Bayesian Bandits, Thompson Sampling

Henry Vu

Feb 9, 2024

Stochastic Bandits: Problem Settings

Given $A = \{1, 2, \ldots, K\}$ the set of action and (possibly) number of rounds $n \geq K$
for $t = 1, 2, \ldots, n$ **do**:

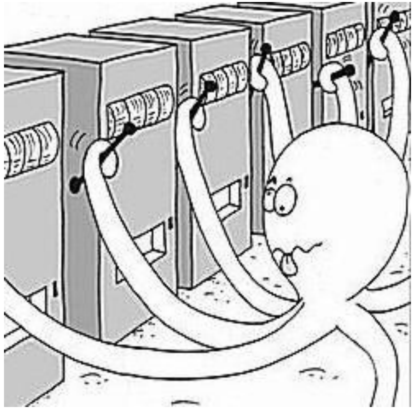
Algorithm pulls arm $A_t \in A$

A reward vector $(X_{t,1}, X_{t,2}, \ldots, X_{t,n})$ is generated

Algorithm observes reward x_{t, A_t}

end for

Goal: Minimizing the static regret, i.e. *gap* between the fixed optimal action and the algorithm's choices



Bayesian Learning

- Notation:
    - X_i: Reward of arm i, a random variable
    - $P(X_i; \theta_i)$: Unknown reward distribution, parameterized by θ_i
    - $R(i) = E[X_i]$: Mean reward of arm i

- Idea: Assume a prior of the reward distribution $P(\theta_i)$. After observing reward sequence $x_{1,i}, x_{2,i}, \ldots, x_{t,i}$, we can update the posterior distribution $P(\theta_i|x_{1,i}, \ldots, x_{t,i})$

- **Bayes' Theorem:**

$$P(\theta_i|x_{1,i}, \ldots, x_{t,i}) \propto P(x_{1,i}, \ldots, x_{t,i}|\theta_i) * P(\theta_i)$$

Bayesian Learning

- Posterior over θ_i, $P(\theta|x_{1,i}, x_{2,i}, \ldots, x_{t,i})$, allows us to estimate:
  - the distribution over the next reward $X_{t+1,i}$ for each arm i

$$P(X_{t+1,i}|x_{1,i}, \ldots, x_{t,i}) = \int_\theta P(X_{t+1,i}; \theta_i) P(\theta_i|x_{1,i}, \ldots, x_{t,i}) d\theta$$

  - Distribution over R(i) if the mean is a function of parameter θ_i

$$P(\theta_i|x_{1,i}, x_{2,i}, \ldots, x_{t,i}) = P(R(i)|x_{1,i}, x_{2,i}, \ldots, x_{t,i})$$

- Guide exploration by sampling from the posterior distribution.

Bayesian Bandits

- Bayesian bandits fall under the category of stochastic bandits.
- In the Bayesian setting, we assume a *prior* distribution for the reward distribution and update the *posterior* after each reward realization.
- Use *Bayesian learning* to select action based on the full distribution instead of a bound like in UCB: $P(R(i)|x_{1, i}, \ldots, x_{t,i})$.

Simple Case: Bernoulli Bandits

- The likelihood function $P(X_i; \theta_i)$ follows a Bernoulli distribution with mean $R(i) = \theta_i$. $x_i = 1$ with probability θ_i and $x_i = 0$ with probability $1 - \theta_i$.

- Reward of each arm $X_{t, i} \in [0, 1]$.

- The likelihood of observing reward x_i for a Bernoulli(θ) distribution is:

  $P(x_i|\theta) = \theta^{xi}(1-\theta)^{1-xi}$

Conjugate Prior:

- If the *posterior* distribution is in the same probability distribution family as the *prior* distribution, then they are called **conjugate distributions**.

