Stanford University

AA228/CS238: Decision Making Under Uncertainty
Fall 2021
Prof. Mykel J. Kochenderfer • Online • email: *mykel@stanford.edu*

QUIZ 2 **Due date: October 30, 2021 (5pm Pacific)**

Quizzes will be taken on Gradescope. You may consult any material (e.g., books, calculators, computer programs, and online resources), but you may not consult other people inside or outside of the class. The quiz is designed to be completed in 60 minutes, but we will grant you 90 minutes total to complete and submit your quiz (including uploading any images, handling any logistical issues, etc.) The timing on Gradescope is a hard cutoff. You can start at 5pm PDT on Thursday. To accommodate those in other timezones and complex working situations, the quizzes will be open until 5pm PDT on Friday. Ed will not allow any public posts during that time. **Out of fairness to all students, only material submitted during the allowed time will be graded.**

Question 1. Suppose we are driving ourselves to the airport to catch an international flight. Twenty minutes into the drive, we have a vague worried feeling that we forgot our passport at home. We have the following three actions:

1. continue driving to the airport,
2. pull over and search our bag for our passport (and going home if it is not there), and
3. head straight home and look for our passport.

We make the following estimates:

Action a	State s	P(on time with passport | s, a)
1	not in bag	0
2	not in bag	0.2
3	not in bag	0.3
1	in bag	1
2	in bag	0.9
3	in bag	0.7

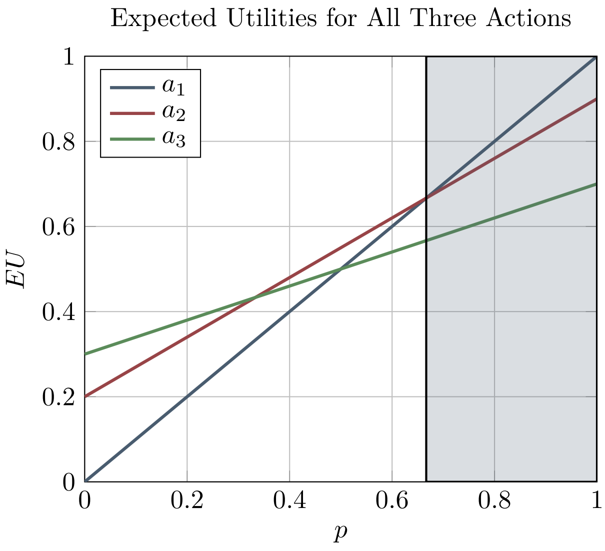
We think that we have our passport in our bag with probability p. For what range of values of p is the best decision to continue to the airport without pulling over or going home first? Assume that our utility depends only on whether we make it to our flight on time with our passport. If we are on time with our passport, our utility is 1. Otherwise, it is 0.

Solution: Compute the expected utility for each action:

$$\begin{aligned}
EU(a_1) &= p(1.0)(1) + (1-p)(0.0)(0) = p \\
EU(a_2) &= p\left[(0.9)(1) + (0.1)(0)\right] + (1-p)\left[(0.2)(1) + (0.8)(0)\right] = 0.7p + 0.2 \\
EU(a_3) &= p\left[(0.7)(1) + (0.3)(0)\right] + (1-p)\left[(0.3)(1) + (0.7)(0)\right] = 0.4p + 0.3
\end{aligned}$$

We are looking for the shaded region of the curve, where the a_1 expected utility dominates:

$$\begin{aligned}
p &> 0.7p + 0.2 \\
0.3p &> 0.2 \\
&\boxed{p > 2/3}
\end{aligned}$$



Question 2. We have a baby and we need to decide whether to sing to them (action a^1) or not (action a^0). The problem is formulated as an infinite horizon discounted MDP with a discount factor of 0.9. There are only two states corresponding to whether the baby is asleep (state s^1) or not (state s^0). We assume the following transition model:

$$T(s^1 \mid s^1, a^1) = 0.9 \tag{1}$$
$$T(s^1 \mid s^0, a^1) = 0.8 \tag{2}$$
$$T(s^1 \mid s^1, a^0) = 0.8 \tag{3}$$
$$T(s^1 \mid s^0, a^0) = 0.1 \tag{4}$$

The immediate reward depends only on whether the baby is asleep. If the baby is asleep, we get a reward of 1; otherwise, we get a reward of 0. What are $U(s^0)$ and $U(s^1)$ for a policy that always has us sing, i.e., chooses action a^1 for both states? What are $U(s^0)$ and $U(s^1)$ for a policy that never has us sing, i.e., chooses action a^0 for both states?

