Lecture 3 - Thompson Sampling for Bandits

Pratik Gajane

September 14, 2022

2AMM20 Research Topics in Data Mining
Eindhoven University of Technology

A Quick Recap of Lecture 1

- Introduction to reinforcement learning (RL).
- Mathematical formulation of a RL problem.
- Formulating RL with multi-armed bandits and its variants.
- Formulating RL with Markov decision processes.

Recap Lecture 2 : Stationary stochastic bandits

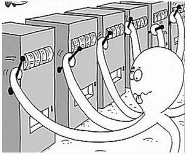


Image source : *Microsoft research*

- At each time step t, the agent selects an action $i(t)$ and then receives a numerical reward $r(t) \sim X_{i(t)}$ with mean $\mu_{i(t)}$.

Recap Lecture 2 : Stationary stochastic bandits

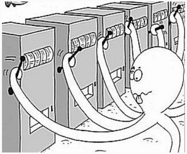


Image source : *Microsoft research*

- At each time step t, the agent selects an action $i(t)$ and then receives a numerical reward $r(t) \sim X_{i(t)}$ with mean $\mu_{i(t)}$.
- Agent's goal : Minimize the expected regret of its policy π

$$\mathfrak{R}_\pi(T) := \underbrace{T\mu_*}_{\text{Optimal expected cumulative reward}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} r(t) \mid \pi\right]}_{\text{Expected cumulative reward of } \pi}$$

where μ_* is the optimal mean reward and T is the horizon.

Recap Lecture 2 : Stationary stochastic bandits

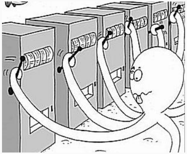


Image source : *Microsoft research*

- At each time step t, the agent selects an action $i(t)$ and then receives a numerical reward $r(t) \sim X_{i(t)}$ with mean $\mu_{i(t)}$.
- Agent's goal : Minimize the expected regret of its policy π

$$\mathfrak{R}_\pi(T) := \underbrace{T\mu_*}_{\text{Optimal expected cumulative reward}} - \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} r(t) \mid \pi\right]}_{\text{Expected cumulative reward of } \pi}$$

where μ_* is the optimal mean reward and T is the horizon.

- Our aim : Construct an algorithm with sub-linear regret (featuring terms like $\sqrt{T}$ or $\log T$, but not T).

Recap Lecture 2 : UCB

Algorithm UCB algorithm Auer et al. [2002]

Parameters : Confidence level δ

for $t = 1, \ldots, K$ **do**

  Choose each arm once.

end for

for $t = K + 1, \ldots$ **do**

  Compute empirical means $\hat{\mu}_1(t-1), \ldots, \hat{\mu}_K(t-1)$.

  Select arm $i(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right]$.

end for

Recap Lecture 2 : UCB

Algorithm UCB algorithm Auer et al. [2002]

Parameters : Confidence level δ

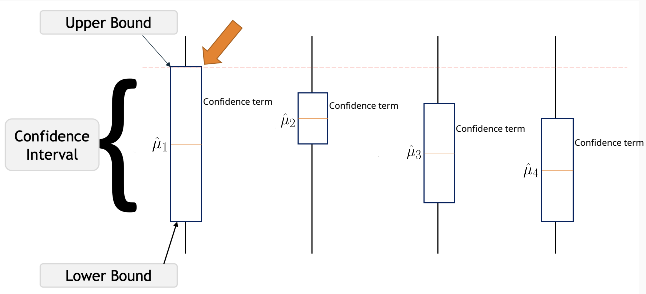
1: **for** $t = 1, \ldots, K$ **do**
2: Choose each arm once.
3: **end for**
4: **for** $t = K + 1, \ldots$ **do**
5: Compute empirical means $\hat{\mu}_1(t-1), \ldots, \hat{\mu}_K(t-1)$.
6: Select arm $i(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right]$.
7: **end for**

- Distribution-dependent regret bound $\displaystyle\sum_{a:\Delta_a>0} \frac{16\log(T)}{\Delta_a} + 3\Delta_a$

  (recall that $\Delta_a = \mu_* - \mu_a$).
- Distribution-free regret bound $O(\sqrt{KT\log(T)})$.

$f(x) = O(g(x))$, if $f(x) < Cg(x)$ for all $x > n$. For more information, click here.

UCB : Solving Bandits from a Frequentist Perspective

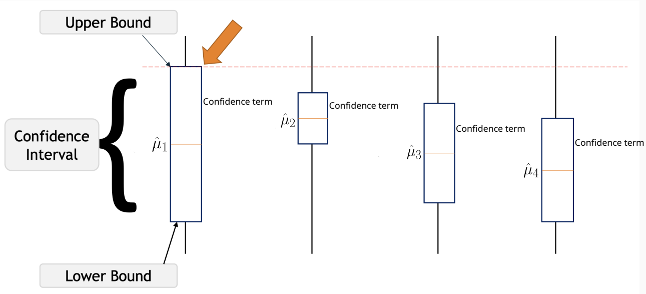


- Build confidence intervals around empirical mean rewards.

  Confidence term for arm $a = \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}$

  Confidence interval for arm $a = \left\{\hat{\mu}_a - \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}, \hat{\mu}_a + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right\}$

UCB : Solving Bandits from a Frequentist Perspective



- Build confidence intervals around empirical mean rewards.

  Confidence term for arm $a = \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}$

  Confidence interval for arm $a = \left\{\hat{\mu}_a - \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}, \hat{\mu}_a + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right\}$

- Arm selection rule using the size of the confidence interval.

  Select arm $i(t) = \arg\max_a \left[\hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{N_a(t-1)}}\right]$.

Lecture 3 : Outline

- Solving Bandits from a Bayesian Perspective
- Thompson Sampling
- Regret Bound for Thompson Sampling

Solving Bandits from a Bayesian Perspective

Solving Bandits from a Bayesian Perspective

Define a prior distribution that incorporates your subjective beliefs about unknown parameters i.e. mean rewards.

Solving Bandits from a Bayesian Perspective

Define a prior distribution that incorporates your subjective beliefs about unknown parameters i.e. mean rewards.

At each time step t,

1. Sample a particular set of parameters from the prior.

Solving Bandits from a Bayesian Perspective

Define a prior distribution that incorporates your subjective beliefs about unknown parameters i.e. mean rewards.

At each time step t,

1. Sample a particular set of parameters from the prior.
2. Select arm $i(t) = \arg\max_j \text{reward}_j \mid \text{parameters}$

Solving Bandits from a Bayesian Perspective

Define a prior distribution that incorporates your subjective beliefs about unknown parameters i.e. mean rewards.
At each time step t,

1. Sample a particular set of parameters from the prior.
2. Select arm $i(t) = \arg\max_j \text{reward}_j \mid \text{parameters}$
3. Observe reward and update posterior.

Solving Bandits from a Bayesian Perspective

Define a prior distribution that incorporates your subjective beliefs about unknown parameters i.e. mean rewards.
At each time step t,

1. Sample a particular set of parameters from the prior.
2. Select arm $i(t) = \arg\max_j \text{reward}_j \mid \text{parameters}$
3. Observe reward and update posterior.
   (Prior at time $t + 1 \leftarrow$ posterior at time t)

Choice of Prior : Beta Prior

Solving bandits from a Bayesian perspective

Choose a prior for the mean reward of each arm.
At each time step,

1. Sample a particular set of parameters from the prior.
2. Select arm $i(t) = \arg\max_j \text{reward}_j \mid \text{parameters}$
3. Observe reward and update posterior.

Choice of Prior : Beta Prior

Solving bandits from a Bayesian perspective

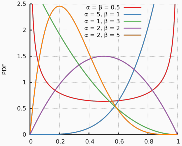
Choose a prior for the mean reward of each arm.
At each time step,

1. Sample a particular set of parameters from the prior.
2. Select arm $i(t) = \arg\max_j \text{reward}_j \mid \text{parameters}$
3. Observe reward and update posterior.

- $Beta(\alpha, \beta)$ is a family of continuous distributions defined on $[0, 1]$.

Probability density function for $Beta(\alpha, \beta)$:

$$f(x, \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du}$$



Choice of Prior : Beta Prior

Solving bandits from a Bayesian perspective

Choose a prior for the mean reward of each arm.
At each time step,

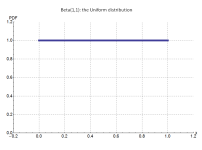
1. Sample a particular set of parameters from the prior.
2. Select arm $i(t) = \arg\max_j \text{reward}_j \mid \text{parameters}$
3. Observe reward and update posterior.

