Live Lecture Notes - Stochastic Optimisation

Dom Hutchinson

January 26, 2021

Question 1) - Sequential Decision Problem as an MDP
This question is from "Chapter 1: Live Lecture" (`LectureSlides2bSO.pdf`).
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

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

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

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

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

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

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

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

where our general mapping functions have signatures

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

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

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

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

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

[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
This question is from "Chapter 2: Live Lecture" (`LectureSlides3bSO.pdf`).
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 not 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* $R(i, j)$. No 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.

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

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

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

$t=1$ 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) - Optimality of a policy for a General FH-MDP
This question is from "Chapter 2: Live Lecture B (Revised)" (`LectureSlides3dSO.pdf`).
Consider a *Generic Finite-Horizon MDP* over horizon $T := \{0, \dots, N-1\}$.

Define, as a backwards-recursion, the *Optimality Equations* $u^*_{N-1}(s), \dots, u^*_0(s)$ as

$$\begin{aligned} u^*_N(s) &= r_N(s) \\ u^*_k(s) &= \max_{a\in A(s)} \left(r_k(s,a) + \sum_{s'\in S} u^*_{k+1}(s')p_k(s'|s,a) \right) \text{ for } k \in [N-1, 0], s \in S \end{aligned}$$

the *Optimal Decision Rule* sets $D^*_0(s), \dots, D^*_{N-1}(s)$ as

$$\begin{aligned} D^*_k(s) &= \text{argmax}_{a\in A(s)} \left(r_k(s,a) + \sum_{s'\in S} u^*_{k+1}(s')p_k(s'|s,a) \right) \text{ for } k \in [N-1, 0], s \in S \\ &= \left\{ a \in A(s) : u^*_k(s) = r_k(s,a) + \sum_{s'\in S} u^*_{k+1}(s')p_k(s'|s,a) \right\} \end{aligned}$$

Let $q^*_0(a|s), \dots, q^*_{N-1}(a|s)$ be any *Markovian Decision Probabilities* which give zero weight to all sub-optimal actions.

$$q^*_t(a|s) = 0 \ \forall \ a \notin D^*_t(s)$$

Let π^* be a *Markovian Policy* based on $q^*_0(a|s), \dots, q^*_{N-1}(a|s)$

$$\pi^* := \{q^*_t(a|s)\}_{t\in T}$$

Show that π^* is an optimal policy.

Answer 3)
Note that, under policy π^*, the agent action Y_t is chosen as

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

By the filtering property of conditional expectations we get

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

To prove π^* is an optimal policy, it is sufficient to show that

$$\mathbb{E}[u^*_0(X_0)] = \mathbb{E}^{\pi^*}\left[\sum_{t=0}^{N-1} r_t(X_t, Y_t) + r_N(X_N)\right]$$

Let R_k denote the expected reward from the last $N-k$ steps

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

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

$$R_k = \mathbb{E}^{\pi^*}\left[u^*_k(X_k)\right] \ \forall \ k \in [0, N] \tag{1}$$

This is show by a backwards recursion.

Initial Step - $k = N$. By definition

$$R_N := \mathbb{E}^{\pi^*}[r_N(X_N)] =: \mathbb{E}^{\pi^*}[u_N^*(X_N)]$$

The result holds.
Inductive Hypothesis - $R_k = \mathbb{E}^{\pi^*}[u_k^*(X_k)]$ holds for an arbitrary $k \in [1, N]$.
Inductive Step - $k - 1$.
Consider R_{k-1}

