Pure Exploration by Solving Games

Rémy Degenne, Wouter M. Koolen and Pierre Ménard

October 27, 2019

CWI
Centrum Wiskunde & Informatica
Inria

Main recipe

Take two adversarial strategies for regret minimization.

Add optimism.

Get one stochastic bandit algorithm for pure exploration.

Paper: R.D., Wouter M. Koolen and Pierre Ménard, Non-Asymptotic Pure Exploration by Solving Games, NeurIPS 2019.

Pure Exploration

Usual Queries

- Best Arm Identification
- Thresholding Bandit

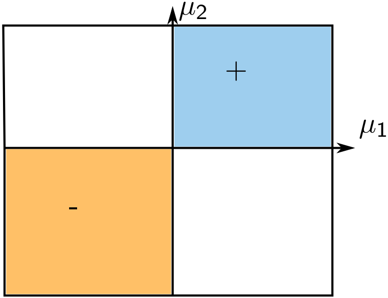
Our setting

- Bandit parametrized by means $\boldsymbol{\mu} \in \mathcal{M} \subset \mathbb{R}^K$.
- Answers $\mathcal{I}$. Correct answer function $i^* : \mathcal{M} \to \mathcal{I}$.
- Fixed confidence $\delta \in [0, 1]$.
- Algorithm stops at time τ_δ, returns $\hat{\imath}$.

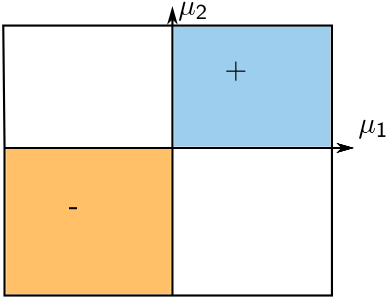
Goal: δ-correct algorithm, such that

$$\forall \boldsymbol{\mu} \in \mathcal{M} \quad \mathbb{P}_{\boldsymbol{\mu}}(\hat{\imath} \neq i^*(\boldsymbol{\mu})) \leq \delta \,, \qquad \mathbb{E}_{\boldsymbol{\mu}} \tau_\delta \text{ is minimal.}$$

Pure Exploration



Pure Exploration



This talk: about sampling rules.
Use GLRT stopping rule from Garivier and Kaufmann, 2016.

Sample complexity: what is "minimal"?

Lower Bound

Any δ-correct algorithm on $\mathcal{M}$ must verify for all $\boldsymbol{\mu} \in \mathcal{M}$,

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

$\neg i = \{\boldsymbol{\lambda} \in \mathcal{M} : i^*(\boldsymbol{\lambda}) \neq i\}$.

Sample complexity: what is "minimal"?

Lower Bound

Any δ-correct algorithm on $\mathcal{M}$ must verify for all $\boldsymbol{\mu} \in \mathcal{M}$,

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

$\neg i = \{\boldsymbol{\lambda} \in \mathcal{M} : i^*(\boldsymbol{\lambda}) \neq i\}$.

Follow the lower bound: attempt 1

Track and Stop

Compute estimated problem $\hat{\boldsymbol{\mu}}_t$.

Compute the solution $\boldsymbol{w}_t^*$ to

$$\operatorname*{argmax}_{w \in \triangle_K} \inf_{\boldsymbol{\lambda} \in \neg i^*(\hat{\boldsymbol{\mu}}_t)} \sum_{k=1}^{K} w^k d(\hat{\mu}_t^k, \lambda^k) .$$

If an arm is sampled less than $\sqrt{t}$, sample it (forced exploration).

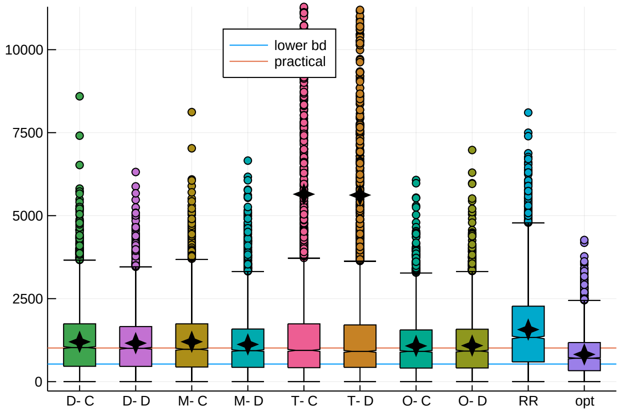
Otherwise, sample arm $k_t = \operatorname{argmin} N_{t-1}^k - (w_t^*)^k$ (tracking).

[Garivier and Kaufmann, Optimal Best Arm Identification with Fixed Confidence, 2016]

Track-and-Stop

- Asymptotically optimal,
- But sometimes only asymptotically.

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_{\boldsymbol{\mu}} \tau_\delta}{\log(1/\delta)} \le \frac{1}{\sup_{\boldsymbol{w}\in\triangle_K} \inf_{\boldsymbol{\lambda}\in\neg i^*(\boldsymbol{\mu})} \sum_{k=1}^K w^k d(\mu^k,\lambda^k)} .$$



Follow the lower bound: attempt 2

with games!

