Exercise Notes - Stochastic Optimisation

Dom Hutchinson

January 26, 2021

Question 1) - Inventory Control

This is from `LectureSlides3cSO.pdf` *and covers chapter 2..* Consider the following *Inventory Control Problem.* Let $n \in \mathbb{N}$ be the maximum number of stored items allowed.

The inventory is controlled over finte time-periods $\{0, \dots, N-1\}$ for $N \geq 2$. At the beginning of each time-period a number of items are demanded by customers and delivered immediately. If more items are demanded than are currently in stock, then the orders are backlogged and satisfied when new items arrive. Backlogged items are satisfied before any new demands. Let m be the maximum number of backlogged items allowed.

Let Z_t denote the number of items demanded by customers at time $t \in T$. We assume Z_t are non-negative IID random variables with $p(k) := \mathbb{P}(Z_t = k)$. Further, we assume Z_t is independent from the number of items stored at the beginning of period $t = 0$ and independent from the number of items currently backlogged.

Let $c(k)$ denote the cost of ordering k new items, $\alpha(k)$ be the cost of storing k un-sold items, $\beta(k)$ be the penalty for having k items in the backlog. If the demand for k items is completely lost then $\gamma(k)$ is paid. We assume $c(k), \alpha(k), \beta(k), \gamma(k)$ are non-negative real numbers for $k \geq 1$ and 0 when $k = 0$.

Our problem, is to determine, at the beginning of each time-period, how many new items to order whilst minimising total cost.

Question 1) (a)

Formulate the described inventory control problem as a finite-horizon *Markov Decision Problem*

Answer 1) (a)

- *Stochsatic System* - Inventory and customers.
- *Agent* - Inventory Manager
- *Decision Epochs* - The start of each time period.
- *Time Horizon* - $T = \{0, \dots, N-1\}$.
- *System States*[1]

  Let X_t denote the system state at epoch t, X_t' be the number of items in the inventory at the beginning of period t and X_t'' be the number of backlogged items a the beginning of period t. We have that if $X_t'' > 0 \implies 0$ and $X_t' > 0 \implies X_t'' = 0$. We want to come up

[1] System states are an encoding of available system information, which is relevant to the selection of Y_t.

with a formulation for X_t, in terms of X_t' and X_t'', which carries all sufficient information from X_t', X_t'' to make a prediction of Y_t given X_t.

$$X_t = \begin{cases} X_t' & \text{if } X_t'' = 0 \\ -X_t'' & \text{if } X_t'' > 0 \end{cases}$$

This means that

$$\begin{aligned} X_t \geq 0 \quad &\implies \quad X_t' = X_t, X_t'' = 0 \\ X_t < 0 \quad &\implies \quad X_t' = 0, X_t'' = -X_t \end{aligned}$$

This shows that X_t' and X_t'' can be uniquely retrieved from X_t.

Since n is the inventory capacity, $X_t' \leq n \implies X_t \leq n$. And, since m is the backlog capacity, $X_t'' \leq m \implies X_t \geq -X_t'' \geq -m$. This means the state space $S = \{-m, \dots, 0, \dots, n\}$.

- *Agent Actions.*

  Let $Y - t$ denote the action the agent takes at epoch t. This is the number of new items ordered at the beginning of time-period t.

  If $X_t + Y_t > 0$ then $X_t + Y_t$ is the number of items stored in the inventory after the arrival of newly ordered items. Since n is the inventory capacity, $X_t + Y_t \leq n$. As $X_t \geq -m$ (See *System States*) we have

  $$Y_t \leq n - X_t \leq n + m$$

  This means the *Action-Space* is $A = \{0, \dots, n + m\}$.

  If $X_t = s$ we have

  $$Y_t \leq n - X_t = n - s$$

  This means the *Admissible Actions* in state s is $A(s) = \{0, \dots, n - s\}$.

- *State Dynamics & Immediate Costs.* The number of items in the invetory at the end of time-period t is $X_t + Y_t - Z_t$. We have three cases