$$\begin{array}{rcl}
R_{k-1} & := & \mathbb{E}^{\pi^*}\left[\displaystyle\sum_{t=k-1}^{N-1} r_t(X_t, Y_t) + r_N(X_N)\right] \\
& = & \mathbb{E}^{\pi^*}[r_{k-1}(X_{k-1}Y_{k-1})] + \mathbb{E}^{\pi^*}\left[\displaystyle\sum_{t=k}^{N-1} r_t(X_t, Y_t) + r_N(X_N)\right] \\
& = & \mathbb{E}^{\pi^*}[r_{k-1}(X_{k-1}, Y_{k-1})] + R_k \text{ by def.} \\
& = & \mathbb{E}^{\pi^*}[r_{k-1}(X_{k-1}, Y_{k-1})] + \mathbb{E}^{\pi^*}[u_k^*(X_k)] \text{ by IH.} \\
& = & \mathbb{E}^{\pi^*}[r_{k-1}(X_{k-1}, Y_{k-1}) + u_k^*(X_k)] \\
& = & \mathbb{E}^{\pi^*}\left[\mathbb{E}^{\pi^*}[r_{k-1}(X_{k-1}, Y_{k-1}) + u_k^*(X_k)\big|X_{k-1}, Y_{k-1}]\right] \text{ by Tower Prpty.} \\
& = & \mathbb{E}^{\pi^*}\left[r_{k-1}(X_{k-1}, Y_{k-1}) + \mathbb{E}^{\pi^*}[u_k^*(X_k)\big|X_{k-1}, Y_{k-1}]\right] \\
& = & \mathbb{E}^{\pi^*}\left[r_{k-1}(X_{k-1}, Y_{k-1}) + \displaystyle\sum_{s' \in S} u_k^*(s')\mathbb{P}^{\pi^*}(X_k = s'|X_{k-1}, Y_{k-1})\right] \\
& = & \mathbb{E}^{\pi^*}\left[r_{k-1}(X_{k-1}, Y_{k-1}) + \displaystyle\sum_{s' \in S} u_k^*(s')p_{k-1}(s'|X_{k-1}, Y_{k-1})\right] \\
& = & \mathbb{E}^{\pi^*}\left[\mathbb{E}^{\pi^*}\left[r_{k-1}(X_{k-1}, Y_{k-1}) + \displaystyle\sum_{s' \in S} u_k^*(s')p_{k-1}(s'|X_{k-1}, Y_{k-1})\middle|X_{k-1}\right]\right] \text{ by Tower Prpty.} \\
& = & \mathbb{E}^{\pi^*}\left[\displaystyle\sum_{a \in A(X_{k-1})}\left[r_{k-1}(X_{k-1}, a) + \displaystyle\sum_{s' \in S} u_k^*(s')p_{k-1}(s'|X_{k-1}, a)\right] q_{k-1}^*(a|X_{k-1})\right] \\
& = & \mathbb{E}^{\pi^*}\left[\displaystyle\sum_{a \in D_{k-1}^*(s)}\left[r_{k-1}(X_{k-1}, a) + \displaystyle\sum_{s' \in S} u_k^*(s')p_{k-1}(s'|X_{k-1}, a)\right] q_{k-1}^*(a|X_{k-1})\right] \text{ by def. } q_{k-1}^*(\cdot) \\
& = & \mathbb{E}^{\pi^*}\left[\displaystyle\sum_{a \in D_{k-1}^*(s)} u_{k-1}^*(s)q_{k-1}^*(a|X_{k-1})\right] \text{ by def. } D_{k-1}^*(s) \\
& = & \mathbb{E}^{\pi^*}\left[u_{k-1}^*(s)\displaystyle\sum_{a \in D_{k-1}^*(s)} q_{k-1}^*(a|X_{k-1})\right] \\
& = & \mathbb{E}^{\pi^*}[u_{k-1}^*(s)]
\end{array}$$

Hence, by mathematical induction, the result holds for all $k \in [0, N]$.

Question 4) - Optimality Equation for Semi-Static FH-MDP
This question is from "Chapter 2: Live Lecture B (Revised)" (`LectureSlides3dSO.pdf`).
Consider a *general Finite-Horizon MDP* and assume the following

$$\begin{array}{rcl} S & = & \{1,\dots,M\} \text{ for } M\in[2,\infty) \\ A(s) & = & A\ \forall\ s\in S \\ p_t(s'|s,a) & = & p_t(s'|\tilde{s},a)\ \forall\ s',s,\tilde{s}\in S \\ r_t(s,a) & \in & [0,r_t(\tilde{s},a)]\ \forall\ s',s,\tilde{s}\in S \text{ where } s\leq\tilde{s} \\ r_N(s) & \in & [0,r_N(\tilde{s})]\ \forall\ s',s,\tilde{s}\in S \text{ where } s\leq\tilde{s} \\ u_N^*(s) & = & r_N(s) \\ u_k^*(s) & = & \max_{a\in A(s)}\left(r_k(s,a)+\sum_{s'\in S}u_{k+1}^*(s')p_k(s'|s,a)\right) \text{ for } k\in[N-1,0] \end{array}$$

This means that the transition probabilities vary depending upon action a, not the system state s. Show that

$$u_t^*(s)\leq u_t^*(\tilde{s})\ \forall\ s,\tilde{s}\in S \text{ where } s\leq\tilde{s}$$

