CS861: Theoretical Foundations of Machine Learning Lecture 17 - 10/13/2023

University of Wisconsin–Madison, Fall 2023

Lect[illegible] bandits, the UCB algorithm

Lecturer: Kirthe[illegible] *Scribed by: Ransheng Guan, Yamin Zhou*

Disclaimer: *These n[illegible]jected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the instructor.*

In this lecture, we will introduce the K–armed bandit and then present the upper confidence bound algorithm. Let us first quickly review the bandit framework.

1. Let $\nu = \{\nu_a, a \in \mathcal{A}\}$ be a bandit model
2. On round t, learner chooses $A_t \in \mathcal{A}$ and observes reward X_t sampled from ν_{A_t}
3. A learner is characterized by a policy $\Pi = (\Pi_t)_{t \in N}$, where Π_t maps the history $\{(A_s, X_s)\}_{s=1}^{t-1}$ to an action in $\mathcal{A}$ (or, for randomized policies, to a (deterministic) distribution over $\mathcal{A}$
4. Let $\mu_a = \mathbb{E}_{X \sim \nu_a}[X]$ be the expected reward of action a, let $a^* \in \arg\max_{a \in \mathcal{A}} \mu_a$ be an optimal action, and let $\mu_* = \mu_{a^*}$ be the optimal value.
5. Regret
$$R_T = R_T(\pi, \nu) = T\mu^* - \mathbb{E}[\sum_{t=1}^{T} X_t]$$
where $\mathbb{E}$ is with respect to the distribution of the action-reward sequence $A_1, X_1, A_2, X_2, ..., A_T, X_T$ induced by the interaction between the policy π and bandit model ν.
We want $R_T \in \mathcal{O}(T)$, i.e. $\lim_{T \to \infty} \frac{R_T}{T} = 0$.

1 K-armed bandits

A K-armed bandit is a stochastic bandit model where the action space consists of K distinct actions. We will write $\mathcal{A} = [K]$. We will assume that each ν_i is σ sub-Gaussian, with σ known. That is,

$$\mathcal{P} = \{\nu = \{\nu_i, i \in [K]\},\ \nu_i \textit{ is } \sigma\textit{-sub-Gaussian } \forall i \in [K]\}$$

For convenient, assume without loss of generality that $1 \geq \mu_1 \geq \mu_2 \geq ... \geq \mu_K \geq 0$, where $\mu_i = \mathbb{E}_{X \sim \nu_i}[X]$. (The learner is not aware of the ordering.) Finally, let $\Delta_i = \mu_* - \mu_i = \mu_1 - \mu_i$ denote the *gap* between the optimal arm and arm i.

2 Explore-then-Commit

One of the simplest algorithms for bandit models is the explore-then-commit algorithm, which simply pulls each arm for a fixed number of rounds at the beginning, and for the remaining rounds, pulls the arm that appeared to be the best. We have stated this algorithm formal in Algorithm 1.

Algorithm 1 Explore-then-Commit Algorithm

Data: time horizon T, number of exploration rewards m $(\leq T/K)$

- *Exploration phase*: Pu the first mK rounds.
  - Let where $\widehat{\mu}_i = \frac{1}{m}\sum_{s=1}^{mK} \mathbb{1}(A_s = i)x_s$
- *Commit phase*: Pull g $T - mK$ rounds

Theorem 1. *Let $\mathcal{P}$ denote the class of σ-subGaussian bandit models, and let $\nu \in \mathcal{P}$, Then the ETC policy satisfies*

$$R_T(\nu) \leq m \sum_{i:\Delta_i>0} \Delta_i + (T - mK) \sum_{i:\Delta_i>0} \Delta_i \exp\left(\frac{-m\Delta_i^2}{4\sigma^2}\right)$$

