DDA4230 Reinforcement learning Thompson sampling

Lecture 7

Lecturer: Guiliang Liu *Scribe: Baoxiang Wang, Jing Dong*

1 Goal of this lecture

The goal is to understand algorithms based on Thompson sampling (TS), in terms of the regret analysis and the underlying Bayesian perspective. For applications, students should also gain some intuition about different algorithms' advantages.
Suggested reading: Chapter 36 of *Bandit algorithms*; *A tutorial on Thompson sampling* by Russo, van Roy, Kazerouni, Osband, and Wen; *Analysis of Thompson Sampling for the Multi-armed Bandit Problem* by Agrawal and Goyal; *Further optimal regret bounds for Thompson sampling* by Agrawal and Goyal; *An information-theoretic analysis of Thompson sampling* by Russo and Van Roy;

2 Recap: ε-greedy, ETC, and UCB

For ε-greedy, by choosing $\varepsilon_t = \min\{1, Ct^{-1}\Delta_{\min}^{-2} m\}$ for some constant C, the regret satisfies

$$\overline{R}_T \le C' \sum_{i \ge 2} \left(\Delta_i + \frac{\Delta_i}{\Delta_{\min}^2} \log \max \left\{ e, \frac{T\Delta_{\min}^2}{m} \right\} \right), \tag{1}$$

where C' is an absolute constant.

For ETC under 2-armed bandits, when $T \ge 4\sqrt{2\pi e}/\Delta^2$, by choosing $k = \lceil \frac{2}{\Delta^2} W(\frac{T^2\Delta^4}{32\pi}) \rceil$, the regret satisfies

$$\overline{R}_T \le \Delta + \frac{2}{\Delta}(\log \frac{T^2\Delta^4}{32\pi} - \log\log \frac{T^2\Delta^4}{32\pi} + \log(1 + \frac{1}{e}) + 2), \tag{2}$$

where $W(y)\exp(W(y)) = y$ denotes the Lambert function.

For UCB, by setting $\delta = T^{-2}$, the regret satisfies

$$\overline{R}_T \le 3\sum_{i=1}^{m} \Delta_i + \sum_{i:\Delta_i > 0} \frac{16 \log T}{\Delta_i}.$$

This result is followed by a series of improvements.

3 Recap: Bayesian statistics and Bernoulli-Beta conjugate

Recall that the reward $r(i)$ of arm i follows some distribution. Assume that the reward distributions of arms belong to the same family with respective parameters, which writes

$$r(i) \sim p(x \mid \theta_i).$$

Recall that when estimating θ, the posterior is

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{\int_{\theta'} p(x \mid \theta')p(\theta')d\theta'} .$$

In Bayesian probability theory, if the posterior distributions $p(\theta \mid x)$ are in the same probability distribution family as the prior probability distribution $p(\theta)$, the prior and posterior are then called conjugate distributions, and the prior is called a conjugate prior for the likelihood function $p(x \mid \theta)$. Some infamous conjugate priors are Gaussian-Gaussian, Bernoulli-Beta, Poisson-Gamma, categorical-Dirichlet. Conjugate priors are convenient in analyses.

The Bernoulli-Beta is important for Thompson sampling for Bernoulli bandits. Recall that the Beta distribution $\text{Beta}(\alpha, \beta)$ with parameter $\theta = \{\alpha, \beta\}$ follows the probability density function of

$$p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} ,$$

where $\Gamma(z) = \int_0^\infty x^{z-1} \exp(-x)dx$, $z \in \mathbb{C}$ is the Gamma function. When $p(\theta) \sim \text{Beta}(\alpha_0, \beta_0)$ and we observe $x_1, \ldots, x_{\alpha'+\beta'} \sim x$ i.i.d. with α' ones and β' zeros, then

$$\begin{aligned}
p(\theta \mid x_1, \ldots, x_{\alpha'+\beta'}) &= \frac{p(x_1, \ldots, x_{\alpha'+\beta'} \mid \theta)p(\theta)}{\int_{\theta'} p(x_1, \ldots, x_{\alpha'+\beta'} \mid \theta')p(\theta')d\theta'} \\
&= \frac{\binom{\alpha'+\beta'}{\alpha'}\theta^{\alpha'}(1-\theta)^{\beta'} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1}}{\int_{\theta'} p(x_1, \ldots, x_{\alpha'+\beta'} \mid \theta')p(\theta')d\theta'} \\
&= \frac{\binom{\alpha'+\beta'}{\alpha'} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}}{\int_{\theta'} p(x_1, \ldots, x_{\alpha'+\beta'} \mid \theta')p(\theta')d\theta'}\theta^{\alpha_0+\alpha'-1}(1-\theta)^{\beta_0+\alpha'-1} \\
&\sim \text{Beta}(\alpha_0 + \alpha', \beta_0 + \beta') .
\end{aligned}$$

