Sutton and Barto RL book Chpt 5 and BK_Simulation_97.pdf

Read Chapter 5 and my paper using this RW

A RW with 3 states

States 0, 1, 2

	0	1	2
	q_0	1-q_0	0
P =	1-q_1/2	q_1	1-q_1/2
	0	1-q_2	q_2

where (q_0=1/3, q_1=1/2, q_2=1/3)

m_0=1+ (1-q_0) *m_1. = (1+0)*q_0+ (1+m_1)**(1-q_0)

m_1=1+0*(1-q_1/2) +q_1*m_1+(1-q_1/2) *m_2

m_2=1 +(1-q_2)*m_1+q_2*m_2

$$m(x) = 1 + \sum_{y \neq 0} p_{xy} m(y) , \quad x \in S \tag{1}$$

$$w(x) = r(x) + \sum_{y \neq 0} p_{xy} w(y) , \quad x \in S \tag{2}$$

$$s(x) = 2r(x)w(x) - r^2(x) + \sum_{y \neq 0} p_{xy} s(y) , \quad x \in S . \tag{3}$$

$$g + h(x) = r(x) + \sum_{y \in S} p_{xy} h(y) , \quad x \in S , \tag{4}$$

For computing w_0, w_1, w_2, g, s_0,s_1,s_2 take

r_0=10,
r_1=20,
r_2=100.

β_1,β_2, … cycles of returns to a `ground state' 0.

Model 1. For the Simplest MDP version

Introduce 2 actions in state 1 (in some states).

action a_{11} as above (ie, p_{1a_11}=(1-q_1/2, q_1,1-q_1/2) and r_{1a_11}=20)

action a_{12} with p_{1a_12}=(1-q_1/3,q_1,1-2*q_1/3) and r_{1a_12}=10)

We have 2 policies, depending on the action in state 1.

Model 2. Introduce 2 actions in all states: 0,1,2

action a_{01} as above (ie, p_{1a_01}, and r_{1a_11} as above)
action a_{02} with p_{1a_12}=(q_0, 1-q_0/2,1-q_0/2) and r_{1a_12}=1)

action a_{11} as above (ie, p_{1a_11}=(1-q_1/2, q_1,1-q_1/2) and r_{1a_11}=20)
action a_{12} with p_{1a_12}=(1-q_1/3, q_1,1-q_2/2) and r_{1a_12}=1)

action a_{21} as above (ie, p_{2a_21}, and r_{2a_21} as above)
action a_{22} with p_{2a_21}=1-q_2/3, 1-2*q_2/3,q_2) and r_{2a_22}=50)

Now we have 8=2^3 policies, depending on the actison in states 0, 1, 2.

In Both Models the Idea is to find optimal policies using

$$\hat{g}_n - U_n^g \leq g \leq \hat{g}_n - L_n^g,$$

Estimation:

Proposition 1: The quantities $\hat{m}_n(x)$, $\hat{w}_n(x)$, $\hat{s}_n(x)$, $\hat{g}_n$, $\hat{h}_n(x)$ *are strongly consistent estimators of* $m(x)$, $w(x)$, $s(x)$, g *and* $h(x)$, *respectively.*

Proposition 3 Gives the C.I. for m,w,s,g,h, e.g.

$$g + h(x) = r(x) + \sum_{y \in S} p_{xy} h(y) , \quad x \in S , \tag{4}$$

where h is a function on S defined up to an additive constant. If the normalization $h(0) = 0$ is adopted, then $h(x)$ can be interpreted as the expected first passage differential reward from x to 0, i.e., $h(x) = \mathbf{E}_x \sum_{t=0}^{\beta_1 - 1} (r(X_t) - g) = w(x) - gm(x)$.

0 (10) => 1 (70) =>0. p_{01}=p_{10}=1

Eqs. (4) :

g+h(0) = r(0) + 1 * h(1)

g+h(1) = r(1) + 1* h(0)

h(0)=0 ===>.

g=10+h(1),
g+h(1)= 70+0

g=10+(70-g) ==>. g=80/2 =40 = lim_n (10+70 +10 +70 +10 +70 +…+10 +70) = (80/2) * n /n

h(1)= 70 -g = 30 .