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
Query:
- most effective drug dose?
- most appealing website layout?
- safest next robot action?

Efficient systems
Sample complexity as function of query and environment

Degenne and Koolen
Multiple Correct Answers
Google '19
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
We study queries w. multiple correct answers. E.g. find an $\epsilon$-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 \( \mu = (\mu_1, \ldots, \mu_K) \).
Set of possible environments: \( \mathcal{M} \).
Formal model

Environment (Multi-armed bandit model)

$K$ distributions parameterised by their means $\mu = (\mu_1, \ldots, \mu_K)$. 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 $\mu = (\mu_1, \ldots, \mu_K)$. 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 $\hat{i} \in [K]$. 

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

Degenne and Koolen '19 6 / 28
Formal model

Environment (Multi-armed bandit model)

$K$ distributions parameterised by their means $\mu = (\mu_1, \ldots, \mu_K)$. 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: $(A_1, X_1), \ldots, (A_\tau, X_\tau), \hat{i}$. 
Formal model

Environment (Multi-armed bandit model)

$K$ distributions parameterised by their means $\mu = (\mu_1, \ldots, \mu_K)$. Set of possible environments: $\mathcal{M}$.

Query

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

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: $(A_1, X_1), \ldots, (A_\tau, X_\tau), \hat{i}$.

**Two objectives:** sample efficiency $\tau$ and correctness $\hat{i} = i^*(\mu)$. 
Goal: PAC learning

Definition

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

$$\mathbb{P}_\mu(\hat{i} \neq i^*(\mu)) \leq \delta$$

for every bandit model $\mu \in \mathcal{M}$. 
Goal: PAC learning

Definition

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

$$\mathbb{P}_\mu(\hat{i} \neq i^*(\mu)) \leq \delta$$

for every bandit model $\mu \in \mathcal{M}$.

Goal: minimise sample complexity $\mathbb{E}_\mu[\tau]$ over all $\delta$-correct strategies.
Examples w. 2 arms

<table>
<thead>
<tr>
<th>Problem name</th>
<th>Best Arm</th>
<th>Minimum Threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Possible answers $\mathcal{I}$</td>
<td>$[K]$</td>
<td>${\text{lo, hi}}$</td>
</tr>
<tr>
<td>Correct answer $i^*(\mu)$</td>
<td>$\arg\max_k \mu_k$</td>
<td>${\text{lo}}$ if $\min_k \mu_k &lt; \gamma$</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
2. Model
3. TaS for BAI
4. Discontinuous single-answer problems
5. Multiple-answer problems
6. Conclusion
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 $\delta$-correct strategy. Then for every bandit model $\mu$,

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

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

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

Intuition (going back to Lai and Robbins [1985]): if observations are likely under both $\mu$ and $\lambda$, yet $i^*(\mu) \neq i^*(\lambda)$, then learner cannot stop and be correct in both.
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 $\delta$-correct strategy. Then for every bandit model $\mu$

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

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 \textbf{alternative} to answer \( i \in \mathcal{I} \) by 
\[ -i = \{ \lambda | i^*(\lambda) \neq i \} \].

\textbf{Theorem (Castro 2014, Garivier and Kaufmann 2016)}

Fix a \( \delta \)-correct strategy. Then for every bandit model \( \mu \)
\[ \mathbb{E}_{\mu} [\tau] \geq T^* (\mu) \ln \frac{1}{\delta} \]

where the \textbf{characteristic time} \( T^* (\mu) \) is given by
\[ \frac{1}{T^* (\mu)} = \max_{w \in \Delta K} \min_{\lambda \in -i^*(\mu)} \sum_{i=1}^{K} w_i \text{KL}(\mu_i || \lambda_i). \]

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

Best Arm identification:  \( i^*(\mu) = \arg\max_i \mu_i \).

