Markovian Bandits[1]

Manyou Ma

manyoum@ece.ubc.ca
Machine Learning Reading Group (MLRG 2019)

Jul 31, 2019

Outline

Motivation

- Consider the modifications below:
  - **Bandit 1:** $\{10, 2, 9, 7, 6, 0, 0, 0, \ldots\}$
  - **Bandit 2:** $\{5, 4, 3, 9, 1, 0, 0, 0, \ldots\}$
- What is the policy that maximized $\lim_{T \to \infty} \mathbb{E}\left[\sum_{t=0}^{T-1} a^t r_{i_t}(\iota_{i_t})\right]$
- **a = 0.1:** "Future is not so important"
  - $10a^0 + 5a^1 + 4a^2 + 3a^3 + 9a^4 + 2a^5 + 8a^6 + \ldots$
- **a = 0.9:** "Future is (almost) as important as the present"
  - $10a^0 + 2a^1 + 8a^2 + 7a^3 + 6a^4 + 5a^5 + 4a^6 + \ldots$
- **a = 0.5:** "Future is somewhat important"
  - $10a^0 + 5a^1 + 2a^2 + 8a^3 + 7a^4 + 6a^5 + 4a^6 + \ldots$

Outline

Markov Decision Process (MDP)

- **State:** Defined on a state space Θ and σ-algebra $\mathcal{X}$ of subsets of Θ which includes every subset of consisting just one element of Θ
- **Action:** When the process is in state x, the set of control may be applied: $\Omega(x)$.
- **Probability:** $P(A \mid x, u)$ is the probability that the state y of the process at time $t+1$ belongs to $A(\in \mathcal{X})$, given at time t the process is in state x and control u is applied.
- **Rewards:** Application of control u at time t with the process in state x yields a reward $a^t R(x, u)$ $(0 < a < 1)$.

MDP Policies

Definition

Any rule, including randomized rules, which for all t specifies the control to be applied at time t as a function of t, the states at times $0, \ldots, t$, and controls applied at times $0, \ldots, t-1$.

Deterministic Policies

Policies that involve no randomization.

Stationary Policies

Policies that involve no explicit time-dependence.

Markov Policies

Policies for which the control chosen at time t is independent of the states and the controls applied at times $0, 1, \ldots, t-1$.

Optimal Policy of MDP

Blackwell's Theorem (1965)

If the control set $\Omega(x)$ is finite and the same for all x, then there is a deterministic stationary Markov policy for which, for any initial state, the total expected reward is the supremum of the total expected reward for the class of all policies.

Optimal Policy

A policy which achieves the supremum of total expected rewards.

Assumptions

It is assumed throughout the presentation that $\Omega(x)$ is finite for all x and the supremum of the total expected reward is finite. (Therefore we can restrict our attention to deterministic stationary policies only.)

Outline

Bandit Process

Definition

A *bandit process* is an MDP for which

- Set of controls: $\Omega(x) = \{0, 1\}$, where
  - **Control 0:** Freezes the process $P(x \mid x, 0) = 1$ and $R(x, 0) = 0, \forall\, x$.
  - **Control 1:** *Continuation* control. No restriction on the transition probabilities and rewards.

Process Time t

The number of times control 1 has been applied to a bandit process, where

- The state at process time t is denoted by $x(t)$.
- The reward between times t and $t + 1$ if control 1 is applied is applied at each stage is $a^t R(x(t), 1)$, abbreviated as $a^t R(t)$.

Standard Bandit Process

A bandit process for which, for some λ, $R(x, 1) = \lambda, \forall x$.

Bandit Process

Freezing Rule f

An arbitrary policy for a Bandit process. Given any freezing rule f,

- $f(t)$ is the number of times control 0 is applied, before the $(t+1)$st application of control 1.

Deterministic Policy

Deterministic stationary Markov policies divide the state space Θ into

- **Stopping Set:** Control 0 is applied
- **Continuation Set:** Control 1 is applied

Stopping Rule and Stopping Time

For some sequentially determined random variable τ, called *stopping time*, we have

- $f(t) = 0, \forall t < \tau$ and $f(\tau) = +\infty$.

