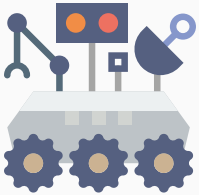
Instance Optimal Pure Exploration in Bandit Models



Wouter M. Koolen

CWI
Centrum Wiskunde & Informatica

UNIVERSITY
OF TWENTE.

Young European Queueing Theorists Workshop
EURANDOM
Wednesday 2nd November, 2022

Warm Thanks

Emilie Kaufmann

Rémy Degenne

Pierre Menard

Han Shao

Aurelién Garivier

Sandeep Juneja

Shubhada Agrawal

Christina Katsimerou

Olivier Cappé

Yoan Russac

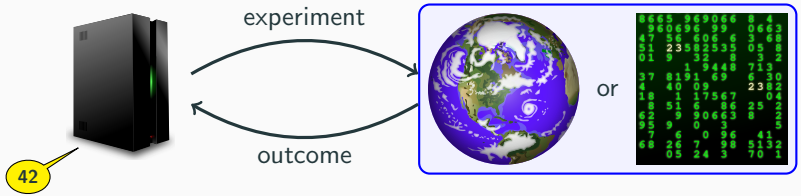
Grand Goal: Interactive Machine Learning

Query:

most effective drug dose?

most appealing website layout?

safest next robot action?



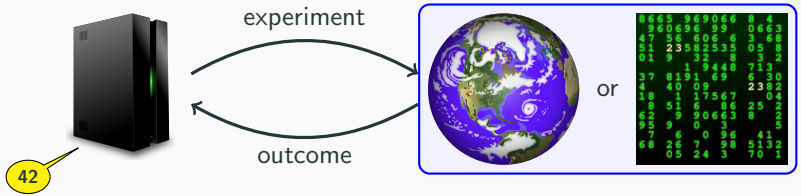
Grand Goal: Interactive Machine Learning

Query:

most effective drug dose?

most appealing website layout?

safest next robot action?



Main scientific questions

- **Efficient** systems
- **Sample complexity** as function of query and environment

Pure Exploration vs Reinforcement Learning

Both are about **learning in uncertain environments**.

Pure Exploration vs Reinforcement Learning

Both are about **learning in uncertain environments**.

Pure Exploration focuses on the statistical problem (learn the truth),
while Reinforcement Learning focuses on behaviour (maximise reward).

Pure Exploration vs Reinforcement Learning

Both are about **learning in uncertain environments**.

Pure Exploration focuses on the statistical problem (learn the truth),
while Reinforcement Learning focuses on behaviour (maximise reward).

Pure Exploration occurs as **sub-module** in some RL algorithms (i.e. Phased Q-Learning by Even-Dar, Mannor, and Mansour, 2002)

Pure Exploration vs Reinforcement Learning

Both are about **learning in uncertain environments**.

Pure Exploration focuses on the statistical problem (learn the truth),
while Reinforcement Learning focuses on behaviour (maximise reward).

Pure Exploration occurs as **sub-module** in some RL algorithms (i.e. Phased Q-Learning by Even-Dar, Mannor, and Mansour, 2002)

Some problems approached with RL are in fact **better modelled** as pure exploration problems. Most notably MCTS for playing games.

Outline

Stochastic Bandit

Experiments

Outcomes

Stochastic Bandit

Experiments 

Outcomes 

Instance (Unknown)

