\right)$$

Ridge leverage scores (2016)

$$\tau_{i,n}(\gamma) = \sum_{j=1}^{n} \frac{\lambda_j}{\lambda_j + \gamma} [\mathbf{U}]^2_{i,j}$$

Implicit Exploration Online Mirror Descent (IXOMD)

IXOMD algorithm $\mu^{t+1} = \mathrm{argmin}_\mu \, \eta\langle \mu, \widehat{\ell}^{\,t}\rangle + D(\mu, \mu^t)$

Loss estimation $\to$ in expectation bound

$$\widehat{\ell}^{\,t}_h(s,a) = \frac{\mathbb{1}_{\{s=s^t_h,\, a=a^t_h\}}}{\mu^t_{1:h}(s,a)}(1 - r^t_h)$$

Implicit Exploration Online Mirror Descent (`IXOMD`)

`IXOMD` algorithm $\mu^{t+1} = \mathrm{argmin}_{\mu}\, \eta\langle \mu, \widehat{\ell}^{\,t}\rangle + D(\mu, \mu^t)$

Loss estimation

$$\mathbb{E}_t \widehat{\ell}_h^{t}(s,a) = \frac{\mu_{1:h}^{t}(s,a)\, p_{1:h}^{\emptyset,t}(s)}{\mu_{1:h}^{t}(s,a)}\left(1 - r_h^t(s,a)\right)$$

Implicit Exploration Online Mirror Descent (IXOMD)

IXOMD algorithm $\mu^{t+1} = \operatorname{argmin}_{\mu} \eta\langle \mu, \widehat{\ell}^{\,t}\rangle + D(\mu, \mu^t)$

Loss estimation with implicit exploration$\rightarrow$ high-probability bound [Kocák et al., 2014, Neu, 2015]

$$\widehat{\ell}^{\,t}(s,a) = \frac{\mathbb{1}_{\{s=s^t, a=a^t\}}}{\mu^t_{1:h}(s,a)+\gamma}(1-r^t_h)$$

Implicit Exploration Online Mirror Descent (IXOMD)

IXOMD algorithm $\mu^{t+1} = \operatorname{argmin}_\mu \eta\langle\mu, \widehat{\ell}^{\,t}\rangle + D(\mu, \mu^t)$

Loss estimation with implicit exploration→ high-probability bound [Kocák et al., 2014, Neu, 2015]

$$\widehat{\ell}^{\,t}(s,a) = \frac{\mathbb{1}_{\{s=s^t, a=a^t\}}}{\mu^t_{1:h}(s,a)+\gamma}(1-r^t_h)$$

Dilated entropy→efficient implementation [Kroer et al., 2015]

$$D(\mu,\mu') = \sum_{s,a,h} \mu_{1:h}(s,a)\log\frac{\mu_h(a|s)}{\mu'_h(a|s)}$$

Implicit Exploration Online Mirror Descent (IXOMD)

IXOMD algorithm $\mu^{t+1} = \operatorname{argmin}_\mu \eta\langle \mu, \widehat{\ell}^{\,t}\rangle + D(\mu, \mu^t)$

Loss estimation with implicit exploration→ high-probability bound [Kocák et al., 2014, Neu, 2015]

$$\widehat{\ell}^{\,t}(s,a) = \frac{\mathbb{1}_{\{s=s^t, a=a^t\}}}{\mu^t_{1:h}(s,a)+\gamma}(1-r^t_h)$$