- $Beta(\alpha, \beta)$ is a family of continuous distributions defined on $[0, 1]$.

Probability density function for $Beta(\alpha, \beta)$:

$$f(x, \alpha, \beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 u^{\alpha-1}(1-u)^{\beta-1} du}$$

- $Beta(1, 1) \equiv$ uniform distribution on $[0, 1]$.



Updating Posterior: Bernoulli Rewards using Beta Prior

Solving bandits from a Bayesian perspective

Choose a prior for the mean reward of each arm.
At each time step,

1. Sample a particular set of parameters from the prior.
2. Select arm $i(t) = \arg\max_j \text{reward} \mid \text{parameters}$
3. **Observe reward and update posterior**.

Updating Posterior : Bernoulli Rewards using Beta Prior

Solving bandits from a Bayesian perspective

Choose a prior for the mean reward of each arm.
At each time step,

1. Sample a particular set of parameters from the prior.
2. Select arm $i(t) = \arg\max_j \text{reward} \mid \text{parameters}$
3. **Observe reward and update posterior**.

- For Bernoulli rewards (i.e. rewards either 0 or 1), interpret $Beta(\alpha, \beta)$ parameters as follows :
  - $\alpha - 1$ as the number of previous $1's$ and
  - $\beta - 1$ as the number of previous $0's$.

Updating Posterior : Bernoulli Rewards using Beta Prior

Solving bandits from a Bayesian perspective

Choose a prior for the mean reward of each arm.
At each time step,

1. Sample a particular set of parameters from the prior.
2. Select arm $i(t) = \arg\max_j$ reward | parameters
3. **Observe reward and update posterior**.

- For Bernoulli rewards (i.e. rewards either 0 or 1), interpret $Beta(\alpha, \beta)$ parameters as follows :
  - $\alpha - 1$ as the number of previous $1's$ and
  - $\beta - 1$ as the number of previous $0's$.
- After observing a Bernoulli reward,

  if the reward is 1,

  then the posterior distribution is $Beta(\alpha + 1, \beta)$

  if the reward is 0,

  then the posterior distribution is $Beta(\alpha, \beta + 1)$.

Updating Posterior : Bernoulli Rewards using Beta Prior

Solving bandits from a Bayesian perspective

Choose a prior for the mean reward of each arm.
At each time step,

1. Sample a particular set of parameters from the prior.
2. Select arm $i(t) = \arg\max_j \text{reward} \mid \text{parameters}$
3. **Observe reward and update posterior**.

- For Bernoulli rewards (i.e. rewards either 0 or 1), interpret $Beta(\alpha, \beta)$ parameters as follows :
  - $\alpha - 1$ as the number of previous $1's$ and
  - $\beta - 1$ as the number of previous $0's$.
- After observing a Bernoulli reward,
  if the reward is 1,
  then the posterior distribution is $Beta(\alpha + 1, \beta)$
  if the reward is 0,
  then the posterior distribution is $Beta(\alpha, \beta + 1)$.

Why Beta prior? Because Beta is the conjugate prior for Bernoulli distribution. For more information, click here.

Thompson Sampling

Thompson Sampling algorithm

Algorithm Thompson sampling with Beta prior for Bernoulli rewards

for $i = 1, \ldots, K$ **do**
    Initialize $\text{Success}_i = 0$ and $\text{Failure}_i = 0$
end for
for $t = 1, \ldots, T$ **do**
    for $i = 1, \ldots, K$ **do**
        Sample $\theta_i(t) \sim \mathit{Beta}(\text{Success}_i + 1, \text{Failure}_i + 1)$
    end for
    Select arm $i(t) = \arg\max_j \theta_j(t)$.
    Observe reward $r(t)$.
    if $r(t) = 1$ **then**
        $\text{Success}_{i(t)} = \text{Success}_{i(t)} + 1$
    else
        $\text{Failure}_{i(t)} = \text{Failure}_{i(t)} + 1$
    end if
end for

Thompson Sampling algorithm

Algorithm Thompson sampling with Beta prior for Bernoulli rewards

```
for i = 1, ..., K do
    Initialize Success_i = 0 and Failure_i = 0
end for
for t = 1, ..., T do
    for i = 1, ..., K do
        Sample θ_i(t) ~ Beta(Success_i + 1, Failure_i + 1)
    end for
    Select arm i(t) = arg max_j θ_j(t).
    Observe reward r(t).
    if r(t) = 1 then
        Success_{i(t)} = Success_{i(t)} + 1
    else
        Failure_{i(t)} = Failure_{i(t)} + 1
    end if
end for
```

Thompson Sampling algorithm

Algorithm Thompson sampling with Beta prior for Bernoulli rewards

```
for i = 1, ..., K do
    Initialize Success_i = 0 and Failure_i = 0
end for
for t = 1, ..., T do
    for i = 1, ..., K do
        Sample θ_i(t) ~ Beta(Success_i + 1, Failure_i + 1)
    end for
    Select arm i(t) = arg max_j θ_j(t).
    Observe reward r(t).
    if r(t) = 1 then
        Success_{i(t)} = Success_{i(t)} + 1
    else
        Failure_{i(t)} = Failure_{i(t)} + 1
    end if
end for
```

Thompson Sampling algorithm

Algorithm Thompson sampling with Beta prior for Bernoulli rewards

```
for i = 1, ..., K do
    Initialize Success_i = 0 and Failure_i = 0
end for
for t = 1, ..., T do
    for i = 1, ..., K do
        Sample θ_i(t) ~ Beta(Success_i + 1, Failure_i + 1)
    end for
    Select arm i(t) = arg max_j θ_j(t).
    Observe reward r(t).
    if r(t) = 1 then
        Success_{i(t)} = Success_{i(t)} + 1
    else
        Failure_{i(t)} = Failure_{i(t)} + 1
    end if
end for
```

Thompson Sampling algorithm

Algorithm Thompson sampling with Beta prior for Bernoulli rewards

for $i = 1, \ldots, K$ **do**
    Initialize $\text{Success}_i = 0$ and $\text{Failure}_i = 0$
end for
for $t = 1, \ldots, T$ **do**
    for $i = 1, \ldots, K$ **do**
        Sample $\theta_i(t) \sim \mathit{Beta}(\text{Success}_i + 1, \text{Failure}_i + 1)$
    end for
    Select arm $i(t) = \arg\max_j \theta_j(t)$.
    Observe reward $r(t)$.
    if $r(t) = 1$ **then**
        $\text{Success}_{i(t)} = \text{Success}_{i(t)} + 1$
    else
        $\text{Failure}_{i(t)} = \text{Failure}_{i(t)} + 1$
    end if
end for

Thompson Sampling algorithm

Algorithm Thompson sampling with Beta prior for Bernoulli rewards

for $i = 1, \ldots, K$ **do**
    Initialize $\text{Success}_i = 0$ and $\text{Failure}_i = 0$
end for
for $t = 1, \ldots, T$ **do**
    for $i = 1, \ldots, K$ **do**
        Sample $\theta_i(t) \sim \mathit{Beta}(\text{Success}_i + 1, \text{Failure}_i + 1)$
    end for
    Select arm $i(t) = \arg\max_j \theta_j(t)$.
    Observe reward $r(t)$.
    if $r(t) = 1$ **then**
        $\text{Success}_{i(t)} = \text{Success}_{i(t)} + 1$
    else
        $\text{Failure}_{i(t)} = \text{Failure}_{i(t)} + 1$
    end if
end for

Thompson Sampling algorithm

Algorithm Thompson sampling with Beta prior for Bernoulli rewards

for $i = 1, \dots, K$ **do**
    Initialize $\text{Success}_i = 0$ and $\text{Failure}_i = 0$
end for
for $t = 1, \dots, T$ **do**
    for $i = 1, \dots, K$ **do**
        Sample $\theta_i(t) \sim \mathit{Beta}(\text{Success}_i + 1, \text{Failure}_i + 1)$
    end for
    Select arm $i(t) = \arg\max_j \theta_j(t)$.
    Observe reward $r(t)$.
    if $r(t) = 1$ **then**
        $\text{Success}_{i(t)} = \text{Success}_{i(t)} + 1$
    else
        $\text{Failure}_{i(t)} = \text{Failure}_{i(t)} + 1$
    end if
end for

Why does Thompson Sampling work?

- Arm selection : Select arm $i(t) = \arg\max_j \theta_j(t)$.

