A Pure Exploration perspective on Game Tree Search

Wouter M. Koolen

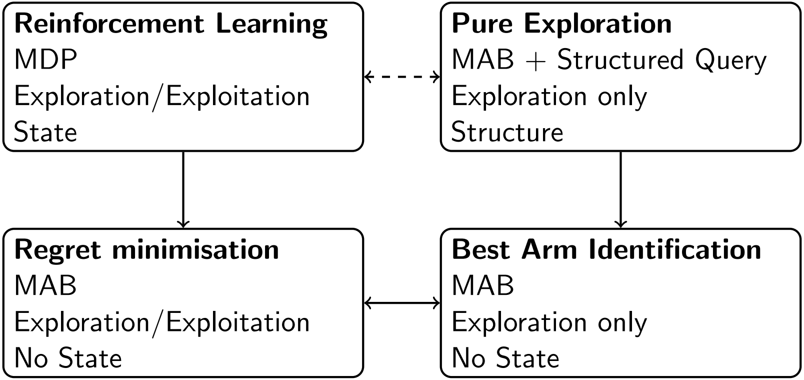
CWI
Centrum Wiskunde & Informatica

BeNeRL
Delft, Monday 19th November, 2018

Menu

- What is Pure Exploration?
- Relation to Reinforcement Learning?
- Why care?
  - Pure Exploration problems occur as **sub-problems** in RL!
  - Novel/interesting/powerful PE learning algorithms! "Fresh"
- My focus: Pure Exploration **Renaissance** (2016)
  1. Track-and-Stop algorithm for Best Arm Identification
  2. BAI-MCTS approach for Game Tree Search
  3. Murphy Sampling for Games Trees of "depth 1.5"

Relation of RL and PE



Example 1: Phased Q-Learning

[Even-Dar, Mannor, and Mansour, 2002]

Early example(s) of Pure Exploration as sub-module in RL

Initialise $V_0(s) := 0$ for each state s.
for phase $i = 1, 2, \ldots$ **do**
  for each state s **do**
    Run **Best Arm Identification** algorithm on

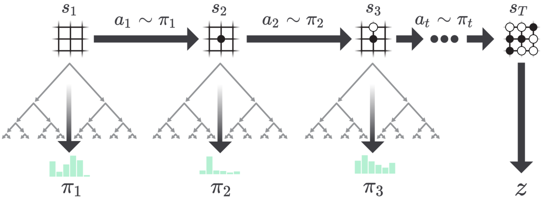
$$a \mapsto r + \gamma V_i(s') \quad \text{where} \quad (r, s') \sim \mathbb{P}\left(r, s' \middle| s, a\right)$$

    Store estimate in $V_{i+1}(s)$.
  end for
end for

Example 2: AlphaZero

MCTS as Policy/Value Improvement Operator

Randomise network parameters θ_1.
for iteration $i = 1, 2, \ldots$ **do**
    for training games $j = 1, 2, \ldots$ **do**



π_t result of **MCTS** from s_t with network θ_i

        Store $(s_1, \pi_1, z), \ldots, (s_T, \pi_T, z)$.
    end for
    Train network θ_{i+1} on stored data.
end for

Formal model

Environment (Multi-armed bandit model)

K distributions parameterised by their means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$.

The **best arm** is

$$i^* \;=\; \operatorname*{argmax}_{i \in [K]} \; \mu_i$$

Formal model

Environment (Multi-armed bandit model)

K distributions parameterised by their means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$.

The **best arm** is

$$i^* = \operatorname*{argmax}_{i \in [K]} \mu_i$$

Strategy

- **Stopping rule** $\tau \in \mathbb{N}$
- In round $t \leq \tau$ **sampling rule** picks $I_t \in [K]$. See $X_t \sim \mu_{I_t}$.
- **Recommendation rule** $\hat{I} \in [K]$.

Formal model

Environment (Multi-armed bandit model)

K distributions parameterised by their means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$.

The **best arm** is

$$i^* = \operatorname*{argmax}_{i \in [K]} \mu_i$$

Strategy

- **Stopping rule** $\tau \in \mathbb{N}$
- In round $t \leq \tau$ **sampling rule** picks $I_t \in [K]$. See $X_t \sim \mu_{I_t}$.
- **Recommendation rule** $\hat{I} \in [K]$.

Realisation of interaction: $(I_1, X_1), \ldots, (I_\tau, X_\tau), \hat{I}$.

