

Lecture 3: January 15, 2024

Lecturer: Yishay Mansour

Scribe: Rotem Nizhar, Tamar Garbuz, Maor Lavi¹

1 Bayesian bandits and Thompson samplings

lecture topics:

- Bayesian model for MAB
- Thompson sampling algorithm
- Analysis in Bayesian setting
- Analysis for stochastic MAB

2 Bayesian model for MAB

We start with the basic model of MAB:

We have T round, K actions, and for each action a there is a distribution D_a and mean $\mu(a)$.

In the Bayesian MAB we add an assumption that the parameters $\vec{\mu} = (\mu(a_1) \dots \mu(a_k))$ and D_a are drawn from the initial prior distribution \mathbb{P} .

Bayesian Regret: we will mean the expected regret under our prior, and not the worst $\vec{\mu}$, namely,

$$BR(T) := E_{\mathbb{P}}[E[R(T)]|\vec{\mu}] = E_{\mathbb{P}}[\mu^* \cdot T - \sum_{t \in [T]} \mu(a_t)]$$

Now we will add some simplifying assumptions:

- One parameter:
There is one-to-one mapping between $\mu(a)$ to D_a ($\mu(a) \leftrightarrow D_a$), meaning it is enough to have distribution over $\vec{\mu}$.
- Profit group denoted by F is finite, meaning that the support of the prior distribution is finite (in that way we avoid integrals).
- A unique optimal action a^* for all $\vec{\mu}$ under the prior.

2.1 Bayesian update in Bayesian bandit

2.1.1 Notation:

History: t -history is denoted by $H_t = ((a_1, r_1), \dots, (a_t, r_t))$, the history is the actions and rewards until time t , this is a random variable that depends on the algorithm, $\vec{\mu}$ and \vec{D} .

feasible t -history: $H_t = ((a'_1, r'_1), \dots, (a'_t, r'_t))$ is called feasible if there is an algorithm such that $P(H = H_t) > 0$ (that is, there is an algorithm that can generate H_t), such algorithm is called **H-consistent**.

¹Based on scribe notes of Omer Amichay and Shahar Lewkowicz from 2021/22

H-induced: An algorithm is called H-induced if for feasible $H_t = ((a'_1, r'_1), \dots, (a'_t, r'_t))$ the algorithm is deterministically choosing a'_i in round $i \forall i \in [t]$.

We are interested for a subset $M \subseteq [0, 1]^k$ in the probability,

$$\mathbb{P}_H(M) \triangleq P_R[\vec{\mu} \in M | H_t = H]$$

This is defined for each H -consistent algorithm.

2.1.2 The posterior does not depend on the algorithm

Lemma 1 *The distribution \mathbb{P}_H is identical for every algorithm that is H-consistent, that is, the distribution $\mathbb{P}_H(M)$ doesn't depend on the algorithm.*

Proof: We will prove for $M = \{\tilde{\mu}\}$. Since any M is a disjoint union of $\tilde{\mu}$, it suffices to prove this when M contains a single profile since if $\mu \neq \mu'$ the events $\vec{\mu} = \mu'$ and $\vec{\mu} = \mu$ are disjoint.

We will prove this by induction on the length of the history: t .

Base case: $t = 0$, in this case $H_0 = \emptyset$ therefore the probability $P[\vec{\mu} \in M | H_0 = H]$ is the prior probability.

Induction hypothesis: $\forall i < t$ the lemma hold for H_i .

Induction step: for $t \geq 1$, we will consider H' as the prefix of H of length $t-1$, means $H = \{H', (a, r)\}$ and define:

$$\pi(a) \triangleq P_R[a_t = a | H_{t-1} = H']$$

This is the probability that the algorithm will choose action a given history H' . Note that there is no dependency on the mean reward vector, only the history.

Notice that:

(2.1.2.1)

$$\begin{aligned} \frac{P_r[\mu = \tilde{\mu}, H_t = H]}{P_R[H_{t-1} = H']} &= \frac{P_r[\mu = \tilde{\mu}, H_{t-1} = H', (a_t, r_t) = (a, r)]}{P_R[H_{t-1} = H']} \\ &= P_R[\mu = \tilde{\mu}, (a_t, r_t) = (a, r) | H_{t-1} = H'] \\ &= P_R[\mu = \tilde{\mu} | H_{t-1} = H'] \cdot P[(a_t, r_t) = (a, r) | \mu = \tilde{\mu}, H_{t-1} = H'] \\ &= \mathbb{P}_{H'}(\tilde{\mu}) \cdot P[a_t = a | \mu = \tilde{\mu}, H_{t-1} = H'] \cdot P[r_t = r | \mu = \tilde{\mu}, H_{t-1} = H', a_t = a] \\ &= \mathbb{P}_{H'}(\tilde{\mu}) D_{\tilde{\mu}(a)}(r) \pi(a) \end{aligned} \tag{1}$$

From the total probability theorem and (2.1.2.1) we get:

(2.1.2.2)

$$\begin{aligned} P[H_t = H] &= \sum_{\tilde{\mu} \in \mathcal{F}} P[H_t = H, \mu = \tilde{\mu}] \\ &= \sum_{\tilde{\mu} \in \mathcal{F}} P[H_{t-1} = H'] \cdot \mathbb{P}_{H'}(\tilde{\mu}) \cdot D_{\tilde{\mu}(a)}(r) \cdot \pi(a) \\ &= \pi(a) P[H_{t-1} = H'] \sum_{\tilde{\mu} \in \mathcal{F}} \mathbb{P}_{H'}(\tilde{\mu}) D_{\tilde{\mu}(a)}(r) \end{aligned} \tag{2}$$

