Notes on Reinforcement Learning

ERMO WEI

Contents

Actor-Critic Method Still working on it . . .

Bellman Equation Due to the Markov property of MDP, we can define the value function of a policy recursively to be

$$\begin{aligned} V^{\pi}(s) &= E_{\pi}\{R_t|s_t = s\} \\ &= E_{\pi}\{\Sigma_{k=0}^{\infty}\gamma^k r_{t+k+1}|s_t = s\} \\ &= E_{\pi}\{r_{t+1} + \gamma\Sigma_{k=0}^{\infty}\gamma^k r_{t+k+2}|s_t = s\} \end{aligned}$$

If policy is stochastic, then we have

$$\begin{aligned} V^{\pi}(s) &= \Sigma_a \pi(s,a)\Sigma_{s'}\Pr(s'|s,a)[R(s,a) + \gamma E_{\pi}\{\Sigma_{k=0}^{\infty}\gamma^k r_{t+k+2}|s_{t+1} = s'\}] \\ &= \Sigma_a \pi(s,a)\Sigma_{s'}\Pr(s'|s,a)[R(s,a) + \gamma V^{\pi}(s')] \\ &= \Sigma_a \pi(s,a)\{R(s,a) + \Sigma_{s'}\Pr(s'|s,a)\gamma V^{\pi}(s')\} \end{aligned}$$

If policy is deterministic, then the Value function of policy π at state s is

$$V^{\pi}(s) = R(s,\pi(s)) + \gamma\Sigma_{s'}\Pr(s'|s,\pi(s))V^{\pi}(s')$$

Thus, the Bellman optimality equation can be defined as

$$V^{*}(s) = \max_a\{R(s,a) + \gamma\Sigma_{s'}\Pr(s'|s,a)V^{*}(s')\}$$

Detail see [1]

Bellman Operator Consider a MDP, Bellman Operator for policy π can be defined as a map from a value function to another value function

$$T^{\pi} : \mathbb{R}^N \to \mathbb{R}^N$$

where N is the cardinality of state space, i.e. $N = |S|$.

Still working on it . . .

Boltzmann Selection Still working on it . . .

Contraction The operator F (function f) is a α-contraction ($0 \le \alpha < 1$, called Lipschitz constant for F) with respect to some norm $\|\cdot\|$ if

$$\forall X, \overline{X} : \|FX - F\overline{X}\| \le \alpha\|X - \overline{X}\| \quad \text{or} \quad \|f(X) - f(\overline{X})\| \le \alpha\|X - \overline{X}\|$$

- Theorem 1. The sequence $X, FX, F^2X, \ldots$ converges for every X. e.g. $X, f(X), f(f(X)), \ldots$ converges for every X

  Proof:

  Useful fact : Cauchy sequences : If for $x_0, x_1, x_2, \ldots,$ we have that

  $$\forall\epsilon, \exists K : \|x_M - x_N\| < \epsilon \quad \text{for} \quad M, N > K$$

  then we call $x_0, x_1, x_2, \ldots,$ a Cauchy sequence.

  If $x_0, x_1, x_2, \ldots,$ is a Cauchy sequence, and $x_i \in \mathbb{R}^n$, then there exists $x^* \in \mathbb{R}^n$ such that $\lim_{i\to\infty} x_i = x^*$.

Proof: Assume $N > M$.

$$\begin{aligned} \|F^M X - F^N X\| &= \|\Sigma_{i=M}^{N-1}(F^i X - F^{i+1} X)\| \\ &\leq \Sigma_{i=M}^{N-1}\|F^i X - F^{i+1} X\| && \text{by triangle inequality} \\ &\leq \Sigma_{i=M}^{N-1}\alpha^i \|X - FX\| && \text{use the condition in Theorem} \\ &= \|X - FX\|\Sigma_{i=M}^{N-1}\alpha^i \\ &= \|X - FX\|\frac{\alpha^M}{1-\alpha} \end{aligned}$$

As $\|X - FX\|\frac{\alpha^M}{1-\alpha}$ goes to zero for M going to infinity, we have that for any $\epsilon > 0$ for $\|F^M X - F^N X\| \leq \epsilon$ to hold for all $M, N > K$, it suffices to pick K large enough. Hence, $X, FX, \ldots$ is Cauchy sequence and converges.

- Theorem 2 (Banach fixed-point theorem). F has a unique fixed point X^* which satisfies $FX^* = X^*$ and all sequences X, FX, $F^2X, \ldots$ converge this unique fixed point X.

  Proof:

  Suppose F has two fixed points. Let's say

  $$\begin{aligned} FX_1 &= X_1 \\ FX_2 &= X_2 \end{aligned}$$

  this implies, $\|FX_1 - FX_2\| = \|X_1 - X_2\|$. At the same time we have from the contractive property of F

  $$\|FX_1 - FX_2\| \leq \alpha\|X_1 - X_2\|$$

  Combining both gives us

  $$\|X_1 - X_2\| \leq \alpha\|X_1 - X_2\|$$

  Hence, $X_1 = X_2$.

