Reinforcement Learning — Deepmind course

Tariq Berrada

September 2020

Contents

1   Introduction

The following are notes from the Reinforcement Learning course given by professor *David Silver* at UCL and hosted by Deepmind. It is actively being completed.

2   Model Free Prediction

In order to estimate the value function of an MDP, we introduce two different methods : Monte Carlo learning and Temporal Difference learning.

2.1   Monte Carlo Learning

This method can be used on episodic MDPs (every episode must eventually finish, it relies on backpropagating information at every episode of learning.// Formally, the goal is to learn the value function v_π from episodes of experience generated by following a policy π.

$$S_1, A_1, R_2, ..., S_k \sim \pi$$

$$v_\pi(s) = \mathbf{E}_\pi[G_t/S_t = s]$$

$$G_t = R_{t+1} + \gamma R_{t+2} + ...$$

2.1.1   First visit Monte Carlo evaluation

Each state gets visited once per episode, to evaluate a state s :

Algorithm 1: First visit Monte Carlo initialization state update

---
Result: Updated state value
if *1st time state s is visited in current episode*
 then
   Increment counter $N(s) \leftarrow N(s) + 1$
   Increment total return $S(s) \leftarrow S(s) + G_t$
   Update state value $V(s) : \frac{S(s)}{N(s)}$
else
   pass
end

Using the law of large numbers, we can prove that $V(s) \rightarrow \frac{S(s)}{N(s)}$ while $N(s) \rightarrow \infty$.

2.1.2   Every visit Monte Carlo evaluation

State can be visited more than once per episode, to evaluate a state s :

Algorithm 2: Every visit Monte Carlo state update

Result: Updated state values
if *On state s*
 then
   Increment counter $N(s) \leftarrow N(s) + 1$
   Increment total return $S(s) \leftarrow S(s) + G_t$
   Update state value $V(s) : \frac{S(s)}{N(s)}$
else
   pass
end

2.1.3   Incremental mean

$$\mu_k = \frac{1}{k} \sum_{j=1}^{k} x_j$$

$$\mu_k = \frac{1}{k}(x_k + \sum_{j=1}^{k-1} x_j)$$

$$\mu_k = \mu_{k-1} + \frac{1}{k}(x_k - \mu_{k-1})$$

Using this formulation, we can modify the Monte Carlo update process and define **Incremental Monte Carlo** :

Algorithm 3: Incremental Monte Carlo

Result: Updated state values
for each state S_t with return G_t
Increment counter $N(S_t) \leftarrow N(S_t) + 1$
Update state value $V(S_t) \leftarrow V(S_t) + \frac{1}{N(S_t)}(G_t - V(S_t))$
Update state value $V(s)$: $\frac{S(s)}{N(s)}$

For non stationnary problems, we can track the running mean and forget about old episodes, this enables us to do an update inspired by the exponential moving average:

$$V(S_t) \leftarrow V(S_t) + \alpha(G_t - V(S_t))$$

2.2 Temporal Difference Learning

Temporal difference learning learns from experience and is adapted to model free incomplete episodes. It makes use of bootstrapping (updates the guess for the current state by making other guesses and taking action on child states).
The goal is to learn v_π from experience under policy π.
In incremental every visit Monte Carlo, we update $V(S_t)$ toward the actual return G_t.

$$V(S_t) \leftarrow V(S_t) + \alpha(G_t - V(S_t))$$

In the simplest temporal difference learning TD(0), we update the state $V(S_t)$ toward the estimated return $R_{t+1} + \gamma.V(S_{t+1})$ which substitutes the real return G_t.

$$V(S_t) \leftarrow V(S_t) + \alpha(R_{t+1} + \gamma V(S_{t+1}) - V(S_t))$$

- $R_{t+1} + \gamma V(S_{t+1})$: TD target

- $R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$: TD error

remark

- TD can learn from incomplete sequences or continuing environments while MC waits until the episode is finished to generate the return.

- TD is more efficient because it has low variance (bias variance tradeoff.

- MC converges to the solution of mean squared error, so the best observed fit minimizes $\sum_{k=1}^{K} \sum_{t=1}^{T_k} (g_t^k - V(S_t^k))^2$.