Therefore:

(2.1.2.3)

$$\begin{aligned} \mathbb{P}_H(\tilde{\mu}) &= \frac{P_R[\mu = \tilde{\mu}, H_t = H]}{P_R[H_t = H]} \\ &= \frac{P_R[H_{t-1} = H'] \cdot \mathbb{P}_{H'}[\tilde{\mu}] \cdot D_{\tilde{\mu}(a)}(r) \cdot \pi(a)}{\pi(a) \cdot P_R[H_{t-1} = H'] \cdot \sum_{\tilde{\mu} \in \mathcal{F}} \mathbb{P}_{H'}(\tilde{\mu}) \cdot D_{\tilde{\mu}(a)}(r)} = \frac{\overbrace{\mathbb{P}_{H'}(\tilde{\mu}) D_{\tilde{\mu}(a)}(r)}^{(2.1.2.1)}}{\underbrace{\sum_{\tilde{\mu} \in \mathcal{F}} \mathbb{P}_{H'}(\tilde{\mu}) D_{\tilde{\mu}(a)}(r)}_{(2.1.2.2)}} \end{aligned} \quad (3)$$

From the induction hypothesis, $\mathbb{P}_{H'}(\tilde{\mu})$ does not depend on the algorithm, therefore $\mathbb{P}_H(\tilde{\mu})$ is independent of the algorithm as well. \square

Notes:

Claim: we can derive from the proof of the lemma that:

$$\mathbb{P}_H(\tilde{\mu}) \propto P_R[\tilde{\mu}] \prod_t D_{\tilde{\mu}(a_t)}(r_t)$$

Corollary: $P_{H'} = P_H$ if H' is a permutation of H because their product will not be affected by the order.

1) We will not always get independence of the algorithm.

Example:

First, we will assume that we look over a set of feasible t -histories \mathcal{H} and the posterior: $P_r[\mu \in M | H_t \in \mathcal{H}]$. Consider our actions set to be $A = \{a_1, a_2, a_3\}$ and $T = 1$, we define ALG_1 that chooses action a_1 and creates the history $H = (a_1, 1)$, and we define ALG_2 that chooses action a_2 and creates the history $H' = (a_2, 1)$.

We can see that under ALG_1 :

$$P_r[\tilde{\mu} \in M | H_t \in \{H, H'\}] = P_r[\tilde{\mu} \in M | H_t = H] = P_H(M)$$

Same for ALG_2 :

$$P_r[\tilde{\mu} \in M | H_t \in \{H, H'\}] = P_r[\tilde{\mu} \in M | H_t = H'] = P_{H'}(M)$$

Therefore the probability $P_R[\tilde{\mu} | H_t \in \{H, H'\}]$ depends on the algorithm.

2) Suppose that we are given a subset of rounds S , we define S – history as:

$$H_S = \{(a_t, r_t) : t \in S\}$$

and the posterior:

$$\mathbb{P}_{H,S}[M] \triangleq P_R[\tilde{\mu} \in M | H_S]$$

For example: consider $S = \{2\}$ at $t = 1$. ALG_1 picks a_1 and ALG_2 picks a_2 , at $t = 2$ they both pick a_1 if they received reward 1 in round 1, else they pick a_2 .

For ALG_1 , $\mathbb{P}_{H,S}[M]$ have a condition on event that we picked a_1 at round 1, same for ALG_2 and a_2 , therefore the distribution depends on the algorithm.

2.1.3 Posterior as a new prior

Definition we will define a concatenation of feasible histories t – history and $(t' - t)$ – history as $H \oplus H'$ which is t' – history.

We would like to show that a concatenation of two histories is equivalent to taking the history of the first t steps and calculating over it the posterior of the next t' - t steps, that is: $P_{H \oplus H'} = (P_H)_{H'}$.

Lemma 2 *Let H be a feasible t -history and H' a feasible $(t'-t)$ -history then*

$$P_{H \oplus H'}(M) = P_{\mu \sim \mathbb{P}_H} [\mu \in M | H_{t'} = H'] \Leftrightarrow P_{H \oplus H'} = (\mathbb{P}_H)_{H'}$$

Proof: We will prove for $M = \{\tilde{\mu}\}$.

We denote ALG as an $H \oplus H'$ - induced algorithm (the choosing of the actions is deterministic).

H_t^{ALG} is the first t steps and H_s^{ALG} is the $s = (t, t')$ steps.

We define two events:

$$\begin{aligned} \mathcal{E}_t &= \{H_t^{ALG} = H\} \\ \mathcal{E}_s &= \{H_s^{ALG} = H'\} \end{aligned}$$

Now we can see that:

$$\begin{aligned} P_{\mu \sim \mathbb{P}_H} [\mu \in M | \mathcal{E}_s] &= \frac{P_{\mu \sim \mathbb{P}_H} [\mu \in M, \mathcal{E}_s]}{P_{\mu \sim \mathbb{P}_H} [\mathcal{E}_s]} \\ &= \frac{P_R[\mu \in M, \mathcal{E}_s | \mathcal{E}_t]}{P_R[\mathcal{E}_s | \mathcal{E}_t]} \\ &= \frac{P_R[\mu \in M, \mathcal{E}_s, \mathcal{E}_t]}{P_R[\mathcal{E}_s, \mathcal{E}_t]} \\ &= P_R[\mu \in M | \mathcal{E}_s, \mathcal{E}_t] \\ &= P_{H \oplus H'}(M) \end{aligned} \tag{4}$$

For a general algorithm we can use the fact that the posterior is independent of the algorithm. \square

2.1.4 Independent Prior

Definition: The prior distribution is independent if the random variables $\mu(a)$ are mutually independent random variables.

