ICRA 2012 Tutorial on Reinforcement Learning
6. Exploration

Pieter Abbeel
UC Berkeley

Jan Peters
TU Darmstadt

Exploration

- Key idea: to learn about unknown things, need to explore

→ Challenge: how to explore efficiently?

- Model-based Exploration Methods

- Model-free Exploration Methods

Random Exploration

- ε greedy
  - Every time step, flip a coin
  - With probability ε, act randomly
  - With probability 1-ε, act according to current policy

Problems with Random Exploration

1. Keep thrashing around once learning is done

2. Exploration is by no means targeting underexplored states

Solutions:

- Lower ε over time (addresses 1, not 2)
- A better solution: exploration functions

Exploration Functions

- Modify rewards: place high bonus reward for state-action pairs that are not well-known

→ The optimal policy in the modified MDP performs targeted exploration

Model-based Exploration: Rmax

[Brafman and Tenneholtz, 2002] Rmax sketch:

Initialize all (x,u) as unknown, and R(x,u) = R_{max}

Initialize T uniformly

Repeat until no more unknown states:

> Find optimal policy for current T, R
>
> Execute this policy
>
> Update T and R based on samples – if a pair (x,u) has been seen sufficiently often, make it "known" and give it its estimated reward

Model-free Exploration

- Insert exploration function into Q-learning:
  - Takes a value estimate and a count, and returns an optimistic utility, e.g. $b(q,n) = q + k/n$ (exact form not too important)

$$Q_{i+1}(s,a) \leftarrow (1-\alpha)Q_i(s,a) + \alpha\left(R(s,a,s') + \gamma \max_{a'} Q_i(s',a')\right)$$

now becomes:

$$Q_{i+1}(s,a) \leftarrow (1-\alpha)Q_i(s,a) + \alpha\left(R(s,a,s') + \gamma \max_{a'} f(Q_i(s',a'), N(s',a'))\right)$$

Exercise X (a)

Which of the following properties are true for Rmax?

(1) Rmax will exploit its knowledge of T of known states to more efficiently reach unknown states

(2) Everything else being equal, Rmax will favor passing through (x,u) with high reward

(3) Rmax can get stuck in a bad part of the state space

(4) The expected number of time steps for Rmax to have all states become known is optimal

Exercise X (b)

Which of the following is a natural generalization of the previous slides exploration function, $b(q,n) = q + k/n$, to Q-learning with linear function approximation, where

$$Q(x,u) = \sum_{i=1}^{n} w_i f_i(x,u)$$

(1) b(x,u) = q(x,u) + k / (number times (x,u) was visited)

(2) $b(x,u) = q(x,u) + \max_u \max_i f_i(x,u)$

(3) $b(x,u) = q(x,u) + \sum_i f_i(x,u)$

(4) $b(x,u) = q(x,u) + f(x,u)^T \Sigma f(x,u)$,

with $\Sigma = (\sum_j f(x^{(j)}, u^{(j)}) f(x^{(j)}, u^{(j)})^T)^{-1}$

Bayesian Exploration

- Exploration thus far:
  - Considered as a goal in and of itself

- In practice often trade-off between exploration and exploitation
  - Bayesian exploration methods try to address this problem in a Bayesian setting
  - Computationally expensive, but
    - Bandits --- Gittins indeices provide exact solution efficiently
    - General – some approximations

      e.g. Near-Bayesian Exploration (Kolter and Ng, 2009)