AI-3001: Advanced Topics in Machine Learning **Aug-Nov 2022**

Lecture 15: Bandit Optimization

Lecturer: Ganesh Ghalme *Scribes: Ganesh Ghalme*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

15.1 Upper Confidence Bound (UCB) based Algorithm

In the UCB1 algorithm for each arm the algorithm maintains a UCB1 estimate and at each round the algorithm plays the arm with the highest UCB1 estimate. Such a UCB1 estimate for an arm $i \in [k]$ at round t is dependent on the empirical mean of the rewards of arm i and a confidence interval associated with arm i. To state it formally let $N_{i,t-1}$ denote the number of times arm i is pulled in $t-1$ rounds. Then the UCB1 estimate for arm $i \in [k]$ at round $t \geq 1$ is $\bar{\mu}_i(t) = 0$ if $N_{i,t-1} = 0$, otherwise $\bar{\mu}_i(t) = \hat{\mu}_{i,N_{i,t-1}}(t-1) + \sqrt{\frac{2\ln(t)}{N_{i,t-1}}}$ where $\hat{\mu}_{i,N_{i,t-1}}(t-1)$ is the empirical mean of the rewards of arm i after being pulled $N_{i,t-1}$ times in $t-1$ rounds and $\sqrt{\frac{2\ln(t)}{N_{i,t-1}}}$ is its associated confidence interval. For ease of notation, we will denote by c_{t,s_i} the confidence interval of arm i at time t when it is pulled s_i times i.e. $c_{t,s_i} = \sqrt{\frac{2\ln(t)}{s_i}}$. Technically for the first k rounds the algorithm plays each arm once to compute a non-zero UCB1 estimate for each arm and for every round $t \geq k+1$ it plays the arm with the highest UCB1 estimate. The total expected regret of UCB1 after T rounds is given by the following theorem, where $\Delta_i = \mu_1 - \mu_i$ for all $i \in [k]$, and $\Delta_i > 0$ as $\mu_1 > \mu_i$ for $i \neq 1$.

Theorem 15.1. *For the MAB problem, the* UCB*1 has expected regret* $\mathbb{E}[\mathcal{R}_{\mathrm{UCB}}(T)] \leq \sum_{i \neq 1} (\frac{8 \ln T}{\Delta_i}) + (1 + \frac{\pi^2}{3}) \sum_{i \in [k]} \Delta_i$.

Proof. To bound the regret of the UCB1 algorithm, we first upper bound $\mathbb{E}[N_{i,T}]$ for $i \neq 1$, i.e. the expected number of pulls of a sub-optimal arm $i \neq 1$ in T rounds. Denote the arm pulled by the algorithm at the t-th round as i_t. In the equation below $\mathbb{1}\{i_t = i\}$ is an indicator random variable that is equal to 1 if $i_t = i$ and is 0 otherwise. In general $\mathbb{1}\{\mathsf{E}\}$ denotes an indicator random variable that is equal to 1 if the event E is true and is 0 otherwise.

$$N_{i,T} = 1 + \sum_{t=k+1}^{T} \mathbb{1}\{i_t = i\}$$

For any positive integer ℓ we may rewrite the above equation as

$$N_{i,T} \leq \ell + \sum_{t=\ell}^{T} \mathbb{1}\{i_t = i, N_{i,t-1} \geq \ell\} \tag{15.1}$$

If $i_t = i$ then $\bar{\mu}_1(t) < \bar{\mu}_i(t)$ i.e. $\hat{\mu}_{1,N_{1,t-1}}(t-1) + c_{t,N_{1,t-1}} < \hat{\mu}_{i,N_{i,t-1}}(t-1) + c_{t,N_{i,t-1}}$. Hence from Equation 15.1

$$N_{i,T} = \ell + \sum_{t=t_\ell}^{T} \mathbb{1}\{\hat{\mu}_{1,N_{1,t-1}}(t-1) + c_{t,N_{1,t-1}} < \hat{\mu}_{i,N_{i,t-1}}(t-1) + c_{t,N_{i,t-1}}\ ,\ N_{i,t-1} \geq \ell\}$$