- TD(0) converges to the maximum likelihood Markov model, which is the solution to the MDP $< S, \mathcal{A}, \hat{\mathcal{P}}, \hat{\mathcal{R}}, \gamma >$ that best fits the data.

$$\hat{\mathcal{P}}_{s,s'}^a = \frac{1}{N(s,a)} \sum_{k=1}^{K} \sum_{t=1}^{T_k} \mathbf{1}(s_t^k, a_t^k, s_{t+1}^k = s, a, s')$$

$$\hat{\mathcal{R}}_s^a = \frac{1}{N(s,a)} \sum_{k=1}^{K} \sum_{t=1}^{T_k} \mathbf{1}(s_t^k, a_t^k = s, a) r_t^k$$

remarks

- Since TD exploits the Markov property, its more effective in Markov environments.

- In bootstrapping, the update involves an estimate (TD).

- In sampling, the update samples an expectation (MC and TD but not DP).

2.2.1   n-step returns

$$n = 1, \ G_t^{(1)} = R_{t+1} + \gamma V(s_{t+1})$$

$$n = 2, \ G_t^{(2)} = R_{t+1} + \gamma V(s_{t+1}) + \gamma^2 V(s_{t+2})$$

The general formula is :

$$G_t^{(n)} = R_{t+1} + \gamma R_{t+2} + ... + \gamma^n V(S_{t+n})$$

if $n = \infty$, the formula is the same as MC learning.

$$n = \infty, \ G_t^{\infty} = R_{t+1}n + ... + \gamma^{T-1} R_T$$

The update rule is:

$$V(S_t) \leftarrow V(S_t) + \alpha(G_t^{(n)} - V(S_t))$$

averaging

Since it can be difficult to find the optimal n to choose in the TD algorithm, we can average n-step returns over different values for n, for example.

$$G = \frac{1}{2}G^{(2)} + \frac{1}{2}G^{(4)}$$

2.2.2   TD-lambda

We can also combine information from all the time-steps, this is called $TD(\lambda)$:

$$G_t^{(\lambda)} = (1 - \lambda) \sum_{n \geq 1} \lambda^{n-1} G_t^{(n)}$$

In *forward view*:

$$V(S_t) \leftarrow V(S_t) + \alpha(G_t^{(\lambda)} - V(S_t))$$

2.2.3   Eligibility traces

How should the algorithms assign importance to problems ? Imagine a situtation where a bell rang 3 times without anything happening, then a light lit, then the agent got a reward.
What was the cause for this reward ? the bells or the light ?
There are two main paradigms:

- **Frequecy Heuristic :**   assigns credits to the most frequent states.

- **Recency Heuristic :**   assigns credits to the most recent state.

Eligibility traces combine both, to get a more intelligent estimate of the causal event.

$$\begin{cases} E_0(s) = 0 \\ E_t(s) = \gamma \lambda E_{t-1}(s) + \mathbf{1}(S_t = s) \end{cases} \tag{1}$$

Backward view TD-lambda

In backwards view, we keep the eligibility trace for every state then update the value for every state in proportion to the TD-error δ_t ad eligibility trace $E_t(s)$.

- $\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$

- $V(s) \leftarrow V(s) + \alpha \delta_t E_t(s)$

When $\lambda = 1$, credit is defered until the end of the episode (MC).

Theorem 1. *The sum of offline updates is identical for forward view and backward view TD-lambda:*

$$\sum_{t=1}^{T} \alpha \delta_t E_t(s) = \sum_{t=1}^{T} \alpha[G_t^{\lambda} - V(S_t)]\mathbf{1}(S_t = s) \tag{2}$$

In forward view, we average over n-step returns.
In backward view, we accumulate eligibility traces.

3 Model Free Control

For model-free prediction, we estimate the value function, in model free control, we optimize the return of an unknown MDP.

There are two main cases where model free control is used :

- The MDP is unknown, but experience can be sampled.

- The MDP is so big that we can only sample it.

There are two main ways to learn a policy:

- On-Policy learning : the agent learns about the policy π using experience sampled from π.

- Off-Policy learning : The agent learns about the policy π by using experience sampled from another policy μ.

In generalized policy iteration, we iterate between two behaviors :

- Policy evaluation : estimates v_π (e.g by iterative policy evaluation).

