Optimism in the Face of Uncertainty

Raunak Kumar

Cornell University

Great Ideas in TCS, Fall 2020

December 13, 2020

Outline

Overview

Overview

- Sequential decision making problems typically involve an exploration-exploitation trade-off.
- The upper confidence bound technique:
  1. Compute an empirical estimate of some desired quantity.
  2. Add an "exploration term" to the empirical estimate.
  3. Exploit this modified estimate instead.

Stochastic Bandits

Setup

- There is a known **time horizon** T.
- The learner has access to a set of K **arms**, denoted by A.

Setup

- There is a known **time horizon** T.
- The learner has access to a set of K **arms**, denoted by A.
- For each arm a:
  - Let $\mathcal{D}_a$ be its **reward distribution** with support in $[0, 1]$.
  - Let $\mu(a) = \mathbb{E}[\mathcal{D}_a]$ be its **mean reward**.

Setup

- There is a known **time horizon** T.
- The learner has access to a set of K **arms**, denoted by A.
- For each arm a:
  - Let $\mathcal{D}_a$ be its **reward distribution** with support in $[0, 1]$.
  - Let $\mu(a) = \mathbb{E}[\mathcal{D}_a]$ be its **mean reward**.
- Let $\mu^* = \max_{a \in A} \mu(a)$ denote the **best mean reward**.
- Let $a^* \in \arg\max_{a \in A} \mu(a)$ denote any **optimal arm**.

Setup

- There is a known **time horizon** T.
- The learner has access to a set of K **arms**, denoted by A.
- For each arm a:
  - Let $\mathcal{D}_a$ be its **reward distribution** with support in $[0, 1]$.
  - Let $\mu(a) = \mathbb{E}[\mathcal{D}_a]$ be its **mean reward**.
- Let $\mu^* = \max_{a \in A} \mu(a)$ denote the **best mean reward**.
- Let $a^* \in \arg\max_{a \in A} \mu(a)$ denote any **optimal arm**.
- The learner does *not* know the true reward distributions.

Problem Protocol

In each round $t \in [T]$,

- The learner chooses an arm $a_t \in A$.
- It earns a reward $r_t \sim \mathcal{D}_{a_t}$.

Examples

- Slot Machines
- Medical Trials
- Dynamic Pricing
- Dynamic Procurement
- ...

Goal

- The learner's goal is to maximize $\mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$.
- If the learner knew the true reward distributions, this would be easy.

Goal

- The learner's goal is to maximize $\mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$.
- If the learner knew the true reward distributions, this would be easy.
  - Simply choose a^* in every round.
- But the learner does *not* know a^*.
  - This leads to the fundamental exploration-exploitation trade-off.
- So, we will measure a learner's performance in terms of its *regret*.

Regret

- **Regret** measures how well a learner performs compared to the best benchmark, which in this case is the best fixed arm.
- The cumulative regret after T rounds is defined as

$$R(T) = \mu^* \cdot T - \sum_{t=1}^{T} \mu(a_t). \tag{1}$$

Regret

- **Regret** measures how well a learner performs compared to the best benchmark, which in this case is the best fixed arm.
- The cumulative regret after T rounds is defined as

$$R(T) = \mu^* \cdot T - \sum_{t=1}^{T} \mu(a_t). \tag{1}$$

- Note that $R(T)$ is a random variable because it depends on the randomness in the rewards and the learner.
  - Therefore, we will usually analyze the **expected regret** $\mathbb{E}[R(T)]$.
- The goal of a learner is to choose actions that minimize regret.

Exploration-Exploitation Trade-Off

A key feature of a multi-armed bandit problem is the trade-off between

- **Exploration:** Find out more information about each arm.
- **Exploitation:** Choose the best arm so far.

Uniform Exploration

Idea

- If we knew the true means, then we'd simply choose a^*.
- Why don't we do the following?
  1. Compute an **empirical estimate** of the true means.
  2. Choose an arm with the **highest empirical mean**.

Uniform Exploration Algorithm

Algorithm 1: Uniform Exploration

Choose each arm N times
For arm $a \in A$, let $\bar{\mu}(a)$ be its empirical mean
Let $\hat{a} \in \arg\max_{a \in A} \bar{\mu}(a)$
Play arm $\hat{a}$ in all remaining rounds.

This algorithm explicitly **explores** in the first KN rounds and then **exploits** in the remaining $T - KN$ rounds.

