Pure Exploration by Solving Games

Rémy Degenne, Wouter M. Koolen and Pierre Ménard

October 27, 2019
Main recipe

Take two adversarial strategies for regret minimization.

Add optimism.

Get one stochastic bandit algorithm for pure exploration.

Pure Exploration

Usual Queries

• Best Arm Identification
• Thresholding Bandit

Our setting

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

Goal: $\delta$-correct algorithm, such that

$$\forall \mu \in \mathcal{M} \quad \mathbb{P}_\mu(\hat{i} \neq i^*(\mu)) \leq \delta, \quad \mathbb{E}_\mu \tau_\delta \text{ is minimal.}$$
This talk: about sampling rules.
Use GLRT stopping rule from Garivier and Kaufmann, 2016.
This talk: about sampling rules. Use GLRT stopping rule from Garivier and Kaufmann, 2016.
Sample complexity: what is “minimal”?

Lower Bound
Any $\delta$-correct algorithm on $\mathcal{M}$ must verify for all $\mu \in \mathcal{M}$,

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

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

Sample complexity: what is “minimal”?

**Lower Bound**

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

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

$$
\neg i = \{ \lambda \in \mathcal{M} : i(\lambda) \neq i \}.
$$
Follow the lower bound: attempt 1

Track and Stop

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

Compute the solution $w_t^*$ to

$$\arg\max_{w \in \Delta_K} \inf_{\lambda \in -i^*(\hat{\mu}_t)} \sum_{k=1}^{K} w^k d(\hat{\mu}_t^k, \lambda^k).$$

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

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

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

- Asymptotically optimal,
- But sometimes only asymptotically.

\[
\liminf_{\delta \to 0} \frac{\mathbb{E}_{\mu} \tau_\delta}{\log(1/\delta)} \leq \frac{1}{\sup_{w \in \triangle_K} \inf_{\lambda \in \iota^*(\mu)} \sum_{k=1}^{K} w^k d(\mu^k, \lambda^k)}.
\]
Follow the lower bound: attempt 2
with games!

A Game
Suppose $\mu, i = i^*(\mu)$ known.

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

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

Algorithms

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

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

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

After $t$ iterations: reward $\sum_{s=1}^{t} d(\mu_{ks}, \lambda_{ks})$.

Algorithms

- Regret-minimizing algorithm for $k$: AdaHedge.
- Regret-minimizing algorithm for $\lambda$: Best-Response.
- Result: value $\frac{1}{t} \sum_{s=1}^{t} \sum_{k=1}^{K} w_s^k d(\mu_k, \lambda_k)$ converges to max-min.
Algorithm for Pure Exploration

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

- Compute $\hat{\mu}_t$, 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.

$$\arg\max_{w \in \Delta_K} \inf_{\lambda \in \arg\neg \lambda^* (\hat{\mu}_t)} \sum_{k=1}^{K} w^k d(\hat{\mu}_k, \lambda^k).$$

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

$$\arg\min_{\lambda \in \neg \lambda^t} \sum_{k=1}^{K} w^k_t d(\hat{\mu}_t^k, \lambda^k).$$

Examples

- Threshholding: closed form vs closed form.
- BAI: (line search)$^2$ vs line-search.
For all $\mu \in \mathcal{M}$,

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

Results
Remarks

Variants

- Solve \textit{max-max-min} at each stage \(\Rightarrow\) lowest sample complexity.
- Use a learner for \(\lambda\) \(\Rightarrow\) 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!