Hint: If you choose to invert a matrix, you can use the following link: https://matrix.reshish.com/inverse.php or any other tool of your choice.

Solution: Use equation (7.10):

$$\mathbf{U}^\pi = (\mathbf{I} - \gamma \mathbf{T}^\pi)^{-1} \mathbf{R}^\pi$$

For always sing:

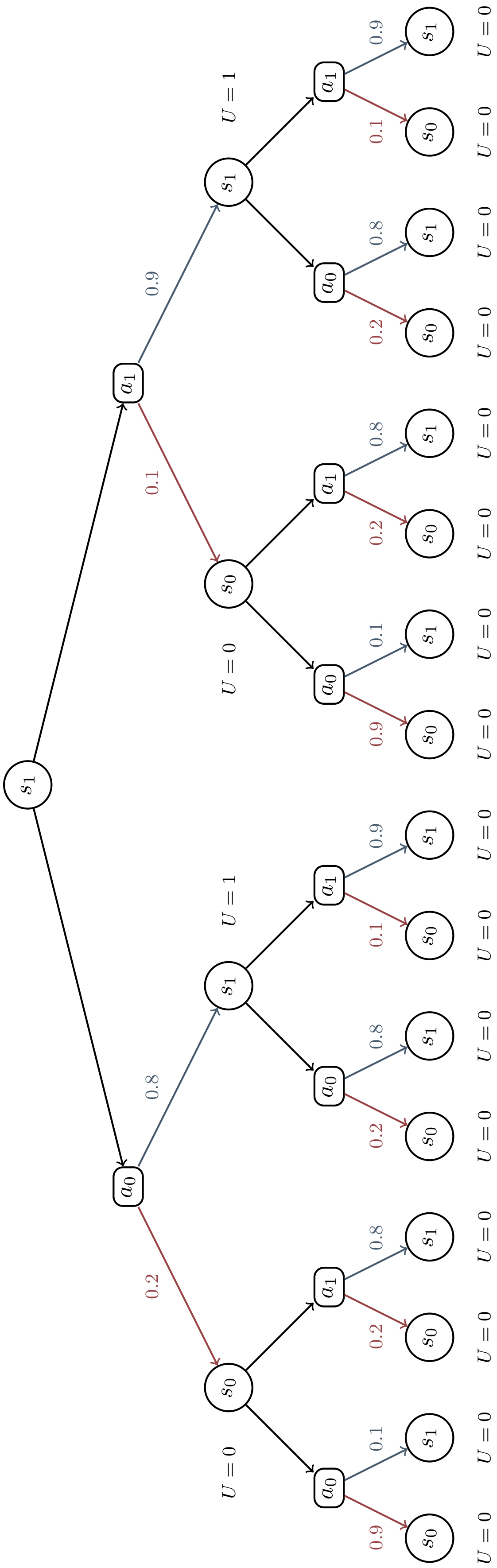
$$\mathbf{U}^\pi = \begin{bmatrix} U(s^0) \\ U(s^1) \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.9 \begin{bmatrix} 0.2 & 0.8 \\ 0.1 & 0.9 \end{bmatrix} \right)^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \boxed{\begin{bmatrix} 7.9121 \\ 9.0110 \end{bmatrix}}$$

For never sing:

$$\mathbf{U}^\pi = \begin{bmatrix} U(s^0) \\ U(s^1) \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.9 \begin{bmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{bmatrix} \right)^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \boxed{\begin{bmatrix} 2.4324 \\ 5.1351 \end{bmatrix}}$$

Question 3. Now use forward search to estimate $U(s^1)$ for the crying baby problem in the previous question. Assume we go to depth 2 and estimate $U(s) = 0$ for all states s at depth 2 (in other words, we assume a horizon of 2).

Solution:


s_1
a_0
a_1
0.2
0.8
0.1
0.9
s_0
s_1
s_0
s_1
$U = 0$
$U = 1$
$U = 0$
$U = 1$
a_0
a_1
a_0
a_1
a_0
a_1
a_0
a_1
0.9
0.1
0.2
0.8
0.2
0.8
0.1
0.9
0.9
0.1
0.2
0.8
0.2
0.8
0.1
0.9
s_0
s_1
s_0
s_1
s_0
s_1
s_0
s_1
s_0
s_1
s_0
s_1
s_0
s_1
s_0
s_1
$U = 0$
$U = 0$
$U = 0$
$U = 0$
$U = 0$
$U = 0$
$U = 0$
$U = 0$
$U = 0$
$U = 0$
$U = 0$
$U = 0$
$U = 0$
$U = 0$
$U = 0$
$U = 0$

We want to find the best action a and its finite horizon expected value U:

$$\begin{aligned} U(s_1) &= \max\left(1 + \gamma\left(0.9(1) + 0\right), 1 + \gamma\left(0.8(1) + 0\right)\right) \\ &= \max\left(1 + (0.9)^2, 1 + (0.9)(0.8)\right) \\ &= \boxed{1.81} \end{aligned}$$

Question 4. We want to use kernel smoothing to estimate the value function from values at m discrete points in the state space. Write the asymptotic time complexity associated with computing $U(s)$ at one particular state s under the assumption that each evaluation of the kernel function requires constant time.

