Stochastic Optimisation - Notes

Dom Hutchinson

December 18, 2020

NOTE - *Markov Chain* typically refers to the discrete setting; whilst *Markov Process* typically refers to the continuous setting.

Contents

1 Multi-Armed Bandit

1.1 The Problem

Example 1.1 - *Motivating Example*
Consider having a group of patients and several treatments they could be assigned to. How best do you go about determining which treatment is best? The obvious approach is to assign some of the patients randomly and then assign the rest to the best treatment, but how much evidence is sufficient? And how likely are you to choose a sub-optimal treatment?

Definition 1.1 - *Multi-Armed Bandit Problem*
An agent is faced with a choice of K actions. Each (discrete) time step the agent plays action i they receive a reward from the random real-valued distribution ν_i. Each reward is independent of the past. The distributions $\nu_1, \dots, \nu_K$ are unknown to the agent.
In the *Multi-Armed Bandit Problem* the agent seeks to maximise a measure of long-run reward.

Remark 1.1 - *Informal Definition of Multi-Armed Bandit Problem*
Given a finite set of actions and a random reward for each action, how best do we learn the reward distribution and maximise reward in the long-run.

Definition 1.2 - *Formal Definition of Multi-Armed Bandit Problem*
Consider a sequence of (unknown) mutually independent random variables $\{X_i(t)\}_{i\in[1,K]}$, with $t \in \mathbb{N}$. Consider $X_i(t)$ to be the distribution of rewards an agent would receive if they performed action i at time t. Since the rewards are independent of the past $X_i(t), X_i(t+1), \dots$ are IID random variables. The *Multi-Armed Bandit Problem* tasks us to find the greatest expected reward from all the actions.

$$\mu^* := \max_{i=1}^{K} \mu_i \quad \text{where } \mu_i = \mathbb{E}(X_i(t))$$

There are a number of ways to formalise this objective.

Definition 1.3 - *Strategy, $I(\cdot)$*
Our agent's strategy $I : \mathbb{N} \to [1, K]$ is a function which determines which action the agent shall make at a given point in time. The strategy can use the knowledge gained from previous actions & their rewards only.

$$I(t) = I\big(t, \underbrace{\{I(s)\}_{\in[1,t)}}_{\text{Prev. Actions}}, \underbrace{\{X_{I(s)}(s)\}_{\in[1,t)}}_{\text{Prev. Rewards}}\big) \in [1, K]$$

Definition 1.4 - *Long-Run Average Reward Criterion, X_**
For a strategy $I(\cdot)$ we define the following measure for *Long-Run Average Reward*

$$X_* = \lim_{T\to\infty} \inf \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}(X_{I(t)})$$

The *Infinum* is taken as there is no guarantee the limit exists (depending on the strategy), typically we will only deal with strategies where this limit exists.
Most strategies as based only on realisations of $\{X_i(s)\}_{s\in[1,t)}$, thus $\mathbb{E}(X_{I(t)}) \leq \mu^*$ and thus $X_* \leq \mu^*$. A strategy $I(\cdot)$ is *Optimal* if $X_* = \mu^*$.

Remark 1.2 - *It is not hard to find an Optimal Strategy in the (very) long run, so we are going to look at Regret Minimisation First.*
Proposition 1.1 - *Mathematical Model & Assumptions for Multi-Armed Bandit Problem*
Model:

- Bandit has K bernoulli arms.
- $X_i(t) \in \mathbb{R}$ is the reward obtained by played arm $i \in [1, K]$ at time step $t \in \mathbb{N}$.

Assumptions:

- $X_1(\cdot), X_2(\cdot), \dots$ are mutually independent sequences.
- For each i $\{X_i(t)\}_{t\in\mathbb{N}}$ is a sequence of iid Bern(μ_i) random variables

We define the following quantities to make analysis easier

- $I(t) \in [1, K]$. The index of the arm played in time t;
- $N_j(t) := \sum_{s=1}^{t} \mathbb{1}(I(s) = j)$. The number of times arm j has been played in the first t rounds;
- $S_j(t) := \sum_{s=1}^{t} X_j(s)\mathbb{1}(I(s) = j)$. The total reward from arm j in the first t rounds. This is a Binomial random variable independent Bin($N_j(t), \mu_j$);
- $\hat{\mu}_{j,n} := \frac{S_j(t)}{N_j(t)}$. The sample mean reward from arm j in the first n plays of arm j.

Definition 1.5 - *Policy*
A *policy* is a family of functions f_t which specify what arm is to be played in round t. f_t should depend on the information available at time t $\big\{I(s), X_{I(s)}(s) : s \in [1, t-1]\big\}$.
Randomised policies are allowed. So, in addition to the history up to time t, f_t can depend upon a $U(t) \sim U[0, 1]$ random variable which is independent of $X_i(\cdot)$. Thus

$$I(t) = f_t\big(\underbrace{I(1), \dots, I(t-1)}_{\text{arms chosen}}, \underbrace{X_{I(1)}(1), \dots, X_{I(t-1)}(t-1)}_{\text{observed rewards}}, \underbrace{U(t)}_{\text{randomness}}\big)$$

We want to find the best policy (ie one which minimises the regret)

1.2 Regret Minimisation

Definition 1.6 - *Regret, R_n*
Regret is a measure of how much reward was lost during the first n time steps. The *Regret* R_n of a strategy $\{I(t)\}_{t\in\mathbb{N}}$ in the first n time steps is given by

$$\begin{aligned} R_n &= \max_{k=1}^{K} \sum_{t=1}^{n} \mathbb{E}[\underbrace{X_k(t)}_{\text{Best Pos}} - \underbrace{X_{I(t)}(t)}_{\text{Actual}}] \\ &= n\mu^* - \sum_{t=1}^{n} \mathbb{E}\big[X_{I(t)}(t)\big] \end{aligned}$$

Regret only involves expectation and thus can be learnt from observations. We want to produce a strategy where *Total Regret* grows sub-linearly.(i.e. $R_T/T \overset{T\to\infty}{\longrightarrow} 0$)

Remark 1.3 - *Minimising the growth rate of R_T with T is quite hard.*
The best achievable regret scales as $R_T \sim c \log T$ (i.e. $R_T / c \log T \overset{T\to\infty}{\longrightarrow} 1$) where c depends on the reward distributions $X_1(t), \dots, X_K(t)$.

Definition 1.7 - *Pseudo-Regret, $\tilde{R}_n$*
Pseudo-Regret $\tilde{R}_n$ is a less popular alternative to *Regret* R_n. The *Pseudo-Regret* R_n of a strategy $\{I(t)\}_{t\in\mathbb{N}}$ in the first n time steps is given by

$$\tilde{R}_n = \max_{k=1}^{K} \sum_{t=1}^{n} \big(X_k(t) - X_{I(t)}(t)\big)$$

Pseudo-Regret includes intrinsic randomness (which is independent of the past) and thus cannot be learnt from observations.

1.3 Best Arm Identification for Bernoulli Distribution

Example 1.2 - *Best Arm Identification for Bernoulli Bandits*
Consider a bandit with two *Bernoulli* arms: $\{X_1(t)\}_{t\in\mathbb{N}}$ IID RVs with distribution Bern(μ_1); and, $\{X_2(t)\}_{t\in\mathbb{N}}$ IID RVs with distribution Bern(μ_2).
Suppose $\mu_1 > \mu_2$ (i.e. arm 1 is better). Let the player play each arm n times and declare the arm with the greatest empirical mean to be the better arm. *What is the probability of choosing the wrong arm (Arm 2)?.*

An error occurs if $\sum_{t=1}^n X_2(t) \geq \sum_{t=1}^n X_1(t)$ and thus we want to calculate the probability of this event.
Define $\{Y(t)\}_{t\in\mathbb{N}}$ st $Y(t) := \{X_2(t) - X_1(t)$. This means $Y(t) \in \{-1, 0, 1\} \subset [-1, 1]$.
To use *Hoeffding's inequality* we need to scale Y to be in $[0, 1]$, so we define $Z(t) := \frac{1}{2}(Y(t) + 1)$. We have $\mathbb{E}(Z(t)) = \frac{1}{2}(1 + \mu_2 - \mu_1)$ and an error occurs if $\sum_{t=1}^n Y(t) > 0 \iff \sum_{t=1}^n Z(t) \geq \frac{n}{2}$. By *Hoeffding's Inequality*

$$\begin{aligned}
\mathbb{P}(\text{error}) &= \mathbb{P}\left(\sum_{i=1}^{n} Z(t) \geq \frac{n}{2}\right) \\
&= \mathbb{P}\left(\left(\sum_{i=1}^{n} Z(t)\right) - \frac{n}{2}(1 + \mu_2 - \mu_1) \geq \frac{n}{2}(\mu_1 - \mu_2)\right) \quad \text{subtracting } \mu \text{ from both sides} \\
&= \mathbb{P}\left(\sum_{i=1}^{n} \left(X_i - \underbrace{\frac{1}{2}(1 + \mu_2 - \mu_1)}_{\mu}\right) \geq n \underbrace{\frac{1}{2}(\mu_1 - \mu_2)}_{t}\right) \quad \text{arranging for Hoeffding's} \\
&\leq \exp\left(-2n \cdot \tfrac{1}{4}(\mu_1 - \mu_2)^2\right) \quad \text{by Hoeffding's Inequality} \\
&= \exp\left(-\tfrac{n}{2}(\mu_1 - \mu_2)^2\right)
\end{aligned}$$

1.4 Heuristic

Remark 1.4 - *How many tests?*
Suppose an agent is comparing two arms and is given a finite time horizon T after in which they must choose the best arm. The obvious strategy is to perform each task N times and then choose the arm with the greatest empirical mean. But, how do we choose N to minimise

regret over time T?

Proposition 1.2 - *Naïve Heuristic (Single Test)*
Consider a 2-armed bandit & the following Heuristic

Play each arm once. Pick the arm with the greatest sample mean reward (breaking ties arbitrarily) and playing that arm on all subsequent rounds.

This heuristic picks the wrong arm with probability $\mu_2(1-\mu_1)$. In this case the wrong arm is played $T-1$ times, giving a bounded regret

$$\mathcal{R}(T) \geq \underbrace{\mu_2(1-\mu_1)}_{\text{prob of wrong choice}} \cdot \underbrace{(\mu_1-\mu_2)}_{\text{Loss}} \cdot \underbrace{(T-1)}_{\#\text{ steps}}$$

This regret grows linearly in T.

Theorem 1.1 - *Chernoff Bound of a Binomial Random Variable*
Let $X \sim \text{Bin}(n, \alpha$ with $n \in \mathbb{N},\ \alpha \in (0,1)$. Then

$$\forall\ \beta > \alpha \quad \mathbb{P}(X \geq Xn) \leq e^{-nK(\beta;\alpha)}$$

where

$$K(\beta;\alpha) := \begin{cases} \beta \ln\left(\frac{\beta}{\alpha}\right) + (1-\beta)\ln\left(\frac{1-\beta}{1-\alpha}\right) & \text{if } \beta \in [0,1] \\ +\infty & \text{otherwise} \end{cases}$$

with $x\ln(x) := 0$ if $x = 0$.
Similarly

$$\forall\ \beta < \alpha \quad \mathbb{P}(X \leq \beta n) \leq e^{-nK(\beta;\alpha)}$$

Note that $K(\cdot;\cdot)$ is known as both *relative entropy* and *Kullback-Leibler Divergence*

Proposition 1.3 - *Better Heuristic (N Tests)*
Consider a 2-armed bandit problem & the following heuristic

Play each arm $N < \frac{T}{2}$. Pick the arm with the greatest sample mean reward (breaking ties arbitrarily) and playing that arm on all subsequent rounds.

Note that $S_1(n)$ & $S_2(n)$ are *binomial* random variables with distributions $\text{Bin}(N,\mu_1)$, $\text{Bin}(N,\mu_2)$ respectively. And, $S_1(n)$ and $S_2(n)$ are independent of each other. Thus for $\beta \in (\mu_2, \mu_1)$

$$\mathbb{P}(S_1(N) < \beta N,\ S_2(N) > \beta N) \quad \leq \quad e^{-N(K(\beta;\mu_1)+K(\beta;\mu_2))} = e^{-NJ(\mu_1,\mu_2)}$$

where

$$J(\mu_1,\mu_2) = \inf_{\beta\in[\mu_2,\mu_1]} \big(K(\beta;\mu_1) + K(\beta;\mu_2)\big)$$

The values of β which solve $J(\cdot;\cdot)$ describe the most likely ways for the event $(S_1(N) < S_2(N)$ to occur (ie the wrong decision is made).

Proposition 1.4 - *Optimal N*
For the situation described in `Proposition 1.2` we want to find N which minimises regret, given a total time horizon of T.

If the right decision is made in the end, regret only occurs during exploration and is equal to $N \cdot (\mu_1 - \mu_2)$ (since the wrong arm is played N times).

However, if the wrong decision is made in the end, regret is equal to $(T - N) \cdot (\mu_1 - \mu_2)$.

Thus, the overall regret up to time T is

$$\begin{aligned}\mathcal{R}(T) &= \underbrace{(T-2N)(\mu_1-\mu_2)\mathbb{P}\big(S_1(N) < S_2(N)\big)}_{\text{if wrong decision made}} + \underbrace{N(\mu_1-\mu_2)}_{\text{guaranteed regret}} \\ &\simeq (\mu_1-\mu_2)(N + Te^{-NJ(\mu_1,\mu_2)})\end{aligned}$$

This expression is minimised for N close to the solution of $1 = TJ(\mu_1, \mu_2)e^{-NJ(\mu_1,\mu_2)}$ (ie when $N = \frac{\ln T}{J(\mu_1,\mu_2)} + O(1)$).
The corresponding regret is

$$\mathcal{R}(T) = \frac{\mu - 1 - \mu_2}{J(\mu_1, \mu_2)} \ln(T) + O(1)$$

If $\mu_1 \simeq \mu_2$ then $J(\mu_1, \mu_2) \simeq (\mu_1 - \mu - 2)^2$ and the above regret becomes $\mathcal{R}(T) = \frac{\ln(T)}{\mu_1 - \mu_2} + O(1)$.

1.5 UCB Algorithm

Remark 1.5 - *UCB Algorithm*
The *Upper Confidence Bound Algorithm* is a *frequentist* algorithm for solving the multi-armed bandit problem.

Remark 1.6 - *Motivation*
The problem with the heuristics in `Proposition 1.2,1.3` is that they treat the sample mean as the true mean (*Certainty Equivalence*), which is not great.

Suppose we observed sample mean reward for arm i of $\hat{\mu}_{i,n}$ after n plays. How far from the true value can μ_i be?

$$\mathbb{P}(\mu_i > \hat{\mu}_{i,n} + x) \leq e^{-2nx^2} \text{ by Hoeffding's Inequality}$$

Suppose the inequality holds with equality (ie greatest possible probability). Then for some chosen $\delta \in [0, 1]$

$$x = \sqrt{\frac{1}{2n} \ln\left(\frac{1}{\delta}\right)} \implies \mathbb{P}(\mu_i > \hat{\mu}_{i,n} + x) = \delta \quad \text{since } \delta = e^{-2nx^2}$$

This suggests a heuristic:

$$\text{Play arm which maximises} \quad \hat{\mu}_{i,N_i(t)} + \sqrt{\frac{1}{2N_i(t)} \ln\left(\frac{1}{\delta}\right)}$$

where you choose $\delta \in [0, 1]$ based on how lucky you feel. This quantity is the upper bound of a $1 - \delta$ confidence interval for the value of μ_i.

This heuristic allows for our choice to be changed any number of times.

Definition 1.8 - *UCB(α) Algorithm*
Consider the set up of the multi-armed bandit problem in `Proposition 1.1` and wlog that $\mu_1 > \mu_2 \geq \cdots \geq \mu_K$.
Consider a k-armed bandit and let $\alpha > 0$.

i). In the first K rounds, play each arm once.

ii). At the end of each round $t \geq K$ compute the *UCB(α)* index of each arm i defined as

$$\hat{\mu}_{i,N_i(t)} + \sqrt{\frac{\alpha \ln(t)}{2N_i(t)}}$$

iii). In round $t+1$ play the arm with the greatest index (breaking ties arbitrarily)

$$I(t+1) = \text{argmax}_{i \in [1,K]} \left\{ \hat{\mu}_{i,N_i(t)} + \sqrt{\frac{\alpha \ln(t)}{2N_i(t)}} \right\}$$

1.5.1 Analysis

Theorem 1.2 - *Upper Bound on Regret*
Consider a K-armed bandit and define $\Delta_i := \mu_1 - \mu_i$.
If the *UCB(α)* algorithm is used, with $\alpha > 1$, then the regret in the first T rounds is bounded above by

$$\mathcal{R} \leq \sum_{i=2}^{K} \left(\frac{\alpha+1}{\alpha-1} \Delta_i + \frac{2\alpha}{\Delta_i} \ln(T) \right)$$

This bounds grows logarithmically in T, which is very good.