Detail see [2]

Data Efficiency for Policy Gradient Method In policy gradient method, when we try to evaluate the expected return of a policy π, we have to run the policy several time to be able to have reasonable estimation of how good the policy is. Let $J(\theta)$ denote the expected return of policy π_θ, and τ denote the trajectory or rollout or history (these terms are interchangeable) of executing the policy π_θ, then we know

$$\begin{aligned} J(\theta) =& E_{P(\tau;\theta)}[\Sigma_{t=0}^{H}\gamma^t R(s_t, u_t)|\pi_\theta] \\ =& \Sigma_\tau P(\tau;\theta)R(\tau) \end{aligned}$$

where $P(\tau;\theta)$ is the probability of having a trajectory τ by following policy π_θ and $R(\tau)$ is just the accumulated reward of that trajectory. In policy gradient method, after we evaluate the policy π_θ, we may want to improve it and use a new policy. Thus, the sample we collected during the process of following policy π_θ are discarded. However, it will preferable if we can reuse the data gathered of following one policy to estimate the value of following another policy. The method "likelihood ratio" estimation make this data reuse possible.

In practice, if we want to evaluate $J(\theta)$, we may want to draw the rollout samples from the distribution induced by policy π_θ. After taking N samples $\{\tau_0, \tau_1, \ldots, \tau_N\}$, we have a unbiased estimator:

$$J(\hat{\theta}) = \frac{1}{N}\Sigma_i R(\tau_i)$$

Imagine, however, instead of π_θ, we only have $\pi_{\theta'}$ available, we can do some trick like this

$$\begin{aligned}J(\theta) =&\Sigma_\tau P(\tau;\theta)R(\tau)\\ =&\Sigma_\tau P(\tau;\theta)\frac{P(\tau';\theta')}{P(\tau';\theta')}R(\tau)\\ =&\Sigma_\tau P(\tau';\theta')\frac{P(\tau;\theta)}{P(\tau';\theta')}R(\tau)\\ =&E_{\tau'}[\frac{P(\tau;\theta)}{P(\tau';\theta')}R(\tau)]\end{aligned}$$

Now, we can estimate the $J(\theta)$ with respect to the the distribution induced by $\pi_{\theta'}$. This method of estimating one expectation with respect to another distribution is called ''importance sampling". So, we can rewrite the $J(\hat{\theta})$ as

$$J(\hat{\theta}) = \frac{1}{N}\Sigma_i R(\tau_i)\frac{P(\tau;\theta)}{P(\tau';\theta')}$$

Note, in the equation above, we have a term $\frac{P(\tau;\theta)}{P(\tau';\theta')}$. Since we don't have the model of the world, it's not possible to directly compute $P(\tau;\theta)$. But if we expand the fraction term, we can see that

$$\frac{P(\tau;\theta)}{P(\tau';\theta')} = \frac{\Pi_{t=0}^T\pi_\theta(u_t|s_t)}{\Pi_{t=0}^T\pi_{\theta'}(u_t|s_t)}$$

Thus, we should be able to calculate the likelihood $\frac{P(\tau;\theta)}{P(\tau';\theta')}$ for any two policies θ and θ'. Actually, we can have a mixture distributions, where $P(\tau;\theta)$ is replaced by $\frac{1}{N}\Sigma_j P(\tau_j|\theta_j)$ where τ_j are trajectory of following θ_j, then we can make use of multiple source of distributions.

ϵ-greedy ϵ-greedy is a common used exploration strategy for Model-free reinforcement learning. This strategy can be seen as a combination of greedy strategy and random strategy. Every time when agent choose an action to perform, it choose greedily with probability 1-ϵ (exploitation), and randomly with probability ϵ (exploration). However, based on the changes of ϵ value, this method have two versions. First, the ϵ value can decrease as the learning process goes on, this is the *decay exploration* method. Another one which are more commonly used is that we fix the value of ϵ.

The difference between those methods is that, the first one can be GLIE if the ϵ-value goes 0 eventually but the second one cannot. Suppose we have a counter of how many times a state have been visited, $n_t(s)$ and a constant c. As long as the ϵ value for the a state $\epsilon_t(s) = \frac{c}{n_t(s)}$ where $0 < c < 1$, this method can be consider GLIE. However, in practice, we usually use fixed value for ϵ and after the convergence of the estimation value of action, we switch to greedy selection.

Ergodic MDP An MDP is said to be ergodic if for each policy π the Markov chain induced by π is ergodic. We are giving the definition of Ergodicity below. But first, we will give some auxiliary definitions. Several definition of Markov chain:

- Reducibility
  A Markov chain is said to be irreducible if it is possible to get to any state from any state.

- Aperiodicity
  A state i has period k if any return to state i must occur in multiples of k time steps. For example, suppose it is possible to return to a state in {6, 8, 10, 12, ...} time steps, then k would be 2, even though 2 does not appear in this list. If $k = 1$, then the state is said to be aperiodic: returns to state i can occur at irregular times.