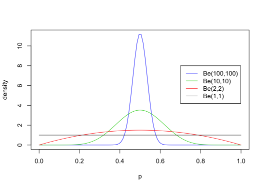
Why does Thompson Sampling work?

- Arm selection : Select arm $i(t) = \arg\max_j \theta_j(t)$.
- Exploration via randomization

  $\theta_i(t) \sim \mathit{Beta}(\mathrm{Success}_i + 1, \mathrm{Failure}_i + 1)$

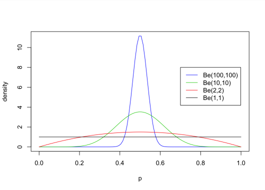
Why does Thompson Sampling work?

- Arm selection : Select arm $i(t) = \arg\max_j \theta_j(t)$.
- Exploration via randomization
  $\theta_i(t) \sim Beta(\text{Success}_i + 1, \text{Failure}_i + 1)$
- Initially the posterior might be poorly concentrated, then the fluctuations in θ's are likely to be large and TS will explore.

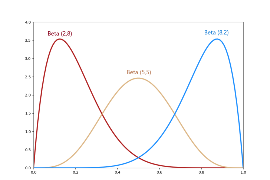


Why does Thompson Sampling work?

- Arm selection : Select arm $i(t) = \arg\max_j \theta_j(t)$.
- Exploration via randomization
  $\theta_i(t) \sim \mathit{Beta}(\text{Success}_i + 1, \text{Failure}_i + 1)$
- Initially the posterior might be poorly concentrated, then the fluctuations in θ's are likely to be large and TS will explore.



- After a large number of observations, the posterior concentrates around the true mean and the rate of exploration decreases.



Regret Bound for Thompson Sampling

Regret Bound for Thompson Sampling

Theorem (Theorem 1 from Agrawal and Goyal [2013])

After T time steps, the expected cumulative regret of Thompson sampling using Beta priors is

$$Regret = \mathfrak{R}(T) \leq (1+\epsilon)^2 \sum_i \frac{\log T}{c} \Delta_i + O\left(\frac{K}{\epsilon^2}\right),$$

where c is a problem-dependent constant.

Proving the Regret Bound : Preliminaries I

- True mean reward of arm i is μ_i.
- By default, 1 is the optimal arm i.e. μ_1 is the optimal mean.

Proving the Regret Bound : Preliminaries I

- True mean reward of arm i is μ_i.
- By default, 1 is the optimal arm i.e. μ_1 is the optimal mean.
- Arm being played at time $t = i(t)$.

Proving the Regret Bound : Preliminaries I

- True mean reward of arm i is μ_i.
- By default, 1 is the optimal arm i.e. μ_1 is the optimal mean.
- Arm being played at time $t = i(t)$.
- $N_i(t) :=$ Number of times arm i is played till $t = \sum_{\tau=1}^{t} \mathbb{I}(i(\tau) = i)$.

Proving the Regret Bound: Preliminaries I

- True mean reward of arm i is μ_i.
- By default, 1 is the optimal arm i.e. μ_1 is the optimal mean.
- Arm being played at time $t = i(t)$.
- $N_i(t) :=$ Number of times arm i is played till $t = \sum_{\tau=1}^{t} \mathbb{I}(i(\tau) = i)$.
- Empirical mean of arm i at $t = \hat{\mu}_i(t) := \frac{1}{N_i(t)} \sum_{\tau=1}^{t} (r(\tau) | i(\tau) = i)$.

Proving the Regret Bound: Preliminaries I

- True mean reward of arm i is μ_i.
- By default, 1 is the optimal arm i.e. μ_1 is the optimal mean.
- Arm being played at time $t = i(t)$.
- $N_i(t) :=$ Number of times arm i is played till $t = \sum_{\tau=1}^{t} \mathbb{I}(i(\tau) = i)$.
- Empirical mean of arm i at $t = \hat{\mu}_i(t) := \frac{1}{N_i(t)} \sum_{\tau=1}^{t} (r(\tau) | i(\tau) = i)$.
- Sampled parameter of arm $i = \theta_i(t)$.

Proving the Regret Bound: Preliminaries II

Recall from the last lecture

Lemma

$$\mathit{Regret} = \mathfrak{R}(T) = \sum_{i=1,\ldots,K,\Delta_i>0} \Delta_i \, \mathbb{E}[N_i(T)].$$

- Suboptimality gap $\Delta_i \coloneqq \mu_* - \mu_i$ where μ_* is the optimal mean reward and μ_i is the mean reward for arm a.
- $N_i(T) \coloneqq$ Number of times arm i is played till $T = \sum_{t=1}^{T} \mathbb{I}(i(t) = i)$.

Proving the Regret Bound: Preliminaries II

Recall from the last lecture

Lemma

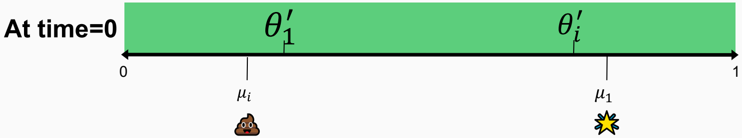
$$Regret = \mathfrak{R}(T) = \sum_{i=1,\ldots,K,\Delta_i>0} \Delta_i \, \mathbb{E}[N_i(T)].$$

- Suboptimality gap $\Delta_i := \mu_* - \mu_i$ where μ_* is the optimal mean reward and μ_i is the mean reward for arm a.
- $N_i(T) :=$ Number of times arm i is played till $T = \sum_{t=1}^{T} \mathbb{I}(i(t) = i)$.
- In order to bound $\mathfrak{R}(T)$, we need to bound $\mathbb{E}[N_i(T)]$.

When does Thompson Sampling Perform Well? I

Arm selection rule of Thompson sampling

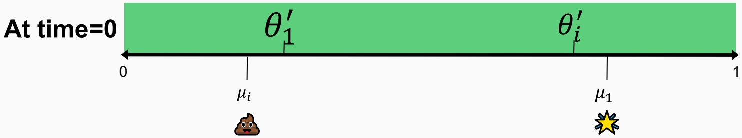
Select arm $i(t) = \arg\max_j \theta_j(t)$.



When does Thompson Sampling Perform Well? I

Arm selection rule of Thompson sampling

Select arm $i(t) = \arg\max_j \theta_j(t)$.

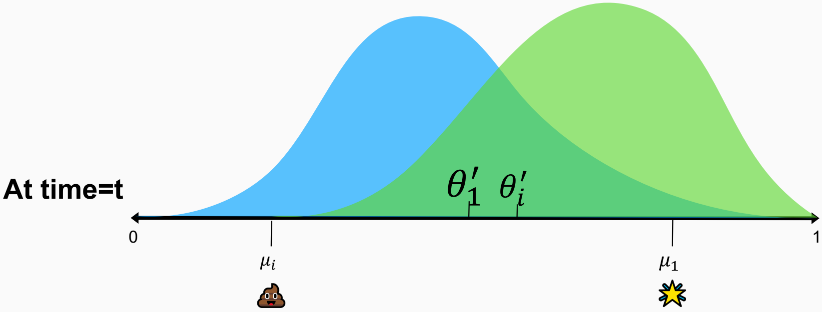


- Initially, all θ's are from the same distribution *Beta*$(1, 1)$ (i.e., the uniform distribution on $[0, 1]$), so not yet! ☹

When does Thompson Sampling Perform Well? II

Arm selection rule of Thompson sampling

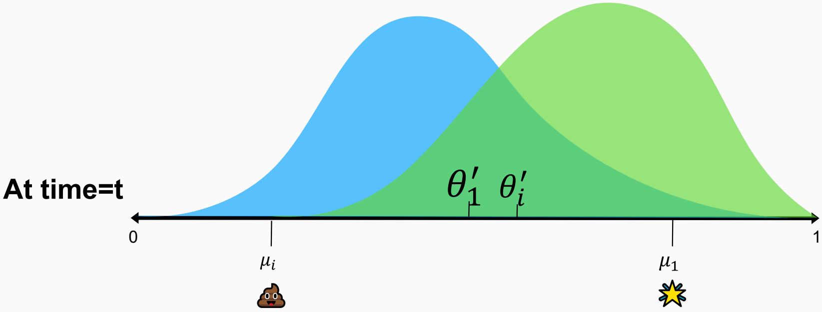
Select arm $i(t) = \arg\max_j \theta_j(t)$.



When does Thompson Sampling Perform Well? II

Arm selection rule of Thompson sampling

Select arm $i(t) = \arg\max_j \theta_j(t)$.

