cs234 hw1

Jon Sondag

2021-03-05

1 Gridworld

(a)

set $r_s = -1$
values: $v_1 = 0$, $v_2 = 1$, $v_3 = 2$, $v_4 = 3$, $v_5 = -5$, $v_6 = 2$, $v_7 = 3$, $v_8 = 4$, $v_9 = 2$, $v_{10} = 3$, $v_{11} = 4$, $v_{12} = 5$, $v_{13} = 1$, $v_{14} = 0$, $v_{15} = -1$, $v_{16} = -2$

(b)

set $r_s = 1$, $r_g = 7$, $r_r = -3$
values: $v_1 = 12$, $v_2 = 11$, $v_3 = 10$, $v_4 = 9$, $v_5 = -3$, $v_6 = 10$, $v_7 = 9$, $v_8 = 8$, $v_9 = 10$, $v_{10} = 9$, $v_{11} = 8$, $v_{12} = 7$, $v_{13} = 11$, $v_{14} = 12$, $v_{15} = 13$, $v_{16} = 14$

(c)

$V^{\pi}_{new} = V^{\pi}_{old} * \dfrac{c}{1-\gamma}$

(d)

$c = 3$
Optimal policy is to move to any unshaded square, forever. Values of unshaded squares become ∞.

(e)

$r_s = 2$, $r_g = 8$, $r_r = -2$
Yes. For some values of γ it may be optmal to travel directly to square 12. For example if $\gamma = 0.01$, the best policy from square 11 is to move directly to square 12 (value: $2 + 0.01 * 8$) rather than moving around forever (value: $2 * \dfrac{1}{1-0.01}$).

(f)

$r_s = -6$, $r_g = 5$, $r_r = -5$, $\gamma = 1$
Yes. Set $r_s = -6$. Then from square 6 it's best to go to square 5 (value: $-6 + -5$) rather than to 12 (value: $-6 * 3 + 5$).

2 Value of Different Policies

Show $V_1^{\pi_1}(x_1) - V_1^{\pi_2}(x_1) = \sum\limits_{t=1}^{H} \mathbf{E}_{\mathbf{x_t}\sim\pi_2}\big(Q_t^{\pi_1}(x_t, \pi_1(x_1,t)) - Q_t^{\pi_1}(x_t, \pi_2(x_t,t))\big)$

Rewriting the RHS of that equation:

$$\sum_{t=1}^{H} \mathbf{E}_{\mathbf{x_t}\sim\pi_2}\big(Q_t^{\pi_1}(x_t, \pi_1(x_1,t)) - Q_t^{\pi_1}(x_t, \pi_2(x_t,t))\big) =$$
$$\mathbf{E}_{\mathbf{x_1}\sim\pi_2}\big(Q_1^{\pi_1}(x_1, \pi_1(x_1,1))\big) +$$
$$\sum_{t=1}^{H-1}\Big(- \mathbf{E}_{\mathbf{x_t}\sim\pi_2}\big(Q_t^{\pi_1}(x_t, \pi_2(x_t,t))\big) + \mathbf{E}_{\mathbf{x_{t+1}}\sim\pi_2}\big(Q_{t+1}^{\pi_1}(x_{t+1}, \pi_1(x_{t+1}, t+1))\big)\Big) +$$
$$\mathbf{E}_{\mathbf{x_H}\sim\pi_2}\big(Q_H^{\pi_1}(x_H, \pi_2(x_H, H))\big)$$

The first term in the sum,

$$\mathbf{E}_{\mathbf{x_1}\sim\pi_2}\big(Q_1^{\pi_1}(x_1, \pi_1(x_1,1))\big)$$
$= Q_1^{\pi_1}(x_1, \pi_1(x_1,1))$ [since x_1 is given]
$= V_1^{\pi_1}(x_1)$

The second term in the sum,

$$\sum_{t=1}^{H-1}\Big(- \mathbf{E}_{\mathbf{x_t}\sim\pi_2}\big(Q_t^{\pi_1}(x_t, \pi_2(x_t,t))\big) + \mathbf{E}_{\mathbf{x_{t+1}}\sim\pi_2}\big(Q_{t+1}^{\pi_1}(x_{t+1}, \pi_1(x_{t+1}, t+1))\big)\Big)$$
$$= \sum_{t=1}^{H-1}\Big(- \mathbf{E}_{\mathbf{x_t}\sim\pi_2}\big(Q_t^{\pi_1}(x_t, \pi_2(x_t,t))\big) + \mathbf{E}_{\mathbf{x_{t+1}}\sim\pi_2}\big(V_{t+1}^{\pi_1}(x_{t+1})\big)\Big)$$
$$= \sum_{t=1}^{H-1}\Big(-\mathbf{E}_{\mathbf{x_t}\sim\pi_2}\big(\sum_{r_t} r_t p(r_t|x_t, \pi_2(x_t,t)) + \sum_{x_{t+1}} p(x_{t+1}|x_t, \pi_2(x_t,t)) V_{t+1}^{\pi_1}(x_{t+1})\big)\big) +$$
$$\mathbf{E}_{\mathbf{x_{t+1}}\sim\pi_2}\big(V_{t+1}^{\pi_1}(x_{t+1})\big)\Big)$$
$$= \sum_{t=1}^{H-1}\Big(-\mathbf{E}_{\mathbf{x_t}\sim\pi_2}\big(\sum_{r_t} r_t p(r_t|x_t, \pi_2(x_t,t))\big) - \mathbf{E}_{\mathbf{x_{t+1}}\sim\pi_2}\big(V_{t+1}^{\pi_1}(x_{t+1})\big) + \mathbf{E}_{\mathbf{x_{t+1}}\sim\pi_2}\big(V_{t+1}^{\pi_1}(x_{t+1})\big)\Big)$$
$$= \sum_{t=1}^{H-1}\Big(- \mathbf{E}_{\mathbf{x_t}\sim\pi_2}\big(\sum_{r_t} r_t p(r_t|x_t, \pi_2(x_t,t))\big)\Big)$$

