ECE 271C Final Exam

Connor Hughes

May $27 - 28^{th}$, 2021

Problem 1

Consider a problem similar to Bertsekas problem 5.10 (problem 4.9 in Bertsekas' 4^{th} edition), but a simplified version in which there are two possible coins.

- Let p_i denote the probability of heads for coin number i
- Let r denote the probability that the coin tosser is using coin number 1. Assume that r starts off with some initial *a priori* value before the first coin toss.

Now suppose that that coin tosser will *switch* coins after each toss with a probability q.

(a) Find an equation that shows how r evolves as the coin tosses are observed.

(b) Discuss how to use dynamic programming to find an optimal policy.

Solution:

Part (a)

As given in the problem, r represents the probability that the coin tosser is currently using coin number 1. We will expand on this notation and use r_k to represent (our belief about) the probability that the coin tosser will use coin number 1 on the **next** flip after k flips have been observed, with r_0 representing our *a priori* belief about the probability that coin number 1 will be used for the first flip.

Furthermore, we will denote the events:

- "Coin 1 is flipped" by C_1
- "Coin 2 is flipped" by C_2
- "the k^{th} flip comes up Heads" by $Z_k = H$
- "the k^{th} flip comes up Tails" by $Z_k = T$

Then, in the *absence* of any chance of the coin tosser switching the coins, from Bayes' Rule we have:

$$r_{k+1} = \begin{cases} P(C_1|H) = \frac{P(H|C_1)P(C_1)}{P(H)} = \frac{P(H|C_1)P(C_1)}{P(H|C_1)P(C_1)+P(H|C_2)P(C_2)} = \frac{p_1 r_k}{p_1 r_k + p_2(1-r_k)} & \text{if } Z_{k+1} = H \\ P(C_1|T) = \frac{P(T|C_1)P(C_1)}{P(T)} = \frac{P(T|C_1)P(C_1)}{P(T|C_1)P(C_1)+P(T|C_2)P(C_2)} = \frac{(1-p_1)r_k}{(1-p_1)r_k+(1-p_2)(1-r_k)} & \text{if } Z_{k+1} = \text{T} \end{cases}$$

However, in order to take into account the possibility of the coin tosser switching coins after the $(k+1)^{th}$ flip, which occurs with probability q independent of k, we must consider two scenarios when computing r_{k+1}. The first is that the coin tosser used coin 1 on the $(k+1)^{th}$ flip and did *not* switch afterward, and the second is that the coin tosser used coin 2 for the $(k+1)^{th}$ flip then switched.

So, the evolution of our beliefs from r_k to r_{k+1} based on our observation of the $(k+1)^{th}$ coin flip is given by the following expression:

$$r_{k+1} = \begin{cases} (1-q)P(C_1|H) + qP(C_2|H) = (1-q)\frac{p_1 r_k}{p_1 r_k + p_2(1-r_k)} + q\frac{p_2(1-r_k)}{p_1 r_k + p_2(1-r_k)} & \text{if } Z_{k+1} = \text{H} \\ (1-q)P(C_1|T) + qP(C_2|T) = (1-q)\frac{(1-p_1)r_k}{(1-p_1)r_k+(1-p_2)(1-r_k)} + q\frac{(1-p_2)(1-r_k)}{(1-p_1)r_k+(1-p_2)(1-r_k)} & \text{if } Z_{k+1} = \text{T} \end{cases}$$

Part (b)

We can formulate this problem for dynamic programming by defining the following:

State Definition: $x_k \in X_k := \{r_k, EOG\}$
where r_k is our belief about the probability that coin number 1 will be used on the $(k+1)^{th}$ flip, just as in part (a), and EOG is a termination or "end-of-game" state which corresponds to having chosen to stop playing the game.

Control Set: $u_k \in U_k := \{C, Q\}$
where $C =$ "Continue playing" and indicates our choice to wager a bet on the next flip and continue the game, and $Q =$ "Quit" and indicates our choice to exit the game (and go to state EOG).

Disturbance: $w_k \in \{H, T\}$ where w_k indicates the outcome of the $(k+1)^{th}$ flip, H represents heads, and T represents tails.