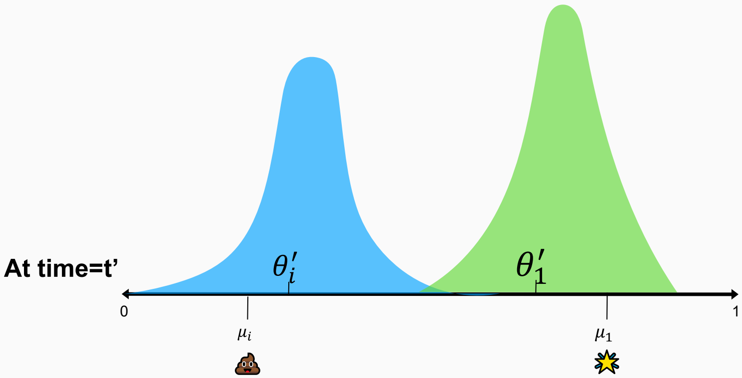


- At t, θ's are too far from μ's, so not yet! ☹

When does Thompson Sampling Perform Well? III

Arm selection rule of Thompson sampling

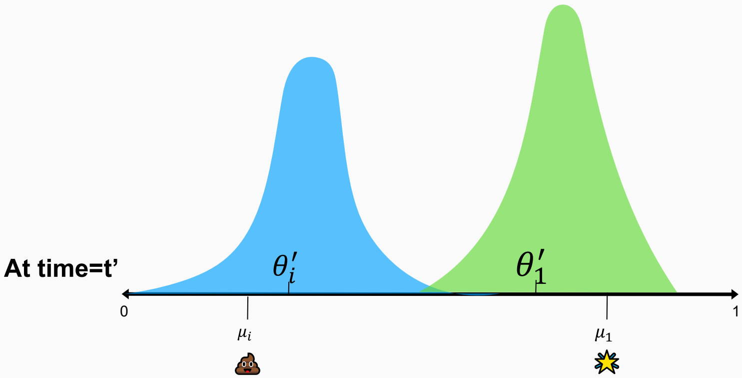
Select arm $i(t) = \arg\max_j \theta_j(t)$.



When does Thompson Sampling Perform Well? III

Arm selection rule of Thompson sampling

Select arm $i(t) = \arg\max_j \theta_j(t)$.



- At t', when θ's are close μ's. 🙂

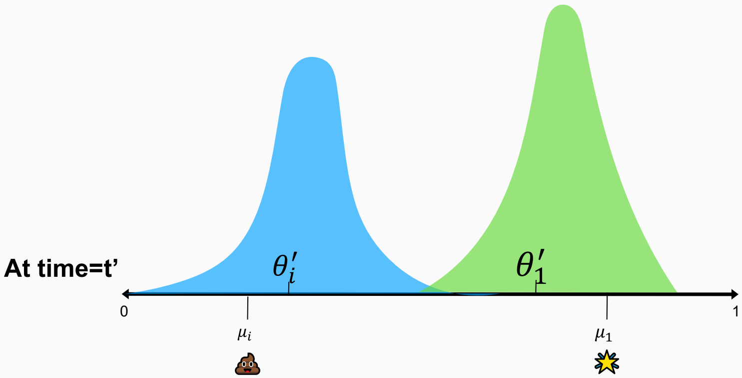
Break

We start again after a break.

When does Thompson Sampling Perform Well?

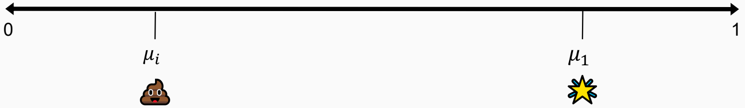
Arm selection rule of Thompson sampling

Select arm $i(t) = \arg\max_j \theta_j(t)$.



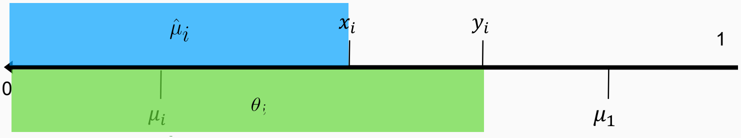
- At t', when θ's are close μ's. 🙂

Proving the Regret Bound : Defining the Good Events



- $E_i^{\theta}(t) \coloneqq$ sampled parameter θ_i is close to μ_i.
- $E_i^{\mu}(t) \coloneqq$ estimated mean $\hat{\mu}_i$ is close to μ_i.

Proving the Regret Bound: Defining the Good Events



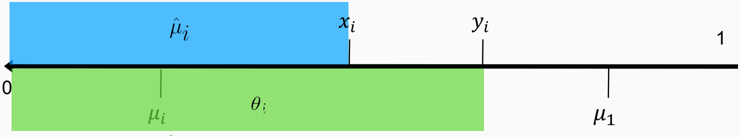
- For each suboptimal arm i, let x_i and y_i be two thresholds such that $\mu_i < x_i < y_i < \mu_1$.
- $E_i^{\theta}(t) :=$ sampled parameter θ_i is close to μ_i,

$$\boxed{E_i^{\theta}(t) := \{\theta_i < y_i\}}.$$

- $E_i^{\mu}(t) :=$ estimated mean $\hat{\mu}_i$ is close to μ_i,

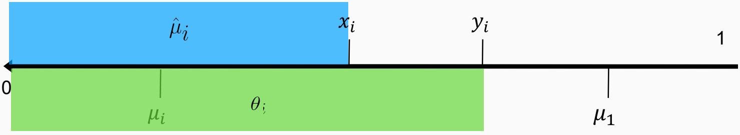
$$\boxed{E_i^{\mu}(t) := \{\hat{\mu}_i < x_i\}}.$$

Proving the Regret Bound : Decomposition into Three Terms



$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}(i(t) = i)$$

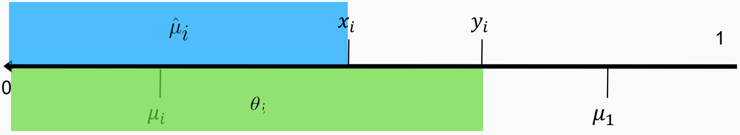
Proving the Regret Bound: Decomposition into Three Terms



$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}(i(t) = i)$$

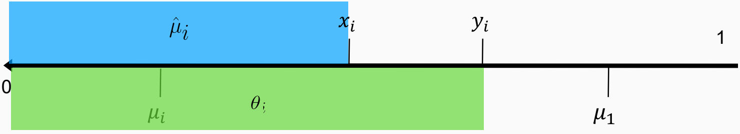
$$= \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \boxed{E_i^{\mu}(t)}, \boxed{E_i^{\theta}(t)}\right) + \dots$$

Proving the Regret Bound : Decomposition into Three Terms



$$
\begin{aligned}
\mathbb{E}[N_i(T)] &= \sum_{t=1}^{T} \mathbb{P}(i(t) = i) \\
&= \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), E_i^{\theta}(t)\right) \\
&\quad + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right) + \dots
\end{aligned}
$$

Proving the Regret Bound: Decomposition into Three Terms



$$
\begin{aligned}
\mathbb{E}[N_i(T)] &= \sum_{t=1}^{T} \mathbb{P}(i(t) = i) \\
&= \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), E_i^{\theta}(t)\right) \\
&\quad + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right) \\
&\quad + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \overline{E_i^{\mu}(t)}\right)
\end{aligned}
$$

Proving the Regret Bound: Analyzing the First Term I

$$\mathbb{E}[N_i(T)] = \boxed{\sum_{t=1}^{T} \mathbb{P}\left(i(t)=i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right)} + \sum_{t=1}^{T} \mathbb{P}\left(i(t)=i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t)=i,\ \overline{E_i^{\mu}(t)}\right)$$

Proving the Regret Bound: Analyzing the First Term I

$$\mathbb{E}[N_i(T)] = \boxed{\sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right)} + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$

- Let "history" $\mathcal{F}_{t-1} = i(1), r(1), i(2), r(2), \ldots, i(t-1), r(t-1)$.

Proving the Regret Bound: Analyzing the First Term I

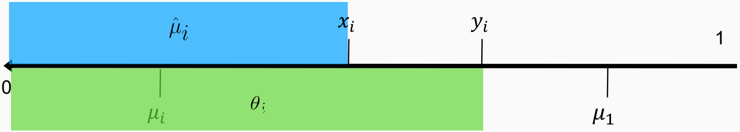
$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \overline{E_i^{\mu}(t)}\right)$$

- Let "history" $\mathcal{F}_{t-1} = i(1), r(1), i(2), r(2), \ldots, i(t-1), r(t-1)$.