  - $X_t + Y_t - Z_t \geq 0$ - *There is surplus in the inventory.*
    This means there are currently $X_{t+1}' = X_t + Y_t - Z_t$ items in storage. Further, there is no excess demand in period t, meaning the number of backlogged items at the start of the next period is $X_{t+1}'' = 0$. No demand was lost in time-period t. We only pay the penalty $\alpha(X_t + Y_t - Z_t)$ for having items in storage. We also pay $c(Y_t)$ for ordering new items. The total cost incurred in time-period t is

    $$c(Y_t) + \alpha(X_t + Y_t - Z_t)$$

    Since $X_{t+1}'' = 0$, by our definition of X_t, we have

    $$X_{t+1} = X_{t+1}' = X_t + Y_t - Z_t$$

  - $X_t + Y_t - Z_t \in [-m, 0)$ - *There is a backlog but within capacity.*
    There are no items in the inventory at the end of period t, hence $X_{t+1}' = 0'$. So no penalty is paid for storing stock.
    Since $X_t + Y_t - Z_t < 0$ the excess demand in period t is $-(X_t + Y - t - Z - t) \leq m$, this is within backlog capacity so no demand is lost. Hence the number of backlogged item at the end of time-period t is $X_{t+1}'' = -(X_t + Y_t - Z_t)$. We only pay a backlog penalty of $\beta(-(X_t + Y_t - Z_t))$ and a fee of $c(Y_t)$ for the new item. The total cost in period t is

    $$c(Y_t) + \beta(X_t + Y_t - Z_t)$$

    Since $X_{t+1}'' = -(X_t + Y_t - Z_t) > 0$, by our definition of X_t, we have

    $$X_{t+1} = -X_{t+1}'' = X_t + Y_t - Z_t$$

– $X_t + Y_t - Z_t < -m$ - *There is a backlog and some demand is lost.*
There are no items in the inventory at the end of period t, hence $X'_{t+1} = 0'$. So no penalty is paid for storing stock.
Since $X_t + Y_t - Z_t < 0$ the excess demand in period t is $-(X_t + Y - t - Z - t) > m$, this is beyond backlog capacity so some demand is lost. Hence the number of backlogged item at the end of time-period t is $X''_{t+1} = m$ and lost demand in period t is $-(X_t + Y_t - Z - t) - m$.
We pay a backlog penalty of $\beta(m)$, a lost demand penalty of $\gamma(-(X_t + Y_t - Z_t) - m)$ and pay $c(Y_t)$ to purchase new items. The total cost of time-period t is

$$c(Y_t) + \beta(m) + \gamma(-(X_t + Y_t - Z_t) - m)$$

As $X''_{t+1} = m > 0$, by our definition of X_t, we have

$$X_{t+1} = -X''_{t+1} = -m$$

By combining these cases we have the state equations

$$X_{t+1} = \begin{cases} X_t + Y_t - Z_t & \text{if } X_t + Y_t - Z_t \geq -m \\ -m & \text{if } X_t + Y_t - Z_t < -m \end{cases} = \max\{-m,\ X_t + Y_t - Z_t\}$$

We have the cost incurrent in each period

$$G_t = g_t(X_t, Y_t, Z_t) = \begin{cases} c(Y_t) + \alpha(X_t + Y_t - Z_t) & \text{if } (X_t + Y_t - Z_t) \geq 0 \\ c(Y_t) + \beta(Z_t - X_t - Y_t) & \text{if } (X_t + Y_t - Z_t) \in [-m, 0) \\ c(Y_t) + \beta(m) + \gamma(Z_t - X_t - Y_t - m) & \text{if } (X_t + Y_t - Z_t) < -m \end{cases}$$

- *Transition Probabilities.*

The definition of transition probabilities is

$$p_t(s'|s, a) = \mathbb{P}^\pi(X_{t+1} = s'|X_t = s, Y_t = a)$$

We can substitute in the state-equation

$$\begin{aligned} p_t(s'|s, a) &= \mathbb{P}^\pi(\max\{-m,\ X_t + Y_t - Z_t\} = s'|X_t = s, Y_t = a) \\ &= \mathbb{P}^\pi(\max\{-m,\ s + a - Z_t\} = s'|X_t = s, Y_t = a) \end{aligned}$$

X_t, Y_t both represent values at the beginning of time-period t and thus independent of Z_t, but dependent upon $Z_0, \ldots, Z_{t-1}$. As Z_t is independent of $Z_0, \ldots, Z_{t-1}$ (from question) we conclude that Z_t is independent of X_t and Y_t. This means the following events are independent

$$\big\{\max\{s + a - Z_t, -m\} = s'\big\} \quad \big\{X_t = s, Y_t = a\big\}$$

Therefore,

$$p_t(s'|s, a) = \mathbb{P}^\pi(\max\{-m,\ s + a - Z_t\} = s')$$

As demand Z_t is independent of our policy π, we further have

$$p_t(s'|s, a) = \mathbb{P}(\max\{-m,\ s + a - Z_t\} = s')$$

We consider three cases

– $s' > s + a$.

We need to compute $p_t(s'|s,a) = \mathbb{P}(\max\{-m,\ s+a-Z_t\} = s')$.