- For Bernoulli likelihood, the *Beta* distribution is its **conjugate prior** since the *posterior* distribution is a Beta distribution.

$$\underbrace{P(\theta_i | x_{1,i}, \ldots, x_{t,i})}_{\text{Posterior: Beta}} \propto \underbrace{P(x_{1,i}, \ldots, x_{t,i} | \theta_i)}_{\text{Likelihood: Bernoulli}} * \underbrace{P(\theta_i)}_{\text{Prior: Beta}}$$

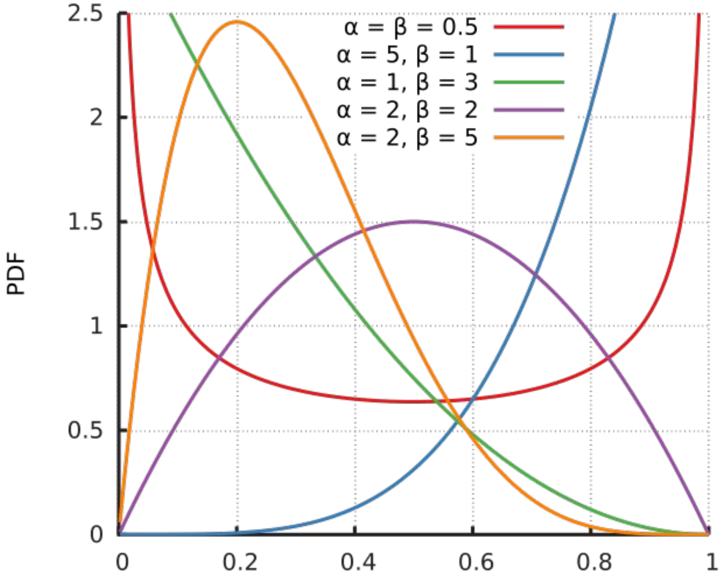
Beta distribution

- $P(\theta_i)$ follows a Beta distribution

  $Beta(\theta_i;\alpha_i,\beta_i) \sim \theta_i^{\alpha_i - 1}(1-\theta_i)^{\beta_i - 1}$

- α_i-1: # of 1's
- β_i-1: # of 0's

- $E[\theta_i] = \alpha_i/(\alpha_i+\beta_i)$



Thompson Sampling

- Given:
  - Set of parameters $\theta = (\theta_1, \ldots, \theta_k)$ and prior distribution $P(\theta_i)$
  - A likelihood function $P(X_i|\theta_i)$
  - Some past observations $x_{1,i}, \ldots, x_{t,i}$

- Idea:
  - Sample from the current posterior $P(\theta_i|x_{1, i}, \ldots, x_{t,i})$ for each arm i
  - Choose a maximizing arm from the sampled rewards
  - Update the posterior distribution of the chosen arm

    $P(\theta_i|x_{1, i}, \ldots, x_{t,i}) \propto P(x_{1, i}, \ldots, x_{t,i}|\theta_i)*P(\theta_i)$

Thompson Sampling for Bernoulli Bandits

Algorithm 5 Thompson Samplings

Input: horizon n, number of arms K, parameters α and β;
for $t = 1, \dots, n$ **do**
    for $i = 1, \dots, K$ **do**
        Sample $\hat{\theta}_i \sim \mathrm{Beta}(\alpha_i, \beta_i)$
    end for
    $A_t \leftarrow \arg\max_i \hat{\theta}_i$
    Take action A_t and observe x_{t,A_t}
    $(\alpha_{x_t}, \beta_{x_t}) \leftarrow (\alpha_{A_t} + x_{t,A_t}, \beta_{A_t} + 1 - x_{t,A_t})$
end for

