Bandit Algorithms for Pure Exploration: Best Arm Identification and Game Tree Search

Wouter M. Koolen

CWI
Centrum Wiskunde & Informatica

Nederlands Mathematisch Congres, Session on Mathematics of Machine Learning
Wednesday 4^{th} April, 2018

Outline

Best Arm Identification (BAI) Problem

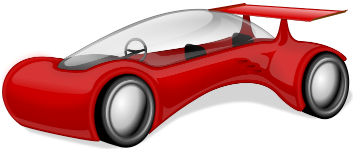
What is the drug with highest effect?

Best Arm Identification (BAI) Problem

What is the drug with highest effect?

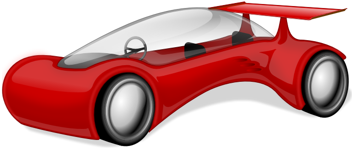
What is the coin with highest expected reward?

Combinatorial Pure Exploration (CPE) Problems

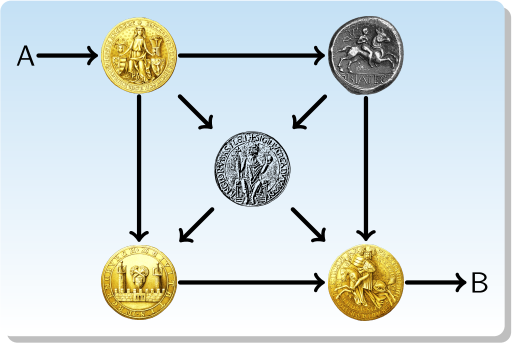


What is the shortest path from A to B?

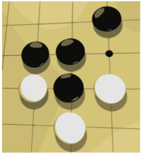
Combinatorial Pure Exploration (CPE) Problems



What is the shortest path from A to B?

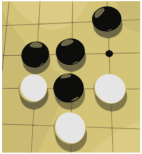


Maximin Action Identification (MMAI) Problem

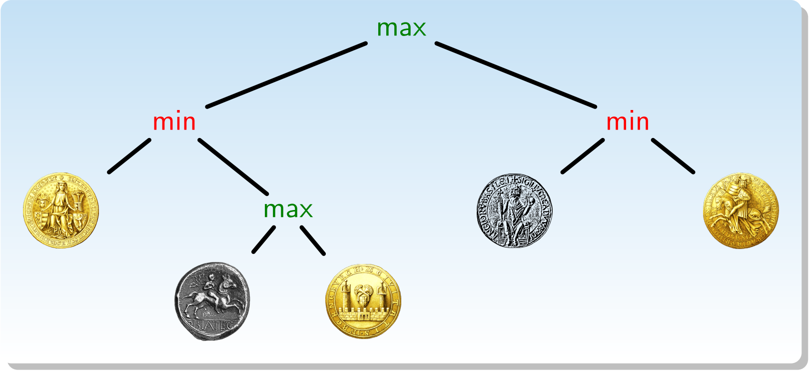
 

What is the optimal move in a given position?

Maximin Action Identification (MMAI) Problem



What is the optimal move in a given position?



Question

Complexity of interactive learning.

Complexity of interactive learning.

- Medical testing [Villar et al., 2015]
- Online advertising and website optimisation [Zhou et al., 2014]
- Monte Carlo planning [Grill et al., 2016], and
- Game-playing AI [Silver et al., 2016]

Pure Exploration

Query: **Which is . . .**

- **the most effective drug dose?**
- **the most appealing website layout?**
- **the safest next robot action?**

Pure Exploration

Query: **Which is . . .**

- **the most effective drug dose?**
- **the most appealing website layout?**
- **the safest next robot action?**

Method

- statistical **experiments** in physical or simulated environment, **interactively** and **adaptively**.

Pure Exploration

Query: **Which is . . .**

- **the most effective drug dose?**
- **the most appealing website layout?**
- **the safest next robot action?**

Method

- statistical **experiments** in physical or simulated environment, **interactively** and **adaptively**.

