Cover Sheet of Hongbo's Note

Journal of Multi-armed Bandit Allocation Indices

Hong-Bo Zhang

College of Engineering and Computer Science, Australian National University,
Canberra, 2601, Australia
u6170245@anu.edu.au

Course Summer Research Project 16
Lecturer Steve, Tony and Shane

Abstract: This is a reading journal of the book "Multi-armed Bandit Allocation Indices" by J. C. Gittins, the 1st edition.

Contents

1. Chapter 2

FABP Family of alternative bandit process. This is a special type of exponentially discounted semi-Markov decision process.

Markov decision process Markov + a set of decision (at each stage) + a set of reward (at each stage) $\longrightarrow$ Markov decision process.

Future reward Future reward will be discounted, so that the total reward obtained is finite. Sometimes, the discount has some explanations, such as the impatience of the player, the pressimstic of the investor, preferrence of finding a bug earlier, and so on.

Payoff *expected* value of *total discounted* reward $\longrightarrow$ payoff

Aim Aim is to find policy to maximize payoff.

Non-discounted Some non-discounted problem can also be solved in this way by (1) the discounted factor $\to 1$ (2) reinterpretation of discount factor.

Semi-Markov The time interval between two successive decision time are themselves random variables.

Discount factor/parameter The β appears in Bellman equation are discount factor, $-\ln\beta$ is discount parameter.

Borel set and σ algreba Borel set is the σ algreba of the set containing all the open set of a topological space. σ on a set is collection of its subsets, which is closed under union, intersection, complement.

Some symbol

- State-space, Θ. For example, $1, 2, 3, 4, 5, 6$ in a dice.
- σ-algreba $\mathcal{X}$ (in Page 14). The measures of state space. For example, empty set, $\{i\}$, $\{i, j\}$, $\{i, j, k\}$ ... are all belong to $\mathcal{X}$. And $\mathcal{X}$ must contain all the subset which only contain one element. For example, $\emptyset, \{1, 2, 3\}, \{4, 5, 6\}, \{1, 2, 3, 4, 5, 6\}$ is a σ algreba of Θ, but it doesn't contain the single-element-subset, so it is not $\mathcal{X}$. It defines the measures of the problem, after this, we can define integal.
- control $\Omega(x)$. The control when the process is in state x. For example, in a special game, if a dice is in the state of 3, its next states must be one of $1, 2, 6$. Then in this case, the $\{\Omega(3) = \text{transfer to } 1, 2 \text{or} 6\}$. More example on controls, $\Omega(3) = \{u1, u2\}$, where $u1 = (30\% to 1, 70\% to 6)$, $u2 = (10\% to 2, 15\% to 5)$
- reward $r(x, u)$.
- $P(A|x, u)$, where $A \in \mathcal{X}$. The probability that the state $y \in A$ of the process immediately after time t, given that at time t the process is in state x and control $u \in \Omega(x)$. For example, $P(\{2\}|3, (\text{transfer to } 1))$ means the probability of finding dice in 2 at immediately after time t, given that at time t the dice is in 3 and control $(transfer to 1)$. Actually, the $P = 0$ in this case.

- $F(B|x,y,u)$. The probability that duration interval $\in$ Borel set B until the next decision happens, given that at time t the process is in state x and control u, leading to a transition of state y. For example, $F\left((1,2)\,|3,1,(\text{transfer to } 1)\right)$ means, when a dice transfers from 3 to 1,the probability of the time interval $\in (1,2)$.
- Obviously, the measures of $\Omega(\cdot), r(\cdot,u), P(A|\cdot,u)$ are $\mathcal{X}$ measureable (the integral interval), and $F(B|\cdot,\cdot,u)$ is $\mathcal{X}^2$ measureable.

Infinite time To ensure that an infinite number of transitions do not occur in a finite time, we require for $\forall\epsilon$, exist *delta*,

$$\int_{\Theta} F\left((\delta,\infty)\,|x,y,u\right) P\left(dy|x,u\right) > \epsilon$$

$P(dy|x,u)$ means at time t, the state is x and control u, and in the immediately next time, the probability of transfering to state $(y, y+dy)$. F means from x applying u, the probability of duration of interval of transfering to y.

Markov decision process defined by F A Markov decision process is a semi-Markov decision process with

$$F\left(\{c\}\,|x,y,u\right) = 1, \forall x, y \text{ and } u.$$

This is just mathematical description of constant time interval between decision times. We can assume $c = 1$.

