The Pure Exploration Renaissance

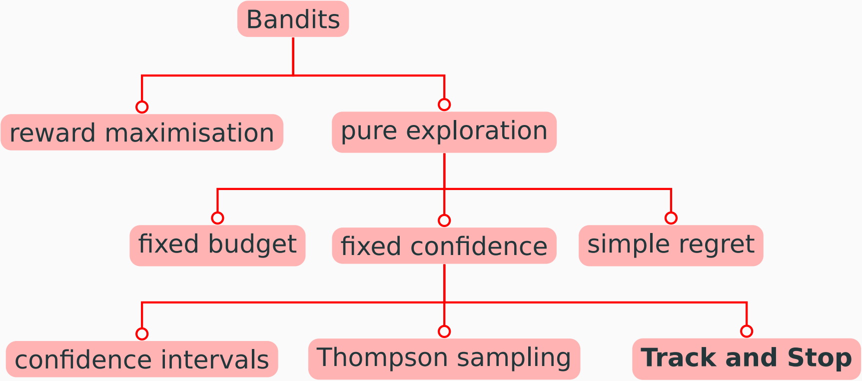
Deep Dive @ Booking.com

Wouter Koolen

December 10th, 2020

Centrum Wiskunde & Informatica

Lay of the Land



Context

Pure Exploration is statistical hypothesis testing on steroids:

- Multiple
- Composite
- Sequential
- Active

Menu

- Introduce Pure Exploration problems.
- Sketch the GLR stopping rule
- Sketch the TaS sampling rule
- Highlight some recent lessons learned.

Pure Exploration Introduction

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

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

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 $t = 1, 2, \ldots$ **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 $\boldsymbol{\mu}$.

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

Best Arm Identification, prototypical solution

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

Best Arm Identification, prototypical solution

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

Fancy Algorithm(δ)

Stop when ...

Sample arm $A_t = \ldots$

Recommend $\hat{I} = \ldots$

Best Arm Identification, 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 (safe)

Fancy Algorithm(δ) *is* δ*-PAC*

Best Arm Identification, 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 (safe)

Fancy Algorithm(δ) *is* δ*-PAC*

Theorem (comput. eff.)

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

Best Arm Identification, 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

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

Best Arm Identification, 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 (lower bd)

Any δ-PAC algorithm needs sample complexity at least

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

Theorem (safe)

Fancy Algorithm(δ) is δ-PAC

Theorem (comput. eff.)

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

Theorem (statistic. eff.)

$\ldots$ has sample complexity

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

Many Variations of BAI Problem in Literature

Assumptions about arm distributions

- Exponential Family:
  Gaussian, Gamma, Poisson, Geometric, ...
- Non-parametric:
  bounded support, sub-Gaussian, $(1+\epsilon)^{\text{th}}$ moment, ...

Many Variations of BAI Problem in Literature

Assumptions about arm distributions

- Exponential Family:
  Gaussian, Gamma, Poisson, Geometric, ...
- Non-parametric:
  bounded support, sub-Gaussian, $(1+\epsilon)^{\text{th}}$ moment, ...

Assume prior knowledge of structured bandit model class

- Linear, Lipschitz, Sparse, Categorical, Unimodal, ...

Many Variations of BAI Problem in Literature

Assumptions about arm distributions

- Exponential Family:
  Gaussian, Gamma, Poisson, Geometric, ...
- Non-parametric:
  bounded support, sub-Gaussian, $(1+\epsilon)^{\text{th}}$ moment, ...

Assume prior knowledge of structured bandit model class

- Linear, Lipschitz, Sparse, Categorical, Unimodal, ...

Identify a near-optimal arm: Learner is (ϵ, δ)-PAC if

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

Many Variations of BAI Problem in Literature

Assumptions about arm distributions

- Exponential Family:
  Gaussian, Gamma, Poisson, Geometric, ...
- Non-parametric:
  bounded support, sub-Gaussian, $(1+\epsilon)^{\text{th}}$ moment, ...

Assume prior knowledge of structured bandit model class

- Linear, Lipschitz, Sparse, Categorical, Unimodal, ...

Identify a near-optimal arm: Learner is (ϵ, δ)-PAC if

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

Feedback graphs (semi-bandit, ...)

A/B Testing Problems

A/B Testing Problems

Input: baseline arm μ_0, candidates $\mu_1, \dots, \mu_K$.