Main scientific questions:

- **sample complexity** of interactive learning
  # experiments as function of query structure and environment
- Design of **efficient** pure exploration systems

Outline

Formal model

Environment (Multi-armed bandit model)

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

The **best arm** is

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

Formal model

Environment (Multi-armed bandit model)

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

The **best arm** is

$$i^* = \underset{i \in [K]}{\text{argmax}} \ \mu_i$$

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

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

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

$$i^* = \underset{i \in [K]}{\text{argmax}} \; \mu_i$$

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

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** $\mathbb{P}_{\boldsymbol{\mu}}\left(\hat{I} \neq i^*(\boldsymbol{\mu})\right)$, 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

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

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

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

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

Families of approaches to BAI

- **Upper and Lower confidence bounds** [Bubeck et al., 2011, Kalyanakrishnan et al., 2012, Gabillon et al., 2012, Kaufmann and Kalyanakrishnan, 2013, Jamieson et al., 2014],
- **Racing or Successive Rejects/Eliminations** [Maron and Moore, 1997, Even-Dar et al., 2006, Audibert et al., 2010, Kaufmann and Kalyanakrishnan, 2013, Karnin et al., 2013],
- **Thompson Sampling** (partly Bayesian) [Russo, 2016]
- Track-and-Stop [Garivier and Kaufmann, 2016].

Outline

Sample Complexity Lower bound

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

Sample Complexity Lower bound

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

Theorem (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*

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

Sample Complexity Lower bound

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

Theorem (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*

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

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

Bernoulli

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

Gaussian ($\sigma^2 = 1/4$)

$$T^*(\boldsymbol{\mu}) \;=\; 223.4 \qquad \boldsymbol{w}^*(\boldsymbol{\mu}) \;=\; (0.45, 0.44, 0.06, 0.03, 0.01)$$

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

Change of Measure Argument

Strategy and model $\boldsymbol{\mu}$ induce distribution on $\Omega = \{(I_t, X_t)_{t \leq \tau}, \hat{I}\}$

Change of Measure Argument

Strategy and model $\boldsymbol{\mu}$ induce distribution on $\Omega = \{(I_t, X_t)_{t \le \tau}, \hat{I}\}$

1. By KL-contraction on $\mathcal{E} = \{\hat{I} \neq i^*(\boldsymbol{\mu})\}$ and δ-correctness, $\boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\mu})$

$$\text{KL}\left(\boldsymbol{\mu}(\Omega) \| \boldsymbol{\lambda}(\Omega)\right) \;\ge\; \text{KL}\left(\boldsymbol{\mu}(\mathcal{E}) \| \boldsymbol{\lambda}(\mathcal{E})\right) \;\ge\; \text{KL}\left(\delta \| 1-\delta\right) \;\to\; \ln \frac{1}{\delta}$$

Change of Measure Argument

Strategy and model $\boldsymbol{\mu}$ induce distribution on $\Omega = \{(I_t, X_t)_{t \le \tau}, \hat{I}\}$

1. By KL-contraction on $\mathcal{E} = \{\hat{I} \neq i^*(\boldsymbol{\mu})\}$ and δ-correctness, $\boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\mu})$

$$\text{KL}\left(\boldsymbol{\mu}(\Omega) \| \boldsymbol{\lambda}(\Omega)\right) \;\ge\; \text{KL}\left(\boldsymbol{\mu}(\mathcal{E}) \| \boldsymbol{\lambda}(\mathcal{E})\right) \;\ge\; \text{KL}\left(\delta \| 1-\delta\right) \;\to\; \ln \frac{1}{\delta}$$

2. Samples X_t are independent given I_t

$$\text{KL}\left(\boldsymbol{\mu}(\Omega) \| \boldsymbol{\lambda}(\Omega)\right) \;=\; \textstyle\sum_{i=1}^K \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau)]\, \text{KL}(\mu_i \| \lambda_i)$$

Change of Measure Argument

Strategy and model $\boldsymbol{\mu}$ induce distribution on $\Omega = \{(I_t, X_t)_{t \le \tau}, \hat{I}\}$

1. By KL-contraction on $\mathcal{E} = \{\hat{I} \neq i^*(\boldsymbol{\mu})\}$ and δ-correctness, $\boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\mu})$

$$\text{KL}\left(\boldsymbol{\mu}(\Omega) \| \boldsymbol{\lambda}(\Omega)\right) \;\ge\; \text{KL}\left(\boldsymbol{\mu}(\mathcal{E}) \| \boldsymbol{\lambda}(\mathcal{E})\right) \;\ge\; \text{KL}\left(\delta \| 1 - \delta\right) \;\to\; \ln \frac{1}{\delta}$$

2. Samples X_t are independent given I_t

$$\text{KL}\left(\boldsymbol{\mu}(\Omega) \| \boldsymbol{\lambda}(\Omega)\right) \;=\; \textstyle\sum_{i=1}^K \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau)] \, \text{KL}(\mu_i \| \lambda_i)$$