Formal model

Environment (Multi-armed bandit model)

K distributions parameterised by their means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$.

The **best arm** is

$$i^* \;=\; \operatorname*{argmax}_{i \in [K]} \; \mu_i$$

Strategy

- **Stopping rule** $\tau \in \mathbb{N}$
- In round $t \le \tau$ **sampling rule** picks $I_t \in [K]$. See $X_t \sim \mu_{I_t}$.
- **Recommendation rule** $\hat{I} \in [K]$.

Realisation of interaction: $(I_1, X_1), \ldots, (I_\tau, X_\tau), \hat{I}$.

Two objectives: **sample efficiency** τ and **correctness** $\hat{I} = i^*$.

Objective

On bandit $\boldsymbol{\mu}$, strategy $(\tau, (I_t)_t, \hat{I})$ has

- **error probability** $\mathbb{P}_{\boldsymbol{\mu}}\big(\hat{I} \neq i^*(\boldsymbol{\mu})\big)$, and
- **sample complexity** $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$.

Idea: constrain one, optimise the other.

Objective

On bandit $\boldsymbol{\mu}$, strategy $(\tau, (I_t)_t, \hat{I})$ has

- **error probability** $\mathbb{P}_{\boldsymbol{\mu}}\big(\hat{I} \neq i^*(\boldsymbol{\mu})\big)$, and
- **sample complexity** $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$.

Idea: constrain one, optimise the other.

Definition

Fix small confidence $\delta \in (0,1)$. A strategy is δ-**correct** if

$$\mathbb{P}_{\boldsymbol{\mu}}\big(\hat{I} \neq i^*(\boldsymbol{\mu})\big) \;\leq\; \delta \qquad \text{for every bandit model } \boldsymbol{\mu}.$$

(Generalisation: output ϵ-best arm)

Objective

On bandit $\boldsymbol{\mu}$, strategy $(\tau, (I_t)_t, \hat{I})$ has

- **error probability** $\mathbb{P}_{\boldsymbol{\mu}}\big(\hat{I} \neq i^*(\boldsymbol{\mu})\big)$, and
- **sample complexity** $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$.

Idea: constrain one, optimise the other.

Definition

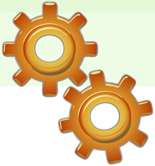
Fix small confidence $\delta \in (0, 1)$. A strategy is δ-**correct** if

$$\mathbb{P}_{\boldsymbol{\mu}}\big(\hat{I} \neq i^*(\boldsymbol{\mu})\big) \;\leq\; \delta \qquad \text{for every bandit model } \boldsymbol{\mu}.$$

(Generalisation: output ϵ-best arm)

Goal: minimise $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ over **all δ-correct strategies**.

Algorithms



- Sampling rule I_t?
- Stopping rule τ?
- Recommendation rule $\hat{I}$?

$$\hat{I} = \operatorname*{argmax}_{i \in [K]} \hat{\mu}_i(\tau)$$

where $\hat{\boldsymbol{\mu}}(t)$ is **empirical mean**.

Instance-Dependent Sample Complexity Lower bound

Define the **alternatives** to $\boldsymbol{\mu}$ by $\text{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} | i^*(\boldsymbol{\lambda}) \neq i^*(\boldsymbol{\mu})\}$.

Instance-Dependent Sample Complexity Lower bound

Define the **alternatives** to $\boldsymbol{\mu}$ by $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} | i^*(\boldsymbol{\lambda}) \neq i^*(\boldsymbol{\mu})\}$.

Theorem (Castro 2014, Garivier and Kaufmann 2016)

Fix a δ-correct strategy. Then for every bandit model $\boldsymbol{\mu}$

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \;\geq\; T^*(\boldsymbol{\mu}) \ln \frac{1}{\delta}$$

where the **characteristic time** $T^*(\boldsymbol{\mu})$ *is given by*

$$\frac{1}{T^*(\boldsymbol{\mu})} \;=\; \max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i \, \mathrm{KL}(\mu_i \| \lambda_i).$$

Instance-Dependent Sample Complexity Lower bound

Define the **alternatives** to $\boldsymbol{\mu}$ by $\mathrm{Alt}(\boldsymbol{\mu}) = \{\boldsymbol{\lambda} | i^*(\boldsymbol{\lambda}) \neq i^*(\boldsymbol{\mu})\}$.