Analysis - Clean Event

- Let the **confidence radius** be $r(a) = \sqrt{\frac{2 \log T}{N}}$.
- Using Hoeffding's inequality,

$$\Pr[|\bar{\mu}(a) - \mu(a)| \leq r(a)] \geq 1 - \frac{2}{T^4}. \tag{2}$$

- Using the union bound,

$$\Pr[\forall a \in A, |\bar{\mu}(a) - \mu(a)| \leq r(a)] \geq 1 - \frac{2}{T^3}. \tag{3}$$

Analysis - Clean Event

- Let the **confidence radius** be $r(a) = \sqrt{\frac{2 \log T}{N}}$.
- Using Hoeffding's inequality,

$$\Pr[|\bar{\mu}(a) - \mu(a)| \leq r(a)] \geq 1 - \frac{2}{T^4}. \tag{2}$$

- Using the union bound,

$$\Pr[\forall a \in A, |\bar{\mu}(a) - \mu(a)| \leq r(a)] \geq 1 - \frac{2}{T^3}. \tag{3}$$

- Define the above to be the **clean event**.
  - The clean event says that all empirical estimates $\approx$ true means.

Analysis - Regret

- Condition on the clean event.
- In the first KN rounds, the regret is at most 1 in each round.
- In the remaining $T - KN$ rounds, the regret is $\Delta(\hat{a}) = \mu(a^*) - \mu(\hat{a})$.

Analysis - Regret

- Condition on the clean event.
- In the first KN rounds, the regret is at most 1 in each round.
- In the remaining $T - KN$ rounds, the regret is $\Delta(\hat{a}) = \mu(a^*) - \mu(\hat{a})$.
- In order to bound $\Delta(\hat{a})$, observe that

$$\mu(a^*) - r(a^*) \leq \bar{\mu}(a^*) \leq \bar{\mu}(\hat{a}) \leq \mu(\hat{a}) + r(\hat{a}). \tag{4}$$

- Therefore,

$$\Delta(\hat{a}) \leq O\left(\sqrt{\frac{\log T}{N}}\right). \tag{5}$$

Analysis - Regret

- So, we have

$$R(T) \leq KN + O\left(\sqrt{\frac{\log T}{N}}\right) \cdot (T - KN) \tag{6}$$

$$\leq KN + O\left(\sqrt{\frac{\log T}{N}}\right) \cdot T. \tag{7}$$

Analysis - Regret

- So, we have

$$R(T) \leq KN + O\left(\sqrt{\frac{\log T}{N}}\right) \cdot (T - KN) \quad (6)$$

$$\leq KN + O\left(\sqrt{\frac{\log T}{N}}\right) \cdot T. \quad (7)$$

- If we choose $N = (T/K)^{2/3} O(\log T)^{1/3}$, then

$$R(T) \leq O\left((K \log T)^{1/3} T^{2/3}\right). \quad (8)$$

Analysis - Regret

Now, we can bound the expected regret as follows:

$$\mathbb{E}[R(T)] = \Pr[\text{clean event}]\mathbb{E}\left[R(T) \mid \text{clean event}\right] \tag{9}$$

$$+ \Pr[\text{dirty event}]\mathbb{E}\left[R(T) \mid \text{dirty event}\right] \tag{10}$$

$$\leq 1 \cdot O\left((K \log T)^{1/3} T^{2/3}\right) + \frac{2}{T^3} \cdot T \tag{11}$$

$$\leq O\left((K \log T)^{1/3} T^{2/3}\right). \tag{12}$$

Discussion

Pros:

- The algorithm is extremely simple.
- It provides a non-trivial regret bound.

Cons:

Discussion

Pros:

- The algorithm is extremely simple.
- It provides a non-trivial regret bound.

Cons:

- Suboptimal.
- The performance in the exploration phase is terrible.
  - Does not explore **adaptively**.

Upper Confidence Bound

Idea

- It's good to choose an arm if
  - it has not been chosen enough number of times yet,
  - or its empirical mean so far is high.
- Do not waste rounds exploring arms that
  - have already been chosen many times,
  - and have a low empirical mean.

Modified Clean Event

- Let the **confidence radius in round** t be

$$r_t(a) = \sqrt{\frac{2 \log T}{n_t(a)}}, \tag{13}$$

where $n_t(a)$ is the number of times arm a has been chosen in the first t rounds.

Modified Clean Event

- Let the **confidence radius in round** t be

$$r_t(a) = \sqrt{\frac{2 \log T}{n_t(a)}}, \tag{13}$$

  where $n_t(a)$ is the number of times arm a has been chosen in the first t rounds.
- Let $\bar{\mu}_t(a)$ denote the **empirical estimate of arm** a **in round** t.
- Then,

$$\Pr[\forall a \in A, t \in [T], |\bar{\mu}_t(a) - \mu(a)| \leq r_t(a)] \geq 1 - \frac{2}{T^2}. \tag{14}$$

- Define the above to be the **clean event**.

