Reinforcement Learning

Jacob Denson[1]

October 19, 2014

[1]Influenced by Lectures by Joseph Modayil

Chapter 1

Markov Decision Processes

Reinforcement learning is an attempt to design intelligences that can both learn from and interact with the environment. In its most general form, an agent must select action in a world that yield the greatest reward for the agent. We begin with a basic problem which provides an introduction to the techniques we use in later chapters, and then we move on to the most general specification of the problem.

The main idea of reinforcement learning is the reward hypothesis – that all goals can be thought of as the maximization of the expected value of some reward signal responding to actions that an agent performs. Given this, the main process by which all agents can become intelligent is to find ways to maximize this reward signal. This is a hypothesis which we will touch on throughout the class. Once concern with this view is that we are not really building intelligence through this method, only building machines that attempt to imitate intelligence. We provide endorphisms in the human brain guiding behaviour as a counter example.

AI takes the intentional stance on this issue, treating all systems as having intensions, regardless of what they do. The intention of evolution is to attempt to build animals with the perfect survival skills. Evolution is a random process which does not really think about its intentions, yet we can see it as having the intentions by the result of the action. This abstracts the problem, and is a useful way of thinking about problems.

Chapter 2

Bandit Problems

Imagine you are at a casino, which has a healthy selection of games. The casino you are in specializes in specific slot machine games. You of course, being a strategic game player, want to find the slot machine that maximizes your profit from gambling. You hope that you will end up with a profit, but you will also be satisfied, if this is not possible, to find the machine that results in you losing the least amount of money. From the start, you have no idea how any slot machines work, so the only way to determine its reward is to get results from pulling its lever. This is the one-armed bandit problem.

One armed bandits provide a good introduction to reinforcement learning because they are non-associative. Actions taken at a current moment have no impact on future actions, and thus are simpler to analyse. We will see that bandit problems are a subset of reinforcement learning problems. A formal specification of the problem is below.

A one-armed bandit problem is a tuple $\langle \mathcal{A}, \mathcal{R} \rangle$, where $\mathcal{A}$ is a non-empty set, called the set of actions, and $\mathcal{R}$ is a function from $\mathcal{A}$ to a probability distribution over the real numbers, called the reward distribution. Given an action a, we denote the distribution mapped from a by $\mathcal{R}$ as

$$\mathcal{R}(\cdot\,|a)$$

We will assume $\mathcal{A}$ is countable, and the range of each distribution mapped from $\mathcal{R}$ discrete.

After a bandit problem has been specified, how do we make a decision on actions we want to take? The key is a policy or decision function. First we decribe the memory of an agent. An experience is a finite sequence from $\mathcal{A} \times \mathbf{R}$ of the form

$$(A_0, R_0), (A_1, R_1), \dots, (A_n, R_n)$$

where for each pair (A_i, R_i), the event that $\mathcal{R}(\cdot\,|A_i)$ takes the value R_i has positive probability. A policy or decision function is a mapping from experiences to a probability distribution over actions. Intuitively, policies learn from experience what actions to take, by changing their policy as actions are learned. We denote the distribution from an experience by

$$\pi(\cdot\,|(a_0, r_0), \ldots, (a_n, r_n))$$

The goal of the one-armed bandit can be written rigorously now: Find a policy such that the reward gained is maximized. How do we formalize the reward? Given a policy π, and experience X we define the reward R_π recursively by the equations

$$R_\pi(X) = \sum_{a,r} \pi(a|X)\mathcal{R}(r|a) + \gamma R(X, (a, r))$$

where γ, a number between 0 and 1, is the discount rate. The smaller the discount rate, the more rewards in the present impact the score, rewarding greedy decisions up front, whereas as the discount rate increases, the future reward becomes more important. The policy thus runs for an infinite amount of periods, the reward converging to an amount. An alternate reward would be to only let the policy execute for a certain number of periods, say T of them. Then the reward equation is R_T, where