Theorem (Castro 2014, Garivier and Kaufmann 2016)

Fix a δ-correct strategy. Then for every bandit model $\boldsymbol{\mu}$

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \;\geq\; T^*(\boldsymbol{\mu}) \ln \frac{1}{\delta}$$

where the **characteristic time** $T^*(\boldsymbol{\mu})$ *is given by*

$$\frac{1}{T^*(\boldsymbol{\mu})} \;=\; \max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i \,\mathrm{KL}(\mu_i \| \lambda_i).$$

Intuition (going back to Lai and Robbins [1985]): if observations are likely under both $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$, yet $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$, then learner cannot stop and be correct in both.

Example

$K = 5$ arms, Bernoulli $\boldsymbol{\mu} = (0, 0.1, 0.2, 0.3, 0.4)$.

$$T^*(\boldsymbol{\mu}) \;=\; 200.4 \qquad \boldsymbol{w}^*(\boldsymbol{\mu}) \;=\; (0.45, 0.46, 0.06, 0.02, 0.01)$$

At $\delta = 0.05$, the time gets multiplied by $\ln \frac{1}{\delta} = 3.0$.

Sampling Rule

Look at the lower bound again. Any good algorithm **must** sample with optimal (**oracle**) proportions

$$\boldsymbol{w}^*(\boldsymbol{\mu}) \;=\; \underset{\boldsymbol{w}\in\triangle_K}{\operatorname{argmax}} \min_{\boldsymbol{\lambda}\in\mathsf{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i \,\mathsf{KL}(\mu_i \| \lambda_i)$$

Sampling Rule

Look at the lower bound again. Any good algorithm **must** sample with optimal (**oracle**) proportions

$$\boldsymbol{w}^*(\boldsymbol{\mu}) \;=\; \operatorname*{argmax}_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i \, \mathrm{KL}(\mu_i \| \lambda_i)$$

Track-and-Stop

Idea: draw $I_t \sim \boldsymbol{w}^*(\hat{\boldsymbol{\mu}}(t))$.

- Ensure $\hat{\boldsymbol{\mu}}(t) \to \boldsymbol{\mu}$ hence $N_i(t)/t \to w_i^*$ by “forced exploration”
- Draw arm with $N_i(t)/t$ below w_i^* (tracking)
- Computation of $\boldsymbol{w}^*$ (reduction to 1d line search)

All in all

Final result: lower and upper bound meet on **every problem instance**.

Theorem (Garivier and Kaufmann 2016)

For **Track-and-Stop** *algorithm, for any bandit* $\boldsymbol{\mu}$

$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}{\ln \frac{1}{\delta}} = T^*(\boldsymbol{\mu})$$

All in all

Final result: lower and upper bound meet on **every problem instance**.

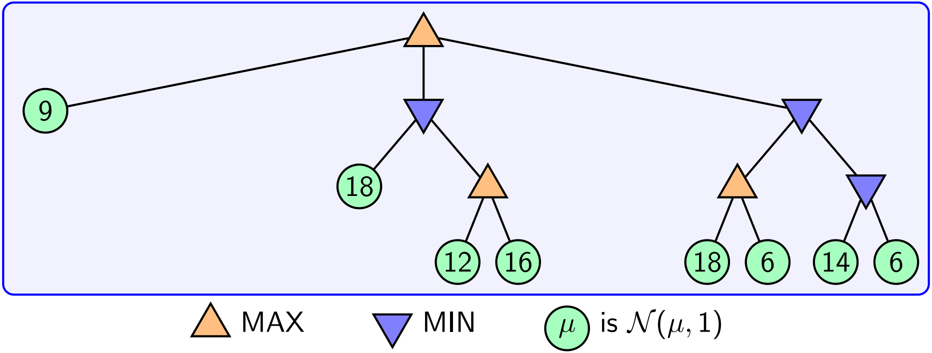
Theorem (Garivier and Kaufmann 2016)

