Example 1

4. Consider the environment $Env_1 = <E, e_0, \tau>$ defined as follows:

$E = \{e_0, e_1, e_2, e_3, e_4, e_5, e_6\}$
$\tau (e_0, \alpha_0) = \{e_1\}$
$\tau (e_0, \alpha_1) = \{e_2, e_4\}$
$\tau (e_1, \alpha_3) = \{e_3\}$
$\tau (e_2, \alpha_2) = \{e_3, e_5\}$
$\tau (e_3, \alpha_2) = \{e_5, e_6\}$

$\tau(e_i, \alpha_i)$ defines the state transition for the environment in state e_i, given action α_i. We will assume that there are two possible agents for this environment Ag_1 and Ag_2. They are defined as:

$Ag_1(e_0) = \alpha_0$ $Ag_2(e_0) = \alpha_1$
$Ag_1(e_1) = \alpha_3$ $Ag_2(e_2) = \alpha_2$
$Ag_1(e_3) = \alpha_2$ $Ag_2(e_3) = \alpha_2$

Assume $|r|$ gives the total number of states in a particular run r. i.e., $|r_1| = 3$ if $r_1 = (e_0, \alpha_0, e_3, \alpha_2, e_5 \,|\, Ag_1, Env_1)$. The probability P, and utility U, of each run is given as follows:

$P(r \,|\, Ag_1, Env_1) = 8/(|r| \times |r|)$ $U(r) = 2 + |r|$
$P(r \,|\, Ag_2, Env_1) = |r|/13$

Example 1

- 4.1 Write down all possible runs for *Ag*1 and *Ag*2.

Ag1

r1 = e0,a0,e1,a3,e3,a2,e5

r2 = e0,a0,e1,a3,e3,a2,e6

Ag2

r3= e0,a1,e2,a2,e3,a2,e5

r4= e0,a1,e2,a2,e3,a2,e6

r5= e0,a1,e2,a2,e5

r6= e0,a1,e4

Example 1

- 4.2 Calculate the expected utility for *Ag*1 and *Ag*2

Ag1 = (8/(4*4) * (2 + 4)) + (8/(4*4) * (2 + 4)) = 6

Ag2 = (4/13 * 6) + (4/13 * 6) + (3/13 * 5) + (2/13 * 4) = 5.46

- 4.3 Which of *Ag*1 and *Ag*2 is optimal with respect to Env1 and *U?*

Ag1

Example 2

- Consider a student who has the choice to buy or not buy a textbook for a course. We'll model this as a decision problem with one Boolean decision node, B, indicating whether the agent chooses to buy the book, and two Boolean chance nodes, M, indicating whether the student has mastered the material in the book, and P, indicating whether the student passes the course. Of course, there is also a utility node, U.

- You might think that P would be independent of B given M, but this course has an openbook final—so having the book helps.

Example 2

- a. Draw the decision network for this problem.
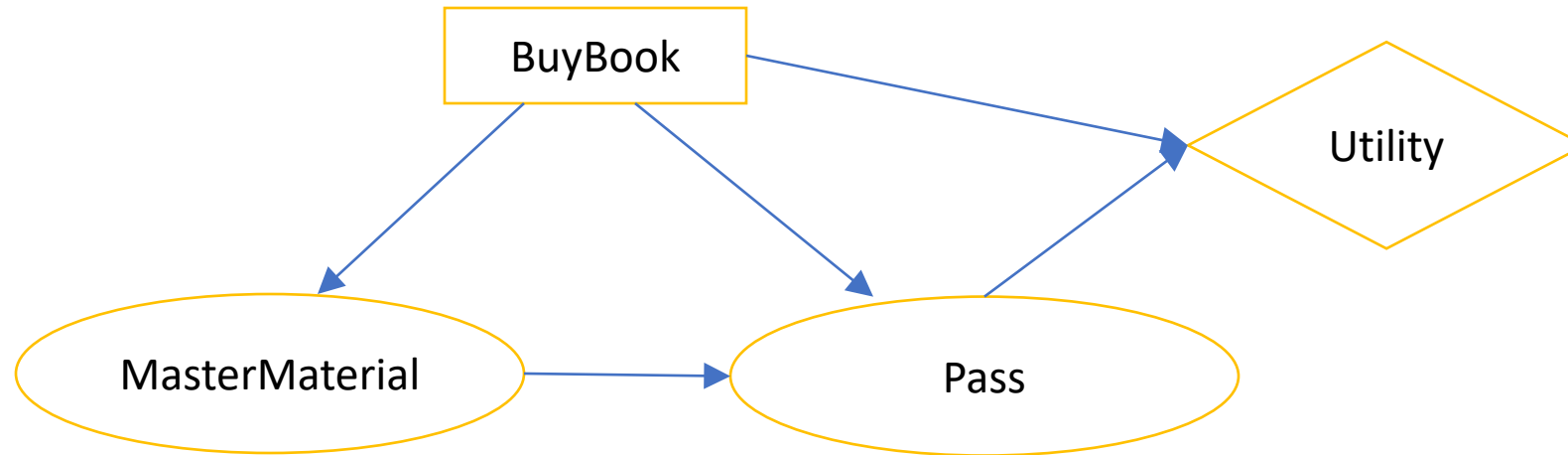
Example 2

- A certain student, Sam, has an additive utility function: 0 for not buying the book and -$100 for buying it; and $2000 for passing the course and 0 for not passing. Sam's conditional probability estimates are as follows:

P(p|b,m) = 0.9

P(p|b,¬m) = 0.5

P(p|¬b,m) = 0.8

P(p|¬b,¬m) = 0.3

P(m|b) = 0.9

P(m|¬b) = 0.7

Example 2

- b. Compute the expected utility of buying the book.