Policy A policy is any rule that specifies the control to be applied at time t. It is a function of (1) time t, (2) state at time t, (3) previous decision time, (4) states at these previous decision times, and (5) control applied at those times. So control applied at time t may depend on the *entire previous history* of the process, not on what happens after time t, which is called *past-measurable*.

Some nomenclature of policy *Optimal* policy, the one maximize the expected total reward over all policies for *every* initial state. *Deterministic* policy, involve no randomization. *Stationary* policy, involve no explicit time-dependence. *Markov* policy, the control chosen at time t depends only on t and the state at time t. Notice, we will study Markov decision process, the "Markov" here has nothing to do with Markov in Markov policy.

Bellman Equation for Markov decision problem

$$R(D,x) = \max_{u\in\Omega(x)}\left[r(x,u) + a\int_{\Theta} R(D,y)P(dy|x,u)\right] \quad (x\in\Theta) \tag{1.1}$$

where D is a Markov decision process, and $R(D,x)$ is the total reward function to infinite future with initial state x. The above equation reads the total reward today equals to the reward today plus the discounted expected total reward of tommorrow. For simplicity, we write $R(D,\cdot)$ as $X(\cdot)$. If the control space $\Omega(x)$ is finite.

Iterative Solving Bellman Equation The above equation (1.1) will have lots of properties, which will not be noted here. It can also be solved iteratively.

$$X_n(x) = \max_{u\in\Omega(x)}\left[r(x,u) + a\int_{\Theta} X_{n-1}(y)P(dy|x,u)\right] \quad (x\in\Theta, n = 1,2,3\ldots)$$

As $n\to\infty$, $X_n \to X$, due to the above map is contraction mapping.

Bellman Equation for Semi-Markov decision problem

$$X(x) = \max_{u\in\Omega(x)} \left[r(x,u) + \int_{\Theta} \int_{t=0}^{\infty} a^t X(y) F(dt|x,y,u) P(dy|x,u) \right] \quad (x \in \Theta, n = 1,2,3\ldots)$$

This equation also reads the total reward at this decision time equals to the reward at this decision time plus discounted expected total reward at next decision time. Since when is the next decision time is unknown, so we average all the possible next decision time by distribution F. This functional can also be solved by iterative method.

Bandit Process A bandit process is one kind of semi-Markov decision process. Its control space set $\Omega(x)$ only consist of two element, 0 and 1, for every state x. Control 0 is *freezing* control, when it is applied, no reward accrues and the state will not change. (recall that reward is a function of control) Control 1 is *continuation* control.

Process time, realiaztion (1) In bandit process, the total time during which the process has not been frozen by control 0 is *process time*. That is, process time is the duration that the process is not frozen. (2) The process time when control 1 is applied for $(i+1)$th time is t_i. The sequences of t_i and $x(t_i)$ constitute a *realization* of bandit process. So, process time is the total time without stopping work of bandit. The time t_i and state $x(i)$ at the ith pull form a sequence which is important. (3) This sequence will not depend on the sequence of control applied, since all the control 0 part has been removed. (4) If control 0 is never used before time t, the process time coincides with real time, the reward is $a^t r(x(t), u = 1)$, or abbreviate to $a^t r(t)$, otherwise, the reward will be $a^{t+f} r(t)$,

Standard bandit process (1) A standard bandit process is a bandit process with just one element in state space, and every point in time is a decision time. Therefore, a standard bandit process is a semi-Markov decision process with state space $|\Theta| = 1$ and control space $\Omega = 0, 1$, since there is only one state, it is not necessary to write control space as $\Omega(x)$. (2) A policy for standard bandit process is a Lebesgue-measurable function $I(t)$ that map $t \in [0, \infty)$ to $0, 1$, where t is time. This will not be applied to general bandit process, since for a general one, we need a Lebesgue-measurable function $I(t, x(t))$ mapping to $0, 1$, since there is more than one state. (3) The total reward under this policy is $\lambda \int a^t I(t) dt$. Here since there is only one state, so the reward $r(x(t), u)$ can be abbreviated to: at time t, if $u = I(t) = 0$, $r = 0$, otherwise, if $u = I(t) = 1$, $r = \lambda$, where $\lambda := r(x_0, u = 1)$, x_0 is the single element in state space.

