Pure Exploration with Multiple Correct Answers

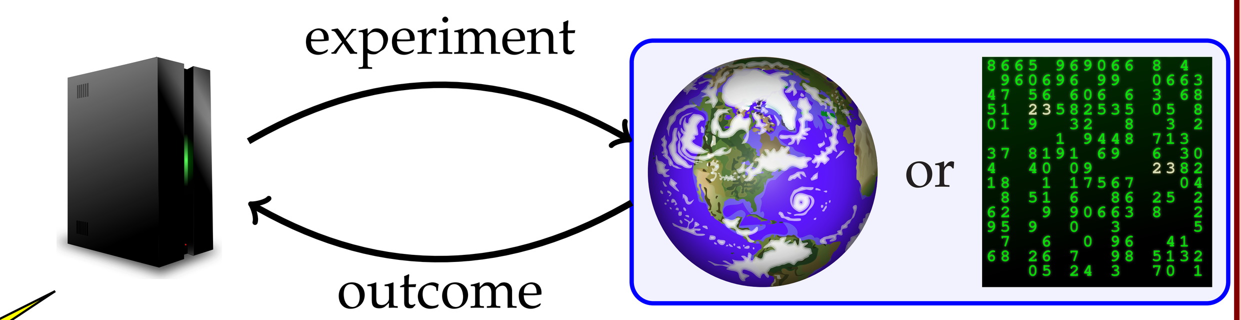
Rémy Degenne CWI

Wouter M. Koolen CWI

Topic: Pure Exploration

Task: **answer query**
- most effective drug dose?
- most appealing website layout?
- safest next robot action?

Setting: interactive learning



Main scientific questions
- **Efficient** systems
- **Sample complexity** as function of query and environment

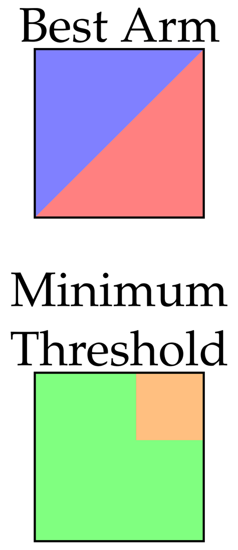
Slogan

Pure exploration with **multiple correct answers** requires **stabilised methods** to pacify **discontinuity**.

Model

K-armed bandit, parameterised by arm means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)$. Set $\mathcal{M}$ of **possible environments**.

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



Strategy:
- **Stopping rule** $\tau \in \mathbb{N}$
- In round $t \le \tau$ **sampling rule** picks arm $A_t \in [K]$ and observes $X_t \sim \mu_{A_t}$.
- **Recommendation rule** $\hat{I} \in [K]$.

Definition. *A strategy is* ***δ-PAC*** *if* $\mathbb{P}_{\boldsymbol{\mu}}(\hat{I} \neq i^*(\boldsymbol{\mu})) \le \delta$ *for every bandit model* $\boldsymbol{\mu} \in \mathcal{M}$.

Goal: minimise **sample complexity** $\mathbb{E}_{\boldsymbol{\mu}}[\tau]$ among δ-PAC strategies.

State of the Art: Lower Bound

Answer $i \in \mathcal{I}$ has **altern.** $\neg i := \{\boldsymbol{\lambda} \in \mathcal{M} | i^*(\boldsymbol{\lambda}) \neq i\}$

Theorem (Castro 2014, Garivier and Kaufmann 2016). *Fix a δ-PAC strategy. Then for every bandit model* $\boldsymbol{\mu} \in \mathcal{M}$

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau] \ge \frac{\ln(1/\delta)}{\max_{\boldsymbol{w} \in \triangle_K} \inf_{\boldsymbol{\lambda} \in \neg i^*(\boldsymbol{\mu})} \sum_{i=1}^K w_i \operatorname{KL}(\mu_i \| \lambda_i)}$$