$$\mathbb{P}\left(\text{  } \right) = 1/6$$

$$\mathbb{P}\left(\text{  } \right) = 4/6$$

$$\mathbb{P}\left(\text{  } \right) = 3/6$$

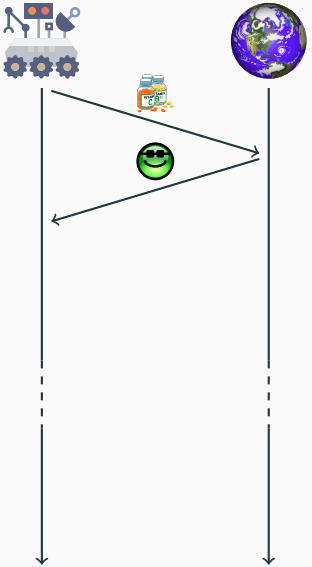
Best Arm Identification Interaction

$$\mathbb{P}\left(\text{  } \middle| \text{  }\right) = 1/6$$

$$\mathbb{P}\left(\text{  } \middle| \text{  }\right) = 4/6$$

$$\mathbb{P}\left(\text{  } \middle| \text{  }\right) = 3/6$$

Best Arm Identification Interaction

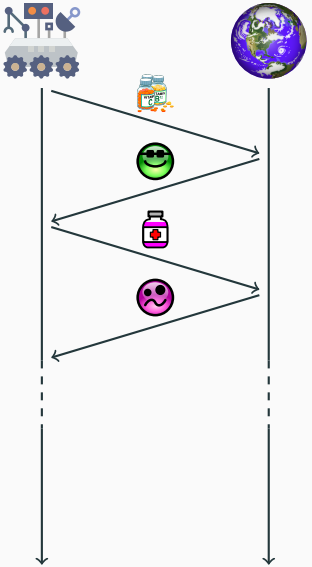


$$\mathbb{P}\left(\text{🟢} \mid \text{🥫}\right) = 1/6$$

$$\mathbb{P}\left(\text{🟢} \mid \text{🧪}\right) = 4/6$$

$$\mathbb{P}\left(\text{🟢} \mid \text{🛒}\right) = 3/6$$

Best Arm Identification Interaction

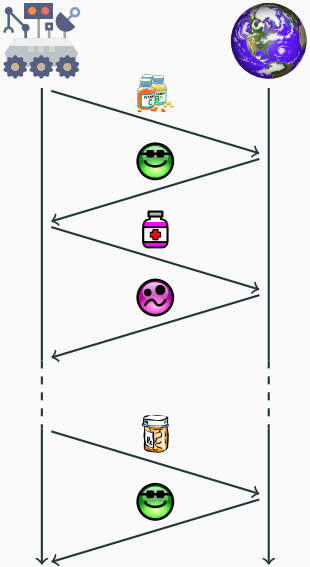


$$\mathbb{P}\left(\text{😊} \mid \text{🫙}\right) = 1/6$$

$$\mathbb{P}\left(\text{😊} \mid \text{💊}\right) = 4/6$$

$$\mathbb{P}\left(\text{😊} \mid \text{🛒}\right) = 3/6$$

Best Arm Identification Interaction

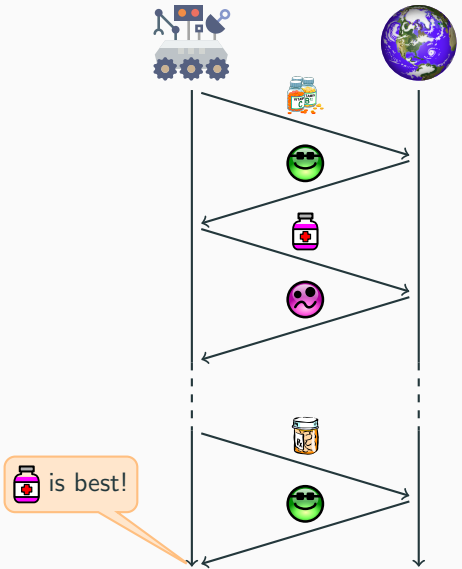


$$\mathbb{P}\left(\text{😎} \mid \text{🥫}\right) = 1/6$$

$$\mathbb{P}\left(\text{😎} \mid \text{💊}\right) = 4/6$$

$$\mathbb{P}\left(\text{😎} \mid \text{🛒}\right) = 3/6$$

Best Arm Identification Interaction

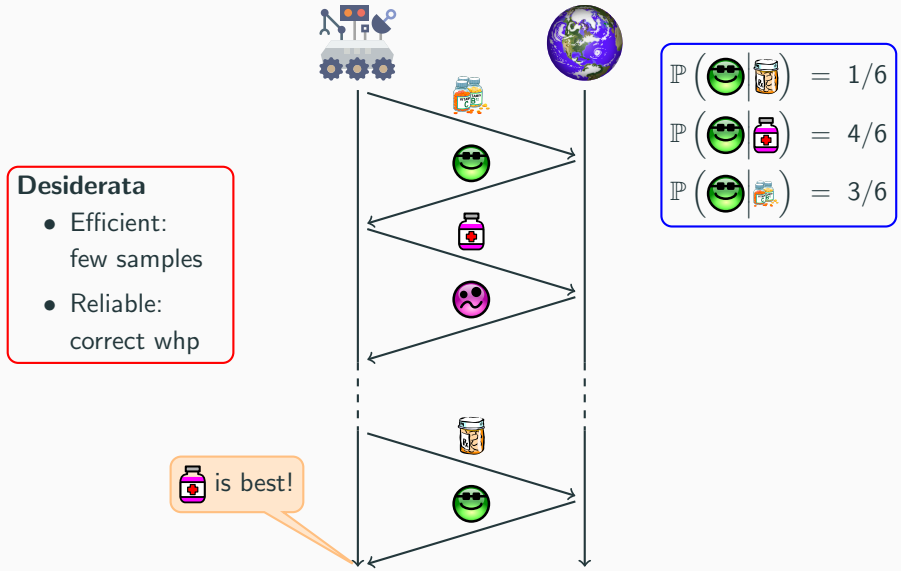


$$\mathbb{P}\left(\text{😊} \mid \text{🧪}\right) = 1/6$$

$$\mathbb{P}\left(\text{😊} \mid \text{💊}\right) = 4/6$$

$$\mathbb{P}\left(\text{😊} \mid \text{🤢}\right) = 3/6$$

Best Arm Identification Interaction



Identification Problems

Problem (Even-Dar, Mannor, and Mansour, 2002)

Which arm has the highest mean

Arms: Bernoulli, Exp. Fam, bounded support, sub-Gaussian, ...

Problem (Yu and Nikolova, 2013)

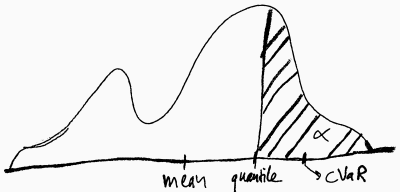
Which arm has the highest α-quantile

Arms: Unrestricted (on $\mathbb{R}$)

Problem (Yu and Nikolova, 2013)

Which arm has the smallest Conditional Value at Risk.

Arms: Exp. Fam (trivial), bounded $(1+\epsilon)^{\text{th}}$ moment



Identification Problems (continued)

Set-valued Queries

Top-M, Thresholding, All ϵ-optimal, all-better-than-control

Structure: standard bandit

Ranking-related queries: Borda/Condorcet winner (Yue et al., 2012)

Structure: duelling bandit

Best arm in stratified population (Russac et al., 2021)

Structure: contextual bandit

Minimax action in extensive form game tree (Teraoka, Hatano, and Takimoto, 2014)

Structure: game tree with stochastic leaves

Shortest Path (Chen et al., 2014)

Structure: graph with stochastic edge costs

Outline

Best Arm Identification (BAI)

Assumption: Bernoulli Multi-Armed Bandit

K Bernoulli arms with unknown means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K) \in [0,1]^K$.

Best Arm Identification (BAI)

Assumption: Bernoulli Multi-Armed Bandit

K Bernoulli arms with unknown means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K) \in [0,1]^K$.

The **best arm** is

$$i^*(\boldsymbol{\mu}) \;=\; \arg\max_i \mu_i$$

Best Arm Identification (BAI)

Assumption: Bernoulli Multi-Armed Bandit

K Bernoulli arms with unknown means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K) \in [0,1]^K$.

The **best arm** is

$$i^*(\boldsymbol{\mu}) \;=\; \arg\max_i \mu_i$$

BAI-MAB Protocol

for $t = 1, 2, \ldots$ **until** Learner decides to stop

- Learner picks arm $A_t \in [K]$
- Learner observes $X_t \sim \text{Bernoulli}(\mu_{A_t})$

Learner recommends $\hat{I} \in [K]$.

Best Arm Identification

BAI-MAB Protocol

for $t = 1, 2, \ldots$ **until** Learner decides to stop

- Learner picks arm $A_t \in [K]$
- Learner observes $X_t \sim \text{Bernoulli}(\mu_{A_t})$

Learner recommends $\hat{I} \in [K]$.