We want to determine the probability of $s' = \max\{-m,\ s+a-Z_t\}$. We have $s + a - Z_t \leq s + a$, as $Z_t \geq 0$ by the question, consequently $\max\{s+a-Z_t, -m\} \leq \max\{s+a, -m\}$.

We have $s + a \geq -m$, as $s \in \{-m, \dots, n\}$ and $a \in \{0, \dots, n-s\}$. Consequently,

$$\max\{s+a, -m\} = s + a$$

Combining these we get

$$\begin{aligned} s+a &< s' = \max\{s+a-Z_t, -m\} \\ &\leq \max\{s+a, -m\} \\ &= s+a \end{aligned}$$

Hence, it is impossible for $s' = \max\{-m,\ s+a-Z_t\}$. Thus

$$p_t(s'|s,a) = \mathbb{P}(\max\{-m,\ s+a-Z_t\} = s') = 0$$

– $s' \in (-m, s+a]$.

We need to compute $p_t(s'|s,a) = \mathbb{P}(\max\{-m,\ s+a-Z_t\} = s')$.

Again, we want to determine the probability of $s' = \max\{-m,\ s+a-Z_t\}$. As $s' > -m$ we have

$$\max\{s+a-Z_t, -m\} = s+a-Z_t \implies s' = s+a-Z_t$$

Hence, the result only holds if $Z_t = s + a - s'$. This gives transition probability

$$\begin{aligned} p_t(s'|s,a) &= \mathbb{P}(\max\{s+a-Z_t, -m\} = s') \\ &= \mathbb{P}(Z_t = s+a-s') \\ &= p(s+a-s') \end{aligned}$$

– $s' = -m$.

We need to compute $p_t(s'|s,a) = \mathbb{P}(\max\{-m,\ s+a-Z_t\} = s')$.

Again, we want to determine the probability of $s' = \max\{-m,\ s+a-Z_t\}$. As $s' = -m$ we have

$$s + a - Z_t \leq -m$$

Hence, the result holds iff $Z_t \geq s + a + m$. This gives transition probability

$$\begin{aligned} p_t(s'|s,a) &= \mathbb{P}(\max(s+a-Z-t, -m) = s') \\ &= \mathbb{P}(Z_t \geq s+a+m) \\ &= \sum_{k=s+a+m}^{\infty} \mathbb{P}(Z_t = k) \\ &= \sum_{k=s+a+m}^{\infty} p(Z_t = k) \end{aligned}$$

- *Equivalent Rewards.*

We want to select a policy $\pi \in HR(T)$ which minimises the expected total cost

$$\mathbb{E}^{\pi}\left[\sum_{t=0}^{N-1} G_t\right] = \mathbb{E}^{\pi}\left[\sum_{t=0}^{N-1} g_t(X_t, Y_t, Z_t)\right]$$

This expectation depends on the demand $Z_0, \dots, Z_{N-1}$, thus minimising it does not fit the framework of a *Markov Decision Problem*. Therefore we transform the expected total cost to an equivalent (but more convenient) form

$$\begin{aligned}\mathbb{E}^\pi\left[\sum_{t=0}^{N-1} G_t\right] &= \sum_{t=0}^{N-1} \mathbb{E}^\pi[G_t] \\ &= \sum_{t=0}^{N-1} \mathbb{E}^\pi\big[\mathbb{E}^\pi[G_t|X_t, Y_t]\big] \text{ by Tower property} \\ &= \sum_{t=0}^{N-1} \mathbb{E}^\pi\big[\mathbb{E}^\pi[g_t(X_t, Y_t, Z_t)|X_t, Y_t]\big]\end{aligned}$$

Define $r_t(s, a)$ and $r_N(s)$ to be

$$\begin{aligned}r_t(s, a) &= -\mathbb{E}^\pi[g_t(X_t, Y_t, Z_t)|X_t = s, Y_t = a] \\ r_N(s) &= 0\end{aligned}$$

Since Z_t is independent of (X_t, Y_t) we have

$$r_t(s, a) = -\mathbb{E}^\pi[g_t(s, a, Z_t)|X_t = s, Y_t = a] = -\mathbb{E}^\pi[g_t(s, a, Z_t)]$$

Substituting these definitions into our formulation for total expected cost yields

$$\mathbb{E}^\pi\left[\sum_{t=0}^{N-1} G_t\right] = -\mathbb{E}^\pi\left[\sum_{t=0}^{N-1} r_t(X_t, Y_t) + r_N(X_N)\right]$$

