Maximin Action Identification: A New Bandit Framework for Games

Aurélien Garivier, Emilie Kaufmann and Wouter M. Koolen

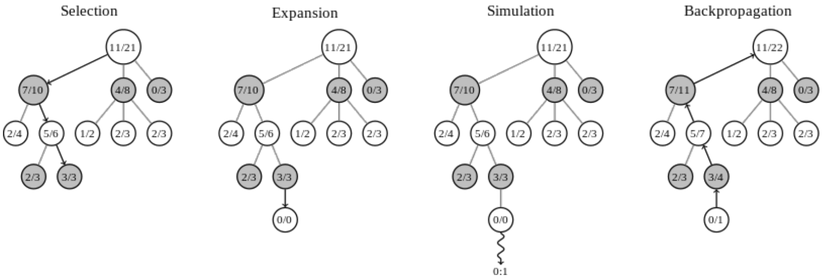
INSTITUT de MATHEMATIQUES de TOULOUSE

CRIStAL Centre de Recherche en Informatique, Signal et Automatique de Lille

CWI Centrum Wiskunde & Informatica

Conference On Learning Theory
Columbia University,
June 23rd, 2016

Monte-Carlo Tree Search for games



We introduce an idealized model:

- perfect rollouts
- depth-two complete tree

and propose new algorithms with sample complexity guarantees

Outline

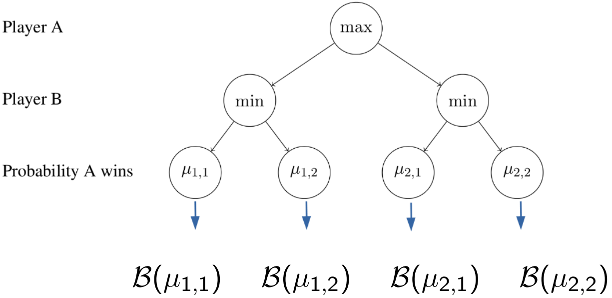
Outline

A PAC learning framework

Consider a two-player game in which

- when A chooses action $i \in \{1, \dots, K\}$
- and then player B choose action $j \in \{1, \dots, K_i\}$,

the probability that A wins is $\mu_{i,j}$.



Best action for A given that B is strategic:

$$i^* \in \underset{i\in\{1,\dots,K\}}{\text{argmax}} \ \min_{j\in\{1,\dots,K_i\}} \mu_{i,j} \quad \text{(maximin action)}$$

Maximin action identification

A bandit model parametrized by $\boldsymbol{\mu} = (\mu_{i,j})_{\substack{1 \leq i \leq K, \\ 1 \leq j \leq K_i}}$

with a different notion of best arm: $i^* = \arg\max_i \min_j \ \mu_{i,j}$

A strategy consists in

- a sampling rule P_t ➜ pair of actions (i,j) chosen at round t

  a rollout $X_t \sim \mathcal{B}(\mu_{P_t})$ is observed
- a stopping rule τ ➜ when did we see enough rollouts ?
- a recommendation rule $\hat{\imath}$ ➜ a guess for the maximin action

Goal: Build a strategy $(P_t, \tau, \hat{\imath})$ such that

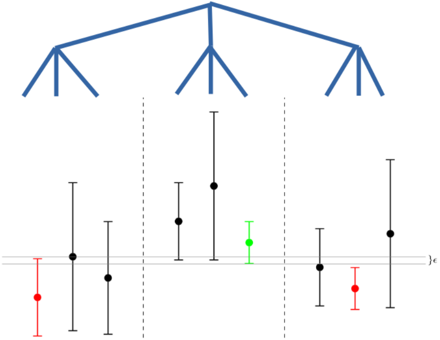
$$\forall \boldsymbol{\mu}, \ \mathbb{P}_{\boldsymbol{\mu}}\left(\min_j \mu_{i^*,j} - \min_j \mu_{\hat{\imath},j} \leq \epsilon\right) \geq 1 - \delta,$$

and $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ is as small as possible.