If α is taken to be large, then the regret grows faster (bad). If α is small, the constant term dominates for smaller values of T (constant term blows up close to 1).

You should choose a value a bit larger than 1 (often $\alpha = 2$).

NOTE this is proved at the end of this subsubsection `Proof 1.4`.

Theorem 1.3 - *When a sub-optimal arm is played*
Consider apply *UCB(α)* to a k-armed bandit and define $\Delta_i := \mu_1 - \mu_i$. Let $s \geq K$ (so we have completed the first stage of UCB) and suppose $I(s+1) = j \neq 1$ (ie arm at time $s+1$ is suboptimal). Then one of the following is true:

i). $\hat{\mu}_{1,N_1(s)} \leq \mu_1 - \sqrt{\frac{\alpha \ln(s)}{2N_1(s)}}$. The sample mean reward on the optimal arm is much smaller than the true mean.

ii). $\hat{\mu}_{j,N_j(s)} \geq \mu_j + \sqrt{\frac{\alpha \ln(s)}{2N_j(s)}}$. The sample mean reward on arm j is much larger than its true mean.

iii). $N_j(s) < \frac{2\alpha \ln(s)}{\Delta_j^2}$. Arm j has been played very few times.

Proof 1.1 - *Theorem 1.3*
This is a proof by contradiction.

Suppose $I(s+1) = j \neq 1$ but that none of the three inequalities holds. Then

$$\begin{aligned}
\underbrace{\hat{\mu}_{1,N_1(s)} + \sqrt{\frac{\alpha \ln(s)}{2N_1(s)}}}_{\text{UCB}(\alpha) \text{ index 1}} &> \mu_1 && \text{by not i)} \\
&= \mu_j + \Delta_j && \text{by def. of } \Delta_j \\
&\geq \mu_j + \sqrt{\frac{2\alpha \ln(s)}{N_j(s)}} && \text{by not iii)} \\
&\geq \hat{\mu}_{1,N_1(s)} - \sqrt{\frac{\alpha \ln(s)}{2N_1(s)}} + \sqrt{\frac{2\alpha \ln(s)}{N_j(s)}} && \text{by not ii)} \\
&\geq \hat{\mu}_{1,N_1(s)} + \left(\sqrt{2} - \tfrac{1}{\sqrt{2}}\right) \sqrt{\frac{\alpha \ln(s)}{N_1(s)}} \\
&= \underbrace{\hat{\mu}_{j,N_j(s)} + \sqrt{\frac{\alpha \ln(s)}{2N_j(s)}}}_{\text{UCB}(\alpha) \text{ index j}}
\end{aligned}$$

But, this implies that the *UCB(α)* index of arm 1 at the end of round s is greater than that of arm j. Hence arm j would not be played in time slot $s+1$. □

Theorem 1.4 - *Counting Lemma*
Let $\{I(t)\}_{t\in\mathbb{N}}$ be a $\{0,1\}$-valued sequence and $N(t) := \sum_{s=1}^{t} I(s)$. Then

$$\forall\, t, u \in \mathbb{N} \quad N(t) \leq u + \sum_{s=u+1}^{t} I(s) \mathbb{1}\{N(s-1) \geq u\}$$

with an empty sum defined to be zero.

Proof 1.2 - *Theorem 1.4*
Fix $t, u \in \mathbb{N}$. There are two possibilities

Case 1 $N(t) \leq u$. (Have not reached u yet)

Case 2 $\exists\ s \in [1, t]$ st $N(s) > u$. (Already reached u). Let s^* denote the smallest such s. Then it must be true that $N(s^* - 1) = u$ and $s^* \geq u + 1$. Hence

$$\begin{aligned}
N(t) &= \sum_{s=1}^{s^*-1} I(s) + \sum_{s=s^*}^{t} I(s) \\
&= N(s^* - 1) + \sum_{s=s^*}^{t} I(s) \underbrace{\mathbb{1}\{N(s-1) \geq u\}}_{\text{true for all in sum}} \\
&\leq u + \sum_{s=u+1}^{t} I(s) \mathbb{1}\{N(s-1) \geq u\} \quad \text{since } s^* \geq u+1
\end{aligned}$$

□

Proof 1.3 - *Theorem 1.2*
Fix $t \in \mathbb{N}$ adn take $u_{t,j} := \left\lceil \frac{2\alpha \ln(t)}{\Delta_j^2} \right\rceil$.
By `Theorem 1.4` we have that

$$N_j(t) \leq u + \sum_{s=u+1}^{t} \mathbb{1}\{(N_j(s-1) \geq u_{t,j}) \ \&\ (I(s) = j)\}$$

Both sides involve random variables. Taking expectations we get

$$\mathbb{E}[N_j(t)] \leq u + \sum_{s=u}^{t-1} \mathbb{P}\big\{(N_j(s) \geq u_{y,j}) \;\&\; (I(s+1)=j)\big\}$$

By `Theorem 1.3` and the definition of u, *IF* $I(s+1)=j$ *and* $N_j(s) \geq u$ *then*

$$\hat{\mu}_{1,N_1(s)} \leq u_1 - \sqrt{\frac{\alpha \ln(s)}{2N_1(s)}} \quad \text{or} \quad \hat{\mu}_{j,N_j(s)} > \mu_j + \sqrt{\frac{\alpha \ln(s)}{2N_j(s)}}$$

Thus

$$\mathbb{E}[N_j(t)] \leq u_{t,j} + \sum_{s=u_{t,j}}^{t-1} \left[\underbrace{\mathbb{P}\left(\hat{\mu}_{1,N_1(s)} \leq \mu_1 - \sqrt{\frac{\alpha \ln(s)}{2N_1(s)}}\right)}_{\hat{\mu}_1 \text{ is unusually small}} + \underbrace{\mathbb{P}\left(\hat{\mu}_{j,N_j(s)} > \mu_j - \sqrt{\frac{\alpha \ln(s)}{2N_j(s)}}\right)}_{\hat{\mu}_j \text{ is unusually large}} \right]$$

By *Hoeffding's Inequality*

$$\begin{aligned}
\mathbb{E}[N_j(t)] &\leq u + \sum_{s=u}^{t-1} 2s^{-\alpha} \\
&\leq u + \int_{u-1}^{\infty} 2s^{-\alpha} ds \quad \text{assumption } \alpha > 1 \text{ required here} \\
&= u + \frac{2(u-1)^{-(\alpha-1)}}{\alpha - 1} \\
&\leq u + \frac{2}{\alpha - 1} \quad \text{since } u \geq 2 \implies (u-1)^{-(\alpha-1)} \leq 1
\end{aligned}$$

Thus

$$\forall\, j \in [2, K] \quad \mathbb{E}[N_j(t)] \leq u + \frac{2}{\alpha - 1} \leq \frac{2\alpha \ln(t)}{\Delta_j^2} + 1 + \frac{2}{\alpha - 1}$$

A regret of $\Delta_j := \mu_1 - \mu_j$ is incurred every time arm j is played. Hence the total regret up to time t is bounded by

$$\begin{aligned}
\mathcal{R}(t) &:= \sum_{i=2}^{K} \Delta_i \mathbb{E}[N_i(t)] \\
&\leq \sum_{i=2}^{K} \left(\frac{2\alpha \ln(t)}{\Delta_i} + \frac{\alpha + 1}{\alpha - 1} \Delta_i \right)
\end{aligned}$$

□

1.5.2 Improvements?

Remark 1.7 - *The regret of UCB grows logarithmically with T. No other algorithm can do better.*
Further, the constant factor of $\ln(T)$ used is almost optimal. This shall now be shown.

Definition 1.9 - *Strongly Consistent*
A strategy for the multi-armed bandit problem is said to be *strongly consistent* if its regret satisfies $\mathcal{R}(T) = o(T^\alpha)$ for all $\alpha > 0$. (i.e. its regret grows slower than any fractional power of T).

The *UCB(α)* algorithm is strongly consistent for all $\alpha > 1$ as its regret grows logarithmically with T.

Theorem 1.5 - *Lai and Robbins*
Consider a K-armed bandit, where the rewards from arm i are iid Bern(μ_i) and rewards from distinct arms are mutually independent. Then, for any *strongly consistent* strategy, the number of times that a sub-optimal arm i is played up to time T, $N_i(T)$ satisifies

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_i(T)]}{\ln(T)} \geq \frac{1}{K(\mu_i;\mu^*)}$$

where $\mu* := \max_{i=1}^K \mu_i$ and $K(q;p)$ is the *KL-Divergence* of a Bern(q) distribution wrt a Bern(p) distribution.

Proposition 1.5 - *Lower bound on Regret*
Here we derive a lower bound for the regret of any strongly consistent strategy from the multi-armed bandit problem.

$$\begin{aligned}\liminf_{T\to\infty} \frac{\mathcal{R}(T)}{\ln(T)} &= \liminf_{T\to\infty} \frac{\sum_{i;\mu_i<\mu^*} *\mu^* - \mu_i)\mathbb{E}[N_i(T)]}{\ln(T)} \\ &\geq \sum_{i;\mu_i<\mu^*} \frac{\mu^* - \mu_i}{K(\mu_i;\mu^*)} \qquad \text{by } \texttt{Theorem 1.6}\end{aligned}$$

Proposition 1.6 - *Comparing to Lower bound of UCB(α)*
We showed that the regret of the *UCB(α)* algorithm satisfies

$$\limsup_{T\to\infty} \frac{\mathcal{R}(T)}{\ln(T)} \leq \sum_{i;\mu_i<\mu^*} \frac{2}{\mu^* - \mu_i}$$

To compare this to the result in `Proposition 1.5` we use *Pinsker's Inequality*. (Proof in homework).

We see thjat the upper bound on the regret achieved by *UCB(α)* is approximately four times greater than the lower bound on the best regret achievable by any algorithm. This is very good.

Theorem 1.6 - *Concentration Inequalities for Sample Means*

$$\begin{aligned}\mathbb{P}\left(\hat{\mu}_{j,N_j(s)} \geq u_j + \sqrt{\frac{\alpha \ln s}{2N_j(s)}}\right) &\leq e^{-\alpha \ln s} = s^{-\alpha} \\ \mathbb{P}\left(\hat{\mu}_{1,N_1(s)} \leq u_1 - \sqrt{\frac{\alpha \ln s}{2N_1(s)}}\right) &\leq e^{-\alpha \ln s} = s^{-\alpha}\end{aligned}$$

Proof 1.4 - *Theorem 1.5*
The proof is immediate from *Hoeffding's Inequality*, which is applicable since the X_j are iid and take values in $\{0,1\} \subseteq [0,1]$.

Remark 1.8 - *Is there an algorithm which achieves lower regret?*
No. There is no algorithm which has regret growing slower than $\ln(T)$.

1.6 Thompson Sampling

Remark 1.9 - *Thompson Sampling*
Thompson Sampling is a *Bayesian* algorithm for the multi-armed bandit problem. It was one of the first algorithms for solving the problem, but remains on of the best as it is asymptotically optimal.

Theorem 1.7 - *Relationship between Beta & Gamma Distribution*
Let $X \sim \text{Gamma}(\alpha, \lambda)$ and $Y \sim \text{Gamma}(\beta, \lambda)$ (ie shared scale parameter but different shape parameters). Then

$$V := \frac{X}{X+Y} \sim \text{Beta}(\alpha, \beta)$$

Proof 1.5 - *Theorem 1.7*
Consider the map $(X, Y) \mapsto \frac{X}{X+Y}$. This maps $\mathbb{R}^2 \to \mathbb{R}$ (reduces dimensions) and thus we cannot directly use the formula to compute the density of V.
We introduce an auxiliary random variable $W = X$. Now $(X, Y) \mapsto (V, W) := g(X, Y) = \left(\frac{X}{X+Y}, X\right)$. Note that X and Y are non-negative random variables. Hence, $V \in [0, 1]$ and $W \in \mathbb{R}^+$.
For $(v, w) \in [0, 1] \times \mathbb{R}^+$ we want to find all solutions of $g(x, y) = (v, w)$. Clearly $(x, y) = \left(w, w\frac{1-v}{v}\right)$ in a unique solution. The joint density of (V, W) is given by the formula

$$f_{V,W}(v, w) = \sum_{(x,y):g(x,y)=(v,w)} \frac{f_{X,Y}(x, y)}{|\det(J_g(x, y))|} \quad (1)$$

Using the density function for *Gamma* random variables and the independence of X and Y

$$f_{X,Y}(x, y) = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} y^{\beta-1} e^{-x-y}$$

Next, we compute the *Jacobian* of g and its determinant

$$J_g(x, y) = \begin{pmatrix} \frac{y}{(x+y)^2} & \frac{-x}{(x+y)^2} \\ 1 & 0 \end{pmatrix} \quad \implies \quad |\det(J_g(x, y))| = \frac{x}{(x+y)^2} = \frac{v^2}{w}$$

Substituting the definition of $f_{X,Y}$, the results of the *Jacobian* and the solution $(x, y) = \left(w, w\frac{1-v}{v}\right)$ into (1) we get

$$\begin{aligned} f_{V,W}(v, w) &= \frac{1}{v^2/w} \frac{1}{\Gamma(\alpha)\Gamma(\beta)} w^{\alpha-1} e^{-w} \left(\frac{w(1-v)}{v}\right)^{\beta-1} e^{-w(1-v)/v} \\ &= \frac{1}{\Gamma(\alpha)\Gamma(\beta)} \frac{w}{v^2} w^{\alpha-1} \left(w\frac{1-v}{v}\right)^{\beta-1} e^{-w/v} \end{aligned}$$

We are only interested in the marginal distribution of V

$$\begin{aligned} f_V(v) &= \int_{w=0}^{\infty} f_{V,W}(v,w)dw \\ &= \frac{(1-v)^{\beta-1}v^{\alpha}-1}{\Gamma(\alpha)\Gamma(\beta)} \int_{w=0}^{\infty} \Big(\frac{w}{v}\Big)^{\alpha+\beta-1} e^{-w/v}\frac{dw}{v} \\ &= \frac{(1-v)^{\beta-1}v^{\alpha-1}}{\Gamma(\alpha)\Gamma(\beta)} \int_{u=0}^{\infty} u^{\alpha+\beta-1}e^{-u}du \qquad \text{where } u := \frac{w}{v} \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} v^{\alpha-1}(1-v)^{\beta-1} \\ &\sim \text{Beta}(\alpha,\beta) \end{aligned}$$

□

Remark 1.10 - *Beta*$(1,1)$ *is equivalent to Uniform*$[0,1]$
Proposition 1.7 - *Thompson Sampling Algorithm*
Consider a K-armed bandit with independent Bernoulli arms with parameters $\mu_1,\dots,\mu_K$. *Thompson Sampling* follows the following process.

i). Define a prior distribution Beta$(1,1)$ for the parameter of each arm.

ii). At the start of round t, sample $\hat{\mu}_i(t)$ for $i \in [1,K]$, from the corresponding prior distributions.

iii). Play the arm with the greatest sample value $I(t) \in \text{argmax}_{i\in[1,K]}\hat{\mu}_i(t)$ (breaking ties arbitrarily).

iv). Compute a posterior distribution for that parameter, based on the observed reward. Beta$(\alpha+1,\beta)$ if a reward is given and Beta$(\alpha,\beta+1)$ it a reward is <u>not</u> given. (The priors for the other arms are the same as their priors as no result was observed).

v). Keep repeating ii)-iv) using the posterior distributions calculated in round t as the priors for round $t+1$.

The endpoint of the algorithm is when time runs out.

Proposition 1.8 - *Choosing the Prior Distribution for Thompson Sampling*
We use *Beta* distributions as the prior for the unknown parameters μ_i of the *Bernoulli* reward distributions. This choice is convenient because, if the prior has a *Beta* distribution then so does the posterior distribution, after observing the reward (i.e. it is a *Conjugate Prior*).

Remark 1.11 - *Beta Distributions are conjugate priors for Bernoulli Random Variables*
Remark 1.12 - *Thompson Sampling on other Reward Distributions*
Thompson Sampling can be extended to non-bernoulli reward distributions. However, the ease of implementation depends on how easy it is to sample from the posterior distribution. *Conjugate priors* should be used to make this easier. Bounds on the regret are not known in all cases.

Theorem 1.8 - *Posterior from a Beta Distribution Prior of a Bernoulli Random Variable*
Let $X \sim$ Bernoulli(μ) and assume a prior distribution $\mu \sim$ Beta(α,β). Then, the posterior distribution of μ given $X=1$ is Beta$(\alpha+1,\beta)$ and given $X=0$ is Beta$(\alpha,\beta+1)$.