Advantage: We can update each action separately, because an action that occurs in time t will not be affected by an action that occurs in time t' .

Given an history H , we denote the set of rounds in which action a was chosen in H by $S_a^H = \{\tau : a_\tau = a\}$. The projection of H onto action a is defined as $proj(H; a) = ((a'_\tau, r'_\tau) : \tau \in S_a^H)$. Note that it is a feasible history.

We define the posterior distribution for action a :

$$\forall M_a \subseteq [0, 1] \quad P_H^a(M_a) \triangleq P_{proj(H,a)}(\mu(a) \in M_a)$$

This means that if we look at an event that depends only on action a , it will depend only on the projection of a .

Lemma 3 *Given an independent prior \mathbb{P} and an event $M_a \subseteq [0, 1]$ for every $a \in A$, then:*

$$\mathbb{P}_H\left(\bigcap_{a \in A} M_a\right) = \prod_{a \in A} P_H^a(M_a)$$

Proof: We will define:

$$\begin{aligned} \mathcal{E}_a &= \{\mu(a) \in M_a\} \\ \mathcal{E}_a^H &= \{proj(H_t; a) = proj(H; a)\} \end{aligned}$$

Denote: $M = \bigcap_{a \in A} M_a$

now:

$$P_R[H_t = H, \mu \in M] = P_R\left[\bigcap_{a \in A} (\mathcal{E}_a \cap \mathcal{E}_a^H)\right] = \prod_{a \in A} P_R[\mathcal{E}_a \cap \mathcal{E}_a^H]$$

We can observe that because the prior is independent, the events $\{\mathcal{E}_a\}_{a \in A}$ are independent too. In addition $\{\mathcal{E}_a^H\}_{a \in A}$ are independent as well, because the algorithm is H -induced, meaning it is deterministic and only the rewards of the actions are random, which depend only on $\mu(a)$, hence the last equality. Additionally:

$$P_R[H_t = H] = P_R\left[\prod_{a \in A} \mathcal{E}_a^H\right] = \prod_{a \in A} P_R[\mathcal{E}_a^H]$$

And now we can see that:

$$\mathbb{P}_H(M) = \frac{P_R[H_t = H, \mu \in M]}{P_R[H_t = H]} = \prod_{a \in A} \frac{P_R[\mathcal{E}_a \cap \mathcal{E}_a^H]}{P_R[\mathcal{E}_a^H]} = \prod_{a \in A} P_R[\mu \in M | \mathcal{E}_a^H] = \prod_{a \in A} P_H^a(M_a)$$

Note: because the algorithm is H-induced we get $P_R[\mu \in M | \mathcal{E}_a^H] = P_H^a(M_a)$ □

3 Thompson Sampling Algorithm

3.1 We will see two versions of the Thompson Sampling Algorithm:

Algorithm 1 Thompson Sampling Algorithm 1

For each $t < T$:

Observe $H_{t-1} = H$ history

Draw action a_t from $P_t(\cdot|H)$ where: $\forall a P_t(a|H) = P_R[a^* = a|H_{t-1} = H]$

In this version, in each round we look at the history until now and calculate the probability for each action to be the best action, based on the history H (we assume that there is always a best action). For each action a , the probability for it to be the one we chose is the probability it is the best action. we can think about it as:

$$P_t(a|H) = P_R[a^* = a|H_{t-1} = H] = \sum_{\substack{\mu: a^* \\ a^* \text{ best in } \mu}} \mathbb{P}_H(\mu)$$

Algorithm 2 Thompson Sampling Algorithm 2

For each $t < T$:

Observe $H_{t-1} = H$ history

Sample $\vec{\mu}_t$ from the posterior distribution \mathbb{P}_H

Choose $\tilde{a}_t = \underset{a}{\operatorname{argmax}} \mu_t(a)$

Here for each step we sample a profile from the posterior distribution, then we choose the best action from the profile we got.

Claim: The algorithms are equivalent

Intuition: In the second algorithm we sample from P_H and choose the best action, the action probability will be the sum of all the weights of all the μ that the action was the best for them, therefore it is equal to:

$$\sum_{\substack{\mu: a^* \\ a^* \text{ best in } \mu}} \mathbb{P}_H(\mu)$$

3.2 Computational aspects

Most of the computational time in the algorithm is spent on calculating the posterior \mathbb{P}_H .

Even though the number of profiles is finite, it will still be hard and the computational time will depend on the amount of profiles, so in general we get $O(|F|)$.

3.2.1 Distributions

Beta-Bernoulli:

We will use for a prior distribution $Beta(1, 1)$ (equals to Uniform distribution over $[0, 1]$).

For $Beta(\alpha, \beta)$ and (s, f) (s -number of successes, f -number of failures) the posterior is

$$Beta(\alpha + s, \beta + f)$$

Meaning that every time we see a success we add 1 to the first parameter and every time we see a failure we add 1 to the second parameter.

Gaussian:

If we use for the prior distribution $\mu \in \mathcal{N}(0, 1)$, then the posterior is:

$$\mathcal{N}\left(\frac{\sum_{i=1}^n x_i}{n+1}, \frac{1}{n+1}\right)$$

Note: the posterior mean is the average of the n samples + time zero for the prior, therefore we add 1, the variance gets smaller with each sample.

We can also notice that if the prior is a Gaussian distribution then the posterior is also a Gaussian distribution. When the prior and posterior belong to the some family of distributions, it is called a "Conjugate prior".