3. Bring out sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau] = \sum_{i=1}^K \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau)]$

$$\sum_{i=1}^K \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau)] \, \text{KL}(\mu_i \| \lambda_i) \;=\; \mathbb{E}_{\boldsymbol{\mu}}[\tau] \sum_{i=1}^K \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]} \, \text{KL}(\mu_i \| \lambda_i)$$

Change of Measure Argument

Strategy and model $\boldsymbol{\mu}$ induce distribution on $\Omega = \{(I_t, X_t)_{t \le \tau}, \hat{I}\}$

1. By KL-contraction on $\mathcal{E} = \{\hat{I} \neq i^*(\boldsymbol{\mu})\}$ and δ-correctness, $\boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\mu})$

$$\text{KL}\left(\boldsymbol{\mu}(\Omega) \| \boldsymbol{\lambda}(\Omega)\right) \;\ge\; \text{KL}\left(\boldsymbol{\mu}(\mathcal{E}) \| \boldsymbol{\lambda}(\mathcal{E})\right) \;\ge\; \text{KL}\left(\delta \| 1 - \delta\right) \;\to\; \ln \frac{1}{\delta}$$

2. Samples X_t are independent given I_t

$$\text{KL}\left(\boldsymbol{\mu}(\Omega) \| \boldsymbol{\lambda}(\Omega)\right) \;=\; \sum_{i=1}^K \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau)] \, \text{KL}(\mu_i \| \lambda_i)$$

3. Bring out sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau] = \sum_{i=1}^K \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau)]$

$$\sum_{i=1}^K \mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau)] \, \text{KL}(\mu_i \| \lambda_i) \;=\; \mathbb{E}_{\boldsymbol{\mu}}[\tau] \sum_{i=1}^K \frac{\mathbb{E}_{\boldsymbol{\mu}}[N_i(\tau)]}{\mathbb{E}_{\boldsymbol{\mu}}[\tau]} \, \text{KL}(\mu_i \| \lambda_i)$$

4. Pick tightest alternative $\boldsymbol{\lambda}$ and best (**oracle**) proportions w_i:

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \text{Alt}(\boldsymbol{\mu})} \; \sum_{i=1}^K w_i \, \text{KL}(\mu_i \| \lambda_i) \;\ge\; \ln \frac{1}{\delta}$$

Outline

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**.

Sampling Rule

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

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

Sampling Rule

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

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

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)

Stopping Rule

Sufficient evidence to stop? Classical hypothesis test [Wald, 1945].

Stopping Rule

Sufficient evidence to stop? Classical hypothesis test [Wald, 1945].

Generalized Likelihood Ratio Test (GLRT)

$$Z_t = \ln \frac{P_{\hat{\boldsymbol{\mu}}(t)}(data)}{\max_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} P_{\boldsymbol{\lambda}}(data)}$$

Stopping Rule

Sufficient evidence to stop? Classical hypothesis test [Wald, 1945].

Generalized Likelihood Ratio Test (GLRT)

$$Z_t \;=\; \ln \frac{P_{\hat{\boldsymbol{\mu}}(t)}(data)}{\max_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} P_{\boldsymbol{\lambda}}(data)}$$

Turns out, GLRT statistic equals

$$Z_t \;=\; \min_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \; \sum_{i=1}^{K} N_i(t)\, \mathrm{KL}(\hat{\mu}_i(t) \| \lambda_i)$$

i.e. lower bound with $\hat{\boldsymbol{\mu}}(t)$ plug-in.

Stopping Rule

Sufficient evidence to stop? Classical hypothesis test [Wald, 1945].

Generalized Likelihood Ratio Test (GLRT)

$$Z_t \;=\; \ln \frac{P_{\hat{\boldsymbol{\mu}}(t)}(data)}{\max_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} P_{\boldsymbol{\lambda}}(data)}$$

Turns out, GLRT statistic equals

$$Z_t \;=\; \min_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}(t))} \; \sum_{i=1}^{K} N_i(t) \, \mathrm{KL}(\hat{\mu}_i(t) \| \lambda_i)$$

i.e. lower bound with $\hat{\boldsymbol{\mu}}(t)$ plug-in.

Roughly: stop when $Z_t \geq \ln \frac{1}{\delta}$. Make precise with careful universal coding (MDL) argument.

All in all

Final result: lower and upper bound meet.