Solution: The linear approximation for kernel smoothing is given by

$$U_{\boldsymbol{\theta}}(s) = \sum_{i=1}^{m} \theta_i \beta_i(s) = \boldsymbol{\theta}^\top \boldsymbol{\beta}(s) \tag{5}$$

where

$$\beta_i(s) = \frac{k\left(s, s_i\right)}{\sum_{j=1}^{m} k\left(s, s_j\right)} \tag{6}$$

Equation 5 is linear in m. The denominator in equation 6 is simply a normalization constant and does not need to be computed on every function call. Thus, the asymptotic time complexity is

$$\boxed{\mathcal{O}\left(m\right)}$$

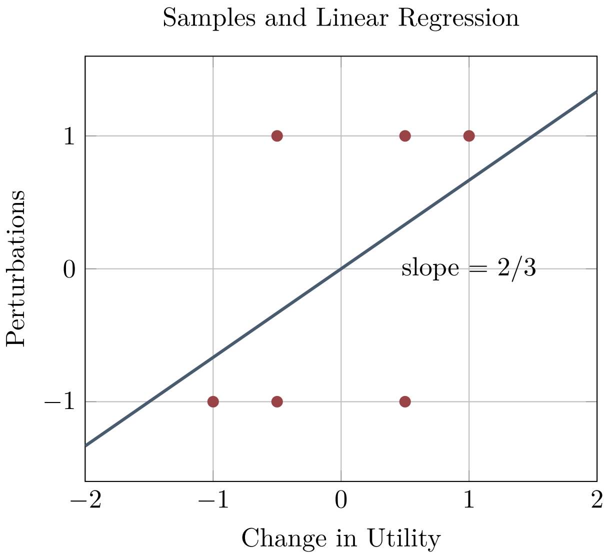
Question 5. Are genetic algorithms guaranteed to find a global optimum in policy space? Why or why not?

Solution: No, genetic algorithms are not guaranteed to find a global optimum in policy space. Although these optimization methods can often avoid becoming stuck in local optima through population recombination, the addition of noisy perturbations means that we might never find a truly global optimum.

Question 6. We have a policy parameterized by a scalar parameter θ. We want to estimate the gradient at $\theta = 5$ using the regression gradient method with a perturbation matrix $\boldsymbol{\Delta\Theta} = [-1, -0.5, -0.5, 0.5, 0.5, 1]$. We do rollouts with these perturbations and get $\boldsymbol{\Delta U} = [-1, -1, 1, 1, -1, 1]$. What is our estimate of the gradient?

Solution: Consider the equation for policy gradient estimation using linear regression:

$$\begin{aligned} \nabla U\left(\boldsymbol{\theta}\right) &\approx \Delta\boldsymbol{\Theta}^{+} \Delta\mathbf{U} \\ &\approx [-1, -0.5, -0.5, 0.5, 0.5, 1]^{+}[-1, -1, 1, 1, -1, 1] \\ &\approx \boxed{0.\overline{66}} \end{aligned}$$



The slope of the least-squares regression line is $\boxed{2/3}$

Question 7. Suppose we estimate the gradient of the value function U to be $[3, 4]$ at $\boldsymbol{\theta} = [1, 3]$. How would the restricted gradient method update $\boldsymbol{\theta}$ if the step length is constrained to 0.5?

Solution: Consider the analytical solution to the restricted gradient update optimization problem:

$$\boldsymbol{\theta}' = \boldsymbol{\theta} + \sqrt{2\epsilon}\frac{\mathbf{u}}{\|\mathbf{u}\|}$$

where $\sqrt{2\epsilon}$ is the step factor. Thus,

$$\begin{aligned}\boldsymbol{\theta}' &= [1, 3] + (0.5)\frac{[3, 4]}{\sqrt{3^2 + 4^2}} \\ &= [1, 3] + (0.5)[0.6, 0.8] \\ &= \boxed{[1.3, 3.4]}\end{aligned}$$