Freezing rule An abitrary policy for a bandit process is *freezing rule*. Obviously, for a standard bandit process, the freezing rule is just a function $I(t)$. Given any freezing rule f for a non-standard bandit process, the random variables f_i is total time for which control 0 is applied before the $(i+1)$th application of control 1. So $f_i \geq f_{i-1}$. In a policy, there will be a sequence of applying control 0 and control 1. For every period of applying control 1, its decision time is t_i. For every period of applying control 0, its duration is $f f_{i-1}$, and $f_i = \sum_{j=-1}^{i-1} f f_j$.

Stopping rule, stopping time Stoppling rule: a policy that before τ control 1 is applied, after τ control 0 is applied. τ is corresponding *stopping time*. In the language of freezing rule, a stopping rule is $f_i = 0$, if $t_i < \tau$, and $f_i = \infty$, if $t_i > \tau$ for all i.

Family of alternative bandit processes (FABP) This is a decision process formed by bringing together a set of n bandit processes with the same discount factor. At each decision time, only one bandit process in that set is applied by control 1, while all the others is applied by control 0.

FABP more details (1) The state space of FABP is the product of the state spaces for individual bandit processes. (2) The realization of FABP is a sequence of decision time and a list of states of individual bandit process at that time. (3) The reward of FABP at each decision time is the one of individual bandit process which is then continued. (which means the individual bandit process which will be operated in the next). (4) control set at each decision time is the control set of all the possible individual bandit process which will be operated in the next. Applying a control the FABP means selecting an individual bandit process to continue. For example, there are 10 individual bandit process, and the current bandit process under operation is 3, and the following possible bandit processes are $1, 7$, then the control set at this decision time is: $\{0, 1\}$ for bandit 1; $\{0, 1\}$ for bandit 7. When applying control 1 to bandit 7, the bandit 7 will be operated, and all the other (such as bandit 1) will be frozen.

Simple FABP, (SFABP) In FABP, usually the control set will not contain the control set of all the individual bandit processes, such as the example given in previous paragraph. However, if a FABP's control set contains the control set of all the individual bandit processes, it names SFABP. Actually, in our fuzzing problem, it is a SFABP. Since at each decision time, every seed has possibility to be fuzzed.

Gittins Index For abitrary bandit process B (it's bandit process, a subclass of more general semi-Markov decision process, and a superclass of FABP). The payoff under freezing rule f will be

$$R_f(B) = \mathbf{E} \sum_{i=0}^{\infty} a^{t_i + f_i} r(t_i)$$

In this formular, t_i is the ith process time (remove frozen time from real time), and f_i is the the summation of frozing time before t_{i+1}, therefore, the real time of t_i is $f_i + t_i$. Furthermore, t_i and f_i are all random variable, since bandit process is a subclass of semi-Markov decision process. Especially, if the freezing rule is null freezing rule, i.e., $\forall i, f_i = 0$, the payoff is $R(B)$. Furthermore, we define

$$\begin{aligned} W_f(B) &= \mathbf{E} \sum_{i=0}^{\infty} a^{f_i} \int_{t_i}^{t_{i+1}} a^t dt \\ \mu_f(B) &= R_f(B)/W_f(B) \\ \mu'(B) &= \sup_{\{f : f_0 = 0\}} \mu_f(B) \end{aligned}$$

W_f has the unit of time, it seems to be a normalization factor. μ_f should be the normalized average payoff per unit time. In the last line $\{f : f_0 = 0\}$ means the freezing rule f is constraint to $f_0 = 0$, which means at real time 0, the process is under continuation control 1. μ' means the maximum value of μ_f for all the freezing rule satisfied $f_0 = 0$. To be more specific, "all freezing rules" means a set of $\{t_i, f_i\}$.

Gittins Index for stopping rule In a simpler case, the freezing rule is stopping rule, all the formular above are reduced to

$$\begin{aligned} R_\tau(B) &= \mathbf{E}\sum_{t_i<\tau} a^{t_i} r(t_i) \\ W_\tau(B) &= \mathbf{E}\int_0^\tau a^t dt \\ \mu_\tau(B) &= \frac{R_\tau(B)}{W_\tau(B)} \\ \mu(B) &= \sup_{\tau>0} \mu_\tau(B) \end{aligned}$$

where τ is stopping time, which is a random variable here. So the expectation, and sup are both resepected to τ. Furthermore, like $\{f : f_0 = 0\}$, $\tau > 0$ guarantee that the process is under continuation control at real time 0. Notice that in this paragraph and the previous one, the reward $r(t_i)$ is abbreviation of $r(x(t_i), u = 1)$, and all the quatities depends on initial state $x(t = 0)$. So in the following, the notation such as $R_f B, x$ will be used to indicate the initial state is x explicitly.

Acknowledgments

References