Stage Reward:

$$g_k(x_k, u_k, w_k) := \begin{cases} 3 & \text{if } u_k = C \text{ and } w_k = T \\ -1 & \text{if } u_k = C \text{ and } w_k = H \\ 0 & \text{if } u_k = Q \end{cases}$$

Terminal Reward: none

State Transition:

$$x_{k+1} = f(x_k, u_k, w_k) := \begin{cases} r_{k+1} & \text{if } u_k = C \\ EOG & \text{if } u_k = Q \text{ or } x_k = EOG \end{cases}$$

where, from part (a) we have:

$$r_{k+1} := \begin{cases} (1-q)\frac{p_1 r_k}{p_1 r_k + p_2(1-r_k)} + q\frac{p_2(1-r_k)}{p_1 r_k + p_2(1-r_k)} & \text{if } w_k = \text{H} \\ (1-q)\frac{(1-p_1)r_k}{(1-p_1)r_k + (1-p_2)(1-r_k)} + q\frac{(1-p_2)(1-r_k)}{(1-p_1)r_k + (1-p_2)(1-r_k)} & \text{if } w_k = \text{T} \end{cases}$$

By definition, this gives the following optimal reward-to-go from state $x_k = r_k$:

$$J_k(x_k) := \max_{u_k \in U_k} \left(\mathop{\mathbb{E}}_{w_k} \left[g_k(x_k, u_k, w_k) + J_{k+1}(f(x_k, u_k, w_k)) \right] \right)$$

Clearly, if we choose to end the game immediately prior to the $(k+1)^{th}$ flip, that is, if $u_k = Q$, then $J_k(x_k) = 0$ because the stage cost g_k will be zero and we will enter the end-of-game state, from which the reward-to-go is zero because future stage rewards will all be zero and there is no terminal cost. So, we can see that the optimal reward-to-go can be rewritten as follows:

$$J_k(x_k) = \max(\mathop{\mathbb{E}}_{w_k} [g_k(x_k, u_k, w_k) + J_{k+1}(f(x_k, u_k, w_k))], 0)$$

which we can expand as below:

$$\begin{aligned} J_k(x_k) &= \max(3P(H) - P(T) + \mathop{\mathbb{E}}_{w_k} [J_{k+1}(f(x_k, u_k, w_k))], 0) \\ &= \max(3P(H) - P(T) + P(H)J_{k+1}(f(x_k, u_k, H)) + P(T)J_{k+1}(f(x_k, u_k, T)), 0) \\ &= \max(P(T)(3 + J_{k+1}(f(x_k, u_k, T))) + P(H)(J_{k+1}(f(x_k, u_k, H)) - 1), 0) \\ &= \max(((1-p_1)r_k + (1-p_2)(1-r_k))(3 + J_{k+1}(f(x_k, u_k, T))) + (p_1 r_k + p_2(1-r_k))(J_{k+1}(f(x_k, u_k, H)) - 1), 0) \end{aligned}$$

We can see that this form of $J_k(x_k)$ reveals the optimal policy for each $x_k = r_k$.
That is, if:

$$0 \leq ((1-p_1)r_k + (1-p_2)(1-r_k))(3 + J_{k+1}(f(x_k, u_k, T))) + (p_1 r_k + p_2(1-r_k))(J_{k+1}(f(x_k, u_k, H)) - 1)$$

then continuing to play the game and wagering a bet on the next coin flip is an optimal control choice. If this inequality is not satisfied, then terminating is optimal.

Since we are allowed to play the game for a maximum of N flips, we can write $J_N(x_N) = 0$ to represent the lack of any terminal reward. Then, from any given state x_k at stage k, we can efficiently solve for the reward-to-go $J_k(x_k)$ using forward dynamic programming and the above recursion, which enables us to evaluate the above inequality and thus determine the optimal policy for the current state.

Problem 2

Consider controlling a scalar linear system with stochastic coefficients:

$$x_{k+1} = a_k x_k + bu$$

where

$$a_k = \begin{cases} 1 & \text{with probability } p_a \\ -1 & \text{with probability (1 - } p_a) \end{cases}$$

and b is constant. At time k, the control can measure x_k, but not a_k.

