CS 747, Autumn 2022: Lecture 5

Shivaram Kalyanakrishnan

Department of Computer Science and Engineering
Indian Institute of Technology Bombay

Autumn 2022

Multi-armed Bandits

1. Understanding Thompson Sampling

2. Other bandit problems

Multi-armed Bandits

1. Understanding Thompson Sampling

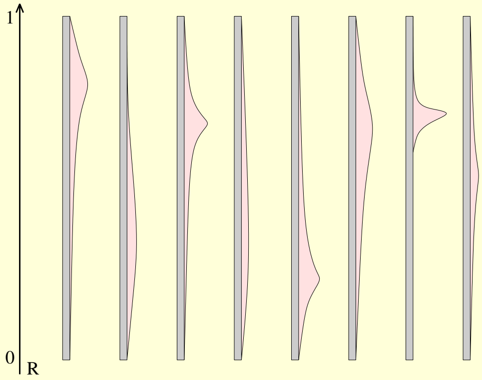
2. Other bandit problems

Thompson Sampling (Thompson, 1933)

- At time t, arm *a* has s_a^t successes (1's) and f_a^t failures (0's).

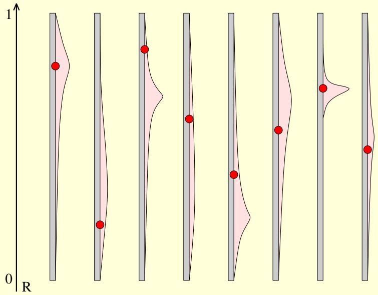
Thompson Sampling (Thompson, 1933)

- At time t, arm a has s_a^t successes (1's) and f_a^t failures (0's).
- $Beta(s_a^t + 1, f_a^t + 1)$ represents a "belief" about p_a.



Thompson Sampling (Thompson, 1933)

- At time t, arm a has s_a^t successes (1's) and f_a^t failures (0's).
- $Beta(s_a^t + 1, f_a^t + 1)$ represents a "belief" about p_a.



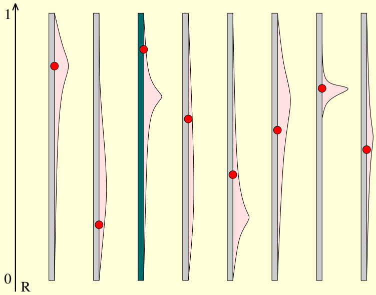
- Computational step: For every arm a, draw a sample

$$x_a^t \sim Beta(s_a^t + 1, f_a^t + 1).$$

- Sampling step: Pull an arm a for which x_a^t is maximum.

Thompson Sampling (Thompson, 1933)

- At time t, arm a has s_a^t successes (1's) and f_a^t failures (0's).
- $Beta(s_a^t + 1, f_a^t + 1)$ represents a "belief" about p_a.



- Computational step: For every arm a, draw a sample

$$x_a^t \sim Beta(s_a^t + 1, f_a^t + 1).$$

- Sampling step: Pull an arm a for which x_a^t is maximum.

Bayesian Inference

- Bayes' Rule of Probability for events A and B:

$$\mathbb{P}\{A|B\} = \frac{\mathbb{P}\{B|A\}\mathbb{P}\{A\}}{\mathbb{P}\{B\}}.$$

Bayesian Inference

- Bayes' Rule of Probability for events A and B:

$$\mathbb{P}\{A|B\} = \frac{\mathbb{P}\{B|A\}\mathbb{P}\{A\}}{\mathbb{P}\{B\}}.$$

- Application: there is an unknown world w from among possible worlds W, in which we live.
- We maintain a belief distribution over $w \in W$.

$$Belief_0(w) = \mathbb{P}\{w\}.$$

Bayesian Inference

- Bayes' Rule of Probability for events A and B:

$$\mathbb{P}\{A|B\} = \frac{\mathbb{P}\{B|A\}\mathbb{P}\{A\}}{\mathbb{P}\{B\}}.$$

- Application: there is an unknown world w from among possible worlds W, in which we live.
- We maintain a belief distribution over $w \in W$.

$$Belief_0(w) = \mathbb{P}\{w\}.$$

- The process by which each w produces evidence e is known.
- Evidence samples $e_1, e_2, \ldots, e_m$ are produced i.i.d. by the unknown world w.