P(p|b) = P(p|b,m)*P(m|b) + P(p|b,~m)*P(~m|b) = 0.9*0.9+0.5*0.1 = 0.81+0.05 = 0.86

P(~p|b) = P(~p|b,m)* P(m|b) + P(p|b,~m)*P(~m|b) = 0.1*0.9+0.5*0.1 = 0.09+0.05 = 0.14

E(b) = P(p|b)*U(p) + P(~p|b)*U(~p) + U(b) = 0.86*2000 + 0 -100 = 1620

Example 2

- b. Compute the expected utility of not buying the book.

P(p|~b) = P(p|~b,m)*P(m|~b) + P(p|~b,~m)*P(~m|~b) = 0.8*0.7 + 0.3*0.3 = 0.56 + 0.09 = 0.64

E(~b) = 0.64*2000 = 1280

Example 2

- c. What should Sam do?

Buy the book, Sam!

Example 3

- Consider an undiscounted MDP having three states, (1, 2, 3), with rewards −1, −2, 0, respectively. State 3 is a terminal state. In states 1 and 2 there are two possible actions: a and b. The transition model is as follows:
  - In state 1, action a moves the agent to state 2 with probability 0.8 and makes the agent stay put with probability 0.2.
  - In state 2, action a moves the agent to state 1 with probability 0.8 and makes the agent stay put with probability 0.2.
  - In either state 1 or state 2, action b moves the agent to state 3 with probability 0.1 and makes the agent stay put with probability 0.9.

- Answer the following questions:

Example 3

- What can be determined qualitatively about the optimal policy in states 1 and 2?
  - Intuitively the agent wants to get to State 3 as soon as possible, because it will pay a cost for each time step it spends in states 1 and 2. However, the only action that reaches state 3 (action b) succeeds with low probability, so the agent should minimize the cost it incurs while trying to reach the terminal state. This suggests that the agent should definitely try action b in state 1; in state 2, it might be better to try action a to get to state 1 (which is the better place to wait for admission to state 3), rather than aiming directly for state 3. The decision in state 2 involves a numerical tradeoff.

Example 3

- Apply policy iteration, showing each step in full, to determine the optimal policy and the values of states 1 and 2. Assume that the initial policy has action b in both states.

Example 3

The application of policy iteration proceeds in alternating steps of value determination and policy update.

- Initialization: $U \leftarrow \langle -1, -2, 0 \rangle, P \leftarrow \langle b, b \rangle$.
- Value determination:

$u_1 = -1 + 0.1u_3 + 0.9u_1, \quad u_2 = -2 + 0.1u_3 + 0.9u_2, \quad u_3 = 0$

Solve the above to get $u_1 = -10$ and $u_2 = -20$.

- Policy Update:

In State 1: $\sum_j T(1, a, j)u_j = 0.8 \times (-20) + 0.2 \times (-10) = -18,$

while $\sum_j T(1, b, j)u_j = 0.1 \times 0 + 0.9 \times (-10) = -9$

So action b is still preferred for State 1.

Example 3

In State 2: $\sum_j T(1, a, j)u_j = 0.8 \times (-10) + 0.2 \times (-20) = -12$

while $\sum_j T(1, b, j)u_j = 0.1 \times 0 + 0.9 \times (-20) = -18$

So action a is preferred for State 2. It changed from initial policy, so proceed.

- Value determination:

$u_1 = -1 + 0.1u_3 + 0.9u_1, \quad u_2 = -2 + 0.8u_1 + 0.2u_2, \quad u_3 = 0$

Solve the above to get $u_1 = -10$ and $u_2 = -12.5$.

- Policy Update:

In State 1: $\sum_j T(1, a, j)u_j = 0.8 \times (-12.5) + 0.2 \times (-10) = -12,$

while $\sum_j T(1, b, j)u_j = 0.1 \times 0 + 0.9 \times (-10) = -9$

So action b is still preferred for State 1.

Example 3

In State 2: $\sum_j T(1, a, j)u_j = 0.8 \times (-10) + 0.2 \times (-12.5) = -10.5,$

while $\sum_j T(1, b, j)u_j = 0.1 \times 0 + 0.9 \times (-12.5) = \quad -11.25$

So action a is preferred for State 2.

It stays unchanged from the previous iteration, and we terminate.

Note that the resulting policy matches our intuition:

when in state 2, try to move to state 1, and when in state

1, try to move to state 3.