Let $\tau \in \mathbb{N} \cup \{\infty\}$ denote the # rounds after which Learner stops.

Best Arm Identification

BAI-MAB Protocol

for $t = 1, 2, \ldots$ **until** Learner decides to stop

- Learner picks arm $A_t \in [K]$
- Learner observes $X_t \sim \text{Bernoulli}(\mu_{A_t})$

Learner recommends $\hat{I} \in [K]$.

Let $\tau \in \mathbb{N} \cup \{\infty\}$ denote the # rounds after which Learner stops.

Definition

Learner is δ-**PAC** if $\mathbb{P}_{\boldsymbol{\mu}}\Big\{\underbrace{\tau < \infty \text{ and } \hat{I} \neq i^*(\boldsymbol{\mu})}_{\text{a mistake}}\Big\} \leq \delta$ for all $\boldsymbol{\mu} \in [0,1]^K$.

Best Arm Identification

BAI-MAB Protocol

for $t = 1, 2, \ldots$ **until** Learner decides to stop

- Learner picks arm $A_t \in [K]$
- Learner observes $X_t \sim \text{Bernoulli}(\mu_{A_t})$

Learner recommends $\hat{I} \in [K]$.

Let $\tau \in \mathbb{N} \cup \{\infty\}$ denote the # rounds after which Learner stops.

Definition

Learner is δ-**PAC** if $\mathbb{P}_{\boldsymbol{\mu}}\Big\{\underbrace{\tau < \infty \text{ and } \hat{I} \neq i^*(\boldsymbol{\mu})}_{\text{a mistake}}\Big\} \leq \delta$ for all $\boldsymbol{\mu} \in [0,1]^K$.

Definition

We call $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ the **sample complexity** of Learner in bandit $\boldsymbol{\mu}$.

Best Arm Identification

BAI-MAB Protocol

for $t = 1, 2, \ldots$ **until** Learner decides to stop

- Learner picks arm $A_t \in [K]$
- Learner observes $X_t \sim \text{Bernoulli}(\mu_{A_t})$

Learner recommends $\hat{I} \in [K]$.

Let $\tau \in \mathbb{N} \cup \{\infty\}$ denote the # rounds after which Learner stops.

Definition

Learner is δ-**PAC** if $\mathbb{P}_{\boldsymbol{\mu}}\Big\{\underbrace{\tau < \infty \text{ and } \hat{I} \neq i^*(\boldsymbol{\mu})}_{\text{a mistake}}\Big\} \leq \delta$ for all $\boldsymbol{\mu} \in [0,1]^K$.

Definition

We call $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ the **sample complexity** of Learner in bandit $\boldsymbol{\mu}$.

Goal: computationally efficient δ-PAC algorithms with minimal sample complexity.

Outline

Instance-Dependent Sample Complexity Lower Bound

Intuition, going back at least to Lai and Robbins (1985)

(Spectacular) difference in behaviour must be due to (spectacular) difference in observations.

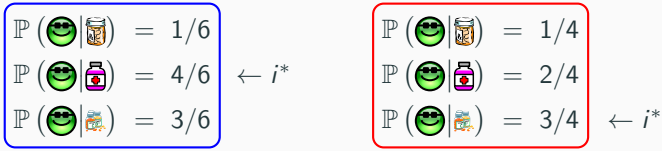
Being δ-PAC on $\boldsymbol{\mu}$ *and* $\boldsymbol{\lambda}$ with $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$ requires gathering enough discriminating data.

Instance-Dependent Sample Complexity Lower Bound

Intuition, going back at least to Lai and Robbins (1985)

(Spectacular) difference in behaviour must be due to (spectacular) difference in observations.

Being δ-PAC on $\boldsymbol{\mu}$ *and* $\boldsymbol{\lambda}$ with $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$ requires gathering enough discriminating data.



If δ-**PAC algorithm** samples t rounds with arm frequencies $(\frac{1}{6}, \frac{3}{6}, \frac{2}{6})$, then

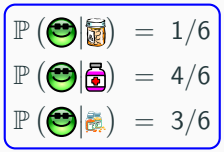
$$t\,\frac{1}{6}\,\mathrm{KL}\left(\frac{1}{6}, \frac{1}{4}\right) + t\,\frac{3}{6}\,\mathrm{KL}\left(\frac{4}{6}, \frac{2}{4}\right) + t\,\frac{2}{6}\,\mathrm{KL}\left(\frac{3}{6}, \frac{3}{4}\right) \;\geq\; \mathrm{KL}(\delta, 1-\delta) \approx \ln \frac{1}{\delta}$$

Instance-Dependent Sample Complexity Lower Bound

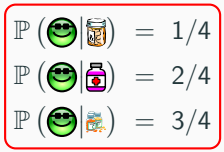
Intuition, going back at least to Lai and Robbins (1985)

(Spectacular) difference in behaviour must be due to (spectacular) difference in observations.

Being δ-PAC on $\boldsymbol{\mu}$ *and* $\boldsymbol{\lambda}$ with $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$ requires gathering enough discriminating data.



$\leftarrow i^*$



$\leftarrow i^*$

If δ-**PAC algorithm** samples t rounds with arm frequencies $(\frac{1}{6}, \frac{3}{6}, \frac{2}{6})$, then

$$t\,\frac{1}{6}\,\mathrm{KL}\left(\frac{1}{6},\frac{1}{4}\right) + t\,\frac{3}{6}\,\mathrm{KL}\left(\frac{4}{6},\frac{2}{4}\right) + t\,\frac{2}{6}\,\mathrm{KL}\left(\frac{3}{6},\frac{3}{4}\right) \;\geq\; \mathrm{KL}(\delta, 1-\delta) \approx \ln\frac{1}{\delta}$$

At typical $\delta = 0.1$: $\quad 0.0796\,t \;\geq\; 1.757 \quad\quad t \;\geq\; \dfrac{1.757}{0.0796} \;=\; 28.9$

Instance-Dependent Sample Complexity Lower Bound

Intuition, going back at least to Lai and Robbins (1985)

(Spectacular) difference in behaviour must be due to (spectacular) difference in observations.

Being δ-PAC on $\boldsymbol{\mu}$ *and* $\boldsymbol{\lambda}$ with $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$ requires gathering enough discriminating data.

