Actor Critic Methods: From Paper to Code

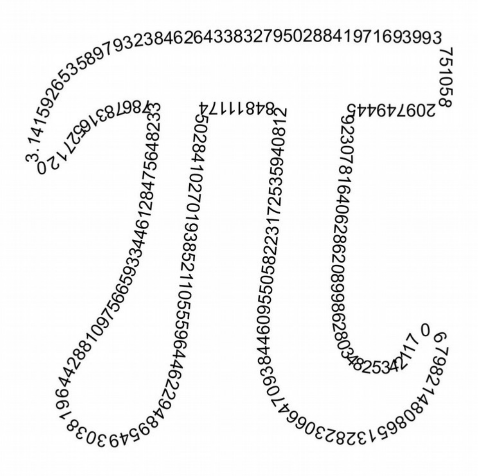
Monte Carlo Control Problem

Action Value Functions

- Without a model can’t transition to next valuable state
- Replace V with Q and use first visit MC
- How to handle explore exploit dilemma?

Exploring Starts

Pair random A, S → exploring starts



Pick random A at episode start

Gives good coverage of state and action space

Relaxing the Assumption of E.S.

- E.S. too limiting for some environments
- Use epsilon soft action selection

Exploratory action → $\frac{\epsilon}{|A(s)|}$

Greedy action → $1-\epsilon+\frac{\epsilon}{|A(s)|}$

- Greed increases over time

Algorithm Overview

```
Initialize Q(s,a) arbitrarily for all s, a; terminal → 0
Initialize arbitrary epsilon soft policy
Initialize list of Returns(s,a) for all states and actions
Repeat for large number of episodes:
    Generate episode using policy
    For each state s and action a in the agent’s memory:
        Calculate the return that followed first visit to s, a
        Append return G to list of Returns(s,a)
        Update Q as the average of Returns(s,a)
    For each state s in the agent’s memory:
        $A^* \leftarrow \underset{a}{argmax}\, Q(s,a)$
        $\pi(a|s) = \begin{cases} 1-\epsilon+\epsilon/|A(s)| & if\ a = A^* \\ \epsilon/|A(s)| & if\ a \neq A^* \end{cases}$
```

200,000 games; plot cum. win ratio over 1000 games $\epsilon \approx 0.001$

Conclusion

- Performance not too bad

- Able to do it without exploring starts

60
TACHYMETER
65
70
500
80
250
90
180
100
150
120
OMEGA
PROFESSIONAL
60
40
20
30
20
10
12
9
3
6
By Cristian Ibarra Santillán