- Recurrence
  A state i is said to be transient if, given that we start in state i, there is a non-zero probability that we will never return to i. State i is recurrent (or persistent) if it is not transient. Recurrent states are guaranteed to have a finite hitting time.

Here we have the definition of Ergodicity.

A state i is said to be ergodic if it is aperiodic and positive recurrent. In other words, a state i is ergodic if it is recurrent, has a period of 1 and it has finite mean recurrence time. If all states in an irreducible Markov chain are ergodic, then the chain is said to be ergodic.

It can be shown that a finite state irreducible Markov chain is ergodic if it has an aperiodic state. **A model has the ergodic property if there's a finite number N such that any state can be reached from any other state in exactly N steps.** For example, we will have $N = 1$ if we have a fully connected transition matrix where all transitions have a non-zero probability. **Additionally, an stationary distribution d^{π} of states exists and is independent start state s_0.**

Detail see [3].

GLIE GLIE stands for "Greedy in the Limit of Infinite Exploration". The learning policies in RL can be divided into two broad categories: a *decay exploration* strategy which become more and more greedy and *persistent exploration* which always maintain a fix exploration rate. The advantage of the first one is that we can eventually converge to the optimal policy. The second one may have the advantage always be adaptive but may not converge to the optimal. (In here, we talk about convergence in the sense that the behavior will become optimal. It is possible that some of the algorithm converge to the correct Q-value but still behave randomly with some probability by using persistent exploration strategy, Q-learning with fix ϵ-greedy for example). We may want to consider this in the context of on-policy&off-policy.

If a *decay exploration* strategy has the following two characters:

1. each action is executed infinitely often in every state that is visited infinitely often, and
2. in the limit, the learning policy is greedy with respect to the Q-value function with probability 1.

Than we can consider this decay exploration strategy GLIE. Some example of GLIE include Boltzmann Selection, ϵ-greedy.

Detail see [4].

Lipschitz Continuity Still working on it . . .

Markov Decision Process blah blah here

Markov Decision Process can be seen as a extension of Markov Chain with additional action set (allowing selection) and reward function (motivation). It can be reduced to Markov chain if we have only one action per state and same reward for all the state.

Monte Carlo method Monte Carlo method is a way of making the prediction in model-free environment. The question it wants to solve is that suppose we have a policy π known, how good is this policy? In this case, we evaluate the policy by giving the method episodes of experience $\{s_1, a_1, r_2, \ldots, s_T\}$ generated by following policy π and wants the value function V^{π} as output.

As we know, the value of being in a state s is the expectation of the discounted rewards received afterwards.

$$V^{\pi}(s) = E_{\pi}[r_{t+1} + \gamma r_{t+2} + \ldots + \gamma^{T-1} r_T]$$

two methods can be used here : first-visit and every-visit method

On-policy and Off-policy An RL algorithm can be essentially divided into two parts, the *learning policy* and *update rule*. The first one is a non-stationary policy that maps experience (state visited, action chosen, reward received) to into a currently choice of action. The second part is how the algorithm uses experience to change its estimation of the optimal value function. In off-policy algorithm, the *update rules* doesn't have relationship with *learning policy*, that is the *update rules* doesn't care the what action agent take. Q-learning can be consider as the off-policy algorithm.

$$Q_{t+1}(s,a) = (1-\alpha)Q_t(s,a) + \alpha(r_t + \gamma \max_{a'} Q(s',a'))$$

We can see that the Q-value is update based on the $\max_{a'} Q(s',a')$, which doesn't depend on the action the agent was taking.

However, if we take a look of SARSA(0), which is very similar to Q-learning.

$$Q_{t+1}(s,a) = (1-\alpha)Q_t(s,a) + \alpha(r_t + \gamma Q(s',a'))$$

We can see the update is based on the Q-value of the next action of the agent. Thus it is an on-policy algorithm. The convergence condition are heavily depend on the *learning policy*, The Q-value of SARSA(0) can only converge to optimality in the limit only if the learning policy behavior optimal in the limit. The SARSA(0) and Q-learning will be same if we use greedy action selection strategy.

Detail see [4].

Policy Gradient Reinforcement Learning Still working on it . . . Ergodic MDP.

Stationary Distribution Still working on it . . .

Stochastic Game Stochastic game can been seen as an extension of MDP.

Temporal Difference Method

References

[1] Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

[2] Keith Conrad. The contraction mapping theorem. *Expository paper. University of Connecticut, College of Liberal Arts and Sciences, Department of Mathematics*, 2014.

[3] Ronald Ortner. Linear dependence of stationary distributions in ergodic markov decision processes. *Operations research letters*, 35(5):619–626, 2007.

[4] Satinder Singh, Tommi Jaakkola, Michael L Littman, and Csaba Szepesvári. Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine Learning*, 38(3):287–308, 2000.