State of the Art: Algorithm

Good algorithm **must** ⟶ **oracle** proportions

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

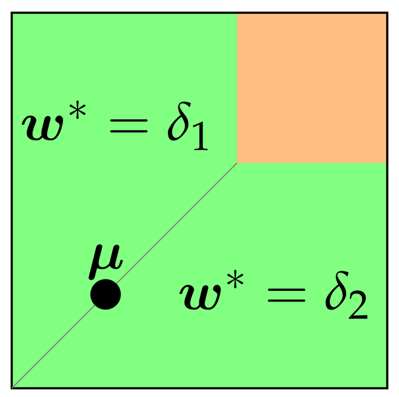
Track-and-Stop [Garivier and Kaufmann, 2016]

Crux: draw $A_t \sim \boldsymbol{w}^*(\hat{\boldsymbol{\mu}}(t))$.
- Ensure $\hat{\boldsymbol{\mu}}(t) \to \boldsymbol{\mu}$ by **forced exploration**
- this ensures $\boldsymbol{w}^*(\hat{\boldsymbol{\mu}}_t) \to \boldsymbol{w}^*(\boldsymbol{\mu})$ **assuming $\boldsymbol{w}^*$ is continuous** ⚠
- Draw arm i with $N_i(t)/t$ below w_i^* (**tracking**)

Discontinuity with Single Answer

Can $\boldsymbol{w}^*$ really be discontinuous? At an instance $\boldsymbol{\mu}$ where the lower bound does not diverge?

Example problem: Minimum Threshold



$\boldsymbol{w}^*(\boldsymbol{\mu}) = \operatorname{conv}(\{\delta_1, \delta_2\})$.

First Result: Continuity Salvaged

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

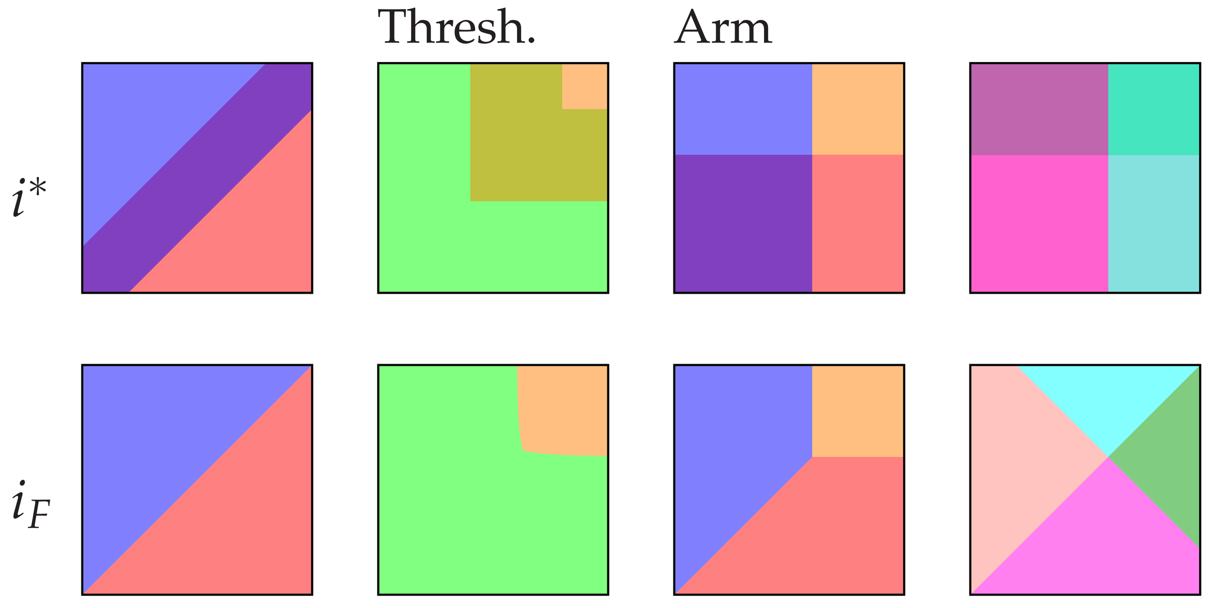
Intuition On bandit model $\boldsymbol{\mu}$, our empirical distribution will be a convex combination of δ_1 and δ_2.