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

\[
T^*(\mu) = 200.4 \quad w^*(\mu) = (0.45, 0.46, 0.06, 0.02, 0.01)
\]

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

\[ w^*(\mu) = \arg\max_{w \in \Delta_K} \min_{\lambda \in \Lambda^*} \sum_{i=1}^{K} w_i \text{KL}(\mu_i \| \lambda_i) \]
Operationalisation of the Oracle Weights

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

\[ w^*(\mu) = \arg\max_{w \in \Delta_K} \min_{\lambda \in -\nu^*(\mu)} \sum_{i=1}^{K} w_i \KL(\mu_i \parallel \lambda_i) \]

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

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

- Ensure \( \hat{\mu}(t) \to \mu \) by “forced exploration”
- assuming \( w^* \) is continuous, this ensures \( w^*(\hat{\mu}_t) \to w^*(\mu) \).
- hence \( N_i(t)/t \to w_i^* \)
- 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
Recall oracle weights are given by

\[ w^*(\mu) = \arg\max_{w \in \Delta} \inf_{\lambda \in \nu^*(\mu)} \sum_a w_a d(\mu_a, \lambda_a) \]
Recall oracle weights are given by

$$w^*(\mu) = \arg\max_{w \in \triangle} \inf_{\lambda \in \neg i^*(\mu)} \sum_a w_a d(\mu_a, \lambda_a)$$

Theorem

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

On bandit model $\mu$, our empirical distribution will be a convex combination of $\delta_1$ and $\delta_2$. 
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 $\hat{\mu}_t \to \mu$.
- Compute $w_t = w^*(\hat{\mu}_t)$.
- Choose arm $A_{t+1}$ to ensure $N_i(t)/t \to w_t$ (Tracking)

**Theorem**

*Track-and-Stop with C-tracking is $\delta$-correct with asymptotically optimal sample complexity.*
Putting it all together

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

**Theorem**

*Track-and-Stop with C-tracking is $\delta$-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 \( i^* : \mathcal{M} \rightarrow 2^\mathcal{I} \).
We now assume a **set-valued** correct answer function $i^*: \mathcal{M} \rightarrow 2^{\mathcal{I}}$.

Examples:

<table>
<thead>
<tr>
<th>Problem</th>
<th>$\epsilon$ Best Arm</th>
<th>Any Low Arm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Answers</td>
<td>$\mathcal{I}$</td>
<td>$[K] \cup {\text{no}}$</td>
</tr>
<tr>
<td>Correct</td>
<td>$i^*(\mu)$</td>
<td>${k</td>
</tr>
<tr>
<td></td>
<td></td>
<td>${\text{no}}$ if $\min_k \mu_k &gt; \gamma$</td>
</tr>
</tbody>
</table>
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 $\delta$-correct algorithm verifies

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

where

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

for any multiple answer instance $\mu$ with sub-Gaussian arm distributions.
Proof ideas

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

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

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

- **Look at likelihood ratio**

$$
L_n = -\ln \frac{d \mathbb{P}^*}{d \mathbb{P} \mu} \leq \sum_{k} q_k \ln \frac{d \mathbb{P}^*}{d \mathbb{P} \lambda_k}.
$$

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

$$
\{L_n > \gamma\} \subseteq \left\{ \sum_{k} q_k \sum_{a} N_n, a d(\mu_a, \lambda^k_a) + \sum_{k} q_k M_n(\mu, \lambda^k) > \gamma \right\}.
$$

$\Rightarrow$ cannot distinguish $\mu$ from at least one $\lambda^k$. 
Matching the lower bound

New problem: **real discontinuity**.

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

Example:

\[
w^* = \delta_1, \quad w^* = \delta_2
\]
Solution: Make it sticky

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

![Diagram showing the approach to making the problem sticky](image-url)
Main Result

When coupled with a good stopping rule,

**Theorem**

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

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

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 $\delta \not\to 0$ regime [Simchowitz et al., 2017].
- Understand sparsity patterns
- Dynamically expanding horizon
Many questions remain open

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

Thank you! And let’s talk!