Dilated entropy→efficient implementation [Kroer et al., 2015]

$$D(\mu,\mu') = \sum_{s,a,h} \mu_{1:h}(s,a)\log\frac{\mu_h(a|s)}{\mu'_h(a|s)}$$

→ $\mu_h(\cdot,\cdot)$ is not a probability distribution but ...consider the random vector $f \in \{0,1\}^{SAH}$

$$f_h(s,a) = \prod_{h'=1}^{h} f_{h'}(a_{h'}|s_{h'}) \text{ where } f_h(a|s) = \mathbb{1}_{\{a=\tilde{a}\}} \text{ with } \tilde{a}\sim\mu_h(\cdot|s)$$

then $\mathbb{E}_\mu[f] = \mu$ and

$$D(\mu,\mu') = \mathrm{KL}(\mathbb{P}^f_\mu, \mathbb{P}^f_{\mu'})\,.$$

IXOMD: efficient implementation

Algorithm IXOMD

Initialize $\mu_h^1(a|s) \leftarrow 1/A$
for $t = 1, \dots, T$ **do**
  for $h = 1, \dots, H$ **do**
    Observe s_h^t execute $a_h^t \sim \mu_h^t(\cdot|s_h^t)$ receive r_h^t
  end for
  for $h = H, \dots, 1$ **do**
    Construct loss estimate $\widehat{\ell}_h^t$
    For each $h \in [H]$ (with $Z_{H+1}^t \leftarrow 0$)

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

    Update only along trajectory

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

  end for
end for

Time complexity $\mathcal{O}(HA)$ per episode and space complexity $\mathcal{O}(AS)$

`IXOMD`: performance guarantee

Theorem

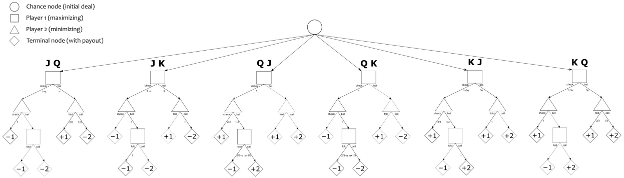
For $\eta = \widetilde{O}\left(\sqrt{1/HAT}\right)$ *and* $\gamma = \widetilde{O}\left(\sqrt{1/TA}\right)$, *for* `IXOMD` *algorithm, with probability at least* $1-\delta$,

$$\mathfrak{R}^T \leqslant \widetilde{\mathcal{O}}(S\sqrt{HAT})$$

Lower bound at least $\Omega(\sqrt{HAST})$ (conjecture from stochastic MDP lower bound).

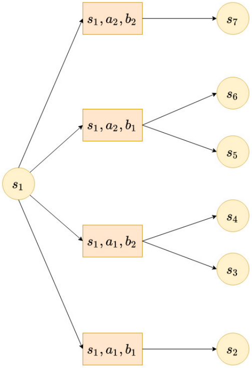
Two-player zero-Sum Imperfect Information Game (IIG)

Example (Kuhn) Poker



Two-player zero-sum imperfect information game (IIG)

- state space $\mathcal{S}$ of size S *and* horizon H
- max-player information set space $\mathcal{X}$ (size X), action space $\mathcal{A}$ (size A)
- min-player information set space $\mathcal{Y}$ (size Y), action space $\mathcal{B}$ (size B)
- reward (of the max-player)/loss (of the min-player) $r_h(s, a, b)$ and transitions $p_h(\cdot|s, a, b)$



$\mathcal{X}$ and $\mathcal{Y}$ partitions of $\mathcal{S}$, note $s \in x(s)$, $s \in y(s)$

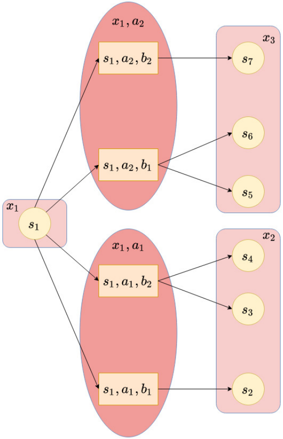
Assumptions:

Tree structure: one sub-trajectory per state

Perfect recall: players do not forget actions or observations

Two-player zero-sum imperfect information game (IIG)

- state space $\mathcal{S}$ of size S *and* horizon H
- max-player information set space $\mathcal{X}$ (size X), action space $\mathcal{A}$ (size A)
- min-player information set space $\mathcal{Y}$ (size Y), action space $\mathcal{B}$ (size B)
- reward (of the max-player)/loss (of the min-player) $r_h(s, a, b)$ and transitions $p_h(\cdot|s, a, b)$