Lemma (Main Lemma. Lemma 1 from Agrawal and Goyal [2013])

For all $t = 1, \ldots, T$ and all suboptimal arms i i.e. $i \neq 1$,

$$\mathbb{P}\left(i(t) = i, E_i^{\mu}(t), E_i^{\theta}(t) \mid \mathcal{F}_{t-1}\right) \leq \textit{Coefficient} \cdot \mathbb{P}\left(i(t) = 1, E_i^{\mu}(t), E_i^{\theta}(t) \mid \mathcal{F}_{t-1}\right)$$



Proving the Regret Bound: Analyzing the First Term I

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$

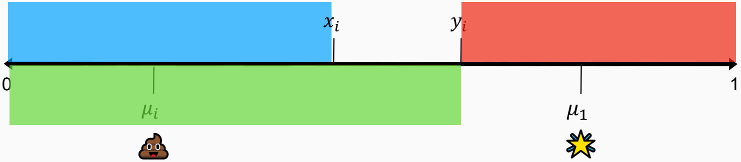
- Let "history" $\mathcal{F}_{t-1} = i(1), r(1), i(2), r(2), \ldots, i(t-1), r(t-1)$.

Lemma (Main Lemma. Lemma 1 from Agrawal and Goyal [2013])

For all $t = 1, \ldots, T$ and all suboptimal arms i i.e. $i \neq 1$,

$$\mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t) \mid \mathcal{F}_{t-1}\right) \leq \frac{(1 - p_{i,t})}{p_{i,t}} \mathbb{P}\left(i(t) = 1,\ E_i^{\mu}(t),\ E_i^{\theta}(t) \mid \mathcal{F}_{t-1}\right)$$

where $p_{i,t} \coloneqq \mathbb{P}(\theta_1(t) > y_i \mid \mathcal{F}_{t-1})$



Proving the Regret Bound : Analyzing the First Term II

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t) \right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)} \right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)} \right)$$

Proving the Regret Bound: Analyzing the First Term II

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$

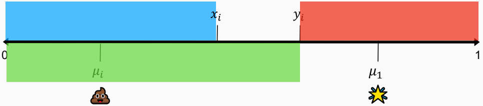
$$\text{First term} \leq \sum_{t=1}^{T} \mathbb{E}\left[\underbrace{\frac{(1-p_{i,t})}{p_{i,t}}}_{\textit{Coefficient}} \cdot \underbrace{\mathbb{P}\left(i(t) = 1,\ E_i^{\mu}(t),\ E_i^{\theta}(t) \mid \mathcal{F}_{t-1}\right)}_{\text{Probability of playing the best arm in the "good" case}}\right]$$

Proving the Regret Bound: Analyzing the First Term II

$$\mathbb{E}[N_i(T)] = \boxed{\sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right)} + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$

$$\text{First term} \leq \sum_{t=1}^{T} \mathbb{E}\left[\underbrace{\frac{(1 - p_{i,t})}{p_{i,t}}}_{\textit{Coefficient}} \cdot \underbrace{\mathbb{P}\left(i(t) = 1,\ E_i^{\mu}(t),\ E_i^{\theta}(t) \mid \mathcal{F}_{t-1}\right)}_{\text{Probability of playing the best arm in the "good" case}}\right]$$

$$p_{i,t} \coloneqq \mathbb{P}(\theta_1(t) > y_i \mid \mathcal{F}_{t-1})$$

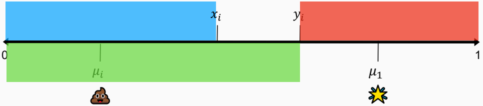


Proving the Regret Bound: Analyzing the First Term II

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$

$$\text{First term} \leq \sum_{t=1}^{T} \mathbb{E}\left[\underbrace{\frac{(1-p_{i,t})}{p_{i,t}}}_{\textit{Coefficient}} \cdot \underbrace{\mathbb{P}\left(i(t) = 1,\ E_i^{\mu}(t),\ E_i^{\theta}(t) \mid \mathcal{F}_{t-1}\right)}_{\text{Probability of playing the best arm in the "good" case}}\right]$$

$$p_{i,t} := \mathbb{P}(\theta_1(t) > y_i \mid \mathcal{F}_{t-1})$$



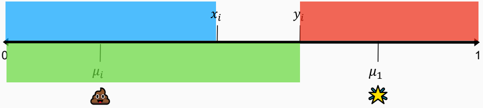
Coefficient decreases exponentially fast with samples of the optimal arm $N_1(t)$.

Proving the Regret Bound: Analyzing the First Term II

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$

$$\text{First term} \leq \sum_{t=1}^{T} \mathbb{E}\left[\underbrace{\frac{(1 - p_{i,t})}{p_{i,t}}}_{\textit{Coefficient}} \cdot \underbrace{\mathbb{P}\left(i(t) = 1,\ E_i^{\mu}(t),\ E_i^{\theta}(t) \mid \mathcal{F}_{t-1}\right)}_{\text{Probability of playing the best arm in the "good" case}}\right]$$

$$p_{i,t} := \mathbb{P}(\theta_1(t) > y_i \mid \mathcal{F}_{t-1})$$



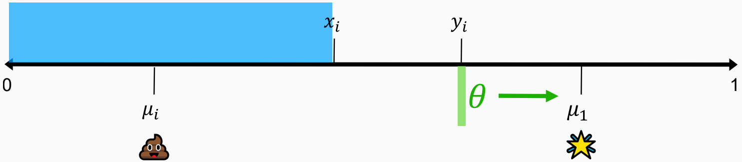
Coefficient decreases exponentially fast with samples of the optimal arm $N_1(t)$.

The term $\sum_{t=1}^{T} \mathbb{P}(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t))$ contributes a constant $O(1)$.

A primer on *big-oh* notation $O()$

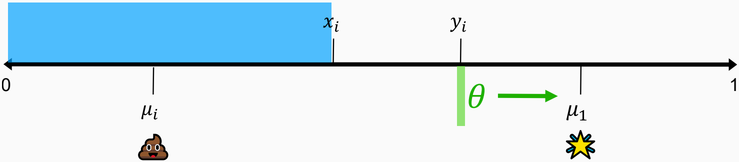
Proving the Regret Bound : Analyzing the Second Term I

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$



Proving the Regret Bound : Analyzing the Second Term I

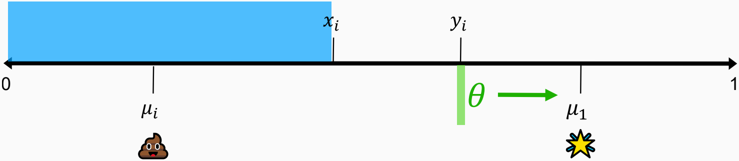
$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \boxed{E_i^{\mu}(t)}, \boxed{E_i^{\theta}(t)}\right) + \boxed{\sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \boxed{E_i^{\mu}(t)}, \boxed{\overline{E_i^{\theta}(t)}}\right)} + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \overline{E_i^{\mu}(t)}\right)$



Proof sketch.

Proving the Regret Bound : Analyzing the Second Term I

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$

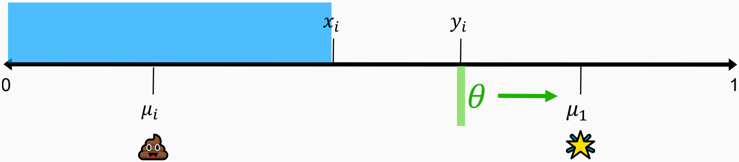


Proof sketch.

- Given that $E_i^{\mu}(t)$ holds, i.e. $\hat{\mu}_i \leq x_i$, the algorithm can only sample $\theta_i > y_i$ before the posterior concentrates around its mean.

Proving the Regret Bound: Analyzing the Second Term I

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$



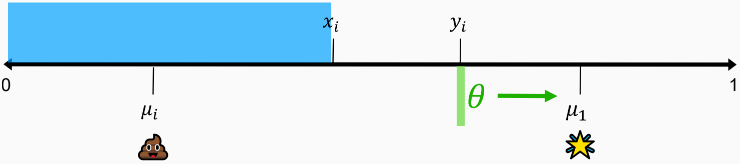
Proof sketch.