Outline

The Maximin-LUCB algorithm

$[\mathrm{LCB}_P(t), \mathrm{UCB}_P(t)]$ confidence interval on μ_P at time t



- Pick one representative per action $P_i = (i, j_i)$,

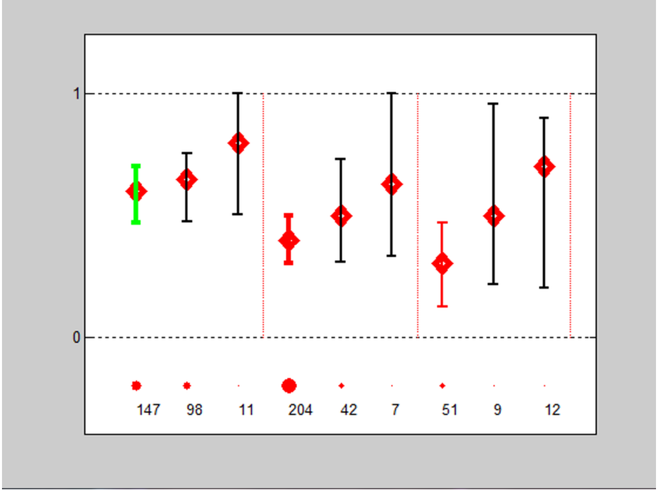
$$j_i = \mathrm{argmin}_j \mathrm{LCB}_{(i,j)}(t)$$

- (BAI step) Letting $\hat{\imath}(t) = \arg\max_i \min_j \hat{\mu}_{(i,j)}(t)$, draw

$$L_t = (\hat{\imath}(t), j_{\hat{\imath}(t)}) \quad \text{and} \quad C_t = \underset{P \in \{(i,j_i)\}_{i \neq \hat{\imath}(t)}}{\arg\max} \mathrm{UCB}_P(t)$$

- Stop if $\mathrm{LCB}_{L_t}(t) > \mathrm{UCB}_{C_t}(t) - \epsilon$

M-LUCB in action !



Sample complexity analysis

$$\mathrm{LCB}_P(t) = \hat{\mu}_P(t) - \sqrt{\frac{\beta(t,\delta)}{2N_P(t)}}, \quad \mathrm{UCB}_P(t) = \hat{\mu}_P(t) + \sqrt{\frac{\beta(t,\delta)}{2N_P(t)}}$$

Theorem

$\epsilon = 0$. Let $\alpha > 1$. There exists $C > 0$ such that for the choice

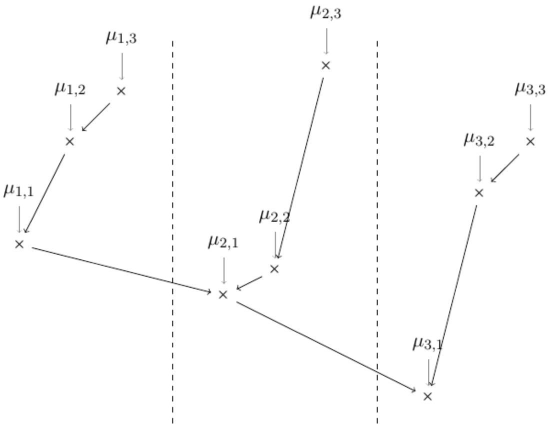
$$\beta(t,\delta) = \log(Ct^{1+\alpha}/\delta),$$

M-LUCB is δ-PAC and

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\log(1/\delta)} \leq 8(1+\alpha)H^*(\boldsymbol{\mu})$$

$$H^*(\boldsymbol{\mu}) := \sum_{(1,j)\in\mathcal{P}_1} \frac{1}{(\mu_{1,j} - \mu_{2,1})^2} + \sum_{(i,j)\in\mathcal{P}\backslash\mathcal{P}_1} \frac{1}{(\mu_{1,1} - \mu_{i,1})^2 \vee (\mu_{i,j} - \mu_{i,1})^2}.$$

The complexity term



$$H^*(\boldsymbol{\mu}) := \sum_{(1,j)\in\mathcal{P}_1} \frac{1}{(\mu_{1,j}-\mu_{2,1})^2} + \sum_{(i,j)\in\mathcal{P}\setminus\mathcal{P}_1} \frac{1}{(\mu_{1,1}-\mu_{i,1})^2 \vee (\mu_{i,j}-\mu_{i,1})^2}.$$

Outline

The M-Racing algorithm

$$I(x,y) := \left[\mathrm{kl}\left(x, \frac{x+y}{2}\right) + \mathrm{kl}\left(y, \frac{x+y}{2}\right)\right] \mathbb{1}_{(x \geq y)}$$