Confidence Bounds

- Define the **upper** and **lower confidence bounds in round** t as

$$\mathrm{UCB}_t(a) = \bar{\mu}_t(a) + r_t(a), \tag{15}$$
$$\mathrm{LCB}_t(a) = \bar{\mu}_t(a) - r_t(a). \tag{16}$$

- Define the **confidence interval in round** t as $[\mathrm{LCB}_t(a), \mathrm{UCB}_t(a)]$.

UCB1 Algorithm

Algorithm 2: UCB1

Try each arm once

In each round t, choose $a_t \in \arg\max_{a \in A} \mathrm{UCB}_t(a)$

UCB1 Algorithm

Algorithm 3: UCB1

1 Try each arm once
2 In each round t, choose $a_t \in \arg\max_{a \in A} \mathrm{UCB}_t(a)$

Note that the selection rule naturally incorporates exploration and exploitation because

$$\mathrm{UCB}_t(a) = \bar{\mu}_t(a) + O\left(\sqrt{\frac{2 \log T}{n_t(a)}}\right). \quad (17)$$

Analysis - Regret

- Condition on the clean event. Then,

$$\bar{\mu}_t(a_t) \leq \mu(a_t) + r_t(a_t). \tag{18}$$

- By the algorithm's selection rule,

$$\mathrm{UCB}_t(a^*) \leq \mathrm{UCB}_t(a_t). \tag{19}$$

Analysis - Regret

- Condition on the clean event. Then,

$$\bar{\mu}_t(a_t) \leq \mu(a_t) + r_t(a_t). \tag{18}$$

- By the algorithm's selection rule,

$$\mathrm{UCB}_t(a^*) \leq \mathrm{UCB}_t(a_t). \tag{19}$$

- Combining the above shows that

$$\begin{aligned} \mu(a^*) &\leq \mathrm{UCB}_t(a^*) && (20)\\ &\leq \mathrm{UCB}_t(a_t) && (21)\\ &= \bar{\mu}_t(a_t) + r_t(a_t) && (22)\\ &\leq \mu(a_t) + 2r_t(a_t). && (23) \end{aligned}$$

Analysis - Regret

- Therefore,

$$\Delta(a_t) = O(r_t(a_t)) = O\left(\sqrt{\frac{2\log T}{n_t(a)}}\right). \tag{18}$$

Analysis - Regret

- Therefore,

$$\Delta(a_t) = O(r_t(a_t)) = O\left(\sqrt{\frac{2\log T}{n_t(a)}}\right). \tag{18}$$

- Consider any arm $a \in A$.
- Let t be the last round when a is played. Then, $n_t(a) = n_T(a)$.

Analysis - Regret

- Therefore,

$$\Delta(a_t) = O(r_t(a_t)) = O\left(\sqrt{\frac{2\log T}{n_t(a)}}\right). \tag{18}$$

- Consider any arm $a \in A$.
- Let t be the last round when a is played. Then, $n_t(a) = n_T(a)$.
- Therefore,

$$\Delta(a) \leq O(r_t(a)) = O(r_T(a)) = O\left(\sqrt{\frac{2\log T}{n_T(a)}}\right). \tag{19}$$

- This shows that if an arm is played many times, then its gap will be small. This is precisely what allows us to bound the regret.

Analysis - Regret

- Let $R(t,a) = \Delta(a) n_t(a)$ denote the regret of arm a in the first t rounds.
- Then, we can write the cumulative regret as

$$R(t) = \sum_{a \in A} O\left(\sqrt{\frac{\log T}{n_t(a)}} n_t(a)\right) = O\left(\sqrt{\log T}\right) \sum_{a \in A} \sqrt{n_t(a)}. \quad (20)$$

Analysis - Regret

- Let $R(t,a) = \Delta(a) n_t(a)$ denote the regret of arm a in the first t rounds.
- Then, we can write the cumulative regret as

$$R(t) = \sum_{a \in A} O\left(\sqrt{\frac{\log T}{n_t(a)}} n_t(a)\right) = O\left(\sqrt{\log T}\right) \sum_{a \in A} \sqrt{n_t(a)}. \quad (20)$$