Minimising this expression is equivalent to maximising the following

$$\mathbb{E}^\pi\left[\sum_{t=0}^{N-1} r_t(X_t, Y_t) + r_N(X_N)\right]$$

This is congrent with the framework of a *Markov Decision Problem*. $r_t(s, a)$ can be viewed as the equivalent reward at epoch t and $r_N(s)$ as the equivalent terminal reward. Since demand Z_t is not affected by our policy π we have

$$r_t(s, a) = -\mathbb{E}[g_t(s, a, Z_t)] = -\sum_{k=0}^{\infty} g_t(s, a, k)p(k)$$

Further, by considering the full breakdown of the cost function we have

$$\begin{aligned}r_t(s, a) &= -c(a) - \sum_{k=0}^{s+a} \alpha(s + a - k)p(k) - \sum_{k=s+a+1}^{s+a+m} \beta(k - s - a)p(k) \\ &- \sum_{k=s+a+m+1}^{\infty} \big(\beta(m) + \gamma(k - s - a - m)\big)p(k)\end{aligned}$$

- *Equivalent Objective.*

  Find a policy $\pi \in HR(T)$ which maximises

$$\mathbb{E}^\pi\left[\sum_{t=0}^{N-1} r_t(X_t, Y_t) + r_N(X_N)\right]$$

Question 1) (b)

By considering the markov decision problem formulated in `2) (a)` and assume the following

- $N = 2, n = 1, m = 1$.
- $p(0) = p(1) = p(2) = p(3) = .25$ and $p(k) = 0$ if $k \geq 0$.
- $c(k) = ck,\ \alpha(k) = \alpha k,\ \beta(k) = \beta k,\ \gamma(k) = \gamma k$ for $k \geq 0$.
- $c = 1,\ \alpha = 2,\ \beta = 3,\ \gamma = 4$.

Find an optimal policy π^*

Answer 1) (b)
From `3) (a)` we can quickly derive this formulation by substituting in the values specified.

- *Number of Epochs* - $N = 2$.
- *Time-Horizon* - $T = \{0, 1\}$.
- *State-Space* - $S = \{-1, 0, 1\}$.
- *Action-Space* - $A = \{0, 1, 2\}$.
- *Admissible Action-Space*

$$\begin{aligned} A(-1) &= \{0, 1, 2\} \\ A(0) &= \{0, 1\} \\ A(1) &= \{0\} \end{aligned}$$

- *Rewards*

$r_t(s, a) =$

$s \backslash a$	0	1	2
-1	-9	-25/4	-5
0	-21/4	-5	ND
1	-3	ND	ND

ND denotes not-defined, since $r_t(s, a)$ is not defined if $a \notin A(s)$.

- *Terminal Reward* $r_2(s) = 0$.
- *Transition Probabilities*

$p_t(s'|s, 0) =$

$s \backslash s'$	-1	0	1
-1	1	0	0
0	3/4	1/4	0
1	1/2	1/4	1/4

$p_t(s'|s, 1)$

$s \backslash s'$	-1	0	1
-1	3/4	1/4	0
0	1/2	1/4	1/4
1	ND	ND	ND

$p_t(s'|s, 2)$

$s \backslash s'$	-1	0	1
-1	1/2	1/4	1/4
0	ND	ND	ND
1	ND	ND	ND

To find the optimal policy we use the *Dynamic Programming Algorithm* which is defined as

$$\begin{aligned} u_t^*(s) &= \max_{a \in A(s)} \left(r_t(s, a) + \sum_{s' \in S} u_{t+1}^*(s') p_t(s'|s, a) \right) \\ d_t^*(s) &\in \text{argmax}_{a \in A(s)} \left(r_t(s, a) + \sum_{s' \in S} u_{t+1}^*(s') p_t(s'|s, a) \right) \end{aligned}$$

where $t = N-1, \dots, 0$ and $u_N^*(s) = r_N(s)$. For simplicity I will use the following to denote the expression we are maximising

$$w_t^*(s,a) := r_t(s,a) + \sum_{s' \in S} u_{t+1}^*(s') p_t(s'|s,a)$$

For this specific scenario we have two epochs to consider

- Epoch $t = 1$

  We need to compute $u_1^*(s), d_1^*(s)$. We have

  $$w_1^*(s,a) = r_1(s,a) + r_2(-1)p_1(-1|s,a) + r_2(0)p_1(0|s,a) + r_2(1)p_1(1|s,a)$$

  $w_1^*(s,a) =$

$s \backslash a$	0	1	2
-1	-9	-25/4	-5
0	-21/4	-4	ND
1	-3	ND	ND

  This gives the following tables of results The optimal strategy depends on what state the system is in.

