Application Notes - Stochastic Optimisation

Dom Hutchinson

December 14, 2020

Question 1)
Consider the following stochastic system. Let $T := \{0, \dots, N-1\}$ be a finite time-horizon, $X_t \in S$ be the system state at epoch $t \in T$, $Y_t \in A$ be the action taken at epoch $t \in T$, $A(s) \subseteq A$ be the admissible actions when in state $s \in S$. The stochastic system has the follow dynamics

$$\begin{aligned} \Psi_t &: S \times A \times B \to S \\ X_{t+1} &= \Psi_t(X_t, Y_t, U_t) \\ \Phi_t &: S \times C \to A \\ Y_{t+1} &= \Phi_t(X_t, V_t) \\ R_t &: S \times A \times D \to \mathbb{R} \\ &\quad \mathcal{R}_t(X_t, Y_t, W_t) \end{aligned}$$

where $U_t \sim \text{Uni}(B)$, $U_t \sim \text{Uni}(C)$, $W_t \sim \text{Uni}(D)$ for some discrete systems B, C, D. Assume that $X_0, \{U_t\}_{t\in T}, \{V_t\}_{t\in T}, \{W_t\}_{t\in T}$ are all mutually independent.

The objective of this task is to maximised the total expected reward from this system

$$\max \mathbb{E}\left[\sum_{t\in T} R_t(X_t, Y_t, W_t)\right]$$

Question 1) (a)
Show that the problem of maximising the expected total reward for this stochastic system is equivalent to the Markov Decision Problem.

Answer 1) (a)
This requires us to show two properties

i). This stochastic system exhibits Markovian Dynamics

$$\mathbb{P}(X_{t+1} = s_{t+1} | X_{0:t} = s_{0:t}, Y_{0:t} = a_{0:t}) = \mathbb{P}(X_{t+1} = s_{t+1} | X_t = s_t, Y_t = a_t)$$

ii). The expected total reward admits the following representation

$$\mathbb{E}\left[\sum_{t\in T} R_t(X_t, Y_t, W_t)\right] = \mathbb{E}\left[\sum_{t\in T} r_t(X_t, Y_t)\right]$$

At epoch $t = 1$ we have

$$\begin{aligned} X_1 &= \Psi_t(X_0, Y - 0, U_0) \\ &= \Psi_1(X_0, \Phi_0(X_0, V_0), U_0) \\ \implies X_1 &= \tilde{\Psi}_1(X_0, U_0, V_0) \end{aligned}$$

for a new function $\tilde{\Psi}_1 : S \times B \times C \to S$. Also, at epoch $t = 1$ we have

$$\begin{aligned} Y_1 &= \Phi_1(X_1, V_1) \\ &= \Phi_1(\tilde{\Psi}_1(X_0, U_0, V_0), V_1) \\ \implies Y_1 &= \tilde{\Phi}_1(X_0, U_0, V_{0:1}) \end{aligned}$$

for a new function $\tilde{\Phi}_1 : S \times B \times C^2 \to A$. We can extend this to the general epoch t

$$\begin{aligned} X_t &= \tilde{\Psi}_t(X_0, U_{0:t-1}, V_{0:t-1}) \\ Y_t &= \tilde{\Phi}_t(X_0, U_{0:t-1}, V_{0:t}) \end{aligned}$$

where our general mapping functions have signatures

$$\begin{aligned} \tilde{\Psi}_t &: S \times B^t \times C^t \to S \\ \tilde{\Phi}_t &: S \times B^t \times C^{t+1} \to A \end{aligned}$$

As we are allowed to assume that $X_0, \{U_t\}_{t\in T}, \{V_t\}_{t\in T}, \{W_t\}_{t\in T}$ are all mutually independent. We have that U_t & $(X_{0:t}, Y_{0;t})$ are mutually independent and W_t & (X_t, Y_t) are mutually independent. [1]

Consider the transition probabilities

$$\begin{aligned} & \mathbb{P}(X_{t+1} = s_{t+1} | X_{0:t} = s_{0:t}, Y_{0:t} = a_{0:t}) \\ = \ & \mathbb{P}(\Psi_t(X_t, Y_t, U_t) = s_{t+1} | X_{0:t} = s_{0:t}, Y_{0:t} = a_{0:t}) \text{ by def. } X_{t+1} \\ = \ & \mathbb{P}(\Psi_t(s_t, a_t, U_t) = s_{t+1} | X_{0:t} = s_{0:t}, Y_{0:t} = a_{0:t}) \text{ by conditions} \\ = \ & \mathbb{P}(\Psi_t(s_t, a_t, U_t) = s_{t+1}) \text{ as } U_t \perp\!\!\!\perp (X_{0:t}, Y_{0:t}) \\ = \ & \mathbb{P}(X_{t+1} = s_{t+1} | X_t = s_t, Y_t = a_t) \end{aligned}$$

This shows that the stochastic system exhibits markovian dynamics.