For **Track-and-Stop** *algorithm, for any bandit* $\boldsymbol{\mu}$

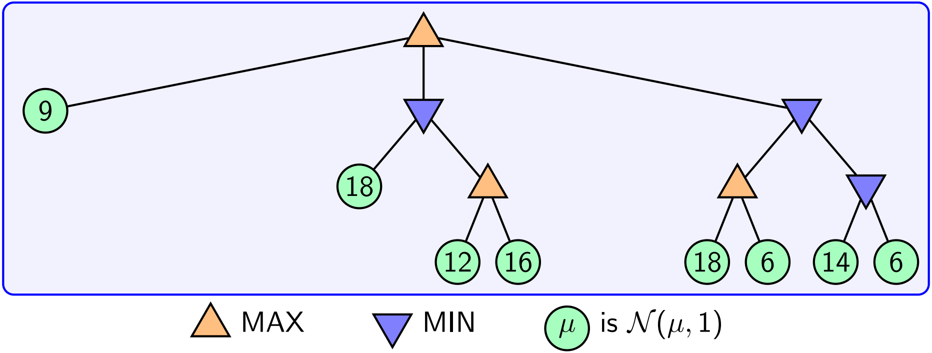
$$\limsup_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau]}{\ln \frac{1}{\delta}} = T^*(\boldsymbol{\mu})$$

Very similar optimality result for **Top Two Thompson Sampling** by Russo [2016]. Here $N_i(t)/t \to w_i^*$ result of posterior sampling.

Challenge Environment: Stochastic Game Tree Search



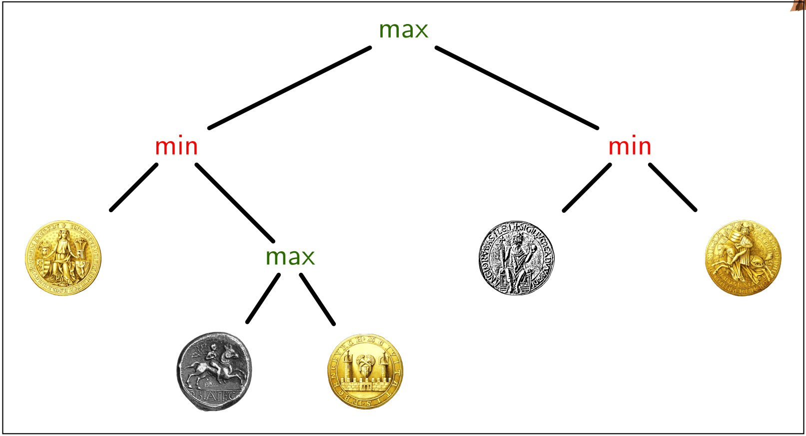
Challenge Environment: Stochastic Game Tree Search



Problem

Determine the optimal move at the root
From **sample access** to the leaf payoffs

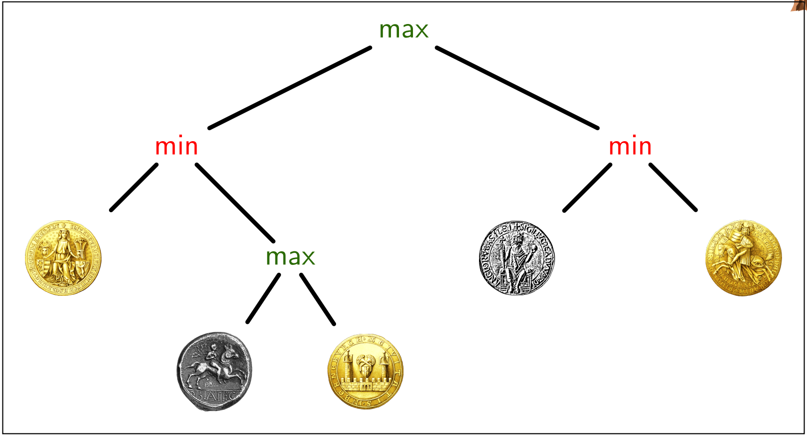
Model [Teraoka et al., 2014]



Maximin Action Identification Problem

Find **best move at root** from samples of **leaves**.

Model [Teraoka et al., 2014]



Maximin Action Identification Problem

Find **best move at root** from samples of **leaves**.

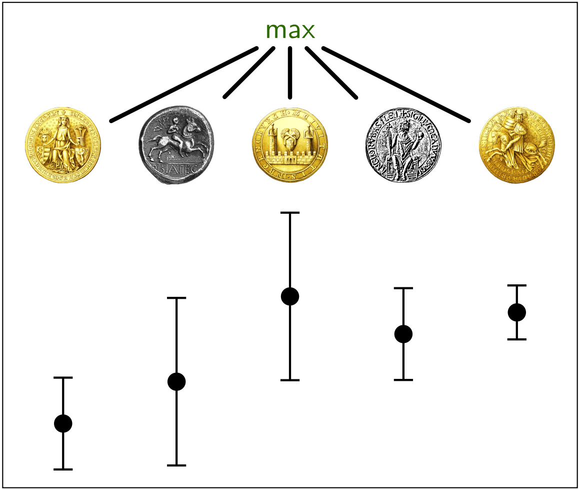
guarantee?