This implies that if our current belief of $\mu_i = \theta$ is $\text{Beta}(\alpha, \beta)$ and we observe a new data $x \sim \text{Ber}(\theta)$, then we update $\alpha \mathrel{+}= 1$ or $\beta \mathrel{+}= 1$ when $x = 1$ or $x = 0$, respectively.

4 Thompson sampling algorithms

4.1 The first algorithm in bandits

We return to where it all began, in bandits, to the first algorithm proposed by Thompson in 1933. The idea is a simple one. Before the game starts, the learner chooses a prior over a set of possible bandit environments. In each round, the learner samples an environment from the posterior and acts according to the optimal action in that environment.

The exploration in Thompson sampling comes from the randomization. If the posterior is poorly concentrated, then the fluctuations in the samples are expected to be large and the policy will likely explore. On the other hand, as more data is collected, the posterior concentrates towards the true environment and the rate of exploration decreases. We discuss finite-armed stochastic bandits, but Thompson sampling has been extended to all kinds of models (see Chapter 36 of the book).

Randomization is crucial and can be useful in both stochastic bandits and adversarial bandits (see Chapters 23 and 32 of the book for examples). We should be wary, however, that injecting noise into our algorithms might come at a cost in terms of variance. What is gained or lost by the randomization in Thompson sampling is still not clear, but we leave this cautionary note as a suggestion to the reader to think about some of the costs and benefits.

Algorithm 1: Thompson sampling (Bernoulli bandits)

Input: Prior α_0, β_0
Output: $a_t, t \in [T]$
Initialize $\alpha_i = \alpha_0$, $\beta_i = \beta_0$, for $i \in [m]$
while $t \leq T - 1$ **do**
  Sample $\theta_i(t) \sim \text{Beta}(\alpha_i, \beta_i)$ independently for $i \in [m]$
  $a_t = \arg\max_{i \in [m]} \theta_i(t)$ with arbitrary tiebreaker
  If $r_t = 1$ then α_{a_t} += 1; If $r_t = 0$ then β_{a_t} += 1;

A general TS algorithm works on any conjugate priors. When the family of the underlying reward distribution is unknown, a Gaussian-Gaussian conjugate (the non-informative prior) can be useful.

Algorithm 2: Thompson sampling

Input: Prior θ_0
Output: $a_t, t \in [T]$
Initialize $\theta_i = \theta_0$, for $i \in [m]$
while $t \leq T - 1$ **do**
  Sample independently for $i \in [m]$, $\theta_i(t) \sim p(\theta \mid \{r_{t'}\}_{\mathbb{1}\{a_{t'}=i, t' \leq t-1\}})$
  $a_t = \arg\max_{i \in [m]} \theta_i(t)$ with arbitrary tiebreaker
  Update the posterior probability distribution of $\theta_{a_t}(t+1)$ by

$$p(\theta_{a_t}(t+1) \mid \{r_{t'}\}_{\mathbb{1}\{a_{t'}=i\}}) = \frac{p(\{r_{t'}\}_{\mathbb{1}\{a_{t'}=i\}} \mid \theta) p(\theta)}{\int_{\theta'} p(\{r_{t'}\}_{\mathbb{1}\{a_{t'}=i\}} \mid \theta') p(\theta') d\theta'}$$

4.2 Analysis of Thompson sampling

Theorem 1 *Assume the rewards of arms are μ_i-Bernoulli. The regret under TS (Bernoulli bandits) is at most*

$$\overline{R}_T \leq \sum_{i:\Delta_i > 0} \frac{\mu_1 - \mu_i}{d_{KL}(\mu_1 \parallel \mu_i)} \log T + o(\log T)\,,$$

where the Kullback-Leibler divergence

$$d_{KL}(\mu_1 \parallel \mu_i) = \mu_1 \log(\frac{\mu_1}{\mu_i}) + (1 - \mu_1) \log(\frac{1 - \mu_1}{1 - \mu_i})\,.$$

As is similar to ETC and UCB, instance-independent regret bound of $O(\sqrt{mT \log T})$ can be obtained.