Question 1) (b)
Identify the elements of the equivalent Markov Decision Problem.

Answer 1) (b)
This requires us to identify the following

i). Transition probabilities

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

ii). Equivalent reward

$$r_t(s, a)$$

We derive the transition probabilities as follows

$$\begin{aligned} p_t(s'|s, a) &:= \mathbb{P}(X_{t+1} = s' | X_t = s, Y_t = a) \\ &= \mathbb{P}(\Psi_t(X_t, Y_t, U_t) = s' | X_t = s, Y + t = a) \text{ by def. } X_{t+1} \\ &= \mathbb{P}(\Psi_t(s, a, U_t) = s') \text{ by conditions} \\ &= \mathbb{E}\left[\mathbb{1}\{\Psi_t(s, a, U_t) = s'\}\right] \\ &= \sum_{u\in B} \mathbb{1}\{\Psi_t(s, a, u) = s'\} \cdot f_{U_t}(u) \end{aligned}$$

[1] Proof is long and given in slides

We have

$$\begin{aligned}\mathbb{E}\left[\sum_{t\in T} R_t(X_t, Y_t, W_t)\right] &= \sum_{t\in T}\mathbb{E}\left[R_t(X_t, Y_t, W_t)\right] \\ &= \sum_{t\in T}\mathbb{E}\left[\mathbb{E}\left[R_t(X_t, Y_t, W_t)|X_t, Y_t\right]\right] \text{ by Tower Property}\end{aligned}$$

Define $r_t(s,a) := \mathbb{E}\left[R_t(X_t, Y_t, W_t)|X_t = s, Y_t = a\right]$. This gives us a representation for expected total reward

$$\mathbb{E}\left[\sum_{t\in T} R_t(X_t, Y_t, W_t)\right] = \mathbb{E}\left[\sum_{t\in T} r_t(X_t, Y_t)\right]$$

Since W_t & (X_t, Y_T) are mutually independent we can get a deterministic expression for $r_t(s,a)$

$$\begin{aligned}r_t(s,a) &= \mathbb{E}\left[R_t(X_t, Y_t, W_t)|X_t = s, Y_t = a\right] \\ &= \mathbb{E}\left[R_t(s, a, W_t)\right] \text{ by conditions} \\ &= \sum_{w\in D} R_t(S, a, w) f_{W_t}(w) \text{ by def. expectation}\end{aligned}$$

Question 2) - Interesting system states X_t

Consider the following *Sequential Decision Problem*. Let $T := \{0, \dots, N-1\}$ and at each epoch the stochastic system can be in one of two conditions C_0 or C_1 (These are <u>not</u> system states). At each epoch the agent can take an action from $A := \{0, 1\}$ and let $A(s) = A$ for all $s \in S$.

Here are the possible interactions between the agent and the stochastic system

(A1) Agent takes action 1 at epoch $t \in T$:

- The system always will be in condition C_1 at epoch $t+1$.

(A0) Agent takes action 0 at epoch t:

- AND the system is in condition C_0 at epoch t: then the system will be in condition C_0 at epoch $t+1$.
- ELSE (if the system is in condition C_1 at epoch t):

  Let k be the number of epochs since action 1 was last taken, then the system will still be in state C_1 at epoch $t+1$ with probability $\pi(k)$, where $\{\pi(k)\}_{k\in\mathbb{N}^0}$ is a decreasing sequence in $[0,1]$ and there is some known $n \in \mathbb{N}$ st $\forall\ k \geq n,\ \pi(k) = 0$.

At each epoch $t \in T$, if the system is in state C_i, $i \in \{0,1\}$ and the agent takes action $j \in A$ then the agent receives *immediate reward* $R(i,j)$. <u>No</u> reward is received at epoch $t = N$

Question 2) (a)

Formulate the describe sequential decision problem as a finite-horizon *Markov Decision Problem*

Answer 2) (a)

This question requires us to identify: the decision epochs; time-horizon; system states; state-space; agent actions; action-space; transition probabilities; and, equivalent rewards.