- Given that $E_i^{\mu}(t)$ holds, i.e. $\hat{\mu}_i \leq x_i$, the algorithm can only sample $\theta_i > y_i$ before the posterior concentrates around its mean.
- Posterior is well-concentrated around its mean when $N_i(t) \geq \frac{\log T}{d(x_i, y_i)}$,

$$d(x_i, y_i) := x_i \log \frac{x_i}{y_i} + (1 - x_i) \log \frac{1 - x_i}{1 - y_i}$$

Proving the Regret Bound : Analyzing the Second Term I

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$



Proof sketch.

- Given that $E_i^{\mu}(t)$ holds, i.e. $\hat{\mu}_i \leq x_i$, the algorithm can only sample $\theta_i > y_i$ before the posterior concentrates around its mean.
- Posterior is well-concentrated around its mean when $N_i(t) \geq \frac{\log T}{d(x_i, y_i)}$,

$$d(x_i, y_i) := x_i \log \frac{x_i}{y_i} + (1 - x_i) \log \frac{1 - x_i}{1 - y_i}$$

- After that, $\mathbb{P}(\theta_i > y_i)$ i.e. $\mathbb{P}\left(\overline{E_i^{\theta}(t)}\right) \leq \frac{1}{T}$.

Proving the Regret Bound: Analyzing the Second Term II

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$

Proving the Regret Bound: Analyzing the Second Term II

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$

- After $N_i(t) > \frac{\log T}{d(x_i, y_i)}$, $\mathbb{P}\left(E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right) \leq \frac{1}{T}$.

Proving the Regret Bound: Analyzing the Second Term II

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \overline{E_i^{\mu}(t)}\right)$$

- After $N_i(t) > \frac{\log T}{d(x_i, y_i)}$, $\mathbb{P}\left(E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right) \leq \frac{1}{T}$.
- Note that $\mathbb{P}(\text{event}) = \mathbb{E}[\mathbb{I}(\text{event})]$.

Proving the Regret Bound: Analyzing the Second Term II

$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \overline{E_i^{\mu}(t)}\right)$

- After $N_i(t) > \frac{\log T}{d(x_i, y_i)}$, $\mathbb{P}\left(E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right) \leq \frac{1}{T}$.
- Note that $\mathbb{P}(\text{event}) = \mathbb{E}[\mathbb{I}(\text{event})]$.

$$\begin{aligned}\text{Second term} \leq\ & \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left(i(t) = i, N_i(t) \leq \frac{\log T}{d(x_i, y_i)}\right)\right] \\ & + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}, N_i(t) > \frac{\log T}{d(x_i, y_i)}\right)\end{aligned}$$

Proving the Regret Bound: Analyzing the Second Term II

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$

- After $N_i(t) > \frac{\log T}{d(x_i, y_i)}$, $\mathbb{P}\left(E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right) \leq \frac{1}{T}$.
- Note that $\mathbb{P}(\text{event}) = \mathbb{E}[\mathbb{I}(\text{event})]$.

$$\begin{aligned}
\text{Second term} &\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left(i(t) = i, N_i(t) \leq \frac{\log T}{d(x_i, y_i)}\right)\right] \\
&\quad + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}, N_i(t) > \frac{\log T}{d(x_i, y_i)}\right) \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left(i(t) = i, N_i(t) \leq \frac{\log T}{d(x_i, y_i)}\right)\right] + \sum_{t=1}^{T} \frac{1}{T}
\end{aligned}$$

Proving the Regret Bound: Analyzing the Second Term II

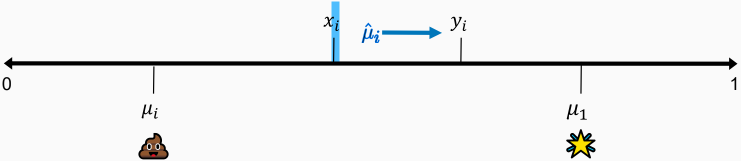
$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$

- After $N_i(t) > \frac{\log T}{d(x_i, y_i)}$, $\mathbb{P}\left(E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right) \leq \frac{1}{T}$.
- Note that $\mathbb{P}(\text{event}) = \mathbb{E}[\mathbb{I}(\text{event})]$.

$$\begin{aligned}
\text{Second term} &\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left(i(t) = i, N_i(t) \leq \frac{\log T}{d(x_i, y_i)}\right)\right] \\
&\quad + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t), \overline{E_i^{\theta}(t)}, N_i(t) > \frac{\log T}{d(x_i, y_i)}\right) \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{I}\left(i(t) = i, N_i(t) \leq \frac{\log T}{d(x_i, y_i)}\right)\right] + \sum_{t=1}^{T} \frac{1}{T} \\
&\leq \frac{\log T}{d(x_i, y_i)} + 1
\end{aligned}$$

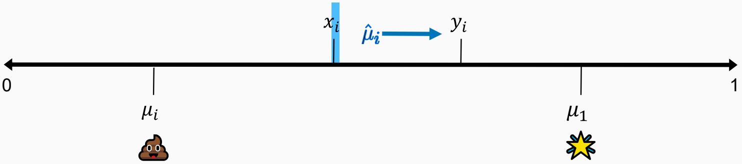
Proving the Regret Bound : Analyzing the Third Term

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$



Proving the Regret Bound: Analyzing the Third Term

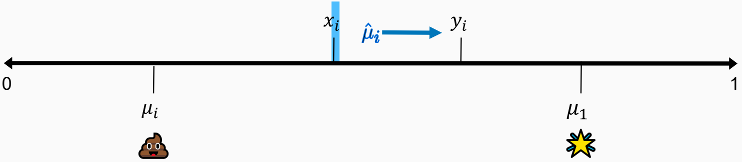
$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$



- We want to know the probability of the empirical mean deviating far from its true mean.

Proving the Regret Bound: Analyzing the Third Term

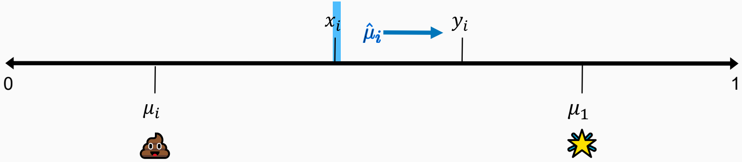
$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$



- We want to know the probability of the empirical mean deviating far from its true mean.
- Recall from last lecture Chernoff-Hoeffding bound provides an upper bound on this probability.

Proving the Regret Bound : Analyzing the Third Term

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \overline{E_i^{\mu}(t)}\right)$$



- We want to know the probability of the empirical mean deviating far from its true mean.
- Recall from last lecture Chernoff-Hoeffding bound provides an upper bound on this probability.

$$\sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \overline{E_i^{\mu}(t)}\right) \le \frac{1}{d(x_i, \mu_i)} + 1$$

Proving the Regret Bound: Putting Everything Together

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$

Proving the Regret Bound: Putting Everything Together

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$

$$\mathbb{E}[N_i(T)] \leq O(1) + \dots$$

Proving the Regret Bound: Putting Everything Together

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \boxed{\sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right)} + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$

$$\mathbb{E}[N_i(T)] \leq O(1) + \boxed{\frac{\log T}{d(x_i, y_i)} + 1} + \dots$$

Proving the Regret Bound : Putting Everything Together

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \boxed{E_i^{\mu}(t)}, \boxed{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \boxed{E_i^{\mu}(t)}, \overline{E_i^{\theta}(t)}\right) + \boxed{\sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \overline{E_i^{\mu}(t)}\right)}$$

$$\mathbb{E}[N_i(T)] \leq O(1) + \frac{\log T}{d(x_i, y_i)} + 1 + \boxed{\frac{1}{d(x_i, \mu_i)} + 1}$$

Proving the Regret Bound : Putting Everything Together

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ E_i^{\mu}(t),\ \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i,\ \overline{E_i^{\mu}(t)}\right)$$

$$\mathbb{E}[N_i(T)] \leq O(1) + \frac{\log T}{d(x_i, y_i)} + 1 + \frac{1}{d(x_i, \mu_i)} + 1$$

Proving the Regret Bound: Putting Everything Together

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \; E_i^{\mu}(t), \; E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \; E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \overline{E_i^{\mu}(t)}\right)$$

$$\mathbb{E}[N_i(T)] \leq O(1) + \frac{\log T}{d(x_i, y_i)} + 1 + \frac{1}{d(x_i, \mu_i)} + 1$$

- Time to set the values of x_i and y_i.

