Pure Exploration Problems

Wouter Koolen

October 26th, 2020

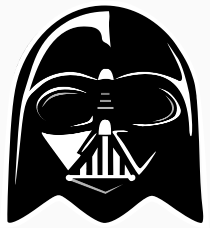
Expectation Management for the Experts

I will discuss

- Instance-dependent results.
- Fixed confidence setting.
- Asymptotic confidence $\delta \to 0$ regime.

Context

Pure Exploration is statistical hypothesis testing ...

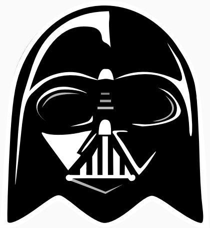


Maximise Reward

Gain Knowledge

Context

Pure Exploration is statistical hypothesis testing ...



Maximise Reward

Gain Knowledge

...on steroids:

- Multiple
- Composite
- Sequential
- Active

Menu

Introduce Pure Exploration problems.

Relate Pure Exploration to Reinforcement Learning.

Highlight some recent lessons learned.

Examples

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 $X_t \sim \mathrm{Bernoulli}(\mu_{A_t})$

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

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

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 $t = 1, 2, \ldots$ **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 $A_t = \dots$
Recommend $\hat{I} = \dots$

Best Arm Identification, prototypical solution

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

Fancy Algorithm(δ)

Stop when $\dots$

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 ...

Sample arm $A_t = \dots$

Recommend $\hat{I} = \dots$

Theorem (lower bd)

Any δ-PAC algorithm needs sample complexity at least

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

Theorem (safe)

Fancy Algorithm(δ) is δ-PAC

Theorem (comput. eff.)

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

Theorem (statistic. eff.)

... has sample complexity

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

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

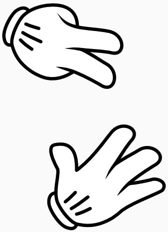
- Linear, Lipschitz, Sparse, Categorical, Unimodal, ...

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

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

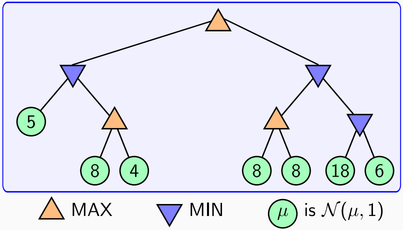
Feedback graphs (semi-bandit, ...)

Major Variations

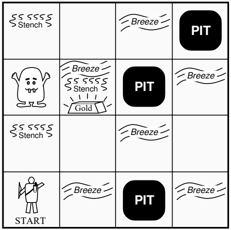


Nash Equilibrium

Shortest Path

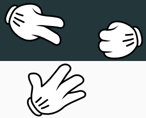


Minimax Tree Search



Markov Decision Process

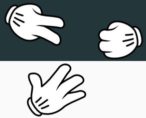
Nash Equilibrium Identification Problem



Assumption: Rectangular Multi-Armed Bandit

$K \times M$ Bernoulli arms with unknown means $\boldsymbol{\mu} \in [0,1]^{K \times M}$.

Nash Equilibrium Identification Problem



Assumption: Rectangular Multi-Armed Bandit

$K \times M$ Bernoulli arms with unknown means $\boldsymbol{\mu} \in [0,1]^{K \times M}$.

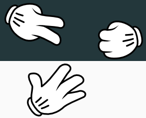
NASH-MAB Protocol

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

- Learner picks arm $(I_t, J_t) \in K \times M$
- Learner observes $X_t \sim \mu_{I_t, J_t}$

Learner recommends $(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}) \in \triangle_K \times \triangle_M$.

Nash Equilibrium Identification Problem



Assumption: Rectangular Multi-Armed Bandit

$K \times M$ Bernoulli arms with unknown means $\boldsymbol{\mu} \in [0,1]^{K \times M}$.

NASH-MAB Protocol

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

- Learner picks arm $(I_t, J_t) \in K \times M$
- Learner observes $X_t \sim \mu_{I_t, J_t}$

Learner recommends $(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}) \in \triangle_K \times \triangle_M$.

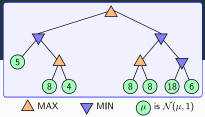
We say that $(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}})$ is an ϵ*-approximate Nash equilibrium for* $\boldsymbol{\mu}$ if

$$\max_{\boldsymbol{p} \in \triangle_K} \boldsymbol{p}^\intercal \boldsymbol{\mu} \hat{\boldsymbol{q}} - \min_{\boldsymbol{q} \in \triangle_M} \hat{\boldsymbol{p}}^\intercal \boldsymbol{\mu} \boldsymbol{q} \ \leq \ \epsilon.$$

Problem

(ϵ, δ)-PAC Nash equilibrium identification.

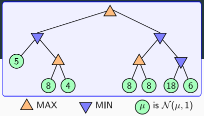
Minimax Action Identification Problem



Assumption: Tree-Structured Multi-Armed Bandit

Given minimax tree $\mathcal{T}$ with unknown Bernoulli μ_ℓ in each leaf ℓ.