3.2.2 How are samples generated

For each version of the algorithm, we will show how the samples are generated for both the Bernoulli and the Gaussian distributions:

Version 1:

1. Bernoulli:
For each action, we initially sample $\mu \sim \text{Beta}(1, 1)$
And at each time step t , we use μ to sample $r_t \sim \text{Ber}(\mu)$
2. Gaussian:
For each action, we initially sample $\mu \sim \mathcal{N}(0, 1)$
And at each time step t , we use μ to sample $r_t \sim \mathcal{N}(\mu, 1)$

Intuition: We sample μ at the initial step from the ground truth, then for each time step we take our sample from the distribution (Bernoulli/Gaussian) with mean μ (note that the parameter μ is sampled only once).

Version 2:

1. Bernoulli:
At each time step t , we sample $\mu_t \sim \text{Beta}(s_t + 1, f_t + 1)$, then we sample $r_t \sim \text{Ber}(\mu_t)$
2. Gaussian:
At each time step t we sample $\mu_t \sim \mathcal{N}\left(\frac{\sum_{\tau=1}^{t-1} r_\tau}{t}, \frac{1}{t}\right)$, then we sample $r_t \sim \mathcal{N}(\mu_t, 1)$

Intuition: In this case we don't know the ground truth so we cant sample from it, instead at each step t we will sample μ_t from the proper posterior distribution (Bernoulli/Gaussian) distribution for that step, now we will sample the reward from the distribution (Bernoulli/Gaussian) with mean μ_t (Every step is independent).

Note: both versions will create the same distribution.

3.3 Beta distribution

$$\text{Beta}(\alpha, \beta) \quad \alpha, \beta > 0$$

$$\text{Density function: } \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\binom{\alpha+\beta}{\alpha}}$$

$$\text{Mean: } E_{x \sim \text{Beta}(\alpha, \beta)}[X] = \frac{\alpha}{\alpha+\beta}$$

There is similarity between Beta distribution and Binomial distribution:

Binomial:

p - the success probability, is given.

s - the number of successes, is calculated.

Beta:

p - the success probability, is calculated.

s - the number of successes, is given.

Formal relation:

$$P_{x \sim \text{Beta}(\alpha, \beta)} [x \leq p] = P_{z \sim \text{Bin}(\alpha + \beta - 1, p)} [z > \alpha]$$

Conjugate prior: In Bayesian probability theory, if the posterior distribution is in the same probability distribution family as the prior probability distribution, the prior and posterior are then called conjugate. (for example: normal distribution)

4 Bayesian Regret Analysis of the Thompson Sampling Algorithm

We will now analyze the regret of the Thompson Sampling Algorithm, and show $BR = O(\sqrt{KT \log(T)})$.

As we have done before, we will define the radius $r_t(a) = \sqrt{\frac{4 \log(T)}{n_t(a)}}$ where $n_t(a)$ is the random value who's value is the amount of times the algorithm sampled the action a until time t , i.e. $n_t(a) = |\{a_\tau = a : \tau \leq t\}|$. Let $\bar{\mu}_t(a)$ be the average of rewards of action a until time t . We will define the upper and lower confidence bounds: $UCB_t(a) = \bar{\mu}_t(a) + r_t(a)$ and $LCB_t(a) = \bar{\mu}_t(a) - r_t(a)$. As we have seen before, with high probability we have $\mu_t(a) \in [LCB_t(a), UCB_t(a)]$.

We will assume that there exists a constant $\gamma > 0$ such that

$$E[UCB_t(a) - \mu_t(a)]^- \leq \frac{\gamma}{TK}$$

and

$$E[\mu_t(a) - LCB_t(a)]^- \leq \frac{\gamma}{TK}$$

where for each random variable Z we define

$$[Z]^- = \begin{cases} 0 & , Z > 0 \\ |Z| & , Z \leq 0 \end{cases}$$

$$[Z]^+ = \begin{cases} z & , Z > 0 \\ 0 & , Z \leq 0 \end{cases}$$

Let's explain this assumption and why it is the case in many scenarios. Generally, we expect that $UCB_t(a) > \mu_t(a) > LCB_t(a)$, which means that the random variable $UCB_t(a) - \mu_t(a)$ will be positive. In the term $E[UCB_t(a) - \mu_t(a)]^-$ we calculate the expectation of this random variable, but in this calculation (i.e., sum or integral) we omit all the positive addends and keep only the negative addends (We multiply by -1 so the result will be positive). We assume that this sum (or integral) is small. There may be cases where $\mu_t(a)$ will be much bigger than $UCB_t(a)$, but the probability for that will be too small, so the expectation $E[UCB_t(a) - \mu_t(a)]^-$ will be small too. The same argument is also valid for the term $E[\mu_t(a) - LCB_t(a)]^-$.

This assumption is correct in cases of Bernoulli or Gaussian prior distributions, with $\gamma = 2$.

$$We will define $W(a, H_t) = \frac{UCB(a, H_t) - LCB(a, H_t)}{2}$$$

We will define BR to be the Bayesian regret of the Thompson Sampling Algorithm. We will also define BR_t to be the Bayesian regret of the algorithm at time t . Clearly,

$$BR = \sum_{t=1}^T BR_t$$

Lemma 4

$$BR \leq 2\gamma + 2 \sum_{t=1}^T E[r(a, H_t)]$$

Proof: A result from the definition of the algorithm is that for each history H we have

$$Pr[a_t = a | H_t = H] = Pr[a^* = a | H_t = H]$$

and therefore we also have

$$E[UCB(a^*, H) | H_t = H] = E[UCB(a, H) | H_t = H]$$

and therefore at time t we have

$$\begin{aligned} BR_t &= E[\mu(a^*) - \mu(a_t)] \\ &= \underset{H \sim H_t}{E} [E[\mu(a^*) - \mu(a_t) | H_t = H]] \\ &= \underbrace{E[UCB(a_t, H_t) - \mu(a_t)]}_{\text{term1}} + \underbrace{E[\mu(a^*) - UCB(a^*, H_t)]}_{\text{term2}} \end{aligned}$$