s	$u_1^*(s)$	$d_1^*(s)$
-1	-5	2
0	-4	1
1	-3	0

- Epoch $t = 0$

  We need to compute $u_0^*(s), d_0^*(s)$. We have

  $$w_0^*(s,a) = r_0(s,a) + u_1^*(-1)p_0(-1|s,a) + u_1^*(0)p_0(0|s,a) + u_1^*(1)p_0(1|s,a)$$

  $w_0^*(s,a) =$

$s \backslash a$	0	1	2
-1	-14	-11	-37/4
0	-10	-33/4	ND
1	-29/4	ND	ND

  This gives the following tables of results The optimal strategy depends on what state the system is in.

s	$u_0^*(s)$	$d_0^*(s)$
-1	-37/4	2
0	-33/4	1
1	-29/4	0

- *Optimal Policy* - $\pi^* = (d_0^*(s), d_1^*(s))$ with $d_0^*(s), d_1^*(s)$ as defined above.
- *Optimal Value Function* - $u_0^*(s)$ as defined above.

Question 2) - Discount Reward Inventory Problem
This is from `LectureSlides4cSO.pdf` *and covers chapter 3..*
This is an *Inventory Control Problem* with discounted-cost (See `Question 1)` for the finite-time version).

Our problem is to decide, at the beginning of each time-period, what number of items to order so the expected discounted cost is minimal.

Question 2) a)
Formulate the inventory control problem as a *Discounted Reward Markov Decision Problem.*

Answer 2) a)

- *Stochastic System* - Inventory and customers.
- *Agent* - Inventory Manager.
- *Decision Epochs* - Beginning of the time-period.
- *Epoch t* - The beginning of time-period t.
- *Time Horizon* - $T = \{0, 1, \dots\}$.
- *System States.*

  Let X_t be the system state at epoch t, X_t' be the number of items in the inventory at the beginning of period t and X_t'' be the number of backlogged items at the beginning of period t.
$$X_t' := \begin{cases} X_t' & \text{if } X_t'' = 0 \\ -X_t'' \text{if } X_t'' > 0 \end{cases}$$
- *State-Space* - $T = \{-m, \dots, 0, \dots, n\}$.
- *Agent-Actions.*

  Let Y_t be the agent action at epoch t. Define Y_t as the number of items ordered at the beginning of time-period t.
- *Action-Space* - $A = \{0, \dots, n+m\}$.
- *Admissible Actions* - $A(s) = \{0, \dots, n-s\}$.
- *Transition Probabilities.*

  Let $s \in S,\ a \in A(s)$. The corresponding transition probabilities are
$$p(s'|s,a) \begin{cases} 0 & \text{if } s' > s+a \\ p(s+a-s') & \text{if } s' \in (-m, s+a] \\ \sum_{k=s+a+m}^{\infty} p(k) & \text{if } s' = m \end{cases}$$
- *Equivalent Rewards.*

  Let $s \in S,\ a \in A(s)$. The corresponding equivalent reward is
$$\begin{aligned} r(s,a) &= = c(a) + \sum_{k=0}^{s+a} \alpha(s+a-k)p(k) - \sum_{k=s+a+1}^{s+a+m} \beta(k-s-a)p(k) \\ &- \sum_{k=s+a+m+1}^{\infty} \left[\beta(m) + \gamma(k-s-a-m)\right] p(k) \end{aligned}$$

- *Equivalent Objective.*

  Find the policy, $\pi \in HR(T)$ which maximises the expected discount reward

$$\mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty}\alpha^t r(X_t, Y_t)\right] \quad \alpha \in (0,1)$$

Question 2) b)
Consider the MDP formulated in `4) a)`. Assume the following

- $n = 1, m = 1, \alpha = .5$.
- $p(0) = p(1) = p(2) = p(3) = \frac{1}{4}$ and $p(k) = 0 \ \forall \ k \geq 4$.
- $c(k) = ck, \alpha(k) = \alpha k, \beta(k) = \beta(k), \gamma(k) = \gamma k \ \forall \ k \geq 0$.
- $c = 1, \alpha = 2, \beta = 3, \gamma = 4$.

Using the policy iteration algorithm, find an algorithm policy.