Methods for Best Arm Identification

Methods for Best Arm Identification

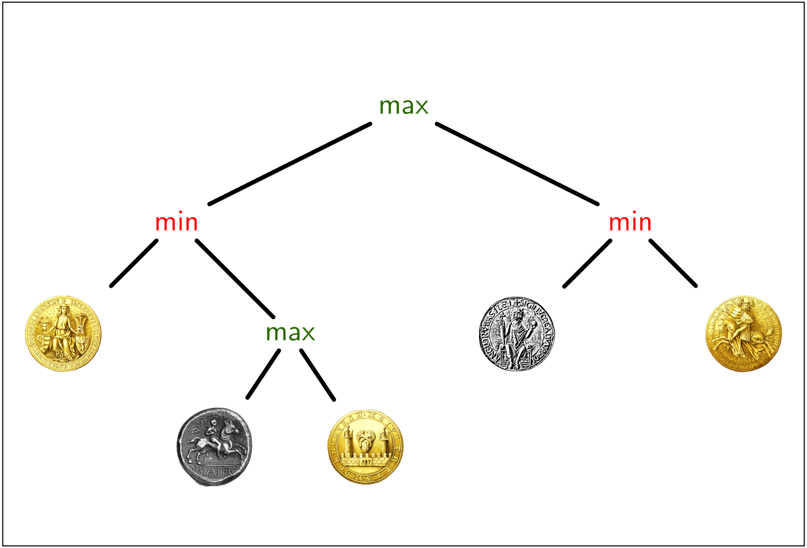
LUCB,
UGapE,
...

Methods for Best Arm Identification

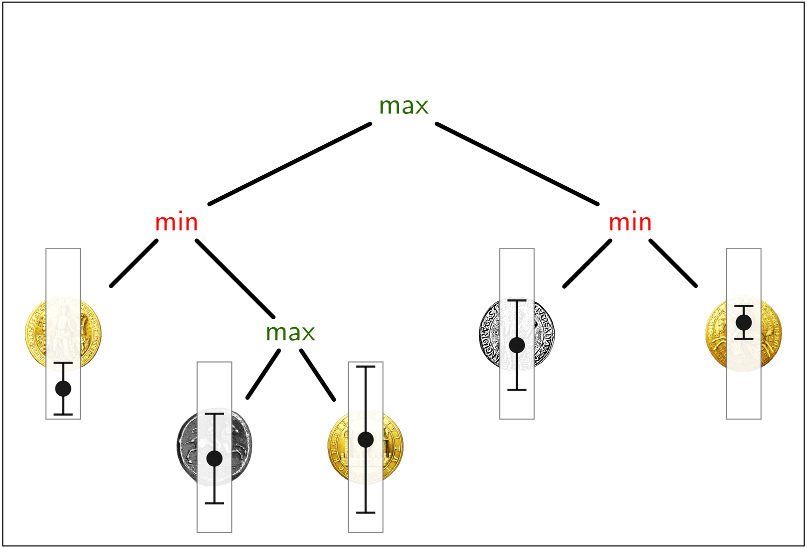


LUCB,
UGapE,
...