We will now bound both *term1* and *term2* in order to get a bound on BR_t . We will start with *term2*:

$$\begin{aligned} \text{term2} &= E[\mu(a^*) - UCB(a^*, H_t)] \leq E[\mu(a^*) - UCB(a^*, H_t)]^+ \\ &\leq \sum_a E[UCB(a, H_t) - \mu(a)]^- \\ &\leq K \frac{\gamma}{TK} = \frac{\gamma}{T} \end{aligned}$$

Before we bound *term1*, please notice that

$$\begin{aligned} E[LCB(a_t, H_t) - \mu(a_t)] &\leq E[LCB(a_t, H_t) - \mu(a_t)]^+ \\ &\leq \sum_a E[\mu(a) - LCB(a, H_t)]^- \\ &\leq K \frac{\gamma}{TK} = \frac{\gamma}{T} \end{aligned}$$

Therefore, we can bound *term1*:

$$\begin{aligned} \text{term1} &= E[UCB(a_t, H_t) - \mu(a_t)] = E[2r(a_t, H_t) + LCB(a_t, H_t) - \mu(a_t)] \\ &= 2E[r(a_t, H_t)] + E[LCB(a_t, H_t) - \mu(a_t)] \\ &\leq 2E[r(a_t, H_t)] + \frac{\gamma}{T} \end{aligned}$$

And now we can finish the proof of the lemma:

$$\begin{aligned}
 BR &= \sum_{t=1}^T BR_t \\
 &= \sum_{t=1}^T \underbrace{E[UCB(a_t, H_t) - \mu(a_t)]}_{\text{term1}} + \underbrace{E[\mu(a^*) - UCB(a^*, H_t)]}_{\text{term2}} \\
 &\leq \sum_{t=1}^T (2E[r(a_t, H_t)] + \frac{\gamma}{T}) + \frac{\gamma}{T} \\
 &= T(\frac{\gamma}{T} + \frac{\gamma}{T}) + 2 \sum_{t=1}^T E[r(a_t, H_t)] \\
 &= 2\gamma + 2 \sum_{t=1}^T E[r(a_t, H_t)]
 \end{aligned}$$

□

Lemma 5

$$\sum_{t=1}^T \sqrt{\frac{1}{n_t(a_t)}} = O(\sqrt{TK})$$

Proof:

$$\begin{aligned}
 \sum_{t=1}^T \sqrt{\frac{1}{n_t(a_t)}} &= \sum_a \sum_{t:a_t=a} \frac{1}{\sqrt{n_t(a_t)}} \\
 &= \sum_a \sum_{j=1}^{n_t(a_t)} \frac{1}{\sqrt{j}} \\
 &= \sum_a O(\sqrt{n_T(a)}) \\
 &\leq \sum_a O(\sqrt{\frac{T}{k}}) \\
 &= K \cdot O(\sqrt{\frac{T}{k}}) = O(\sqrt{TK})
 \end{aligned}$$

The reason behind the inequality is as follows: recall that $\sum_a n_T(a) = T$. The maximum of the term $\sum_a O(\sqrt{n_T(a)})$ is reached when for each action a we have $n_T(a) = \frac{T}{K}$. □

Using the last two lemmas we have just proven, we are ready to prove our main theorem.

Theorem 6

$$BR = O(\sqrt{KT \log(T)})$$

Proof:

$$\begin{aligned}
BR &\leq 2\gamma + 2 \sum_{t=1}^T E[r(a_t, H_t)] \\
&= 2\gamma + 2 \sum_{t=1}^T \sqrt{\frac{4 \log(T)}{n_t(a_t)}} \\
&= 2\gamma + 4 \log(T) \cdot \sum_{t=1}^T \sqrt{\frac{1}{n_t(a_t)}} \\
&= 2\gamma + 4 \log(T) \cdot O(\sqrt{TK}) \\
&= O(\sqrt{KT \log(T)})
\end{aligned}$$

□

5 Thompson Sampling Algorithm - Worse Case

Until now we dealt with the case where we had an initial prior distribution \mathbb{P} , which we assumed that it is being used to draw the parameters $\vec{\mu} = (\mu(a_1) \dots \mu(a_k))$ and D_a at the beginning. In this section, we will omit this assumption and discuss the performance of Thompson Sampling Algorithm when there is no assumption regarding a prior, i.e. there will be no bayesian assumptions.

However, we will assume that all the rewards are Bernoulli: For each action a , the reward at each sample is a Bernoulli random variable with an unknown estimation $\mu(a)$ which remains the same for the entire running. Of course, that means that all the rewards are always either 0 (we will call this reward “fail”) or 1 (we will call this reward “success”).

Please note that the Thompson Sampling Algorithm is defined using a known prior \mathbb{P} , so in order to continue using it, we will have to define a prior \mathbb{P} ourselves which will be used for the definition of the algorithm. We will do it now: the prior will be $Beta(1, 1)$. In other words, we “assume” that for each action a , D_a is a Bernoulli random variable with an unknown estimation $\mu(a)$ and this unknown estimation is sampled at the beginning of the algorithm and remains the same for the entire time. Now the algorithm is well defined again.

