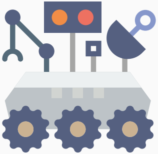
Pure Exploration Problems
Information Theory and Equilibria



Wouter M. Koolen

CWI
Centrum Wiskunde & Informatica

Tuesday 19th April, 2022

Outline

1. Problems
2. Good Learning Strategies
3. Lower Bounds: Information Theory
4. Design of Algorithms: Equilibria
5. Conclusion

This talk is about

Understanding Interactive Learning

What if the learning system can decide which data to collect?

This talk is about

Understanding Interactive Learning

What if the learning system can decide which data to collect?

- **How many** experiments are needed?
- **Which** experiments to pick?
- **How** to learn from the data collected?

This talk is about

Understanding Interactive Learning

What if the learning system can decide which data to collect?

- **How many** experiments are needed?
- **Which** experiments to pick?
- **How** to learn from the data collected?

Today: Active Sequential Hypothesis Testing. Applications to

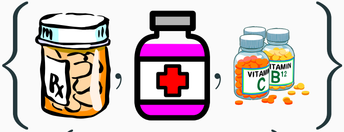
- Medical testing
- A/B testing (e-commerce)
- Simulation-based planning
- Reinforcement learning
- ...

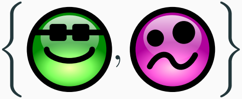
Stochastic Bandit

Experiments $\{$ , ,  $\}$

Outcomes $\{$ ,  $\}$

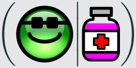
Stochastic Bandit

Experiments 

Outcomes 

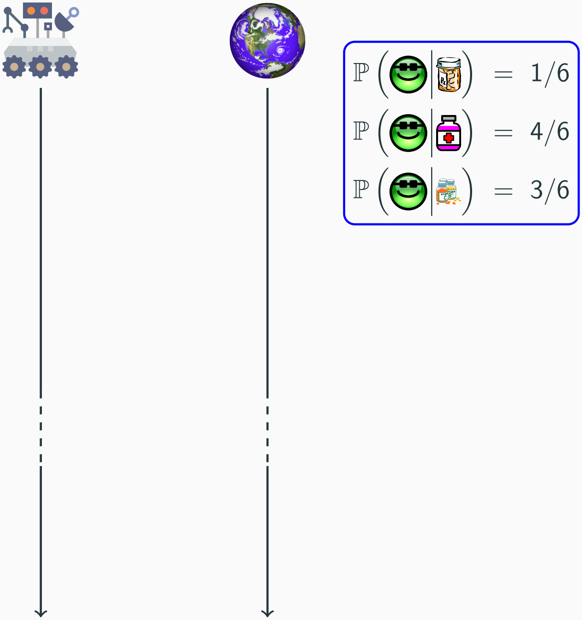
Instance (Unknown)