Bayesian Inference

- Bayes' Rule of Probability for events A and B:

$$\mathbb{P}\{A|B\} = \frac{\mathbb{P}\{B|A\}\mathbb{P}\{A\}}{\mathbb{P}\{B\}}.$$

- Application: there is an unknown world w from among possible worlds W, in which we live.
- We maintain a belief distribution over $w \in W$.

$$Belief_0(w) = \mathbb{P}\{w\}.$$

- The process by which each w produces evidence e is known.
- Evidence samples $e_1, e_2, \ldots, e_m$ are produced i.i.d. by the unknown world w.
- How to continuously refine our belief distribution based on incoming evidence?

$$Belief_m(w) = \mathbb{P}\{w|e_1, e_2, \ldots, e_m\}$$

Bayesian Inference

$$Belief_{m+1}(w) = \mathbb{P}\{w | e_1, e_2, \ldots, e_{m+1}\}$$

Bayesian Inference

$$\begin{aligned}
Belief_{m+1}(w) &= \mathbb{P}\{w|e_1, e_2, \ldots, e_{m+1}\} \\
&= \frac{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}|w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}}
\end{aligned}$$

Bayesian Inference

$$
\begin{aligned}
Belief_{m+1}(w) &= \mathbb{P}\{w | e_1, e_2, \ldots, e_{m+1}\} \\
&= \frac{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1} | w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}} \\
&= \frac{\mathbb{P}\{e_1, e_2, \ldots, e_m | w\}\mathbb{P}\{e_{m+1} | w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}}
\end{aligned}
$$

Bayesian Inference

$$\begin{aligned}
Belief_{m+1}(w) &= \mathbb{P}\{w|e_1, e_2, \ldots, e_{m+1}\} \\
&= \frac{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}|w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}} \\
&= \frac{\mathbb{P}\{e_1, e_2, \ldots, e_m|w\}\mathbb{P}\{e_{m+1}|w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}} \\
&= \frac{\mathbb{P}\{e_1, e_2, \ldots, e_m, w\}\mathbb{P}\{e_{m+1}|w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}}
\end{aligned}$$

Bayesian Inference

$$
\begin{aligned}
\mathit{Belief}_{m+1}(w) &= \mathbb{P}\{w|e_1, e_2, \ldots, e_{m+1}\} \\
&= \frac{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}|w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}} \\
&= \frac{\mathbb{P}\{e_1, e_2, \ldots, e_m|w\}\mathbb{P}\{e_{m+1}|w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}} \\
&= \frac{\mathbb{P}\{e_1, e_2, \ldots, e_m, w\}\mathbb{P}\{e_{m+1}|w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}} \\
&= \frac{\mathbb{P}\{w|e_1, e_2, \ldots, e_m\}\mathbb{P}\{e_1, e_2, \ldots, e_m\}\mathbb{P}\{e_{m+1}|w\}}{\mathbb{P}\{e_1, e_2, \ldots, e_{m+1}\}}
\end{aligned}
$$

Bayesian Inference

$$
\begin{aligned}
Belief_{m+1}(w) &= \mathbb{P}\{w|e_1, e_2, \dots, e_{m+1}\} \\
&= \frac{\mathbb{P}\{e_1, e_2, \dots, e_{m+1}|w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \dots, e_{m+1}\}} \\
&= \frac{\mathbb{P}\{e_1, e_2, \dots, e_m|w\}\mathbb{P}\{e_{m+1}|w\}\mathbb{P}\{w\}}{\mathbb{P}\{e_1, e_2, \dots, e_{m+1}\}} \\
&= \frac{\mathbb{P}\{e_1, e_2, \dots, e_m, w\}\mathbb{P}\{e_{m+1}|w\}}{\mathbb{P}\{e_1, e_2, \dots, e_{m+1}\}} \\
&= \frac{\mathbb{P}\{w|e_1, e_2, \dots, e_m\}\mathbb{P}\{e_1, e_2, \dots, e_m\}\mathbb{P}\{e_{m+1}|w\}}{\mathbb{P}\{e_1, e_2, \dots, e_{m+1}\}} \\
&= \frac{Belief_m(w)\mathbb{P}\{e_{m+1}|w\}}{\sum_{w' \in W} Belief_m(w')\mathbb{P}\{e_{m+1}|w'\}}.
\end{aligned}
$$

Bayesian Inference in Thompson Sampling

- View each arm *a*'s mean p_a as world *w*, estimated from rewards (evidence).