- *Number of Epochs.*

  $N = 21$. Stated in question.

- *Time-Horizon.*

  $T+ = \{0, \dots, N-1\}$. Stated in question.

- *Agent actions.*

  Let Y_t denote the action the agent takes at epoch t.

- *Action-Space.*

  $A = \{0, 1\}$. Stated in question.

- *Admissible Actions.*

  $A(s) = A$ for all $s \in S$.

- *State-Space.*[2]

  Let X_t be the system state at epoch t ($X_t \notin \{C_0, c_1\}$, X'_t be the system condition at epoch t ($X'_t \in \{C_0, c_1\}$ and X''_t denote the number of decision epochs between epoch t and the last epoch in which action 0 was taken. Since X'_t, X''_t encode all relevant system information, we want to devise a definition of X_t which is a deterministic encoding of X'_t, X''_t.

  By considering the definitions of X'_t, X''_t, we can derive the following conclusions from the interactions described in the question

  - If ($Y_t = 0$ and $X''_t \geq n$): $\pi(X''_t) = 1 \implies X'_{t+1} = C_0$.
  - If ($X''_t \geq n+1$): Then $X''_{t-1} \geq n$ and action 0 was taken last turn $\implies X'_t = C_0$,
  - If ($Y_t = 0$ and $X'_t = C_0$): $X'_{t+1} = C_0$ as stated in question.
  - If ($X'_t = C_0$): It remains in C_0 until action 1 is taken.

  From these conclusions we state, if $X''_t \geq n \implies X''_t$ is not relevant for the selection of Y_t. Further, it is not relevant to the prediction of X_{t+1} given Y_t.

  We now define the system states X_t as

$$X_t = \begin{cases} X''_t & \text{if } X'_t = C_1 \\ n+1 & \text{if } X'_t = C_0 \end{cases}$$

  This is justified by considering what information is sufficient to make a prediction given possible combinations of X'_t, X''_t. This means the state-space is $S = \{0, \dots, n+1\}$.

- *Transition Probabilities*

  The definition of transition probabilities is

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

  We need to compute three cases

  i). $a = 1$ (ie $Y_t = 1$).
  In this case $X'_{t+1} = C_1, X''_{t+1} = 0 \implies X_{t+1} = 0$. Giving

$$p_t(s'|s, 1) \equiv \mathbb{P}^\pi[X_{t+1} = s'|X_t = s, Y_t = 1] = \begin{cases} 1 & \text{if s'=0} \\ 0 & \text{otherwise} \end{cases}$$

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

ii). $a = 0, s = n + 1$ (ie $Y_t = 0, X_t = n + 1$).
In this case $X'_t = C_0$. Giving

$$p_t(s'|n+1,0) \equiv \mathbb{P}^\pi[X_{t+1} = s'|X_t = n+1, Y_t = 0] = \begin{cases} 1 & \text{if s'=n+1} \\ 0 & \text{otherwise} \end{cases}$$

iii). $a = 0, s \leq n$ (ie $Y_t = 0, X_t = s \leq n$).
In this case $X'_{t+1} = C_1,\ X''_t = X_t = s$. We have that X'_{t+1} takes either C_0 or C_1 so we need to consider two probabilities

$$\begin{aligned} p_t(s+1|s,0) &= \mathbb{P}^\pi(X'_{t+1} = C_1|X'_t = C_1, X''_t = s, Y_t = 0) = \pi(s) \\ p_t(n+1|s,0) &= \mathbb{P}^\pi(X'_{t+1} = C_0|X'_t = C_1, X''_t = s, Y_t = 0) = 1 - \pi(s) \end{aligned}$$

We can summarise these two expression as the following

$$p_t(s'|s,a) = \begin{cases} \pi(s) & \text{if } s' = s+1 \\ 1-\pi(s) & \text{if } s' = n+1 \\ 0 & \text{otherwise} \end{cases}$$

- *Equivalent Rewards.*
  If $X_t \leq n$ then $X'_t = C_1 \implies r_t = R(1, Y_t)$.
  If $X_t = n + 1$ then $X'_t = C_0 \implies r_t = R(0, Y_t)$.
  This can be summarised as

$$r_t(s,a) = \begin{cases} R(1,a) & \text{if } s \leq n \\ R(0,a) \text{if } s = n+1 \end{cases}$$

- *Terminal Award.*
  $r_N(s) = 0$. Stated in the question.
- *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 2) (b)
By considering the markov decision problem formulated in `2) (a)` and assuming the following

- $N = 2,\ n = 1$
- $\pi(0 = .5)$
- $R(0,0) = -5,\ R(0,1) = -7,\ R(1,0) = 0,\ R(1,1) = -2.$

Find an optimal policy π^*

Answer 2) (b)
From `2) (a)` we can quickly derive this formulation by substituting in the values specified.