$\mathcal{X}$ and $\mathcal{Y}$ partitions of $\mathcal{S}$, note $s \in x(s)$, $s \in y(s)$

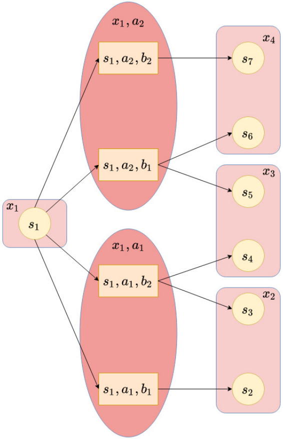
Assumptions:

Tree structure: one sub-trajectory per state

Perfect recall: players do not forget actions or observations

Two-player zero-sum imperfect information game (IIG)

- state space $\mathcal{S}$ of size S *and* horizon H
- max-player information set space $\mathcal{X}$ (size X), action space $\mathcal{A}$ (size A)
- min-player information set space $\mathcal{Y}$ (size Y), action space $\mathcal{B}$ (size B)
- reward (of the max-player)/loss (of the min-player) $r_h(s, a, b)$ and transitions $p_h(\cdot|s, a, b)$



$\mathcal{X}$ and $\mathcal{Y}$ partitions of $\mathcal{S}$, note $s \in x(s)$, $s \in y(s)$

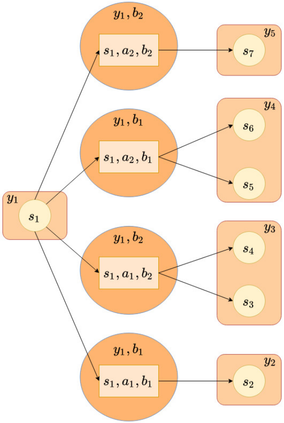
Assumptions:

Tree structure: one sub-trajectory per state

Perfect recall: players do not forget actions or observations

Two-player zero-sum imperfect information game (IIG)

- state space $\mathcal{S}$ of size S *and* horizon H
- max-player information set space $\mathcal{X}$ (size X), action space $\mathcal{A}$ (size A)
- min-player information set space $\mathcal{Y}$ (size Y), action space $\mathcal{B}$ (size B)
- reward (of the max-player)/loss (of the min-player) $r_h(s, a, b)$ and transitions $p_h(\cdot|s, a, b)$



$\mathcal{X}$ and $\mathcal{Y}$ partitions of $\mathcal{S}$, note $s \in x(s)$, $s \in y(s)$

Assumptions:

Tree structure: one sub-trajectory per state

Perfect recall: players do not forget actions or observations

Nash Equilibrium

Value of profile (μ, ν) [von Stengel, 1996]

$$V^{\mu,\nu} = \sum_{s,a,b,h} p_{1:h}^{\mu,\nu}(s,a,b) r_h(s,a,b)$$

$$= \sum_{x,a,y,b,h} \underbrace{\mu_{1:h}(x,a)}_{\text{max-player}} \underbrace{\nu_{1:h}(y,b)}_{\text{min-player}} \left(\underbrace{\sum_{s \in x \cap y} p_{1:h}^{\emptyset}(s) r_h(s,a,b)}_{\text{nature}} \right)$$

Nash Equilibrium $V^\star = \max_\mu \min_\nu V^{\mu,\nu} = \min_\nu \max_\mu V^{\mu,\nu}$

Exploitability gap of profile (μ, ν)

$$\Delta(\mu,\nu) = \max_{\mu'} V^{\mu',\nu} - \min_{\nu'} V^{\mu,\nu'} \geqslant 0$$

$\rightarrow$ if $\Delta(\mu,\nu) = 0$ then $V^{\mu,\nu} = V^\star$ and (μ,ν) Nash equilibrium
$\rightarrow$ Goal: minimize $\Delta(\mu,\nu)$