Bayesian Inference in Thompson Sampling

- View each arm a's mean p_a as world w, estimated from rewards (evidence).
- $Belief_0$ over p_a is typically set to $Uniform(0, 1)$, but need not.

Bayesian Inference in Thompson Sampling

- View each arm a's mean p_a as world w, estimated from rewards (evidence).
- $Belief_0$ over p_a is typically set to $Uniform(0, 1)$, but need not.
- If e_{m+1} is a 1-reward, we must set for $x \in [0, 1]$

$$Belief_{m+1}(x) = \frac{Belief_m(x) \cdot x}{\int_{y=0}^{1} Belief_m(y) \cdot y}.$$

Bayesian Inference in Thompson Sampling

- View each arm a's mean p_a as world w, estimated from rewards (evidence).
- $Belief_0$ over p_a is typically set to $Uniform(0, 1)$, but need not.
- If e_{m+1} is a 1-reward, we must set for $x \in [0, 1]$

$$Belief_{m+1}(x) = \frac{Belief_m(x) \cdot x}{\int_{y=0}^{1} Belief_m(y) \cdot y}.$$

- If e_{m+1} is a 0-reward, we must set for $x \in [0, 1]$

$$Belief_{m+1}(x) = \frac{Belief_m(x) \cdot (1 - x)}{\int_{y=0}^{1} Belief_m(y) \cdot (1 - y)}.$$

Bayesian Inference in Thompson Sampling

- View each arm a's mean p_a as world w, estimated from rewards (evidence).
- $Belief_0$ over p_a is typically set to $Uniform(0, 1)$, but need not.
- If e_{m+1} is a 1-reward, we must set for $x \in [0, 1]$

$$Belief_{m+1}(x) = \frac{Belief_m(x) \cdot x}{\int_{y=0}^{1} Belief_m(y) \cdot y}.$$

- If e_{m+1} is a 0-reward, we must set for $x \in [0, 1]$

$$Belief_{m+1}(x) = \frac{Belief_m(x) \cdot (1 - x)}{\int_{y=0}^{1} Belief_m(y) \cdot (1 - y)}.$$

- We achieve exactly that by taking

$$Belief_m(x) = Beta_{s+1, f+1}(x) dx$$

when the first m pulls yield s 1's and f 0's!

Principle of Selecting Arm to Pull

- We have a belief distribution for each arm's mean.
- Together, these distributions represent a belief distribution over bandit instances.
- We sample a bandit instance I from the joint belief distribution, and
- We act optimally w.r.t. I.

Principle of Selecting Arm to Pull

- We have a belief distribution for each arm's mean.
- Together, these distributions represent a belief distribution over bandit instances.
- We sample a bandit instance I from the joint belief distribution, and
- We act optimally w.r.t. I.

- Alternative view: the probability with which we pick an arm is our belief that it is optimal. For example, if $A = \{1, 2\}$, the probability of pulling 1 is

$$\mathbb{P}\{x_1^t > x_2^t\} = \int_{x_1=0}^{1} \int_{x_2=0}^{x_1} Beta_{s_1^t+1, f_1^t+1,}(x_1) Beta_{s_2^t+1, f_2^t+1,}(x_2) dx_2 dx_1.$$

Multi-armed Bandits

1. Understanding Thompson Sampling
2. Other bandit problems

Other Bandit Problems

- In this course, we have covered
  - stochastic multi-armed bandits,
  - minimisation of expected cumulative regret.

  There are many other variations/formulations.

Other Bandit Problems

- In this course, we have covered
  - stochastic multi-armed bandits,
  - minimisation of expected cumulative regret.