$$R_T = \sum_{a,r} \pi(a)\mathcal{R}(r|a)(r + R_{T-1}((a, r)))$$

$$R_k(X) = \sum_{a,r} \pi(a|X)\mathcal{R}(r|a)(r + R_{k-1}(X, (a, r)))$$

$$R_0(X) = \sum_{a,r} \pi(a|X)\mathcal{R}(r|a)r$$

Given either of these rewards, we denote the policy that maximises the reward to be π_* Of course, if we know $\mathcal{R}$ exactly, it is easy to find π_*. Simply define π_* to map absolutely to the action which has the highest expected reward. The real problem is finding a policy given that we know nothing about $\mathcal{R}$. The best we can hope for is that our policy converges to π_*, either in the infinite strategy or as the number of periods allowed in the problem extends to infinity. This forms the definition of regret. What the policy should really minimize is the regret of the policy, the difference between the agents action at each interval and the optimal action. In a good policy, the regret should converge to zero over time.

We begin by attempting to estimate the expected reward of an action. For an action a, denote the expected reward, or action value, as $q_*(a)$. If we keep an unbiased estimator of $q_*(a)$ for each a, then if we sample every action with non-zero probability at every interval, the estimator will converge to $q_*(a)$. To be simple, given an experience $(A_0, R_0), \ldots, (A_t, R_t)$. Given that the experience contains the subsequence $(a, R_{k_0}), (a, R_{k_1}), \ldots, (a, R_{k_m})$, denote by $Q_t(a)$ the sample average defined by the formula

$$Q_t(a) = \frac{R_{k_0} + R_{k_1} + \ldots R_{k_m}}{m}$$

This estimator is unbiased, and thus as $m \to \infty$, $Q_t(a) \to q_*(a)$. We normally define $Q_0(a) = 0$ for all actions a. Storing all these values becomes computationally infeasible. Thus the following equations helps us calculate the value recursively. If $R_{k_{m-1}}$ was discovered at time t', we may iteratively construct Q_t by

$$\begin{aligned} Q_t(a) &= \frac{R_{k_0} + R_{k_1} + \ldots R_{k_m}}{m} \\ &= \frac{m-1}{m}\frac{R_{k_0} + \ldots R_{k_{m-1}}}{m-1} + \frac{R_{k_m}}{m} \\ &= \frac{m-1}{m} Q_{t'}(a) + \frac{R_{k_m}}{m} + \frac{Q_{t'}(a)}{m} - \frac{Q_{t'}(a)}{m} \\ &= Q_{t'} + \frac{1}{m}[R_{k_m} - Q_{t'}(a)] \end{aligned}$$

This may seem to solve our problem completely, but in order for our sample of each action to approach infinity, we must select each action with non-zero probability at each step. It then follows that we cannot use our converging estimate to our advantage completely - if we switch our policy to choose only the action a such that $Q_t(a)$ is the maximum of all other actions, then our convergence stops – we cannot assure Q_t tends to q_* for any action but action a, which is the only action we will continue to sample. Thus we must form a balance of acting in the greedy way, choosing an action with the highest sample average, and acting in an exploratory way.

A naive policy, the greedy method, always takes the action with the highest sample average. The problem with this method is, if the reward for some action gets distorted low enough, it may never be taken again, even if the actual expected reward is much higher. This normally occurs if the variance of actions is too high. Given a low-variance reward function, the greedy policy may choose the right action, but no guarentees are set. One other helpful trick may be to set $Q_0(a)$ optimistically, that is, not the default value of zero but some value higher than all rewards for an action. Then all values are at least tried one before they begin to settle down, so some exploration is done. Again, there are no guarentees on how successful this will be. A smarter policy, the ε greedy method, attempts to fix this issue by guarenteeing convergence at the cost of complete optimality.