Instance-Dependent Sample Complexity Lower Bound

Intuition, going back at least to Lai and Robbins (1985)

(Spectacular) difference in behaviour must be due to (spectacular) difference in observations.

Being δ-PAC on $\boldsymbol{\mu}$ *and* $\boldsymbol{\lambda}$ with $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$ requires gathering enough discriminating data.

Define the **alternative** to $\boldsymbol{\mu}$ by $\mathrm{Alt}(\boldsymbol{\mu}) \coloneqq \{\text{bandit } \boldsymbol{\lambda} | i^*(\boldsymbol{\lambda}) \neq i^*(\boldsymbol{\mu})\}$.

Instance-Dependent Sample Complexity Lower Bound

Intuition, going back at least to Lai and Robbins (1985)

(Spectacular) difference in behaviour must be due to (spectacular) difference in observations.

Being δ-PAC on $\boldsymbol{\mu}$ *and* $\boldsymbol{\lambda}$ with $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$ requires gathering enough discriminating data.

Define the **alternative** to $\boldsymbol{\mu}$ by $\mathsf{Alt}(\boldsymbol{\mu}) \coloneqq \{\text{bandit } \boldsymbol{\lambda} | i^*(\boldsymbol{\lambda}) \neq i^*(\boldsymbol{\mu})\}$.

Theorem (Castro 2014; Garivier and Kaufmann 2016)

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

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

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

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

Characteristic Time and Oracle Proportions

Sample complexity lower bound at bandit $\boldsymbol{\mu}$ governed by **characteristic time**

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

Characteristic Time and Oracle Proportions

Sample complexity lower bound at bandit $\boldsymbol{\mu}$ governed by **characteristic time**

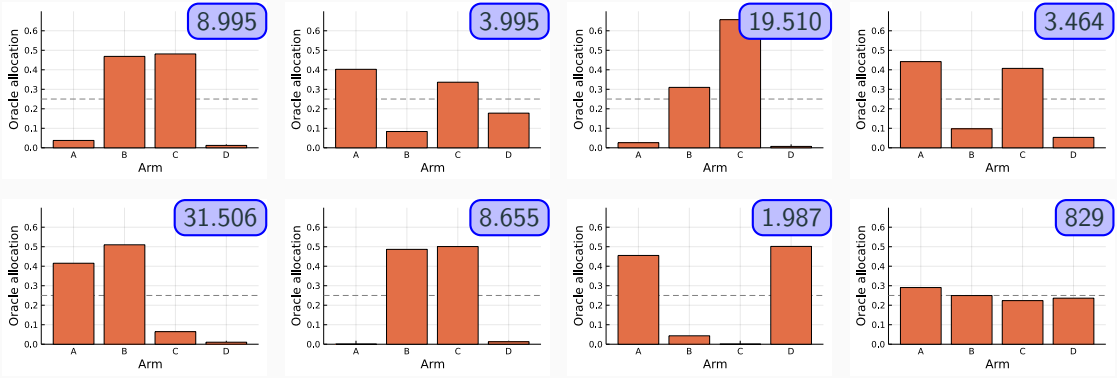
$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq \left(\max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \mathsf{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i \, \mathsf{KL}(\mu_i, \lambda_i) \right)^{-1} \ln \frac{1}{\delta}$$

Moreover, matching algorithms must sample arms with **oracle proportions**

$$\boldsymbol{w}^*(\boldsymbol{\mu}) \coloneqq \arg \max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \mathsf{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i \, \mathsf{KL}(\mu_i, \lambda_i)$$

Outline

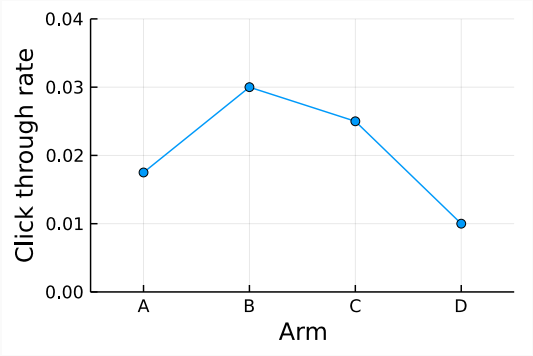
Examples: variations of Best Arm question

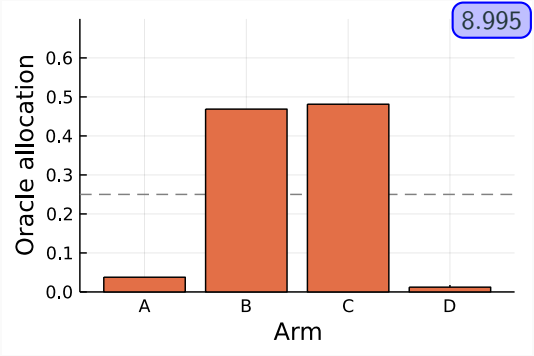


- Sample complexities vastly different between questions
- Optimal allocation depends strongly on the specific question being asked

Best Arm Identification (BAI)

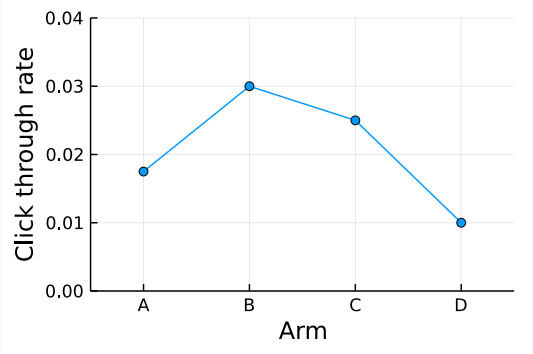
$$i^*(\boldsymbol{\mu}) \coloneqq \arg\max_{a \in \mathcal{A}} \mu_a \qquad \text{where } \mathcal{A} = \{A, B, C, D\}$$

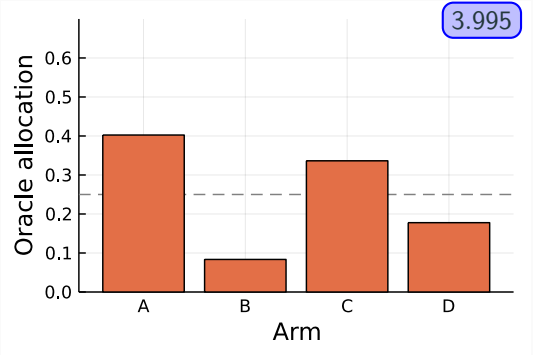




All-Better-than-the-Control (ABC)