μ_P has statistical evidence to be larger that μ_Q at round r
$\Leftrightarrow \quad rI(\hat{\mu}_P(r), \hat{\mu}_Q(r)) > \log(Ct^2/\delta)$, written $\mu_P \gg_r \mu_Q$

M-Racing samples at each round r a set of active arms, and possibly removes arms from it in two possible ways:

- High arms elimination: eliminate (i,j) if
  $\exists j' : \mu_{(i,j)} \gg_r \mu_{(i,j')}$
- Action elimination: eliminate $(\tilde{\imath}, \tilde{\jmath}) = \arg\min_{P \in \mathcal{R}} \ \hat{\mu}_P(r)$, together with $(\tilde{\imath}, j)$ for all j if
  $\exists i$: for all active (i,j), $\mu_{(i,j)} \gg_r \mu_{(\tilde{\imath},\tilde{\jmath})}$

➜ **Improved sample complexity guarantees**,
for $\epsilon > 0$, expressed with $I(\mu_P, \mu_Q) > (\mu_P - \mu_Q)^2$

Some numerical results

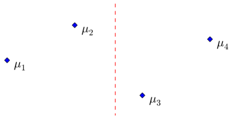
$$\boldsymbol{\mu} = \begin{bmatrix} 0.4 & 0.5 \\ 0.3 & 0.35 \end{bmatrix}$$

	$\mathbb{E}[\tau_{1,1}]$	$\mathbb{E}[\tau_{1,2}]$	$\mathbb{E}[\tau_{2,1}]$	$\mathbb{E}[\tau_{2,2}]$
M-LUCB	1762	198	1761	462
M-KL-LUCB	762	92	733	237
M-Chernoff	**315**	**59**	**291**	**136**
M-Racing	324	152	301	298
KL-LUCB	351	64	3074	2768

Outline

A lower bound revealing a surprising behavior

2 actions by player:



Theorem

Any δ-PAC algorithm satisfies

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq T^*(\boldsymbol{\mu}) \log(1/(2.4\delta)),$$

where

$$T_*^{-1}(\boldsymbol{\mu}) = \max_{w \in \Sigma_4} \inf_{\boldsymbol{\mu}' : \mu_1' \wedge \mu_2' < \mu_3' \wedge \mu_4'} \left(\sum_{a=1}^{4} w_a \, \mathrm{kl}(\mu_a, \mu_a') \right)$$

<u>Particular case:</u> if $\mu_4 > \mu_2$,

$$w^*(\boldsymbol{\mu}) = \underset{w \in \Sigma_4}{\operatorname{argmax}} \inf_{\boldsymbol{\mu}' : \mu_1' \wedge \mu_2' < \mu_3' \wedge \mu_4'} \left(\sum_{a=1}^{4} w_a \, \mathrm{kl}(\mu_a, \mu_a') \right)$$

can be computed and $w_4^*(\boldsymbol{\mu}) = 0$!

Conclusion and perspectives

For depth-two MCTS:

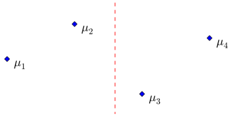
- we devise the first algorithms based BAI tools (rather than UCBs)...
- ... and provide the first sample complexity guarantees in a PAC learning framework

Future work:

- optimal strategies remain to be characterized
- ... we need to go deeper !
- fixed-budget setting

Lower bound and optimal algorithm ?

2 actions by player:



$$w^*(\boldsymbol{\mu}) = \underset{w \in \Sigma_4}{\text{argmax}} \inf_{\boldsymbol{\mu}' \in \text{Alt}(\boldsymbol{\mu})} \left(\sum_{a=1}^{4} w_a \, \text{kl}(\mu_a, \mu'_a) \right)$$

Assuming, in general, that $w^*(\boldsymbol{\mu})$ is unique and well-behaved, with

$$\hat{Z}(t) = \inf_{\boldsymbol{\mu}' \in \text{Alt}(\hat{\boldsymbol{\mu}}(t))} \sum_{a=1}^{4} N_a(t) \text{kl}(\hat{\mu}_a(t), \mu'_a),$$

a strategy such that $\frac{N_a(t)}{t} \to w_a^*(\boldsymbol{\mu})$ and

$$\tau = \inf\{t \in \mathbb{N} : \hat{Z}(t) \geq \log(Ct/\delta)\},$$

would satisfy $\tau_\delta \leq P^*(\boldsymbol{\mu}) \log(1/\delta) + o(\log(1/\delta))$, a.s.