- Since square root is a concave function and $\sum_{a \in A} n_t(a) = t$,

$$R(t) = O\left(\sqrt{Kt \log T}\right). \quad (21)$$

Analysis - Regret

- Let $R(t,a) = \Delta(a) n_t(a)$ denote the regret of arm a in the first t rounds.
- Then, we can write the cumulative regret as

$$R(t) = \sum_{a \in A} O\left(\sqrt{\frac{\log T}{n_t(a)}} n_t(a)\right) = O\left(\sqrt{\log T}\right) \sum_{a \in A} \sqrt{n_t(a)}. \quad (20)$$

- Since square root is a concave function and $\sum_{a \in A} n_t(a) = t$,

$$R(t) = O\left(\sqrt{Kt \log T}\right). \quad (21)$$

- We can bound the expected regret as before and we have that for all rounds $t \in [T]$,

$$\mathbb{E}[R(t)] = O\left(\sqrt{Kt \log T}\right). \quad (22)$$

Discussion

Pros:

- Regret bound is **optimal**.
- The UCB trick is **widely applicable**.

MDP and UCBVI

(Quick Overview)

Markov Decision Processes

A **Markov decision process (MDP)** M is a tuple (S, A, P, r, T, μ), where

- S is a set of **states**,
- A is a set of **actions**,
- $P : S \times A \to \Delta(S)$ is a set of **transition probabilities**,
- $R : S \times A \to [0, 1]$ is a **reward function**,
- $T \in \mathbb{N}$ is the **time horizon**,
- $\mu \in \Delta(S)$ is an **initial state distribution**.

A stationary, randomized **policy** $\pi : S \to \Delta(A)$ is a mapping from states to distribution over actions.

Markov Decision Processes

The **dynamics** of an MDP are as follows:

- Sample an initial state $s_0 \sim \mu$.
- In each round $t = 0, 1, \ldots, T-1$:
    1. Choose an action $a_t \sim \pi(\cdot|s_t)$
    2. Observe reward $r_t = R(s_t, a_t)$
    3. Transition to the next state $s_{t+1} \sim P(\cdot|s_t, a_t)$

The goal of a learner is to learn a policy that maximizes $\mathbb{E}\left[\sum_{t=0}^{T-1} r_t\right]$.

MDP - Planning

If the MDP is **known**, i.e., the learner knows P and r, then the problem is “easy” to solve using dynamic programming.

MDP - Planning

If the MDP is **known**, i.e., the learner knows P and r, then the problem is “easy” to solve using dynamic programming.

What if the MDP is **unknown**?

MDP - Learning

- For simplicity, assume that the reward is known, but the **transition probabilities** are **unknown**.
- In each **episode** $n \in [N]$,
    - The learner chooses some policy π^n.
    - This policy is executed on $s_0^n \sim \mu$ for T rounds.
- The goal is to minimize the **regret** between the values of the optimal policy and the sequence of policies executed by the learner:

$$\mathbb{E}\left[\text{regret}\right] = \mathbb{E}\left[\sum_{n=1}^{N} V^* - V^{\pi^n}\right]. \tag{23}$$

Upper Confidence Bound Value Iteration (UCBVI)

- Think about value iteration (VI) as a black box that accepts an MDP as input and outputs the optimal policy for this MDP.
- The MDP is specified by its transition probabilities and reward function.

Upper Confidence Bound Value Iteration (UCBVI)

Algorithm 4: UCBVI

for $n = 1, 2, \ldots, N$ **do**

  Let $N_t^n(s, a)$ be the number of times we saw the state-action pair (s, a) in round t in the first $n - 1$ episodes

  Let $N_t^n(s, a, s')$ be the number of times we saw the state-action pair (s, a) in round t in the first $n - 1$ episodes and transitioned to state s'

  For all s, a, s', t, estimate the transition probabilities as

$$\hat{P}_t^n(s'|s, a) = \frac{N_t^n(s, a, s')}{N_t^n(s, a)}. \tag{24}$$

  Compute $\pi^n = \mathrm{VI}\left(\{\hat{P}_t^n, r_t + b_t^n\}_{t=1}^{T-1}\right)$

  Execute π^n

end

Upper Confidence Bound Value Iteration (UCBVI)

- The b_t^n terms are defined as

$$b_t^n(s,a) = O\left(T\sqrt{\frac{\ln(SATN/\delta)}{N_t^n(s,a)}}\right). \tag{24}$$

- As before, this term allows us to trade-off between exploration and exploitation.

Takeaway

Takeaway

- Sequential decision making problems typically involve an exploration-exploitation trade-off.
- The upper confidence bound technique:
  1. Compute an empirical estimate of some desired quantity.
  2. Add an "exploration term" to the empirical estimate.
  3. Exploit this modified estimate instead.

The End