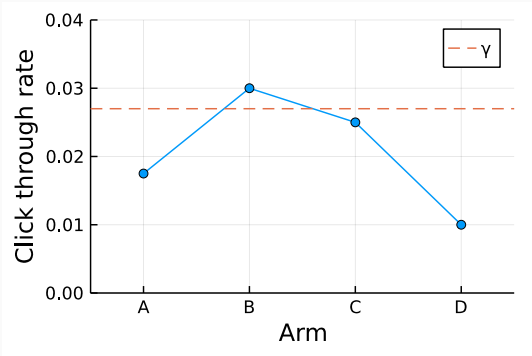
$$i^*(\boldsymbol{\mu}) \coloneqq \left\{ a \in \{B, C, D\} \;\middle|\; \mu_a \geq \mu_A \right\}$$

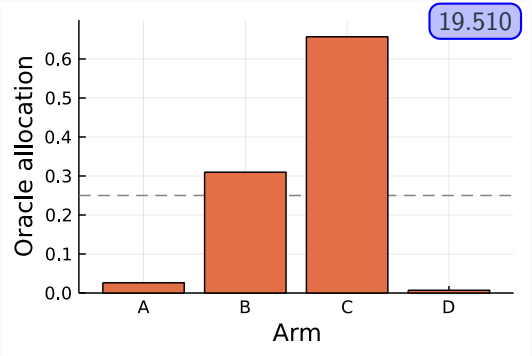




All-Better-than-Threshold

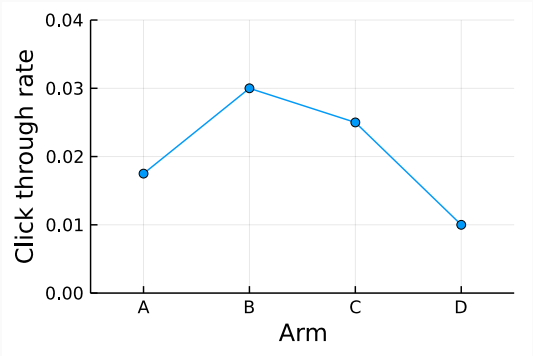
$$i^*(\mu) \coloneqq \{a \in \mathcal{A} \mid \mu_a \geq \gamma\}$$

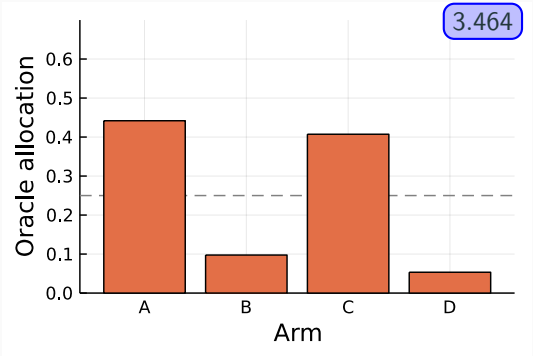




Top-2

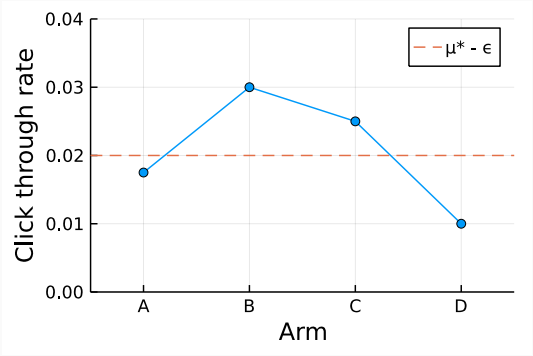
$$i^*(\boldsymbol{\mu}) \coloneqq \left\{ a \in \mathcal{A} \;\middle|\; \mu_a \geq \mu_{(2)} \right\}$$ where $\mu_{(1)} \geq \mu_{(2)} \geq \ldots$

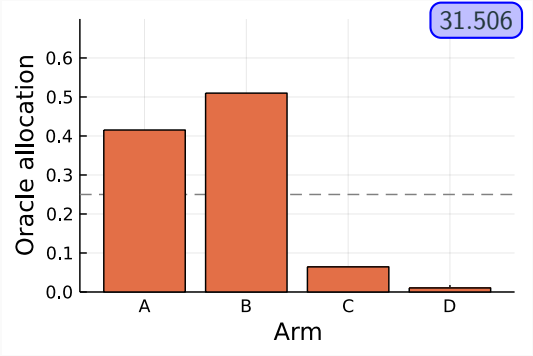




Near-optimal arms

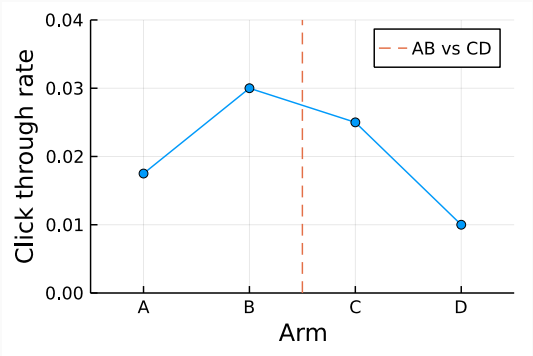
$$i^*(\mu) \coloneqq \left\{ a \in \mathcal{A} \;\middle|\; \mu_a \geq \mu^* - \epsilon \right\} \qquad \text{where } \mu^* = \max_{a \in \mathcal{A}} \mu_a$$

