Markov Bandit Process & Gittins Index

Henry Vu

March 29, 2024

Outline

- Introduction
  - Motivating Example
- Markov Bandit Process, Objective Function
- Gittins Index Theorem

Example: Beta-Bernoulli Bandits

- An (α_i, β_i) prior corresponds to a success probability of $\alpha_i/(\alpha_i + \beta_i)$ in the current step, and the arm i becomes an (α_i+1, β_i) in the event of a success, and an (α_i, β_i+1) arm in the even of a failure.

- Notion of **discount factor** γ, i.e. "present value of tomorrow's reward."
  If the reward \$1 tomorrow, it is worth \$$\gamma$ to you today. If you are going to earn \$1 the day after tomorrow, it is worth \$$\gamma^2$ to you today.

- γ is usually set to 1 - 1/T when T, the time horizon, is known.
  e.g. $T = 10, \gamma = 0.9$
  $T = 10000, \gamma = 0.9999$

Gittins Index Theorem

- There exists a function g of three variables, $g(\alpha, \beta, \gamma)$, such that an optimum strategy for maximizing total expected discounted reward in the multi-armed bandit problem with Beta priors is to play the arm i with the largest value of $g(\alpha_i, \beta_i, \gamma)$.

- Function g is known as the **Gittins Index**. At each period, we just need to
  - Find the Gittins Index of arm i;
  - Play the arm with the highest Gittins Index.

Two-armed Bandits

- Suppose arm 1 has fixed success probability p ($0 < p < 1$), arm 2 has priors (α, β).

- Idea: If we can find p such that we are *indifferent* between play arm 1 and arm 2, then we can assign p as the Gittins Index $g(\alpha, \beta, \gamma)$ of arm 2.

- Let us define the value function, $V(p; \alpha, \beta, \gamma)$, to be the **maximum expected discounted reward** of any strategy that starts with arm 1 and 2 above.

Two-armed Bandits

- Suppose we play arm 1 at the first period. Then, E[total discounted reward] = $p + p\gamma + p\gamma^2 + \dots = p/(1-\gamma)$. So $V(p; \alpha, \beta, \gamma) \geq p/(1-\gamma)$.

- Thus, the indifference point between the two arms would be the p for which the value function $V(p; \alpha, \beta, \gamma)$ is exactly **equal** to $p/(1-\gamma)$.

- How to compute $V(p; \alpha, \beta, \gamma)$?

Computing Value Function

- If arm 1 is played in the first period, then arm 1 will always be played because p does not change. Therefore, the expected discounted reward will be $p/(1-\gamma)$.

- If arm 2 is played in the first period, we must add the value this period and the expected value of two possibilities next period: success and failure. We have $\alpha/(\alpha+\beta)$ probability of success and $\beta/(\alpha+\beta)$ probability of failure. We also need to multiply next period's reward by γ.

$$\frac{\alpha}{\alpha+\beta} + \gamma\left(\frac{\alpha}{\alpha+\beta}V(p;\alpha+1,\beta,\gamma) + \frac{\beta}{\alpha+\beta}V(p;\alpha,\beta+1,\theta)\right)$$

The Bellman Equation

- The maximization becomes:

$$V(p;\alpha,\beta,\gamma) = \max\left\{\frac{p}{1-\gamma}, \frac{\alpha}{\alpha+\beta} + \gamma\left(\frac{\alpha}{\alpha+\beta}V(p;\alpha+1,\beta,\gamma) + \frac{\beta}{\alpha+\beta}V(p;\alpha,\beta+1,\theta)\right)\right\}$$

- Now, to find the Gittins Index of an arm with priors (α, β), we solve for p such that V(p; α, β, γ) = p/(1−γ). The Gittins Index of that arm is g(α, β, γ) = p.

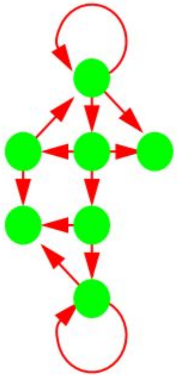
General Case

- Multiple arms.
- Generalization of states, actions and rewards.
- Generalization of the objective function.

$\Rightarrow$ Markov Bandit

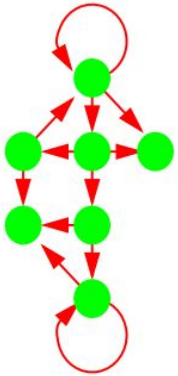
Markov Decision Process (MDP)

- A Markov Decision Process (MDP) model contains:
  - A set of possible world states S
  - A set of possible actions A
  - A real valued reward function **R(s, a)**
  - A description **T** of each action's effects in each state



- We assume the Markov Property: the effects of an action taken in a state depend only on that state and not on the prior history

MDP: Representing Actions and Solutions



- Deterministic Actions:
  T: S x A → S For each state and action, we specify a new stage.