Problem (Best Arm Identification)

$$i^*(\boldsymbol{\mu}) = \underset{i \in \{0,\dots,K\}}{\arg\max}\ \mu_i$$

A/B Testing Problems

Input: baseline arm μ_0, candidates $\mu_1, \ldots, \mu_K$.

Problem (Best Arm Identification)

$$i^*(\boldsymbol{\mu}) = \underset{i \in \{0,\ldots,K\}}{\arg\max}\ \mu_i$$

Problem (A/B test, decision version)

$$i^*(\boldsymbol{\mu}) = \mathbf{1}\left\{\max_{i \in \{1,\ldots,K\}} \mu_i > \mu_0\right\}$$

A/B Testing Problems

Input: baseline arm μ_0, candidates $\mu_1, \ldots, \mu_K$.

Problem (A/B test, identification version)

$$i^*(\boldsymbol{\mu}) = \begin{cases} \{0\} & \mu_0 > \max_{i \in \{1,\ldots,K\}} \mu_i \\ \{i \in \{1,\ldots,K\} | \mu_i > \mu_0\} & \textit{o.w.} \end{cases}$$

this is a multiple-answer problem

A/B Testing Problems

Input: baseline arm μ_0, candidates $\mu_1, \ldots, \mu_K$.

Problem (A/B test, identification version)

$$i^*(\boldsymbol{\mu}) = \begin{cases} \{0\} & \mu_0 > \max_{i \in \{1,\ldots,K\}} \mu_i \\ \{i \in \{1,\ldots,K\} | \mu_i > \mu_0\} & o.w. \end{cases}$$

this is a multiple-answer problem

Problem (A/B test, thresholding version)

$$i^*(\boldsymbol{\mu}) = \{i \in \{1,\ldots,K\} | \mu_i > \mu_0\}$$

this is a ***set-valued*** *single-answer problem*

More Advanced Testing Problems

Input: arms $\mu_{i,j}$, $i \in \{1, \ldots, K\}$, $j \in \{1, \ldots, M\}$

Problem (Maximin Action Identification)

$$i^*(\boldsymbol{\mu}) = \operatorname*{arg\,max}_{i \in \{1,\ldots,K\}} \min_{j \in \{1,\ldots,M\}} \mu_{i,j}$$

GLRT Stopping

Stopping

When can we stop?

Stopping

When can we stop?

When can we stop and give answer $\hat{\imath}$?

Stopping

When can we stop?

When can we stop and give answer $\hat{\imath}$?

There is no plausible bandit model $\boldsymbol{\lambda}$ on which $\hat{\imath}$ is wrong.

Stopping

When can we stop?

When can we stop and give answer $\hat{\imath}$?

There is no plausible bandit model $\boldsymbol{\lambda}$ on which $\hat{\imath}$ is wrong.

Definition

Generalized Likelihood Ratio (GLR) measure of evidence

$$\mathrm{GLR}_n(\hat{\imath}) \;:=\; \ln \frac{\sup_{\boldsymbol{\mu}:\hat{\imath}\in i^*(\boldsymbol{\mu})} P\left(X^n | A^n, \boldsymbol{\mu}\right)}{\sup_{\boldsymbol{\lambda}:\hat{\imath}\notin i^*(\boldsymbol{\lambda})} P\left(X^n | A^n, \boldsymbol{\lambda}\right)}$$

Stopping

When can we stop?

When can we stop and give answer $\hat{\imath}$?

There is no plausible bandit model $\boldsymbol{\lambda}$ on which $\hat{\imath}$ is wrong.

Definition

Generalized Likelihood Ratio (GLR) measure of evidence

$$\mathrm{GLR}_n(\hat{\imath}) \;:=\; \ln \frac{\sup_{\boldsymbol{\mu}:\hat{\imath}\in i^*(\boldsymbol{\mu})} P\left(X^n \middle| A^n, \boldsymbol{\mu}\right)}{\sup_{\boldsymbol{\lambda}:\hat{\imath}\notin i^*(\boldsymbol{\lambda})} P\left(X^n \middle| A^n, \boldsymbol{\lambda}\right)}$$

Idea: stop when $\mathrm{GLR}_n(\hat{\imath})$ is big for some answer $\hat{\imath}$.