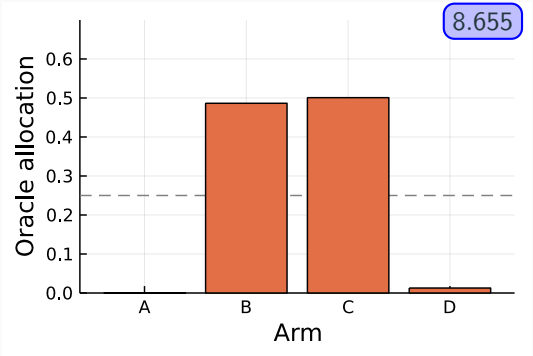




Winning Side

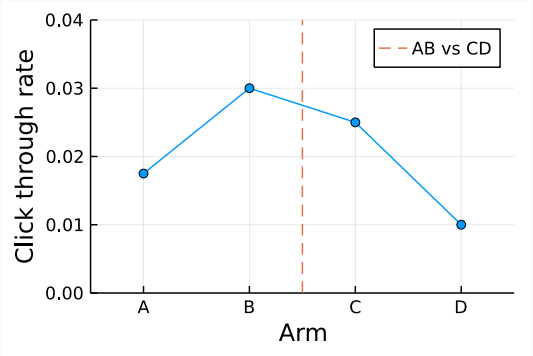
$$i^*(\boldsymbol{\mu}) \;\coloneqq\; \arg\max\left\{\max\left\{\mu_A, \mu_B\right\}, \max\left\{\mu_C, \mu_D\right\}\right\}$$

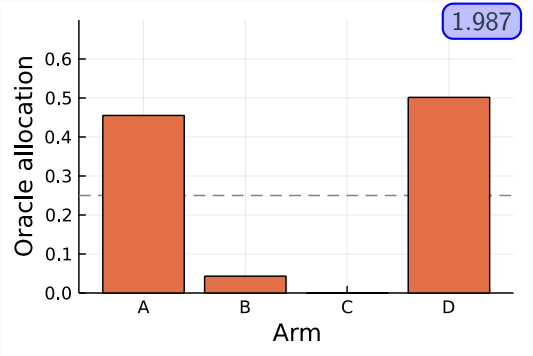




Robust best arm

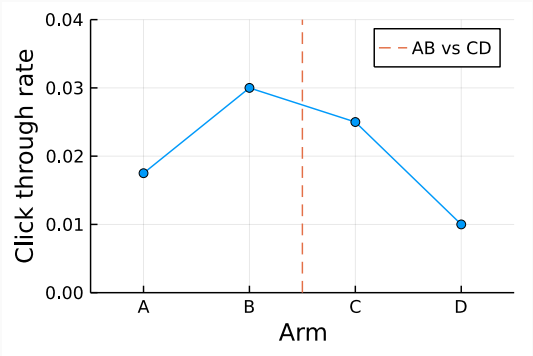
$$i^*(\boldsymbol{\mu}) \;\coloneqq\; \arg\max\left\{\min\left\{\mu_A, \mu_B\right\}, \min\left\{\mu_C, \mu_D\right\}\right\}$$

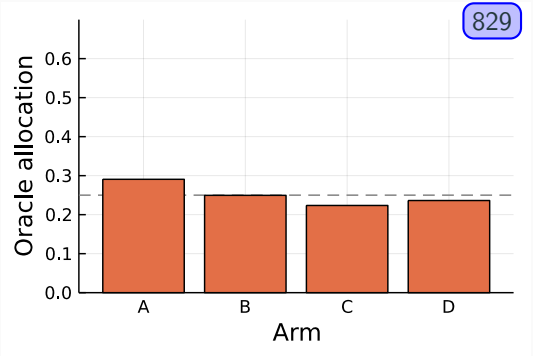




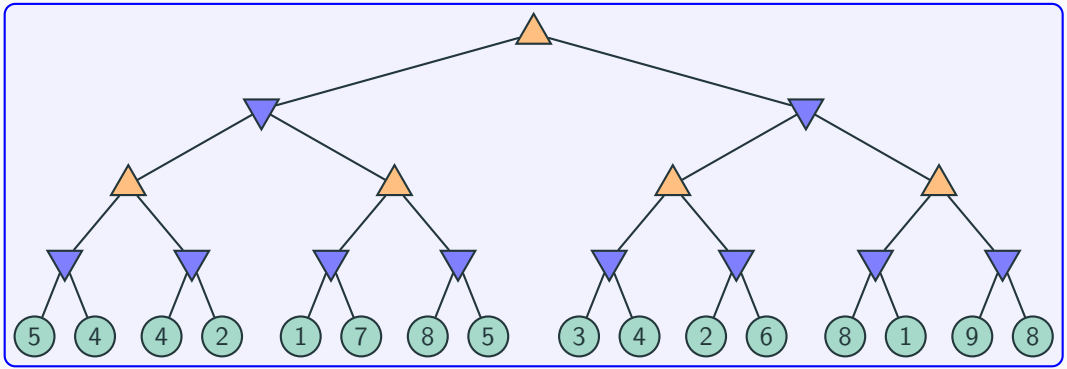
Largest Profit

$$i^*(\boldsymbol{\mu}) \coloneqq \arg\max \{\mu_A - \mu_B, \mu_C - \mu_D\}$$



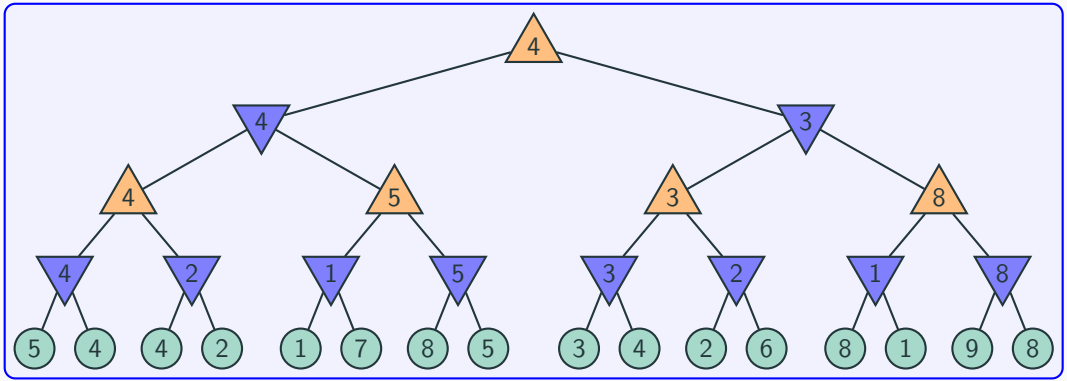


Tree Search Example: Backward Induction Computation



MAX MIN k is Bernoulli($k/10$)