⇒ **need to rethink Tracking!**

Theorem. *Track-and-Stop with* ***C-tracking*** *is δ-PAC with* ***asymptotically optimal*** *sample complexity. Track-and-Stop with* ***D-tracking*** *may* ***fail*** *to converge.*

Multiple-answer Problems

Set-valued correct answer function $i^* : \mathcal{M} \to 2^{\mathcal{I}}$.



Rethinking the Lower Bound

Single-answer lower bound is based on **KL contraction**. With **multiple correct answers**, this gives the **wrong leading constant**. ⇒ we do direct proof.

Theorem. *Any δ-PAC algorithm verifies*

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]}{\ln(1/\delta)} \ge D(\boldsymbol{\mu})^{-1},$$

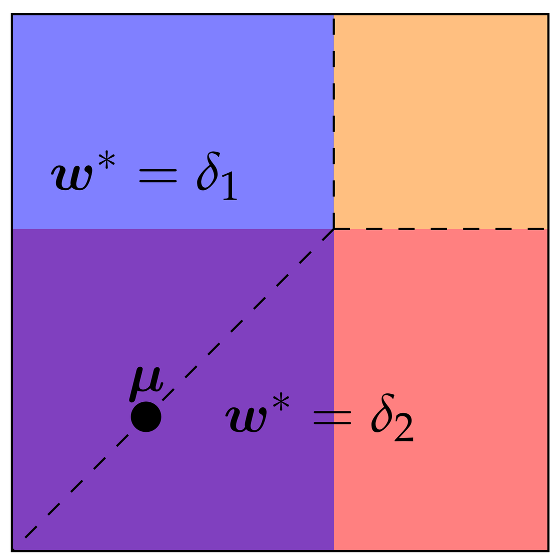
where $D(\boldsymbol{\mu}) = \max_{i \in i^*(\boldsymbol{\mu})} \max_{\boldsymbol{w} \in \triangle_K} \inf_{\boldsymbol{\lambda} \in \neg i} \sum_{k=1}^K w_k d(\mu_k, \lambda_k)$

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

On Matching the Lower Bound

Complication: $\boldsymbol{w}^*(\boldsymbol{\mu})$, the set of maximisers of $D(\boldsymbol{\mu})$, is **unsalvageably discontinuous**.

Example: Any Low Arm

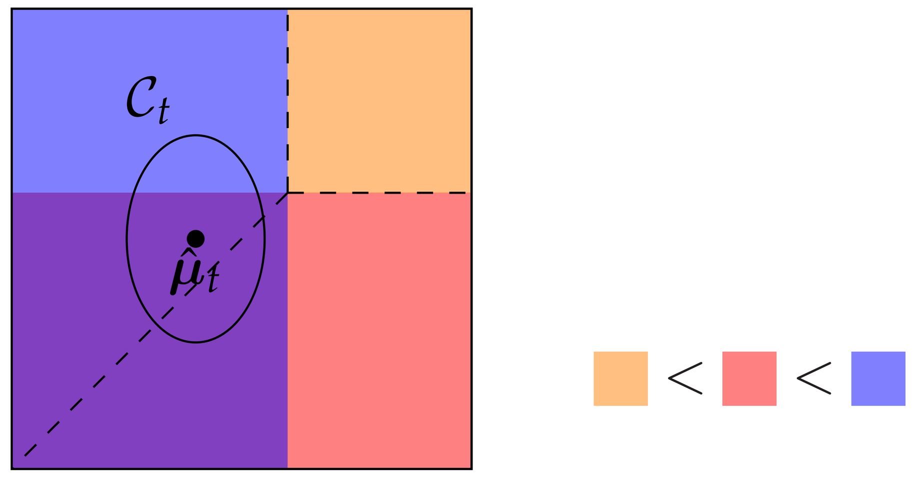


At $\boldsymbol{\mu}$, the oracle weights are $\boldsymbol{w}^*(\boldsymbol{\mu}) = \{\delta_1, \delta_2\}$. Tracking $\boldsymbol{w}^*(\hat{\boldsymbol{\mu}}_t)$ will play from convex hull.