The proof of the regret bound can be obtained by either using probability or using techniques in information theory. We refer the proofs to the papers listed in suggested reading.

Due to time and space limits, we are unable to present the original, probabilistic proof in full. We will instead show a weaker bound of $O\left(\left(\sum_i \frac{\log T}{\Delta_i^2}\right)^2\right)$.

We now use N to denote the total number of arms.

Proof: For simplicity, we assume that arm 1 is the optimal arm. At every time step t, we divide the set of suboptimal arms into saturated and unsaturated arms. We say that an arm $i \neq 1$ is in the saturated set $C(t)$ at time t, if it has been played at least $L_i = \frac{24 \log T}{\Delta_i^2}$ times before time t. We bound the regret due to playing unsaturated and saturated suboptimal arms separately.

In the following derivation, by an interval of time we mean a set of contiguous time steps. Let random variable I_j denote the interval between (and excluding) the j^{th} and $(j+1)^{th}$ plays of the first arm. We say that event $M(t)$ holds at time t, if $\theta_1(t)$ exceeds $\mu_i + \frac{\Delta_i}{2}$ of all the saturated arms (for t such that $C(t)$ is empty, we define $M(t)$ to hold trivially), i.e.,

$$M(t) : \theta_1(t) > \max_{i \in C(t)} \mu_i + \frac{\Delta_i}{2} .$$

Let random variable γ_j denote the number of occurrences of event $M(t)$ in interval I_j :

$$\gamma_j = |\{t \in I_j : M(t) = 1\}| .$$

Events $M(t)$ divide I_j into sub-intervals in a natural way: For $\ell = 2$ to γ_j, let random variable $I_j(\ell)$ denote the sub-interval of I_j between the $(\ell-1)^{th}$ and ℓ^{th} occurrences of event $M(t)$ in I_j (excluding the time steps in which the event $M(t)$ occurs). We also define $I_j(1)$ and $I_j(\gamma_j + 1)$: If $\gamma_j > 0$ then $I_j(1)$ denotes the sub-interval in I_j before the first occurrence of event $M(t)$ in I_j; and $I_j(\gamma_j + 1)$ denotes the sub-interval in I_j after the last occurrence of event $M(t)$ in I_j. For $\gamma_j = 0$ we have $I_j(1) = I_j$. We define event $E(t)$ as

$$E(t) : \{\theta_i(t) \in [\mu_i - \Delta_i/2, \mu_i + \Delta_i/2], \forall i \in C(t)\} .$$

In words, $E(t)$ denotes the event that all saturated arms have $\theta_i(t)$ tightly concentrated around their means. Intuitively, from the definition of saturated arms, $E(t)$ should hold with high probability. We prove this in the lemma below.

Lemma 2 *For all t, $\mathbb{P}(E(t)) \geq 1 - \frac{4(N-1)}{T^2}$. For all t, j, and $s \leq j$, $\mathbb{P}(E(t) \mid s(j) = s) \geq 1 - \frac{4(N-1)}{T^2}$, where $s(j)$ denotes the number of successes in first j plays of arm 1.*

Proof: To prove the second statement of this lemma, we are required to lower bound the probability of $\mathbb{P}(E(t) \mid s(j) = s)$ all t, j, and $s \leq j$, by $1 - \frac{4(N-1)}{T^2}$. Recall that event $E(t)$ holds if the following is true:

$$\left\{\forall i \in C(t), \theta_i(t) \in \left[\mu_i - \frac{\Delta_i}{2}, \mu_i + \frac{\Delta_i}{2}\right]\right\} .$$

We define $E_i^+(t)$ as the event $\left\{\theta_i(t) \leq \mu_i + \frac{\Delta_i}{2} \text{ or } i \notin C(t)\right\}$, and $E_i^-(t)$ as the event $\{\theta_i(t) \geq \mu_i - \frac{\Delta_i}{2} \text{ or } i \notin C(t)\}$. Then, we can bound $\mathbb{P}(\overline{E(t)} \mid s(j))$ as

$$\mathbb{P}(\overline{E(t)} \mid s(j)) \leq \sum_{i=2}^{N} \mathbb{P}\left(\overline{E_i^+(t)} \mid s(j)\right) + \mathbb{P}\left(\overline{E_i^-(t)} \mid s(j)\right) .$$