Learning in Imperfect Information Games

Learning procedure: For episode $t \in [T]$:

- Max-player chooses a policy $(\mu_h^t)_{h\in[H]}$
- Min-player chooses a policy $(\nu_h^t)_{h\in[H]}$
- For $h \in [H]$:
  - Max/min-players observe information set x_h^t/y_h^t of the current state s_h^t
  - Max-player takes action $a_h^t \sim \mu_h^t(\cdot|x_h^t)$
  - Min-player takes action $b_h^t \sim \mu_h^t(\cdot|y_h^t)$
  - get reward/loss $r_h^t = r_h(s_h^t, a_h^t, b_h^t)$
  - next state $s_{h+1}^t \sim p_h(\cdot|s_h^t, a_h^t, b_h^t)$
- output a profile μ^T, ν^T

$\rightarrow$ bandit feedback: trajectory $(x_1^t, a_1^t, y_1^t, b_1^t, r_1^t, \ldots, x_H^t, a_H^t, y_H^t, b_H^t, r_H^t)$

$\rightarrow$ Loss $\Delta(\mu^T, \nu^T)$

From regret minimization to solving games

Consider a sequence of profile $(\mu^t, \nu^t)_{t\in[T]}$

Regrets of max-player and min-player

$$\mathfrak{R}^T_{\max} = \max_{\mu} \sum_{t=1}^{T} V^{\mu,\nu^t} - V^{\mu^t,\nu^t} \qquad \mathfrak{R}^T_{\min} = \max_{\nu} \sum_{t=1}^{T} V^{\mu^t,\nu^t} - V^{\mu^t,\nu}$$

Average policy $(\bar{\mu}^T, \bar{\nu}^T)$ of the sequence $(\mu^t, \nu^t)_{t\in[T]}$, where

$$\bar{\mu}^T_h(a|x) = \frac{\sum_{t=1}^T \mu^t_{1:h}(x,a)}{\sum_{a'} \sum_{t=1}^T \mu^t_{1:h}(x,a')} \qquad \bar{\nu}^T_{1:h}(b|y) = \frac{\sum_{t=1}^T \nu^t_{1:h}(y,b)}{\sum_{b'} \sum_{t=1}^T \nu^t_{1:h}(y,b')}$$

From regret minimization to solving games

Consider a sequence of profile $(\mu^t, \nu^t)_{t\in[T]}$

Regrets of max-player and min-player

$$\mathfrak{R}^T_{\max} = \max_{\mu} \sum_{t=1}^{T} V^{\mu,\nu^t} - V^{\mu^t,\nu^t} \qquad \mathfrak{R}^T_{\min} = \max_{\nu} \sum_{t=1}^{T} V^{\mu^t,\nu^t} - V^{\mu^t,\nu}$$

Average policy $(\bar{\mu}^T, \bar{\nu}^T)$ of the sequence $(\mu^t, \nu^t)_{t\in[T]}$, where

$$\bar{\mu}^T_{1:h}(x,a) = \frac{1}{T}\sum_{t=1}^{T} \mu^t_{1:h}(x,a) \qquad \bar{\nu}^T_{1:h}(y,b) = \frac{1}{T}\sum_{t=1}^{T} \nu^t_{1:h}(y,b)$$

From regret minimization to solving games

Consider a sequence of profile $(\mu^t, \nu^t)_{t\in[T]}$

Regrets of max-player and min-player

$$\mathfrak{R}^T_{\max} = \max_{\mu} \sum_{t=1}^{T} V^{\mu,\nu^t} - V^{\mu^t,\nu^t} \qquad \mathfrak{R}^T_{\min} = \max_{\nu} \sum_{t=1}^{T} V^{\mu^t,\nu^t} - V^{\mu^t,\nu}$$