Dynamic Allocation Index

Expected Total Reward

- **Freezing Rule** f: $R_f(D) = \mathbb{E}\sum_{t=0}^{\infty} a^{t+f(t)}R(t)$, $R(D) = \sup_f R_f(D)$
- **Stopping Rule** τ: $R_\tau(D) = \mathbb{E}\sum_{t=0}^{\tau-1} a^t R(t)$, $R(D) = \sup_\tau R_\tau(D)$

Expected Rewards per **Unit of Discounted Time**

- **Freezing Rule** f: $W_f(D) = \mathbb{E}\sum_{t=0}^{\infty} a^{t+f(t)}$, $\nu_f(D) = \frac{R_f(D)}{W_f(D)}$, and $\nu(D) = \sup_{f:f(0)=0} \nu_f(D)$
- **Stopping Rule** τ: $W_\tau(D) = \mathbb{E}\sum_{t=0}^{\tau-1} a^t$, $\nu_\tau(D) = \frac{R_\tau(D)}{W_\tau(D)}$, and $\nu(D) = \sup_{\tau>0} \nu_\tau(D)$

- All these quantities depend on initial state of the bandit process $x(0)$
- $\nu(D, x)$ is defined as the *Dynamic Allocation Index*.

Outline

Forward Induction Theorems

Superprocess $(\mathcal{M}, g)$

Given any Markov decision process $\mathcal{M}$, with a deterministic stationary Markov policy g, a bandit process may be defined by at each time t either apply

- Freeze control 0, or the control given by g.

Forwards Induction Policy

For the Markov decision process $\mathcal{M}$, whose state at time zero is x_0,

- Find a policy γ_1 and a stopping time σ_1, such that the discounted average reward per unit time of the superprocess $(\mathcal{M}, g)$ up to the stopping time τ is maximized over g and τ, by setting $(g, \tau) = (\gamma_1, \sigma_1)$. That is $\nu_{\gamma_1\sigma_1}(\mathcal{M}) = \nu(\mathcal{M}, x_0)$
- Let x_1 be the random state of the superprocess $(\mathcal{M}, \gamma_1)$ at time σ_1. Define the policy γ_2 and σ_2 such that $\nu_{\gamma_2\sigma_2}(\mathcal{M}, x_1) = \nu(\mathcal{M}, x_1)$
- And so on.

Theorems

A Family of Alternative Bandit Processes

Formed by bringing together a set of n bandit processes, with the constraint that

- **Control 1** must be applied to just one bandit process at a time
- **Control 0** is applied to the other $(n-1)$ bandit processes.

The Forwards Induction Theorem

For a simple family of alternative bandit problem processes a policy is optimal if and only if it coincides almost always with a forward induction policy.

DAI Theorem

For a simple family of alternative bandit processes a policy is optimal if and only if

- At each stage the bandit process selected for continuation is almost always one of those whose dynamic allocation index (DAI) is then maximal.

Outline

More Properties of DAI

Definition

The DAI for a bandit process D in state x may be written more precisely as

$$\nu(D,x) = \sup_{\{f:f(0)=0\}} \left[\frac{\mathbb{E}\left[\sum_{t=0}^{\infty} a^{t+f(t)} R(x(t),1) \,|\, x(0)=x\right]}{\mathbb{E}\left[\sum_{t=0}^{\infty} a^{t+f(t)} \,|\, x(0)=x\right]} \right] \tag{1}$$

$$= \sup_{\{\tau>0\}} \nu_\tau(D,x) = \sup_{\{\tau>0\}} \left[\frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} a^{t} R(x(t),1) \,|\, x(0)=x\right]}{\mathbb{E}\left[\sum_{t=0}^{\tau-1} a^{t} \,|\, x(0)=x\right]} \right] \tag{2}$$

Remarks

Expression (2) correspond to the case where the stopping set is used. We will focus on this case, first.

Lemma

The supreme in (3) is attained by setting $\Theta_0 = \{y \in \Theta : \nu(D, y) < \nu(D, x)\}$.

Proof.

Dropping the condition $x(0) = x$ from the notation, for any non-random $s \in \mathbb{Z}^+$ and τ,

$$\nu_\tau(D, x) = \left[\frac{\mathbb{E} \sum_{t=0}^{\sigma-1} a^t R(x(t), 1) + \mathbb{E}\left[\mathbb{E} \sum_{t=s}^{\tau-1} a^t R(x(t), 1) \mid x(s)\right]}{\mathbb{E} \sum_{t=0}^{\sigma-1} a^t + \mathbb{E}\left[\mathbb{E} \sum_{t=s}^{\tau-1} a^t \mid x(s)\right]} \right], \tag{3}$$

where $\sigma = \min(s, \tau)$. If $\tau > s$, then

$$\mathbb{E}\left[\sum_{t=s}^{\tau-1} a^t R(x(t), 1) \mid x(s)\right] / \mathbb{E}\left[\sum_{t=s}^{\tau-1} a^t \mid x(s)\right] = \nu_{\tau-s}(D, x(s)) \leq \nu(D, x(s)), \tag{4}$$

Lemma

Proof.