- Policy improvement : generates $\pi' \geq \pi$ by acting greedy (for example by greedy policy improvement).

3.1 Monte Carlo Control

We know that greedy policy improvement requires a model of the MDP since knowledge of the values of the states that an agent might be end up in given an action show in the function to maximize.

$$\pi'(s) = \text{argmax}_{a \in \mathcal{A}} \mathcal{P}_s^a + \mathcal{P}_{ss'}^a V(s') \tag{3}$$

Greedy policy improvement over $Q(s, a)$ is model-free :

$$\pi'(s) = \text{argmax}_{a \in \mathcal{A}} Q(s, a) \tag{4}$$

3.1.1 ϵ-greedy exploration

To ensure continual exploration, we can use an ϵ-greedy policy : /

$$\pi(a/s) = \begin{cases} \frac{\epsilon}{m} + 1 - \epsilon \text{ if } a^* = \text{argmax}_{a \in \mathcal{A}} Q(s, a) \\ \frac{\epsilon}{m} \text{ otherwise} \end{cases} \tag{5}$$

For actions that have a probability $\neq 0$, we take a greedy action with $p = 1 - \epsilon$ and a random action with $p = \epsilon$.

Theorem 2. *For any ϵ-greedy policy π, the ϵ-greedy policy π^* with respect to q_π is an improvement $v_{\pi'} \geq v_\pi$.*

3.2 GLIE

Definition 1. *GLIE - Greedy in the limit with infinite exploration*

1. *All state-action pairs are explored infinitely many times $\lim_{k \to \infty} N_k(s, a) = \infty$.*

2. *The policy converges on a greedy policy $\lim_{k \to \infty} \pi_k(a/s) = \mathbf{1}, \ a = argmax_{a' \in \mathbf{A}} q_k(s, a')$.*

Algorithm 4: GLIE Monte Carlo Control

Result: Updated state values

Sample k^{th} episode using π: $\{S_1, \ A_1, \ R_2, \ ..., \ S_T\} \sim \pi$

For each state S_t and action A_t in the episode :

$$N(S_t, A_t) \leftarrow N(S_t, A_t) + 1$$

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{1}{N(S_t, A_t)}(G_t - Q(S_t, A_t))$$

Then the policy can be improved using the new action-value function.

$$\epsilon \leftarrow \frac{1}{k}$$

$$\pi \leftarrow \epsilon\text{-greedy}(Q)$$

Theorem 3. *GLIE Monte-Carlo converges to the optimal action value function $Q(s, a) \rightarrow q_*(s, a)$.*

In practise, we can see that MC learning is unsuitable because it presents a very high variance. TD learning can result in lower variance while it is also online and therefore enables the exploitation of incomplete sequences.

3.3 SarSa

Definition 2. *SarSa*
$$Q(s, a) \leftarrow Q(s, a) + \alpha(R + \gamma Q(S', A') - Q(S, A))$$

In every timestep, Q is evaluated for one point (s, a), $Q \sim q_\pi$, followed by an ϵ-greedy policy improvement.

Algorithm 5: SarSa for on-policy control

Result: Updated state values

Arbitrarily initialize $Q(s, a) \ \forall s \in \delta, a \in A(s)$.

Set $Q(\text{terminal state}) = 0$

repeat

    for each episode, initialize S

    Choose A from S using policy derived from Q (e.g ϵ-greedy)

    for *each state S_t and action A_t in the episode* **do**

        Take action A, observe R, S'.

        Choose A' from S' using policy derived Q (e.g ϵ-greedy)

        Update ac
$$Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma Q(S', A') - Q(S, A)]$$
$$S \leftarrow S'$$
$$A \leftarrow A'$$

    end

until *S is not terminal*;

Theorem 4. *Sarsa converges to the optimal action-value function, $Q(s, a) \rightarrow q_*(s, a)$ under the following conditions :*

1. *GLIE sequence of policies $\pi_t(a/s)$*

2. *Robbins-Monro sequence of step-sizes α_t : $\sum_{t \geq 1} \alpha_t = \infty$; $\sum_{t \geq 1} \alpha_t^2 < \infty$*

4 Conclusion