Here, t_ℓ is the time at which arm i is pulled for ℓ number of times. If arm i is not pulled for ℓ times in the entire run of an algorithm, we will use the upper bound of ℓ.

$$\leq \ell + \sum_{t=t_\ell}^{T} \mathbb{1}\{ \min_{0<s_1<t} \hat{\mu}_{1,s_1}(t-1) + c_{t,s_1} < \max_{\ell \leq s_i < t} \hat{\mu}_{i,s_i}(t-1) + c_{t,s_i} \}$$
$$\leq \ell + \sum_{t=t_\ell}^{T} \sum_{s_1=1}^{t} \sum_{s_i=\ell}^{t} \mathbb{1}\{\hat{\mu}_{1,s_1}(t-1) + c_{t,s_1} < \hat{\mu}_{i,s_i}(t-1) + c_{t,s_i}\}$$

At time t, $\hat{\mu}_{1,s_1}(t-1) + c_{t,s_1} < \hat{\mu}_{i,s_i}(t-1) + c_{t,s_i}$ implies that at least one of the following events is true

$$\{\hat{\mu}_{1,s_1}(t-1) \leq \mu_1 - c_{t,s_1}\} \tag{15.2}$$
$$\{\hat{\mu}_{i,s_i}(t-1) \geq \mu_i + c_{t,s_i}\} \tag{15.3}$$
$$\{\mu_1 < \mu_i + 2c_{t,s_i}\} \tag{15.4}$$

The probability of the events in Equations 15.2 and 15.3 can be bounded using Hoeffding's inequality as:

$$\mathbb{P}(\{\hat{\mu}_{1,s_1}(t-1) \leq \mu_1 - c_{t,s_1}\}) \leq t^{-4}$$
$$\mathbb{P}(\{\hat{\mu}_{i,s_i}(t-1) \geq \mu_i + c_{t,s_i}\}) \leq t^{-4}$$

The event in equation 15.4 $\{\mu_1 < \mu_i + 2c_{t,s_i}\}$ can be written as $\{\mu_1 - \mu_i - 2\sqrt{\frac{2\ln t}{s_i}} < 0\}$. Substituting $\Delta_i = \mu_1 - \mu_i$ and if $s_i \geq \lceil \frac{8 \ln T}{\Delta_i^2} \rceil \geq \lceil \frac{8 \ln t}{\Delta_i^2} \rceil$ then

$$\mathbb{P}\Big(\Big\{\Delta_i - 2\sqrt{\frac{2\ln t}{s_i}} < 0\Big\}\Big) = 0 \tag{15.5}$$

Thus if $\ell = \lceil \frac{8 \ln T}{\Delta_i^2} \rceil$ then

$$N_{i,T} \leq \lceil \frac{8 \ln T}{\Delta_i^2} \rceil + \sum_{t=\frac{8\ln T}{\Delta_i^2}}^{T} \sum_{s_1=1}^{t} \sum_{s_i=\frac{8\ln T}{\Delta_i^2}}^{t} \mathbb{1}\{\hat{\mu}_{1,s_1}(t-1) + c_{t,s_1} < \hat{\mu}_{i,s_i}(t-1) + c_{t,s_i}\}$$
$$\mathbb{E}[N_{i,T}] \leq \lceil \frac{8 \ln T}{\Delta_i^2} \rceil + \sum_{t=\frac{8\ln T}{\Delta_i^2}}^{T} \sum_{s_1=1}^{t} \sum_{s_i=\frac{8\ln T}{\Delta_i^2}}^{t} 2t^{-4}$$
$$\leq \quad \lceil \frac{8 \ln T}{\Delta_i^2} \rceil + \sum_{t=\frac{8\ln T}{\Delta_i^2}}^{\infty} \sum_{s_1=1}^{t} \sum_{s_i=\frac{8\ln T}{\Delta_i^2}}^{t} 2t^{-4} \leq \frac{8\ln T}{\Delta_i^2} + 1 + \frac{\pi^2}{3}$$