Answer 4)
Fix $s,\tilde{s}\in S$ with $s\leq\tilde{s}$ and $t\in T$. By the question we have

$$\begin{array}{rcl} p_t(s'|s,a) & = & p_t(s'|\tilde{s},a) \\ r_t(s,a) & \leq & r_t(\tilde{s},a) \end{array}$$

Thus

$$\begin{array}{rrcl} & \sum_{s'\in S}u_{t+1}^*(s')p_t(s'|s,a) & = & \sum_{s'\in S}u_{t+1}^*(s')p_t(s'|\tilde{s},a) \\ \implies & r_t(s,a)\sum_{s'\in S}u_{t+1}^*(s')p_t(s'|s,a) & \leq & r_t(\tilde{s},a)\sum_{s'\in S}u_{t+1}^*(s')p_t(s'|\tilde{s},a) \end{array}$$

By taking the maximum of both sides we get

$$\begin{array}{rcl} \max_{a\in A}\left\{r_t(s,a)\sum_{s'\in S}u_{t+1}^*(s')p_t(s'|s,a)\right\} & \leq & \max_{a\in A}\left\{r_t(\tilde{s},a)\sum_{s'\in S}u_{t+1}^*(s')p_t(s'|\tilde{s},a)\right\} \\ \implies u_t^*(s) & \leq & u_t^*(\tilde{s}) \end{array}$$

Question 5) - Optimality of a policy for DR-MDP
This question is from "Chapter 3: Live Lecture" (`LectureSlides4bSO.pdf`).
Consider a general *Discounted Reward MDP*. Notably this means, $r_t(s,a) = \alpha^t r(s,a)$ for some $\alpha \in (0,1)$.

Let $v^*(s)$ be the unique solution to the *Bellman Equation* for Discounted Reward MDPs

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

Let $D^*(s)$ be the set of optimal agent actions in state s

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

Let $q^*(a|s)$ be a *Markovian Decision Function* which gives zero weight to sub-optimal actions

$$q^*(a|s) = 0 \; \forall \; a \notin D^*(s)$$

Let π^* be the stationary *Markovian Policy* based on the $q^*(a|s)$ (ie π^* applies $q^*(a|s)$ in all epochs).

Show that π^* is an *Optimal Policy*.

Answer 5)
Note that under π^* the agent action Y_t is chosen as

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

By the filtering property of conditional expectations, we get

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

The maximum expected reward is $\mathbb{E}[v^*(X_0)]$, thus to show that π^* is optimal, it is sufficient to show that

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

Let $w^*(s,a)$ denote the function to maximised by the *Bellman Equation*

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

We can restate the *Bellman Equation* and $D^*(s)$ as

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

Thus

$$\begin{array}{rcl} a \in D^*(s) \implies w^*(s,a) & = & \displaystyle\max_{a \in A(s)} w^*(s,a) \\ & = & v^*(s) \end{array}$$

By the question we have that $a \notin D^*(s) \implies q^*(a|s) = 0$, thus

$$\begin{array}{rcl}
\displaystyle\sum_{a\in A(s)} w^*(s,a)q^*(a|s) &=& \displaystyle\sum_{a\in D^*(s)} w^*(s,a)q^*(a|s) \text{ by def. } q^*() \\
&=& \displaystyle\sum_{a\in D^*(s)} v^*(s)q^*(a|s) \text{ by above} \\
&=& \displaystyle v^*(s)\sum_{a\in D^*(s)} q^*(a|s) \\
&=& v^*(s) \\
\implies \displaystyle\sum_{a\in A(s)} w^*(s,a)q^*(a|s) &=& v^*(s)
\end{array}$$

Assume that $\{(X_t,Y_t)\}_{t\in T}$ was generated by π^* and set $s = X_t$. Then

$$\begin{array}{rcl}
v^*(X_t) &=& \displaystyle\sum_{a\in A(X_t)} w^*(X_t,a)q^*(a|X_t) \\
&=& \mathbb{E}^{\pi^*}[w^*(X_t,Y_t)|X_t] \\
&=& \displaystyle\mathbb{E}^{\pi^*}\left[r(X_t,Y_t) + \alpha\sum_{s'\in S} v^*(s')p(s'|X_t,Y_t)\middle|X_t\right] \text{ by def. } w^*(X_t,Y_t) \\
&=& \mathbb{E}^{\pi^*}\left[r(X_t,Y_t) + \alpha\mathbb{E}^{\pi^*}[v^*(X_{t+1})|X_t,Y_t]\middle|X_t\right]
\end{array}$$