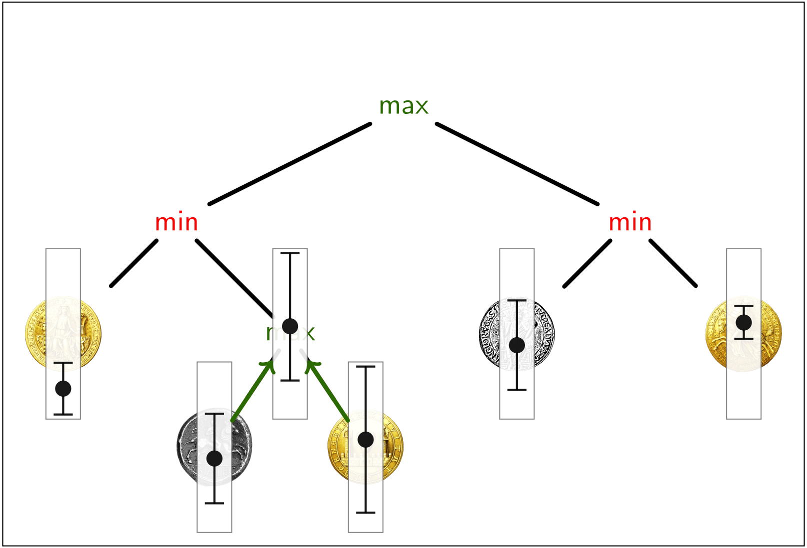
Reduction of MCTS to BAI



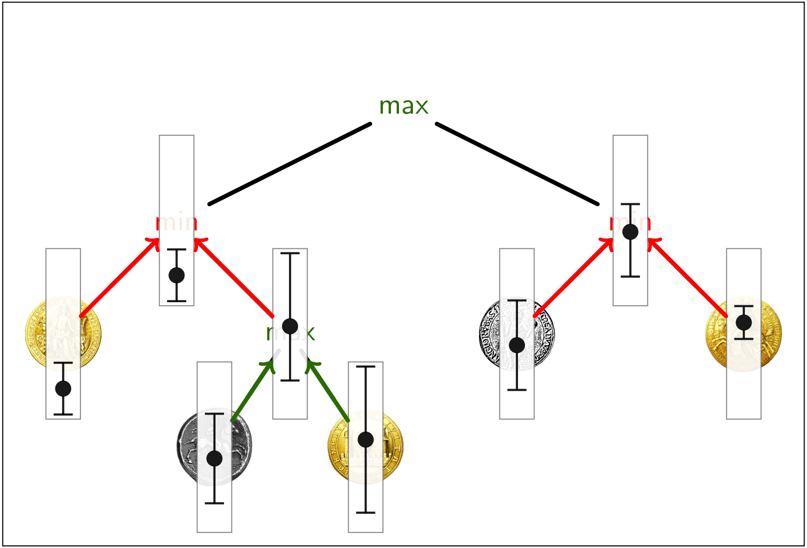
Reduction of MCTS to BAI



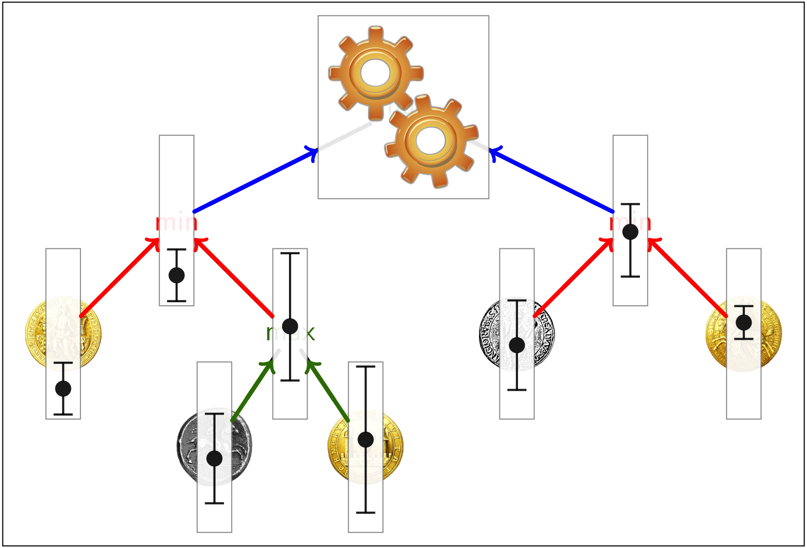
Reduction of MCTS to BAI



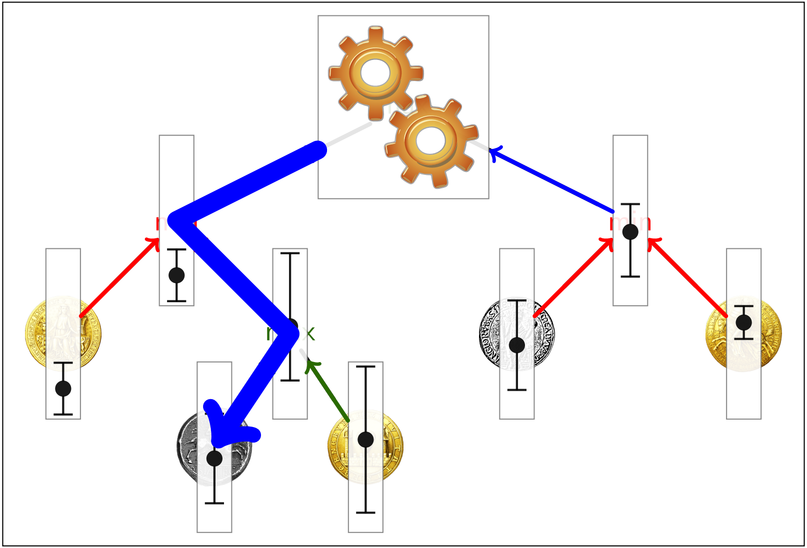
Reduction of MCTS to BAI



Reduction of MCTS to BAI



Reduction of MCTS to BAI



Results

Correctness

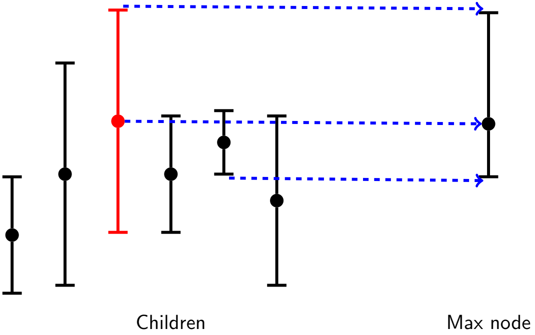
(ϵ, δ)-PAC algorithms

Efficiency

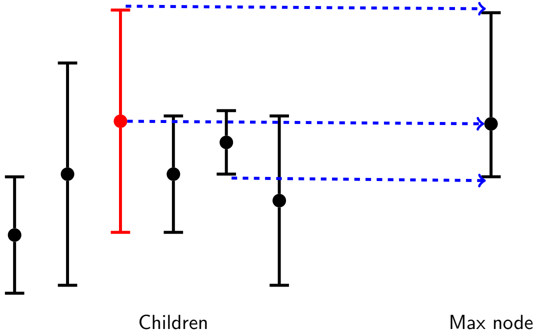
Sample complexity function of **leaf gaps** Δ_ℓ

$$O\left(\sum_{\ell \in \mathcal{L}} \frac{1}{\Delta_\ell^2 \vee \epsilon^2} \log\left(\frac{1}{\delta}\right)\right)$$