Answer 2) b)
For these specific conditions we have the following

- *Number of Epochs* - $N = \infty$.
- *Time-Horizon* - $T = \{0, 1, \dots\}$.
- *State-Space* - $S = \{-1, 0, 1\}$.
- *Action-Space* - $A = \{0, 1, 2\}$.
- *Admissible Actions*

$$\begin{aligned} A(-1) &= \{0,1,2\} \\ A(0) &= \{0,1\} \\ A(1) &= \{0\} \end{aligned}$$

- *Rewards*

$r(s,a) =$

$s \backslash a$	0	1	2
-1	-9	-25/4	-5
0	-21/5	-4	ND
1	-3	ND	ND

- *Transition Probabilties*

$p(s'|s,0) =$

$s \backslash s'$	-1	0	1
-1	1	0	0
0	3/4	1/4	0
1	1/2	1/4	1/4

$p(s'|s,1) =$

$s \backslash s'$	-1	0	1
-1	3/4	1/4	0
0	1/2	1/4	1/4
1	ND	ND	ND

$p(s'|s,2) =$

$s \backslash s'$	-1	0	1
-1	1/2	1/4	1/4
0	ND	ND	ND
1	ND	ND	ND

- *Policy Evaluation Step*

  In the k^{th} iteration, we compute value function $v_k(s)$ where $v_k(s)$ is the solution to the following equations

$$\begin{aligned} v(s) &= (T_d v)(s) \\ &= r(s, d_k(s)) + \alpha \sum_{s' \in S} v(s') p(s'|s, d_k(s)) \end{aligned}$$

- *Policy Improvement Step*

  In the k^{th} iteration, we compute the decision function $d_{k+1}(s)$

$$d_{k+1}(s) \in \text{argmax}_{a \in A(s)} \left(r(s,a) + \alpha \sum_{s' \in S} v_k(s') p(s'|s,a) \right)$$

- *Alternative Form of Policy Improvement Step*

$$\begin{aligned} w_{k+1}(s,a) &= r(s,a) + \alpha \sum_{s' \in S} v_k(s') p(s'|s,a) \\ d_{k+1}(s) &\in \text{argmax}_{a \in A(s)} w_{k+1}(s,a) \end{aligned}$$

- *Initialisation.*

  We set

$$\begin{aligned} d_0(-1) &= 2 \\ d_0(0) &= 1 \\ d_0(1) &= 0 \end{aligned}$$

  $d_0(s)$ can be any *Markovian Decision Function* which satisfies $d_0(s) \in A(s) \ \forall \ s \in S$.

- *Iteration - $k = 0$.*

  Policy Evaluation

  We compute solution $v_0(s)$ for the following system of equations

$$\begin{aligned} v(s) &= (T_{d_0} v)(s) \\ &= r(s, d_0(s)) + \alpha \sum_{s'=-1}^{1} v(s') p(s'|s, d_0(s)) \end{aligned}$$

  where $v(s)$ is unknown and $s \in \{-1, 0, 1\} = S$. We can expand this equation as

$$\begin{aligned} v(-1) &= r(-1, d_0(-1)) + \alpha \sum_{s'=-1}^{1} v(s') p(s'|-1, d_0(-1)) \\ v(0) &= r(0, d_0(0)) + \alpha \sum_{s'=-1}^{1} v(s') p(s'|0, d_0(0)) \\ v(1) &= r(1, d_0(1)) + \alpha \sum_{s'=-1}^{1} v(s') p(s'|1, d_0(1)) \end{aligned}$$

  This has solutions

$$\begin{aligned} v_0(-1) &= -37/4 \\ v_0(0) &= -33/4 \\ v_0(1) &= -29/4 \end{aligned}$$

  Policy Improvement

We compute $d_1(s)$ using the following system of equations

$$\begin{aligned} w_1(s,a) &= r(s,a) + \alpha \sum_{s'=-1}^{1} v_0(s')p(s'|s,a) \\ d_1(s) &\in \text{argmax}_{a \in A(s)} w_1(s,a) \end{aligned}$$

The table below summarises the values for these equations

$w_1(s,a) =$

$s \backslash a$	0	1	2
-1	-109/8	-43/4	-37/4
0	-39/4	-33/4	ND
1	-29	ND	ND

$d_1(s) =$

s	$d_1(s)$
-1	2
0	1
1	0

Stopping Criterion - As $d_1(s) = d_0(s)\ \forall\ s \in S$, then $d_1(s)$ is optimal.

- *Optimal Solution.*

  The *Optimal Markovian Decision Function* is $d_0(s)$ and the *Optimal Value Function* is $v_0(s)$.

Question 2) c)
Formulate a linear problem equivalent to the MDP solved in `4) b)`.