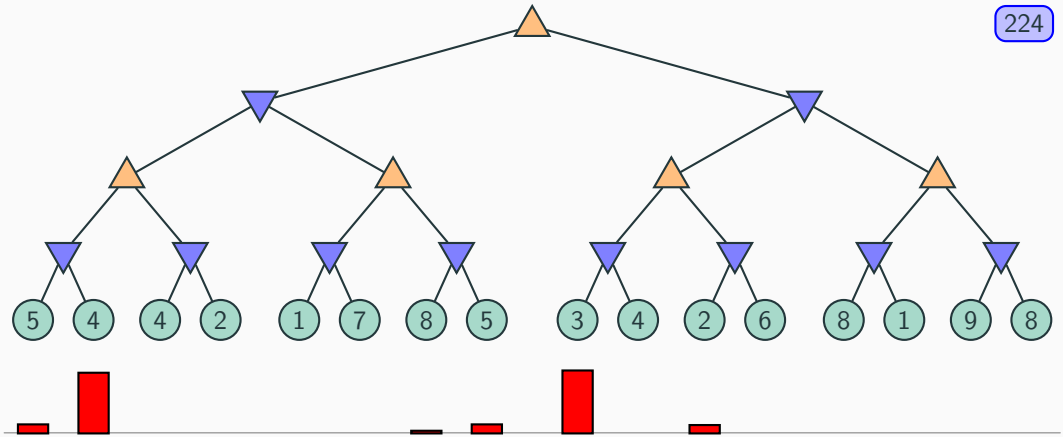
Tree Search Example: Backward Induction Computation



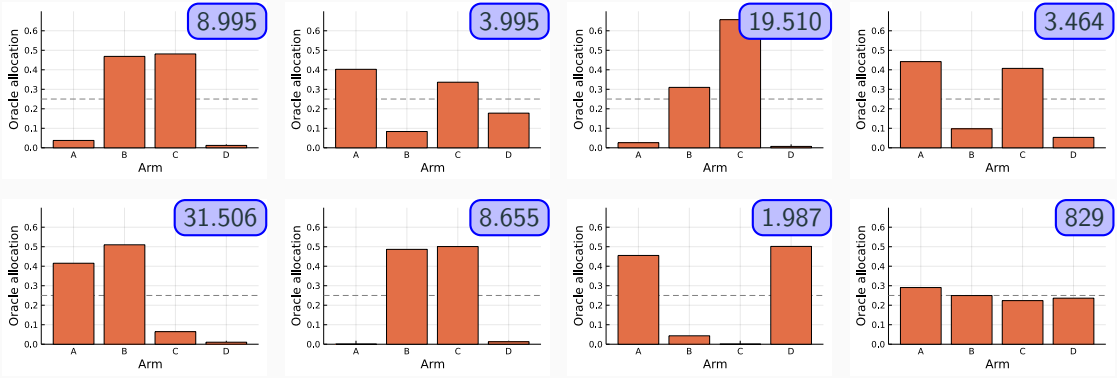
MAX MIN k is Bernoulli($k/10$)

Tree Search Example: Best action at root

$$i^*(\boldsymbol{\mu}) \coloneqq \arg\max_i \min_j \max_k \min_l \mu_{ijkl} \in \{\text{left}, \text{right}\}$$



Overview of Optimal Sampling Allocations



- Sample complexities vastly different between questions
- Optimal allocation depends strongly on the specific question being asked

Outline

Lower Bounds Inspire Strategies

Recall sample complexity lower bound at bandit $\boldsymbol{\mu}$ governed by

$$\max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \mathsf{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i \, \mathsf{KL}(\mu_i, \lambda_i)$$

Lower Bounds Inspire Strategies

Recall sample complexity lower bound at bandit $\boldsymbol{\mu}$ governed by

$$\max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \mathsf{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i \, \mathsf{KL}(\mu_i, \lambda_i)$$

Matching algorithms must sample arms with **argmax** proportions $\boldsymbol{w}^*(\boldsymbol{\mu})$.

Lower Bounds Inspire Strategies

Recall sample complexity lower bound at bandit $\boldsymbol{\mu}$ governed by

$$\max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \mathsf{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i \, \mathsf{KL}(\mu_i, \lambda_i)$$

Matching algorithms must sample arms with **argmax** proportions $\boldsymbol{w}^*(\boldsymbol{\mu})$.

Main issue: Bandit instance $\boldsymbol{\mu}$ **unknown**

Lower Bounds Inspire Strategies

Recall sample complexity lower bound at bandit $\boldsymbol{\mu}$ governed by

$$\max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \mathsf{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i \, \mathsf{KL}(\mu_i, \lambda_i)$$

Matching algorithms must sample arms with **argmax** proportions $\boldsymbol{w}^*(\boldsymbol{\mu})$.

Main issue: Bandit instance $\boldsymbol{\mu}$ **unknown**

Approach: plug in estimate $\hat{\boldsymbol{\mu}}_t$ (Garivier and Kaufmann, 2016)

Saddle Point Techniques

Import online learning techniques

$$\max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \mathsf{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i \, \mathsf{KL}(\mu_i, \lambda_i)$$

Approx. solve saddle point problem iteratively: $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots \to \boldsymbol{w}^*(\boldsymbol{\mu})$

Saddle Point Techniques

Import online learning techniques

$$\max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \mathsf{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i \, \mathsf{KL}(\mu_i, \lambda_i)$$

Approx. solve saddle point problem iteratively: $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots \to \boldsymbol{w}^*(\boldsymbol{\mu})$

Main pipeline (Degenne, Koolen, and Ménard, 2019):

- Pick arm $A_t \sim \boldsymbol{w}_t$
- Plug-in estimate $\hat{\boldsymbol{\mu}}_t$ (so problem is **shifting**).
- Advance the saddle point solver one iteration per bandit interaction.
- Add optimism to gradients to induce exploration ($\hat{\boldsymbol{\mu}}_t \to \boldsymbol{\mu}$).
- Compose regret bound, concentration and optimism to get finite-confidence guarantee.