Solution: Make it Sticky

New Sticky-Track-and-Stop Sampling Rule:

Find least (in "sticky order") oracle answer in confidence region $\mathcal{C}_t$. Track its oracle weights at $\hat{\boldsymbol{\mu}}_t$.



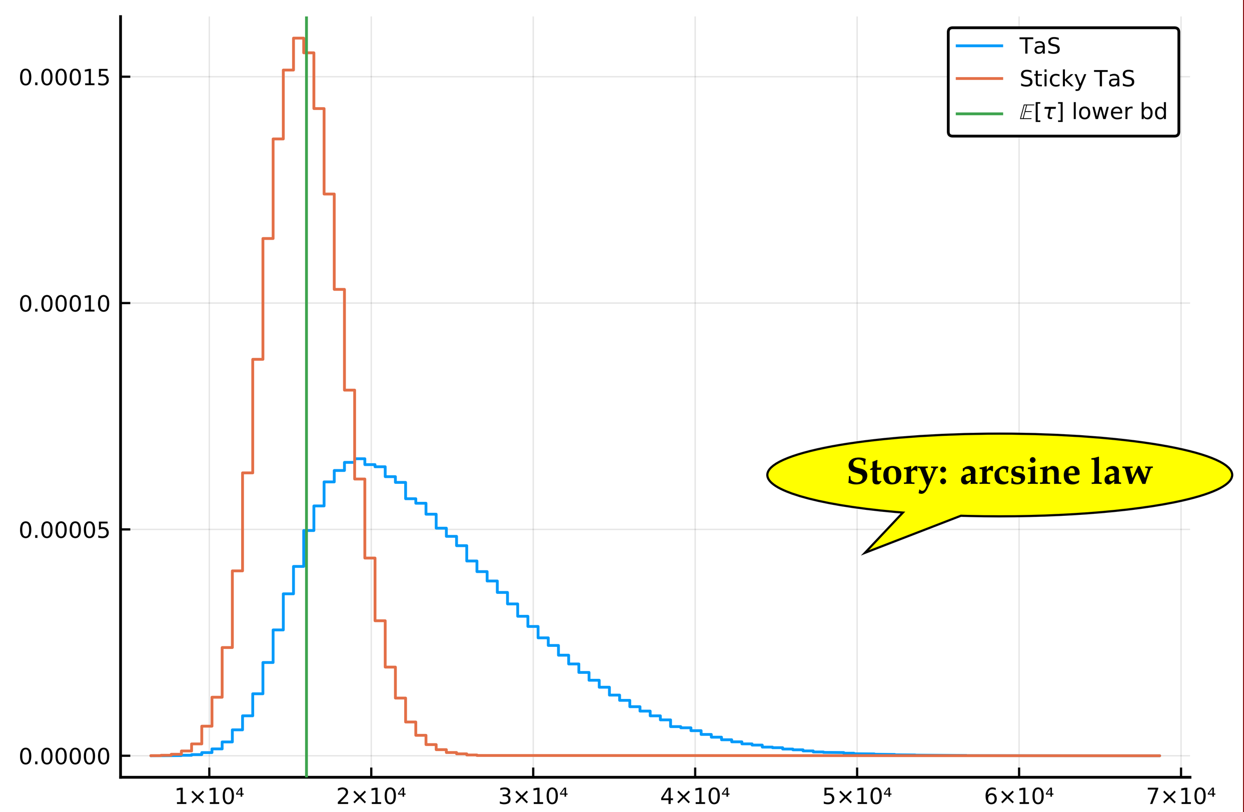
Main Result

When coupled with a good stopping rule,

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

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

Non-Sticky is Actually Dangerous



Histogram of stopping time, $\delta = e^{-80}$.

Where to go from here

- Practical efficiency
- Avoid forced exploration
- Regret (WIP), RL …
- Moderate confidence $\delta \not\to 0$ regime; bounds
- Understand sparsity patterns
- Dynamically expanding planning horizon