By the filtering property of conditional expectations

$$\begin{array}{rrcl}
& v^*(X_t) &=& \mathbb{E}^{\pi^*}\left[r(X_t,Y_t)\middle|X_t\right] + \alpha\mathbb{E}^{\pi^*}\left[\mathbb{E}^{\pi^*}[v^*(X_{t+1})|X_t,Y_t]\middle|X_t\right] \\
&&=& \mathbb{E}^{\pi^*}\left[r(X_t,Y_t)\middle|X_t\right] + \alpha\mathbb{E}^{\pi^*}\left[v^*(X_{t+1})\middle|X_t\right] \\
&&=& \mathbb{E}^{\pi^*}\left[r(X_t,Y_t) + \alpha v^*(X_{t+1})\middle|X_t\right] \\
\implies & \mathbb{E}^{\pi^*}[v^*(X_t)] &=& \mathbb{E}\left[\mathbb{E}^{\pi^*}\left[r(X_t,Y_t) + \alpha v^*(X_{t+1})\middle|X_t\right]\right] \text{ by tower prpty.} \\
&&=& \mathbb{E}^{\pi^*}[r(X_t,Y_t) + \alpha v^*(X_{t+1})] \\
&&=& \mathbb{E}^{\pi^*}[r(X_t,Y_t)] + \alpha\mathbb{E}^{\pi^*}[v^*(X_{t+1})] \\
\implies & \mathbb{E}^{\pi^*}[r(X_t,Y_t)] &=& \mathbb{E}^{\pi^*}[v^*(X_t)] - \alpha\mathbb{E}^{\pi^*}[v^*(X_{t+1})] \\
\implies & \displaystyle\mathbb{E}^{\pi^*}\left[\sum_{t=0}^{\infty}\alpha^t r(X_t,Y_t)\right] &=& \displaystyle\sum_{t=0}^{\infty}\alpha^t\mathbb{E}^{\pi^*}[r(X_t,Y_t)] \\
&&=& \displaystyle\sum_{t=0}^{\infty}\alpha^t\left(\mathbb{E}^{\pi^*}[v^*(X_t)] - \alpha\mathbb{E}^{\pi^*}[v^*(X_{t+1})]\right) \\
&&=& \displaystyle\sum_{t=0}^{\infty}\alpha^t\mathbb{E}^{\pi^*}[v^*(X_t)] - \sum_{t=0}^{\infty}\alpha^{t+1}\mathbb{E}^{\pi^*}[v^*(X_{t+1})] \\
&&=& \displaystyle\sum_{t=0}^{\infty}\alpha^t\mathbb{E}^{\pi^*}[v^*(X_t)] - \sum_{t=1}^{\infty}\alpha^{t}\mathbb{E}^{\pi^*}[v^*(X_{t})] \\
&&=& \mathbb{E}^{\pi^*}[v^*(X_0)] \\
&&=& \mathbb{E}[v^*(X_0)]
\end{array}$$

Question 6) - Uniqueness of Solution to Bellman Equation for AR-MDP
This question is from "Chapter 4: Live Lecture" (`LectureSlides5bSO.pdf`).
Consider a general *Average-Reward MDP*, note that this means the objective is to find π which maximises

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

Let $(r*,w^*(s))$ be a solution to the *Bellman Equations* for *Average Reward MDPs*

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

Let $(\tilde{r}^*,\tilde{w}^*(s))$ be another solution to the *Bellman Equations* for *Average Reward MDPs*

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

Assume that $\{X_t\}_{t\geq}$ is an irreducible Markov chain under any stationary, Markovian, deterministic policy

Question 6) (a)
Show that $\tilde{r}^*=r^*$

Answer 6) (a)
Let $d^*(s)$ be a Markovian decision function which only chooses actions which maximise the *Bellman Equations* using the solutions $(r^*,w^*(s))$.

$$d^*(s)\in\text{argmax}_{a\in A(s)}\left[r(s,a)+\sum_{s'\in S}w^*(s')p(s'|s,a)\right]$$

Let π^* be the stationary policy which applies $d^*(s)$ in every epoch. Let $\tilde{d}^*(s)$ be a Markovian decision function which only chooses actions which maximise the *Bellman Equations* using the other solutions $(\tilde{r}^*,\tilde{w}^*(s))$.

$$\tilde{d}^*(s)\in\text{argmax}_{a\in A(s)}\left[r(s,a)+\sum_{s'\in S}\tilde{w}^*(s')p(s'|s,a)\right]$$

Let $\tilde{\pi}^*$ be the stationary policy which applies $\tilde{d}^*(s)$ in every epoch.

