Apply of Bellman's equation for Q function

Peng Xie

Munich Institute of Robotics and Machine Intelligence

January 31, 2023

Table of contents

Why this equation is important

Its limitations

Its origin

Applications

Is there anything that can be improved about it?

Why this equation is important

$Q_i(x, u) = R(x_i, u_i) + \mathbb{E}_e [\max_{u'} Q_{i+1}(f(x_i, u_i, e_i), u')]$

▶ tool for solving optimal value problems

▶ the basis of policy iteration in RL

▶ It allows us to express the value of one state as the value of another state [1]

▶ the dynamic optimization problem is turned into simple sub-problems

▶ simplify the reinforcement learning or Markov decision problems. Just make complex problems simple!

Its limitations

Linear equations, theoretically the solution can be solved:

$$\mathrm{v} = \mathcal{R} + \gamma \mathcal{P} \mathrm{v}$$
$$(\mathrm{I} - \gamma \mathcal{P}) \mathrm{v} = \mathcal{R}$$
$$\mathrm{v} = (\mathrm{I} - \gamma \mathcal{P})^{-1} \mathcal{R}$$

▶ complexity is $\mathrm{O}\left(\mathrm{n}^3\right)$

▶ only suitable for small-scale MRPS.

▶ large-scale MRP usually needs to use other iterative method.

Its origin

- Markov Decision Processes (MDP)
  - ▶ Elements
  - ▶ MDP problem
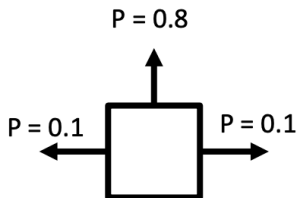 - ▶ Optimal Policy

Dynamic Programming (DP)
  - ▶ Solving MDP using DP
  - ▶ Policy Iteration
  - ▶ Value Iteration

$V_{k+1}(s) = \max_a [r(s, a) + \gamma \sum_{s'} p(s' \mid s, a) V_k(s')]v$

  - ▶ immediate reward $r(s, a)$
  - ▶ discounted value of successor state $\gamma \sum_{s'} p(s' \mid s, a) V_k(s')$

Applications

Rewards The agent moves in the selected direction with probability 0.8 and in the perpendicular directions with prob. 0.1. If the agent hits the wall, it stays in the same cell.

Application

$V_k(s)$

k=1

0	0	0.72	1.81
0		0	-99.9
0	0	0	0

k=5

0.81	1.60	2.47	3.74
0.27		0.30	-99.5
0	0.03	0.12	0.00

k=10

2.68	3.53	4.40	5.81
2.02		1.09	-98.8
1.39	0.90	0.74	0.12

k=1000

5.47	6.31	7.19	8.67
4.80		3.34	-96.7
4.16	3.65	3.22	1.53

k=1000

5.47	6.31	7.19	8.67
4.80		3.34	-96.7
4.16	3.65	3.22	1.53

$V_k(s)$

$$\pi(s) = \arg\max_a \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a)V(s') \right]$$

→	→	→	↑
↑		←	←
↑	←	←	↓

Optimal Policy

Improvement

The solution of large-scale MRP usually needs to use iterative method. Except the DP, some other iteration methods are also used:

▶ Monte-Carlo evaluation [2]
   ▶ agent run multiple
   ▶ sampling average value and update
▶ Temporal-Difference learning [3]
   ▶ don't need complete environment
   ▶ improvement on existing estimates without entire knowledge.

Thanks for your attention

📑 EN Barron and H Ishii.
The bellman equation for minimizing the maximum cost.
NONLINEAR ANAL. THEORY METHODS APPLIC.,
13(9):1067–1090, 1989.

📑 James B Hittner, Kim May, and N Clayton Silver.
A monte carlo evaluation of tests for comparing dependent
correlations.
The Journal of general psychology, 130(2):149–168, 2003.

📑 Gerald Tesauro et al.
Temporal difference learning and td-gammon.
Communications of the ACM, 38(3):58–68, 1995.