(a) Use dynamic programming to derive the optimal cost-to-go and optimal policy to minimize:

$$q_N {x_N}^2 + \sum_{k=0}^{N-1} q{x_k}^2 + {u_k}^2$$

(b) Now suppose that the coefficient b is no longer constant, but is also random, where

$$b_k = \begin{cases} 2 & \text{with probability } p_b \\ 0 & \text{with probability (1 - } p_b) \end{cases}$$

(c) Briefly discuss how your solutions would change if the controls were constrained by

$$|u_k| \leq 1$$

(d) Finally, consider again the case where b is constant and the a_k equal either 1 or -1, but now the a_k coefficients are correlated to their past. The new model for a_k is as follows:

$$a_k = a_{k-1} \quad \text{with probability } p$$

At time k, including $k = 0$, the control can measure x_k and a_{k-1}. Derive the optimal cost-to-go and optimal policy.

Solution:

Part (a)

We can formulate this problem for dynamic programming as follows:
State Definition: $x_k \in X_k := \mathbb{R}$

Control Set: $u_k \in U_k := \mathbb{R}$

Disturbance: $a_k \in \{-1, 1\}$

Stage Cost:

$$g_k(x_k, u_k) := qx_k^2 + u_k^2$$

Terminal Cost: $\quad J_N(x_N) := q_N x_N^2$

State Transition:

$$x_{k+1} = f(x_k, u_k, a_k) = a_k x_k + bu_k$$

where

$$a_k = \begin{cases} 1 & \text{with probability } p_a \\ -1 & \text{with probability (1 - } p_a) \end{cases}$$

Then, at stage $N-1$ we can see that

$$\begin{aligned} J_{N-1}(x_{N-1}) &= \min_{u_{N-1}} (qx_{N-1}^2 + u_{N-1}^2 + \mathop{\mathbb{E}}_{a_{N-1}}[J_N(x_N)]) \\ &= \min_{u_{N-1}} (qx_{N-1}^2 + u_{N-1}^2 + q_n(p_a(x_{N-1} + u_{N-1})^2 + (1-p_a)(-x_{N-1} + u_{N-1})^2) \\ &= \min_{u_{N-1}} ((q + q_N)x_{N-1}^2 + 2q_N b(2p_a - 1)(x_{N-1}u_{N-1}) + ((1 + q_N b^2)u_{N-1}^2 \end{aligned}$$

which we can see is quadratic with positive coefficients for the squared terms, as expected, so the value of u_{N-1} which achieves the above minimum is given by:

$$u_{N-1} = \frac{-q_N b(2p_a - 1)x_{N-1}}{1 + q_N b^2}$$

which, when substituted in gives:

$$J_{N-1}(x_{N-1}) = -2\frac{(q_N b(2p_a - 1))^2}{1 + q_N b^2} x_{N-1} + (q + q_N + \frac{(q_N b(2p_a - 1))^2}{1 + q_N b^2})x_{N-1}^2$$

then, leveraging the principle of optimality, we have that at stage $N-2$:

$$\begin{aligned} J_{N-2}(x_{N-2}) &= \min_{u_{N-2}} (qx_{N-2}^2 + u_{N-2}^2 + \mathop{\mathbb{E}}_{a_{N-2}}[J_{N-1}(x_{N-1})]) \\ &= \min_{u_{N-2}} (qx_{N-2}^2 + u_{N-2}^2 + \mathop{\mathbb{E}}_{a_{N-2}}[-2\frac{(q_N b(2p_a - 1))^2}{1 + q_N b^2} x_{N-1} + (q + q_N + \frac{(q_N b(2p_a - 1))^2}{1 + q_N b^2})x_{N-1}^2]) \\ &= \min_{u_{N-2}} (qx_{N-2}^2 + u_{N-2}^2 + p_a[-2\frac{(q_N b(2p_a - 1))^2}{1 + q_N b^2}(x_{N-2} + bu_{N-2}) \\ &\quad + (q + q_N + \frac{(q_N b(2p_a - 1))^2}{1 + q_N b^2})(x_{N-2} + bu_{N-2})^2] \\ &\quad\quad + (1 - p_a)[-2\frac{(q_N b(2p_a - 1))^2}{1 + q_N b^2}(-x_{N-2} + bu_{N-2}) \\ &\quad\quad\quad + (q + q_N + \frac{(q_N b(2p_a - 1))^2}{1 + q_N b^2})(-x_{N-2} + bu_{N-2})^2]) \end{aligned}$$