GLR Stopping

For any plausible answer $\hat{\imath} \in i^*(\hat{\boldsymbol{\mu}}(n))$, the GLR_n simplifies to

$$\mathrm{GLR}_n(\hat{\imath}) \;=\; \inf_{\boldsymbol{\lambda} : \hat{\imath} \notin i^*(\boldsymbol{\lambda})} \sum_{a=1}^{K} N_a(n)\, \mathrm{KL}(\hat{\mu}_a(n), \lambda_a)$$

where $\mathrm{KL}(x, y)$ is the Kullback-Leibler divergence in the exponential family.

GLR Stopping, Treshold

What is a suitable threshold for GLR_n so that we do not make mistakes?

GLR Stopping, Treshold

What is a suitable threshold for GLR_n so that we do not make mistakes?

A mistake is made when $\text{GLR}_n(\hat{\imath})$ is big while $\hat{\imath} \notin i^*(\boldsymbol{\mu})$.

GLR Stopping, Treshold

What is a suitable threshold for GLR_n so that we do not make mistakes?

A mistake is made when $\text{GLR}_n(\hat{\imath})$ is big while $\hat{\imath} \notin i^*(\boldsymbol{\mu})$.

But then

$$\text{GLR}_n(\hat{\imath}) = \inf_{\boldsymbol{\lambda}:\hat{\imath} \notin i^*(\boldsymbol{\lambda})} \sum_{a=1}^{K} N_a(n) \, \text{KL}(\hat{\mu}_a(n), \lambda_a) \leq \boxed{\sum_{a=1}^{K} N_a(n) \, \text{KL}(\hat{\mu}_a(n), \mu_a)}.$$

GLR Stopping, Treshold

What is a suitable threshold for GLR_n so that we do not make mistakes?

A mistake is made when $\text{GLR}_n(\hat{\imath})$ is big while $\hat{\imath} \notin i^*(\boldsymbol{\mu})$.

But then

$$\text{GLR}_n(\hat{\imath}) = \inf_{\boldsymbol{\lambda}: \hat{\imath} \notin i^*(\boldsymbol{\lambda})} \sum_{a=1}^K N_a(n) \, \text{KL}(\hat{\mu}_a(n), \lambda_a) \le \boxed{\sum_{a=1}^K N_a(n) \, \text{KL}(\hat{\mu}_a(n), \mu_a)}.$$

Good anytime deviation inequalities exist for that upper bound.

Theorem (Kaufmann and Koolen, 2018)

$$\mathbb{P}\left(\exists n : \boxed{\sum_{a=1}^K N_a(n) \, \text{KL}(\hat{\mu}_a(n), \mu_a)} - \sum_n \ln \ln N_a(n) \ge C(K, \delta) \right) \le \delta$$

for $C(K, \delta) \approx \ln \frac{1}{\delta} + K \ln \ln \frac{1}{\delta}$.

GLR Stopping Conclusion

- Tight criterion for stopping.
- We will see asymptotically matches lower bound.
- Often relatively easy to compute
- Typically reduces sample complexity by factor ≈ 2 for BAI problems compared to confidence-interval based stopping (Why? Confidence region)
- Performs very well in practise

Track-and-Stop Algorithm Template

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 $\text{Alt}(\boldsymbol{\mu}) \coloneqq \{\text{bandit } \boldsymbol{\lambda} | i^*(\boldsymbol{\lambda}) \neq i^*(\boldsymbol{\mu})\}$.

Instance-Dependent Sample Complexity Lower Bound

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

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 $\text{Alt}(\boldsymbol{\mu}) \coloneqq \{\text{bandit } \boldsymbol{\lambda} | i^*(\boldsymbol{\lambda}) \neq i^*(\boldsymbol{\mu})\}$.

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

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

Example

$K = 5$ Bernoulli arms, $\boldsymbol{\mu} = (0.4, 0.3, 0.2, 0.1, 0.0)$.

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

At confidence $\delta = 0.05$ we have $\ln \frac{1}{\delta} = 3.0$ and hence $\mathbb{E}_{\boldsymbol{\mu}}[\tau] \geq 601.2$.

Lower Bounds Inspire Strategies

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

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

Lower Bounds Inspire Strategies

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

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

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

Lower Bounds Inspire Strategies

Track-and-Stop scheme (Garivier and Kaufmann, 2016)

At each time step t

- compute plug-in **oracle solution** $\boldsymbol{w}^*(\hat{\boldsymbol{\mu}}_t)$
- sample arm A_t to track (ensure $N_a(t)/t \to w_a^*(\hat{\mu}_t)$)
- force exploration to ensure $\hat{\boldsymbol{\mu}}_t \to \boldsymbol{\mu}$.