Average policy $(\bar\mu^T, \bar\nu^T)$ of the sequence $(\mu^t, \nu^t)_{t\in[T]}$, where

$$\bar\mu^T_{1:h}(x,a) = \frac{1}{T}\sum_{t=1}^{T} \mu^t_{1:h}(x,a) \qquad \bar\nu^T_{1:h}(y,b) = \frac{1}{T}\sum_{t=1}^{T} \nu^t_{1:h}(y,b)$$

Regret to exploitability gap

$$\Delta(\bar\mu^T, \bar\nu^T) \leqslant \frac{\mathfrak{R}^T_{\max} + \mathfrak{R}^T_{\min}}{T}$$

IXOMD for IIG

Idea Use IXOMD to minimize the regrets $\mathfrak{R}^T_{\mathrm{max}}$ and $\mathfrak{R}^T_{\mathrm{min}}$ and output $(\bar{\mu}^T, \bar{\nu}^T)$

From the point of view of the max player

$$V^{\mu,\nu^t} = \sum_{x,a} \mu_{1:h}(x,a) \underbrace{\left(\sum_{y,b} \nu_{1:h}(y,b) \sum_{s\in x\cap y} p^{\emptyset}_{1:h}(s) r_h(s,a,b) \right)}_{\text{nature for max-player}}$$

IXOMD for IIG

Idea Use IXOMD to minimize the regrets $\mathfrak{R}^T_{\max}$ and $\mathfrak{R}^T_{\min}$ and output $(\bar{\mu}^T, \bar{\nu}^T)$

From the point of view of the max player

$$V^{\mu,\nu^t} = \sum_{x,a} \mu_{1:h}(x,a) \underbrace{\left(\sum_{y,b} \nu_{1:h}(y,b) \sum_{s \in x \cap y} p^{\emptyset}_{1:h}(s) r_h(s,a,b) \right)}_{\text{nature for max-player}}$$

Theorem

If max-player and min-player run IXOMD for T games then with probability at least $1-\delta$

$$\Delta(\bar{\mu}^T, \bar{\nu}^T) \leqslant \widetilde{\mathcal{O}}\left(\sqrt{\frac{H}{T}} (\sqrt{A}X + \sqrt{B}Y) \right)$$

$\texttt{IXOMD}$ for IIG

Idea Use $\texttt{IXOMD}$ to minimize the regrets $\mathfrak{R}^T_{\max}$ and $\mathfrak{R}^T_{\min}$ and output $(\bar{\mu}^T, \bar{\nu}^T)$

From the point of view of the max player

$$V^{\mu,\nu^t} = \sum_{x,a} \mu_{1:h}(x,a) \underbrace{\left(\sum_{y,b} \nu_{1:h}(y,b) \sum_{s \in x \cap y} p^{\emptyset}_{1:h}(s) r_h(s,a,b) \right)}_{\text{nature for max-player}}$$

Theorem

If max-player and min-player run $\texttt{IXOMD}$ for T games then with probability at least $1-\delta$

$$\Delta(\bar{\mu}^T, \bar{\nu}^T) \leqslant \widetilde{\mathcal{O}}\left(\sqrt{\frac{H}{T}} (\sqrt{A}X + \sqrt{B}Y) \right)$$

$\rightarrow$ Model-free algorithm for trajectory bandit feedback

$\rightarrow$ Time complexity?

IXOMD for IIG

Idea Use IXOMD to minimize the regrets $\mathfrak{R}^T_{\max}$ and $\mathfrak{R}^T_{\min}$ and output $(\bar{\mu}^T, \bar{\nu}^T)$

From the point of view of the max player

$$V^{\mu,\nu^t} = \sum_{x,a} \mu_{1:h}(x,a) \underbrace{\left(\sum_{y,b} \nu_{1:h}(y,b) \sum_{s \in x \cap y} p^{\emptyset}_{1:h}(s) r_h(s,a,b) \right)}_{\text{nature for max-player}}$$

Theorem

If max-player and min-player run IXOMD for T games then with probability at least $1-\delta$

$$\Delta(\bar{\mu}^T, \bar{\nu}^T) \leqslant \widetilde{\mathcal{O}}\left(\sqrt{\frac{H}{T}} (\sqrt{A}X + \sqrt{B}Y) \right)$$

→ Model-free algorithm for trajectory bandit feedback

→ Time complexity?