In the last inequality we use $\sum_{t=\lceil \frac{8\ln T}{\Delta_i^2} \rceil}^{\infty} \ \sum_{s_1=1}^{t} \ \sum_{s_i=\lceil \frac{8\ln t}{\Delta_i^2} \rceil}^{t} 2t^{-4} \leq \sum_{t=1}^{\infty} 2t^{-2} = \frac{\pi^2}{3}$. Recall from Section **??**, Equation **??**, that

$$\mathbb{E}[\mathcal{R}_{\text{UCB}}(T)] = \sum_{i\in[k]} \Delta_i \cdot \mathbb{E}[N_{i,T}] \leq \sum_{i\neq 1} \frac{8\ln T}{\Delta_i} + (1 + \frac{\pi^2}{3}) \cdot \sum_{i\in[k]} \Delta_i$$

$\square$

15.1.0.1 *Distribution-free Regret Bound for* UCB*1*

Theorem 15.2. *For the MAB problem, the* UCB*1 has expected (distribution-free) regret* $\mathbb{E}[\mathcal{R}_{\text{UCB}}(T)] = O(\sqrt{T \ln T})$.

Proof. Recall from Section 15.1 that the expected cumulative regret of the UCB1 algorithm in any round T is given by

$$\mathbb{E}[\mathcal{R}_{\text{UCB}}(T)] = \sum_{i \in [k]} \Delta_i \cdot \mathbb{E}[N_{i,T}].$$

To bound the above quantity, we begin by defining the event

$$C := \left\{ |\hat{\mu}_i(t) - \mu_i| \leq \sqrt{\frac{2 \ln T}{N_{i,t}}}, \forall i \in [k], \forall t \leq T \right\}.$$

By applying Hoeffding's inequality, and taking union bound, we get

$$\mathbb{P}(\bar{C}) \leq \frac{2kT}{T^4} \leq \frac{2}{T^2}.$$

Next, we will bound the value of $\mathbb{E}[\mathcal{R}_{\text{UCB}}(T)]$ by conditioning on C and $\bar{C}$. Let us first bound $\mathbb{E}[\mathcal{R}_{\text{UCB}}(T)|C]$. Assume the event C holds and some arm $i_t \neq 1$ is pulled in round $t \in [T]$. Then, by definition of UCB1 algorithm, we have $\bar{\mu}_1(t) < \bar{\mu}_i(t)$. Then,

$$\begin{aligned} \mu_1 - \mu_{i_t} &\leq \mu_1 - \mu_{i_t} + \bar{\mu}_i(t) - \bar{\mu}_1(t) \\ &= (\mu_1 - \bar{\mu}_1(t)) + (\bar{\mu}_i(t) - \mu_{i_t}) \end{aligned}$$

Since event C holds, we have

$$\mu_1 - \bar{\mu}_1(t) = \mu_1 - \hat{\mu}_1(t-1) - \sqrt{\frac{2 \ln T}{N_{i,t-1}}} \leq 0.$$

and

$$\bar{\mu}_i(t) - \mu_{i_t} = \hat{\mu}_{i_t}(t-1) - \mu_{i_t} + \sqrt{\frac{2 \ln T}{N_{i_t,t-1}}} \leq 2 \cdot \sqrt{\frac{2 \ln T}{N_{i_t,t-1}}}.$$

Therefore,

$$\mu_1 - \mu_{i_t} \leq 2 \cdot \sqrt{\frac{2 \ln T}{N_{i_t,t-1}}} \tag{15.6}$$

Now, consider any arm $i \in [k]$ and consider the last round $t_i \leq t$ when this arm was last pulled. Since the arm has not been pulled between t_i and t, we know $N_{i,t_i} = N_{i,t-1}$. Hence, applying the inequality in Equation 15.6 to arm i in round t_i, we get

$$\mu_1 - \mu_i \leq 2 \cdot \sqrt{\frac{2 \ln T}{N_{i,t-1}}}, \text{ for all} t \leq T$$