- *Decision Epochs* - $N = 2$.

- *Time-Horizon* - $T = \{0, 1\}$.
- *Action-Space* - $A = \{0, 1\}$.
- *Admissible Actions* - $A(s) = \{0, 1\}\ \forall\ s \in S$.
- *State-Space* - $S = \{0, 1, 2\}$
- *Transition Probabilities*

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

s\s'	0	1	2
0	0	.5	.5
1	0	0	1
2	0	0	1

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

s\s'	0	1	2
0	1	0	0
1	1	0	0
2	1	0	0

- *Rewards.* $r_t(s, a) =$

s\a	0	1
0	0	-2
1	0	-2
2	-5	-7

- *Terminal Award* - $r_2(s) = 0$.

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 $t = N - 1, \ldots, 0$ 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

$t = 1$ We need to compute $u_1^*(s), d_1^*(s)$. We have

$$\begin{aligned} w_1^*(s, a) &= r_1(s, a) + u_2^*(0) p_1(0|s, a) + u_2^*(1) p_1(1|s, a) + u_2^*(2) p_1(2|s, a) \\ &= r_1(s, a) \text{ since } u_2^*(s) = 0\ \forall\ s \end{aligned}$$

where $s \in S = \{0, 1, 2\}$.

Consider the following table for the value of $w_1^*(s, a)$

$w_1^*(s, a) =$

$s\backslash a$	0	1
0	0	-2
1	0	-2
2	-5	-7

We can use this to determine $u_1^*(s), d_1^*(s)$ for each state s

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

This shows that action 0 is optimal for all states in epoch $t = 1$

$t = 0$ We need to compute $u_0^*(s), d_0^*(s)$. We have

$$w_0^*(s,a) = r_0(s,a) + u_1^*(0)p01(0|s,a) + u_1^*(1)p_0(1|s,a) + u_1^*(2)p_0(2|s,a)$$

where $s \in S = \{0, 1, 2\}$.

Consider the following table for the value of $w_1^*(s,a)$

$w_0^*(s,a) =$

$s \backslash a$	0	1
0	$-\frac{5}{2}$	-2
1	-5	-2
2	-10	-7

We can use this to determine $u_0^*(s), d_0^*(s)$ for each state s

s	u_0^*	$d_0^*(s)$
0	-2	1
1	-2	1
2	-7	1

This shows that action 1 is optimal for all states in epoch $t = 0$

This shows that the optimal strategy is $\pi^* = (1, 0)$

Question 3) - Inventory Control

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 3) (a)

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

Problem

Answer 3) (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*[3]

  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 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

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

  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

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

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 $Z_0, \dots, Z_{t-1}$. As Z_t is independent of $Z_0, \dots, Z_{t-1}$ (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 $Z_0, \ldots, Z_{N-1}$, 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 3) (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 3) (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*

$$r_t(s,a) = \begin{array}{c|ccc} s\backslash a & 0 & 1 & 2 \\ -1 & -9 & -25/4 & -5 \\ 0 & -21/4 & -5 & \text{ND} \\ 1 & -3 & \text{ND} & \text{ND} \end{array}$$

  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

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

  The optimal strategy depends on what state the system is in.

- 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

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

  The

optimal strategy depends on what state the system is in.

- *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 4) - Discount Reward Inventory Problem
This is from `LectureSlides4cSO.pdf` *and covers chapter 3..*
This is an *Inventory Control Problem* with discounted-cost (See `Question 3)` 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 4) a)
Formulate the inventory control problem as a *Discounted Reward Markov Decision Problem.*

Answer 4) 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

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

- *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 4) b)
Consider the MDP formulated in `4) a)`. Assume the following

- $n = 1, m = 1, \alpha = .5$.
- $p(0) = p(1) = p(2) = p(3) = \dfrac{1}{4}$ 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 4) 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 4) c)
Formulate a linear problem equivalent to the MDP solved in `4) b)`.