Stop using GLRT stopping rule, recommend single non-rejected arm.

Lower Bounds Inspire Strategies

Track-and-Stop scheme (Garivier and Kaufmann, 2016)

At each time step t

- compute plug-in **oracle solution** $\boldsymbol{w}^*(\hat{\boldsymbol{\mu}}_t)$
- sample arm A_t to track (ensure $N_a(t)/t \to w_a^*(\hat{\mu}_t)$)
- force exploration to ensure $\hat{\boldsymbol{\mu}}_t \to \boldsymbol{\mu}$.

Stop using GLRT stopping rule, recommend single non-rejected arm.

Theorem (Asymptotic Instance-Optimality)

The sample complexity of Track-and-Stop for BAI is bounded by

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

Analysis

Convergence $\hat{\boldsymbol{\mu}}_t \to \boldsymbol{\mu}$ and continuity of $\boldsymbol{\mu} \mapsto \boldsymbol{w}^*(\boldsymbol{\mu})$ ensures sampling proportion $N_a(t)/t$ approximates oracle $w_a^*(\boldsymbol{\mu})$.

Asymptotic Optimality

Why interested in asymptotically optimal algorithms?

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

Asymptotic Optimality

Why interested in asymptotically optimal algorithms?

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

Asymptotic Optimality

Why interested in asymptotically optimal algorithms?

- State-of-the-art performance in practise (some problems)
  - Best Arm Identification
  - Minimax Game Tree Search
- Different (“fresh”) structure compared to other techniques (confidence intervals, elimination, Thompson sampling, ...)
- TaS reduces the identification problem to efficiently computing $\boldsymbol{w}^*(\boldsymbol{\mu})$.

TaS Conclusion

- “Track” instance-optimal sampling rule.
- Often relatively easy to compute
- Works very well for BAI. Not many experiments beyond.
- Performs very well in practise

Two Interesting Points

Sticky Track-and-Stop for Multiple Correct Answers

Q: Is $\boldsymbol{\mu} \mapsto \boldsymbol{w}^*(\boldsymbol{\mu})$ always continuous? No!

Sticky Track-and-Stop for Multiple Correct Answers

Q: Is $\boldsymbol{\mu} \mapsto w^*(\boldsymbol{\mu})$ always continuous? No!

For single-answer problems, can escalate to **set-valued mappings** and **upper hemi-continuity**. Tracking requires care.

Sticky Track-and-Stop for Multiple Correct Answers

Q: Is $\boldsymbol{\mu} \mapsto \boldsymbol{w}^*(\boldsymbol{\mu})$ always continuous? No!

For single-answer problems, can escalate to **set-valued mappings** and **upper hemi-continuity**. Tracking requires care.

For multiple-answer problems (including ϵ-BAI), continuity is unsalvageable.

Sticky Track-and-Stop for Multiple Correct Answers

Q: Is $\boldsymbol{\mu} \mapsto \boldsymbol{w}^*(\boldsymbol{\mu})$ always continuous? No!

For single-answer problems, can escalate to **set-valued mappings** and **upper hemi-continuity**. Tracking requires care.

For multiple-answer problems (including ϵ-BAI), continuity is unsalvageable.

Contributions in (Degenne and Koolen, 2019)

- A lower-bound with multiple correct answers (now $\max \max \inf$).
- A new algorithm *Sticky Track-and-Stop* that asymptotically matches the lower bound.
- Explicit example where vanilla TaS fails (arcsine law)

Saddle Point Techniques

Standard technique: approximately solve saddle point problem

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

iteratively using two online learners.

Saddle Point Techniques

Standard technique: approximately solve saddle point problem

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

iteratively using two online learners.

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

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

Saddle Point Techniques

Standard technique: approximately solve saddle point problem

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

iteratively using two online learners.

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

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

Implementation available in tidnabbil library.
Analogue for regret in (Degenne, Shao, and Koolen, 2020)

Thanks!

References

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*, pp. 14564–14573.

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*, pp. 14465–14474.

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)*.

- Garivier, A. and E. Kaufmann (2016). “Optimal Best arm Identification with Fixed Confidence”. In: *Proceedings of the 29th Conference On Learning Theory (COLT)*.
- Kaufmann, E. and W. M. Koolen (Oct. 2018). “Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals”. Preprint.
- Lai, T. L. and H. Robbins (1985). “Asymptotically efficient adaptive allocation rules”. In: *Advances in Applied Mathematics* 6.1, pp. 4–22.