Notice the quotation mark we used in the word “assume” in the paragraph above. Until now, we assumed that there is a prior. Not only we used this assumption when we defined the Thompson algorithm, but we also used this assumption when we bounded it’s Bayesian regret. However, here we use the assumption regarding the prior $Beta(1, 1)$ only when we define the algorithm but we do not assume that it holds. In other words, the estimations of the rewards of the actions might be drowned using a completely different distribution. It is not trivial that we will manage to prove anything interesting regarding the Pseudo regret of the algorithm in such a case.

We will define $\Theta_i(t)$ to be the sample of action i at time t .

$S_i(t)$ will be the amount of successes of action i until time t (excluding).

$F_i(t)$ will be the amount of fails of action i until time t (excluding).

$n_i(t)$ will be the amount of times we played action i until time t (excluding). Clearly, $n_i(t) = S_i(t) + F_i(t)$.

a_t will be the action chosen at time t .

Now we will define the Thompson Sampling Algorithm according to the prior knowledge we declared:

We would like to analyze this algorithm for an arbitrary profile $\vec{\mu} = (\mu_1 \dots \mu_k)$. We will assume without

Algorithm 3 Thompson Sampling Algorithm for Bernoulli Prior Knowledge

for each action i :

init $S_i = 0$ and $F_i = 0$

for $1 \leq t \leq T$:

for each action i :

draw $\Theta_i(t)$ from $Beta(S_i + 1, F_i + 1)$
 $a_t = \arg \max_i \Theta_i(t)$

play the action a_t and observe the reward r_t

update $S_{a_t} = S_{a_t} + r_t$ and $F_{a_t} = F_{a_t} + (1 - r_t)$

loss of generality that $\mu^* = \mu_1 > \mu_2 > \dots > \mu_k$. Indeed, we do not assume a prior in our analysis.

We will define $\hat{\mu}_i(t) = \frac{S_i(t)}{n_i(t)+1}$ and $\Delta_i = \mu_1 - \mu_i$.

In addition, for every $i \neq 1$ (every non optimal action) we will define the following events:

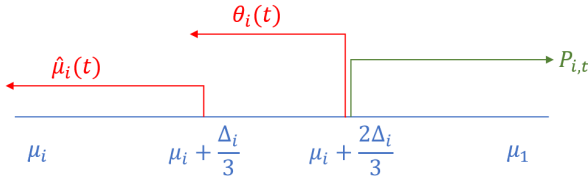
$$G_i^\mu(t) = \{\hat{\mu}_i(t) \leq \mu_i + \frac{\Delta_i}{3}\}$$

$$G_i^\Theta(t) = \{\Theta_i(t) \leq \mu_i + \frac{\Delta_i}{3}\}$$

We will also define the probability:

$$\begin{aligned} P_{i,t} &= Pr[\Theta_1(t) > \mu_i + \frac{2\Delta_i}{3} | H_{t-1}] \\ &= Pr[\Theta_1(t) > \mu_1 - \frac{\Delta_i}{3} | H_{t-1}] \end{aligned}$$

In the illustration below we can see the idea behind those 3 events:



One should note the differences between the events declared here and the events we had in Lesson 1 in the algorithms we learned there (Successive Elimination, UCB Algorithm, etc.). There, we had one type of event for each action - we wanted the estimation of the expectation to be not too far from the actual/real expectation (i.e., we wanted $|\hat{\mu}_i(t) - \mu_i|$ to be small. Here, we have the event $G_i^\mu(t)$, which is similar to what we have seen so far. The “new” event is $G_i^\Theta(t)$. The idea in $G_i^\Theta(t)$ is that we don’t want the $\Theta_i(t)$ that we sample to be too far from our estimation $\hat{\mu}_i(t)$, which is where we sample $\Theta_i(t)$ from (actually, we sample $\Theta_i(t)$ from a Beta distribution with mean value $\hat{\mu}_i(t)$). In other words, we sample $\Theta_i(t)$ from a Beta distribution with an expectation value $\hat{\mu}_i(t)$ (and in the event $G_i^\Theta(t)$ we want the value we sample $\Theta_i(t)$ to be not too far from $\hat{\mu}_i(t)$) but the estimated expectation value itself $\hat{\mu}_i(t)$ might be far from the actual/real expectation value μ_i (and in the event $G_i^\mu(t)$ we want this expectation not to be far from our the real value μ_i).

If both the events $G_i^\mu(t)$ and $G_i^\Theta(t)$ happen, we have

$$\Theta_i(t) \leq \hat{\mu}_i(t) + \frac{\Delta_i}{3} \leq (\mu_i + \frac{\Delta_i}{3}) + \frac{\Delta_i}{3} = \mu_i + \frac{2\Delta_i}{3} = \mu_1 - \frac{\Delta_i}{3}$$

so if we can have also guarantee that $P_{i,t}$ is high, we will get that

$$\Theta_i(t) \leq \mu_1 - \frac{\Delta_i}{3} < \Theta_1(t)$$

and that is exactly what we want, because we want action 1 to be chosen, and it happens when $\Theta_i(t) < \Theta_1(t)$ for all $2 \leq i \leq k$.

Lemma 7 For each time $1 \leq t \leq T$, action $2 \leq i \leq k$ and history H_{t-1} :

$$Pr[a_t = i, G_i^\mu(t), G_i^\ominus(t) | H_{t-1}] \leq \frac{1 - P_{i,t}}{P_{i,t}} Pr[a_t = 1, G_i^\mu(t), G_i^\ominus(t) | H_{t-1}]$$

The meaning of this lemma is that if $P_{i,t}$ is close to 1, then the probability of choosing a non-optimal i will be small.