Proving the Regret Bound: Putting Everything Together

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \overline{E_i^{\mu}(t)}\right)$$

$$\mathbb{E}[N_i(T)] \leq O(1) + \frac{\log T}{d(x_i, y_i)} + 1 + \frac{1}{d(x_i, \mu_i)} + 1$$

- Time to set the values of x_i and y_i.
- Set x_i and y_i such that for some $\epsilon = [0, 1]$, $d(x_i, \mu_1) = \frac{d(\mu_i, \mu_1)}{1+\epsilon}$ and $d(x_i, y_i) = \frac{d(\mu_i, \mu_1)}{(1+\epsilon)^2}$

Proving the Regret Bound : Putting Everything Together

$$\mathbb{E}[N_i(T)] = \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), E_i^{\theta}(t)\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, E_i^{\mu}(t), \overline{E_i^{\theta}(t)}\right) + \sum_{t=1}^{T} \mathbb{P}\left(i(t) = i, \overline{E_i^{\mu}(t)}\right)$$

$$\mathbb{E}[N_i(T)] \leq O(1) + \frac{\log T}{d(x_i, y_i)} + 1 + \frac{1}{d(x_i, \mu_i)} + 1$$

- Time to set the values of x_i and y_i.
- Set x_i and y_i such that for some $\epsilon = [0, 1]$,
  $d(x_i, \mu_1) = \frac{d(\mu_i, \mu_1)}{1+\epsilon}$ and $d(x_i, y_i) = \frac{d(\mu_i, \mu_1)}{(1+\epsilon)^2}$

$$\mathbb{E}[N_i(T)] \leq (1+\epsilon)^2 \frac{\log T}{d(\mu_i, \mu_1)} + O\left(\frac{K}{\epsilon^2}\right)$$

Proving the Regret Bound: Final Step

Expected cumulative regret after T time steps is

$$\begin{aligned} \mathfrak{R}(T) &= \sum_i \Delta_i \, \mathbb{E}[N_i(T)] \\ &\leq (1+\epsilon)^2 \sum_i \frac{\log T}{d(\mu_i, \mu_1)} \Delta_i + O\left(\frac{K}{\epsilon^2}\right) \quad \square \end{aligned}$$

Distribution-free Regret Bound for Thompson Sampling

Theorem (Theorem 2 from Agrawal and Goyal [2013])

After T time steps, the expected cumulative regret of Thompson sampling using Beta priors is

$$Regret = \mathfrak{R}(T) \leq O(\sqrt{KT \log(T)})$$

Summary

- Solving Bandits using a Bayesian Perspective.

Summary

- Solving Bandits using a Bayesian Perspective.
- Thompson Sampling and its Regret Bound.

Summary

- Solving Bandits using a Bayesian Perspective.
- Thompson Sampling and its Regret Bound.
- Proof for the Regret Bound.

References

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, pages 99–107, 2013. URL http://proceedings.mlr.press/v31/agrawal13a.pdf.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2–3):235–256, may 2002. ISSN 0885-6125. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

Extra Material

- For more insights into Thompson Sampling, watch this video (till minute 32).
- Some resources on frequentist and Bayesian perspective : Stanford Encyclopedia of Philosophy articles - Interpretations of Probability by Alan Hájek, and Philosophy of Statistics by Jan-Willem Romeijn, a StackExchange question.
- For the purpose of producing useful and self-consistent results, any frequentist interpretation can generally be given a Bayesian interpretation, and vice versa.

Next lecture

- Non-stationary Stochastic Bandits.
- Adversarial Bandits.
- Dueling Bandits (and a lower bound).
- Contextual Bandits.

Main Lemma

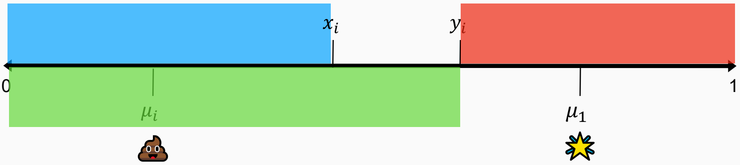
- Conditioned on any history, $\mathbb{P}$(playing any suboptimal arm at t) $\leq$ linear function of $\mathbb{P}$(playing the optimal arm at t).

Lemma (Lemma 1 from Agrawal and Goyal [2013])

For all $t = 1, \ldots, T$ and all suboptimal arms i i.e. $i \neq 1$,

$$\mathbb{P}\left(i(t) = i, \boxed{E_i^{\mu}(t)}, \boxed{E_i^{\theta}(t)} \mid \mathcal{F}_{t-1}\right) \leq \frac{(1 - p_{i,t})}{p_{i,t}} \mathbb{P}\left(i(t) = 1, \boxed{E_i^{\mu}(t)}, \boxed{E_i^{\theta}(t)} \mid \mathcal{F}_{t-1}\right)$$

where $p_{i,t} := \mathbb{P}(\boxed{\theta_1(t) > y_i} \mid \mathcal{F}_{t-1})$



Proving the Main Lemma I

Lemma (Main Lemma)

For all $t = 1, \ldots, T$ and all suboptimal arms i i.e. $i \neq 1$,

$$\mathbb{P}\left(i(t) = i, \boxed{E_i^\mu(t)}, \boxed{E_i^\theta(t)} \mid \mathcal{F}_{t-1}\right)$$
$$\leq \frac{(1 - p_{i,t})}{p_{i,t}} \mathbb{P}\left(i(t) = 1, \boxed{E_i^\mu(t)}, \boxed{E_i^\theta(t)} \mid \mathcal{F}_{t-1}\right)$$

Proving the Main Lemma I

Lemma (Main Lemma)

For all $t = 1, \ldots, T$ and all suboptimal arms i i.e. $i \neq 1$,

$$\mathbb{P}\left(i(t) = i, \boxed{E_i^{\mu}(t)}, \boxed{E_i^{\theta}(t)} \mid \mathcal{F}_{t-1}\right) \leq \frac{(1 - p_{i,t})}{p_{i,t}} \mathbb{P}\left(i(t) = 1, \boxed{E_i^{\mu}(t)}, \boxed{E_i^{\theta}(t)} \mid \mathcal{F}_{t-1}\right)$$

Proof.

- History till time $t - 1$ i.e. $\mathcal{F}_{t-1}$ determines the value of $\boxed{E_i^{\mu}(t)}$.

Proving the Main Lemma I

Lemma (Main Lemma)

For all $t = 1, \ldots, T$ and all suboptimal arms i i.e. $i \neq 1$,

$$\mathbb{P}\left(i(t) = i, \boxed{E_i^{\mu}(t)}, \boxed{E_i^{\theta}(t)} \mid \mathcal{F}_{t-1}\right) \leq \frac{(1 - p_{i,t})}{p_{i,t}} \mathbb{P}\left(i(t) = 1, \boxed{E_i^{\mu}(t)}, \boxed{E_i^{\theta}(t)} \mid \mathcal{F}_{t-1}\right)$$

Proof.

- History till time $t - 1$ i.e. $\mathcal{F}_{t-1}$ determines the value of $\boxed{E_i^{\mu}(t)}$.
- If $\mathcal{F}_{t-1}$ is such that $\boxed{E_i^{\mu}(t)}$ is false, then LHS is 0 and the lemma is trivially true as the RHS will also be 0.

Proving the Main Lemma I

Lemma (Main Lemma)

For all $t = 1, \ldots, T$ and all suboptimal arms i i.e. $i \neq 1$,

$$\mathbb{P}\left(i(t) = i, \boxed{E_i^{\mu}(t)}, \boxed{E_i^{\theta}(t)} \mid \mathcal{F}_{t-1}\right) \leq \frac{(1 - p_{i,t})}{p_{i,t}} \mathbb{P}\left(i(t) = 1, \boxed{E_i^{\mu}(t)}, \boxed{E_i^{\theta}(t)} \mid \mathcal{F}_{t-1}\right)$$

Proof.

- History till time $t-1$ i.e. $\mathcal{F}_{t-1}$ determines the value of $\boxed{E_i^{\mu}(t)}$.
- If $\mathcal{F}_{t-1}$ is such that $\boxed{E_i^{\mu}(t)}$ is false, then LHS is 0 and the lemma is trivially true as the RHS will also be 0.
- So we try to prove the lemma when $\mathcal{F}_{t-1}$ is such that $\boxed{E_i^{\mu}(t)}$ is true,

Proving the Main Lemma I

Lemma (Main Lemma)

For all $t = 1, \ldots, T$ and all suboptimal arms i i.e. $i \neq 1$,

$$\mathbb{P}\left(i(t) = i, \boxed{E_i^{\mu}(t)}, \boxed{E_i^{\theta}(t)} \mid \mathcal{F}_{t-1}\right) \leq \frac{(1 - p_{i,t})}{p_{i,t}} \mathbb{P}\left(i(t) = 1, \boxed{E_i^{\mu}(t)}, \boxed{E_i^{\theta}(t)} \mid \mathcal{F}_{t-1}\right)$$

Proof.