We have that

$$\begin{array}{rcl}
\lim_{N\to\infty}\sup\mathbb{E}^\pi\left[\frac{1}{N}\sum_{t=0}^{N-1}r(X_t,Y_t)\right] & \leq & r^*\ \forall\ \pi\\
\lim_{N\to\infty}\sup\mathbb{E}^{\pi^*}\left[\frac{1}{N}\sum_{t=0}^{N-1}r(X_t,Y_t)\right] & = & r^*\ \forall\ \pi\\
\lim_{N\to\infty}\sup\mathbb{E}^\pi\left[\frac{1}{N}\sum_{t=0}^{N-1}r(X_t,Y_t)\right] & \leq & \tilde{r}^*\ \forall\ \pi\\
\lim_{N\to\infty}\sup\mathbb{E}^{\tilde{\pi}^*}\left[\frac{1}{N}\sum_{t=0}^{N-1}r(X_t,Y_t)\right] & = & \tilde{r}^*\ \forall\ \pi
\end{array}$$

Setting $\pi = \pi^*$ we have that

$$\begin{aligned} r^* &= \lim_{N\to\infty} \mathbb{E}^{\pi^*}\left[\frac{1}{N}\sum_{t=0}^{N-1} r(X_t,Y_t)\right] \\ &= \lim_{N\to\infty} \sup \mathbb{E}^{\pi^*}\left[\frac{1}{N}\sum_{t=0}^{N-1} r(X_t,Y_t)\right] \\ &\leq \tilde{r}^* \\ \implies r^* &\leq \tilde{r}^* \end{aligned}$$

Setting $\pi = \tilde{\pi}^*$ we have that

$$\begin{aligned} \tilde{r}^* &= \lim_{N\to\infty} \mathbb{E}^{\tilde{\pi}^*}\left[\frac{1}{N}\sum_{t=0}^{N-1} r(X_t,Y_t)\right] \\ &= \lim_{N\to\infty} \sup \mathbb{E}^{\tilde{\pi}^*}\left[\frac{1}{N}\sum_{t=0}^{N-1} r(X_t,Y_t)\right] \\ &\leq r^* \\ \implies \tilde{r}^* &\leq r^* \end{aligned}$$

Thus

$$\tilde{r}^* = r^*$$

Question 6) (b)
Show that $\exists c \in \mathbb{R}$ st

$$\tilde{w}^*(s) = w^*(s) + c \ \forall \ s \in S$$

Answer 6) (b)
Note that $r^*, w^*(s), d^*(s)$ satisfy the following

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

Thus

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

and this is an equality if $a = d^*(s)$.

Setting $a = d^*(s)$ we get

$$\begin{aligned} r(s,d^*(s)) + \sum_{s'\in S} w^*(s')p(s'|s,d^*(s)) &= r^* + w^*(s) \\ \implies \qquad r(s,d^*(s)) - r^* &= w^*(s) - \sum_{s'\in S} w^*(s')p(s'|s,d^*(s)) \end{aligned}$$

Similarly, note that $\tilde{r}^*, \tilde{w}^*(s), d^*(s)$ satisfy the following

$$\begin{aligned} \tilde{r}^* + \tilde{w}^*(s) &= \max_{a\in A(s)}\left(r(s,a) + \sum_{s'\in S} \tilde{w}^*(s')p(s'|s,a)\right) \\ \tilde{d}^*(s) &= \text{argmax}_{a\in A(s)}\left(r(s,a) + \sum_{s'\in S} \tilde{w}^*(s')p(s'|s,a)\right) \end{aligned}$$

Thus

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

and this is an equality if $a=\tilde{d}^*(s)$.
Setting $a=\tilde{d}^*(s)$ we get

$$\begin{array}{rrcl} & r(s,\tilde{d}^*(s))+\sum_{s'\in S}\tilde{w}^*(s')p(s'|s,\tilde{d}^*(s)) & \leq & \tilde{r}^*+\tilde{w}^*(s) \\ & & = & r^*+\tilde{w}^*(s) \text{ by a)} \\ \implies & \tilde{w}^*(s)-\sum_{s'\in S}\tilde{w}^*(s)p(s'|s,\tilde{d}^*(s)) & \geq & r(s,d^*(s))-r^* \end{array}$$

By combining this inequality with the earlier expression we get

$$[\tilde{w}^*(s)-w^*(s)]-\sum_{s'\in S}[\tilde{w}^*(s')-w^*(s')]p(s'|s,d^*(s))\geq 0 \tag{2}$$