How to compute average policy $(\bar{\mu}^T, \bar{\nu}^T)$?

Computing the average policy

Idea Use that the **policy does not change outside of the trajectory** between two updates

Algorithm Compute average policy max-player

for $t = 1, \dots, T$ **do**
  $\mathring{\mu}^t_{1:0} \leftarrow t$
  for $h = 1, \dots, H$ **do**
    Observe x^t_h execute $a^t_h \sim \mu^t_h(\cdot|x^t_h)$ receive r^t_h
    $\tau \leftarrow \tau^t(x^t_h)$ and
    $\mathring{\mu}^t_{1:h}(x^t_h, a) \leftarrow \mathring{\mu}^\tau_{1:h}(x^t_h, a) + \left(\mathring{\mu}^t_{1:h-1} - \sum_b \mathring{\mu}^\tau_{1:h}(x^t_h, b)\right)\mu^t_h(a|x^t_h)$
    $\tau^t(x^t_h) \leftarrow t$ and $\mathring{\mu}^t_{1:h-1} \leftarrow \mathring{\mu}^t_{1:h-1}(x^t_h, a^t_h)$
  end for
end for
for h, x, a **do**
  $\tau \leftarrow \tau^t(x_h)$
  $\mathring{\mu}^T \leftarrow \mathring{\mu}^T_{1:h-1}(x_{h-1}, a_{h-1})$ with x_{h-1}, a_{h-1} predecessor of x,
  $\mathring{\mu}^T_{1:h}(x, a) \leftarrow \mathring{\mu}^\tau_{1:h}(x_h, a) + \left(\mathring{\mu}^T - \sum_b \mathring{\mu}^\tau_{1:h}(x, b)\right)\mu^t_h(a|x^T)$
end for
Return policy

$$\bar{\mu}^T_h(a|x) = \frac{\mathring{\mu}^T_{1:h}(x, a)}{\sum_b \mathring{\mu}^T_{1:h}(x, b)}$$

Computing the average policy

Idea Use that the **policy does not change outside of the trajectory** between two updates

Algorithm Compute average policy max-player

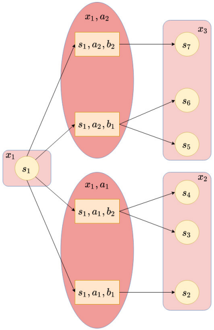
for $t = 1, \ldots, T$ **do**
  $\mathring{\mu}^t_{1:0} \leftarrow t$
  for $h = 1, \ldots, H$ **do**
    Observe x^t_h execute $a^t_h \sim \mu^t_h(\cdot|x^t_h)$ receive r^t_h
    $\tau \leftarrow \tau^t(x^t_h)$ and
    $\mathring{\mu}^t_{1:h}(x^t_h, a) \leftarrow \mathring{\mu}^\tau_{1:h}(x^t_h, a) + \left(\mathring{\mu}^t_{1:h-1} - \sum_b \mathring{\mu}^\tau_{1:h}(x^t_h, b)\right)\mu^t_h(a|x^t_h)$
    $\tau^t(x^t_h) \leftarrow t$ and $\mathring{\mu}^t_{1:h-1} \leftarrow \mathring{\mu}^t_{1:h-1}(x^t_h, a^t_h)$
  end for
end for
for h, x, a **do**
  $\tau \leftarrow \tau^t(x_h)$
  $\mathring{\mu}^T \leftarrow \mathring{\mu}^T_{1:h-1}(x_{h-1}, a_{h-1})$ with x_{h-1}, a_{h-1} predecessor of x,
  $\mathring{\mu}^T_{1:h}(x, a) \leftarrow \mathring{\mu}^\tau_{1:h}(x_h, a) + \left(\mathring{\mu}^T - \sum_b \mathring{\mu}^\tau_{1:h}(x, b)\right)\mu^t_h(a|x^T)$
end for
Return policy

$$\bar{\mu}^T_h(a|x) = \frac{\mathring{\mu}^T_{1:h}(x, a)}{\sum_b \mathring{\mu}^T_{1:h}(x, b)}$$