(Continued.) From (3) and (4) it follows that if the probability of the event

$$E_s = \{\tau > s \cap \nu(D, x(s)) < \nu_\tau(D, x)\} \tag{5}$$

is positive, and define a random variable ρ, where

$$\rho = \begin{cases} s & \text{if event } E_s \text{ occurs,} \\ \tau & \textit{otherwise}. \end{cases} \tag{6}$$

Then, we have

$$\nu_\rho(D, x) > \nu_\tau(D, x). \tag{7}$$

Thus, if τ is such that the supremum is attained in (2) we must have $\mathbb{P}(\cup_{s=1}^{\infty} E_s) = 0$.

Lemma

Proof.

(Continued.) $\mathbb{P}\{\cup_{s=1}^{\infty} E_s\} = 0$, where $E_s = \{\tau > s \cap \nu(D, x(s)) < \nu_\tau(D, x)\}$ is equivalent to

- The probability that, starting in state x, the bandit process D passes through a state which belongs to the set $\Theta_s = \{y \in \Theta : \nu(D, y) < \nu(D, x)\}$ before process time τ is zero.
- Thus the stopping set Θ_0 which defines τ must include Θ_s (except for a subset that will not be reached before τ).

A similar argument shows that $\mathbb{P}\{\nu(D, x(\tau)) > \nu(D, x) \mid x(0) = x\} = 0$.

- Since otherwise $\nu_\tau(D, x)$ could be increased by increasing τ in an appropriate fashion for those realizations of D for which $\nu(D, x(\tau)) > \nu(D, x)$.

Also, $\nu_\tau(D, x)$ is unaffected by the inclusion or exclusion from Θ_0 of states belonging to the set $\Theta_e = \{y \in \Theta : \nu(D, y) = \nu(D, x)\}$. □

Outline

Two Simple Cases for Calculating DAI

Consider any bandit process D and an arbitrary state x. Dropping D from the notations, we have

$$\nu(x) = \sup_{\tau>0} \nu_\tau(x) = \sup_{\tau>0} \frac{R_\tau(x)}{W_\tau(x)} = \sup_{\sigma\geq 0} \frac{R(x,1) + a\mathbb{E}[R_\sigma(x(1)) \,|\, x(0) = x]}{1 + a\mathbb{E}[W_\sigma(x(1)) \,|\, x(0) = x]}, \tag{8}$$

where τ and σ are stopping ties, and τ is restricted to be positive.

Case 1: The Deteriorating Case

- We have $\mathbb{P}\{\nu(x(1)) \leq \nu(x(0)) \,|\, x(0) = 0\} = 1$
- Since $\nu(x(1)) = \sup_{\sigma>0} \frac{R_\sigma(x(1))}{W_\sigma(x(1))}$, we have $\nu(x) = R(x,1)$.
- One-step look ahead policy is optimal.

Two Simple Cases for Calculating DAI

$$\nu(x) = \sup_{\tau>0} \nu_\tau(x) = \sup_{\tau>0} \frac{R_\tau(x)}{W_\tau(x)} = \sup_{\sigma\geq 0} \frac{R(x,1) + a\mathbb{E}[R_\sigma(x(1)) \,|\, x(0)=x]}{1 + a\mathbb{E}[W_\sigma(x(1)) \,|\, x(0)=x]}. \tag{9}$$

Case 2: The Improving Case

- We have $\mathbb{P}\{\nu(x(s)) \geq \nu(x(s-1)) \cdots \geq \nu(x(1)) \geq \nu(x(0)) \,|\, x(0)=x\} = 1$, for some non-random integer s.
- From the above equation, we have

$$\nu(x) = \sup_{\sigma>0} \frac{\mathbb{E}\left[\sum_{t=0}^{s-1} a^t R(x(t),1) \,|\, x(0)=x\right] + a^s \mathbb{E}\left[R_\sigma(x(s)) \,|\, x(0)=x\right]}{1 + a + \cdots + a^{s-1} + a^s \mathbb{E}[W_\sigma(x(s)) \,|\, x(0)=x]}, \tag{10}$$

- This will simplify if the defining condition holds for all s and if we set $s = \infty$.

