

Pure Exploration with Multiple Correct Answers



Rémy Degenne

Wouter M. Koolen

CWI

Centrum Wiskunde & Informatica



Outline

- 1 Introduction
- 2 Model
- 3 TaS for BAI
- 4 Discontinuous single-answer problems
- 5 Multiple-answer problems
- 6 Conclusion

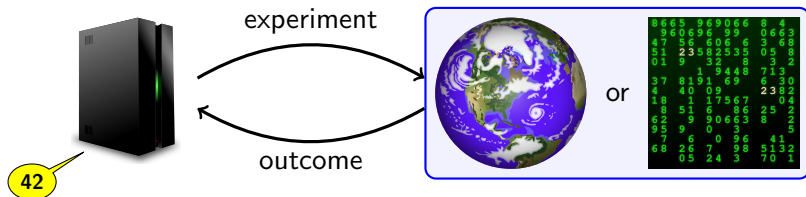
Topic: Pure Exploration

Query:

most effective drug dose?

most appealing website layout?

safest next robot action?



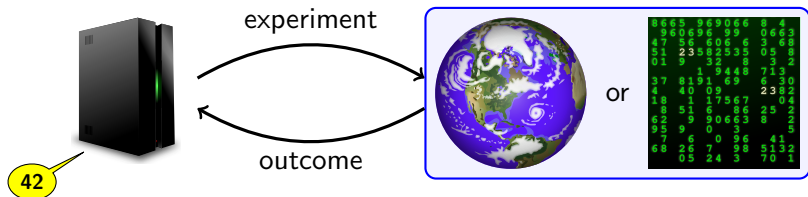
Topic: Pure Exploration

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**

This Talk

- We study queries w. **multiple correct answers**.
E.g. find an ϵ -optimal drug.
- The leading existing approach **fails** due to non-continuity.
- We propose a stabilisation called “Sticky Track-and-Stop”



Outline

- 1 Introduction
- 2 Model**
- 3 TaS for BAI
- 4 Discontinuous single-answer problems
- 5 Multiple-answer problems
- 6 Conclusion

Formal model

Environment (Multi-armed bandit model)

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

Set of possible environments: \mathcal{M} .

Formal model

Environment (Multi-armed bandit model)

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

Set of possible environments: \mathcal{M} .

Query

Set of possible answers \mathcal{I} . **Correct answer** function $i^* : \mathcal{M} \rightarrow \mathcal{I}$.

Formal model

Environment (Multi-armed bandit model)

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

Set of possible environments: \mathcal{M} .

Query

Set of possible answers \mathcal{I} . **Correct answer** function $i^* : \mathcal{M} \rightarrow \mathcal{I}$.

Strategy

- **Stopping rule** $\tau \in \mathbb{N}$
- In round $t \leq \tau$ **sampling rule** picks $A_t \in [K]$. See $X_t \sim \mu_{A_t}$.
- **Recommendation rule** $\hat{I} \in [K]$.

Formal model

Environment (Multi-armed bandit model)

K distributions parameterised by their means $\mu = (\mu_1, \dots, \mu_K)$.

Set of possible environments: \mathcal{M} .

Query

Set of possible answers \mathcal{I} . **Correct answer** function $i^* : \mathcal{M} \rightarrow \mathcal{I}$.

Strategy

- **Stopping rule** $\tau \in \mathbb{N}$
- In round $t \leq \tau$ **sampling rule** picks $A_t \in [K]$. See $X_t \sim \mu_{A_t}$.
- **Recommendation rule** $\hat{I} \in [K]$.

Realisation of interaction: $(A_1, X_1), \dots, (A_\tau, X_\tau), \hat{I}$.

Formal model

Environment (Multi-armed bandit model)

K distributions parameterised by their means $\mu = (\mu_1, \dots, \mu_K)$.

Set of possible environments: \mathcal{M} .

Query

Set of possible answers \mathcal{I} . **Correct answer** function $i^* : \mathcal{M} \rightarrow \mathcal{I}$.

Strategy

- **Stopping rule** $\tau \in \mathbb{N}$
- In round $t \leq \tau$ **sampling rule** picks $A_t \in [K]$. See $X_t \sim \mu_{A_t}$.
- **Recommendation rule** $\hat{I} \in [K]$.

Realisation of interaction: $(A_1, X_1), \dots, (A_\tau, X_\tau), \hat{I}$.

Two objectives: **sample efficiency** τ and **correctness** $\hat{I} = i^*(\mu)$.