Proof 1.6 - *Theorem 1.8*

Let f_μ denote the density of μ. By out assumption of a Beta(α, β) prior we have

$$f_\mu(p) = \frac{\Gamma(\alpha+\beta)}{\gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \quad \text{for } p \in [0,1]$$

Also $\mathbb{P}(X=1|\mu=p) = p = 1 - \mathbb{P}(X=0|\mu=p)$.

Hence, by *Bayes's Theorem*, the posterior density conditional on $X=1$ satisfies

$$\begin{aligned} f_\mu(p|X=1) &\propto f_\mu(p)\mathbb{P}(X=1|p) \\ &\propto p^\alpha(1-p)^{\beta-1} \end{aligned}$$

Where the constant of proportionality is determined by the fact the density integrates to 1.

We can recognise this distribution as the pdf of a Beta$(\alpha+1, \beta)$. Similarly, it can be shown that $f_\mu(\cdot|X=0)$ is the density of a Beta$(\alpha, \beta+1)$ random variable. □

Theorem 1.9 -
Let $X \sim \text{Beta}(\alpha, \beta)$ for $\alpha, \beta \in \mathbb{N}$ and $Y \sim \text{Bin}(\alpha+\beta-1, p)$ for $p \in (0,1)$. Then

$$\mathbb{P}(X > p) = \mathbb{P}(Y \leq \alpha - 1)$$

Proof 1.7 - *Theorem 1.9*
Consider this result. Let $\{N_t\}_{t\geq 0}$ be a poisson process with intensity $\lambda > 0$ and $n, t \in \mathbb{N}$.Then, conditional on the event $N_t = n$, the unordered increment times on $[0,t]$ are mutually independent and uniformly distributed on $[0.t]$.

Let $X \sim \text{Beta}(\alpha, \beta)$ then we can write $X = \frac{V}{V+W}$ where $V \sim \text{Gamma}(\alpha,)$, $W \sim \text{Gamma}(\beta, \gamma)$. If α, β are integered values then we can interpret V as the time of the α^{th} increment of a unit rate Poisson process and $(V+W)$ as the time of the $(\alpha+\beta)^{th}$ increment.

Consequently, conditional on $(V+W) = \tau$, the poisson process N_t has exactly $\alpha+\beta-1$ increments in the interval $[0, \tau)$. By the fact result above, the unordered times of these increments are iid uniformly distributed on $[0, \tau)$.

The event $\{X > p\}$ is the same as the event $\{V > p\tau\}$, conditional on $(V+W) = \tau$. This means that at most $\alpha = 1$ increments occur in $[0, p\tau]$. As the increments are IID uniform in $[0, \tau)$ and there are $\alpha+\beta-1$ in total, the number of increments in $[0, p\tau]$ has a Bin$(\alpha+\beta-1, p)$ distribution.

THus the events $\{X > p\}$, $\{V > p\tau | V+W = \tau\}$ and $\{Y \leq \alpha - 1\}$ all have the same probability. □

1.6.1 Analysis

Remark 1.13 - *Analysis of Thompson Sampling is Hard*
The main challenge is to deal with the situation where there is an initial run of bad luck on the optimal arm. This causes the posterior for the optimal arm to be biased towards small values. Hence, the optimal arm is not played very often meaning it takes a long time to recover from the initial bad luck.

For contrast, we only need to worry about plays of sub-optimal arms after they have been played sufficiently often, by which time the posterior is concentrated around the true parameter value.

Theorem 1.10 - *Bound on Regret*
The regret of *Thompson Sampling* applied to a multi-armed bandit with two *Bernoulli* arms is bounded as

$$\mathcal{R}(T) \leq \frac{40\ln(T)}{\Delta} + c$$

where Δ is the arm gap and c is some arbitrary constant which depends on Δ (but, importantly, not T).

The proof to this is not given in full, but some useful lemmas are shown.

Remark 1.14 - *Posterior Distribution over Time*
As the number of times each arm is played increases, its posterior distribution concentrates increasingly sharply around the true parameter value.

Proposition 1.9 - *Number of times wrong arm is played*
In the analysis of a multi-armed bandit with two *Bernoulli arms* we assume that the second arm (with parameter μ_2) is the worse of the two arms. Thus to bound regret it suffices to bound the number of times the second arm is played.

Fix a time horizon T and define $L = \left\lceil \frac{24\ln(T)}{\Delta^2} \right\rceil$, $\tau = \inf\{t \in [0,T] : N_2(t) \geq L\}$. Then, for $\tau \leq t \leq T$

$$\mathbb{P}\left(\theta_2(t) \geq \mu_2 + \frac{\Delta}{2}\right) \leq \frac{2}{T^3}$$

Where $\theta_i(t)$ is the value sampled from the prior of μ_i at time t.

Proof 1.8 - *Proposition 1.9*
By the definition of τ if $t \geq \tau$ then $N_2(t) \geq L$. Thus

$$\begin{aligned}
& \mathbb{P}\left(\theta_2(t) \geq \mu_2 + \frac{\Delta}{2}\right) \\
= \; & \mathbb{P}\left(\theta_2(t) \geq \mu_2 + \frac{\Delta}{2}, \frac{S_2(t)}{N_2(t)} \leq \mu_2 + \frac{\Delta}{4}\right) + \mathbb{P}\left(\theta_2(t) \geq \mu_2 + \frac{\Delta}{2}, \frac{S_2(t)}{N_2(T)} > \mu_2 + \frac{\Delta}{4}\right) \\
\leq \; & \mathbb{P}\left(\theta_2(t) \geq \mu_2 + \frac{\Delta}{2} \middle| \frac{S_2(t)}{N_2(t)} \leq \mu_2 + \frac{\Delta}{4}\right) + \mathbb{P}\left(\theta_2(t) \geq \mu_2 + \frac{\Delta}{2}, \frac{S_2(t)}{N_2(T)} > \mu_2 + \frac{\Delta}{4}\right) \quad (1)
\end{aligned}$$

We now bound these two terms.

Firstly, conditional on the number of times the second arm is player $N_2(T)$, the total reward from these plays $S_2(t)$ is the sum of $N_2(t)$ independent Bern(μ_2) random variables. Hence, using *Hoeffding's Inequality* we have

$$\mathbb{P}\left(\frac{S_2(t)}{N_2(t)} > \mu_2 + \frac{\Delta}{4} \middle| N_2(t)\right) \leq \exp\left(-2N_2(t)\frac{\Delta^2}{16}\right)$$

As we have assumed that $N_2(t) \geq L \geq \frac{1}{\Delta^2}(24\ln(T))$ we can conclude that

$$\mathbb{P}\left(\frac{S_2(t)}{N_2(t)} > \mu_2 + \frac{\Delta}{4}\right) \leq \exp\left(-3\ln(T)\right) = \frac{1}{T^3} \quad (2)$$

Next, we note that conditional on $S_2(t)$ and $N_2(t)$, the distribution of $\theta_2(t)$ is Beta($S_2(t) + 1, N_2(t) - S_2(t) + 1$). Consequently, by `Theorem 1.9`, we have that

$$\mathbb{P}\left(\theta_2(t) \geq \mu_2 + \frac{\Delta}{2}\right) = \mathbb{P}\left(\text{Bin}\left(N_2(t)+1, \mu_2 + \frac{\Delta}{2}\right) \leq S_2(t)\right)$$

Applying *Hoeffding's inequality* to the RH term, we see that for $N_2(t) \geq L$ we have

$$\begin{aligned}\mathbb{P}\left(\theta_2(t) \geq \mu_2 + \frac{\Delta}{2}\middle|\frac{S_2(t)}{N_2(t)} \leq \mu_2 + \frac{\Delta}{4}\right) &\leq \exp\left(-2\big(N_2(t)+1\big)\frac{\Delta^2}{16}\right)\\ &\leq \exp\left(-\frac{L\Delta^2}{8}\right)\\ &\leq \frac{1}{T^3} \qquad (3)\end{aligned}$$

Substituting (2) and (3) into (1), we can conclude that if $t \geq \tau$ (ie $N_2(t) \geq L$) then

$$\mathbb{P}\left(\theta_2(t) \geq \mu_2 + \frac{\Delta}{2}\right) \leq \frac{2}{T^3}$$

□

1.7 Genie Analysis

Remark 1.15 - *Genie*
Here we analyse a simpler version of Thompson Sampling for a 2-armed bandit. We assume that the value of μ_1 is known, but it is not known whether it is greater than μ_2.

This means we now only have a prior/posterior for μ_2 and we compare $\theta_2(t)$ (the value sampled from the prior of μ_2) to the true value of μ_1.

It is likely that this scenario should be more successful than the standard scenario, thus we can only find an upper bound on the regret of the normal scenario.

Theorem 1.11 - *Times Sub-optimal arm is played*
Fix $T \in \mathbb{N}$ and define $L := \left\lceil \frac{2\ln(T)}{\Delta^2} \right\rceil$, $\tau := \inf\{t \in [1,T] : N_2(t) \geq L\}$. τ is the first time that arm two has been played at least L times. The number of plays of arm two after τ is bounded as

$$\forall\ t \geq \tau \quad \mathbb{P}(\theta_2(t) \geq \mu_1) \leq \frac{2}{T}$$

This means $\mathbb{E}[\text{plays of arm two after time } \tau] = (T-\tau)\frac{2}{T} \leq 2$

Proof 1.9 - *Theorem 1.11*
Define the events

$$A_t := \{\theta_2(t) \geq \mu_1\} \quad B_t := \left\{\frac{S_2(t)}{N_2(t)} \leq \mu_2 + \frac{\Delta}{2}\right\}$$

A_t is the event that the sample from the prior of μ_2 is greater than μ_1. B_t is the event the observed rewards from arm 2 at closer to μ_2 than μ_1.

If $t \geq \tau$, then $N_2(t) \geq L$ and Hoeffding's inequality yields

$$\begin{aligned}\mathbb{P}(B_t^c) &\leq \exp\left(-2N_t\tfrac{\Delta^2}{4}\right)\\ &\leq \exp\left(-\tfrac{\Delta^2 L}{2}\right) \qquad \text{by def. } L\\ &\leq e^{-\ln(T)}\\ &= \tfrac{1}{T}\end{aligned}$$

We can bound $\mathbb{P}(A_t)$ as follows

$$\begin{aligned}\mathbb{P}(A_t) &= \mathbb{P}(A_t \cap B_t) + \mathbb{P}(A_t \cap B_t^c)\\ &= \mathbb{P}(A_t|B_t)\mathbb{P}(B_t) + \mathbb{P}(A_t|B_t^c)\mathbb{P}(B_t^c)\\ &\leq \mathbb{P}(A_t|B_t) + \mathbb{P}(B_t^c)\end{aligned}$$

We now want to bound $\mathbb{P}(A_t|B_t)$.
Since $\theta_2(t+1)$ is sampled from the posterior distribution of μ_2 after t rounds we have

$$\theta_2(t+1) \sim \text{Beta}\left(1 + \underbrace{S_2(t)}_{\#\text{ successes}}, 1 + \underbrace{N_2(t) - S_2(t)}_{\#\text{ failures}}\right)$$

Hence, by `Theorem 1.9`

$$\mathbb{P}\big(\theta_2(t+1) \geq \mu_1 \big| S_2(t), N_2(t)\big) = \mathbb{P}\big(\text{Bin}(N_2(t)+1, \mu_1) \leq S_2(t)\big)$$

By Hoeffding's inequality, if $S_2(t) < \mu_1 \cdot N_2(t)$, then

$$\mathbb{P}\big(\text{Bin}(N_2(t)+1, \mu_1) \leq S_2(t)\big) \leq \exp\left(-2N_2(t)\left(\mu_1 - \frac{S_2(t)}{N_2(t)}\right)^2\right)$$

Conditioning on the event B_t, we have

$$\begin{aligned} S_2(t) &\leq \left(\mu_2 + \tfrac{\Delta}{2}\right) N_2(t) \\ &= \left(\mu_1 - \tfrac{\Delta}{2}\right) N_2(t) \end{aligned}$$

Hence

$$\begin{aligned} \mathbb{P}(A_t|B_t, N_2(t)) &= \mathbb{P}(\theta_2(t+1) \geq \mu_1 | B_t, N_2(t)) \\ &\leq \exp\left(-\frac{2N_2(t)\Delta^2}{4}\right) \end{aligned}$$

Recall, by the definition of τ, $\forall\ t \geq \tau,\ N_2(t) \geq L$.
Hence, $\forall\ t \geq \tau$

$$\begin{aligned} \mathbb{P}(A_t|B_t) &\leq \exp\left(-\tfrac{L\Delta^2}{2}\right) \\ &\leq e^{-\ln T} \\ &= \tfrac{1}{T} \end{aligned}$$

We have already showed $\forall\ t \geq \tau,\ \mathbb{P}(B_t^c) \leq \frac{1}{T}$.

Combining these results, we get

$$\forall\ t \geq \tau, \quad \mathbb{P}(A_t) \leq \mathbb{P}(A_t|B_t) + \mathbb{P}(B_t^c) \leq \frac{2}{T}$$

This is claim on `Theorem 1.11` □

Proposition 1.10 - *Bound of Regret*
Using `Theorem 1.11` above we can bound the regret as

$$\mathcal{R}(T) \leq \Delta \cdot (L+2)$$

where $L+2$ is the most time arm two is played in the first T time steps.

2 Stochastic Dynamic Optimisation

2.1 Introduction

Remark 2.1 - *AKA*
Stochastic Dynamic Programming or Sequential Decision Making

Definition 2.1 - *Types of Optimisation Problem*

Stochastic The quantity to be optimised involves random parameters

Dynamic Actions are taken in stages, over a period of time, so later actions have more information available to them.

Functional A *Decision Function* maps from system-states to the action to be taken.

Definition 2.2 - *Stochastic Dynamic Optimisation*
Stochastic Dynamic Optimisation is a branch of probability and applied mathematics which deals with the problem of making optimal (or nearly-optimal) decisions and actions in stochastic systems.

Definition 2.3 - *Stochastic System*
A *Stochastic System* is a dynamic system whose model involves uncertain parameters which can be represented by random variables.

Definition 2.4 - *Types of Stochastic Dynamic Processes*

- *Sequential Decision Process* - An agent is able to observe the state of a stochastic system and use this information to make a decision on what action to take. The action causes the system to evolve and the agent can then observe again before making their next decision. These are *Functional Optimisation Problems*.

Proposition 2.1 - *Formulation of a Sequential Decision Problem*
In a *Sequential Decision Problem* the agent in a *Sequential Decision Process* is tasked with choosing a sequence of actions st that the system performs optimally wrt a predeterminied performance criterion.

The actions of the agent are information by the available relevant information contained in the current state, past states and past actions. Actions can be considered as functions of these sources of information.

Here are the main elements of a *Sequential Decision Problem*

- *Decision Epochs* - The periods in which decisions are taken and their effect realised.
- *System States* - An encoding of the available system information, which is relevant to picking the agents actions Y_t.
- *Agent's Actions* - Actions taken by an agent which directly affect the system.
- *Immediate Rewards/Costs* - The reward/cost the agent receives/incurs by taking this action.
- *Mathematical Model of Stochastic System* - Collection of mathematical equations describing the system within which the process is occurring. Particularly, the affect of actions.

Example 2.1 - *Motivating Example - Inventory Control Problem*
Stochastic Dynamic Optimisation was born out of the *Inventory Control Problem* (A *Sequential Decsion Problem*).

The single-item *Inventory Control Problem* considers the amount of a single item stored. A quantity of this item is ordered (by us) and sold (to customers) at discrete time periods in $[0, N)$ and the amount ordered is stochastic. We are tasked to meet this stochastic demand, while keeping costs incurred from ordering and storing to a minimum.

Let X_t denote the stock available at time t, Y_r be the stock ordered by us at time t, and Z_t be the demand from customers at time t. We assume that $Z_0, \dots, Z_{N-1}$ are IID random variables with some known distribution and that excess demand is backlogged (and fulfilled as soon as additional inventory is available).

We can derive the following equation for stock levels at each time period

$$X_{t+1} = X_t + Y_t - Z_t \text{ for } t \in [1, N)$$

if $X_t < 0$ there is a backlog of orders.

Let $c > 0$ be the cost to order each item, so the total order cost at time t is (cY_t). Let $g(|x|)$ represent the cost of holding x stock end of the time period; either as a storage cost (for $x > 0$) or a penalty for having a backlog ($x < 0$). Note that $g(\cdot)$ should be a non-negative function. The total cost over N time period is

$$\texttt{TotalCost} = \sum_{t=0}^{N-1} (cY_t + g(\underbrace{X_t + Y_t - Z_t}_{\equiv X_{t+1}}))$$

We want to minimise `TotalCost` by making appropriate choices of $Y_0, \dots, Y_{N-1}$ (given $Y_i \geq 0 \; \forall \; i$).