. Thus, the regret in t rounds is bounded by

$$\mathcal{R}(t) = \sum_{i \in [k]} \Delta_i \cdot N_{i,t} \leq 2\sqrt{2 \ln T} \cdot \sum_{i \in [k]} \sqrt{N_{i,t}}.$$

Square root is a concave function, and hence from Jensen's inequality, we obtain

$$\sum_{i \in [k]} \sqrt{N_{i,t}} \leq \sqrt{kt}.$$

Therefore, we have

$$\mathbb{E}[\mathcal{R}_{\text{UCB}}(T)|C] \leq 2\sqrt{2kt\ln T}.$$

Hence, the expected cumulative regret in t rounds can be bounded as

$$\begin{aligned}\mathbb{E}[\mathcal{R}_{\text{UCB}}(T) &= \mathbb{E}[\mathcal{R}_{\text{UCB}}(T)|C]\mathbb{P}(C) + \mathbb{E}[\mathcal{R}_{\text{UCB}}(T)|\bar{C}]\bar{\mathbb{C}} \\ &\leq 2\sqrt{2kt\ln T} + t\cdot\frac{2}{T^2} \\ &= O(\sqrt{kt\ln T}), \quad \forall t \leq T\end{aligned}$$

Thus, the distribution-free regret bound of UCB1 algorithm at some time T is $O(\sqrt{KT\ln T})$. □

15.2 Thompson Sampling Algorithm

Next we consider the Thompson sampling algorithm. We will not prove the regret guarantee of the Thompson Sampling algorithm here. Students can refer to [Agarwal et al and Kauffman et al] for the detailed proof. We provide important ideas in this lecture.

Algorithm 1: Thompson Sampling for Bernoulli Bandits

Input: number of arms K
Initialize: $S_i(0)$ $=F_i(0)$ $=0$ $\forall i$
for $t = 1, 2\cdots$ **do**

- **Sample:**
  $\lambda_i(t) \sim Beta(1 + S_i(t-1), 1 + F_i(t-1))$ $\forall i$
- **Play arm:**
  $i_t = \arg\max_i \lambda_i(t)$
- **Observe Reward:**
  $X_{i_t}(t) \in \{0, 1\}$
- **Update:**
  - $S_{i_t}(t) = S_{i_t}(t-1) + X_{i_t}(t)$
  - $F_{i_t}(t) = F_{i_t}(t-1) + 1 - X_{i_t}(t)$
  - $S_i(t) = S_i(t-1),\ F_i(t) = F_i(t-1)$ $\forall i \neq i_t$

Theorem 15.3. *For any $\epsilon > 0$, there exists a problem dependent constant $C(\epsilon, \mu_1, \mu_2, \ldots, \mu_K)$ such that the regret of Thompson sampling algorithm is given as:*

$$R_T(\text{THOMPSONSAMPLING}) \leq (1+\epsilon)\sum_{i,\mu_i\neq\mu^*}\frac{\Delta_i(\ln T + \ln\ln T)}{D(\mu_i||\mu^*)} + C(\epsilon, \mu_1, \mu_2, \ldots, \mu_K).$$

15.2.1 Intuition

TODO!

15.3 Other Bandit Algorithms (Beyond bounded rewards)

TODO! KL-UCB, Bayes-UCB, (α, ψ)-UCB, UCB-Normal, MOSS (todo: introduce instance independent regret), Guassian Bandits, Best arm identification (pure exploration setting), infinite arms, Linear Bandits (LinUCB), Combinatorial bandits and semi-bandit feedback, knapsack-bandits, sleeping bandits ...

15.4 EXP4 (Contextual Bandits framework)

TODO!

15.5 Markovian Bandits

TODO! Gittins index, Whittle index, different interpretations, indexability