Let $p^*(s'|s)$ denote the transition kernel when using $d^*(\cdot)$

$$p^*(s'|s)=p(s'|s,d^*(s))$$

Assume that $\{(X_t,Y_t)\}_{t\geq 0}$ is generated by π^*. Then we know the following

i). $Y_t=d^*(X_t)\ \forall\ t\in T$.

ii). $\{X_t\}_{t\geq 0}$ is a homogeneous Markov chain.

iii). $p^*(s'|s)$ is the transition kernel for $\{X_t\}_{t\geq 0}$

Thus, $\{X_t\}_{t\geq 0}$ is an irreducible Markov chain. Meaning

i). $\{X_t\}_{t\geq 0}$ has a unique invariant pmf $\mu^*(s)$.

ii). $\mu^*(s)>0\ \forall\ s\in S$.

By the definition of an invariant pmf, we have

$$\mu^*(s)=\sum_{s'\in S}p^*(s|s')\mu(s')$$

Consider 2 and calculate

$$\begin{array}{rl} = & \sum_{s\in S}\mu^*(s)\left\{[\tilde{w}^*(s)-w^*(s)]-\sum_{s'\in S}[\tilde{w}^*(s')-w^*(s')]p^*(s'|s)\right\} \\ = & \sum_{s\in S}\mu^*(s)[\tilde{w}^*(s)-w^*(s)]-\sum_{s\in S}\mu^*(s)\sum_{s'\in S}[\tilde{w}^*(s')-w^*(s')]p^*(s'|s) \\ = & \sum_{s\in S}\mu^*(s)[\tilde{w}^*(s)-w^*(s)]-\sum_{s'\in S}[\tilde{w}^*(s')-w^*(s')]\sum_{s\in S}\mu^*(s)p^*(s'|s) \\ = & \sum_{s\in S}\mu^*(s)[\tilde{w}^*(s)-w^*(s)]-\sum_{s'\in S}[\tilde{w}^*(s')-w^*(s')]\mu^*(s') \text{ by def. invariant pmf} \\ = & 0 \end{array}$$

Since $\mu^*(s)>0\ \forall\ s$ we have that

$$[\tilde{w}^*(s)-w^*(s)]-\sum_{s'\in S}[\tilde{w}^*(s')-w^*(s')]p^*(s'|s)=0$$

Define the following functions

$$\begin{array}{rcl} f(s) & = & 0 \\ \check{f}(s) & = & 0 \\ \check{f}'(s) & = & \tilde{w}^*(s) - w^*(s) \\ \bar{f} & = & \displaystyle\sum_{s\in S} f(s)\mu^*(s) \end{array}$$

Using these functions, we can restate the expression above as

$$\begin{array}{rcl} f(s) - \bar{f} & = & \check{f}'(s) - \displaystyle\sum_{s\in S} \check{f}'(s')p^*(s'|s) \\ \text{and } f(s) - \bar{f} & = & \check{f}(s) - \displaystyle\sum_{s\in S} \check{f}(s')p^*(s'|s) \end{array}$$

These are the Poisson equation for $\{X_t\}_{t\geq 0}$ and $f(s)$ where $\check{f}(s), \check{f}'(s)$ are solutions to the Poisson equations.

As they are both solutions, then

$$\exists c \in \mathbb{R},\ \check{f}'(s) - \check{f}(s) = c\ \forall\ s \in S$$

This means that

$$\exists\ c \in \mathbb{R},\ \tilde{w}^*(s) = w^*(s) + c\ \forall\ s \in S$$

Question 7) - AR-MDPs are DR-MDPs where $\alpha = 1$
This question is from "Revision(Live) Lecture 1" (`RevisionSlides1SO.pdf`).
Consider a general *Infinite-Horizon MDP* with static transition probabilities and rewards.

Assume the state sequence $\{X_t\}_{t\geq 0}$ is an irreducible Markov chain under any stationary, Markovian, deterministic policy.