Observe that

$$\mathbb{P}\left(\overline{E_i^+(t)} \mid s(j)\right) = \mathbb{P}\left(\theta_i(t) > \mu_i + \frac{\Delta_i}{2}, k_i(t) \geq L_i \mid s(j)\right) ,$$

where $k_i(t)$ is the number of plays of arm i until time $t-1$. We define $A_i(t)$ as the event

$$A_i(t) : \frac{S_i(t)}{k_i(t)} \leq \mu_i + \frac{\Delta}{4} ,$$

where $S_i(t), k_i(t)$ denote the number of successes and number of plays respectively of the i^{th} arm until time $t-1$. We will upper bound the probability of $\mathbb{P}\left(\overline{E_i^+(t)} \mid s(j)\right)$ for all $t, j, i \neq 1$, using,

$$\begin{aligned}\mathbb{P}\left(\overline{E_i^+(t)} \mid s(j)\right) &= \mathbb{P}\left(\theta_i(t) > \mu_i + \frac{\Delta_i}{2}, k_i(t) \geq L_i \mid s(j)\right) \\ &\leq \mathbb{P}\left(\overline{A_i(t)}, k_i(t) \geq L_i \mid s(j)\right) + \mathbb{P}\left(\theta_i(t) > \mu_i + \frac{\Delta_i}{2}, k_i(t) \geq L_i, A_i(t) \mid s(j)\right) .\end{aligned}$$

For every $i = 1, \ldots, N$ we define variables $\{Z_{i,m}\}$, and $\bar{Z}_{i,M}$. $Z_{i,m}$ denotes the output of the m-th play of the i-th arm. And,

$$\bar{Z}_{i,M} = \frac{1}{M} \sum_{m=1}^{M} Z_{i,m} .$$

Note that for all i, m, $Z_{i,m}$ is a Bernoulli variable with mean μ_i, and all $Z_{i,m}$, $i = 1, \ldots, N$, $m = 1, \ldots, T$ are independent of each other.

Now, instead of bounding the first term $\mathbb{P}\left(\overline{A_i(t)}, k_i(t) \geq L_i \mid s(j)\right)$, we prove a bound on $\mathbb{P}\left(\overline{A(t)}, k_2(t) \geq L \mid Z_{1,1}, \ldots, Z_{1,j}\right)$. Note that the latter bound is stronger, since $s(j)$ is

simply $\sum_{m=1}^{j} Z_{1,m}$. For all $t, i \neq 1$,

$$\begin{aligned}
\mathbb{P}\left(\overline{A_i(t)}, k_i(t) \geq L_i \mid Z_{1,1}, \ldots, Z_{1,j}\right) &= \sum_{\ell=L}^{T} \mathbb{P}\left(\bar{Z}_{i,k_i(t)} > \mu_i + \frac{\Delta_i}{4}, k_i(t) = \ell \mid Z_{1,1}, \ldots, Z_{1,j}\right) \\
&= \sum_{\ell=L}^{T} \mathbb{P}\left(\bar{Z}_{i,\ell} > \mu_i + \frac{\Delta_i}{4}, k_i(t) = \ell \mid Z_{1,1}, \ldots, Z_{1,j}\right) \\
&\leq \sum_{\ell=L}^{T} \mathbb{P}\left(\bar{Z}_{i,\ell} > \mu_i + \frac{\Delta_i}{4} \mid Z_{1,1}, \ldots, Z_{1,j}\right) \\
&= \sum_{\ell=L}^{T} \mathbb{P}\left(\bar{Z}_{i,\ell} > \mu_i + \frac{\Delta_i}{4}\right) \\
&\leq \sum_{\ell=L}^{T} e^{-2\ell\Delta_i^2/16} \\
&\leq \frac{1}{T^2}\,.
\end{aligned}$$