Answer 4) 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 5) - Optimality of π^* for Discounted Reward MDP
Consider the following generic *Discounted Reward MDP*

- *Time-Horizon* - $T = \{0, 1, \dots\}$.
- *State-Space* - S.
- *Action-Space* - A.
- *Admissible Actions* - $A(s)$.
- *Transition Probabilities* - $p_t(s'|s,a) = p(s'|s,a)\ \forall\ t \in T$.
- *Rewards* - $r_t(s,a) = \alpha^t r(s,a)$ where $\alpha \in (0,1)$.
- *Objective* - Find $\pi \in HR(T)$ which maximises

$$\mathbb{E}^{\pi}\left(\sum_{t=0}^{\sum} \alpha^t r(X_t, Y_t)\right)$$

Let $v^*(s)$ be the optimal value function which is the unique solution to the Bellam equation

$$v^*(s) = \max_{a \in A(s)} \left(r(s,a) + \alpha \sum_{s' \in S} v^*(s')p(s'|s,a) \right)$$

Let $D^*(s)$ be the set of optimal actions for a given state s

$$
\begin{aligned}
D^*(s) &= \text{argmax}_{a \in A(s)} \left(r(s,a) + \alpha \sum_{s' \in S} v^*(s')p(s'|s,a) \right) \\
&= \left\{ a \in A(s) : v^*(s) = r(s,a) + \alpha \sum_{s' \in S} v^*(s')p(s'|s,a) \right\}
\end{aligned}
$$

Let $q^*(a|s)$ be any *Markovian Decision Probability* which satisfies the condition

$$q^*(a|s) = 0 \ \forall \ a \notin D^*(s), \ \forall \ s \in S$$

Let π^* be the *Stationary Markovian Policy* based on the *Markovian Decision Probability* $q^*(a|s)$. (ie π^* applies $q^*(a|s)$ at each epoch).

Show that π^* is optimal.

Answer 5)

Note that, under policy π^*, the agent action Y_t at each epoch $t \in T$ is selected according to the decision probability $q^*(a|s)$

$$\mathbb{P}^{\pi^*}\left(Y_t = a|X_{0:t}, Y_{0:t-1}\right) = q^*(a|X_t)$$

By the filtering property of conditional expectations, we get

$$\begin{aligned}
\mathbb{P}^{\pi^*}(Y_t = a|X_t) &= \mathbb{E}^{\pi^*}\left(\mathbb{P}^{\pi^*}(Y_t = a|X_{0:t}, Y_{0:t-1})\big|X_t\right) \\
&= \mathbb{E}^{\pi^*}(q^*(a|X_t)|X_t) \\
&= q^*(a|X_t) \qquad [1]
\end{aligned}$$

From the notes, we know that the maximum expected discounted reward is

$$\mathbb{E}[v^*(X_0)]$$

Thus, to show π^* is optimal, it is sufficient to show

$$\mathbb{E}^{\pi^*}\left[\sum_{t=0}^{\infty} \alpha^t r(X_t, Y_t)\right] = \mathbb{E}[v^*(X_0)]$$

Let $w^*(s,a)$ denote the following

$$w^*(s,a) := r(s,a) + \alpha \sum_{s' \in S} v^*(s) p(s'|s,a)$$

This means the *Bellman Equation* and $D^*(s)$ can be rewritten as

$$\begin{aligned}
v^*(s) &= \max_{a \in A(s)} w^*(s,a) \\
D^*(s) &= \text{argmax}_{a \in A(s)} w^*(s,a)
\end{aligned}$$

If $a \in D^*(s)$ then $w^*(s,a) = \max_{a \in A(s)} w^*(s,a) = v^*(s)$.

$$a \in D^*(s) \implies [v^*(s) = w^*(s,a)]$$

By the definition of $q^*(a|s)$, we have

$$\begin{aligned}
a \notin D^*(s) &\implies q^*(a|s) = 0 \\
\implies \sum_{a \in A(s)} w^*(s,a) q^*(a|s) &= \sum_{a \in D^*(s)} w^*(s,a) q^*(a|s) + \sum_{a \in A(s) \setminus D^*(s)} w^*(s,a) \underbrace{q^*(a|s)}_{=0} \\
\implies \sum_{a \in A(s)} w^*(s,a) q^*(a|s) &= \sum_{a \in D^*(s)} w^*(s,a) q^*(a|s) \\
&= \sum_{a \in D^*(s)} v^*(s) q^*(a|s) \\
&= v^*(s) \sum_{a \in D^*(s)} q^*(a|s) \\
&= v^*(s) \\
\implies v^*(s) &= \sum_{a \in A(s)} w^*(s,a) q^*(a|s) \qquad [1]
\end{aligned}$$