We cannot directly minimise the expression of `TotalCost` as it value depends on the stochastic random variables $Z_0, \dots, Z_{N-1}$. Thus we try to minimise the expected value $\mathbb{E}[\texttt{TotalCost}]$

$$\mathbb{E}[\texttt{TotalCost}] := \mathbb{E}\left(\sum_{t=0}^{N-1} (cY_t + g(X_t + Y_t - Z_t)) \right)$$

We can take either a: Static (`Proposition 2.3`); or Dynamic Approach (`Proposition 2.4`).

Proposition 2.2 - *Formulation of a Sequential Decision Problem*
Here are the specification of the components of a sequential decision problem, as specified in `Proposition 2.1`.

- *Decision Epochs* - Beginnings of the time periods $0, \dots, N-1$
- *System States* - Stock levels $X_0, \dots, X_{N-1}$
- *Agent's Actions* - Ordering new stock $Y_0, \dots, Y_{N-1}$
- *Immediate Costs* - Storage costs/penalty for underfuling orders $cY_t + g(X_t + Y_t - Z_t)$
- *Mathematical Model of Stochastic System* - A single state equation $X_{t+1} = X_t + Y_t - Z_t$ and the distributions of demand $Z_0, \dots, Z_{N-1}$.

Proposition 2.3 - *Static Approach to `Example 2.1`*
In the *Static Approach* to the *Inventory Control Problem* $Y_0, \dots, Y_{N-1}$ are treated as deterministic variables. We choose the values for $Y_0, \dots, Y_{N-1}$ at the very start (before any transactions occur) and thus are chosen before stock levels $X_0, \dots, X_{N-1}$ or sale quantities $(Z_0, \dots, Z_{N-1})$ are available. Thus this approach is inevitably sub-optimal.

Proposition 2.4 - *Dynamic Approach to `Example 2.2`*
Since there no penalty in not choosing the order quantity Y_t before time t, there is no need to take a static approach. Moreover, it is highly advantageous to delay choosing the order quantity until as much relevant information (e.g. Demand & Stock Levels) is known as possible.

In this *Dynamic Approach* we can consider Y_t to be a function of current and past stock levels $X_0, \dots, X_t$

$$Y_t = d_t(X_0, \dots, X_t)$$

We want to find *Decision Functions* $d_0(\cdot), \dots, d_{N-1}(\cdot)$ st that the *Expected Total Cost* is minimised.

2.1.1 Induced Stochastic Process

Definition 2.5 - *Induced Stochastic Process* $\{(X_t, Y_t)\}_{t\in T}$
An *Induced Stochastic Process* is the set of actions and states over a time horizon T, $\{(X_t, Y_t)\}_{t\in T}$. An *Induced Stochsatic Process* is fully specified by

i). The pmf of X_0.

ii). The transition probabilities $\{p_t(\cdot|\cdot)\}_{t\in T}$.

iii). A decision policy $\pi := \{q_t(\cdot|\cdot)\}_{t\in T}$.

Theorem 2.1 - *Marginal Distribution of Actions & States*
The induced stochastic process satisfies the following equation for all time-horizons T, state-spaces S and action-spaces A

$$\mathbb{P}(X_{0:t} = s_{0:k}, Y_{0,t-1} = a_{0,t-1}) = \mathbb{P}(X_0 = s_0) \prod_{k=0}^{t-1} \underbrace{p_k(s_{k+1}|s_k, a_k)}_{\text{transition}} \underbrace{q_k(a_k|s_{0:k}, a_{0:k-1})}_{\text{decision}}$$

In the case of a deterministic decision rule the term $q_k(a_k|s_0, \dots, s_k, a_0, \dots, a_{k-1})$ should be replaced by $\mathbb{1}\{d_k(s_{0:k}, a_{0:k-1}) = a_k\}$

Proof 2.1 - *Theorem 2.1*

$$\begin{aligned}
& \mathbb{P}^\pi(X_{0:t} = s_{0:t}, Y_{0:t-1} = a_{0:t-1}) \\
= \;& \mathbb{P}^\pi(X_0 = s_0) \prod_{k=1}^{t} \mathbb{P}^\pi(X_k = s_k, Y_{k-1} = a_{k-1}|X_{0:k-1} = s_{0:k-1}, Y_{0:k-2} = a_{0:k-2}) \text{ by chain rule} \\
= \;& \mathbb{P}(X_0 = s_0) \prod_{k=1}^{t} \frac{\mathbb{P}^\pi(X_{0:k} = s_{0:k}, Y_{0:k-1} = a_{0:k-1})}{\mathbb{P}^\pi(X_{0:k-1} = s_{0:k-1}, Y_{0:k-2} = a_{0:k-2})} \text{ by def. of conditional} \\
= \;& \mathbb{P}(X_0 = s_0) \prod_{k=0}^{t-1} \frac{\mathbb{P}^\pi(X_{0:k+1} = s_{0:k+1}, Y_{0:k} = a_{0:k})}{\mathbb{P}^\pi(X_{0:k} = s_{0:k}, Y_{0:k-1} = a_{0:k-1})} \\
= \;& \mathbb{P}(X_0 = s_0) \prod_{k=0}^{t-1} \left(\frac{\mathbb{P}^\pi(X_{0:k+1} = s_{0:k+1}, Y_{0:k} = a_{0:k})}{\mathbb{P}^\pi(X_{0:k} = s_{0:k}, Y_{0:k} = a_{0:k})} \cdot \frac{\mathbb{P}^\pi(X_{0:k} = s_{0:k}, Y_{0:k} = a_{0:k})}{\mathbb{P}^\pi(X_{0:k} = s_{0:k}, Y_{0:k-1} = a_{0:k-1})} \right)
\end{aligned}$$

I shall simplify both terms of the product separately. Consider the first

$$\begin{aligned}
& \frac{\mathbb{P}^\pi(X_{0:k+1} = s_{0:k+1}, Y_{0:k} = a_{0:k})}{\mathbb{P}^\pi(X_{0:k} = s_{0:k}, Y_{0:k} = a_{0:k})} \\
= \;& \frac{\mathbb{P}^\pi(X_{k+1} = s_{k+1}, X_{0:k} = s_{0:k}, Y_{0:k} = a_{0:k})}{\mathbb{P}^\pi(X_{0:k} = s_{0:k}, Y_{0:k} = a_{0:k})} \text{ by rearrangement} \\
= \;& \mathbb{P}^\pi(X_{k+1} = s_{k+1}|X_{0:k} = s_{0:k}, Y_{0:k} = a_{0:k}) \text{ by def. of conditional} \\
= \;& \mathbb{P}^\pi(X_{k+1} = s_{k+1}|X_k = s_k, Y_k = a_k) \text{ by Markov property} \\
= \;& p_K(s_{k+1}|s_k, a_k)
\end{aligned}$$

This is the *Transition Probability* at epoch k. Now consider the second term

$$\begin{aligned} & \frac{\mathbb{P}^\pi(X_{0:k} = s_{0:k}, Y_{0:k} = a_{0:k})}{\mathbb{P}^\pi(X_{0:k} = s_{0:k}, Y_{0:k-1} = a_{0:k-1})} \\ = & \frac{\mathbb{P}^\pi(Y_k = a_k, X_{0:k} = s_{0:k}, Y_{0:k-1} = a_{0:k-1})}{\mathbb{P}^\pi(X_{0:k} = s_{0:k}, Y_{0:k-1} = a_{0:k-1})} \text{ by rearrangement} \\ = & \mathbb{P}^\pi(Y_k = a_k | X_{0:k} = s_{0:k}, Y_{0:k-1} = a_{0:k-1}) \text{ by def. of conditional} = \\ & q_k(a_k | s_{0:k}, a_{0:k-1}) \end{aligned}$$

This is the *Decision Probability* at epoch k. By substituting these two simplifications into the original expression, we get

$$\mathbb{P}^\pi(X_{0:t} = s_{0:t}, Y_{0:t-1} = a_{0:t-1}) = P(X_0 = s_0) \prod_{k=0}^{t-1} p_k(s_{k+1} | s_k, a_k) q_k(a_k | s_{0:k}, a_{0:k-1})$$

□

Theorem 2.2 - *Markov Chains in Markov Decision Process*
Let the policy π be Markovian. Then, $\{X_t\}_{t \in T}$ and $\{(X_t, Y_t)\}_{t \in T}$ are *Markov Chains*. Further, if the transition and decision probabilities are stationary (ie independent of t), then these chains are homogeneuous.

NOTE - We only prove that the sequence of states $\{X_t\}_{t \in T}$ is a Markov Chain, see `Proof 2.2`.

Theorem 2.3 - *Transition Kernel of Sequence of States* $\{X_t\}_{t \in T}$
The sequence states $\{X_t\}_{t \in T}$ is a markov chain with transition kernel

$$\mathbb{P}(X_{t+1} = s' | X_t = s) = \sum_{a \in A} p_t(s' | s, a) q_t(a | s)$$

NOTE - This is proved in `Proof 2.2`

Proof 2.2 - *Sequence of States* $\{X_t\}_{t \in T}$ *is a Markov Chain*
To show that $\{X_t\}_{t \in T}$ is a markov chain it is sufficient to show

$$\mathbb{P}(X_{t+1} = s_{t+1} | X_{0:t} = s_{0:t}) = \mathbb{P}(X_{t+1} = s_{t+1} | X_t = s_t)$$

(ie is independent of $s_{0:t-1}$).
By the definition of conditional probabilities we have

$$\mathbb{P}(X_{t+1} = s_{t+1} | X_{0:t} = s_{0:t}) = \frac{\mathbb{P}(X_{0:t+1} = s_{0:t+1})}{\mathbb{P}(X_{0:t} = s_{0:t})}$$

By marginalising we can re-express the numerator as

$$\mathbb{P}(X_{0:t+1} = s_{0:t+1}) = \sum_{a_{0:t} \in A^{t+1}} \mathbb{P}(X_{0:t+1} = s_{0:t+1}, Y_{0:t} = a_{0:t})$$

Further, we can expand each term of this summation as

$$\begin{aligned} & \mathbb{P}(X_{0:t+1} = s_{0:t+1}, Y_{0:t} = a_{0:t}) \\ = & \frac{\mathbb{P}(X_{0:t+1} = s_{0:t+1}, Y_{0:t} = a_{0:t})}{\mathbb{P}(X_{0:t} = s_{0:t}, Y_{0:t} = a_{0:t})} \cdot \frac{\mathbb{P}(X_{0:t} = s_{0:t}, Y_{0:t} = a_{0:t})}{\mathbb{P}(X_{0:t} = s_{0:t}, Y_{0:t-1} = a_{0:t-1})} \cdot \mathbb{P}(X_{0:t} = s_{0:t}, Y_{0:t-1} = a_{0:t-1}) \\ = & \mathbb{P}(X_{t+1} = s_{t+1} | X_{0:t} = s_{0:t}, Y_{0:t} = a_{0:t}) \mathbb{P}(Y_t = a_t | X_{0:t} = s_{0:t}, Y_{0:t-1} = a_{0:t-1}) \mathbb{P}(X_{0:t} = s_{0:t}, Y_{0:t-1} = a_{0:t-1}) \\ = & \end{aligned}$$

Substituting this into the previous expression we get

$$\mathbb{P}(X_{0:t+1} = s_{0:t+1}) = \sum_{a_{0:t} \in A^{t+1}} p_t(s_{t+1}|s_t, a_t) q_t(a_t|s_t) \mathbb{P}(X_{0:t} = s_{0:t}, Y_{0:t-1} = a_{0:t-1})$$

We can make this expression recursive, by splitting the summation into two stages: one over $X_{0:t-1}$; the other over X_t.

$$\begin{aligned}
& \mathbb{P}(X_{0:t+1} = s_{0:t+1}) \\
= & \sum_{a_{0:t} \in A^{t+1}} p_t(s_{t+1}|s_t, a_t) q_t(a_t|s_t) \mathbb{P}(X_{0:t} = s_{0:t}, Y_{0:t-1} = a_{0:t-1}) \\
= & \sum_{a_t \in A} \sum_{a_{0:t-1} \in A^t} p_t(s_{t+1}|s_t, a_t) q_t(a_t|s_t) \mathbb{P}(X_{0:t} = s_{0:t}, Y_{0:t-1} = a_{0:t-1}) \\
= & \left(\sum_{a_t \in A} p_t(s_{t+1}|s_t, a_t) q_t(a_t|s_t) \right) \left(\sum_{a_{0:t-1} \in A^t} \mathbb{P}(X_{0:t} = s_{0:t}, Y_{0:t-1} = a_{0:t-1}) \right) \\
= & \left(\sum_{a_t \in A} p_t(s_{t+1}|s_t, a_t) q_t(a_t|s_t) \right) \mathbb{P}(X_{0:t} = s_{0:t}) \text{ by the recursive definition}
\end{aligned}$$

This expression can be substituted into the expression of the conditional probability

$$\begin{aligned}
\mathbb{P}(X_{t+1} = s_{t+1}|X_{0:t} = s_{0:t}) &= \frac{\mathbb{P}(X_{0:t+1} = s_{0:t+1})}{\mathbb{P}(X_{0:t} = s_{0:t})} \\
&= \frac{\left(\sum_{a_t \in A} p_t(s_{t+1}|s_t, a_t) q_t(a_t|s_t)\right) \mathbb{P}(X_{0:t} = s_{0:t})}{\mathbb{P}(X_{0:t} = s_{0:t})} \\
&= \sum_{a_t \in A} p_t(s_{t+1}|s_t, a_t) q_t(a_t|s_t)
\end{aligned}$$

This final expression is independent of $s_{0:t-1}$, thus $\{X_t\}_{t \in T}$ is a markov chain. Further, it has transition kernel

$$\mathbb{P}(X_{t+1} = s'|X_t = s) = \sum_{a \in A} p_t(s'|s, a) q_t(a|s)$$

Remark 2.2 -
In *Markov Decision Processes* expressions of the following form

$$\mathbb{P}(X_0 = s_0, \dots, X_t = X_t, Y_0 = a_0, \dots, Y_t = a_t) \quad \mathbb{E}[f(X_0, \dots, X_t, Y_0, \dots, Y_t)]$$

depend on the applied policy π. To emphasis this the following notation is used $\mathbb{P}^\pi(\cdot)$ and $\mathbb{E}^\pi[\cdot]$.

$$\mathbb{P}^\pi(X_0 = s_0, \dots, X_t = X_t, Y_0 = a_0, \dots, Y_t = a_t) \quad \mathbb{E}^\pi[f(X_0, \dots, X_t, Y_0, \dots, Y_t)]$$

2.2 Markov Decision Processes

Definition 2.6 - *Markov Decision Process (MDP)*
A *Markov Decision Process* (MDP) is a *Sequential Decision Problem* where the underlying stochastic system evolves in a Markovian Fashion. There are two main components to a *Markov Decision Process*:

i). A stochastic system.

ii). An agent.

Proposition 2.5 - *Process of an MDP*
A *Markov Decision Process* follows the following steps

i). At a specified point in time, an agent observes the state of a system.

ii). Based on the observed state, the agent selects an action.

iii). The selected action produces two results

- An immediate reward for the agent.
- A new evolved state for the system.

iv). The process repeats with the updated system.

2.2.1 Problem Formulation

Definition 2.7 - *Markov Decision Problem*
In a *Markov Decision Problem* the agent is tasked with maximising the total reward recieved over a given time horizon T.

$$\texttt{TotalReward} := \sum_{t \in T} r_t(X_t, Y_t)$$

A *Markov Decision Problem* has the following elements

i). *Decision Epochs* T - The points in time where the agent selects and makes their action.

ii). *System States* S - A numerical representation of the condition of the stochastic system.

iii). *Actions* A - The product of the agent's decision. Actions affect the behaviour of the system.

iv). *Immediate Rewards* $r_t(s, a)$ - What the agents receives for their action. This is a measure of the quality of the action taken (More reward=Better Action).

v). *Transition Probabilities* $p_t(\cdot|s, a)$ - The probability of the system being in a particular state at the next time step $t + 1$, given it's current state s and the action a the agent takes. These have *Markovian Properties*.