The equality in the third last line holds because for all i, i', m, m', $Z_{i,m}$ and $Z_{i',m'}$ are independent of each other, which means that $\bar{Z}_{i,\ell}$ is independent of $Z_{1,m}$ for all $m = 1, \ldots, j$. The inequality in the second last line is by applying the Chernoff bounds, since $\bar{Z}_{i,\ell}$ is simply the average of ℓ i.i.d. Bernoulli variables, each with mean μ_2. It will be useful to define $W(\ell, z)$ as a random variable distributed as Beta$(\ell z + 1, \ell - \ell z + 1)$. Note that if at time t, the number of plays of arm i is $k_i(t) = \ell$, then $\theta_i(t)$ is distributed as Beta $\big(\ell\bar{Z}_{i,\ell} + 1, \ell - \ell\bar{Z}_{i,\ell} + 1\big)$, i.e., same as $W\left(\ell, \bar{Z}_{i,\ell}\right)$. Then,

$$\begin{aligned}
&\mathbb{P}\left(\theta_i(t) > \mu_i + \frac{\Delta}{2}, A_i(t), k_i(t) \geq L_i \mid Z_{1,1}, \ldots, Z_{1,j}\right) \\
&= \sum_{\ell=L_i}^{T} \mathbb{P}\left(\theta_i(t) > \mu_i + \frac{\Delta_i}{2}, A_i(t), k_i(t) = \ell \mid Z_{1,1}, \ldots, Z_{1,j}\right) \\
&\leq \sum_{\ell=L_i}^{T} \mathbb{P}\left(\theta_i(t) > \frac{S_i(t)}{k_i(t)} - \frac{\Delta_i}{4} + \frac{\Delta_i}{2}, k_i(t) = \ell \mid Z_{1,1}, \ldots, Z_{1,j}\right) \\
&= \sum_{\ell=L_i}^{T} \mathbb{P}\left(W\left(\ell, \bar{Z}_{i,\ell}\right) > \bar{Z}_{i,\ell} + \frac{\Delta_i}{4}, k_i(t) = \ell \mid Z_{1,1}, \ldots, Z_{1,j}\right) \\
&\leq \sum_{\ell=L_i}^{T} \mathbb{P}\left(W\left(\ell, \bar{Z}_{i,\ell}\right) > \bar{Z}_{i,\ell} + \frac{\Delta_i}{4} \mid Z_{1,1}, \ldots, Z_{1,j}\right) \\
&= \sum_{\ell=L_i}^{T} \mathbb{P}\left(W\left(\ell, \bar{Z}_{i,\ell}\right) > \bar{Z}_{i,\ell} + \frac{\Delta_i}{4}\right) \\
&\leq \sum_{\ell=L_i}^{T} \mathbb{E}\left[F^B_{\ell, \bar{Z}_{i,\ell} + \frac{\Delta_i}{4}}\left(\ell\bar{Z}_{i,\ell}\right)\right]
\end{aligned}$$

$$\leq \sum_{\ell=L_i}^{T} \exp\left\{-\frac{2\Delta_i^2\ell^2/16}{\ell}\right\}$$
$$\leq \frac{1}{T^2}\,.$$

Here, we used the observation that for all $i, i', m, m', Z_{i,m}$ and $Z_{i',m'}$ are independent of each other, which means $\bar{Z}_{i,\ell}$ and $W\left(\ell, \bar{Z}_{i,\ell}\right)$ are independent of $Z_{1,m}$ for all $m = 1, \ldots, j$. The third last inequality follows from the observation that

$$F^B_{n+1,p}(r) = (1-p)F^B_{n,p}(r) + pF^B_{n,p}(r-1) \leq (1-p)F^B_{n,p}(r) + pF^B_{n,p}(r) = F^B_{n,p}(r)\,.$$

The second last inequality follows from the Chernoff-Hoeffding bound. Substituting the above inequality, we obtain.

$$\mathbb{P}\left(\overline{E_i^+(t)} \mid s(j)\right) \leq \frac{2}{T^2}$$

Similarly, we could obtain

$$\mathbb{P}\left(\overline{E_i^-(t)} \mid s(j)\right) \leq \frac{2}{T^2}$$

Summing over $i = 2, \ldots, N$, we get

$$\mathbb{P}(\overline{E(t)} \mid s(j)) \leq \frac{4(N-1)}{T^2}\,,$$

which implies the second statement of the lemma. The first statement follows immediately. □