$\rightarrow$ Time complexity $\mathcal{O}\big(H(A+B)T + XA + YB\big)$ space complexity, $\mathcal{O}\big(XA + YB\big)$

Comparison

Algorithm	Rate	Time complexity	Game
IXOMD **[this work]**	$\widetilde{\mathcal{O}}\left(\sqrt{\frac{H}{T}}\left(\sqrt{A}X+\sqrt{B}Y\right)\right)$	$\mathcal{O}(H(A+B)T+AX+BY)$	**Unknown**
CFR [Zinkevich et al., 2007]		$\mathcal{O}((A+B)ST)$	Known
MCCFR [Lanctot et al., 2009, Farina et al., 2020]		$\mathcal{O}((AX+BY)T)$	Known



Conclusion

Algorithm	Convergence	Time complexity	Space complexity
IXOMD for (tree) MDPs	$\widetilde{\mathcal{O}}(S\sqrt{HAT})$ regret	$\mathcal{O}(HAT)$	$\mathcal{O}(SA)$
IXOMD for games	$\widetilde{\mathcal{O}}\big(\sqrt{H/T}(\sqrt{A}X+\sqrt{B}Y)\big)$ rate	$\mathcal{O}(H(A+B)T+AX+BY)$	$\mathcal{O}(AX+BY)$

Open questions

- Optimal regret bound for adversarial tree MDP?
- Optimal rate for IIG with bandit feedback?
  - If game known + first order feedback $\mathcal{O}(\text{poly}(H,X,A,Y,B)/T)$.
- Structure on the game (linear/deep approximation)?
- Last iterate?

Bibliography I

Brown, G. W. (1949). Some notes on computation of games solutions.
Technical report, RAND CORP SANTA MONICA CA.

Brown, N. and Sandholm, T. (2015).
Regret-based pruning in extensive-form games.
In NIPS, pages 1972–1980.

Burch, N., Moravcik, M., and Schmid, M. (2019).
Revisiting CFR+ and Alternating Updates.
Journal of Artificial Intelligence Research, 64:429–443.

Farina, G., Kroer, C., and Sandholm, T. (2019).
Optimistic Regret Minimization for Extensive-Form Games via Dilated Distance-Generating Functions.
In Advances in Neural Information Processing Systems.

Farina, G., Kroer, C., and Sandholm, T. (2020).
Stochastic Regret Minimization in Extensive-Form Games.
In International Conference on Machine Learning.

Farina, G., Kroer, C., and Sandholm, T. (2021a).
Bandit Linear Optimization for Sequential Decision Making and Extensive-Form Games.
In AAAI Conference on Artificial Intelligence.

Bibliography II

Farina, G., Kroer, C., and Sandholm, T. (2021b).
Faster Game Solving via Predictive Blackwell Approachability: Connecting Regret Matching and Mirror Descent.
In AAAI Conference on Artificial Intelligence.

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

Hart, S. and Mas-Colell, A. (2000).
A Simple Adaptive Procedure Leading to Correlated Equilibrium.
Econometrica, 68(5):1127–1150.

Heinrich, J., Lanctot, M., and Silver, D. (2015).
Fictitious self-play in extensive-form games.
In International conference on machine learning, pages 805–813. PMLR.

Kocák, T., Neu, G., Valko, M., and Munos, R. (2014).
Efficient learning by implicit exploration in bandit problems with side observations.
In Advances in Neural Information Processing Systems.

Koller, D., Megiddo, N., and Von Stengel, B. (1996).
Efficient Computation of Equilibria for Extensive Two-Person Games.
Games and Economic Behavior, 14(2):247–259.