Definition 2.8 - *Time-Horizon T*
The *Time-Horizon* is the set T of all *Decision Epochs*. There are two types of *Time-Horizon*

i). *Continuous-Time* where the time-horizon is specified over an interval $T := [t_1, t_2]$

ii). *Discrete-Time* where the time-horizon is specified for specific points in time $T := \{t_0, \dots, t_N\}$ with $t_0 < \cdots < t_N$. As the intervals between epochs is irrelevant T is often defined as a subset of natural numbers $T := \{0, \dots, N\}$. There are two further categories of *Discrete Time-Horizons*

- *Finite Time* - $T := \{0, , \dots, N\}$.
- *Infinite Time* - $T := \{0, 1, \dots\}$.

N.B. - Only discrete-time time-horizons are within the scope of this course.

Definition 2.9 - *State-Space S*
The *State-Space* is the set S of all states the stochastic system can take. There are three types of *State Space*

i). *Continuous-State* - The state space S is uncountable.

ii). *Discrete-State* - The state space S is countably infinite.

iii). *Finite-State* - The state space S is countably finite $S := \{s_1, \dots, s_N\}$ with $N \geq 2$.

N.B. - Only finite-state state-spaces are within the scope of this course.

Definition 2.10 - *Action-Space A*
The *Action-Space* is the set A of all actions available to the agent. There are three types of *Action Space*

i). *Continuous-Action* - The action space A is uncountable.

ii). *Discrete-Action* - The action space A is countably infinite.

iii). *Finite-Action* - The action space A is countably finite $A := \{a_1, \dots, a_N\}$ with $N \geq 2$.

It is possible that not all actions in A are available for all states in S. We let $A(s)$ denote the set of available actions in state s.

N.B. - Only finite-action action-spaces are within the scope of this course.

Definition 2.11 - *Transition Probabilities* $p_t(\cdot|s, a)$
Transition Probabilities $p_t(\cdot|s, a)$ are a family of parametric probability mass functions on state-space S. Each transition probability is parameterised by the epoch t and are conditional on the system state s and action taken a.

The *Transition Probabilities* have *Markovian Properties*

$$\begin{aligned}\mathbb{P}(X_{t+1} = s_{t+1}|X_0 = s_0, Y_0 = a_0, \dots, X_t = s_t, Y_t = a_t) &= \mathbb{P}(X_{t+1} = s_{t+1}|X_t = s_t, Y_t = a_t) \\ &=: p_t(s_{t+1}|s_t, a_t)\end{aligned}$$

i.e. The probability of transition a given state s_{t+1} only depends on the current state s_t and action a_t (and nothing earlier).

$$\begin{aligned}\{X_{t+1}|X_0, Y_0, \dots, X_t, Y_t\} &\sim \{X_{t+1}|X_t, Y_t\} \\ &\sim p_t(\cdot|X_t, Y_t)\end{aligned}$$

Definition 2.12 - *Decision Rules*
A *Decision Rule* is a procedure for selecting an action at the specified decision epoch. In the process of selecting an action, the decision rule takes into account the current system state, past system state and past agent actions.

There are four types of *Decision Rule*

i). *History Dependent Randomised (HR)* - The decision rule $q_t(\cdot)$ is a conditional probability mass function on the action-space A which represent the probability of taking a given action given all currently available information.

$$\begin{aligned}
\mathbb{P}(Y_0 = a_0 | X_0 = s_0) &= q_0(a_0|s_0) \\
\mathbb{P}(Y_t = a_t | X_0 = s_0, Y_0 = a_0, \dots, X_{t-1} = s_{t-1}, Y_{t-1} = a_{t-1}, X_t = s_t) &= q_t(a_t|s_0, \dots, s_t, a_0, \dots, a_{t-1}) \\
\implies \{Y_0|X_0\} &\sim q_0(\cdot|X_0\} \\
\{Y_t|X_0, \dots, X_t, Y_0, \dots, Y_{t-1}\} &\sim q_t(\cdot|X_0, \dots, X_t, Y_0, \dots, Y_{t-1})
\end{aligned}$$

ii). *History Dependent Deterministic (HD)* - The decision rule $d_t(\cdot)$ is a deterministic function of all currently available information

$$Y_t := d_t(X_0, \dots, X_t, Y_0, \dots, Y_{t-1})$$

iii). *Markovian Randomised (MR)* - The decision rule $q_t(\cdot)$ is a conditional probability mass function on the action-space A which is conditional only on the current state (ie is independent of past states and actions)

$$\begin{aligned}
\mathbb{P}(Y_t = a_t | X_0 = s_0, Y_0 = a_0, \dots, X_{t-1} = s_{t-1}, Y_{t-1} = a_{t-1}, X_t = s_t) &= q_t(a_t|s_t) \\
\implies \{Y_t|X_0, \dots, X_t, Y_0, \dots, Y_{t-1}\} &\sim q_t(\cdot|X_t)
\end{aligned}$$

iv). *Markovian Deterministic (MD)* - The decision rule $d(t)$ is a deterministic function of the current state

$$Y_t = d_t(X_t)$$

Remark 2.3 - *M*
arkovian Decision rules are memoryless.

Definition 2.13 - *Decision Policy* π
A *Decision Policy* π is a procedure which specifies a decision rule for any decision epoch. A *Decision Policy* is a sequence of decision rules

$$\pi = \{q_t(\cdot|\cdot)\}_{t \in T}$$

where $q_t(\cdot|cdot)$ is the decision probability at epoch t.

There are two classes of decision policies

i). *Stationary* - The decision rule is the same for all decision epochs (ie $q_t(\cdot|\cdot)$ is independent of t).

ii). *Non-Stationary* - The decision rule varies between epochs (ie $q_t(\cdot|\cdot)$ is dependent on t).

Remark 2.4 - *Quantifying Total Reward*
As `TotalReward` is a random quantity, it cannot be maximised directly. Thus we target finding the policy π which gives the maximum expected reward, given the rewards $\{r_t(\cdot)\}_{t \in T}$ and transition probabilities $\{p_t(\cdot|\cdot)\}_{t \in T}$

$$\mathbb{E}^\pi[\texttt{TotalReward}] := \mathbb{E}^\pi \left(\sum_{t \in T} r_t(X_t, Y_t) \right)$$

2.2.2 Finite Horizon Problems

Definition 2.14 - *Finite Horizon Markov Decision Problems*
In a *Finite Horizon MDP* the agent makes decisions at a finite number of epochs $N < \infty$.

Proposition 2.6 - *Elements of Finite Horizon MDPs*
A *Finite Horizon MDP* over N epochs has the following steps

- Time Horizon $T = \{0, \dots, N-1\}$.
- Transition Probabilities $p_0(s'|s,a), \dots, p_{N-1}(s'|s,a)$.
- Rewards $r_0(s,a), \dots, r_{N-1}(s,a), r_N(s)$.

Proposition 2.7 - *Objective in a Finite Horizon MDP*
For a *Finite Horizon MDP* with transition probabilities $p_0(s'|s,a), \dots, p_{N-1}(s'|s,a)$ and the rewards $r_0(s,a), \dots, r_{N-1}(s,a), r_N(s)$, our objective is to find a *History Dependent Randomised* policy over time-horizon T $\pi \in HR(T)$ which achieves the maximum expected reward

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

Definition 2.15 - *Terminal Reward* $r_X(s)$
The *Terminal Reward* is recieved for the final state the agent ends up in (ie the state they are in after N actions). As no action is taken at epoch $t = N$, $r_N(\cdot)$ depends only on the state s.

Proposition 2.8 - *Stochastic Systems for a Finite Horizon MDP*
As the sequence of states $\{X_t\}_{t \in T}$ is a markov chain we can deduce the probability of transitioning to a given state, given all available information.

$$\forall\ t \in T \quad (X_{t+1}|X_{0:t}, Y_{0:t}) \sim^{[1]} (X_{t+1|X_t,Y_t}) \sim p_t(X_{t+1}|X_t, Y_t)$$

[1] by Markov property.

Two-Stage MDP

Definition 2.16 - *Two-Stage MDP*
A *Two-Stage MDP* is a *Finite Horizon MDP* over only $N = 2$ epochs. It has the following elements

- Time Horizon $T = \{0, 1\}$.
- Transition Probabilities $p_0(s'|s,a), p_1(s'|s,a)$.
- Rewards $r_0(s,a), r_1(s,a), r_2(s)$.

Proposition 2.9 - *Stochastic System for a Two-Stage MDP*

$$(X_2|X_{0:1}, Y_{0:1}) \sim (X_2|X_1, Y_1) \sim p_1(\cdot|X_1, Y_1)$$
$$(X_1|X_0, Y_0) \sim p_0(\cdot|X_0, Y_0)$$

Proposition 2.10 - *Objective in a Two-Stage MDP*
For a *Two-Stage MDP* with transition probabilities $p_0(s'|s,a), p_1(s'|s,a)$ and the rewards $r_0(s,a), r_1(s,a), r_2(s)$, our objective is to find a *History Dependent Randomised* policy over time-horizon T $\pi \in HR(T)$ which achieves the maximum expected reward

$$\begin{aligned} R^\pi &:= \mathbb{E}^\pi \left[r_0(X_0, Y_0) + r_1(X_1, Y_1) + u_2^*(X_2) \right] \\ &= \mathbb{E}^\pi \left[r_0(X_0, Y_0) + r_1(X_1, Y_1) \right] + \mathbb{E}^\pi \left[u_2^*(X_2) \right] \end{aligned}$$

where $u_2^*(X_2) := r_2(s)$.

Proposition 2.11 - *Optimising a Two-Stage MDP - Epoch $t = 1$*
For a *Two-Stage MDP* we want to maximise the expected total reward R^π

$$R^\pi := \mathbb{E}^\pi \left[r_0(X_0, Y_0) + r_1(X_1, Y_1) \right] + \mathbb{E}^\pi \left[u_2^*(X_2) \right]$$

Note that by the *Tower Property*

$$\mathbb{E}^\pi [u_2^*(X_2)] = \mathbb{E}^\pi \left[\mathbb{E}^\pi (u_2^*(X_2) | X_1, Y_1) \right]$$

Thus,

$$\begin{aligned} R^\pi &= \mathbb{E}^\pi \left[r_0(X_0, Y_0) + r_1(X_1, Y_1) \right] + \mathbb{E}^\pi \left[\mathbb{E}^\pi (u_2^*(X_2) | X_1, Y_1) \right] \\ &= \mathbb{E}^\pi \left[r_0(X_0, Y_0) + r_1(X_1, Y_1) + \mathbb{E}^\pi (u_2^*(X_2) | X_1, Y_1) \right] \end{aligned}$$

By the definition of conditional expectation and transition probabilities we have

$$\begin{aligned} \mathbb{E}^\pi (u_2^*(X_2) | X_1, Y_1) &= \sum_{s' \in S} u_2^*(s') \mathbb{P}^\pi (X_2 = s' | X_1, Y_1) \\ &= \sum_{s' \in S} u_2^*(s') p_1(s' | X_1, Y_1) \end{aligned}$$

This can be substituted back into the expression of expected total reward to get an expression which is sufficient to choose an optimal action at epoch $t = 1$

$$R^\pi = \mathbb{E}^\pi \left[r_0(X_0, Y_0) + r_1(X_1, Y_1) + \sum_{s' \in S} u_2^*(s') p_1(s' | X_1, Y_1) \right]$$

At epoch $t = 1$ X_0, Y_0, X_1 are all known. So, the expression $r_0(X_0, Y_0) + r_1(X_1, Y_1) + \sum_{s' \in S} u_2^*(s') p_1(s' | X_1, Y_1)$ can be interpreted as a deterministic function of Y_1 (the action played at epoch $t = 1$)

$$\begin{aligned} Y_1^* &:= \text{argmax}_{a \in A(s)} \left(r_0(X_0, Y_0) + r_1(X_1, a) + \sum_{s' \in S} u_2^*(s') p_1(s' | X_1, a) \right) \\ &= \text{argmax}_{a \in A(s)} \left(r_1(X_1, a) + \sum_{s' \in S} u_2^*(s') p_1(s' | X_1, a) \right) \text{[1]} \end{aligned}$$

I now define the following notation

$$\begin{aligned} u_1^*(s) &:= \max_{a \in A(s)} \left(r_1(s, a) + \sum_{s' \in S} u_2^*(s') p_1(s' | s, a) \right) \\ d_1^*(s) &:= \text{argmax}_{a \in A(s)} \left(r_1(s, a) + \sum_{s' \in S} u_2^*(s') p_1(s' | s, a) \right) \end{aligned}$$

[1] As terms independent of a do not affect the maximisation.

$u_1^*(s)$ is the maximum expected reward which can be achieved after being in state s at epoch $t = 1$, $d_1^*(s)$ is optimal decision function (ie optimal action) given the agent starts epoch $t = 0$ in state s (There may be multiple such actions). Thus

$$Y_1^* = d_1^*(X_1)$$

As $u_1^*(s)$ is the optimal value we can use it to bound the expected reward expression

$$\forall\ s \in S,\ \forall\ a \in A(s) \quad r_1(s,a) + \sum_{s' \in S} u_2^*(s') p_1(s'|s,a) \leq u_1^*(s)$$

This inequality becomes an equality when $a = d_1^*(s)$.
Consider where $s = X_1$ and $a = Y_1$, we can conclude

$$r_1(X_1, Y_1) + \sum_{s' \in S} u_2^*(s') p_1(s'|X_1, Y_1) \leq u_1^*(X_1)$$

This inequality becomes an equality when $Y_1 = d_1^*(X_1)$.

To consider the first action Y_0 we add $r_0(X_0, Y_0)$ to both sides and take expectations, giving

$$\mathbb{E}^\pi \left(r_0(X_0, Y_0) + r_1(X_1, Y_1) + \sum_{s' \in S} u_2^*(s') p_1(s'|X_1, Y_1) \right) \leq \mathbb{E}^\pi (r_0(X_0, Y_0) + u_1^*(X_1))$$

Again, this inequality becomes an equality when $Y_1 = d_1^*(X_1)$.

As the LHS is just the definition for total expected reward R^π we can restate the bound as a bound on R^π, nothing the inequality is maximised for $Y_1 = d_1^*(X_1)$

$$R^\pi \leq \mathbb{E}^\pi (r_0(X_0, Y_0) + u_1^*(X_1))$$

Proposition 2.12 - *Optimising a Two-Stage MDP - Epoch $t = 0$*
Let $\tilde{R}^\pi := \mathbb{E}^\pi (r_0(X_0, Y_0) + u_1^*(X_1))$ denote the bound derived in `Proposition 2.11`. From `Proposition 2.11` we have that $R^\pi \leq \tilde{R}^\pi$ and that this is an equality if $Y_1 = d_1^*(X_1)$.

By the *Tower Property*, we have

$$\mathbb{E}^\pi (U_1^*(X_1)) = \mathbb{E}^\pi \left[\mathbb{E}^\pi (u_1^*(X_1) | X_0, Y_0) \right]$$

Using this we can restate $\tilde{R}^\pi$ as

$$\begin{aligned} \tilde{R}^\pi &= \mathbb{E}^\pi (r_0(X_0, Y_0)) + \mathbb{E}^\pi (u_1^*(X_1)) \\ &= \mathbb{E}^\pi (r_0(X_0, Y_0)) + \mathbb{E}^\pi \left[\mathbb{E}^\pi (u_1^*(X_1) | X_0, Y_0) \right] \end{aligned}$$

By the definition of conditional expectations and transition probabilities we have

$$\begin{aligned} \mathbb{E}^\pi (u_1^*(X_1) | X_0, Y_0) &= \sum_{s' \in S} u_1^*(s') \mathbb{P}^\pi (X_1 = s' | X_0, Y_0) \\ &= \sum_{s' \in S} u_1^*(s') p_0(s' | X_0, Y_0) \\ \implies \qquad \tilde{R}^\pi &= \mathbb{E}^\pi \left[r_0(X_0, Y_0) + \sum_{s' \in S} u_1^*(s') p_0(s' | X_0, Y_0) \right] \end{aligned}$$

At epoch $t = 0$ the value of X_0 is known, thus the following is a deterministic function of Y_0

$$r_0(X_0, Y_0) + \sum_{s' \in S} u_1^*(s') p_0(s' | X_0, Y_0)$$

This means the optimal action Y_0^* at epoch $t = 0$ is the action which maximises this function

$$Y_0^* := \text{argmax}_{a \in A(s)} \left(r_0(X_0, a) + \sum_{s' \in S} u_1^*(s') p_0(s'|X_0, a) \right)$$

Define $u_0^*(s)$ to be the maximum reward after epoch $t = 0$ given the agent starts the epoch in state s and d_0^* be the optimal decision function (ie optimal action) at epoch $t = 0$ given the agent starts the epoch in state s.

$$\begin{aligned} u_0^*(s) &:= \max_{a \in A(s)} \left(r_0(s, a) + \sum_{s' \in S} u_1^*(s') p_0(s'|s, a) \right) \\ d_0^*(s) &:= \text{argmax}_{a \in A(s)} \left(r_0(s, a) + \sum_{s' \in S} u_1^*(s') p_0(s'|s, a) \right) \end{aligned}$$

Using this $Y_0^* = d_0^*(X_0)$ and we can bound the expected reward

$$\forall\ s \in S, a \in A(s) \quad r_0(s, a) + \sum_{s' \in S} u_1^*(s') p_0(s'|s, a) \leq u_0^*(s)$$

Note that is becomes an equality if $a = d_0^*(s)$.

