Notes: Model-Free Episodic Control[1]

esgl Hu

December 8, 2017

1 Episodic Control

Episodic control is a complementary approach that can rapidly re-enact observed, successful policies. Episodic control records highly rewarding experiences and follows a policy that replays sequences of actions that previously yielded high returns.

Domain of applicability of episodic control may be hopelessly limited by the complexity of the world. In real environments the same exact situation is rarely. if ever, revisited, In RL terms, repeated visits to the exactly the same state are also extremely rare.

2 Algorithm

$Q^{\mathrm{EC}}(s,a)$ is a table that each entry in it contains the highest return over obtained by taking action a from state s. At the end of each episode, Q^{EC} is updated according to the return received as follow:

$$Q^{\mathrm{EC}}(s_t,a_t) \leftarrow \begin{cases} R_t & \text{if } (s_t,a_t) \notin Q^{\mathrm{EC}}, \\ \max\{Q^{\mathrm{EC}}(s_t,a_t), R_t\} & \text{otherwise}, \end{cases} \tag{1}$$

where r_t is the discounted return received after taking action a_t in state s_t. Note that (1) is not suited to a general purpose RL learning update: **since the stored value can never decrease, it is not suited to rational action selection in stochastic environment.**

Tabular RL methods suffer from two key deficiencies:

1. for large probles they consume a large amount of memory;
2. they lack a way to generalise across similar states.

To address the first problem, we limit the size of the table by removing the least recently updated entry once a maximum size has been reached.

For states that have never been visited, Q^{EC} is approximated by averaging the value of the k nearest states. Thus if s is a novel state then Q^{EC} is estimated as

$$\widehat{Q^{\mathrm{EC}}}(s,a) = \begin{cases} \frac{1}{k}\sum_{i=1}^{k} Q^{\mathrm{EC}}(s^{(i)},a) & \text{if } (s,a) \notin Q^{\mathrm{EC}}, \\ Q^{\mathrm{EC}}(s,a) & \text{otherwise}, \end{cases} \tag{2}$$

where $s^{(i)}, 1 = 1, ..., k$ are the k states with the smallest distance to state s

Algorithm 1 Model-Free Episodic Control

```
for each episodic do
    for $t = 1, 2, 3, ..., T$ do
        Recieve observation $o_t$ from environment.
        Let $s_t = \phi(o_t)$
        Estimate return for each action $a$ via (2)
        Let $a_t = \arg\max_a \widehat{Q^{\mathrm{EC}}}(s_t, a)$
        Take action $a_t$, receive reward $r_{t+1}$
    end for
    for $t = T, T-1, ..., 1$ do
        Update $Q^{\mathrm{EC}}(s_t, a_t)$ using $R_t$ according to (1)
    end for
end for
```

References

[1] Charles Blundell, Benigno Uria, Alexander Pritzel, Yazhe Li, Avraham Ruderman, Joel Z Leibo, Jack Rae, Daan Wierstra, and Demis Hassabis. Model-free episodic control. 2016.