Goal: PAC learning

Definition

Fix small confidence $\delta \in (0, 1)$. A strategy is δ -**correct** if

$$\mathbb{P}_{\mu}(\hat{I} \neq i^*(\mu)) \leq \delta \quad \text{for every bandit model } \mu \in \mathcal{M}.$$

Goal: PAC learning

Definition

Fix small confidence $\delta \in (0, 1)$. A strategy is δ -**correct** if

$$\mathbb{P}_{\mu}(\hat{I} \neq i^*(\mu)) \leq \delta \quad \text{for every bandit model } \mu \in \mathcal{M}.$$

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

Examples w. 2 arms

Problem name

Best Arm

Minimum Threshold

Possible answers \mathcal{I}

$[K]$

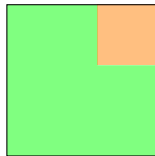
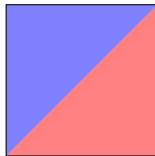
$\{\text{lo}, \text{hi}\}$

Correct answer $i^*(\mu)$

$\text{argmax}_k \mu_k$

$\{\text{lo}\}$ if $\min_k \mu_k < \gamma$

$\{\text{hi}\}$ if $\min_k \mu_k > \gamma$





Outline

- 1 Introduction
- 2 Model
- 3 TaS for BAI**
- 4 Discontinuous single-answer problems
- 5 Multiple-answer problems
- 6 Conclusion

Instance-Dependent Sample Complexity Lower bound

Define the **alternative** to answer $i \in \mathcal{I}$ by $\neg i = \{\lambda | i^*(\lambda) \neq i\}$.

Instance-Dependent Sample Complexity Lower bound

Define the **alternative** to answer $i \in \mathcal{I}$ by $\neg i = \{\lambda | i^*(\lambda) \neq i\}$.

Theorem (Castro 2014, Garivier and Kaufmann 2016)

Fix a δ -correct strategy. Then for every bandit model μ

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

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

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

Instance-Dependent Sample Complexity Lower bound

Define the **alternative** to answer $i \in \mathcal{I}$ by $\neg i = \{\lambda | i^*(\lambda) \neq i\}$.

Theorem (Castro 2014, Garivier and Kaufmann 2016)

Fix a δ -correct strategy. Then for every bandit model μ

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

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

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

Intuition (going back to Lai and Robbins [1985]): if observations are likely under both μ and λ , yet $i^*(\mu) \neq i^*(\lambda)$, then learner cannot stop and be correct in both.

Example

Best Arm identification: $i^*(\boldsymbol{\mu}) = \operatorname{argmax}_i \mu_i$.

$K = 5$ arms, Bernoulli $\boldsymbol{\mu} = (0, 0.1, 0.2, 0.3, 0.4)$.

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

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

Operationalisation of the Oracle Weights

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

$$w^*(\mu) = \operatorname{argmax}_{w \in \Delta_K} \min_{\lambda \in \neg i^*(\mu)} \sum_{i=1}^K w_i \operatorname{KL}(\mu_i \| \lambda_i)$$

Operationalisation of the Oracle Weights

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

$$w^*(\mu) = \operatorname{argmax}_{w \in \Delta_K} \min_{\lambda \in \neg i^*(\mu)} \sum_{i=1}^K w_i \operatorname{KL}(\mu_i \| \lambda_i)$$

Track-and-Stop [Garivier and Kaufmann, 2016]

Idea: draw $A_t \sim w^*(\hat{\mu}(t))$.

- Ensure $\hat{\mu}(t) \rightarrow \mu$ by “forced exploration”
- **assuming w^* is continuous**, this ensures $w^*(\hat{\mu}_t) \rightarrow w^*(\mu)$.
- hence $N_i(t)/t \rightarrow w_i^*$
- Draw arm with $N_i(t)/t$ below w_i^* (tracking)



Outline

- 1 Introduction
- 2 Model
- 3 TaS for BAI
- 4 Discontinuous single-answer problems**
- 5 Multiple-answer problems
- 6 Conclusion

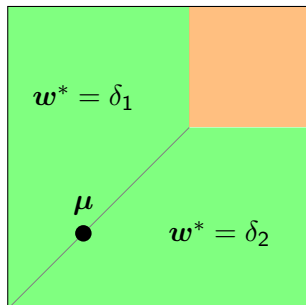
About that continuity assumption?

Can w^* be discontinuous?

About that continuity assumption?

Can w^* be discontinuous?