Saddle Point Techniques

Import online learning techniques

$$\max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \mathsf{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i \, \mathsf{KL}(\mu_i, \lambda_i)$$

Approx. solve saddle point problem iteratively: $\boldsymbol{w}_1, \boldsymbol{w}_2, \ldots \to \boldsymbol{w}^*(\boldsymbol{\mu})$

Main pipeline (Degenne, Koolen, and Ménard, 2019):

- Pick arm $A_t \sim \boldsymbol{w}_t$
- Plug-in estimate $\hat{\boldsymbol{\mu}}_t$ (so problem is **shifting**).
- Advance the saddle point solver one iteration per bandit interaction.
- Add optimism to gradients to induce exploration ($\hat{\boldsymbol{\mu}}_t \to \boldsymbol{\mu}$).
- Compose regret bound, concentration and optimism to get finite-confidence guarantee.

Theorem (Instance-Optimality)

For every $\delta \in (0,1)$, the sample complexity is bounded by $\mathbb{E}_{\boldsymbol{\mu}}[\tau] \leq T^*(\boldsymbol{\mu}) \ln \frac{1}{\delta} + o(\ln \frac{1}{\delta})$.

Sketch of Argument

As long as we do not stop (and concentration holds):

$$
\begin{aligned}
\ln \frac{1}{\delta} &\geq \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\hat{\boldsymbol{\mu}}_t)} \sum_{k=1}^{K} N_t^k \, \mathrm{KL}(\mu^k, \lambda^k) && \text{(stop rule)} \\
&\approx \inf_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{s=1}^{t} \sum_{k=1}^{K} w_s^k \, \mathrm{KL}(\mu^k, \lambda^k) && \text{(tracking)} \\
&\geq \sum_{s=1}^{t} \sum_{k=1}^{K} w_s^k \, \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}_s} \mathrm{KL}(\mu^k, \lambda^k) - R_t^{\lambda} && \text{(regret } \boldsymbol{\lambda}\text{)} \\
&\geq \max_k \sum_{s=1}^{t} \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}_s} \mathrm{KL}(\mu^k, \lambda^k) - R_t^{\lambda} - R_t^k && \text{(regret } k\text{)} \\
&\geq t \inf_{\boldsymbol{q} \in \mathcal{P}(\mathrm{Alt}(\boldsymbol{\mu}))} \max_k \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}} \mathrm{KL}(\mu^k, \lambda^k) - O(\sqrt{t}) \\
&= t \, \boxed{\max_{\boldsymbol{w} \in \triangle_K} \min_{\boldsymbol{\lambda} \in \mathrm{Alt}(\boldsymbol{\mu})} \sum_{i=1}^{K} w_i \, \mathrm{KL}(\mu_i, \lambda_i)} - O(\sqrt{t})
\end{aligned}
$$

Outline

Conclusion

Canonical Path to Instance Optimality

- State-of-the-art performance in practise (some problems)
  - Best Arm Identification
  - All-better-than-Control
  - Minimax Game Tree Search

Conclusion

Canonical Path to Instance Optimality

- State-of-the-art performance in practise (some problems)
  - Best Arm Identification
  - All-better-than-Control
  - Minimax Game Tree Search
- Different ("fresh") structure compared to other techniques (confidence intervals, elimination, Thompson sampling, ...)

Conclusion

Canonical Path to Instance Optimality

- State-of-the-art performance in practise (some problems)
  - Best Arm Identification
  - All-better-than-Control
  - Minimax Game Tree Search
- Different ("fresh") structure compared to other techniques (confidence intervals, elimination, Thompson sampling, ...)
- Reduces identification problems to online learning (efficiently computing gradients/best response).

Conclusion

Canonical Path to Instance Optimality

- State-of-the-art performance in practise (some problems)
  - Best Arm Identification
  - All-better-than-Control
  - Minimax Game Tree Search
- Different ("fresh") structure compared to other techniques (confidence intervals, elimination, Thompson sampling, ...)
- Reduces identification problems to online learning (efficiently computing gradients/best response).
- Foundation for
  - Linear bandits
  - Contextual bandits
  - Optimal policy learning (reinforcement learning)

Wish list:

- Instance optimality for (ϵ, δ)-case currently deeply asymptotic

Thanks!

References i

- Castro, R. M. (Nov. 2014). "Adaptive sensing performance lower bounds for sparse signal detection and support estimation". In: **Bernoulli** 20.4, pp. 2217–2246.
- Chen, S., T. Lin, I. King, M. Lyu, and W. Chen (2014). "Combinatorial Pure Exploration of Multi-Armed Bandits". In: **Advances in Neural Information Processing Systems**.
- Degenne, R., W. M. Koolen, and P. Ménard (Dec. 2019). "Non-Asymptotic Pure Exploration by Solving Games". In: **Advances in Neural Information Processing Systems (NeurIPS) 32**. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Curran Associates, Inc., pp. 14492–14501.
- Even-Dar, E., S. Mannor, and Y. Mansour (2002). "PAC Bounds for Multi-armed Bandit and Markov Decision Processes". In: **Computational Learning Theory, 15th Annual Conference on Computational Learning Theory, COLT 2002, Sydney, Australia, July 8-10, 2002, Proceedings**. Ed. by J. Kivinen and R. H. Sloan. Vol. 2375. Lecture Notes in Computer Science. Springer, pp. 255–270.

References ii

- Garivier, A. and E. Kaufmann (2016). "Optimal Best arm Identification with Fixed Confidence". In: **Proceedings of the 29th Conference On Learning Theory (COLT)**.
- Lai, T. L. and H. Robbins (1985). "Asymptotically efficient adaptive allocation rules". In: **Advances in Applied Mathematics** 6.1, pp. 4–22.
- Russac, Y., C. Katsimerou, D. Bohle, O. Cappé, A. Garivier, and W. M. Koolen (Dec. 2021). "A/B/n Testing with Control in the Presence of Subpopulations". In: **Advances in Neural Information Processing Systems (NeurIPS) 34**.
- Teraoka, K., K. Hatano, and E. Takimoto (2014). "Efficient Sampling Method for Monte Carlo Tree Search Problem". In: **IEICE Transactions** 97-D.3, pp. 392–398.
- Yu, J. Y. and E. Nikolova (2013). "Sample complexity of risk-averse bandit-arm selection". In: **Twenty-Third International Joint Conference on Artificial Intelligence**.
- Yue, Y., J. Broder, R. Kleinberg, and T. Joachims (2012). "The K-armed dueling bandits problem". In: **Journal of Computer and System Sciences, Special Issue for COLT 2009**, pp. 1538–1566.