- Stochastic Actions:
  T: S x A → Prob(S) For each state and action we specify a probability distribution over next states. Represents the distribution P(s'|s, a).

- Solutions: A policy π: S → A determines what action to take in each state.

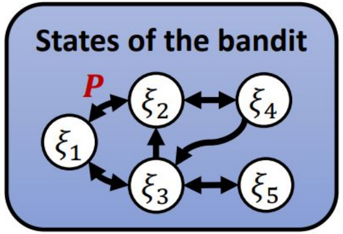
Markov Bandit Process

- An MDP on a countable state space, where $s(t) \in \{s_1, \ldots, s_K\}$ is the state of the bandit at discrete decision time $t \in \{0, 1, 2, \ldots\}$.

- Controls applied at decision time t:
  - $u(t) = 0$ **freezes** the process and gives no reward;
  - $u(t) = 1$ **continues** the process and gives instantaneous reward

$\gamma^t R(s(t))$

State Transitions are instantaneous with $P(\xi'|\xi)$ **when $u(t) = 1$.**



$\gamma \in (0, 1)$ is the discount factor

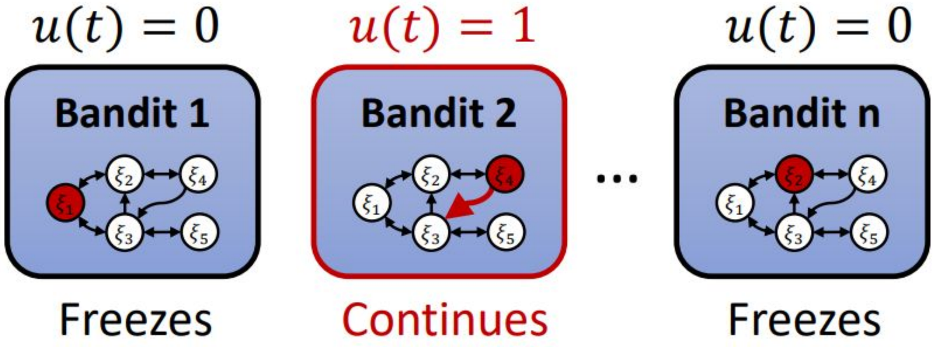
$r(.) > 0$ is the reward

Simple Family of Alternative Bandit Processes

- **n Markov Bandit Processes** with state space $S = S_1 \times S_2 \times \ldots \times S_n$.
- State of the selected bandit i_t at each decision t is $s_i(t)$.
- Control $u(t) = 1$ is applied to a single bandit i_t at each decision time t. Transition probability $P_{it}(s'|s_{it}(t))$
- Control $u(t) = 0$ is applied to all other bandits. These bandits remain in the **same state**.
- Reward obtained is $\gamma^t r_{it}(s_{it}(t))$.

Simple Family of Alternative Bandit Processes

- n Markov Bandits
- At time t, apply u(t) = 1 to bandit 2 and u(t) = 0 to all other bandits



Objective Function

- Maximize the expected discounted sum of rewards

$$J_\pi(\vec{s}) = \lim_{T \to \infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r_{i_t}(s_{i_t}(t)) \middle| \vec{s}(0) = \vec{s}\right]$$

- At time t, we know the states of each arm i (vector s), the transition probabilities, the discount factor and the reward function $r_i(.)$.

Example

- Consider 2 bandits, each evolving according to a **deterministic** state sequence
  - Bandit 1 : {10, 2, 8, 7, 6, 0, 0, …}
  - Bandit 2 : {5, 4, 3, 9, 1, 0, 0, …}

- The policy that maximizes $\lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r_{i_t}(s_{i_t}(t))\right]$ is:

  If γ = 0.1: $10\gamma^0 + 5\gamma^1 + 4\gamma^2 + 3\gamma^3 + 9\gamma^4 + 2\gamma^5 + 8\gamma^6 + \ldots$

  If γ = 0.9: $10\gamma^0 + 2\gamma^1 + 8\gamma^2 + 7\gamma^3 + 6\gamma^4 + 5\gamma^5 + 4\gamma^6 + \ldots$

Index Policy

- We are trying to maximize:

$$J_\pi(\vec{s}) = \lim_{T\to\infty} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r_{i_t}(s_{i_t}(t)) \middle| \vec{s}(0) = \vec{s}\right]$$

- **Index Theorem:** The **optimal policy** for this problem is an **Index policy**.