Then, we can see that the cost to go from stage k is given by:

$$\begin{aligned} J_k(x_k) = \min_{u_k}&((N - k - 1)qx_k^2 + 2qb(2p_a - 1)^k x_k u_k + (1 + (N - k - 2)qb^2)u_k^2 \\ &- 2\frac{(q_N b(2p_a - 1))^2}{1 + q_N b^2} x_k + (q + q_N + \frac{(q_N b(2p_a - 1))^2}{1 + q_N b^2})x_k^2 \end{aligned}$$

Part (b)

In this setting, we can formulate the problem for dynamic programming just as in part (a), but now our state transition is given by:

State Transition:

$$x_{k+1} = f(x_k, u_k, a_k, b_k) = a_k x_k + b_k u_k$$

where

$$a_k = \begin{cases} 1 & \text{with probability } p_a \\ -1 & \text{with probability (1 - } p_a\text{)} \end{cases}$$

and

$$b_k = \begin{cases} 2 & \text{with probability } p_b \\ 0 & \text{with probability (1 - } p_b\text{)} \end{cases}$$

which gives the following cost-to-go from state x_k:

$$J_k(x_k) = \min_{u_k}(g_k(x_k, u_k) + \mathop{\mathbb{E}}_{a_k, b_k}[J_{k+1}(f(x_k, u_k, a_k, b_k))])$$

Part (c)

In the case where our controls are constrained by:

$$|u_k| \leq 1$$

the optimal costs-to-go can only increase or stay the same. For solutions which begin with a very large magnitude of the state x_0 this constraint on our control will likely result in an increase in the optimal cost-to-go, as the optimal sequence of control inputs might likely include stages for which the optimal u_k is greater in magnitude than 1. Prohibiting such controls by imposing the given constraint will result in a decrease in control performance. However, for solutions which begin at states with small magnitudes, it is likely that the optimal solution will not include any control inputs which have a magnitude greater than 1, in which case this constraint will not impact the solution at all.

Part (d)

In the case where $a_k = a_{k-1}$ with probability p, we can formulate the problem for dynamic programming just as before, except that now we have:

State Transition:

$$x_{k+1} = f(x_k, u_k, a_k, b_k) = a_k x_k + b_k u_k$$

where

$$a_k = \begin{cases} a_{k-1} & \text{with probability } p_a \\ -a_{k-1} & \text{with probability (1 - } p_a\text{)} \end{cases}$$

and

$$b_k = \begin{cases} 2 & \text{with probability } p_b \\ 0 & \text{with probability (1 - } p_b\text{)} \end{cases}$$

As before, the cost-to-go from state x_k has the following form:

$$J_k(x_k) = \min_{u_k}(g_k(x_k, u_k) + \mathop{\mathbb{E}}_{a_k, b_k}[J_{k+1}(f(x_k, u_k, a_k, b_k))])$$

Problem 3

Consider a variation of the asset selling problem discussed in class. In this setting, we are able to keep old offers; however, old offers are discounted by a factor of $\delta \in (0, 1)$ each day. At day $k + 1$, we can accept any of the following offers:

$$\{\delta^k w_0,\ \delta^{k-1} w_1,\ ...,\ \delta^1 w_{k-1},\ w_k\}$$

where w_k represents the offer that comes in at the end of day k. Accordingly, the seller does not have access to bid w_{k+1} when making a decision at day $k + 1$. As with the problem in class, we are forced to sell the asset at day N. If we sell the asset at day $k < N$ for price x, we can invest the money and earn a rate of return r and our payoff is

$$(1+r)^{N-k}x$$

where $r > 0$.