Observe that since arm i, if saturated, can be played at time t only if $\theta_i(t)$ is greater than $\theta_1(t)$, arm i, if saturated, can be played at time t where $M(t)$ holds only if $\theta_i(t) > \mu_i + \Delta_i/2$. Thus, unless the high probability event $E(t)$ is violated, $M(t)$ denotes a play of an unsaturated arm at time t, and γ_j essentially denotes the number of plays of unsaturated arms in interval I_j. And, the number of plays of saturated arms in interval I_j is at most

$$\sum_{\ell=1}^{\gamma_j+1} |I_j(\ell)| + \sum_{t\in I_j} I(\overline{E(t)})\,.$$

We are interested in bounding the regret due to playing saturated arms, which depends not only on the number of plays, but also on which saturated arm is played at each time step. Let $V_j^{\ell,a}$ denote the number of steps in $I_j(\ell)$, for which a is the best saturated arm, i.e.

$$V_j^{\ell,a} = \left|\left\{t \in I_j(\ell) : \mu_a = \max_{i\in C(t)} \mu_i\right\}\right|\,.$$

Note that we resolve the ties for best saturated arm using an arbitrary, but fixed, ordering on arms.

Recall that $M(t)$ holds trivially for all t such that $C(t)$ is empty. Therefore, there is at least one saturated arm at all $t \in I_j(\ell)$, and hence $V_j^{\ell,a}, a = 2, \ldots, N$ are well defined and cover the interval $I_j(\ell)$

$$|I_j(\ell)| = \sum_{a=2}^{N} V_j^{\ell,a}$$

Next, we will show that the regret due to playing any saturated arm at a time step t in one of the $V_j^{\ell,a}$ steps is at most $3\Delta_a + I(\overline{E(t)})$. The idea is that if all saturated arms have their $\theta_i(t)$ tightly concentrated around their means μ_i, then either the arm with the highest mean (i.e., the best saturated arm a) or an arm with mean very close to μ_a will be chosen to be played during these $V_j^{\ell,a}$ steps. That is, if a saturated arm i is played at a time t among one of the $V_j^{\ell,a}$ steps, then, either $E(t)$ is violated, i.e., $\theta_{i'}(t)$ for some saturated arm i' is not close to its mean, or

$$\mu_i + \Delta_i/2 \geq \theta_i(t) \geq \theta_a(t) \geq \mu_a - \Delta_a/2\,,$$

which implies that

$$\Delta_i = \mu_1 - \mu_i \leq \mu_1 - \mu_a + \frac{\Delta_a}{2} + \frac{\Delta_i}{2} \Rightarrow \Delta_i \leq 3\Delta_a\,.$$

Therefore, regret due to play of a saturated arm at a time t in one of the $V_j^{\ell,a}$ steps is at most $3\Delta_a + I(\overline{E(t)})$. With slight abuse of notation let us use $t \in V_j^{\ell,a}$ to indicate that t is one of the $V_j^{\ell,a}$ steps in $I_j(\ell)$. Then, the expected regret due to playing saturated arms in interval I_j is bounded as

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}^s\left(I_j\right)\right] &\leq \mathbb{E}\left[\sum_{\ell=1}^{\gamma_j+1}\sum_{a=2}^{N}\sum_{t\in V_j^{\ell,a}}\left(3\Delta_a + I(\overline{E(t)})\right)\right] + \sum_{t\in I_j} I(\overline{E(t)}) \\ &= \mathbb{E}\left[\sum_{\ell=1}^{\gamma_j+1}\sum_{a=2}^{N} 3\Delta_a V_j^{\ell,a}\right] + 2\mathbb{E}\left[\sum_{t\in I_j} I(\overline{E(t)})\right]\,.\end{aligned}$$

Notice that the second term can be bounded by Lemma 2. For the first term, we show the following lemma.