$\mathbb{P}$  $= 1/6$

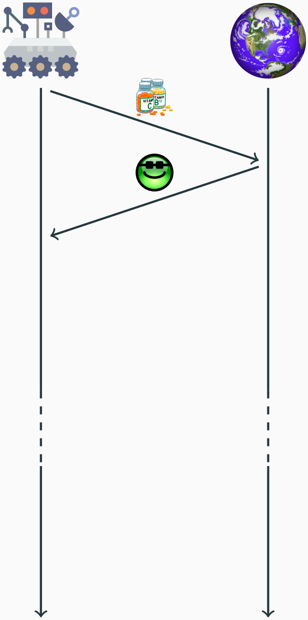
$\mathbb{P}$  $= 4/6$

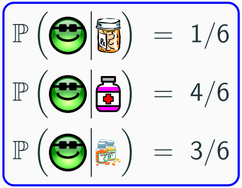
$\mathbb{P}$  $= 3/6$

Best Arm Identification Interaction

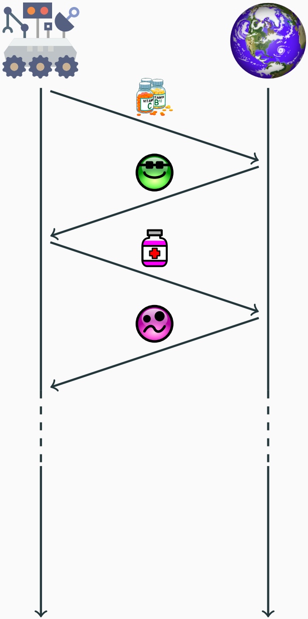


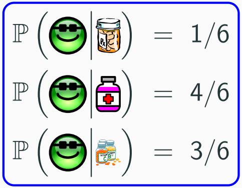
Best Arm Identification Interaction



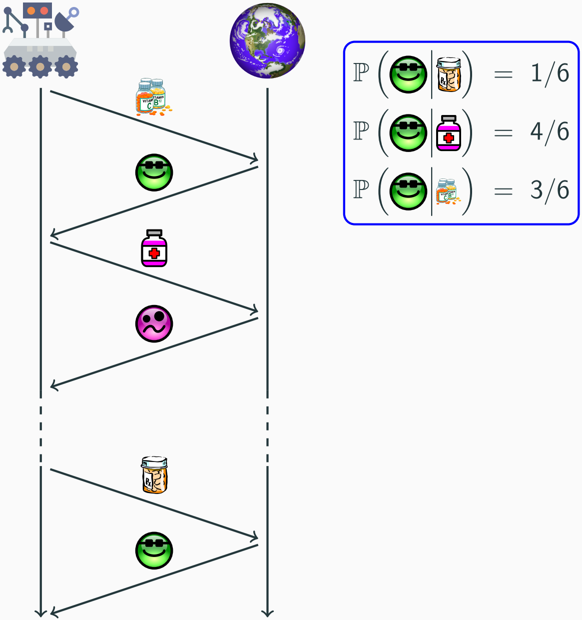


Best Arm Identification Interaction

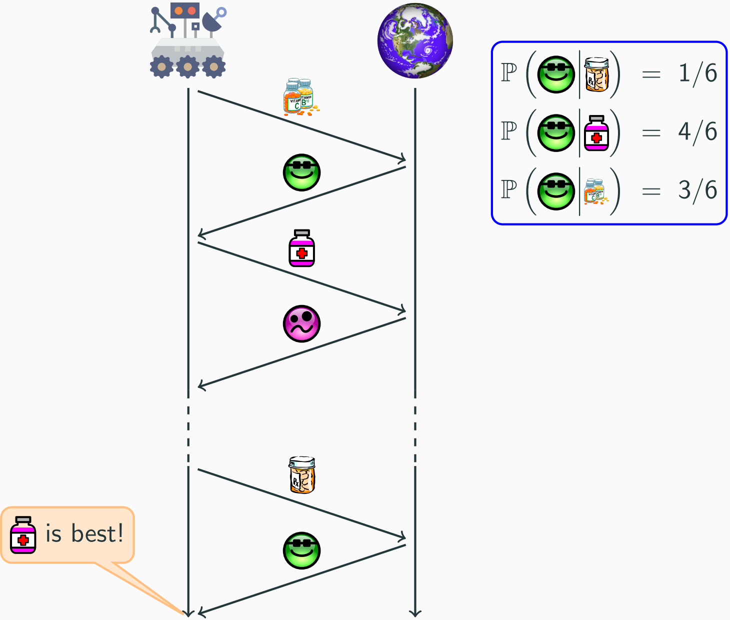




Best Arm Identification Interaction



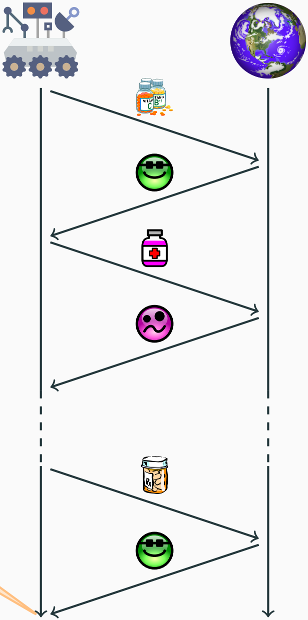
Best Arm Identification Interaction

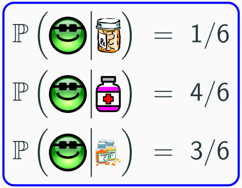


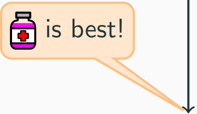
Best Arm Identification Interaction

Desiderata

- Efficient: few samples
- Reliable: correct whp







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)