Example: Minimum Threshold



Continuity restored

Recall oracle weights are given by

$$w^*(\mu) = \operatorname{argmax}_{w \in \Delta} \inf_{\lambda \in \neg i^*(\mu)} \sum_a w_a d(\mu_a, \lambda_a)$$

Continuity restored

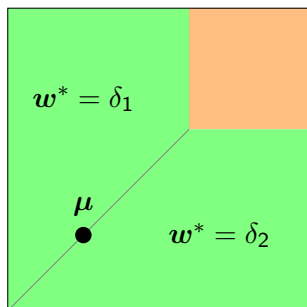
Recall oracle weights are given by

$$w^*(\mu) = \operatorname{argmax}_{w \in \Delta} \inf_{\lambda \in \neg i^*(\mu)} \sum_a w_a d(\mu_a, \lambda_a)$$

Theorem

w^* , when viewed as a **set-valued** function, is upper hemicontinuous. Moreover, its output is always a convex sets.

Intuition



On bandit model μ , our empirical distribution will be a convex combination of δ_1 and δ_2 .

Putting it all together

TaS:

- Forced exploration to ensure $\hat{\mu}_t \rightarrow \mu$.
- Compute $w_t = w^*(\hat{\mu}_t)$.
- Choose arm A_{t+1} to ensure $N_i(t)/t \rightarrow w_t$ (Tracking)

Putting it all together

TaS:

- Forced exploration to ensure $\hat{\mu}_t \rightarrow \mu$.
- Compute $w_t = w^*(\hat{\mu}_t)$.
- Choose arm A_{t+1} to ensure $N_i(t)/t \rightarrow w_t$ (Tracking)

Theorem

Track-and-Stop with C-tracking is δ -correct with asymptotically optimal sample complexity.

Putting it all together

TaS:

- Forced exploration to ensure $\hat{\mu}_t \rightarrow \mu$.
- Compute $w_t = w^*(\hat{\mu}_t)$.
- Choose arm A_{t+1} to ensure $N_i(t)/t \rightarrow w_t$ (Tracking)

Theorem

Track-and-Stop with C-tracking is δ -correct with asymptotically optimal sample complexity.

Theorem

Track-and-Stop with D tracking may fail to converge.



Outline

- 1 Introduction
- 2 Model
- 3 TaS for BAI
- 4 Discontinuous single-answer problems
- 5 Multiple-answer problems**
- 6 Conclusion

Updated Problem

We now assume a **set-valued** correct answer function $i^* : \mathcal{M} \rightarrow 2^{\mathcal{I}}$.

Updated Problem

We now assume a **set-valued** correct answer function $i^* : \mathcal{M} \rightarrow 2^{\mathcal{I}}$.

Examples:

Problem

ϵ Best Arm

Any Low Arm

Answers \mathcal{I}

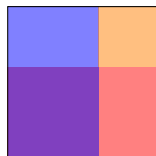
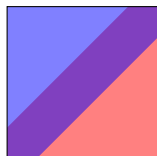
$[K]$

$[K] \cup \{\text{no}\}$

Correct $i^*(\mu)$

$\{k \mid \mu_k \geq \max_j \mu_j - \epsilon\}$

$\{k \mid \mu_k \leq \gamma\}$ if $\min_k \mu_k < \gamma$
 $\{\text{no}\}$ if $\min_k \mu_k > \gamma$



Rethinking the lower bound

For single-answer problems, lower bound is based on KL contraction.

Rethinking the lower bound

For single-answer problems, lower bound is based on KL contraction.

With multiple correct answers, this gives the wrong leading constant.

Rethinking the lower bound

For single-answer problems, lower bound is based on KL contraction.

With multiple correct answers, this gives the wrong leading constant.

Theorem

Any δ -correct algorithm verifies

$$\liminf_{\delta \rightarrow 0} \frac{\mathbb{E}_{\mu}[\tau_{\delta}]}{\log(1/\delta)} \geq T^*(\mu) := D(\mu)^{-1}$$

where

$$D(\mu) = \max_{i \in i^*(\mu)} \max_{w \in \Delta_K} \inf_{\lambda \in -i} \sum_{k=1}^K w_k d(\mu_k, \lambda_k)$$

for any multiple answer instance μ with sub-Gaussian arm distributions.

Proof ideas

- Min-max swap: For any answer $i \in \mathcal{I}$,

$$\max_{w \in \Delta_K} \inf_{\lambda \in \mathcal{I}} \sum_{k=1}^K w_k d(\mu_k, \lambda_k) = \inf_{\mathbb{P}} \max_{k \in [K]} \mathbb{E}_{\lambda \sim \mathbb{P}} [d(\mu_k, \lambda_k)].$$

\Rightarrow get \mathbb{P}^* . Say weights q_1, \dots, q_K on $\lambda^1, \dots, \lambda^K$.