Let $(r^*, w^*(\cdot))$ be solutions to the *Bellman Equation* for *Average Reward MDPs*

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

Let $v^*_\alpha(\cdot)$ be a solution to the *Bellman Equation* for *Discounted Reward MDPs*, with discount factor $\alpha \in (0,1)$

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

Question 7) (a)
Show that

$$r^* = \lim_{\alpha\to 1}(1-\alpha)v^*_\alpha(s) \ \forall\ s \in S$$

and show

$$\exists\ c \in \mathbb{R} \text{ st } w^*(s) = \lim_{\alpha\to 1}\left(v^*_\alpha(s) - \frac{r^*}{1-\alpha}\right) + c \ \forall\ s \in S$$

Answer 7) (a)
By the question, $\exists\ \varepsilon \in (0,1)$ and a *Markovian Deterministic Decision Function* $d^*(\cdot)$ st

$$d^*(s) \in \text{argmax}_{a\in A(s)} \left(r(s,a) + \alpha \sum_{s'\in S} v^*_\alpha(s')p(s'|s,a) \right) \ \forall\ \alpha \in (\varepsilon, 1),\ s \in S$$

$d^*(s)$ is an optimal *Markovian Decision Function* for a *Discounted Reward MDP* when $\alpha \in (\varepsilon, 1)$.

Let π^* be a stationary policy which applies $d^*(s)$ every epoch. This is an optimal policy for a *Discounted Reward MDP* when $\alpha \in (\varepsilon, 1)$.

Assume $\{(X_t, Y_t)\}_{t\in T}$ is a generated by policy π^* and define $p^*(s'|s)$ to be the transition kernel when using $d^*(s)$.

$$p^*(s'|s) := p(s'|s, d^*(s))$$

By the definition of value functions for *Discounted Reward MDPs* we have that

$$v^*_\alpha(s) = v^{\pi^*}_\alpha(s) \ \forall\ s \in S,\ \alpha \in (\varepsilon, 1)$$

We can deduce that

i). $Y_t = d^*(X_t)\ \forall\ t \in T$.

ii). $\{X_t\}_{t\geq 0}$ is a homogeneous Markov chain.

iii). $p^*(s'|s)$ is the transition kernel for $\{X_t\}_{t\geq 0}$

This means that $\{X_t\}_{t\geq 0}$ is an irreducible Markov chain and thus has a unique invariant pmf $\mu^*(s)$.

Define the following functions

$$\begin{array}{rcl} r^*(s) & := & r(s,d^*(s)) \\ \bar{r}^* & = & \sum_{s\in S} r^*(s)\mu^*(s) \end{array}$$

Let $\tilde{u}^*(s)$ be a function which satisfies the Poisson equation for $\{X_t\}_{t\geq 0}$, associated with $r^*(s)$

$$\tilde{u}^*(s) - \sum_{s'\in S} \tilde{u}^*(s)p^*(s'|s) = r^*(s) - \bar{r}^*,\ \forall\ s\in S$$

Define $u^*(s)$ to be the zero-mean version of $\tilde{u}^*(s)$.

$$u^*(s) = \tilde{u}^*(s) - \sum_{s'\in S}\tilde{u}^*(s')\mu^*(s')$$

$u^*(s)$ is still a solution of the Poisson equation, and thus its expected value wrt $\mu^*(s)$ is zero

$$\sum_{s\in S} u^*(s)\mu^*(s) = 0$$

Let $v_\alpha(s)$ be the α resolvent of $X_t\}_{t\geq 0}$, wrt $r^*(s)$, and $\tilde{v}^\alpha$ be the residual of the first order *Laurent Expansion* of $v_\alpha(s)$

$$\begin{array}{rcl} v_\alpha(s) & = & \mathbb{E}^{\pi^*}\left[\sum_{t=0}^\infty \alpha^t r^*(X_t)\big|X_0 = s\right] \\ \tilde{v}_\alpha(s) & = & v_\alpha(s) - \left[\frac{\bar{r}^*}{1-\alpha} + \mu^*(s)\right] \end{array}$$

Then

$$\lim_{\alpha\to 1}\tilde{v}_\alpha(s) = 0\ \forall\ s\in S$$

By rearranging the definition of $\tilde{v}_\alpha(s)$ we have the following

$$\begin{array}{rcl} v_\alpha(s) & = & \frac{\bar{r}^*}{1-\alpha} + u^*(s) + \tilde{v}_\alpha(s) \\ \bar{r}^* & = & (1-\alpha)v_\alpha(s) - (1-\alpha)[u^*(s)+\tilde{v}_\alpha(s)] \\ u^*(s) & = & \left[v_\alpha(s) - \frac{\bar{r}^*}{1-\alpha}\right] - \tilde{v}_\alpha(s) \end{array}$$

By taking limits we have

$$\begin{array}{rcl} \bar{r}^* & = & \lim_{\alpha\to 1}(1-\alpha)v_\alpha(s)\ \forall\ s\in S \\ u^*(s) & = & \lim_{\alpha\to 1}\left[v_\alpha(s) - \frac{\bar{r}^*}{1-\alpha}\right]\ \forall\ s\in S \end{array}$$

