Pure Exploration with Multiple Correct Answers

Rémy Degenne **Wouter M. Koolen**

CWI
Centrum Wiskunde & Informatica

Outline

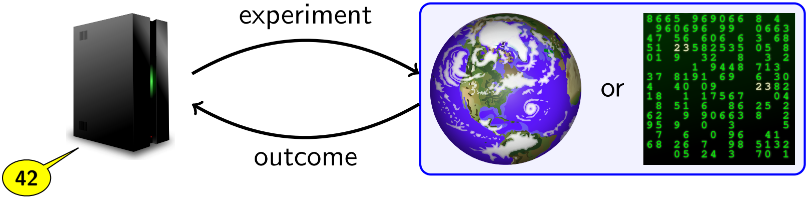
Topic: Pure Exploration

Query:

most effective drug dose?

most appealing website layout?

safest next robot action?



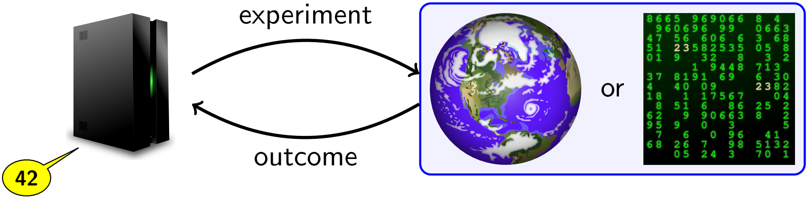
Topic: Pure Exploration

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

Formal model

Environment (Multi-armed bandit model)

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

Formal model

Environment (Multi-armed bandit model)

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

Query

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

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 $i^* : \mathcal{M} \to \mathcal{I}$.

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** $\hat{I} \in [K]$.

Formal model

Environment (Multi-armed bandit model)

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

Formal model

Environment (Multi-armed bandit model)

K distributions parameterised by their means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$.
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** τ and **correctness** $\hat{I} = i^*(\boldsymbol{\mu})$.

Goal: PAC learning

Definition

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

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

Goal: PAC learning

Definition

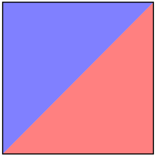
Fix small confidence $\delta \in (0, 1)$. A strategy is δ-**correct** if

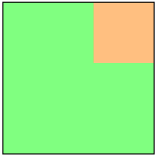
$$\mathbb{P}_{\boldsymbol{\mu}}\big(\hat{I} \neq i^*(\boldsymbol{\mu})\big) \;\leq\; \delta \qquad \text{for every bandit model } \boldsymbol{\mu} \in \mathcal{M}.$$

Goal: minimise sample complexity $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ 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 $i^*(\boldsymbol{\mu})$	$\text{argmax}_k \mu_k$	$\{\text{lo}\}$ if $\min_k \mu_k < \gamma$ $\{\text{hi}\}$ if $\min_k \mu_k > \gamma$





Outline

Instance-Dependent Sample Complexity Lower bound

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

Instance-Dependent Sample Complexity Lower bound

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

Theorem (Castro 2014, Garivier and Kaufmann 2016)

Fix a δ-correct strategy. Then for every bandit model $\boldsymbol{\mu}$

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

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

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

Instance-Dependent Sample Complexity Lower bound

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

Theorem (Castro 2014, Garivier and Kaufmann 2016)

Fix a δ-correct strategy. Then for every bandit model $\boldsymbol{\mu}$

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

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

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

Intuition (going back to Lai and Robbins [1985]): if observations are likely under both $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$, yet $i^*(\boldsymbol{\mu}) \neq i^*(\boldsymbol{\lambda})$, 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 $\boldsymbol{\mu} = (0, 0.1, 0.2, 0.3, 0.4)$.

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

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

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

Operationalisation of the Oracle Weights

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

$$\boldsymbol{w}^*(\boldsymbol{\mu}) \;=\; \underset{\boldsymbol{w}\in\triangle_K}{\operatorname{argmax}} \; \min_{\boldsymbol{\lambda}\in\neg i^*(\boldsymbol{\mu})} \; \sum_{i=1}^K w_i \, \mathrm{KL}(\mu_i \| \lambda_i)$$

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

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

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

Outline

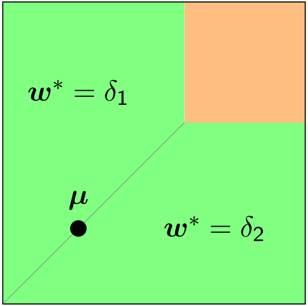
About that continuity assumption?