Proof: If either $G_i^\mu(t)$ or $G_i^\ominus(t)$ does not happen, then we need to prove that $0 \leq \frac{1 - P_{i,t}}{P_{i,t}} \cdot 0$, so there is nothing for us to prove. Therefore, we can condition on $G_i^\mu(t)$ and $G_i^\ominus(t)$, meaning that we need to prove that

$$Pr[a_t = i | G_i^\mu(t), G_i^\ominus(t), H_{t-1}] \leq \frac{1 - P_{i,t}}{P_{i,t}} Pr[a_t = 1 | G_i^\mu(t), G_i^\ominus(t), H_{t-1}]$$

Note that $G_i^\mu(t)$ is determined by H_{t-1} , because $\hat{\mu}_i(t)$ is determined by the rewards we have seen so far. Therefore it is sufficient to prove that

$$Pr[a_t = i | G_i^\ominus(t), H_{t-1}] \leq \frac{1 - P_{i,t}}{P_{i,t}} Pr[a_t = 1 | G_i^\ominus(t), H_{t-1}]$$

For $a_t = i$ when $G_i^\ominus(t)$ happens, it is required that $\forall j : \Theta_j(t) \leq \mu_i + \frac{2\Delta_i}{3}$. Therefore

$$\begin{aligned} Pr[a_t = i | G_i^\ominus(t), H_{t-1}] &\leq Pr[\forall j : \Theta_j(t) \leq \mu_i + \frac{2\Delta_i}{3} | G_i^\ominus(t), H_{t-1}] \\ &= Pr[\Theta_1(t) \leq \mu_i + \frac{2\Delta_i}{3} | G_i^\ominus(t), H_{t-1}] \cdot Pr[\forall j \neq 1 : \Theta_j(t) \leq \mu_i + \frac{2\Delta_i}{3} | G_i^\ominus(t), H_{t-1}] \\ &= Pr[\Theta_1(t) \leq \mu_i + \frac{2\Delta_i}{3} | H_{t-1}] \cdot Pr[\forall j \neq 1 : \Theta_j(t) \leq \mu_i + \frac{2\Delta_i}{3} | G_i^\ominus(t), H_{t-1}] \\ &= (1 - P_{i,t}) \cdot Pr[\forall j \neq 1 : \Theta_j(t) \leq \mu_i + \frac{2\Delta_i}{3} | G_i^\ominus(t), H_{t-1}] \end{aligned}$$

Moreover,

$$\begin{aligned} Pr[a_t = 1 | G_i^\ominus(t), H_{t-1}] &\geq Pr[\forall j \neq 1 : \Theta_1(t) > \mu_i + \frac{2\Delta_i}{3} \geq \Theta_j(t) | G_i^\ominus(t), H_{t-1}] \\ &= Pr[\Theta_1(t) > \mu_i + \frac{2\Delta_i}{3} | G_i^\ominus(t), H_{t-1}] \cdot Pr[\forall j \neq 1 : \mu_i + \frac{2\Delta_i}{3} \geq \Theta_j(t) | G_i^\ominus(t), H_{t-1}] \\ &= Pr[\Theta_1(t) > \mu_i + \frac{2\Delta_i}{3} | H_{t-1}] \cdot Pr[\forall j \neq 1 : \Theta_j(t) \leq \mu_i + \frac{2\Delta_i}{3} | G_i^\ominus(t), H_{t-1}] \\ &= P_{i,t} \cdot Pr[\forall j \neq 1 : \Theta_j(t) \leq \mu_i + \frac{2\Delta_i}{3} | G_i^\ominus(t), H_{t-1}] \end{aligned}$$

and therefore

$$\begin{aligned} Pr[a_t = i | G_i^\ominus(t), H_{t-1}] &\leq (1 - P_{i,t}) \cdot Pr[\forall j \neq 1 : \Theta_j(t) \leq \mu_i + \frac{2\Delta_i}{3} | G_i^\ominus(t), H_{t-1}] \\ &\leq \frac{1 - P_{i,t}}{P_{i,t}} \cdot Pr[a_t = 1 | G_i^\ominus(t), H_{t-1}] \end{aligned}$$

□

The pseudo regret equals to $\sum_{i=2}^k E[n_i(T)] \cdot \Delta_i$. We will analyze now $E[n_i(T)]$.

$$\begin{aligned} E[n_i(T)] &= \sum_{t=1}^T Pr[a_t = i] \\ &= \underbrace{\sum_{t=1}^T Pr[a_t = i, G_i^\mu(t), G_i^\Theta(t)]}_{term1} + \underbrace{\sum_{t=1}^T Pr[a_t = i, G_i^\mu(t), \neg G_i^\Theta(t)]}_{term2} + \underbrace{\sum_{t=1}^T Pr[a_t = i, \neg G_i^\mu(t)]}_{term3} \end{aligned}$$

We will bound *term1*, *term2* and *term3* in order to get a bound for $E[n_i(T)]$. In *term1* we deal with the case that both the 2 good events happen. In *term2* we deal with the case where the estimation of the expectation $\hat{\mu}_i(t)$ is close to the real value μ_i but the $\Theta_i(t)$ we sample from the Bernoulli variable with expectation $\hat{\mu}_i(t)$ is too far from this expectation. In *term3* we deal with the case where our estimation of the expectation $\hat{\mu}_i(t)$ is too far from the real expectation μ_i .