- Look at likelihood ratio

$$L_n = -\ln \frac{d\mathbb{P}^*}{d\mathbb{P}_\mu} \leq \sum_k q_k \ln \frac{d\mathbb{P}_\mu}{d\mathbb{P}_{\lambda^k}}.$$

It follows that for any $\gamma \in \mathbb{R}$ we have

$$\{L_n > \gamma\} \subseteq \left\{ \underbrace{\sum_k q_k \sum_a N_{n,a} d(\mu_a, \lambda_a^k)}_{\leq \text{value}} + \underbrace{\sum_k q_k M_n(\mu, \lambda^k)}_{\text{martingale}} > \gamma \right\}.$$

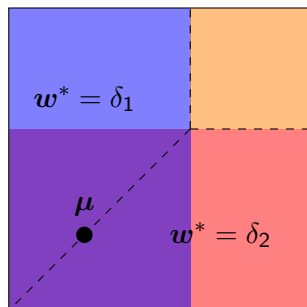
\Rightarrow cannot distinguish μ from at least one λ^k .

Matching the lower bound

New problem: **real discontinuity**.

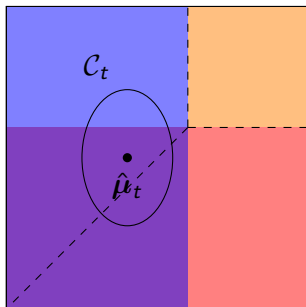
$$\max_{i \in i^*(\mu)} \max_{w \in \Delta_K} \inf_{\lambda \in -i} \sum_{k=1}^K w_k d(\mu_k, \lambda_k)$$

Example:



Solution: Make it sticky

Sampling rule: find least (in sticky order) oracle answer in the aggressive confidence region \mathcal{C}_t . Track its oracle weights at $\hat{\mu}_t$.



$$\text{orange} < \text{red} < \text{blue}$$

Main Result

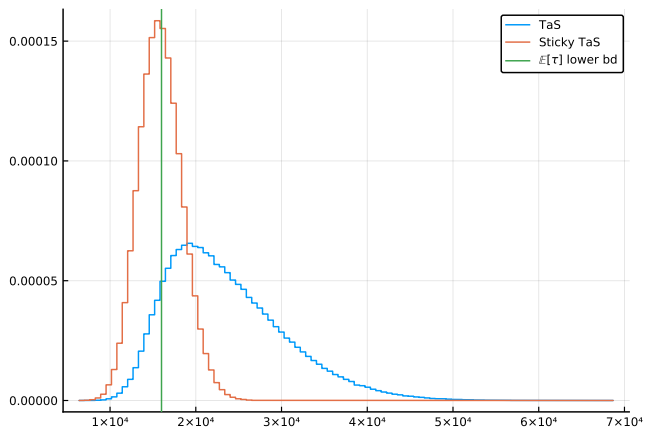
When coupled with a good stopping rule,

Theorem

Sticky Track-and-Stop is asymptotically optimal, i.e. it verifies for all $\mu \in \mathcal{M}$,

$$\lim_{\delta \rightarrow 0} \frac{\mathbb{E}_{\mu}[\tau_{\delta}]}{\log(1/\delta)} \rightarrow \frac{1}{D(\mu)}.$$

How bad is “Teflon” TaS?



Story: arcsine law.



Outline

- 1 Introduction
- 2 Model
- 3 TaS for BAI
- 4 Discontinuous single-answer problems
- 5 Multiple-answer problems
- 6 Conclusion**

Conclusion

- Pure Exploration currently going through a renaissance
- Instance-optimal identification algorithms
 - ▶ Best Arm
 - ▶ Combinatorial best action
 - ▶ Game Tree Search
 - ▶ ...
- Moving toward more complex queries. RL on the horizon ...
- Useful submodules

Many questions remain open

- Practically efficient algorithms
- Remove forced exploration
- Moderate confidence $\delta \not\rightarrow 0$ regime [Simchowitz et al., 2017].
- Understand sparsity patterns
- Dynamically expanding horizon

Many questions remain open

- Practically efficient algorithms
- Remove forced exploration
- Moderate confidence $\delta \not\rightarrow 0$ regime [Simchowitz et al., 2017].
- Understand sparsity patterns
- Dynamically expanding horizon

Thank you! And let's talk!