A Game

Suppose $\boldsymbol{\mu}$, $i = i^*(\boldsymbol{\mu})$ known.

- k-Player plays in $\{1, \ldots, K\}$.
- λ-Player plays in $\neg i$.
- zero-sum. reward for k-player: $d(\mu^k, \lambda^k)$.

After t iterations: reward $\sum_{s=1}^{t} d(\mu^{k_s}, \lambda_s^{k_s})$.

Algorithms

- Regret-minimizing algorithm for k: AdaHedge.
- Regret-minimizing algorithm for λ: Best-Response.
- Result: value $\frac{1}{t}\sum_{s=1}^{t} d(\mu^{k_s}, \lambda_s^{k_s})$ converges to max-min.

Follow the lower bound: attempt 2

with games!

A Game

Suppose $\boldsymbol{\mu}$, $i = i^*(\boldsymbol{\mu})$ known.

- k-Player plays in $\{1, \dots, K\}$.
- λ-Player plays in $\neg i$.
- zero-sum. reward for k-player: $d(\mu^k, \lambda^k)$.

After t iterations: reward $\sum_{s=1}^{t} d(\mu^{k_s}, \lambda_s^{k_s})$.

Algorithms

- Regret-minimizing algorithm for k: AdaHedge.
- Regret-minimizing algorithm for λ: Best-Response.
- Result: value $\frac{1}{t} \sum_{s=1}^{t} \sum_{k=1}^{K} w_s^k d(\mu^k, \lambda_s^k)$ converges to max-min.

Algorithm for Pure Exploration

At stage $t \in \mathbb{N}$,

- Compute $\hat{\boldsymbol{\mu}}_t$, define candidate answer i_t.
- Define game with optimistic reward $\max_{\xi \in [\hat{\mu}_t^k \pm \ldots]} d(\xi, \lambda^k)$.
- Do 1 iteration of each learner on optimistic game.
- Sample the arm prescribed by the k-player (tracking).

And stop according to GLRT stopping rule.

Computational Complexity

Track-and-Stop: solves one "max-min" at each stage.

$$\operatorname*{argmax}_{\boldsymbol{w}\in\triangle_K} \inf_{\boldsymbol{\lambda}\in\neg i^*(\hat{\boldsymbol{\mu}}_t)} \sum_{k=1}^{K} w^k d(\hat{\mu}_t^k, \lambda^k) .$$

AdaHedge + Best-response: solves one "min" at each stage.

$$\operatorname*{argmin}_{\boldsymbol{\lambda}\in\neg i_t} \sum_{k=1}^{K} w_t^k d(\hat{\mu}_t^k, \lambda^k) .$$

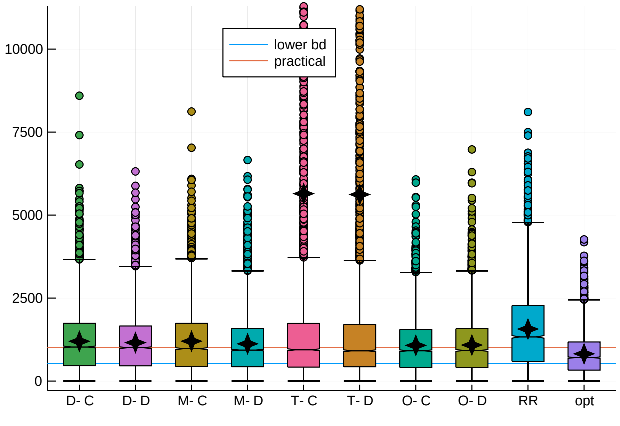
Examples

- Threshlolding: closed form vs closed form.
- BAI: (line search)^2 vs line-search.
- Many Problems (sparse, lipschitz, unimodal): complicated? vs convex.

Results

For all $\boldsymbol{\mu} \in \mathcal{M}$,

$$\mathbb{E}_{\boldsymbol{\mu}} \tau_\delta \leq \frac{\log(1/\delta)}{\max \inf \sum_{k=1}^{K} w^k d(\mu^k, \lambda^k)} \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{\log(1/\delta)}}\right)\right) .$$



Remarks

Variants

- Solve max-max-min at each stage $\Rightarrow$ lowest sample complexity.
- Use a learner for λ $\Rightarrow$ no tracking needed:
  - Follow the perturbed leader: always available but t samples at stage t,
  - Easy if union of few simple convex regions.

Open problem

What if only few samples are available?
What if we want $\delta = 1/4$?

Conclusion

- Pure Exploration is a very broad setting.
- The game point of view is successful.
- Many other applications possible in bandits.
- The small confidence regime is still unclear.

Conclusion

- Pure Exploration is a very broad setting.
- The game point of view is successful.
- Many other applications possible in bandits.
- The small confidence regime is still unclear.

Thank you!