Our first policy which attempts to solve this problem is the ε greedy method, for some real number ε such that $0 \leq \varepsilon \leq 1$. The idea is that we act greedily most of the time, whereas for some proportion of the time related to ϵ, we act randomly, ensuring we sample enough time to get convergence of our sample averages. The ε greedy policy π is defined for each action a by

$$\pi(a|(A_0,R_0),\dots,(A_t,R_t)) = \begin{cases} \varepsilon + \frac{1-\varepsilon}{|\mathcal{A}|} & : Q_t(a) \text{ is the maximum over all a} \\ \frac{1-\varepsilon}{|\mathcal{A}|} & : \text{otherwise} \end{cases}$$

No action has non-zero probability, so Q_t converges to q_* as the number of actions taken approaches infinity. This means that the ε greedy strategy will act optimally a proportionally to $\varepsilon + \frac{1-\varepsilon}{|\mathcal{A}|}$. If we decrease the value of ε, our policy will act more correctly, but it will probably take longer for Q_t to converge to q_*. (NEEDS FURTHER PROOF). Then the reward for π converges to

$$(1-\varepsilon)\max_{a\in\mathcal{A}} q_*(a) + \frac{\varepsilon}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} q_*(a)$$

as the length of the experience extends to infinity.

ε policies are a simple method that converge to a close to optimal policy. However, if there is one action a with a very low expected value, the reward for π will be strongly affected. The problem is that the action with low value expectation is selected at equal probability with any other action. The softmax method attempts to fix this. We select an action more probably if it has a higher sample average. A softmax method takes as a parameter a temperature τ, and calculates a policy by:

$$\pi(a|(S_0,A_0),(S_1,A_1),\dots,(S_n,A_n)) = \frac{e^{Q_t(a)/\tau}}{\sum_{a'} e^{Q_t(a')/\tau}}$$

Then as we increase the length of the experience, our reward converges to

$$\sum_a \frac{e^{q_*(a)/\tau}}{\sum_{a'} e^{q_*(a')/\tau}} q_*(a)$$

As τ decreasing in value, the disparity between the selection of actions with differing sample average increases, causing the policy to act more greedily. Besides the convergence of the policy, there are no mathematical guarentees on which τ to use that have been developed, which makes it hard to pick which τ is useful. This means unless the quality of softmax is absolutely needed, it is hard to recommend softmax over an ε greedy algorithm do to the more complicated definition of softmax.

A more advanced and recent solution to the problem is to compute confidence intervals for the sample averages collected. Then we act greedily, our policy selecting the action a such that the equation

$$Q_t(a) + c\sqrt{\frac{log(t)}{K_t(a)}}$$

is maximized, where $K_t(a)$ is the upper confidence bound for the sample average of the action a.

A final problem we will discuss with bandit problems is the problem of nonstationarity. Regardless of how long a policy interacts with a bandit, the distributions mapped from $\mathcal{R}(a)$ remain the same. However, what if these values do change according to some rule, that is, if there is instead a sequence of functions $(\mathcal{R}_0, \mathcal{R}_1, \dots)$, which the policy plays with at each interval. If these functions have no relation, there is no chance a policy could learn from them. Given that these chain in some systematic way, there may be a way to carry over experience from one reward kernal to the next. Our policy do not want convergence though, as q_* changes and thus there is no meaningful convergence we would want. Thus out normal sample average Q_t will not perform well. We take inspiration for a new sample average calculator from the equation

$$Q_t = Q_{t'} + \frac{1}{k}[R_k - Q_{t'}]$$

It is of the form

$$\text{'new average'} = \text{'old average'} + \text{'step size'}[\text{'new value'} - \text{'old average'}]$$

If we adjust the step size to a non-zeroconstant value α, our sample average will not settle down to some value, but will still have some convergence to the mean. The full formula will then become

$$Q_{t+1} = (1-\alpha)^t Q_0 + \sum_{m=1}^{n} (1-\alpha)^{t-i} R_{k_m}$$

We call this the exponential recency weighted average. Things in the past become exponentially less inportant than things in the present. This allows us to get some convergence while not settling down completely.