5.1 Bounding term 1

$$\begin{aligned} \sum_{t=1}^T Pr[a_t = i, G_i^\mu(t), G_i^\Theta(t)] &= \sum_{t=1}^T E[Pr[a_t = i, G_i^\mu(t), G_i^\Theta(t) | H_{t-1}]] \\ &\leq \sum_{t=1}^T E\left[\frac{1 - P_{i,t}}{P_{i,t}} Pr[a_t = 1, G_i^\mu(t), G_i^\Theta(t) | H_{t-1}]\right] = \sum_{t=1}^T E\left[\frac{1 - P_{i,t}}{P_{i,t}} I(a_t = 1, G_i^\mu(t), G_i^\Theta(t))\right] \end{aligned}$$

Where the inequality is by using Lemma 7.

Note that everytime that this indicator $I(\cdot) = 1$ then $n_1(t)$ also increases by 1. We will look at those times and denote them by τ_j . So:

$$\begin{aligned} \sum_{t=1}^T E\left[\frac{1 - P_{i,t}}{P_{i,t}} I(a_t = 1, G_i^\mu(t), G_i^\Theta(t))\right] &= \sum_{j=1}^T E\left[\frac{1 - P_{i,\tau_j}}{P_{i,\tau_j}} \underbrace{\sum_{t=\tau_j}^{\tau_{j+1}-1} I(a_t = 1, G_i^\mu(t), G_i^\Theta(t))}_1\right] \\ &\leq \sum_{j=1}^T E\left[\frac{1 - P_{i,\tau_j}}{P_{i,\tau_j}}\right] \end{aligned}$$

The second inequality is because from time τ_j (including) to time τ_{j+1} (excluding) there is always exactly one time where we have played action 1, which means that the corresponding indicator will be equal to 1 and all the other indicators will be equal to 0, so the sum of these indicators will be 1.

Lemma 8

$$E\left[\frac{1 - P_{i,\tau_j}}{P_{i,\tau_j}}\right] \leq \begin{cases} O\left(\frac{1}{\Delta_i}\right) & j \leq \frac{8}{\Delta_i} \\ O\left(\frac{1}{\Delta_i^2}\right) \exp\left(-\frac{\Delta_i^2}{2}\right) & j \geq \frac{8}{\Delta_i} \end{cases}$$

We will not prove this lemma in the lectures (it appears in the paper). Now we can use the lemma to bound term one:

$$\begin{aligned} \sum_{t=1}^T Pr[a_t = i, G_i^\mu(t), G_i^\Theta(t)] &\leq \sum_{t=1}^T E\left[\frac{1 - P_{i,t}}{P_{i,t}} \cdot I[a_t = 1, G_i^\mu(t), G_i^\Theta(t)]\right] \\ &\leq \sum_{j=1}^T E\left[\frac{1 - P_{i,\tau_j}}{P_{i,\tau_j}}\right] \leq O\left(\frac{1}{\Delta_i^2} + \frac{1}{\Delta_i}\right) \end{aligned}$$

5.2 Bounding term 2

We will now show what happens in the case where the estimation of the expectation $\hat{\mu}_i(t)$ is close to the real value μ_i but the $\Theta_i(t)$ we sample from the Bernoulli variable with expectation $\hat{\mu}_i(t)$ is too far from the expectation:

Lemma 9

$$\forall i \neq 1. \sum_{t=1}^T \Pr[a_t = i, G_i^\mu(t), -G_i^\Theta(t)] = O\left(\frac{\log(T)}{\Delta_i^2}\right)$$

Proof idea:

Define $L_i = \frac{c \cdot \log(T)}{\Delta_i}$. We allow L_i times to take action i (so the sum of probabilities will be L_i) and after L_i times that we use action i we can bound the following term using Hoeffding's inequality:

$$E\left[\sum_{t=1}^T I[n_i(t) > L_i, \hat{\mu}_i(t) \leq \mu_i + \frac{\Delta_i}{3}] \cdot \Pr[\Theta_i(t) > \mu_i + \frac{2 \cdot \Delta_i}{3} | H_{t-1}]\right] \leq \frac{1}{T}$$

There is a need to continue the proof by presenting (and using) a connection to the Binomial distribution (recall that the algorithm uses this distribution). We will not see this in the lesson (the full proof is in the paper).

5.3 Bounding term 3

In this case we chose a non-optimal action and our estimation of the expectation $\hat{\mu}_i(t)$ is too far from the real expectation μ_i

Lemma 10

$$\forall i \neq 1. \sum_{t=1}^T \Pr[a_t = i, -G_i^\mu(t)] = O\left(\frac{1}{\Delta_i^2}\right)$$

Proof:

$$\sum_{t=1}^T \Pr[a_t = i, -G_i^\mu(t)] \leq \sum_{t=1}^T \Pr[-G_i^\mu(t)]$$

By Hoeffding's Inequality we get:

$$\leq \sum_{j=1}^T \exp\left(\frac{-\Delta_i^2 \cdot j}{9}\right) = O\left(\frac{1}{\Delta_i^2}\right)$$

□

5.4 Bounding the regret

By summing over the bounds of each term, we get:

$$E[n_i(T)] = O\left(\frac{1}{\Delta_i^4}\right) + O\left(\frac{\log(T)}{\Delta_i^2}\right) + O\left(\frac{1}{\Delta_i^2}\right) = O\left(\frac{1}{\Delta_i^4} + \frac{\log(T)}{\Delta_i^2}\right)$$

Therefore, the Pseudo Regret will be:

$$\sum_{i=2}^k E[n_i(T)] \cdot \Delta_i = \sum_{i=2}^k O\left(\frac{1}{\Delta_i^3} + \frac{\log(T)}{\Delta_i}\right)$$

6 References

1. Slivkins, Chapter 3:
 - (a) Bayesian bandits
 - (b) Bayesian regret Thompson Sampling
2. Worse case: Agrawal Goyal, JACM, 2017