Outline

General Method for Calculating DAI

- When problem does not simplify: use the standard bandit as a calibration device
- Consider simple family of alternative bandit processes $\{D, \lambda\}$, formed by
  - An arbitrary bandit process D
  - And a standard bandit process with the parameter λ.
- Optimal policies for $\{D, \lambda\}$ are DAI policies. Start by
  - Continuing D if $\nu(D) > \lambda$
  - Continuing the standard process if $\nu(D) < \lambda$.
  - When $\nu(0) = \lambda$, optimal policy can start either way.
- **Goal:** Find a value of λ such that an optimal policy for $\{D, \lambda\}$ can either start with D or the standard policy.

Maximum Total Expected Reward

$$R(\{D, \lambda\}, x) = \max(\lambda/(1-a), R(x,1) + \mathbb{E}\left[R(\{D, \lambda\}, y)\right]. \tag{11}$$

Outline

The Scheduling Problem

Problem Setup

- j jobs to be carried out by a single machine.
- Times taken to process the jobs: independent integer-valued random variables
- Jobs must be processed one at a time.
- At the beginning of each time unit
  - Any job may be selected for processing
  - Whether or not the job processed during the preceding time unit has been completed
  - There is no penalty involved in switching from one job to another
- Probability that $(t+1)$ time units are required to complete the processing of job i, conditioned on more than t time units being needed is $p_i(t)$.
- Reward for finishing job i at time s is $a^s V_i$ $(0 < a < 1; V_i > 0, i = 1, 2, \ldots, n)$.
- **Problem:** Decide which job to process next, to maximize the total expected reward.

The Scheduling Problem

Let D be a bandit process such that

- $\Theta = \{C\} \cap \mathbb{Z}$, where state C signifies that the job has been completed.
- $$\mathbb{P}(s_{t+1}|s_t = C) = \begin{cases} 1, & \text{if } s_{t+1} = C \\ 0, & \text{otherwise} \end{cases}, \quad \mathbb{P}(s_{t+1}|s_t \neq C) = \begin{cases} p(t), & \text{if } s_{t+1} = C \\ 1 - p(t), & \text{if } s_{t+1} = s_t + 1 \\ 0, & \text{otherwise} \end{cases}$$
- $$R(s_t, u = 1) = \begin{cases} 0, & \text{if } s_t = C \\ p_t V, & \text{otherwise.} \end{cases}$$

Deteriorating Bandit Process

- If $p(t)$ is a non-increasing function of t.
- Job to to continued at any time is one of those for which $p_i(t_i)V_i$ is the largest.

The Scheduling Problem

(Modified) Improving Bandit Process

- If $p(t)$ is a non-decreasing function of t, where $\nu(C) = 0$ and $\mathbb{P}\{\nu(x(s)) \geq \nu(x(s-1)) \cdots \geq \nu(x(1)) \geq \nu(x(0)) \geq 0 \,|\, x(0) = x, x(s) \neq C\} = 1$,
- By definition,

$$\nu(x) = \frac{\mathbb{E}\left[\sum_{t=0}^{\tau-1} a^t R(x(t), 1) \,|x(0) = x)\right]}{\mathbb{E}\left[1 + a + \ldots + a^{\tau-1} \,|\, x(0) = x\right]}, \tag{12}$$

where $\tau = \min\{s : x(s) = C\}$. Rewrite it into the form

- $$\nu(x) = \frac{V(1-a)\mathbb{E}\left[a^{\tau-1} \,|\, x(0) = x\right]}{1 - \mathbb{E}\left[a^{\tau} \,|\, x(0) = x\right]} \tag{13}$$

The Scheduling Problem

(Modified) Improving Bandit Process

- For an arbitrary job with no restriction on the function $p(t)$, τ for which the supremum is achieved should not be greater than the time taken to complete the job.
- Uncompleted jobs: state coincides with process time, so that the stopping set wich defines τ must be reached at some non-random (and possibly infinite) time r. Thus, τ is of the form $\min\{r, \min[s : x(s) = C]\}$. Therefore, we have $\nu(C) = 0$, and

$$\nu(x) = \sup_{r>0} \frac{V(1-a)\{\mathbb{E}a^{\rho-1} - \mathbb{P}(\rho > r)\mathbb{E}(a^{\rho-1}|\rho > r)\}}{1 - \mathbb{E}a^{\rho} + \mathbb{P}(\rho > r)\{\mathbb{E}(a^{\rho}|\rho > r) - a^{r}\}}, \tag{14}$$

where $\rho = \min\{s : x(s) = C\}$.