Theorem

For **Track-and-Stop** *algorithm*

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

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

Outline

Beyond asymptotic bounds

Okay, so good algorithms have

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

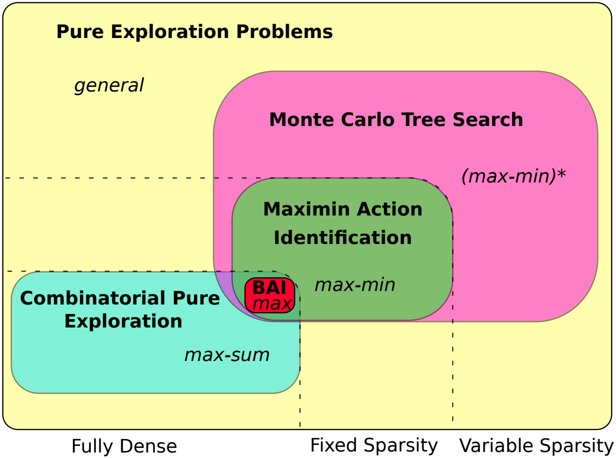
What about lower-order terms? "Moderate confidence" regime!

- Dependence on $\ln \ln \frac{1}{\delta}$.
- Dependence on $\ln K$ (i.e. Fano).

[Simchowitz et al., 2017, Chen et al., 2017b]

Beyond Best Arm

Practical and fundamental question: solving more complex pure exploration problems.



Combinatorial Pure Exploration

- Best k-set
- Shortest path
- Spanning tree
- ...

Combinatorial Pure Exploration

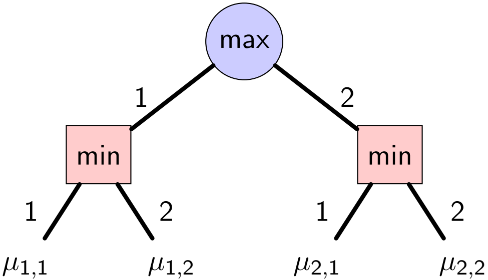
- Best k-set
- Shortest path
- Spanning tree
- ...

Combinatorial collection $\mathcal{F}$ of subsets of $[K]$.

$$i^* = i^*(\boldsymbol{\mu}) = \underset{S \in \mathcal{F}}{\text{argmax}} \sum_{i \in S} \mu_i.$$

Track-and-stop-like algorithms [Chen et al., 2017a]. Can compute oracle weights. Dense.

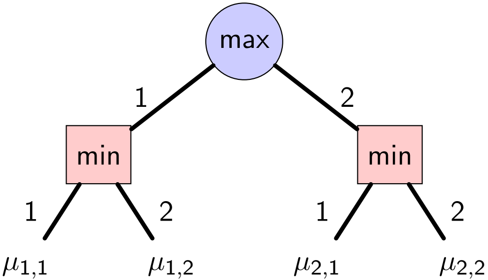
Game Tree Search



Goal: find **maximin action**

$$i^* \quad := \quad \arg\max_i \min_j \mu_{i,j}$$

Game Tree Search



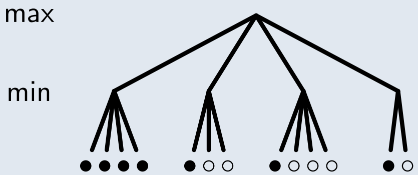
Goal: find **maximin action**

$$i^* \ := \ \arg\max_i \min_j \mu_{i,j}$$

Range of algorithms: Teraoka et al. [2014], Garivier et al. [2016], Kaufmann and Koolen [2017], Huang et al. [2017]

Sparsity in the Lower Bound (depth 2)

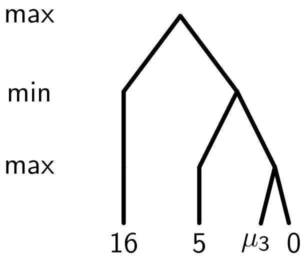
Sparsity Pattern [Kaufmann and Koolen, 2017]

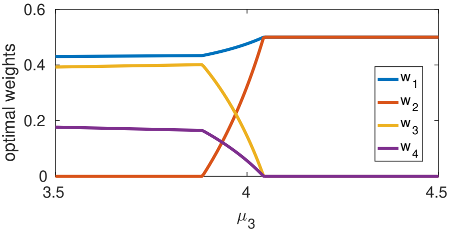


Oracle weights $\boldsymbol{w}^*$ supported on only 7 of the 13 leaves.

Open problem: algorithms incorporating appropriate pruning?

Sparsity in the Lower Bound (depth 3)





Oracle weights $\boldsymbol{w}^* = (w_1, w_2, w_3, w_4)$ as a function of μ_3

Open Problem: Characterisation of sparsity patterns.
Computation.

Conclusion

BAI: Invert lower bounds to obtain **algorithms**.
Challenge: landscape of all pure exploration problems