Start with $\alpha, \beta = [1]_1^K$. So the arms are uniformly distributed

Comparison

Greedy Method

- Sample

$$\hat{\theta}_i \leftarrow \alpha_i/(\alpha_i + \beta_i)$$

- Action Selection

$$A_t \leftarrow \arg\max_i \hat{\theta}_i$$

⇒ No exploration, only consider the best so far

Thompson Sampling

- Sample

$$\hat{\theta}_i \sim \mathrm{Beta}(\alpha_i, \beta_i)$$

- Action Selection

$$A_t \leftarrow \arg\max_i \hat{\theta}_i$$

⇒ Some exploration

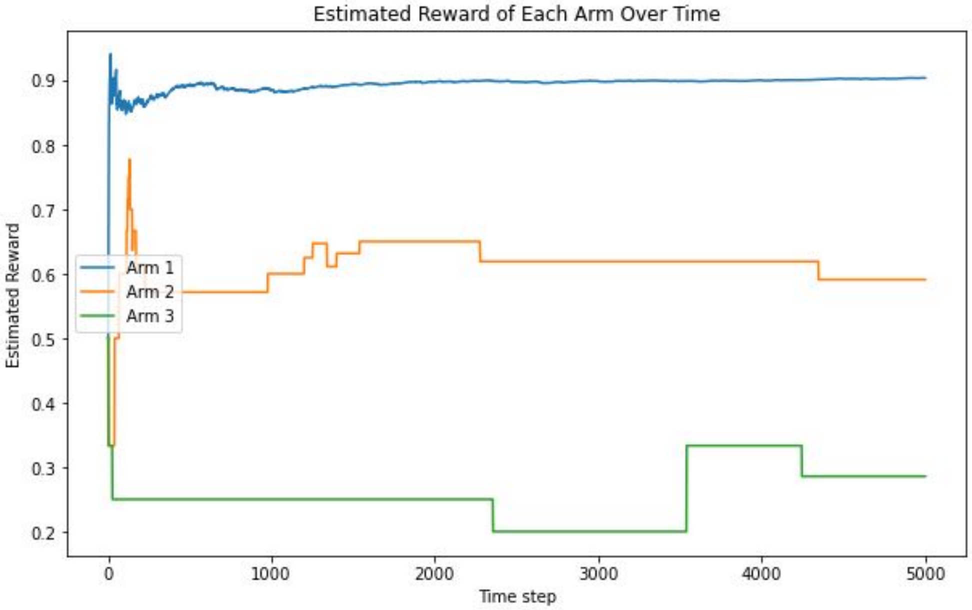
Bernoulli Thompson Sampling

- Posterior distribution update for arm i after observing reward $x_{t,i}$:

$$
\begin{aligned}
P(\theta_i|x_{t,i}) &\propto P(x_{t,i}|\theta_i)P(\theta_i) \\
&\propto \theta_i^{x_{t,i}}(1-\theta_i)^{1-x_{t,i}}\theta_i^{\alpha_i-1}(1-\theta_i)^{\beta_i-1} \\
&= \theta_i^{\alpha_i-1+x_{t,i}}(1-\theta_i)^{\beta_i-1+1-x_{t,i}} \\
&\propto \mathrm{Beta}(\alpha_i+x_{t,i}, \beta_i+1-x_{t,i})
\end{aligned}
$$

Simple Example

- 3 armed bandits with Bernoulli reward distribution as follows [0.9, 0.5, 0.2]



Regret Bound

- For the simple case of Bernoulli bandits established above, the regret of Thompson Sampling when the priors are initialized with a uniform distribution is:

$$\max_{\theta'} \mathbb{E}[\text{Regret}_n | \theta = \theta'] = O\left(\sqrt{Kn\log(n)}\right)$$

Thompson Sampling for Contextual Bandits with Linear Payoffs

Shipra Agrawal
Microsoft Research

Navin Goyal
Microsoft Research

February 4, 2014

General Thompson Sampling

Algorithm 3 Greedy($\mathcal{X}, p, q, r$)

for $t = 1, 2, \ldots$ **do**
  #estimate model:
  $\hat{\theta} \leftarrow \mathbb{E}_p[\theta]$

  #select and apply action:
  $x_t \leftarrow \text{argmax}_{x \in \mathcal{X}} \mathbb{E}_{q_{\hat{\theta}}}[r(y_t) | x_t = x]$
  Apply x_t and observe y_t

  #update distribution:
  $p \leftarrow \mathbb{P}_{p,q}(\theta \in \cdot | x_t, y_t)$
end for

Algorithm 4 Thompson($\mathcal{X}, p, q, r$)

for $t = 1, 2, \ldots$ **do**
  #sample model:
  Sample $\hat{\theta} \sim p$

  #select and apply action:
  $x_t \leftarrow \text{argmax}_{x \in \mathcal{X}} \mathbb{E}_{q_{\hat{\theta}}}[r(y_t) | x_t = x]$
  Apply x_t and observe y_t

  #update distribution:
  $p \leftarrow \mathbb{P}_{p,q}(\theta \in \cdot | x_t, y_t)$
end for

Next Time

- Gittins Index

- Restless Bandits

Thank you!