Minimax Action Identification Problem



Assumption: Tree-Structured Multi-Armed Bandit

Given minimax tree $\mathcal{T}$ with unknown Bernoulli μ_ℓ in each leaf ℓ.

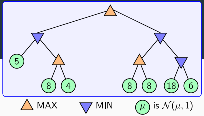
MINIMAX-MAB Protocol

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

- Learner picks a leaf L_t of $\mathcal{T}$
- Learner observes $X_t \sim \mu_{L_t}$

Learner recommends child $\hat{c}$ of the root of $\mathcal{T}$.

Minimax Action Identification Problem



Assumption: Tree-Structured Multi-Armed Bandit

Given minimax tree $\mathcal{T}$ with unknown Bernoulli μ_ℓ in each leaf ℓ.

MINIMAX-MAB Protocol

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

- Learner picks a leaf L_t of $\mathcal{T}$
- Learner observes $X_t \sim \mu_{L_t}$

Learner recommends child $\hat{c}$ of the root of $\mathcal{T}$.

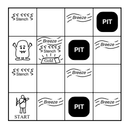
The optimal move at the root of $\mathcal{T}$ is

$$c^*(\boldsymbol{\mu}) \; = \; \arg\max_i \min_j \max_k \ldots \mu_{(i,j,k,\ldots)}$$

Problem

δ-PAC identification of optimal move at the root of $\mathcal{T}$.

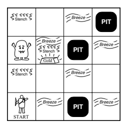
Optimal Policy Identification Problem



Assumption: MDP

Known state space $\mathcal{S}$, action space $\mathcal{A}$ and start state $s_0 \in \mathcal{S}$.
Unknown mean reward $\mu_{s,a}$ and transition dynamics $P(s'|s,a)$.

Optimal Policy Identification Problem



Assumption: MDP

Known state space $\mathcal{S}$, action space $\mathcal{A}$ and start state $s_0 \in \mathcal{S}$.
Unknown mean reward $\mu_{s,a}$ and transition dynamics $P(s'|s,a)$.

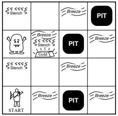
BPI-RL Protocol

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

- Learner picks a state $S_t \in \mathcal{S}$ and action $A_t \in \mathcal{A}$
- Learner observes $R_t \sim \mu_{S_t,A_t}$ and $S'_t \sim P(\cdot|S_t, A_t)$

Learner recommends policy $\hat{\pi} : \mathcal{S} \to \mathcal{A}$.

Optimal Policy Identification Problem



Assumption: MDP

Known state space $\mathcal{S}$, action space $\mathcal{A}$ and start state $s_0 \in \mathcal{S}$.
Unknown mean reward $\mu_{s,a}$ and transition dynamics $P(s'|s,a)$.

BPI-RL Protocol

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

- Learner picks a state $S_t \in \mathcal{S}$ and action $A_t \in \mathcal{A}$
- Learner observes $R_t \sim \mu_{S_t,A_t}$ and $S'_t \sim P(\cdot|S_t, A_t)$

Learner recommends policy $\hat{\pi} : \mathcal{S} \to \mathcal{A}$.

The *value function* of policy π at discount factor $\gamma \in (0,1)$ solves $V^\pi(s) = \mu_{s,\pi(s)} + \gamma \, \mathbb{E}_{s' \sim P(\cdot|s,\pi(s))} \left[V^\pi(s')\right]$. The *optimal policy* is $\pi^* = \arg\max_\pi V^\pi(s_0)$.

Problem

δ-PAC identification of optimal policy π^*.

The Canonical Path to Instance-Optimal Algorithms

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

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

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

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

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

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 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

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

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 $w^*(\boldsymbol{\mu})$.

Three 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 $\boldsymbol{\mu} \mapsto \boldsymbol{w}^*(\boldsymbol{\mu})$ 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 $\max \max \min$).
- 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

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

iteratively using two online learners.

Saddle Point Techniques

Standard technique: approximately solve saddle point problem

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

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

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

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.

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

Martingales, GLRT and T^*

(Agrawal, Koolen, and Juneja, 2020) look at identifying the arm of minimum CVaR in a **non-parametric** setting with heavy tails.

The dual of the lower bound problem gives a natural collection of martingales. The mixture martingale method then gives the deviation inequalities for the stopping rule threshold.

Open Problems

Open problems

- “Pure Exploration Compiler”
  - query
  - structure (prior knowledge) assumptions
- Moderate confidence regime (i.e. dependence in problem parameters other than δ)
- Scaling back up: subroutines for planning systems

Thanks!

References

Agrawal, S., W. M. Koolen, and S. Juneja (Aug. 2020). "Optimal Best-Arm Identification Methods for Tail-Risk Measures". In: *ArXiv*.

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)*.

Lai, T. L. and H. Robbins (1985). “Asymptotically efficient adaptive allocation rules”. In: *Advances in Applied Mathematics* 6.1, pp. 4–22.