Non-Asymptotic Pure Exploration by Solving Games

Rémy Degenne, Wouter M. Koolen CWI Centrum Wiskunde & Informatica and Pierre Ménard Inria

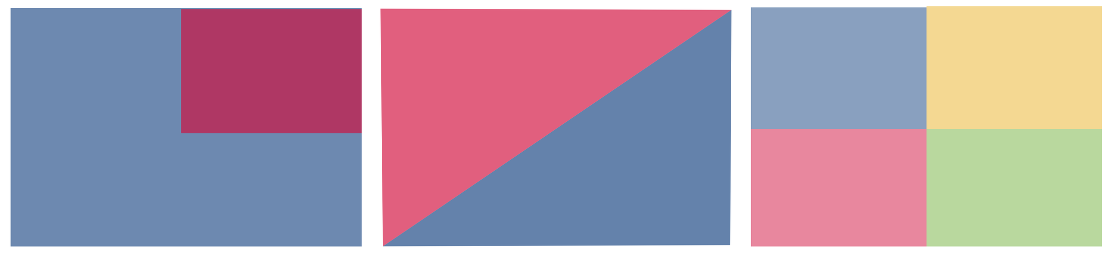
Topic: Pure Exploration

We want to answer a question about the parameters of a stochastic bandit.

- K arms with unknown distribution parameters (vector $\boldsymbol{\mu}$ of means).
- A query. Ex: is there an arm with mean $\mu < 0$? $\rightarrow$ correct answer at $\boldsymbol{\mu}$ is given by function $i^*(\boldsymbol{\mu})$
- At each stage, choose an arm and get an observation from the arm.
- Decide when to stop and return an answer.

Goals:

- answer correctly with probability $> 1 - \delta$.
- small sample complexity: stop at τ_δ s.t. $\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta]$ is small.



$\exists k, \mu^k < 0$? argmax$_k$ μ^k ? signs of all μ^k ?

Exploration as a Game

Lower Bound, with value of a game:

$$\mathbb{E}_{\boldsymbol{\mu}}[\tau_\delta] \geq \log(\frac{1}{\delta}) / \max_{\boldsymbol{w} \in \triangle_K} \inf_{\boldsymbol{\lambda} \in \neg i^*(\boldsymbol{\mu})} \sum_{k=1}^{K} w^k d(\mu^k, \lambda^k)$$

Optimal fixed sampling:
$\boldsymbol{w}^* = \text{argmax}_{\boldsymbol{w} \in \triangle_K} \inf_{\boldsymbol{\lambda} \in \neg i^*(\boldsymbol{\mu})} \sum_{k=1}^{K} w^k d(\mu^k, \lambda^k)$

Previous work: Track and Stop

- my estimate $\hat{\boldsymbol{\mu}}_t$ has answer YES.
- the optimal way to sample at $\hat{\boldsymbol{\mu}}_t$ from the lower bound is $\boldsymbol{w}_t^*$.
- I sample to track $\boldsymbol{w}_t^*$ (+ forced exploration)

Is asymptotically optimal (but sometimes very asymptotically).
Need $\text{argmax}_{\boldsymbol{w} \in \triangle_K} \inf_{\boldsymbol{\lambda} \in \neg i^*(\hat{\boldsymbol{\mu}}_t)} \sum_{k=1}^{K} w^k d(\hat{\mu}_t^k, \lambda^k)$ at every time step.
Relies on forced exploration, not adaptive to data.

Solving bandit pure exploration problems with games is computationally efficient and has optimal asymptotic sample complexity.

Our Strategy

Emulate Nature with a second algorithm.
Get two algorithms playing against each other.

- Alg: my estimate $\hat{\boldsymbol{\mu}}_t$ has answer YES.
- Nature: but here is $\boldsymbol{\lambda}_t$ with answer NO which could have generated the same data with relatively high probability.
- Alg: then I sample k_t s.t. if $\boldsymbol{\mu} \approx \hat{\boldsymbol{\mu}}_t$ (+optimism), I get maximal evidence for $\boldsymbol{\mu} \neq \boldsymbol{\lambda}_t$.

Why it works

As long as we do not stop:

$$\log\frac{1}{\delta} \geq \inf_{\boldsymbol{\lambda} \in \neg i_t} \sum_{k=1}^{K} N_t^k d(\mu^k, \lambda^k) \qquad \text{(stop rule)}$$

$$\approx \inf_{\boldsymbol{\lambda} \in \neg i^*} \sum_{s=1}^{t} \sum_{k=1}^{K} w_s^k d(\mu^k, \lambda^k) \qquad \text{(tracking)}$$

$$\geq \sum_{s=1}^{t} \sum_{k=1}^{K} w_s^k \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}_s} d(\mu^k, \lambda^k) - R_t^\lambda \qquad \text{(regret } \boldsymbol{\lambda}\text{)}$$

$$\geq \max_k \sum_{s=1}^{t} \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}_s} d(\mu^k, \lambda^k) - R_t^\lambda - R_t^k \qquad \text{(regret } k\text{)}$$

$$\geq t \inf_{\boldsymbol{q} \in \mathcal{P}(\neg i^*)} \max_k \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}} d(\mu^k, \lambda^k) - O(\sqrt{t})$$

But $\boldsymbol{\mu}$ unknown $\rightarrow$ optimism to explore efficiently.