We leave off this chapter with a complication in the problem. Suppose you are playing a one-armed bandit that changes $\mathcal{R}$ as in the non-stationary problem, but that we can identify which positions use the same $\mathcal{R}$ by some characteristic, say, the slot machine has the same colour when you pull it. We could then attempt to predict each individual colour's sample average, and from each colour learn what the best action would be. An additional challenge would be the added complexity in that which slot you pull may determine the next colour you see. Then the best action you would take may be impacted by which colour you see next. We will see that these complications form the bulk of the full 'reinforcement learning' problem.

Chapter 3

Reinforcement Learning: A Short Introduction to Markov Decision Processes

The next step to adding to the learning process is the introduction of state, the 'colours' of the bandit problem discussed at the end of the last chapter represent what we mean by the word state, a summary of experiences relevant enough to predict the present to the best of our abilities. Actions thus have further consequences than immediate reward. We call this associate learning problem the reinforcement learning problem.

The reinforcement learning task consists of an agent/learner/decision-maker interacting with its environment, everything that is outside of the agents direct control. The two interact continually through the environments presentation of rewards to the agents actions in the environment, providing 'reinforcement' that allows the agent to improve its behaviour over time, the agents primary motive. Over time, the environment changes, but the environment should give enough information at the current time or via previous interactions to allow an agent to form a state representation of the agents current standing with the markov property - the presense of enough relavant information to make intelligent decisions based on previous information about the same state. We define the reinforcement learning task rigorously through the introduction of a markov decision problem, or MDP.

A Markov Decision Problem, or MDP, is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P} \rangle$, where S is a non-empty set of states, A is a non-empty set of actions, R is a mapping from an initial state, an action, and a resultant state, $S \times A \times S$, to a distribution over the real numbers called the reward distribution (think of the reward distribution from the bandit problem), and P is a mapping from $S \times A$ to a probability distribution over states, called the transition probability kernel. We write

$$R(\cdot\,|s,a) \qquad\qquad\qquad\qquad P(\cdot\,|s,a)$$

for the reward and probability distributions mapped from state s and action a. Each state s has associated with it a subset of A called the admissible actions at state s. We write this as $\mathcal{A}_s$. The set of states, actions, and rewards are assumed to be countable in this course, though it should be said that this need not always be assumed. An experience in a markov decision process is a sequence with elements in S, A, and R, in that order, normally denoted

$$(S_0, A_0, R_1, S_1, A_1, R_2, \dots, S_n, A_n, R_{n+1})$$

We write R_{t+1} instead of R_t in the sequence to represent the fact that R_t is created leading into the next time period. A policy π is then a mapping from experience to a probability distribution over the admissible actions in the most recent state seen. The aim of a policy is to maximise its return, the expectation of reward the policy may see. We denote the expected reward after t steps by $\mathbf{E}_\pi[G_t]$. We define the expectation by the following equations based on the initial state S_0:

$$\mathbf{E}_\pi[G_t] = \sum_{k=0}^{\infty} \gamma^k R_k$$

$$R_0(S_0, A_0, R_1, \dots, S_n) = \sum_{a,s,r} \pi(a|S_0, A_0, \dots, S_n) P(s|S_n, a) \mathcal{R}(r|S_n, a, s) r$$

$$R_t(S_0, \dots, S_n) = \sum_{a,s,r} \pi(a|S_0, A_0, \dots, S_n) P(s|S_n, a) R_{t-1}(S_0, \dots, S_n, a, r, s)$$

Intuitively, given some initial state S_0, the policy picks an action, gets a reward from it, and moves to a new state. Given the new information, it picks an action, gets a reward, and the process continues. R_i then becomes the expected reward from the t'th action given we are using policy π. The optimal function π_* maximises this value.

Before we end this short chapter