(a) Formulate the asset selling problem as a dynamic programming problem.

(b) Assume that bids are uniformly distributed between 0 and 1 for all stages, i.e., bids can take on all real numbers between 0 and 1. Derive the optimal policy at stage $N-1$.

(c) Prove that the optimal policy at stage $N-1$ is a threshold policy. Denote the threshold by α^*_{N-1}.

(d) Consider the time instance $N-2$. Prove that if $x_{N-2} \geq \frac{1}{\delta}\alpha^*_{N-1}$ then the optimal decision is to sell.

(e) What is the optimal decision if $x_{N-2} < \frac{1}{\delta}\alpha^*_{N-1}$?

Solution:

Part (a)

We can formulate this problem for dynamic programming as follows:

State Definition: $x_k := \max(\{\delta^{k-1}w_o, \delta^{k-2}w_1, ..., \delta w_{k-2}, w_{k-1}\})$ or T
where x_k represents the best offer available at day k, and T corresponds to the "terminated" state, wherein the asset has already been sold.

Control Set: $u_k \in U_k := \{S, H\}$
where S = sell the asset for the best price available, and H = hold the asset and wait for a better offer.

Disturbance: $w_k \in \mathbb{R}$
the offer received at the end of day k

Stage Reward:

$$g_k(x_k, u_k) := \begin{cases} (1+r)^{N-k}x_k & \text{if } u_k = S \\ 0 & \text{if } u_k = H \text{ or } x_k = T \end{cases}$$

where p indicates the price for which the asset is sold on day k.

Terminal Reward:

$$J_N(x_N) := \begin{cases} x_N & \text{if } x_k \neq T \\ 0 & \text{if } x_k = T \end{cases}$$

where p indicates the price for which the asset is sold on day N.

State Transition:

$$x_{k+1} = f(x_k, u_k, w_k) := \begin{cases} \max\{\delta x_k, w_k\} & \text{if } u_k = H \\ T & \text{if } u_k = S \end{cases}$$

This formulation gives the following reward-to-go from state x_k:

$$J_k(x_k) = \max_{u_k}(g_k(x_k, u_k) + \mathop{\mathbb{E}}_{w_k}[J_{k+1}(f(x_k, u_k, w_k)])$$

Part (b)

At stage $N-1$ we have the following expression for the optimal reward-to-go:

$$\begin{aligned} J_{N-1}(x_{N-1}) &= \max_{u_{N-1}\in\{S,H\}}(g_{N-1}(x_{N-1},u_{N-1}) + \mathop{\mathbb{E}}_{w_{N-1}}[J_N(f(x_{N-1},u_{N-1},w_{N-1})]) \\ &= \max((1+r)x_{N-1}, \mathop{\mathbb{E}}_{w_{N-1}}(x_N)]) \\ &= \max((1+r)x_{N-1}, \mathop{\mathbb{E}}_{w_{N-1}}(\max(\delta x_{N-1}, w_{N-1}))]) \end{aligned}$$

So, we can see that the optimal policy at stage $N-1$ is to Sell immediately (rather than Hold and Sell on day N), only if

$$(1+r)x_{N-1} > \mathop{\mathbb{E}}_{w_{N-1}}(\max(\delta x_{N-1}, w_{N-1}))]$$

which, assuming that the bids are uniformly distributed between 0 and 1, can be written as

$$\begin{aligned} (1+r)x_{N-1} > \int_0^{\delta x_{N-1}} &\delta x_{N-1}\, dx + \int_{\delta x_{N-1}}^1 x\, dx \\ &= (\delta x_{N-1})^2 + \frac{1}{2} - \frac{(\delta x_{N-1})^2}{2} \\ &= \frac{1}{2} + \frac{(\delta x_{N-1})^2}{2} \end{aligned}$$

or, equivalently

$$0 > \frac{1}{2(1+r)} - (1+r)x_{N-1} + \frac{\delta^2 x_{N-1}^2}{2(1+r)}$$

which we can see is quadratic, and has the following roots:

$$\begin{aligned} x_1, x_2 &= \frac{(1+r) \pm \sqrt{(1+r)^2 - 4(\frac{\delta^2}{4(1+r)^2})}}{\frac{1}{(1+r)}} \\ &= (1+r)^2 \pm \sqrt{(1+r)^4 - \delta^2} \end{aligned}$$