Outline

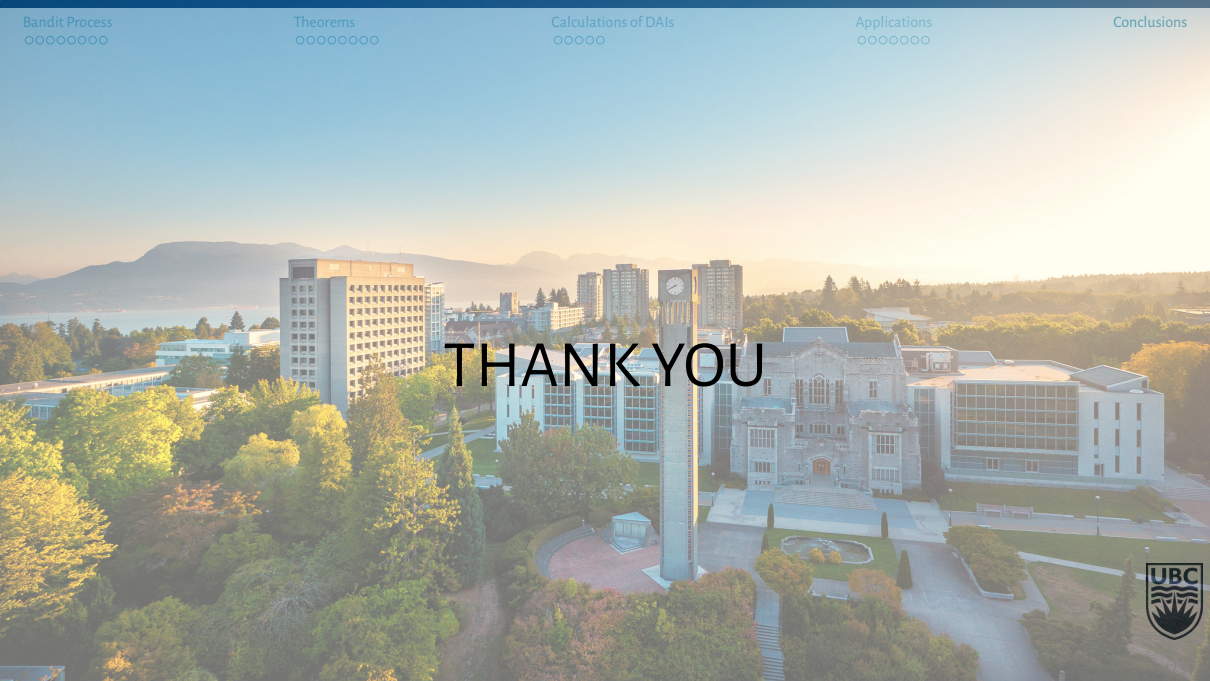
Multi-Armed Bandit Problem

Problem Setup

- n arms which may be pulled repeatedly in any order.
- Each pull takes one time unit and only one arm maybe pulled at a time
- A pull may result a success or a failure. Bernoulli process with unknown probability θ_i ($i = 1, 2, \ldots, n$): sequence of successes and failures result from pulling arm i
- Successful pull at time t yields a reward a^t ($0 < a < 1$), while unsuccessful pull yields 0.
- $t = 0, \theta_i \sim \frac{(\alpha_i(0)+\beta_i(0)+1)!}{(\alpha(0)!\beta(0)!)}\theta_i^{\alpha_i(0)}(1-\theta_i)^{\beta_i(0)}$, i.e., Beta dist. with $(\alpha_i(0), \beta_i(0))$.
- If in the first t pulls there are r successes $(\alpha_i(t), \beta_i(t)) = (\alpha_i(0) + r, \beta_i(0) + t - r)$
- If the $(t+1)$st pull on arm i takes place at time s, the expected reward, conditioned on the record of successes and failures up to then, is a^s times the expected value of a beta variate with parameters $(\alpha_i(t), \beta_i(t))$, which is $((\alpha_i(t) + 1)/(\alpha_i(t) + \beta_i(t) + 2)$.

Conclusions

- Bandit Process
- Dynamic Allocation Index
- Forward Induction Policy
- Stochastic Scheduling Policy


THANK YOU

References

[1] J. C. Gittins, "Bandit processes and dynamic allocation indices," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 41, no. 2, pp. 148–164, 1979.