Bibliography III

Kroer, C., Farina, G., and Sandholm, T. (2018).
Solving Large Sequential Games with the Excessive Gap Technique.
In Advances in Neural Information Processing Systems.

Kroer, C., Waugh, K., Kilinç-Karzan, F., and Sandholm, T. (2015).
Faster First-Order Methods for Extensive-Form Game Solving .
In ACM Conference on Economics and Computation, pages 817–834.

Kroer, C., Waugh, K., Kılınç-Karzan, F., and Sandholm, T. (2020).
Faster algorithms for extensive-form game solving via improved smoothing functions.
Mathematical Programming, 179(1):385–417.

Kuhn, H. W. (1953).
Extensive Games and the Problem of Information.
Annals of Mathematics Studies, 28:193–216.

Kuhn, H. W. (2016).
11. extensive games and the problem of information.
In Contributions to the Theory of Games (AM-28), Volume II, pages 193–216. Princeton University Press.

Lanctot, M., Waugh, K., Zinkevich, M., and Bowling, M. (2009).
Monte Carlo Sampling for Regret Minimization in Extensive Games.
In Advances in Neural Information Processing Systems.

Bibliography IV

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Perolat, J., Silver, D., and Graepel, T. (2017).
A unified game-theoretic approach to multiagent reinforcement learning.
In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

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

Munos, R., Perolat, J., Lespiau, J.-B., Rowland, M., De Vylder, B., Lanctot, M., Timbers, F., Hennes, D., Omidshafiei, S., Gruslys, A., et al. (2020).
Fast Computation of Nash Equilibria in Imperfect Information Games.
In International Conference on Machine Learning, pages 7119–7129.

Nash Jr, J. F. (1950).
Equilibrium points in N-person games.
Proceedings of the National Academy of Sciences of the United States of America, 36(1):48–49.

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

Osborne, M. J. and Rubinstein, A. (1994).
A Course in Game Theory.
The MIT Press.

Bibliography V

Ponsen, M., De Jong, S., and Lanctot, M. (2011).
Computing Approximate Nash Equilibria and Robust Best-Responses Using Sampling.
Journal of Artificial Intelligence Research, 42:575–605.

Romanovsky, J. V. (1962).
Reduction of a game with complete memory to a matricial game.
Dokl. Akad. Nauk SSSR, 144:62–64.

Shapley, L. S. (1953).
Stochastic Games.
Proceedings of the National Academy of Sciences of the United States, 39(10):1095–1100.

Tammelin, O. (2014).
Solving Large Imperfect Information Games Using CFR+.
arXiv preprint arXiv:1407.5042.

von Stengel, B. (1996).
Efficient Computation of Behavior Strategies.
Games and Economic Behavior, 14(2):220–246.

Zhang, B. H. and Sandholm, T. (2021).
Finding and Certifying (Near-) Optimal Strategies in Black-Box Extensive-Form Games.
In AAAI Conference on Artificial Intelligence.

Bibliography VI

Zinkevich, M., Johanson, M., Bowling, M., and Piccione, C. (2007).
Regret Minimization in Games with Incomplete Information.
Advances in neural information processing systems, 20:1729–1736.

Sketch of proof

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

BIAS of the IX estimator

$$\text{BIAS 1} \leqslant H\sqrt{2T\iota} + \gamma TXA \qquad\qquad \text{BIAS 2} \leqslant X\frac{\iota}{2\gamma}$$

Regret

$$\begin{aligned}\text{REGRET} &= \sum_{t=1}^{T}\frac{D(\mu^\dagger,\mu^t) - D(\mu^\dagger,\mu^{t+1})}{\eta} + \frac{1}{\eta}D(\mu^t\|\mu^{t+1})\\ &\leqslant \frac{X\log A}{\eta} + \eta(1+H)TXA + \frac{\eta(1+H)H\iota}{2\gamma}\end{aligned}$$