Answer 2) c)
The *Equivalent Linear Program* is to minimise

$$\sum_{s'=-1}^{s} \gamma(s')v(s') \quad \text{wrt } v(-1), v(0), v(1)$$

subject to the condition that

$$r(s,a) + \alpha \sum_{s'=-1}^{1} v(s')p(s'|s,a) \leq v(s)$$

where $\gamma(-1), \gamma(0), \gamma(1) \in (0, \infty)$ and are constants.
This condition can be expanded and restated as

$$\begin{aligned} r(-1,0) + \alpha \sum_{s'=1}^{1} v(s')p(s'|-1,0) &\leq v(-1) \\ r(-1,1) + \alpha \sum_{s'=1}^{1} v(s')p(s'|-1,1) &\leq v(-1) \\ r(-1,2) + \alpha \sum_{s'=1}^{1} v(s')p(s'|-1,2) &\leq v(-1) \\ r(0,0) + \alpha \sum_{s'=1}^{1} v(s')p(s'|0,0) &\leq v(0) \\ r(0,1) + \alpha \sum_{s'=1}^{1} v(s')p(s'|0,1) &\leq v(0) \\ r(1,0) + \alpha \sum_{s'=1}^{1} v(s')p(s'|1,0) &\leq v(0) \end{aligned}$$

Question 3) - Average Reward Inventory Problem
This is from `LectureSlides5cSO.pdf` *and covers chapter 4..*
This is an *Inventory Control Problem* with discounted-cost (See `Question 1)` for the finite-time version).

Our problem is to decide, at the beginning of each time-period, what number of items to order so the expected average cost is minimal.

Question 3) (a)
Formulate the considered inventory control problem as an *Average Reward MDP*

Answer 3) (a)

- *Decision Epochs* - Beginning of each time-period.
- *Time-Horizon* - $T = \{0, 1, \dots\}$.
- *System States.*

  Let X_t be the system state in epoch t; X'_t be the number of items kept in the inventory at the beginning of period t; X''_t be the number of backlogged items at the beginning of period t.

  $$X_t := \begin{cases} X'_t & \text{if } X''_t = 0 \\ -X''_t & \text{if } X''_t > 0 \end{cases}$$

- *State-Space* - $S = \{-m, \dots, 0, n\}$.
- *Agent Actions.*

  Let Y_t be the number of items ordered at the beginning of time-period t.

- *Action-Space* - $A = \{0, \dots, n + m\}$.
- *Admissible Actions* - $A(s) = \{0, \dots, n - s\}$.
- *Transition Probabilities.*

  $$p(s'|s, a) = \begin{cases} 0 & \text{if } s' > s + a \\ p(s + a - s') & \text{if } s' \in (-m, s + a] \\ \sum_{k=s+a+m}^{\infty} p(k) & \text{if } s' = m \end{cases}$$

- *Equivalent Rewards.*

  $$\begin{aligned} r(s, a) &= -c(a) - \sum_{k=0}^{s+a} \alpha(s + a - k)p(k) \\ &- \sum_{k=s+a+1}^{s+a+m} \beta(k - s - a)p(k) \\ &- \sum_{k=s+a+m+1}^{\infty} \big(\beta(m) + \gamma(k - s - a - m)\big)p(k) \end{aligned}$$

- *Equivalent Objective.*

Find $\pi \in HR(T)$ st

$$\pi = \text{argmax}_\pi \lim_{N\to\infty} \inf \mathbb{E}^\pi \left[\frac{1}{N} \sum_{t=0}^{N-1} r(X_t, Y_t) \right]$$

Question 3) (b)
Consider the *MDP* formulated in `3) (a)`. Assume the following

- $n = 1,\ m = 1,\ \alpha = 0.5.$
- $p(k) = \begin{cases} \frac{1}{4} & \text{if } k \in [1,4] \\ 0 & \text{otherwise} \end{cases}.$
- $c(k) = k,\ \alpha(k) = 2k,\ \beta(k) = 3k,\ \gamma(k) = 4k$ for $k \geq 0.$

Use the *Policy Iteration Algorithm* to find an optimal policy

Answer 3) (b)

- *Time-Horizon* - $T = \{0, 1, \dots\}$.
- *State-Space* - $S = \{-1, 0, 1\}$.
- *Action-Space* - $A = \{0, 1, 2\}$.
- *Admissible Actions.*

$$\begin{aligned} A(-1) &= \{0,1,2\} \\ A(0) &= \{0,1\} \\ A(1) &= \{0\} \end{aligned}$$