Can $\boldsymbol{w}^*$ be discontinuous?

About that continuity assumption?

Can $\boldsymbol{w}^*$ be discontinuous?

Example: Minimum Threshold



Continuity restored

Recall oracle weights are given by

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

Continuity restored

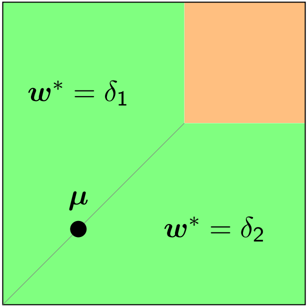
Recall oracle weights are given by

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

Theorem

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

Intuition



On bandit model $\boldsymbol{\mu}$, our empirical distribution will be a convex combination of δ_1 and δ_2.

Putting it all together

TaS:

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

Putting it all together

TaS:

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

Updated Problem

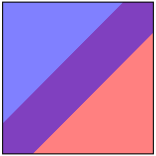
We now assume a **set-valued** correct answer function $i^* : \mathcal{M} \to 2^{\mathcal{I}}$.

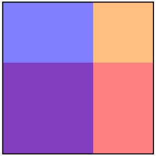
Updated Problem

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

Examples:

Problem	ϵ Best Arm	Any Low Arm
Answers $\mathcal{I}$	$[K]$	$[K] \cup \{\text{no}\}$
Correct $i^*(\boldsymbol{\mu})$	$\{k \mid \mu_k \geq \max_j \mu_j - \epsilon\}$	$\begin{cases} \{k \mid \mu_k \leq \gamma\} & \text{if } \min_k \mu_k < \gamma \\ \{\text{no}\} & \text{if } \min_k \mu_k > \gamma \end{cases}$





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

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

where

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

for any multiple answer instance $\boldsymbol{\mu}$ with sub-Gaussian arm distributions.

Proof ideas

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

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

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

- Look at likelihood ratio

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

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

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

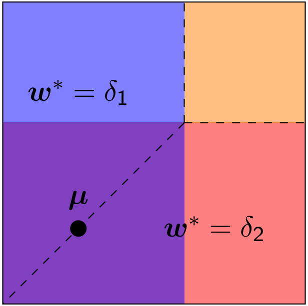
  $\Rightarrow$ cannot distinguish $\boldsymbol{\mu}$ from at least one $\boldsymbol{\lambda}^k$.

Matching the lower bound

New problem: **real discontinuity**.

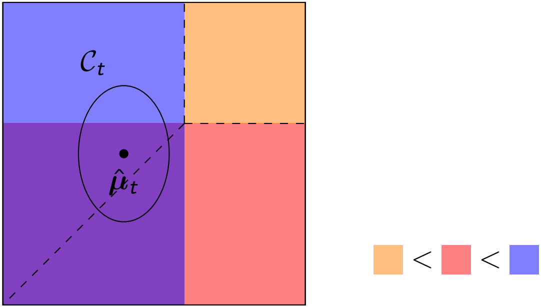
$$\max_{i \in i^*(\boldsymbol{\mu})} \max_{\boldsymbol{w} \in \triangle_K} \inf_{\boldsymbol{\lambda} \in \neg i} \sum_{k=1}^{K} w_k d(\mu_k, \lambda_k)$$

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 $\hat{\boldsymbol{\mu}}_t$.



Main Result

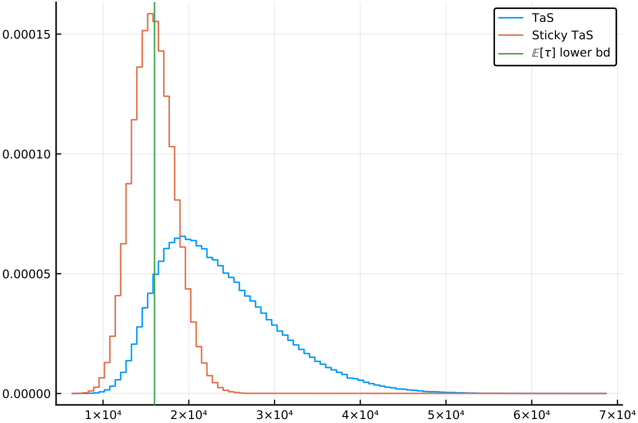
When coupled with a good stopping rule,

Theorem

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

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

How bad is “Teflon” TaS?



Story: arcsine law.

Outline

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!