If we choose $m \asymp K^{-1/3}T^{2/3}$, then

$$\sup_{\nu\in\mathcal{P}} R_T(\nu) \in \tilde{\mathcal{O}}(K^{1/3}T^{2/3})$$

The proof of this theorem will be in HW2

3 The Upper Confidence Bound Algorithm

The UCB algorithm is based on the principle of *optimism in the face of uncertainty*, where on each round we will pretend that the bandit model is as nice as *statistically plausible*. To state the algorithm, we will first define upper confidence bounds on each arm at the end of round t as follows:

$$N_{i,t} = \sum_{s=1}^{t} \mathbb{1}(A_s = i)$$

$$\hat{\mu}_{i,t} = \frac{1}{N_{i,t}} \sum_{s=1}^{t} \mathbb{1}(A_s = i)X_s \qquad \text{(undefined if } N_{i,t} = 0)$$

$$e_{i,t} = \sigma\sqrt{\frac{2\log(1/\delta_t)}{N_{i,t}}} \quad \text{where} \quad \delta_t = \frac{1}{T^2 t} \qquad \text{(undefined if } N_{i,t} = 0)$$

Then, $\hat{\mu}_{i,t} + e_{i,t}$ is an upper confidence bound for μ_i, and $\hat{\mu}_{i,t} - e_{i,t}$ is a lower confidence bound for μ_i.

We can now state the upper confidence bound algorithm, which stipulates that we choose the arm with the highest upper confidence bound $\hat{\mu}_{i,t-1} + e_{i,t-1}$ on each round. Intuitively, when you maximize $\hat{\mu}_{i,t-1} + e_{i,t-1}$, the $\hat{\mu}_{i,t-1}$ favors *exploitation*, and $e_{i,t-1}$ favors *exploration*. The algorithm is stated formally in Algorithm 2 below.

Algorithm 2 The Upper Confidence Bound Algorithm

Data: time horizon T, number of exploration rewards $m(\leq T/K)$

for $t = 1, \cdots, k$ **do**

  Pull arm t, i.e. $A_t = t$ and observe $X_t \sim \nu_t$

end

for $t = k+1, \cdots, T$ **do**

  Pull $A_t = \arg\max_{i\in[K]} \hat{\mu}_{i,t-1} + e_{i,t-1}$ and observe $X_t \sim \nu_{A_t}$ ▷ break ties arbitrarily

end

Theorem 2. *Let $\mathcal{P}$ denote the class of σ-subGaussian bandit models, and let $\nu \in \mathcal{P}$. Then the UCB policy satisfies*

$$\ldots K + \sum_{i;\Delta_i>0} \frac{24\sigma^2 \log(T)}{\Delta_i}.$$

Moreover,

$$\ldots \sigma\sqrt{96KT\log(T)} \in \tilde{\mathcal{O}}\left(\sqrt{KT}\right).$$

Here, the first boun[illegible] *p-dependent bound* while the second bound can be viewed as a *gap-independent boun*[illegible]*d*. If the gaps $\Delta_i = \mu_1 - \mu_i$ are large, then $R_T \in \mathcal{O}(\log(T))$. Otherwise $R_T \in \tilde{\mathcal{O}}\left(\sqrt{KT}\right)$

Before we prove our theorem, we will first state the following decomposition of the regret.

Lemma 1 (Regret decomposition). *(Applies to any policy, not just UCB)*

$$R_T(\nu) = \sum_{i,\Delta_i>0} \Delta_i \,\mathbb{E}[N_{i,T}],$$

where the expectation $\mathbb{E}$ is with respect to the action reward sequence $A_1, X_1, A_2, X_2, \ldots, A_t, X_t$