The third term in the sum,

$$\mathbf{E}_{\mathbf{x_H}\sim\pi_2}\big(Q_H^{\pi_1}(x_H, \pi_2(x_H, H))\big)$$
$$= \mathbf{E}_{\mathbf{x_H}\sim\pi_2}\big(\sum_{r_t} r_t p(r_t|x_t, \pi_2(x_t,t))\big)$$

Adding up the three terms we have:

$$V_1^{\pi_1}(x_1) + \sum_{t=1}^{H-1}\Big(-\mathbf{E}_{\mathbf{x_t}\sim\pi_2}\big(\sum_{r_t} r_t p(r_t|x_t, \pi_2(x_t,t))\big)\Big) + \mathbf{E}_{\mathbf{x_H}\sim\pi_2}\big(\sum_{r_H} r_H p(r_H|x_H, \pi_2(x_H,H))\big)$$
$$= V_1^{\pi_1}(x_1) + \sum_{t=1}^{H}\Big(- \mathbf{E}_{\mathbf{x_t}\sim\pi_2}\big(\sum_{r_t} r_t p(r_t|x_t, \pi_2(x_t,t))\big)\Big)$$
$$= V_1^{\pi_1}(x_1) - \sum_{t=1}^{H}\Big(\mathbf{E}_{\mathbf{x_t}\sim\pi_2}\big(\sum_{r_t} r_t p(r_t|x_t, \pi_2(x_t,t))\big)\Big)$$

The second term there is the definition of the value function $V_1^{\pi_2}$, so we get:
$V_1^{\pi_1}(x_1) - V_1^{\pi_2}(x_1)$

3 Fixed Point

(a)

We have defined:
$V_2 = BV_1$
and from lecture 2:
$\|BV' - BV''\|_\infty \leq \gamma \|V' - V''\|_\infty$

So for the base case n == 1:
$\|V_2 - V_1\|_\infty = \|BV_1 - BV_0\|_\infty \leq \gamma \|V_1 - V_0\|_\infty$

For the inductive case, assume that for $n-1$:
$\|V_n - V_{n-1}\|_\infty \leq \gamma^{n-1} \|V_1 - V_0\|_\infty$
Then, $\|V_{n+1} - V_n\|_\infty \leq \gamma \|V_n - V_{n-1}\| = \gamma\gamma^{n-1} \|V_1 - V_0\|_\infty = \gamma^n \|V_1 - V_0\|_\infty$

(b)

By definition of ∞ norm:
$\|V_{n+c} - V_n\|_\infty \leq \|V_{n+c} - V_{n+c-1}\|_\infty + \|V_{n+c-1} - V_{n+c-2}\|_\infty + ... + \|V_{n+1} - V_n\|_\infty$
The rhs of the previous equation $\leq \gamma^{n+c-1} \|V_1 - V_0\|_\infty + \gamma^{n+c-2} \|V_1 - V_0\|_\infty + ... + \gamma^n \|V_1 - V_0\|_\infty = \gamma^n \|V_1 - V_0\|_\infty \sum_{i=0}^{c-1} \gamma^i \leq \frac{\gamma^n}{1-\gamma} \|V_1 - V_0\|_\infty$

(c)

For $\epsilon > 0$, set $n = log_\gamma(\epsilon \|V_1 - V_0\|)$
Then $\|V_n - V_{n-1}\| < \epsilon$ and we have a Cauchy sequence

(d)

If the fixed point is not unique, there are values V_a, V_b such that $\|V_a - V_b\|_\infty > 0$, for fixed points a, b.

Since a and b are fixed points, $BV_a = V_a$ and $BV_b = V_b$. But in that case $\|BV_a - BV_b\|_\infty = \|V_a - V_b\| \nleq \gamma \|V_a - V_b\|$ (the last inequality failing when $\gamma < 1$), and we have a contradiction.

4 Value of Different Policies

(a)

[coding]

(b)

[coding]

(c)

Stochasticity increases the number of iterations required.

In this environment stochasticity makes the resulting policy more conservative: instead of aggressively moving towards the goal state, the agent now makes an effort to avoid "hole" terminal states, which it might fall into due to bad luck.