Setting $s = X_0, a = Y_0$ we can conclude

$$r_0(X_0, Y_0) + \sum_{s' \in S} u_1^*(s') p_0(s'|X_0, Y_0) \leq u_0^*(X_0)$$

This bound is an equality if $a = d_0^*(X_0)$. We can therefore bound the expected reward as

$$\begin{aligned} & \mathbb{E}^\pi \left[r_0(X_0, Y_0) + \sum_{s' \in S} u_1^*(s') p_0(s'|X_0, Y_0) \right] \leq \mathbb{E}^\pi \left[u_0^*(X_0) \right] \\ \implies & \qquad \tilde{R}^\pi \leq \mathbb{E}^\pi \left[u_0^*(X_0) \right] \end{aligned}$$

which is optimised when $Y - 0 = d_0^*(X_0)$.

As the LHS is the definition of $\tilde{R}^\pi$ we can bound the total expected reward as

$$R^\pi \leq \tilde{R}^\pi \leq \mathbb{E}^\pi [u_0^*(X_0)]$$

which is optimised if $Y_0 = d_0^*(X_0)$ and $Y_1 = d_1^*(X_1)$.

Remark 2.5 - *Conclusions*

- The expected total reward R^π cannot exceed $\mathbb{E}^\pi[u_0^*(X_0)]$, but it can equal it.
- The optimal expected total reward is achieved when $Y_0 = d_0^*(X_0), Y_1 = d_1^*(X_1)$. Thus the optimal policy π^* is

  $$\pi^* := (d_0^*(s), d_1^*(s))$$

  Note that the construction of π^* is from functions $u_0^*(s), u_1^*(s), d_0^*(s), d_1^*(s)$ which are defined for $t \in \{0, 1\}$ as

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

  where $u_N^*(s) := r_N(s)$

Remark 2.6 - *Optimal Policy for any Finite-Horizon MDP*
We can generalise the optimal policy for a *Two-Stage MDP*, given in `Remark 2.5`, to get an optimal decision policy for any *Finite-Horizon MDP*.

For a *Finite-Horizon MDP* over N epochs the optimal decision policy is

$$\pi^* := (d_0^*(s), \dots, d_{N-1}^*(s))$$

where for $t \in \{0, \dots, N-1\}$

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

General Finite-Horizon MDPs

Definition 2.17 - *Optimality Equation $u_K^*(s)$*
The *Optimality Equation* for a General Finite-Horizon MDP over N epochs is defined by the following backward recursion

$$u_k^*(s) := \max_{a \in A(s)} \left(r_k(s,a) + \sum_{s' \in S} u_{k+1]^*(s') p_k(s'|s,a)} \right) \text{ for } k \in [0, N-1], s \in S, u_N^*(\cdot) := r_N(\cdot)$$

AKA *Dynamic Programming Algorithm*

Definition 2.18 - *Optimal Decision Rules $d_k^*(\cdot)$*
The *Optimality Decision Rules* for a General Finite-Horizon MDP over N epochs is defined by the following backward recursion

$$d_k^*(s) := \text{argmax}_{a \in A(s)} \left(r_k(s,a) + \sum_{s' \in S} u_{k+1}^*(s') p_k(s'|s,a) \right) \text{ for } k \in [0, N-1], s \in S$$

Definition 2.19 - *Optimal Policy π^**
The *Optimal Policy* for a General Finite-Horizon MDP over N epochs is defined as

$$\pi^* := \{d_t^*(s)\}_{t \in T}$$

Theorem 2.4 - *Bound on Expected Total Reward*
For any *History Dependent Randomised Policy* $\pi \in HR(T)$ over time-horizon T the following inequality holds for the expected Total Reward

$$\underbrace{\mathbb{E}^\pi \left(\sum_{t=0}^{N-1} r_t(X_t, Y_t) + r_N(X_N) \right)}_{\text{Expected Total Reward}} \leq \mathbb{E}^\pi(u_0^*(X_0)) = \sum_{s \in S} u_0^*(s) \mathbb{P}(X_0 = s)$$

This inequality becomes an equality if the policy is optimal $\pi = \pi^*$.

Proof 2.3 - *Theorem 2.4*
For notational ease I define the following

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

Note that R_0^π equivalent to the expected total reward. To proved `Theorem 2.4` it is sufficient to show that $\forall\ k \in [0, N],\ \pi \in HR(T)$

$$\begin{array}{lrcl} (1) & R_k^\pi & \leq & \mathbb{E}^\pi[u_k^*(X_k)] \\ (2) & R_k^{\pi^*} & = & \mathbb{E}^\pi[u_k^*(X_k)] \end{array}$$

I shall show (1) and (2) by backwards induction on k.

Base Case - $k = N$.
Note that by the definition of $u_N^*(s)$ and R_N^π, we have

$$u_N^*(s) = r_N(s) \quad \text{and} \quad R_N^\pi =^\pi [r_N(X_N)]$$

Hence (1) and (2) are true for $k = N$.

Inductive Hypothesis - Assume that (1) and (2) how for some $k \in [1, N]$.

Inductive Case - To prove the theorem, it is sufficient to show that (1) and (2) are true for $k-1$ given they are true for k. (ie given the IH holds).
Consider the following derivation

$$\begin{aligned} R_{k-1}^\pi &= \mathbb{E}^\pi \left[\sum_{t=k-1}^{N-1} r_t(X_t, Y_t) + r_N(X_N) \right] \text{ by def.} \\ &= \mathbb{E}^\pi \left[r_{k-1}(X_{k-1}, Y_{k-1}) \right] + \mathbb{E}^\pi \left[\sum_{t=k}^{N} r_t(X_t, Y_t) + r_N(X_N) \right] \\ &= \mathbb{E}^\pi (r_{k-1}(X_{k-1}, Y_{k-1})) + R_k^\pi \text{ by def.} \\ &\leq \mathbb{E}^\pi [r_{k-1}(X_{k-1}, Y_{k-1})] + \mathbb{E}^\pi (u_k^*(X_k)) \text{ by IH} \end{aligned}$$

By the *Tower Property*, we have

$$\mathbb{E}^\pi (u_k^*(X_k)) = \mathbb{E}^\pi \left[\mathbb{E}^\pi (u_k^*(X_k | X_{k-1}, Y_{k-1})) \right]$$

Substituting this expression into the inequality above we find

$$\begin{aligned} R_{k-1}^\pi &\leq \mathbb{E}^\pi (r_{k-1}(X_{k-1}, Y_{k-1})) + \mathbb{E}^\pi (u_k^*(X_k)) \\ &= \mathbb{E}^\pi [r_{k-1}(X_{k-1}, Y_{k-1})] + \mathbb{E}^\pi \left[\mathbb{E}^\pi (u_k^*(X_k | X_{k-1}, Y_{k-1})) \right] \\ &= \mathbb{E}^\pi \left[r_{k-1}(X_{k-1}, Y_{k-1}) + \mathbb{E}^\pi (u_k^*(X_k | X_{k-1}, Y_{k-1})) \right] \end{aligned}$$

By the definition of conditional expectations and transition probabilities, we have

$$\begin{aligned} \mathbb{E}^\pi [u_k^*(X_k) | X_{k-1}, Y_{k-1}] &= \sum_{s' \in S} u_k^*(s') \mathbb{P}^\pi (X_k = s' | X_{k-1}, Y_{k=1}) \\ &= \sum_{s' \in S} u_k^*(s') p_{k-1}(s' | X_{k-1}, Y_{k-1}) \end{aligned}$$

Substituting this expression into the inequality above we find

$$\begin{aligned} R^{\pi}_{k-1} &\leq \mathbb{E}^{\pi}\big[r_{k-1}(X_{k-1}, Y_{k-1}) + \mathbb{E}^{\pi}(u^*_k(X_k|X_{k-1}, Y_{k-1}))\big] \\ &= \mathbb{E}^{\pi}\left[r_{k-1}(X_{k-1}, Y_{k-1}) + \sum_{s'\in S} u^*_k(s')p_{k-1}(s'|X_{k-1}, Y_{k-1})\right] \quad (3) \end{aligned}$$

Consider the *Optimality Equations* for epoch $t = k-1$

$$\begin{aligned} u^*_{k-1}(s) &= \max_{a\in A(s)}\left(r_{k-1}(s,a) + \sum_{s'\in S} u^*_k(s')p_{k-1}(s'|s,a)\right) \\ d^*_{k-1}(s) &= \text{argmax}_{a\in A(s)}\left(r_{k-1}(s,a) + \sum_{s'\in S} u^*_k(s')p_{k-1}(s'|s,a)\right) \end{aligned}$$

As u^*_{k-1} is the max then we have the following upper bound $\forall\ as \in S, a \in A(s)$

$$r_{k-1}(s,a) + \sum_{s'\in S} u^*_k(s')p_{k-1}(s'|s,a \leq u^*_{k-1}(s) \quad (4)$$

this inequality is an equality when $\bar{d^*_{k-1}}(s)$.
By setting $s = X_{k-1}$ and $a = Y_{k-1}$ in (4) we have that

$$r_{k-1}(X_{k-1}, Y_{k-1}) + \sum_{s'\in S} u^*_k(s')p_{k-1}(s'|X_{k-1}, Y_{k-1} \leq u^*_{k-1}(x_{k-1})$$

Taking expectations of both sides gives us

$$\mathbb{E}^{\pi}\left[r_{k-1}(X_{k-1}, Y_{k-1}) + \sum_{s'\in S} u^*_k(s')p_{k-1}(s'|X_{k-1}, Y_{k-1}\right] \leq \mathbb{E}^{\pi}\left[u^*_{k-1}(x_{k-1})\right] \quad (5)$$

Under policy π^* $Y_{k-1} = d^*_{k-1}(X_{k-1})$, meaning this inequality is an equality when $\pi = \pi^*$.
Combining (3) and (5) we get, the following inequality (which is an equality when $\pi = \pi^*$)

$$\begin{aligned} R^{\pi}_{k-1} &\leq \mathbb{E}^{\pi}\left[r_{k-1}(X_{k-1}, Y_{k-1}) + \sum_{s'\in S} u^*_k(s')p_{k-1}(s'|X_{k-1}, Y_{k-1})\right] \\ &\leq \mathbb{E}^{\pi}(u^*_{k-1}(X_{k-1})) \end{aligned}$$

Hence, given (1) and (2) hold for k then they hold for $k-1$. Thus, by the principle of mathematical induction, (1) and (2) hold $\forall\ k \in [0, N-1]$.

Remark 2.7 - *The Initial State X_0 is not affect by any policy π*
This can be emphasized by dropping policy notation

$$\begin{aligned} &\mathbb{P}(X_0 = s) \text{ instead of } \mathbb{P}^{\pi}(X_0 = s) \\ &\mathbb{E}(u^*_0(X_0)) \text{ instead of } \mathbb{E}^{\pi}(u^*_0(X_0)) \end{aligned}$$

Definition 2.20 - *Value Function $v^{\pi}(\cdot)$*
The *Value Function* associated with a *History Dependent Randomised Policy* $\pi \in HR(T)$ is the expected total reward, given the initial state $s \in S$ (ie $X_0 = s$)

$$v^{\pi}(s) := \mathbb{E}^{\pi}\left(\sum_{t=0}^{N-1} r_t(X_t, Y_t) + r_N(X_N)\big|X_0 = s\right)$$

Definition 2.21 - *Optimal Value Function $v^*(\cdot)$*
The *Optimal Value Function* is the maximum expected total reward, given the initial state $s \in S$ (ie $X_0 = s$)

$$\begin{aligned} v^*(s) &:= \max_{\pi \in HR(T)} v^\pi(s) \\ &= \max_{\pi \in HR(T)} \mathbb{E}^\pi \left(\sum_{t=0}^{N-1} r_t(X_t, Y_t) + r_N(X_N) \big| X_0 = s \right) \end{aligned}$$

Note that this considers all possible *History Dependent Randomised Policies*.

Theorem 2.5 - *Restatements of Optimal Value Function*
Here are two equivalent expressions of the *Optimal Value Function* $v^*(\cdot)$

i). $v^*(s) = u_0^*(s) \ \forall \ s \in S$

ii). $v^*(s) = v^{\pi^*}(s) \ \forall \ s \in S$

Proof 2.4 - *Theorem 2.5 i)*
Proof follows immediately from `Theorem 2.4` as it makes no assumptions about the distribution of X_0.

Proof 2.5 - *Theorem 2.5 ii)*
By `Theorem 2.4` we have that

$$\begin{aligned} \mathbb{E}^\pi \left[\sum_{t=0}^{N-1} r_t(X_t, Y_t) + r_N(X_N) \right] &\leq \mathbb{E}[u_0^*(X_0)] \\ &= \sum_{s' \in S} u_0^*(s') \mathbb{P}(X_0 = s') \end{aligned}$$

By the definition of the *Value Functions*, we have

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

By combining `Theorem 2.4` and this result we get that

$$\begin{aligned} \sum_{s' \in S} u_0^*(s') \mathbb{P}(X_0 = x') &\geq \mathbb{E}^\pi \left[\sum_{t=0}^{N-1} r_t(X_t, Y_t) + r_N(X_N) \right] \text{ by } \texttt{Theorem 2.4} \\ &= \mathbb{E}^\pi \left(\mathbb{E}^\pi \left(\sum_{t=0}^{N-1} r_t(X_t, Y_t) + r_N(X_N) \big| X_0 \right) \right) \text{ by Tower Property} \\ &= \mathbb{E}^\pi \left[v^\pi(X_0) \right] \\ &= \sum_{s' \in S} v^\pi(s') \mathbb{P}(X_0 = s') \end{aligned}$$

Note that this inequality is an equality when $\pi = \pi^*$ and is true regardless of the distribution of X_0. Due to the independence from the distribution of X_0 we can assume, without loss of correctness, that X_0 follows *Dirac Distribution* centred at some chosen state $s \in S$. Hence

$$\mathbb{P}(X_0 = s') = \mathbb{1}\{s' = s\}$$

Introducing this definition of the distribution of X_0 into the previous inequality we have

$$\begin{aligned} v^\pi(s) &= \sum_{s' \in S} v^\pi(s') \overbrace{\mathbb{1}\{s' = s\}}^{\mathbb{P}^\pi(X_0 = s')} \\ &\leq \sum_{s' \in S} u_0^*(s') \mathbb{1}\{s' = s\} \\ &= u_0^*(s) \end{aligned}$$

Thus we have that $\forall\ s \in S$

$$v^\pi(s) \leq u_0^*(s) \quad \text{and} \quad v^{\pi^*}(s) = u_0^{\pi^*}(s) \quad (1)$$

By the definition of the *Optimal value Function* $v^*(\cdot)$ and the inequality above, we have that $\forall\ s \in S$

$$v^*(s) = \max_\pi(v^\pi(s)) \leq u_0^*(s) \text{and}$$
$$v^*(s) = \max_\pi(v^\pi(s)) \geq v^{\pi^*}(s)$$

Combining these results with (1) we get

$$u_0^*(s) = v^{\pi^*}(s) \leq v^*(s) \leq u_0^*(s) \quad \forall\ s \in S$$

Hence

$$u_0^*(s) = v^{\pi^*}(s) = v^*(s) \quad \forall\ s \in S$$

Optimality Principle

Definition 2.22 - *Tail Subproblem*
The *Tail Subproblem of Length* $l \in [1, N]$ is a subproblem of a *Finite Time-Horizon MDP* over N epochs. The subproblem has the following elements

- L decision epochs.
- Time-Horizon $T_L = \{N - L, \dots, N - 1\}$.
- Transition Probabilities $p_{N-L}(s'|s, a), \dots, p_{N-1}(s'|s, a)$.
- Rewards $r_{N-L}(s, a), \dots, r_{N-1}(s, a), r_N(s)$.

All these elements are from the main problem (ie this is a subproblem). Further, this subproblem is equivalent to the last L steps of the *Finite Time-Horizon MDP* over N epochs. The practical difference between the problem and the subproblem is that the initial epoch for the subproblem is $t = N - L$, rather than $t = 0$.