- *Immediate Rewards.*

$r(s,a) =$

$s\backslash a$	0	1	2
-1	-9	-25/4	-5
0	-21/4	-4	ND
1	-3	ND	ND

- *Transition Probabilities.*

$p(s'|s,0) =$

$s\backslash s'$	-1	0	1
-1	1	0	0
0	3/4	1/4	0
1	1/2	1/4	1/4

$p(s'|s,1) =$

$s\backslash s'$	-1	0	1
-1	3/4	1/4	0
0	1/2	1/4	1/4
1	ND	ND	ND

$p(s'|s,2) =$

$s\backslash s'$	-1	0	1
-1	1/2	1/4	1/4
0	ND	ND	ND
1	ND	ND	ND

Now I apply the *Policy Iteration Algorithm*

- *Initialisation* - Define the following Markovian Decision Function

$$\begin{aligned} d_0(-1) &= 2 \\ d_0(0) &= 1 \\ d_0(1) &= 0 \end{aligned}$$

- *Iteration* $k = 0$.
- *Policy Evaluation* - We compute a solution $\mu_0(s)$ to the following system of equations

$$\begin{aligned} \sum_{s' \in S} p_0(s|s')\mu(s') &= \mu(s) \ \forall \ s \in S \\ \sum_{s' \in S} \mu(s') &= 1 \end{aligned}$$

where $\mu(s)$ is unknown and $p_0(s'|s) = p(s'|s, d_0(s))$. We can derive the following three equations

$$\begin{aligned} i) \quad & p(-1|-1)\mu(-1) + p(-1|0)\mu(0) + p(-1|1)\mu(1) &&= \mu(-1) \\ ii) \quad & p(0|-1)\mu(-1) + p(0|0)\mu(0) + p(0|1)\mu(1) &&= \mu(0) \\ iii) \quad & p(1|-1)\mu(-1) + p(1|0)\mu(0) + p(1|1)\mu(1) &&= \mu(1) \\ \\ \implies & \frac{1}{2}\mu(-1) + \frac{1}{2}\mu(0) + \frac{1}{2}\mu(1) &&= \mu(-1) \text{ from } i) \\ \implies & \mu(0) + \mu(1) &&= \mu(-1) \\ \implies & \mu(-1) &&= \frac{1}{2} \text{ as sum to 1} \\ \\ \implies & \frac{1}{4}\mu(-1) + \frac{1}{4}\mu(0) + \frac{1}{4}\mu(1) &&= \mu(-1) \text{ from } ii) \\ \implies & \mu(-1) + \mu(1) &&= 3\mu(0) \\ \implies & \frac{1}{2} + \mu(0) + \left(3\mu(0) - \frac{1}{2}\right) &&= 1 \text{ as sum to 1} \\ \implies & \mu(0) &&= \frac{1}{4} \\ \implies & \mu(1) &&= \frac{1}{4} \text{ as sum to 1} \end{aligned}$$

We compute solutions $w_0(s)$ to the system of equations

$$\begin{aligned} w(s) - \sum_{s' \in S} p_0(s|s')w(s') &= r_0(s) - \bar{r}(0) \ \forall \ s \in S \\ \sum_{s' \in S} w(s')\mu_0(s') &= 0 \end{aligned}$$

where $w(s)$ is unknown and $r_0(s), \bar{r}_0$ are defined as

$$\begin{aligned} r_0(s) &= r(s, d_0(s)) = \begin{cases} -5 & \text{if } s = -1 \\ -4 & \text{if } s = 0 \\ -3 & \text{if } s = 1 \end{cases} \\ \bar{r}_0 &= \sum_{s' \in S} r_0(s')\mu(s') \\ &= (-5) \cdot \frac{1}{2} + (-4) \cdot \frac{1}{4} + (-3) \cdot \frac{1}{4} = -9/2 \end{aligned}$$

I can't be asked to do it, but the solution is

$$\begin{aligned} w_0(-1) &= -3/4 \\ w_0(0) &= -1/4 \\ w_0(1) &= -5/4 \end{aligned}$$

- *Policy Improvement.*

  Compute $d_1(s)$ using the following equations

$$\begin{aligned} u_1(s,a) &= r(s,a) + \sum_{s' \in S} w_0(s') p(s'|s,a) \\ d_1(s) &\in \text{argmax}_{a \in A(s)} u_1(s,a) \end{aligned}$$

  This gives values

$u_1(s,a) =$

$s \backslash a$	0	1	2
-1	-39/4	-27/4	-5
0	-23/4	-4	ND
1	-3	ND	ND
ND			

$u_1(s,a) =$

s	$d_1(s)$
-1	2
0	1
1	0

- As $d_1(s) = d_0(s)\ \forall\ s \in S$ the stopping condition is met.

Question 3) (c)
Formulate a linear program equivalent to the average reward problem solve in `3) (b)`.

Answer 3) (c)
Minimise t wrt $r, w(-1), w(0), w(1)$
Subject to

$$r(s,a) + \sum_{s' \in S} w(s') p(s'|s,a) \leq r + w(s) \quad s \in S, a \in A(s)$$