Proof

$$\begin{aligned}
R_T &= \sum_{t=1}^{T} (\mu_1 - \mathbb{E}[X_t]) \\
&= \sum_{t=1}^{T} \left(\mu_1 - \mathbb{E}[\sum_{i=1}^{K} \mathbb{1}(A_t = i) X_t]\right) \\
&= \sum_{t=1}^{T} \sum_{i=1}^{K} \mathbb{E}[(\mu_1 - X_t)\mathbb{1}(A_t = i)] \\
&= \sum_{i=1}^{K} \sum_{t=1}^{T} \mathbb{E}[\mathbb{E}[(\mu_1 - X_t)\mathbb{1}(A_t = i)|A_t]] \\
&= \sum_{i=1}^{K} \sum_{t=1}^{T} \mathbb{E}[\mathbb{1}(A_t = i)\,\mathbb{E}[(\mu_1 - X_t)|A_t]] \\
&= \sum_{i=1}^{K} \sum_{t=1}^{T} \mathbb{E}[\mathbb{1}(A_t = i)(\mu_1 - \mu_{At})] \quad \textit{(Integrating out observation)} \\
&= \sum_{i=1}^{K} \sum_{t=1}^{T} \mathbb{E}[\mathbb{1}(A_t = i)(\mu_1 - \mu_i)] \quad \textit{(it will be 0 if } A_1 \neq i\textit{)} \\
&= \sum_{i=1}^{K} \sum_{t=1}^{T} \mathbb{E}[\Delta_i \mathbb{1}(A_t = i)] \\
&= \sum_{i=1}^{K} \Delta_i \,\mathbb{E}[\sum_{t=1}^{T} \mathbb{1}(A_t = i)] \\
&= \sum_{i=1}^{K} \Delta_i \,\mathbb{E}[N_{i,t}]
\end{aligned}$$

The last step follows from the fact that

$$\frac{1}{N_{i,t}}\sum_{s=1}^{t}\mathbb{1}(A_s = i)X_s$$

$\square$

Proof Proof of Theor[...] that each arm samples rewards $y_{i,r_{r\in\mathbb{N}}}$ and we observe these samples one-by-one as [...]efore, we can write $\hat{\mu}_{i,t} = \frac{1}{N_{i,t}}\sum_{r=1}^{N_{i,t}} y_{i,r}$.

We now define the following good events, G_1, G_i, $\forall i$ s.t. $\Delta_i > 0$.

$$G_1 = \{\forall t > K,\ \mu_1 < \hat{\mu}_{1,t} + e_{1,t}\}$$

$$G_i = \{\forall t > K,\ \mu_i > \hat{\mu}_{i,t} - e_{i,t}\}$$

where G_1 indicates that the true mean is below the UCB, and G_i indicates that the true mean is above the LCB.

Claim 1. *We have,* $\mathbb{P}(G_1^c) \leq \frac{1}{T}$, *and* $\mathbb{P}(G_i^c) \leq \frac{1}{T}$

$$\begin{aligned}
\mathbb{P}(G_1^c) &= \mathbb{P}(\exists t > K,\ s.t.\ \mu_1 \geq \hat{\mu}_{1,t} + e_{1,t}) \\
&\leq \sum_{t>K} \mathbb{P}\left(\mu_1 > \hat{\mu}_{1,t} + e_{1,t}\right) \\
&= \sum_{t>K} \mathbb{P}\left(\mu_1 > \frac{1}{N_{1,t}}\sum_{r=1}^{N_{1,t}} y_{1,r} + \sigma\sqrt{\frac{2\log(1/\delta_t)}{N_{1,t}}}\right) \\
&\leq \sum_{t>K} \mathbb{P}\left(\exists s \in [t-K+1]\ s.t.\ \mu_1 > \frac{1}{s}\sum_{r=1}^{s} y_{1,r} + \sigma\sqrt{\frac{2\log(1/\delta_t)}{s}}\right) \\
&\leq \sum_{t>K}\sum_{s=1}^{t-K+1} \mathbb{P}\left(\frac{1}{s}\sum_{r=1}^{s}(y_{1,r} - \mu_1) < -\sigma\sqrt{\frac{2\log(1/\delta_t)}{s}}\right) \\
&\leq \sum_{t>K}\sum_{s=1}^{t-K+1} \exp\left(-\frac{s}{2\sigma^2}\cdot\sigma^2\cdot\frac{2\log(1/\delta_t)}{s}\right) \\
&= \sum_{t>K}\sum_{s=1}^{t-K+1} \frac{1}{T^2 t} \qquad \text{as} \quad \delta_t = \frac{1}{T^2 t} \\
&\leq \sum_{t>K} \frac{1}{T^2} \leq \frac{1}{T}
\end{aligned}$$

$\square$

Remark The trick we used in the fourth and fifth steps only works in K−armed bandits. For other bandit models, we usually use martingales.