Since $Y_t = d^*(X_t)$ we have that

$$\begin{array}{rcl} v_\alpha^{\pi^*}(s) & = & \mathbb{E}^{\pi^*}\left[\sum_{t=0}^\infty \alpha^t r(X_t,Y_t)\Big|X_0 = s\right] \\ & = & \mathbb{E}^{\pi^*}\left[\sum_{t=0}^\infty \alpha^t r(X_t,d^*(X_t))\Big|X_0 = s\right] \\ & = & \mathbb{E}^{\pi^*}\left[\sum_{t=0}^\infty \alpha^t r^*(X_t)\Big|X_0 = s\right] \\ & = & v_\alpha(s) \end{array}$$

Thus
$$\bar{r}^* = \lim_{\alpha\to 1}(1-\alpha)v^*_\alpha(s)$$
By the *Bellman Equations* we have
$$\begin{array}{rrcl}
 & v^*_\alpha(s) & = & \max_{a\in A(s)}\left\{r(s,a)+\alpha\sum_{s'\in S}v^*_\alpha(s')p(s'|s,a)\right\}\\
 & & = & \max_{a\in A(s)}\left\{r(s,a)+\alpha\sum_{s'\in S}v_\alpha(s')p(s'|s,a)\right\}\\
 & & = & \max_{a\in A(s)}\left\{r(s,a)+\alpha\sum_{s'\in S}\left[\frac{\bar{r}^*}{1-\alpha}+u^*(s')+\tilde{v}_\alpha(s')\right]p(s'|s,a)\right\}\\
 & & = & \max_{a\in A(s)}\left\{r(s,a)+\frac{\alpha\bar{r}^*}{1-\alpha}+\alpha\sum_{s'\in S}\left[u^*(s')+\tilde{v}_\alpha(s')\right]p(s'|s,a)\right\}\\
 & & = & \frac{\alpha\bar{r}^*}{1-\alpha}+\max_{a\in A(s)}\left\{r(s,a)+\alpha\sum_{s'\in S}\left[u^*(s')+\tilde{v}_\alpha(s')\right]p(s'|s,a)\right\}\\
\implies & v^*_\alpha(s)-\frac{\alpha\bar{r}^*}{1-\alpha} & = & \max_{a\in A(s)}\left\{r(s,a)+\alpha\sum_{s'\in S}\left[u^*(s')+\tilde{v}_\alpha(s')\right]p(s'|s,a)\right\}\\
\implies & v^*_\alpha(s)-\frac{\alpha\bar{r}^*}{1-\alpha} & = & \\
\implies & \frac{\bar{r}^*}{1-\alpha}+u^*(s)+\tilde{v}_\alpha(s)-\frac{\alpha\bar{r}^*}{1-\alpha} & = & \\
\implies & \bar{r}^*+u^*(s)+\tilde{v}_\alpha(s) & = &
\end{array}$$
Noting that $\lim_{\alpha\to 1}\tilde{v}_\alpha(s)=0$, we find that as $\alpha\to 1$
$$\bar{r}^*+u^*(s)=\max_{a\in A(s)}\left\{r(s,a)+\sum_{s'\in S}u^*(s')p(s'|s,a)\right\}$$
By defining $w^*(s)=u^*(s)\ \forall\ s\in S$ we get the expression of the *Bellman Equation* for *Average Reward MDPs*
$$\bar{r}^*+w^*(s)=\max_{a\in A(s)}\left\{r(s,a)+\sum_{s'\in S}w^*(s')p(s'|s,a)\right\}$$
This means that $(r^*,w^*(s))$ and $(\bar{r}^*,u^*(s))$ are solutions to the equivalent bellman equations.

As shown before for *Average Reward MDPs*, $r^*=\bar{r}^*\implies r^*=\lim_{\alpha\to 1}(1-\alpha)v^*_\alpha(s)$ and $\exists\ c\in\mathbb{R}$ st $w^*(s)=u^*(s)+c$, further
$$\begin{array}{rcl}
w^*(s) & = & u^*(s)+c\\
 & = & \lim_{\alpha\to 1}\left[v_\alpha(s)-\frac{\bar{r}^*}{1-\alpha}\right]+c \text{ by result of } u^*(s)\\
 & = & \lim_{\alpha\to 1}\left[v^*_\alpha(s)-\frac{\bar{r}^*}{1-\alpha}\right]+c \text{ by result of } v_\alpha(s)
\end{array}$$