Assume that $\{(X_t, Y_t)\}_{t\in T}$ is generated with policy π^*. Setting $s = X_t$ is the above formulation of [2] and using [1], we get

$$\begin{aligned}
v^*(X_t) &= \sum_{a\in A(X_t)} w^*(X_t,a)q^*(a|X_t) \\
&= \sum_{a\in A(X_t)} w^*(X_t,a)\mathbb{P}^{\pi^*}[Y_t = a|X_t] \\
&= \mathbb{E}^{\pi^*}[w^*(X_t,Y_t)|X_t] \\
&= \mathbb{E}^{\pi^*}\left[r(X_t,Y_t) + \alpha\sum_{s'\in S} v^*(s')p(s'|X_t,Y_t)|X_t\right] \\
&= \mathbb{E}^{\pi^*}\left[r(X_t,Y_t) + \alpha\sum_{s'\in S} v^*(s')\mathbb{P}^{\pi^*}[X_{t+1} = x'|X_t,Y_t]|X_t\right] \\
&= \mathbb{E}^{\pi^*}\left[r(X_t,Y_t) + \alpha\mathbb{E}^{\pi^*}[v^*(X_{t+1})|X_t,Y_t]|X_t\right] \\
&= \mathbb{E}^{\pi^*}[r(X_t,Y_t)|X_t] + \alpha\mathbb{E}^{\pi^*}\left[\mathbb{E}^{\pi^*}[v^*(X_{t+1})|X_t,Y_t]|X_t\right] \\
&= \mathbb{E}^{\pi^*}[r(X_t,Y_t)|X_t] + \alpha\mathbb{E}^{\pi^*}[v^*(X_{t+1})|X_t] \\
&= \mathbb{E}^{\pi^*}[r(X_t,Y_t) + \alpha v^*(X_{t+1})|X_t] \\
\implies v^*(X_t) &= \mathbb{E}^{\pi^*}[r(X_t,Y_t) + \alpha v^*(X_{t+1})|X_t] \\
\implies \mathbb{E}^{\pi^*}[v^*(X_t)] &= \mathbb{E}^{\pi^*}\left[\mathbb{E}^{\pi^*}[r(X_t,Y_t) + \alpha v^*(X_{t+1})|X_t]\right] \\
&= \mathbb{E}^{\pi^*}\left[r(X_t,Y_t) + \alpha v^*(X_{t+1})\right] \text{ by tower property} \\
&= \mathbb{E}^{\pi^*}[r(X_t,Y_t)] + \alpha\mathbb{E}^{\pi^*}[v^*(X_{t+1})] \\
\implies \mathbb{E}^{\pi^*}[r(X_t,Y_t)] &= \mathbb{E}^{\pi^*}[v^*(X_t)] - \alpha\mathbb{E}^{\pi^*}[v^*(X_{t+1})]
\end{aligned}$$

Using this result, we can deduce the following about the total expected reward

$$\begin{aligned}
\mathbb{E}^{\pi^*}\left[\sum_{t=0}^{\infty}\alpha^t r(X_t,Y_t)\right] &= \sum_{t=0}^{\infty}\alpha^t\mathbb{E}^{\pi^*}\left[r(X_t,Y_t)\right] \\
&= \sum_{t=0}^{\infty}\alpha^t\left\{\mathbb{E}^{\pi^*}[v^*(X_t)] - \alpha\mathbb{E}^{\pi^*}[v^*(X_{t+1})]\right\} \\
&= \sum_{t=0}^{\infty}\alpha^t\mathbb{E}^{\pi^*}[v^*(X_t)] - \sum_{t=0}^{\infty}\alpha^{t+1}\mathbb{E}^{\pi^*}[v^*(X_{t+1})] \\
&= \sum_{t=0}^{\infty}\alpha^t\mathbb{E}^{\pi^*}[v^*(X_t)] - \sum_{t=1}^{\infty}\alpha^t\mathbb{E}^{\pi^*}[v^*(X_t)] \\
&= \mathbb{E}^{\pi^*}[v^*(X_0)] \\
&= \mathbb{E}[v^*(X_0)]
\end{aligned}$$

□