How to build confidence intervals on min/max nodes

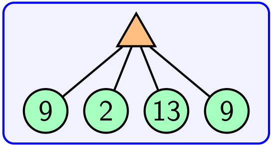


How to build confidence intervals on min/max nodes

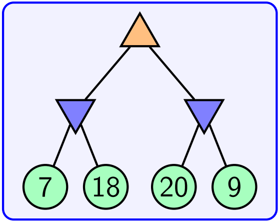


More principled approach in [Kaufmann, Koolen, and Garivier, 2018].
Many equal intervals $\Rightarrow$ higher lower bound.

Simplify

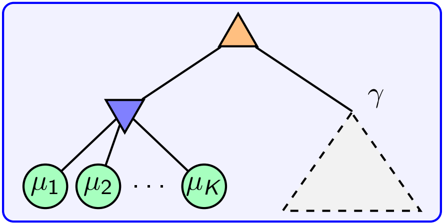


Best Arm Identification
[Garivier and Kaufmann, 2016]
Solved

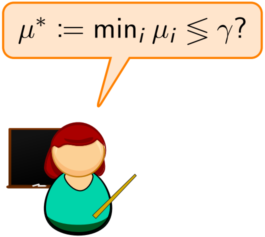


Depth 2
[Garivier, Kaufmann, and Koolen, 2016]
Open

Simple Instance: Minimum Threshold Identification



Fix threshold γ.