The object of this subproblem is to find the *History Dependent Randomised Policy* $\pi \in HR(T_l)$ which maximises the expected total reward

$$\text{argmax}_{\pi \in HR(T_L)} \mathbb{E}^\pi \left[\left(\sum_{t=N-L}^{N-1} r_t(X_t, Y_t) \right) + r_N(X_N) \right]$$

Remark 2.8 - *Equivalence of MDP and Tail Subproblem*

If the tail subproblem is translated from epochs $N-L,\dots,N-1$ to epochs $0,\dots,L-1$ it is equivalent an MDP over L epochs (The elements of the problem need to translated accordingly).

This means the optimality equation for the tail subproblem is just a re-indexing of the optimality equations for the main problem.

Proposition 2.13 - *Optimality Equation for Tail Subproblem $u^*_{l,k}(\cdot)$*
The *Optimality Equation* for the tail subproblem of length L is defined with the following backwards recursion

$$u^*_{L,k}(s) := \max_{a\in A(s)}\left(r_k(s,a)+\sum_{s'\in S}u^*_{l,k+1}(s')p_k(s'|s,a)\right)$$

for $k\in[N-L,N-1], s\in S$ and with $u^*_{l,N}(\cdot):=r_N(\cdot)$.

Proposition 2.14 - *Optimality Decision Rules for Tail Subproblem $d^*_{l,k}(\cdot)$*
The *Optimality Decision Rules* for the tail subproblem of length L is defined with the following backwards recursion

$$d^*_{L,k}(s) := \text{argmax}_{a\in A(s)}\left(r_k(s,a)+\sum_{s'\in S}u^*_{l,k+1}(s')p_k(s'|s,a)\right)$$

for $k\in[N-L,N-1], s\in S$.

Proposition 2.15 - *Optimality Decision Policy for Tail Subproblem π^*_l*
The *Optimality Decision Policy* for the tail subproblem of length L is defined

$$\pi^*_l = \{d^*_{l,t}(s)\}_{t\in T_L}$$

Remark 2.9 - *Proeprties of Optimality Equations, Decision Rules and Policy*

- The optimality equation for the tail subproblem $u^*_{l,k}(\cdot)$ is a subrecursion of the optimality equation for the main problem $u^*_k(\cdot)$.
- The initial condition in the optimality equation for the tail subproblem $u^*_{l,N}(\cdot)=r_N(\cdot)$ is the same as the initial condition for the main problem $u^*_N(\cdot)=r_N(\cdot)$.

Given these two properties, the equations generated by the optimality equation for the tail subproblem are identical to the corresponding functions generated by the optimality equation for the main problem.

Theorem 2.6 - *Optimality Equations for Sub & Main Problem are Identical*
Consider a tail subproblem of length L, then $\forall\ s\in S,\ k\in[N-L,N-1]$

$$\begin{aligned} u^*_{l,k}(s) &= u^*_k(s) \\ d^*_{l,k}(s) &= d^*_k(s) \end{aligned}$$

Definition 2.23 - *Optimal Value Function $v^*_l(\cdot)$*

The *Optimal value Function* for the tail subproblem of length L is defined as

$$v_l*(s) := \max_{\pi \in HR(T_L)} \mathbb{E}^{\pi} \left[\sum_{t=N-L}^{N-1} r_t(X_t, Y_t) + r_N(X_N) \big| X_{N-t} = s \right] \quad \text{with } s \in S$$

The value of $v_l^*(s)$ can be interpreted as the maximum expected total reward recieved from epochs $N-L$ to N when the system is in state s at the start of epoch $t = N-L$.

Theorem 2.7 - *Optimal Value Function as Optimality Equation*
Consider a tail subproblem of length L, then $\forall\ s \in S,\ L \in [1, N]$ the following holds

$$v_L^*(s) = u_{L,N-L}^*(s)$$

Theorem 2.8 - *Optimality Principle*
Consider a tail subproblem of length L, then the *Optimality Principle States* that $\forall\ s \in S, L \in [1, N]$ the following hold

$$\begin{aligned} v_L^*(s) &= u_{N-L}^*(s) \\ \pi_L^* &= \{d_t^*(s)\}_{t \in T_L} \end{aligned}$$

Remark 2.10 - *Optimality Principle*
Here are some remarks about the *Optimality Principle*

- The *Optimality Principlei* is a straightforward consequence of `Theorems 2.6 & 2.7`. Due to i), the *Dynamic Programming Algorithm*, starting with $L = N$ and going backwards towards $L = 0$, solves the tail suboptimal problem of length L.
- Due to ii), the *Optimal Policy* for the tail subproblem of length L π_L^* can be interpreted as the truncation of optimal policy for the main problem π^* to time horizon T_L
- The *Dynamic Programming Algorithm* can be written as the following with $k \in [1, N], s \in S$ and $v_0^*(\cdot) := r_N(\cdot)$

$$\begin{aligned} v_k^*(s) &= \max_{a \in A(s)} \left(r_{N-k}(s,a) + \sum_{s' \in S} v_{k-1}^*(s') p_{N-k}(s'|s,a) \right) \\ d_k^*(s) &= \text{argmax}_{a \in A(s)} \left(r_{N-k}(s,a) + \sum_{s' \in S} v_{k-1}^*(s') p_{N-k}(s'|s,a) \right) \end{aligned}$$

2.2.3 Discounted Reward Infinite-Horizon MDPs

Problem Formulation

Definition 2.24 - *Formulation of Discounted Reward Infinite-Horizon MDP*
Here is the specification of the components of a *Discounted Reward Infinite-Horizon MDP*

- *Number of Decision Epochs* - $N = \infty$.
- *Time-Horizon* - $T = \{0, 1, \dots\}$.

- *Transition Probabilities* - $p_t(s'|s,a) = p(s'|s,a)\ \forall\ t \in T$.[2]
- *Rewards* - $r_t(s,a) = \alpha^t r(s,a)$ where $\alpha \in (0,1),\ t \in T$.[3]

Definition 2.25 - *Stochastic System of a Discounted Reward Infinite-Horizon MDP*

$$\begin{aligned} X_{t+1}|X_{0:t}, Y_{0:t} &\sim X_{t+1}|X_t, Y_t \\ &\sim p_t(\cdot|X_t, Y_t) \\ &= p(\cdot|X_t, Y_t) \end{aligned}$$

Definition 2.26 - *Objective of a Discounted Reward Infinite-Horizon MDP*
Given the transition probabilities $p(s'|s,a)$, the reward $r(s,a)$ and the discounting factor α, the agent is tasked to find a *History Dependent Randomised Policy* $\pi \in HR(T)$ st the expected total reward is maximised

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

Remark 2.11 - *Time-Importance of Rewards*
In many applications[4], the importance of immediate rewards decreases in time. This means the present and near-future rewards are the most important to maximise.

This is encoded into a *Discount Reward MDP* as the discount rate α and thus α^t characterises the important of the reward received in epoch t.

Approximating Finite-Horizon MDPs

Proposition 2.16 - *Discounted Reward Infinite-Horizon MDPs approximate Finite-Horizon MDPs*
Finite-Horizon MDPs are their optimal solutions are computationally expensive, hence approximations are desirable. This can be done using a *Discout Reward MDP*.

Consider a *Finite-Horizon MDP* as defined in `Section 2.2.2` and assume the following

i). $N \gg 1$.

ii). The parameters of the stochastic system and the parameters of the immediate rewards change slowly wrt time.

By assumption ii) we can make the following approximations

$$\begin{aligned} p_t(s'|s,a) &\approx p(s'|s,a) \\ r_t(s,a) &\approx r(s,a) \end{aligned}$$

[2] Same transition probabilities at all points in time.
[3] α is known as the *Discounting Factor*. The reward decreases with time but base value $r(s,a)$ is independent of time.
[4] Economics.

By assumption i), the terminal reward can be neglected compared to the total reward over epochs $0, \dots, N-1$

$$\begin{aligned}
\implies \qquad \left|\sum_{t=0}^{N-1} r_t(X_t, Y_t)\right| &\gg |r_N(X_N)| \\
\implies \quad \mathbb{E}^\pi\left(\sum_{t=0}^{N-1} r_t(X_t, Y_t) + r_N(X_N)\right) &\approx \mathbb{E}^\pi\left(\sum_{t=0}^{N-1} r_t(X_t, Y_t)\right) \\
&\approx \mathbb{E}^\pi\left(\sum_{t=0}^{N-1} r(X_t, Y_t)\right) \\
&\approx \lim_{N\to\infty} \mathbb{E}^\pi\left(\sum_{t=0}^{N-1} r(X_t, Y_t)\right) \\
&= \mathbb{E}^\pi\left(\sum_{t=0}^{\infty} r(X_t, Y_t)\right)
\end{aligned}$$

Remark 2.12 - *Approximation - Is total reward well-defined?*
Consider the last expression of the expected total reward in `Proposition 2.16`

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

This may not be well-defined. To overcome this, we multiply the reward value $r(X_t, Y_t)$ by the discount α^t

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

Since the state-space S and the action-space A each have a finite number of elements, there is a finite-upper bound to the reward

$$c := \max_{s\in S, a\in A(s)} |r(s,a)| < \infty$$

Consequently, we get

$$\begin{aligned}
\sum_{t=0}^{\infty} \alpha^t |r(X_t, Y_t)| &\leq \sum_{t=0}^{\infty} \alpha^t c \\
&= \frac{c}{1-\alpha} < \infty
\end{aligned}$$

Hence, the expected discounted reward 1 is well-defined and finite.

Remark 2.13 - *Approximation - How good is the approximation?*
Since $N \gg 1, \alpha \approx 1$ we have that

$$\begin{aligned}
\mathbb{E}^\pi\left(\sum_{t=0}^{N-1} r_t(X_t, Y_t) + r_N(X_N)\right) &\approx \mathbb{E}^\pi\left(\sum_{t=0}^{N-1} r(X_t, Y_t)\right) \text{ since } N \gg 1 \\
&\approx \mathbb{E}^\pi\left(\sum_{t=0}^{N-1} \alpha^t r(X_t, Y_t)\right) \text{ since } \alpha \approx 1 \\
&\approx \mathbb{E}^\pi\left(\sum_{t=0}^{\infty} \alpha^t r(X_t, Y_t)\right) \text{ since } N \gg 1
\end{aligned}$$

Given this, we can conclude that *Discount Reward MDPs* accurately approximate *Finite-Horizon MDPs* under the following assumptions

i). $N \gg 1$.

ii). $\alpha \in (0,1),\ \alpha \approx 1$.

iii). The parameters of the stochastic system and the parameters of immediate rewards change slowly in time.

Deriving Optimality Equation

Definition 2.27 - *Value Function, $v^\pi(\cdot)$*
The *Value Function* $v^\pi(s)$ associate with a policy $\pi \in HR(T)$ is the expected discounted reward given policy π is applied and the system starts is state $s \in S$ (ie $X_0 = s$)

$$v^\pi(s) := \mathbb{E}^\pi \left(\sum_{t=0}^{\infty} \alpha^t r(X_t, Y_t) \big| X_0 = s \right) \quad \text{where } s \in S$$

Definition 2.28 - *Optimal Value Function, $v^*(\cdot)$*
The *Optimal Value Function* $v^*(s)$ is the maximum possible expected discounted reward, given the system starts in state s (ie $X_0 = s$), when all policies $\pi \in HR(T)$ are considered

$$v^*(s) := \max_{\pi \in HR(T)} v^\pi(s) = \max_{\pi \in HR(T)} \mathbb{E}^\pi \left(\sum_{t=0}^{\infty} \alpha^t r(X_t, Y_t) \big| X_0 = s \right) \quad \text{where } s \in S$$

Remark 2.14 - *Notation - $v_N^*(\cdot)$*
We introduce notation v_N^* for the maximum expected discounted reward in the first N epochs

$$v_N^*(s) = \max_{\pi \in HR(T_N)} \mathbb{E}^\pi \left(\sum_{t=0}^{N-1} \alpha^t r(X_t, Y_t) \big| X_0 = s \right) \quad \text{where } s \in S$$

where $T_N := \{0, \dots, N-1\}$.
$v_N^*(s)$ is the optimal value function for the *MDP* with the following elements[5]

- *Number of Deicsion-Epochs*: N.
- *Time-Horizon*: T_N.
- *Transition Probabilities*: $p_t(s'|s,a) = p(s'|s,a)$ where $t \in T_N$.
- *Rewards*: $r_t(s,a) = \alpha^t r(s,a)$ where $t \in T_N$.
- *Terminal Reward* is zero for all states.

Theorem 2.9 - $v_{N+1}^*(\cdot)$

$$v_{N+1}^*(s) = \max_{a \in A(s)} \left(r(s,a) + \alpha \sum_{s' \in S} v_N^*(s') p(s'|s,a) \right) \quad \forall\ s \in S,\ N \geq 1$$

This can be interpreted as the optimality principle for $FHMDP(T_{N+1})$.

[5] This MDP is called the *Finite-Horizon Subproblem with Time-Horizon T_N* and is denoted *FHMDP(T_N)*

Proof 2.6 - *Theorem 2.9*
Since $v^*_{N+1,l}$ is the optimal value function for a $FHMDP(T_{N+1})$, thus we start with the optimality equation of this subproblem.

$$\begin{aligned} u^*_{N-1}(s) &= \max_{a\in A(s)}\left(r_l(s,a) + \sum_{s'\in S} u^*_{N+1,l+1}(s')p_l(s'|s,a)\right) \\ &= \max_{a\in A(s)}\left(\alpha^l \sum_{s'\in S} u^*_{N+1,l+1}(s')p(s'|s,a)\right) \end{aligned}$$

where $l = N,\dots,0$[6] and $y^*_{N+1,N+1}(s) = 0 \ \forall\ s = 0$.

`Theorem 2.5` states that the value generated in the last iteration of the *Value Function*, is the value of the *Optimal Value Function*, thus

$$v^*_{N+1}(s) = u^*_{N+1,0}(s)$$

Now consider the optimality equation for $FHMDP(T_N)$, which is defined as

$$\begin{aligned} u^*_{N,k}(s) &= \max_{a\in A(s)}\left(r_k(s,a) + \sum_{s'\in S} u^*_{N,k+1}(s')p_k(s'|s,a)\right) \\ &= \max_{a\in A(s)}\left(\alpha^k r(s,a) + \sum_{s'\in S} u^*_{N,k+1}(s')p(s'|s,a)\right) \end{aligned}$$

Similarly to above,

$$v^*_N(s) = u^*_{N,0}(s)$$

Consider the follow auxillary function

$$\begin{aligned} \tilde{u}^*_k(s) &= \alpha u^*_{N,k-1}(s) \quad \text{where } k\in[1,N+1] \\ \implies \quad u^*_{N,k}(s) &= \frac{1}{\alpha}\tilde{u}^*_{k+1}(s) \quad \text{where } k\in[0,N] \\ \text{and} \quad \tilde{u}^*_{N+1}(s) &= \alpha u^*_{N,N}(s) = 0 \quad \forall\ s\in S \end{aligned}$$

$$\begin{aligned} \implies \quad \tilde{u}^*_1(s) &= \alpha u^*_{N,0}(s) \\ &= \alpha v^*_N(s) \end{aligned}$$

We can derive the following definition

$$\begin{aligned} \frac{1}{\alpha}\tilde{u}^*_{k+1}(s) &= u^*_{N,k}(s) \\ &= \max_{a\in A(s)}\left(\alpha^k r(s,a) + \sum_{s'\in S} u^*_{N,k+1}(s')p(s'|s,a)\right) \\ &= \max_{a\in A(s)}\left(\alpha^k r(s,a) + \frac{1}{\alpha}\sum_{s'\in S} \tilde{u}^*_{k+2}(s')p(s'|s,a)\right) \\ \implies \quad \tilde{u}^*_{k+1}(s) &= \max_{a\in A(s)}\left(\alpha^{k+1} r(s,a) + \sum_{s'\in S} \tilde{u}^*_{k+2}(s')p(s'|s,a)\right) \quad \text{where } k = N-1,\dots,0 \end{aligned}$$

Comparing this expression with the first expression for $u^*_{N-1}(s)$ in terms of α, we conclude that this is just a sub-recursion of the earlier expression. Thus

$$\tilde{u}^*_l(s) = u^*_{N+1,l}(s) \quad \text{where } l\in[1,N+1]$$

[6] This being a backwards recursive definition is important

Setting $l = 1$ we can derive

$$\begin{aligned} u^*_{N+1,1}(s) &= \tilde{u}^*_1(s) \\ &= \alpha v^*_N(s) \end{aligned}$$

We can use this expression, and previously derived expression, we can conclude that

$$\begin{aligned} v^*_{N+1}(s) &= u^*_{N+1,0}(s) \\ &= \max_{a\in A(s)} \left(r(s,a) + \sum_{s'\in S} u^*_{N+1,1}(s')p(s'|s,a) \right) \\ &= \max_{a\in A(s)} \left(r(s,a) + \alpha \sum_{s'\in S} v^*_N(s')p(s'|s,a) \right) \end{aligned}$$

□

Proposition 2.17 - *Discounted Reward is Convergent*
Discounted reward is convergent

$$\sum_{t=0}^{\infty} \alpha^t |r(X_t, Y_t)| < \infty$$

Consequently, it is reasonable to expect the optimal value function to converge.

