Reinforcement Learning for Markovian Bandit: Is Posterior Sampling more Scalable than Optimism?

LIG WAX: May 12, 2022

Kimang KHUN
Supervised by Nicolas GAST and Bruno GAUJAL
Table of Content

• Problem Statement
• Learning Approaches
  • Optimism
  • Posterior Sampling
• Our Result
• Conclusion
Problem Statement

Agent

Environment \((r^*, P^*)\)
\[ r^* := (r_1, \ldots, r_n) \]
\[ P^* := (P_1, \ldots, P_n) \]

Goal: learning how to map state vectors to arms so as to maximize a numerical reward in an unknown and uncertain environment.

Activate one arm

State vector

Reward of active arm
Problem Statement

Goal: learning how to map state vectors to arms so as to maximize a numerical reward in an unknown and uncertain environment.

Exploitation: act greedily based on the observations collected so far.

Exploration: collect more observations.
Problem Statement

**Goal**: learning how to map state vectors to arms so as to maximize a numerical reward in an unknown and uncertain environment.

**Exploitation**: act greedily based on the observations collected so far.

**Exploration**: collect more observations.

**Challenge**: best trade-off between exploitation and exploration.
The Optimism Principle

OPTIMISM
It's the best way to see life.

Oh my! I'm flying!
The Optimism Principle

Optimism in Face of Uncertainty:
When you are uncertain, consider the best possible environment (reward-wise).
The Optimism Principle

Optimism in Face of Uncertainty:
When you are uncertain, consider the **best possible environment** (reward-wise).

[Diagram of Parameter Space with point \((r^*, P^*)\) and coordinate axes for \(r\) and \(P\).]
Optimism in Face of Uncertainty:
When you are uncertain, consider the best possible environment (reward-wise).
The Optimism Principle

Optimism in Face of Uncertainty:
When you are uncertain, consider the **best possible environment (reward-wise)**.
The Optimism Principle

Optimism in Face of Uncertainty:
When you are uncertain, consider the best possible environment (reward-wise).
The Optimism Principle

Optimism in Face of Uncertainty:
When you are uncertain, consider the **best possible environment (reward-wise)**.
Optimism in Face of Uncertainty:
When you are uncertain, consider the best possible environment (reward-wise).
The Optimism Principle

Optimism in Face of Uncertainty:
When you are uncertain, consider the **best possible environment** (reward-wise).

If the **best possible environment** is **correct**
=> no reward lost
**Exploitation**

If the **best possible environment** is **wrong**
=> learn useful information
**Exploration**

⇒ Build confidence set for each pair \((r_i, P_i)\)
⇒ Choose \((\bar{r_i}, \bar{P_i})_{i \in [n]}\) such that \((\bar{r}, \bar{P})\) is the **best possible environment**
Posterior Sampling

\[ \mathbb{P}(H \mid O) = \frac{\mathbb{P}(O \mid H) \mathbb{P}(H)}{\mathbb{P}(O)} \]
Posterior Sampling (a.k.a Thompson Sampling [Thompson, 1933])

Posterior Sampling:
Hypothesis: the environment is sampled from a certain distribution.
Sample an environment from posterior distribution and act greedily.

Parameter Space

Prior distribution

\( (r^*, P^*) \)

\( (r, P) \)

reward
Posterior Sampling (a.k.a Thompson Sampling [Thompson, 1933])

Posterior Sampling:
Hypothesis: the environment is sampled from a certain distribution. Sample an environment from posterior distribution and act greedily.

Parameter Space

Posterior distribution

\( r \)

\( P \)

\((r^*, P^*)\)
Posterior Sampling (a.k.a Thompson Sampling [Thompson, 1933])

Posterior Sampling:
Hypothesis: the environment is sampled from a certain distribution. Sample an environment from posterior distribution and act greedily.

Parameter Space

Posterior distribution

$r^*, P^*$

$(r, P)$

reward
Posterior Sampling (a.k.a Thompson Sampling [Thompson, 1933])

Posterior Sampling:
Hypothesis: the **environment** is sampled from a **certain distribution**. **Sample an environment** from posterior distribution and **act greedily**.

Parameter Space

- Posterior distribution
- $(r^*, P^*)$
- $r$
- $P$

reward
Posterior Sampling (a.k.a Thompson Sampling [Thompson, 1933])

Posterior Sampling:
Hypothesis: the environment is sampled from a certain distribution.
Sample an environment from posterior distribution and act greedily.

More observations
=> posterior concentrates on the true environment

Exploitation

Few observations
=> uncertainty in the estimate

Exploration

⇒ Choose prior distribution $\phi_i$ for each arm $i$
⇒ Compute posterior $\phi_i(\cdot | O)$ and sample each pair $(r_i, P_i) \sim \phi_i(\cdot | O)$
Our Result

• Runtime:
  • When \((r, P)\) is given, an optimal solution (Gittins index policy) can be computed in \((2/3)nS^3 + O(nS^2)\) [Gast et al., 2022]
  • The imaginary environment of both approaches is a Markovian bandit, Gittins index policy is applicable

• Learning Performance:
  • Keeping the estimate of \((r_i, P_i)_{i\in[n]}\) is linear in \(n\)

=> Both approaches are scalable.
Conclusion

• We show how the Optimism and Posterior Sampling approaches can be used to learn Markovian bandit problem.
• We conclude that both approaches are scalable in the number of arms.

Future Work

• What if the non-active arms also change state?