For $t = 1, \ldots, \tau$
- Pick leaf A_t
- See $X_t \sim \mu_{A_t}$

Recommend $\hat{m} \in \{<, >\}$

Goal: **fixed confidence** $\mathbb{P}_{\boldsymbol{\mu}} \{\text{error}\} < \delta$ and small **sample complexity** $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$

Lower Bound

Generic lower bound [Castro, 2014, Garivier and Kaufmann, 2016] shows **sample complexity** for **any** δ-correct algorithm is at least

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \;\geq\; T^*(\boldsymbol{\mu}) \ln \frac{1}{\delta}.$$

Lower Bound

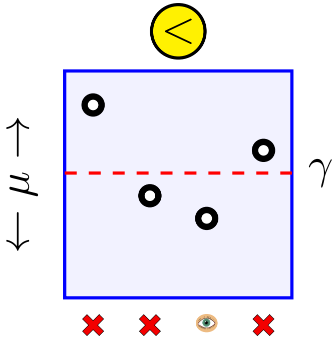
Generic lower bound [Castro, 2014, Garivier and Kaufmann, 2016] shows **sample complexity** for **any** δ-correct algorithm is at least

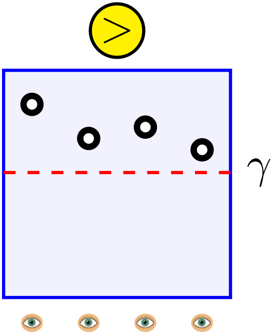
$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \;\geq\; T^*(\boldsymbol{\mu}) \ln \tfrac{1}{\delta}.$$

For our problem the **characteristic time** and **oracle weights** are

$$T^*(\boldsymbol{\mu}) \;=\; \begin{cases} \dfrac{1}{\mathrm{KL}(\mu^*, \gamma)} & \mu^* < \gamma, \\ \displaystyle\sum_a \dfrac{1}{\mathrm{KL}(\mu_a, \gamma)} & \mu^* > \gamma, \end{cases} \qquad w_a^*(\boldsymbol{\mu}) \;=\; \begin{cases} \mathbf{1}_{a=a^*} & \mu^* < \gamma, \\ \dfrac{\frac{1}{\mathrm{KL}(\mu_a,\gamma)}}{\sum_j \frac{1}{\mathrm{KL}(\mu_j,\gamma)}} & \mu^* > \gamma. \end{cases}$$

Dichotomous Oracle Behaviour! Sampling Rule?





Sampling Rules

- **Lower Confidence Bounds**
  Play $A_t = \arg\min_a \ \mathrm{LCB}_a(t)$
- **Thompson Sampling** (Π_{t-1} is posterior after $t-1$ rounds)
  Sample $\boldsymbol{\theta} \sim \Pi_{t-1}$, then play $A_t = \arg\min_a \ \theta_a$.
- **Murphy Sampling** **condition on low minimum mean**
  Sample $\boldsymbol{\theta} \sim \Pi_{t-1}\left(\cdot \middle| \min_a \theta_a < \gamma\right)$, then play $A_t = \arg\min_a \ \theta_a$.

Intuition for Murphy Sampling

- When $\mu^* < \gamma$ conditioning is immaterial: $\boldsymbol{\theta} \approx \boldsymbol{\mu}$ and MS $\equiv$ TS.
- When $\mu^* > \gamma$ conditioning results in $\boldsymbol{\theta} \approx (\mu_1, \ldots, \gamma, \ldots, \mu_K)$. Index a lowered to γ with probability $\propto \frac{1}{\mathrm{KL}(\mu_a, \gamma)}$ [Russo, 2016].

Murphy Sampling Rule [KKG, NIPS'18]

Theorem

Asymptotic optimality: $N_a(t)/t \to w_a^*(\boldsymbol{\mu})$ *for all* $\boldsymbol{\mu}$

Sampling rule	$<$	$>$
Thompson Sampling	✓	✗
Lower Confidence Bounds	✗	✓
Murphy Sampling	✓	✓

Lemma

Any anytime sampling strategy $(A_t)_t$ *ensuring* $\frac{\boldsymbol{N}_t}{t} \to \boldsymbol{w}^*(\boldsymbol{\mu})$ *and* **good** *stopping rule* τ_δ *guarantee* $\limsup_{\delta \to 0} \frac{\tau_\delta}{\ln \frac{1}{\delta}} \leq T^*(\boldsymbol{\mu})$.

Conclusion

- Pure Exploration currently going through a renaissance
- Instance-optimal identification algorithms
  - Best Arm
  - Game Tree Search
  - ...
- Moving toward more complex queries. RL on the horizon ...
- Useful submodules

Conclusion

- Pure Exploration currently going through a renaissance
- Instance-optimal identification algorithms
  - Best Arm
  - Game Tree Search
  - ...
- Moving toward more complex queries. RL on the horizon ...
- Useful submodules

Thank you! And let's talk!