Lemma 3 *For all j,*

$$\begin{aligned}&\mathbb{E}\left[\sum_{\ell=1}^{\gamma_j+1}\sum_{a} V_j^{\ell,a}\Delta_a\right] \\ \leq&\mathbb{E}\left[\mathbb{E}\left[(\gamma_j+1)\mid s(j)\right]\sum_{a=2}^{N}\Delta_a\mathbb{E}\left[\min\left\{X\left(j,s(j),\mu_a+\frac{\Delta_a}{2}\right),T\right\}\mid s(j)\right]\right]\,,\end{aligned}$$

where $X(j,s(j),y)$ is defined as the number of trials until an independent sample from $\mathrm{Beta}(s+1,j-s+1)$ *distribution exceeds y.*

Proof: Observe that

$$\mathbb{E}\left[\sum_{\ell=1}^{\gamma_j+1} V_j^{\ell,a}\mid s(j)\right] = \mathbb{E}\left[\sum_{\ell=1}^{T} V_j^{\ell,a}\cdot \mathrm{I}\left(\gamma_j \geq \ell-1\right)\mid s(j)\right]\,.$$

Let $\mathcal{F}_{\ell-1}$ denote the history until before the beginning of interval $I_j(\ell)$ (i.e., the values of $\theta_i(t)$ and the outcomes of playing the arms until the time step before the first time step of $I_j(\ell)$). Note that the value of random variable $\mathrm{I}(\gamma_j \geq \ell - 1)$ is fully determined by $\mathcal{F}_{\ell-1}$. Therefore,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\ell=1}^{\gamma_j+1} V_j^{\ell,a} \mid s(j)\right] \\
=&\mathbb{E}\left[\sum_{\ell=1}^{T} \mathbb{E}\left[V_j^{\ell,a} \cdot \mathrm{I}(\gamma_j \geq \ell - 1) \mid s(j), \mathcal{F}_{\ell-1}\right] \mid s(j)\right] \\
=&\mathbb{E}\left[\sum_{\ell=1}^{T} \mathbb{E}\left[V_j^{\ell,a} \mid s(j), \mathcal{F}_{\ell-1}\right] \cdot \mathrm{I}(\gamma_j \geq \ell - 1) \mid s(j)\right].
\end{aligned}$$

Recall that $V_j^{\ell,a}$ is the number of contiguous steps t for which a is the best arm in saturated set $C(t)$ and i.i.d variables $\theta_1(t)$ have value smaller than $\mu_a + \frac{\Delta_a}{2}$. Observe that given $s(j) = s$ and $\mathcal{F}_{\ell-1}$, $V_j^{\ell,a}$ is the length of an interval which ends when the value of an i.i.d. Beta $(s+1, j-s+1)$ distributed variable exceeds $\mu_a + \frac{\Delta_a}{2}$ (i.e., $M(t)$ happens), or if an arm other than a becomes the best saturated arm, or if we reach time T. Therefore, given $s(j), \mathcal{F}_{\ell-1}, V_j^{\ell,a}$ is stochastically dominated by $\min\left\{X\left(j, s(j), \mu_a + \frac{\Delta_a}{2}\right), T\right\}$. That is, for all a,

$$\begin{aligned}
\mathbb{E}\left[V_j^{\ell,a} \mid s(j), \mathcal{F}_{\ell-1}\right] &\leq \mathbb{E}\left[\min\left\{X\left(j, s(j), \mu_a + \frac{\Delta_a}{2}\right), T\right\} \mid s(j), \mathcal{F}_{\ell-1}\right] \\
&= \mathbb{E}\left[\min\left\{X\left(j, s(j), \mu_a + \frac{\Delta_a}{2}\right), T\right\} \mid s(j)\right].
\end{aligned}$$

Substituting the term, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\ell=1}^{\gamma_j+1} V_j^{\ell,a} \mid s(j)\right] \\
&\leq \mathbb{E}\left[\sum_{\ell=1}^{T} \mathbb{E}\left[\min\left\{X\left(j, s(j), \mu_a + \frac{\Delta_a}{2}\right), T\right\} \mid s(j)\right] \cdot \mathrm{I}(\gamma_j \geq \ell - 1) \mid s(j)\right] \\
&= \mathbb{E}\left[\min\left\{X\left(j, s(j), \mu_a + \frac{\Delta_a}{2}\right), T\right\} \mid s(j)\right] \cdot \mathbb{E}\left[\sum_{\ell=1}^{T} \mathrm{I}(\gamma_j \geq \ell - 1) \mid s(j)\right] \\
&= \mathbb{E}\left[\min\left\{X\left(j, s(j), \mu_a + \frac{\Delta_a}{2}\right), T\right\} \mid s(j)\right] \cdot \mathbb{E}\left[\gamma_j + 1 \mid s(j)\right].
\end{aligned}$$