  There are many other variations/formulations.

- Incorporating risk/variance in the objective.
  - Arm 1 gives rewards 0 and 100, each w.p. $1/2$.
  - Arm 2 gives rewards 48 and 50, each w.p. $1/2$.
  - Which arm would you prefer?

Other Bandit Problems

- In this course, we have covered
  - stochastic multi-armed bandits,
  - minimisation of expected cumulative regret.

  There are many other variations/formulations.

- Incorporating risk/variance in the objective.
  - Arm 1 gives rewards 0 and 100, each w.p. $1/2$.
  - Arm 2 gives rewards 48 and 50, each w.p. $1/2$.
  - Which arm would you prefer?

- What if the arms' (true) means vary over time?
  - Nonstationary setting, seen for example, in on-line ads.
  - Approach depends on nature of drift/change in rewards.
  - In practice, one might only trust most recent data from arms.
  - In practice, the set of arms can itself change over time!

Other Bandit Problems

- Pure exploration.
  - Separate "testing" and "live" phases.
  - In testing phase, rewards don't matter.
  - PAC formulation: W.p. at least $1 - \delta$, must return an ϵ-optimal arm, while incurring a small number of pulls.
  - Simple regret formulation: Given a budget of T pulls, must output an arm a such that p_a is large, or equivalently, simple regret $= p^\star - p_a$ is small).

Other Bandit Problems

- Pure exploration.
  - Separate "testing" and "live" phases.
  - In testing phase, rewards don't matter.
  - PAC formulation: W.p. at least $1 - \delta$, must return an ϵ-optimal arm, while incurring a small number of pulls.
  - Simple regret formulation: Given a budget of T pulls, must output an arm a such that p_a is large, or equivalently, simple regret $= p^\star - p_a$ is small).

- Limited number of feedback stages.
  - Suppose you are given budget T, but your algorithm can look at history only $s < T$ times?
  - UCB, Thompson Sampling, etc. are fully sequential ($s = T$).
  - How to manage with fewer "stages" s?

Other Bandit Problems

- What if the number of arms is large (thousands, millions)?
  - If arms can be described using features, mean reward is often treated as a (linear) function of these features.
  - Quantile-regret: look for "good", rather than "optimal" arms.

Other Bandit Problems

- What if the number of arms is large (thousands, millions)?
  - If arms can be described using features, mean reward is often treated as a (linear) function of these features.
  - Quantile-regret: look for “good”, rather than “optimal” arms.

- What if we are interacting with many bandits simultaneously?
  - Contextual bandits: If the bandits themselves can be described using features (a “context”), data from one can be used to generate estimates about others.

Other Bandit Problems

- What if the number of arms is large (thousands, millions)?
  - If arms can be described using features, mean reward is often treated as a (linear) function of these features.
  - Quantile-regret: look for “good”, rather than “optimal” arms.
- What if we are interacting with many bandits simultaneously?
  - Contextual bandits: If the bandits themselves can be described using features (a “context”), data from one can be used to generate estimates about others.
- What if the rewards do not come from a fixed random process?
  - Adversarial bandits make no assumption on the rewards.
  - Possible to show sub-linear regret when compared against playing a single arm for the entire run.
  - Necessary to use a randomised algorithm.

Multi-armed Bandits

- The exploration-exploitation dilemma
- Definitions: Bandit, Algorithm
- ϵ-greedy algorithms
- Evaluating algorithms: Regret
- Achieving sub-linear regret
- A lower bound on regret
- UCB, KL-UCB algorithms
- Thompson Sampling algorithm
- Concentration bounds
- Analysis of UCB
- Understanding Thompson Sampling
- Other bandit problems

Multi-armed Bandits

- The exploration-exploitation dilemma
- Definitions: Bandit, Algorithm
- ϵ-greedy algorithms
- Evaluating algorithms: Regret
- Achieving sub-linear regret
- A lower bound on regret
- UCB, KL-UCB algorithms
- Thompson Sampling algorithm
- Concentration bounds
- Analysis of UCB
- Understanding Thompson Sampling
- Other bandit problems

- **Next class:** Markov Decision Problems