Results

- Non-asymptotic sample complexity guarantees
- Asymptotically optimal
- Need only $\text{argmin}_{\boldsymbol{\lambda} \in \neg i^*(\hat{\boldsymbol{\mu}}_t)} \sum_{k=1}^{K} N_{t-1}^k d(\hat{\mu}_t^k, \lambda^k)$.
  $\rightarrow$ better computational complexity.
  Up to 100$\times$ faster than T-and-S on best arm identification.

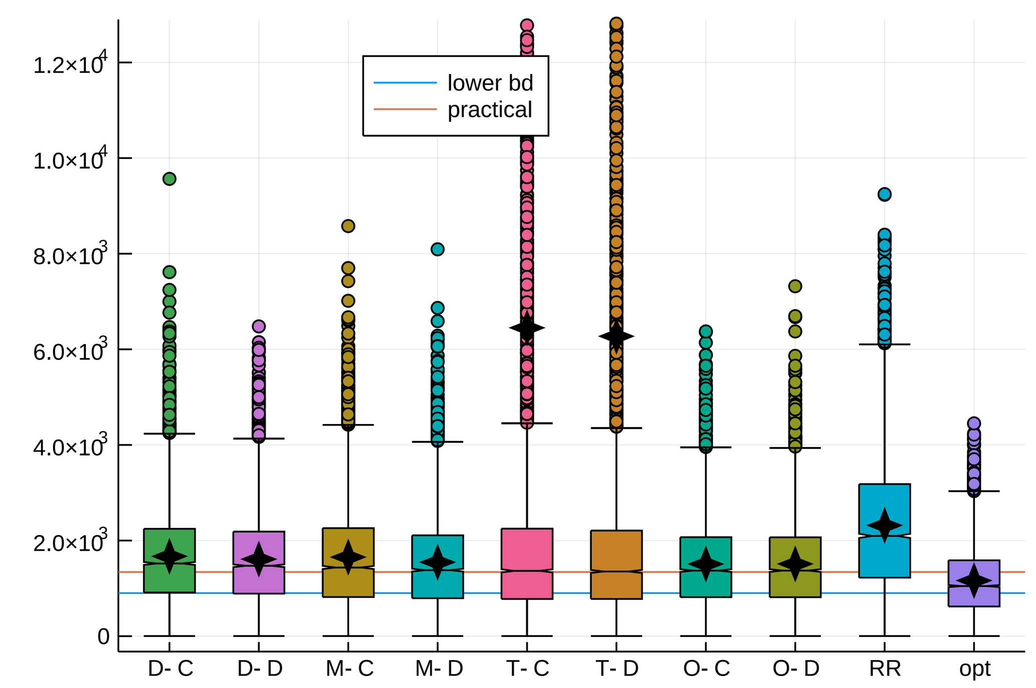


Figure 1:Track and Stop can fail, even for $\delta = e^{-10}$. Our algorithms are (provably) good.

Algorithm

Inputs:

- Algorithms $\mathcal{A}^k$ and $\mathcal{A}^{\boldsymbol{\lambda}}$, full information adversarial regret minimization algorithms.
- stopping threshold $\beta(t, \delta) \approx \log \frac{\log t}{\delta}$, exploration bonus $f(t) \approx \log t$.

Algorithm:

- Sample each arm once and form estimate $\hat{\boldsymbol{\mu}}_K$. For $t = K + 1, \ldots$
  - For $k \in [K]$, let $[\alpha_t^k, \beta_t^k] = \{\xi : N_{t-1}^k d(\hat{\mu}_{t-1}^k, \xi) \leq f(t-1)\}$. (KL confidence intervals)
  - Let $i_t = i^*(\hat{\boldsymbol{\mu}}_{t-1})$.
  - Stop and output $\hat{\imath} = i_t$ **if** $\inf_{\boldsymbol{\lambda} \in \neg i_t} \sum_k N_{t-1}^k d(\hat{\mu}_{t-1}^k, \lambda^k) > \beta(t, \delta)$. (GLRT Stopping rule)
  - Get $\boldsymbol{w}_t$ and $\boldsymbol{q}_t$ from $\mathcal{A}_{i_t}^k$ and $\mathcal{A}_{i_t}^{\boldsymbol{\lambda}}$.
  - For $k \in [K]$, let $U_t^k = \max_{\xi \in \{\alpha_t^k, \beta_t^k\}} \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}_t} d(\xi, \lambda^k)$. (Optimism)
  - Feed $\mathcal{A}_{i_t}^k$ the loss $\ell_t^{\boldsymbol{w}}(\boldsymbol{w}) = -\sum_{k=1}^{K} w^k U_t^k$.
  - Feed $\mathcal{A}_{i_t}^{\boldsymbol{\lambda}}$ the loss $\ell_t^{\boldsymbol{\lambda}}(\boldsymbol{q}) = \mathbb{E}_{\boldsymbol{\lambda} \sim \boldsymbol{q}} \sum_{k=1}^{K} w_t^k d(\hat{\mu}_{t-1}^k, \lambda^k)$.
  - Pick arm $k_t = \text{argmin}_k N_{t-1}^k - \sum_{s=1}^{t} w_s^k$. (Cumulative tracking)
  - Observe sample $X_t \sim \nu_{k_t}$. Update $\hat{\boldsymbol{\mu}}_t$.

I want to know more!

arXiv

Take a picture
to download the full paper.