This immediately implies,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{a=2}^{N} \Delta_a \mathbb{E}\left[\sum_{\ell=1}^{\gamma_j+1} V_j^{\ell,a} \mid s(j)\right]\right] \\
\leq&\mathbb{E}\left[\sum_{a=2}^{N} \Delta_a \mathbb{E}\left[\min\left\{X\left(j, s(j), \mu_a + \frac{\Delta_a}{2}\right), T\right\} \mid s(j)\right] \cdot \mathbb{E}\left[\gamma_j + 1 \mid s(j)\right]\right],
\end{aligned}$$

as the lemma desires. □

Then, we can decompose the first term as

$$3\mathbb{E}\left[\sum_{j=0}^{\sum_i L_i} \mathbb{E}\left[(\gamma_j+1) \mid s(j)\right] \sum_a \Delta_a \mathbb{E}\left[\min\left\{X\left(j, s(j), y_a\right), T\right\} \mid s(j)\right]\right]$$
$$\leq 3\mathbb{E}\left[\left(\sum_{j=0}^{\sum_i L_i} \mathbb{E}\left[(\gamma_j+1) \mid s(j)\right]\right)\left(\sum_{j=0}^{\sum_i L_i} \sum_a \Delta_a \mathbb{E}\left[\min\left\{X\left(j, s(j), y_a\right), T\right\} \mid s(j)\right]\right)\right].$$

Recall that γ_j is (approximately) the total number of plays of unsaturated arms in interval I_j. Therefore, the first term in the product above is bounded by the total number of plays of unsaturated arms, i.e. $O\left(\sum_{i=2}^N L_i\right)$. For the second term, we use the following lemma.

Lemma 4 *Consider any positive $y < \mu_1$, and let $\Delta' = \mu_1 - y$. Also, let $R = \frac{\mu_1(1-y)}{y(1-\mu_1)} > 1$, and let D denote the KL divergence between μ_1 and y, i.e., $D = y \ln \frac{y}{\mu_1} + (1-y) \ln \frac{1-y}{1-\mu_1}$.*

$$\mathbb{E}[\mathbb{E}[\min\{X(j, s(j), y), T\} \mid s(j)]] \leq \begin{cases} 1 + \frac{2}{1-y} + \frac{\mu_1}{\Delta'} e^{-Dj} & j < \frac{y}{D} \ln R \\ 1 + \frac{R^y}{1-y} e^{-Dj} + \frac{\mu_1}{\Delta'} e^{-Dj} & \frac{y}{D} \ln R \leq j < \frac{4 \ln T}{\Delta'^2} \\ \frac{16}{T} & j \geq \frac{4 \ln T}{\Delta'^2}, \end{cases}$$

where the outer expectation is taken over $s(j)$ distributed as Binomial(j, μ_1).

The proof of the lemma is omitted. We observe that $\mathbb{E}\left[\mathbb{E}\left[\min\left\{X\left(j, s(j), y_a\right), T\right\} \mid s(j)\right]\right]$ is bounded by $O\left(\frac{1}{\Delta_a}\right)$. Therefore, the second term is bounded by $O\left(\sum_{i=2}^N L_i\right)$ as well. This gives a bound of $O\left(\left(\sum_i L_i\right)^2\right) = O\left(\left(\sum_i \frac{\log T}{\Delta_i^2}\right)^2\right)$ on the above, and thus on the contribution of the first term towards the regret. The total contribution of the second term can be bounded by a constant using Lemma 2.

Since an unsaturated arm u becomes saturated after L_u plays, regret due to unsaturated arms is at most $\sum_{u=2}^N L_u \Delta_u = 24(\ln T)\left(\sum_{u=2}^N \frac{1}{\Delta_u}\right)$. Summing the regret due to saturated and unsaturated arms, we obtain the weaker bound of $O\left(\left(\sum_i \frac{\log T}{\Delta_i^2}\right)^2\right)$ on regret. □

4.3 Applications

Students should implement ε-greedy, ETC, UCB, and TS and try them on synthetic and real datasets to gain some intuition about their behavior.

Acknowledgement

This lecture notes partially use material from *Bandit algorithms* and the papers listed in suggested reading.