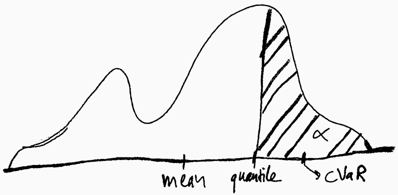
Which arm has the highest α-quantile

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



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

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

$$\mathbb{P}_{\boldsymbol{\mu}}\Big\{\underbrace{\tau < \infty \text{ and } \hat{I} \neq \arg\max_i \mu_i}_{\text{a mistake}}\Big\} \leq \delta \quad \text{for all } \boldsymbol{\mu} \in [0,1]^K.$$

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

$$\mathbb{P}_{\boldsymbol{\mu}}\Big\{\underbrace{\tau < \infty \text{ and } \hat{I} \neq \arg\max_i \mu_i}_{\text{a mistake}}\Big\} \leq \delta \quad \text{for all } \boldsymbol{\mu} \in [0,1]^K.$$

Definition

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

Best Arm Identification

BAI-MAB Protocol

for $t = 1, 2, \dots$ **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

$$\mathbb{P}_{\boldsymbol{\mu}}\Big\{\underbrace{\tau < \infty \text{ and } \hat{I} \neq \arg\max_i \mu_i}_{\text{a mistake}}\Big\} \leq \delta \quad \text{for all } \boldsymbol{\mu} \in [0,1]^K.$$

Definition

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

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

Pure Exploration Prototypical Solution

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

Pure Exploration Prototypical Solution

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

Fancy Algorithm(δ)

Stop when ...

Sample arm $A_t = \dots$

Recommend $\hat{I} = \dots$

Pure Exploration Prototypical Solution

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

Fancy Algorithm(δ)

Stop when $\ldots$

Sample arm $A_t = \ldots$

Recommend $\hat{I} = \ldots$

Theorem (safe)

Fancy Algorithm(δ) *is* δ*-PAC*

Pure Exploration Prototypical Solution

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

Fancy Algorithm(δ)

Stop when ...

Sample arm $A_t = \ldots$

Recommend $\hat{I} = \ldots$

Theorem (safe)

Fancy Algorithm(δ) is δ-PAC

Theorem (comput. eff.)

... runs in time $O(\ldots)$

Pure Exploration Prototypical Solution

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

Fancy Algorithm(δ)

Stop when ...

Sample arm $A_t = \ldots$

Recommend $\hat{I} = \ldots$

Theorem (safe)

Fancy Algorithm(δ) is δ-PAC

Theorem (comput. eff.)

... runs in time $O(\ldots)$

Theorem (statistic. eff.)

... has sample complexity

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \;\leq\; f(\boldsymbol{\mu}) \ln \tfrac{1}{\delta} + o(\ln \tfrac{1}{\delta}).$$

Pure Exploration Prototypical Solution

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

Fancy Algorithm(δ)

Stop when ...

Sample arm $A_t = \dots$

Recommend $\hat{I} = \dots$

Theorem (lower bd)

Any δ-PAC algorithm needs sample complexity at least

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

Theorem (safe)

Fancy Algorithm(δ) is δ-PAC

Theorem (comput. eff.)

... runs in time $O(\dots)$

Theorem (statistic. eff.)

... has sample complexity

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \;\leq\; f(\boldsymbol{\mu}) \ln \tfrac{1}{\delta} + o(\ln \tfrac{1}{\delta}).$$

Outline

1. Problems
2. Good Learning Strategies
3. Lower Bounds: Information Theory
4. Design of Algorithms: Equilibria
5. Conclusion

Instance-Dependent Sample Complexity Lower Bound

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

A (spectacular) difference in behaviour must be due to a (spectacular) difference in the observations.

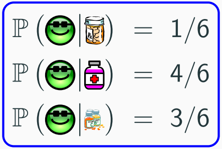
So being δ-PAC on $\boldsymbol{\mu}$ and also on $\boldsymbol{\lambda}$ with $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$ requires collecting enough discriminating information.

Instance-Dependent Sample Complexity Lower Bound

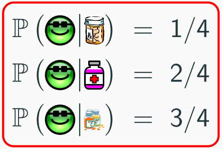
Intuition, going back at least to Lai and Robbins (1985)

A (spectacular) difference in behaviour must be due to a (spectacular) difference in the observations.

So being δ-PAC on $\boldsymbol{\mu}$ and also on $\boldsymbol{\lambda}$ with $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$ requires collecting enough discriminating information.