Since $(1+r)^4 > \delta^2$, but $\delta > 0$ and $(1+r)^2 = \sqrt{(1+r)^4}$, we can see that both roots are positive and real. Also, the larger of the two positive roots is clearly greater than 1, which means that x_{N-1} cannot exceed this value. Furthermore, observing the signs of the terms in the inequality on the RHS of the above inequality, we can see that it is negative between the two roots and thus we can see that for all

$$x_{N-1} > (1+r)^2 - \sqrt{(1+r)^4 - \delta^2}$$

the optimal policy at day $N-1$ is to Sell immediately. Conversely, for all

$$x_{N-1} < (1+r)^2 - \sqrt{(1+r)^4 - \delta^2}$$

the optimal policy at day $N-1$ is to Hold. Thus, we have found a threshold.

Part (c)

See derivation in part (b) above, we have already argued the existence of a threshold at day $N-1$, and we have found $\alpha^*_{N-1} = (1+r)^2 - \sqrt{(1+r)^4 - \delta^2}$.

Part (d)

We can express the reward-to-go from stage $N-2$ as follows:

$$J_{N-2}(x_{N-2}) = \max_{u_{N-2}\in\{S,H\}}(g_{N-2}(x_{N-2},u_{N-2}) + \mathop{\mathbb{E}}_{w_{N-2}}[J_{N-1}(f(x_{N-2},u_{N-2},w_{N-2})])$$

Then, we know that if we choose NOT to sell at day $N-2$, that

$$x_{N-1} = \max(\delta x_{N-2}, w_{N-2})$$

and if $x_{N-2} \geq \frac{1}{\delta}\alpha^*_{N-1}$, then we know $x_{N-1} \geq \alpha^*_{N-1}$, and thus by the reasoning given in part (c) we will choose to sell at day $N-1$ if we do not do so at day $N-2$. This means we can express the reward-to-go from day $N-2$ as

$$\begin{aligned} J_{N-2}(x_{N-2}) &= \max((1+r)^2 x_{N-2}, \underset{w_{N-2}}{\mathbb{E}}[(1+r)(f(x_{N-2}, u_{N-2}, w_{N-2})]) \\ &= (1+r)\max((1+r)x_{N-2}, \underset{w_{N-2}}{\mathbb{E}}[(f(x_{N-2}, u_{N-2}, w_{N-2})]) \\ &= (1+r)\max((1+r)x_{N-2}, \underset{w_{N-2}}{\mathbb{E}}[\max(\delta x_{N-2}, w_{N-2})]) \end{aligned}$$

which, since we are assuming that bids are independent and uniformly distributed between 0 and 1 for *all* stages, means that the maximization shown on the RHS of these expressions is exactly the same problem we addressed in part (b). Thus, we will sell at day $N-2$ if $x_{N-2} \geq \alpha^*_{N-1}$. So, since $\frac{1}{\delta}\alpha^*_{N-1} > \alpha^*_{N-1}$, we have shown that if $x_{N-2} \geq \frac{1}{\delta}\alpha^*_{N-1}$ then the optimal policy is to sell at day $N-2$.

Part (e)

While we have proven in part (d) that the optimal policy is to sell if $x_{N-2} \geq \frac{1}{\delta}\alpha^*_{N-1}$, this does not necessarily mean that $\frac{1}{\delta}\alpha^*_{N-1}$ is the threshold for selling/holding at day $N-2$. It is possible that there is a value for x_{N-2} between α^*_{N-1} and $\frac{1}{\delta}\alpha^*_{N-1}$, above which it is optimal to sell and below which it is optimal to hold. Since this problem exhibits the monotonicity property, we know that the threshold at day $N-2$ was between the two values stated here. Unfortunately, I do not have time to derive the exact value of the threshold for day $N-2$.