BANDITS WITH MANY OPTIMAL ARMS

Rianne de Heide, James Cheshire, Pierre Ménard and Alexandra Carpentier

MODEL

The stochastic multi-armed bandit

- A learner interacts with environment in rounds
- At each round, the learner chooses an arm to play, and receives a reward from the associated probability distribution
- Common assumptions:
  - There is a single optimal arm
  - The number of arms is small

We lift both assumptions

NOTATION

- Potentially infinite set $\mathcal{A}$ called the reservoir
- Each arm $a \in \mathcal{A}$ is associated with a probability distribution ν_a supported on $[0,1]$ with mean μ_a
- Highest mean $\mu^* = \max_{a \in \mathcal{A}}$ and second highest mean $\mu_{sub} = \sup_{a \in \mathcal{A}: \mu_a \neq \mu^*} \mu_a$
- Minimal gap $\Delta = \mu^* - \mu_{sub}$; we assume $\Delta > 0$
- There exists a partition $\mathcal{A} = \mathcal{A}^* \cup \mathcal{A}_{sub}$
- Proportion p^* of optimal arms
- $\mathfrak{B}_{\Delta,p^*}$: set of bandit problems of which the proportion of optimal arms is at least p^* and the suboptimality gap is at least Δ

SETTING

We fix the time horizon T. At each round t, the learner chooses an arm a_t by either playing a past arm or picking a new arm from the reservoir $\mathcal{A}$. The learner gets a reward $Y_t \sim \nu_{a_t}$.
We aim for either minimising cumulative regret:

$$R(T) = \sum_{t=1}^{T} \mu^* - \mu_{a_t},$$

or for identifying the best arm, while minimising the probability of outputting a suboptimal arm:

$$e(T) = \mathbb{P}(\hat{a}_T \notin \mathcal{A}^*).$$

BEST-ARM IDENTIFICATION

Algorithm: Elimination

- **Input:** $\bar{c}$
- **Initialise:** set $i \leftarrow 1$
- **while** $i < \log T / \bar{c}$ **do**
  Sample each arm in $\mathcal{A}_i$ a number of $t_i = \lfloor \bar{c}T/(|\mathcal{A}_i| \log T) \rfloor$ of times and compute their empirical means $(\hat{\mu}_i(a))_{a \in \mathcal{A}_i}$.
  Put in $\mathcal{A}_{i+1}$ the $1 \vee \lfloor |\mathcal{A}_i|/2 \rfloor$ arms that have highest empirical means and add op top of that $\lfloor |\mathcal{A}_i|/4 \rfloor$ new arms taken at random from $\mathcal{A}$.
  $i \leftarrow i+1$
  end
  Return any $\hat{a}_T \in \mathcal{A}_i$.

Upper bound

Theorem Set $\bar{c} = \log(4/3)$. Elimination satisfies

$$\mathbb{P}(\hat{a}_T \in \mathcal{A}^*) \geq 1 - 2\log(T) \exp\left(-c \frac{\Delta^2 p^* T}{\log T}\right),$$

where $c = \bar{c}/19200$.

Lower bound

Theorem Consider $\Delta \in (0, 1/4)$ and $p^* \in [0, 1/4]$. For any bandit algorithm, there exists a bandit problem in $\mathfrak{B}_{\Delta,p^*}$ such that

$$e(T) \geq \frac{1}{4} \exp\left(-Tp^* \frac{\Delta^2}{32}\right).$$

CUMULATIVE REGRET MINIMISATION

Algorithm: Sampling UCB

- **Input:** $\gamma \in (0,1), L \geq 1$
- **Initialise:** Pick $\mathcal{L}$, with $|\mathcal{L}| = L$ arms from $\mathcal{A}$. Sample each arm once.
- **for** $t = L+1$ to T do
  Compute for each arm $a \in \mathcal{L}$ the quantity

$$U_a^t = \hat{\mu}_a^t + \sqrt{\frac{\frac{\gamma^2}{4(1-\gamma)} + \log(\frac{\pi^2}{6}) + 2\log(N_a^t)}{2N_a^t}}$$

  Play $a_t = \arg\max_{a \in \mathcal{L}} U_a^t$
  end

Upper bound

Theorem For $T \geq 2, \gamma \in (0,1)$ and $L = \lceil 4\log(T)/(p^*\gamma^2) \rceil$, the expected cumulative regret of Sampling UCB is upper bounded as:

$$\mathbb{E}R(T) \leq O\left(\frac{\log(T)\log(1/\Delta)}{p^*\Delta}\right).$$

Lower bound

Theorem Consider $\Delta \in (0, 1/4)$ and $p^* \in (0, 1/4]$. For any bandit algorithm, there exists a bandit problem in $\mathfrak{B}_{\Delta,p^*}$ such that

$$\mathbb{E}R(T) \geq \min\left(\frac{1}{60} \frac{\log(\Delta^2 T/16)}{p^*\Delta}, \sqrt{T}\right).$$

ADAPTING TO p^*

In regret minimisation it is impossible to adapt to p^*, as follows from the following theorem.

Theorem Let $p^* \leq 1/4$ and $c > 0$ such that

$$T \geq 4\left(\frac{c\log(T)}{p^*\Delta^2}\right)^2.$$

For any bandit algorithm $\mathfrak{A}$ such that for all bandit problems $\mathfrak{B}_{\Delta,p^*}$, we have

$$\mathbb{E}R(T) \leq \frac{c\log(T)}{p^*\Delta},$$

one has that $\forall q^* \leq \frac{4p^*}{c}$ there exists a problem in $\mathfrak{B}_{\Delta,q^*}$ such that

$$\mathbb{E}R(T) \geq \frac{\sqrt{T}\Delta}{4}.$$

In best-arm identification, adapting to p^* is possible, as is done in our elimination algorithm.

AUTHORS

Rianne de Heide
INRIA Lille and CWI Amsterdam
r.de.heide@cwi.nl

James Cheshire
Otto von Guericke University Magdeburg
james.cheshire@ovgu.de

Pierre Ménard
Otto von Guericke University Magdeburg
pierre.menard@ovgu.de

Alexandra Carpentier
University of Potsdam
carpentier@uni-potsdam.de