$\leftarrow i^*$ (second row)



$\leftarrow i^*$ (third row)

If δ-**PAC algorithm** samples t rounds with arm freqs. 1/5, 3/5, 2/5, then

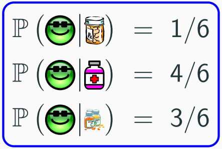
$$t\,\frac{1}{5}\,\mathrm{KL}\left(\frac{1}{6},\frac{1}{4}\right)+t\,\frac{3}{5}\,\mathrm{KL}\left(\frac{4}{6},\frac{2}{4}\right)+t\,\frac{2}{5}\,\mathrm{KL}\left(\frac{3}{6},\frac{3}{4}\right) \;\geq\; \mathrm{KL}(\delta,1-\delta)\approx \ln\frac{1}{\delta}$$

Instance-Dependent Sample Complexity Lower Bound

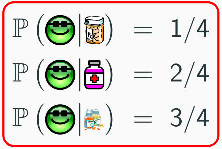
Intuition, going back at least to Lai and Robbins (1985)

A (spectacular) difference in behaviour must be due to a (spectacular) difference in the observations.

So being δ-PAC on $\boldsymbol{\mu}$ and also on $\boldsymbol{\lambda}$ with $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$ requires collecting enough discriminating information.



$\leftarrow i^*$



$\leftarrow i^*$

If **δ-PAC algorithm** samples t rounds with arm freqs. 1/5, 3/5, 2/5, then

$$t\,\frac{1}{5}\,\mathrm{KL}\left(\frac{1}{6},\frac{1}{4}\right) + t\,\frac{3}{5}\,\mathrm{KL}\left(\frac{4}{6},\frac{2}{4}\right) + t\,\frac{2}{5}\,\mathrm{KL}\left(\frac{3}{6},\frac{3}{4}\right) \;\geq\; \mathrm{KL}(\delta, 1-\delta) \approx \ln\frac{1}{\delta}$$

At typical $\delta = 0.1$: $\qquad 0.0956\,t \;\geq\; 1.757 \qquad t \;\geq\; \dfrac{1.757}{0.0956} \;=\; 18.4$

Instance-Dependent Sample Complexity Lower Bound

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

A (spectacular) difference in behaviour must be due to a (spectacular) difference in the observations.

So being δ-PAC on $\boldsymbol{\mu}$ and also on $\boldsymbol{\lambda}$ with $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$ requires collecting enough discriminating information.

Instance-Dependent Sample Complexity Lower Bound

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

A (spectacular) difference in behaviour must be due to a (spectacular) difference in the observations.

So being δ-PAC on $\boldsymbol{\mu}$ and also on $\boldsymbol{\lambda}$ with $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$ requires collecting enough discriminating information.

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

Instance-Dependent Sample Complexity Lower Bound

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

A (spectacular) difference in behaviour must be due to a (spectacular) difference in the observations.

So being δ-PAC on $\boldsymbol{\mu}$ and also on $\boldsymbol{\lambda}$ with $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$ requires collecting enough discriminating information.

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

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*

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

Outline

1. Problems
2. Good Learning Strategies
3. Lower Bounds: Information Theory
4. Design of Algorithms: Equilibria
5. Conclusion

Lower Bounds Inspire Strategies

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

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

Lower Bounds Inspire Strategies

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

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

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

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

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

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

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

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

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

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

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

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

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)

Thanks!

References i

Agrawal, S., W. M. Koolen, and S. Juneja (Dec. 2021). "Optimal Best-Arm Identification Methods for Tail-Risk Measures". In: **Advances in Neural Information Processing Systems (NeurIPS) 34**.

Castro, R. M. (Nov. 2014). "Adaptive sensing performance lower bounds for sparse signal detection and support estimation". In: **Bernoulli** 20.4, pp. 2217–2246.

Degenne, R. and W. M. Koolen (Dec. 2019). "Pure Exploration with Multiple Correct Answers". 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. 14591–14600.

- 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.
- Degenne, R., H. Shao, and W. M. Koolen (July 2020). "Structure Adaptive Algorithms for Stochastic Bandits". In: **Proceedings of the 37th International Conference on Machine Learning (ICML)**.
- 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 iii

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.

Yu, J. Y. and E. Nikolova (2013). “Sample complexity of risk-averse bandit-arm selection”. In: **Twenty-Third International Joint Conference on Artificial Intelligence**.