- History till time $t - 1$ i.e. $\mathcal{F}_{t-1}$ determines the value of $\boxed{E_i^{\mu}(t)}$.
- If $\mathcal{F}_{t-1}$ is such that $\boxed{E_i^{\mu}(t)}$ is false, then LHS is 0 and the lemma is trivially true as the RHS will also be 0.
- So we try to prove the lemma when $\mathcal{F}_{t-1}$ is such that $\boxed{E_i^{\mu}(t)}$ is true, i.e. prove that $\mathbb{P}\left(i(t) = i \mid \boxed{E_i^{\theta}(t)}, \mathcal{F}_{t-1}\right) \leq \frac{(1-p_{i,t})}{p_{i,t}} \mathbb{P}\left(i(t) = 1 \mid \boxed{E_i^{\theta}(t)}, \mathcal{F}_{t-1}\right)$.

Proving the Main Lemma II

To prove: $\mathbb{P}\left(i(t) = i \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right) \leq \frac{(1-p_{i,t})}{p_{i,t}} \mathbb{P}\left(i(t) = 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$.

- $\boxed{E_i^\theta(t)}$ is $\theta_i(t) \leq y_i$.

Proving the Main Lemma II

To prove: $\mathbb{P}\left(i(t)=i \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right) \leq \frac{(1-p_{i,t})}{p_{i,t}} \mathbb{P}\left(i(t)=1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$.

- $\boxed{E_i^\theta(t)}$ is $\theta_i(t) \leq y_i$.
- Selection rule of TS : select arm $i(t) = \arg\max_j \theta_j(t)$.

Proving the Main Lemma II

To prove: $\mathbb{P}\left(i(t) = i \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right) \leq \frac{(1-p_{i,t})}{p_{i,t}} \mathbb{P}\left(i(t) = 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$.

- $\boxed{E_i^\theta(t)}$ is $\theta_i(t) \leq y_i$.
- Selection rule of TS : select arm $i(t) = \arg\max_j \theta_j(t)$.
- So given $\boxed{E_i^\theta(t)}$, $i(t) = i$ only if $\theta_b(t) \leq y_i, \forall$ arms b

Proving the Main Lemma II

To prove: $\mathbb{P}\left(i(t) = i \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right) \leq \frac{(1-p_{i,t})}{p_{i,t}} \mathbb{P}\left(i(t) = 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$.

- $\boxed{E_i^\theta(t)}$ is $\theta_i(t) \leq y_i$.
- Selection rule of TS : select arm $i(t) = \arg\max_j \theta_j(t)$.
- So given $\boxed{E_i^\theta(t)}$, $i(t) = i$ only if $\theta_b(t) \leq y_i, \forall$ arms b

$$\mathbb{P}\left(i(t) = i \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$$

Proving the Main Lemma II

To prove: $\mathbb{P}\left(i(t) = i \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right) \leq \frac{(1-p_{i,t})}{p_{i,t}} \mathbb{P}\left(i(t) = 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$.

- $\boxed{E_i^\theta(t)}$ is $\theta_i(t) \leq y_i$.
- Selection rule of TS : select arm $i(t) = \arg\max_j \theta_j(t)$.
- So given $\boxed{E_i^\theta(t)}$, $i(t) = i$ only if $\theta_b(t) \leq y_i, \forall$ arms b

$$\mathbb{P}\left(i(t) = i \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right) \leq \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$$

Proving the Main Lemma II

To prove: $\mathbb{P}\left(i(t) = i \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right) \leq \frac{(1-p_{i,t})}{p_{i,t}} \mathbb{P}\left(i(t) = 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$.

- $\boxed{E_i^\theta(t)}$ is $\theta_i(t) \leq y_i$.
- Selection rule of TS : select arm $i(t) = \arg\max_j \theta_j(t)$.
- So given $\boxed{E_i^\theta(t)}$, $i(t) = i$ only if $\theta_b(t) \leq y_i, \forall$ arms b

$$
\begin{aligned}
\mathbb{P}\left(i(t) = i \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right) &\leq \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right) \\
&= \mathbb{P}\left(\theta_1(t) \leq y_i \mid \mathcal{F}_{t-1}\right) \\
&\quad \cdot \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \neq 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)
\end{aligned}
$$

Proving the Main Lemma II

To prove: $\mathbb{P}\left(i(t) = i \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right) \leq \frac{(1-p_{i,t})}{p_{i,t}} \mathbb{P}\left(i(t) = 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$.

- $\boxed{E_i^\theta(t)}$ is $\theta_i(t) \leq y_i$.
- Selection rule of TS : select arm $i(t) = \arg\max_j \theta_j(t)$.
- So given $\boxed{E_i^\theta(t)}$, $i(t) = i$ only if $\theta_b(t) \leq y_i, \forall$ arms b

$$
\begin{aligned}
\mathbb{P}\left(i(t) = i \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right) &\leq \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right) \\
&= \mathbb{P}\left(\theta_1(t) \leq y_i \mid \mathcal{F}_{t-1}\right) \\
&\quad \cdot \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \neq 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right) \\
&= (1 - p_{i,t}) \cdot \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \neq 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)
\end{aligned}
$$

Using $p_{i,t} := \mathbb{P}(\theta_1(t) > y_i \mid \mathcal{F}_{t-1})$

Proving the Main Lemma III

$$
\begin{aligned}
&\mathbb{P}\left(i(t) = i \mid \boxed{E_i^{\theta}(t)}, \mathcal{F}_{t-1}\right) \\
&\leq \frac{(1 - p_{i,t})}{p_{i,t}} \cdot p_{i,t} \cdot \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \neq 1 \mid \boxed{E_i^{\theta}(t)}, \mathcal{F}_{t-1}\right)
\end{aligned}
$$

Proving the Main Lemma III

$$
\begin{aligned}
&\mathbb{P}\left(i(t) = i \mid \boxed{E_i^{\theta}(t)}, \mathcal{F}_{t-1}\right) \\
&\leq \frac{(1 - p_{i,t})}{p_{i,t}} \cdot p_{i,t} \cdot \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \neq 1 \mid \boxed{E_i^{\theta}(t)}, \mathcal{F}_{t-1}\right)
\end{aligned}
$$

$$
p_{i,t} \cdot \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \neq 1 \mid \boxed{E_i^{\theta}(t)}, \mathcal{F}_{t-1}\right)
$$

Proving the Main Lemma III

$$\mathbb{P}\left(i(t) = i \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$$
$$\leq \frac{(1 - p_{i,t})}{p_{i,t}} \cdot p_{i,t} \cdot \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \neq 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$$

$$p_{i,t} \cdot \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \neq 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$$
$$\leq \mathbb{P}(\theta_1(t) > y_i \mid \mathcal{F}_{t-1}) \cdot \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \neq 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$$

Proving the Main Lemma III

$$\mathbb{P}\left(i(t) = i \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$$
$$\leq \frac{(1 - p_{i,t})}{p_{i,t}} \cdot p_{i,t} \cdot \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \neq 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$$

$$p_{i,t} \cdot \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \neq 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$$
$$\leq \mathbb{P}(\theta_1(t) > y_i \mid \mathcal{F}_{t-1}) \cdot \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \neq 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$$
$$= \mathbb{P}\left(\theta_b(t) \leq y_i < \theta_1(t), \forall b \neq 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$$

Proving the Main Lemma III

$$\mathbb{P}\left(i(t) = i \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$$
$$\leq \frac{(1-p_{i,t})}{p_{i,t}} \cdot p_{i,t} \cdot \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \neq 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$$

$$p_{i,t} \cdot \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \neq 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$$
$$\leq \mathbb{P}(\theta_1(t) > y_i \mid \mathcal{F}_{t-1}) \cdot \mathbb{P}\left(\theta_b(t) \leq y_i, \forall b \neq 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$$
$$= \mathbb{P}\left(\theta_b(t) \leq y_i < \theta_1(t), \forall b \neq 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right)$$
$$\leq \mathbb{P}\left(i(t) = 1 \mid \boxed{E_i^\theta(t)}, \mathcal{F}_{t-1}\right) \quad \square$$