- **Index Policy:** There exists a function $G_i(s_i)$, computed for each bandit, such that at time step t, the optimal policy continues the bandit $i_t = \text{argmax}_i \{G_i(s_i)\}$. G_i is the index function of arm i; at time step t choose the arm with the **highest** index

Derivation of Index Function

- Consider a single arm i with a **playing charge** λ. At time t, if we haven't stopped playing, we can choose to continue and pay λ to receive reward $r_i(s_i(t))$.

- Optimal Stopping:

$$J(s_i) = \sup_{\tau>0} \mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t (r_i(s_i(t) - \lambda) \middle| s_i(0) = s_i\right]$$

- For every s_i, there is a λ such that there is a null reward for playing:

$$J(s_i) = \sup_{\tau>0} \mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t (r_i(s_i(t) - \lambda) \middle| s_i(0) = s_i\right] = 0$$

Derivation of Index Function

- $J(s_i) = \sup_{\tau>0} \mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t (r_i(s_i(t) - \lambda) \middle| s_i(0) = s_i\right] = 0$ is linear and decreasing

  on λ, and therefore has a single root λ which is the **Gittins Index**, $G_i(s_i)$, given by:

$$G_i(s_i) = \sup_{\tau>0} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t r_i(s_i(t)) \middle| s_i(0) = s_i\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t \middle| s_i(0) = s_i\right]}$$

- $G_i(s_i)$ is called the **fair charge** during state s_i.

- When charge $\lambda = G_i(s_i)$, we are indifferent between continuing and stopping.

Gittins Index

- Going back to the **n Markov Bandit Setting** with **no charge**, the Gittins Index of each arm i is:

$$G_i(s_i) = \sup_{\tau > 0} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t r_i(s_i(t)) \middle| s_i(0) = s_i\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t \middle| s_i(0) = s_i\right]}$$

  where τ is the stopping time

- Numerator is the **discounted reward** up to time τ.

- Denominator is the **discounted time** up to time τ.

Proof of Gittins Index

- Supposed that at time $t = 0$ we are in state s_i with a fair charge of $G_i(s_i)$.
- If we set $\lambda = G_i(s_i)$ and play bandit i **optimally**, then the expected payoff is 0.
- What if at stopping time τ, we reset the charge? i.e. set $\lambda = G_i(s_i')$

Proof of Gittins Index

- As the game continues, the charge is reset several times
- Let $\lambda_i(t)$ be the current charge.
- Since $G_i(s_i) = \sup_{\tau > 0} \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t r_i(s_i(t)) \middle| s_i(0) = s_i\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} \gamma^t \middle| s_i(0) = s_i\right]}$ is the supremum over time, $\lambda_i(t)$ non-increasing and is equal to the minimum $G_i(s_i)$ so far.
  $\lambda_i(t)$ is also called the **prevailing charge**.

Proof of Gittins Index

- Consider **n bandits**, each with a different initial state s_i. We set each initial charge $\lambda_i = G_i(s_i)$ for all arms i and update them as in the previous slide.

- Consider a **policy** $\boldsymbol{\pi}$* that selects the bandits with highest $\lambda_i(t)$ at time t:
  - $\boldsymbol{\pi}$* has null profit and incurs the highest sum of discounted charges (since the charges are non-increasing).
  - Since Reward - Charge = Profit = 0 ⇒ $\boldsymbol{\pi}$* incurs highest discounted expected reward.

- $\boldsymbol{\pi}$ is equivalent to choosing bandits with highest **Gittins Index** ⇒ G.I is optimal

Proof of Gittins Index

- By definition of λ_i, the expected profit is:

$$E_\pi\left[\sum_{t=0}^{\infty}\gamma^t\left(r_{i_t}(x_{i_t}(t)) - \lambda_{i_t}(x_{it}(t))\right)\middle| x(0)\right] \leq 0$$

- By definition of $\boldsymbol{\pi}$*,

$$E_\pi\left[\sum_{t=0}^{\infty}\gamma^t r_{i_t}(x_{i_t})\middle| x(0)\right] \leq E_\pi\left[\sum_{t=0}^{\infty}\gamma^t \lambda_{i_t}(x_{i_t})\middle| x(0)\right] \leq E_{\pi^*}\left[\sum_{t=0}^{\infty}\gamma^t \lambda_{i_t}(x_{i_t})\middle| x(0)\right]$$

- Equality at **Gittins Index**, i.e. LHS is maximized

Next Time

- Peter Whittle demonstrated that the index emerges as a Lagrange multiplier from a dynamic programming formulation of the problem called retirement process and conjectured that the same index would be a good heuristic in a more general setup named **Restless bandit**

Thank you!