Proposition 2.18 - *Deriving Optimality Equation*
Given the convergence of discounted reward, we can derive the following

$$\begin{aligned} \lim_{N\to\infty} v^*_N(s) &= \lim_{N\to\infty} \left(\max_\pi \mathbb{E}^\pi \left(\sum_{t=0}^{N} \alpha^t r(X_t, Y_t) \middle| X_0 = s \right) \right) \\ &= \max_\pi \mathbb{E}^\pi \left(\lim_{N\to\infty} \sum_{t=0}^{N} \alpha^t r(X_t, Y_t) \middle| X_0 = s \right) \\ &= \max_\pi \mathbb{E}^\pi \left(\sum_{t=0}^{\infty} \alpha^t r(X_t, Y_t) \middle| X_0 = s \right) \\ &= v^*(s) \end{aligned}$$

This means that $\forall\ s \in S$ the following is true

$$\begin{aligned} v^*(s) &= \lim_{N\to\infty} v^*_{N+1}(s) \\ &= \lim_{N\to\infty} \left(\max_{a\in A(s)} \left(r(s,a) + \alpha \sum_{s'\in S} v^*_N(s')p(s'|s,a) \right) \right) \\ &= \max_{a\in A(s)} \left(r(s,a) + \alpha \sum_{s'\in S} \left(\lim_{N\to\infty} v^*_N(s') \right) p(s'|s,a) \right) \\ &= \max_{a\in A(s)} \left(r(s,a) + \alpha \sum_{s'\in S} v^*(s')p(s'|s,a) \right) \\ \implies v^*(s) &= \max_{a\in A(s)} \left(r(s,a) + \alpha \sum_{s'\in S} v^*(s')p(s'|s,a) \right) \end{aligned}$$

This final expression is the *Optimality Equation* for *Discounted Reward Infinite-Time MDPs.*[7]

[7] This equation is also known as the *Bellman Equation.*

Analysing Optimality Equation

Remark 2.15 - *Notation*

π_d - The stationary policy based on the *Markovian Decision Function* $d : S \to A$.

V - The set of all functions mapping $S \to \mathbb{R}$.[8]

$\|v\|$ - The uniform norm in V defined as $\|v\| := \max_{s \in S} |v(s)|$ for $v \in V$.

T - A transform of the value function $T : V \to V$ defined as

$$(Tv)(s) := \max_{a \in A(s)} \left(r(s,a) + \alpha \sum_{s' \in S} v(s') p(s'|s,a) \right) \text{ for } v \in V,\ s \in S$$

Tv - The transform T of function $v \in V$.[9]

Remark 2.16 - *Compact Bellamn Equation*
The *Bellman Optimality Equation* can be written in the following compact way

$$Tv = v$$

Theorem 2.10 - *T is a Contractive Mapping*
Consider transform T and any elements $v', v'' \in V$.

$$\|Tv' - Tv''\| \leq \alpha \|v' - v''\|$$

Since $\alpha \in (0, 1)$ this means T is a *Contractive Mapping* with *Contraction Factor* α.

Proof 2.7 - *Theorem 2.10*
This proof is based on the inequality

$$\left| \max_{j \in [1,n]} x_k - \max_{k \in [1,n]} y_k \right| \leq \max_{k \in [1,n]} |x_k - y_k \quad \text{where } x_1, \dots, x_n, y_1, \dots, y_n \in \mathbb{R} \tag{2}$$

Let $v', v'' \in V$ and $s \in S$. By the definition of T and 2 we have

$$\begin{aligned}
|(Tv')(s) - (Tv'')(s)| &= \left| \max_{a \in A(s)} \left(r(s,a) + \alpha \sum_{s' \in S} v'(s') p(s'|s,a) \right) - \max_{a \in A(s)} \left(r(s,a) + \alpha \sum_{s' \in S} v''(s') p(s'|s,a) \right) \right| \\
&\leq \max_{a \in A(s)} \left| \left(r(s,a) + \alpha \sum_{s' \in S} v'(s') p(s'|s,a) \right) - \left(r(s,a) + \alpha \sum_{s' \in S} v''(s') p(s'|s,a) \right) \right| \\
&= \max_{a \in A(s)} \left| \alpha \sum_{s' \in S} \big(v'(s') - v''(s') \big) p(s'|s,a) \right| \\
&\leq \max_{a \in A(s)} \alpha \sum_{s' \in S} \big(v'(s') - v''(s') \big) p(s'|s,a)
\end{aligned}$$

[8] V is a linear space.
[9] $(Tv)(\cdot)$ is a function of $s \in S$, $Tv : S \to \mathbb{R}$.

We can use the definition of the *norm* $\|\cdot\|$ to bound the distance between $v'(s'), v''(s')$

$$\begin{aligned} |v'(s') - v''(s')| &\leq \max_{s'' \in S} |v'(s'') - v''(s'')| \\ &= \|v' - v''\| \end{aligned}$$

Applying this inequality to the above expression, gives us

$$\begin{aligned} |(Tv')(s) - (Tv'')(s)| &\leq \alpha \max_{a \in A(s)} \sum_{s' \in S} \big(v'(s') - v''(s')\big) p(s'|s,a) \\ &\leq \alpha \max_{a \in A(s)} \sum_{s' \in S} \|v' - v''\| p(s'|s,a) \\ &= \alpha \|v' - v''\| \max_{a \in A(s)} \sum_{s' \in S} p(s'|s,a) \\ &= \alpha \|v' - v''\| \\ \implies |(Tv')(s) - (Tv'')(s)| &\leq \alpha \|v' - v''\| \end{aligned}$$

Since s is an arbitrary element of S and by the def of max, we have

$$\max_{s \in S} \big|(Tv')(s) - (Tv'')(s)\big| \leq \alpha \|v' - v''\|$$

By the definition of the *norm* $\|\cdot\|$ we have

$$\begin{aligned} \|Tv' - Tv''\| &= \max_{s \in S} \big|(Tv')(s) - (Tv'')(s)\big| \\ &\leq \alpha \|v' - v''\| \\ \implies \|Tv' - Tv''\| &\leq \alpha \|v' - v''\| \end{aligned}$$

□

Definition 2.29 - *Transform T_d*
For a *Markovian Decision Function* $d : S \to A$, we define the transform $T_d : V \to V$ as

$$(T_d V)(s) := r\big(s, d(s)\big) + \alpha \sum_{s' \in S} v(s') p\big(s'|s, d(s)\big) \text{ where } v \in V,\ s \in S$$

$t_d v$ is *affine* in v.

Theorem 2.11 - *Bounding on the Distance between Transformations*
For all $v', v'' \in V$ and *Markovian Decision Functions* $d : S \to A$, we have

$$\|T_d v' - T_d v''\| \leq \alpha \|v' - v''\|$$

Theorem 2.12 - *Banach Fixed-Point Theorem applied to T*
The following is the *Banach Fixed-Point Theorem* applied to transform T.[10]

i). Let $v_0 \in V$ and $\{v_k\}_{k \geq 0}$ be recursively defined by $v_{k+1}(s) = (Tv_k)(s)$. Then,

$$v(s) = \lim_{k \to \infty} v_k(s) \quad \forall\ s \in S$$

Moreover, we have $(Tv)(s) = v(s)\ \forall\ s \in S$ wrt transform T

$$\|v_k - v\| \leq \frac{\alpha^k}{1 - \alpha} \|v_1 - v_0\| \quad \forall\ k \geq 1$$

This shows that convergence occurs at an exponential rate $\frac{\alpha^k}{1-\alpha}$ and the limit of this value $\|v_1 - v_0\|$ is the unique solution to the *Fixed Point Equation* of T[11].

[10] $(Tv)(s) = v(s)$ is the *Fixed Point Equation* and $v_{k+1} = Tv_k$ is the *Fixed Point Recursion*, both wrt transform T.
[11] Is this the correct limit?

ii). If $(Tv')(s) = v'(s)$ for some $v' \in S$ and $\forall\ s \in S$, then

$$v(s) = v'(s) \quad \forall\ s \in S$$

Proof 2.8 - *Theorem 2.12 i)*
By `Theorem 2.10` and the *Fixed Point Recursion* of T, $v_{k+1} = Tv_k$ we get

$$\begin{aligned} \|v_{k+1} - v_k\| &= \|Tv_k - Tv_{k-1}\| \\ &\leq \alpha\|v_k - v_{k-1}\| \end{aligned}$$

By considering the iterations of the above with $k = 1, 2, 3, \dots$ we get

$$\begin{aligned} \|v_2 - v_1\| &\leq \alpha\|v - 1 - v_0\| \\ \|v_3 - v_2\| &\leq \alpha\|v - 2 - v_1\| \\ &\leq \alpha^2\|v_1 - v_0\| \\ \|v_4 - v_3\| &\leq \alpha\|v - 3 - v_2\| \\ &\leq \alpha^3\|v_1 - v_0\| \\ &\vdots \end{aligned}$$

Hence, get can state the general formula that

$$\|v_{k+1} - v_k\| \leq \alpha^k\|v_1 - v_0\| \quad \text{k}\geq 0\text{k}\geq 0_{\text{k}\geq 0_{(3)}}$$

Hence

$$\begin{aligned} &\sum_{k=0}^{\infty} \|v_{k+1} - v_k\| \\ &\qquad \leq \sum_{k=0}^{\infty} \alpha^k\|v_1 - v_0\| \\ &\qquad = \frac{\|v_1 - v_0\|}{1 - \alpha} < \infty \end{aligned}$$

Consequently, the series $\left\{\sum_{k=0}^{\infty}(v_{k+1} - v_k)\right\}_{k\geq 1}$ is well-defined and finite.
Let v be the element in V which is defined as

$$v := v_0 + \sum_{k=0}^{\infty}(v_{k+1} - v_k)$$

As the series is well-defined, so is v.
We have

$$\begin{aligned} v - v_k &= \left(v_0 + \sum_{i=0}^{\infty}(v_{i+1} - v_i)\right) - \left(v_0 + \sum_{i=0}^{k-1}(v_{i+1} - v_i)\right) \\ &= \sum_{i=k}^{\infty}(v_{i+1} - v_i) \end{aligned}$$

Combining this result with 3, we get

$$\begin{aligned} \|v - v_k\| &= \left\| \sum_{i=k}^{\infty} (v_{i+1} - v_i) \right\| \\ &\leq \sum_{i=k}^{\infty} \|v_{i+1} - v_i\| \\ &\leq \sum_{i=k}^{\infty} \alpha^i \|v_1 - v_0\| \\ &= \frac{\alpha^k}{1-\alpha} \|v_1 - v_0\} \\ \implies \|v - v_k\| &= \frac{\alpha^k}{1-\alpha} \|v_1 - v_0\} \end{aligned}$$

Note that

$$\lim_{k \to \infty} \|v_k - v\| = 0$$

Thus,

$$\lim_{k \to \infty} v_k = v$$

By `Theorem 2.10` and the *Fixed Point Recursion* of T, $v_{k+1} = Tv_k$, we can conclude that

$$\begin{aligned} \|Tv - v\| &= \|(Tv - v_{k+1}) + (v_{k+1} - v)\| \\ &= \|(Tv - Tv_k) + (v_{k+1} - v)\| \\ &\leq \|Tv - Tv_k\| + \|v_{k+1} - v\| \\ &\leq \alpha \|v_k - v\| + \|v_{k+1} - v\| \\ \implies \|Tv - v\| &\leq \alpha \|v_k - v\| + \|v_{k+1} - v\| \end{aligned}$$

Letting $k \to \infty$ in this inequality, we get

$$\begin{aligned} \|Tv - v\| &\leq \lim_{k \to \infty} \left(\alpha \|v_k - v\| + \|v_{k+1} - v\| \right) \\ &= \alpha \lim_{k \to \infty} \|v_k - v\| + \lim_{k \to \infty} \|v_{k+1} - v\| \\ &= \alpha \cdot 0 + 0 \\ &= 0 \implies \\ \|Tv - v\| &= 0 \\ \implies Tv &= v \end{aligned}$$

□

Proof 2.9 - *Theorem 2.12 ii)*
By `Theorem 2.10` and that $Tv = v$, $Tv' = v'$ by `Theorem 2.12i)`, we get that

$$\begin{aligned} \|v - v'\| &= \|Tv - Tv'\| \\ &\leq \alpha \|v - v'\| \\ \implies (1-\alpha)\|v - v'\| &\leq 0 \\ \implies \|v - v'\| &\leq 0 \text{ since } 1 - \alpha > 0 \\ \implies \|v - v'\| &= 0 \text{ by def. } \|\cdot\| \\ \implies v &= v' \end{aligned}$$

□

Theorem 2.13 - *Banach Fixed Point Theorem applied to T_d*
Let $d : S \to A$ be a *Markov Decision Function*. Then the following hold

i). $\exists!\ v_d \in V$ st $(T_d v_d)(s) = v_d(s)\ \forall\ s \in S$.[12]

ii). $v_d(s) = v^{\pi_d}(s)\ \forall\ s \in S$ where π_d is the stationary policy based on *Markovian Decision Problem d*.

This can re-interpretted as the *Fixed Point Equation* of T_d, $T_d v = v$, has a unique solution v_d which is the value function associated with π_d, $v_d = v^{\pi_d}$.

Proof 2.10 - *Theorem 2.13 i)*
This is proved by using `Theorem 2.11` and repeating, word-for-word, the argument of `Theorem 2.12` which concludes that $T_d v = v$ has a unique solution v_d.

Proof 2.11 - *Theorem 2.13 ii)*
Assume that $\{(X_t, Y_t)\}_{t\geq 0}$ is generated by policy π_d. By `Theorem 2.2` we know the following properties

i). $Y_t = d(X_t)\ \forall\ t \geq 0$.

ii). $\{X_t\}_{t\geq 0}$ is an *Homogeneous Markov Chain*.

iii). $p(s'|s, d(s))$ is the transition kernel of $\{X_t\}_{t\geq 0}$.

Consequently

$$\begin{aligned} \mathbb{P}^{\pi_d}(X_{t+1} = s'|X_1, X_0) &= \mathbb{P}^{\pi_d}(X_{t+1} = s'|X_1) \quad (1)\\ &= p(s'|X_1, d(X_1)) \\ \mathbb{P}^{\pi_d}(X_{t+1} = s'|X_1) &= \mathbb{P}^{\pi_d}(X_t = s'|X_0 = s) \end{aligned}$$

Step 1 - Representation of v^{π_d}.
By the definition of value functions for Discount Reward MDPs $v^{\pi_d}(\cdot)$, we have

$$\begin{aligned} v^{\pi_d}(s) &= \mathbb{E}^{\pi_d}\left(\sum_{t=0}^{\infty} \alpha^t r(X_t, Y_t)\Big| X_0 = s\right) \\ &= \mathbb{E}^{\pi_d}\big(r(X_0, d(X_0))\big|X_0 = s\big) + \sum_{t=1}^{\infty} \alpha^t \mathbb{E}^{\pi_d}\big(r(X_t, d(X_t))\big|X_0 = s\big) \\ &= r(s, d(s)) + \sum_{t=1}^{\infty} \alpha^t \mathbb{E}^{\pi_d}\big(r(X_t, d(X_t))\big|X_0 = s\big) \\ \implies v^{\pi_d}(s) &= r(s, d(s)) + \sum_{t=1}^{\infty} \alpha^t \mathbb{E}^{\pi_d}\big(r(X_t, d(X_t))\big|X_0 = s\big) \end{aligned}$$

Setting $t = k + 1 \implies k = t - 1$ we get

$$v^{\pi_d}(s) = r(s, d(s)) + \sum_{k=0}^{\infty} \alpha^{k+1} \mathbb{E}^{\pi_d}\big(e(X_{k+1}, d(X_{k+1}))\big|X_0 = s\big)$$

Due to the *filtering proeprty* of conditional expectations, we have

$$\mathbb{E}^{\pi_d}\big(r(X_{k+1}, d(X_{k+1}))\big|X_0 = s\big) = \mathbb{E}^{\pi_d}\big(\mathbb{E}^{\pi_d}\big(r(X_{k+1}, d(X_{k+1}))\big|X_1, X_0\big)\big|X_0 = s\big)$$

[12] This is *Banach's Fixed Point Theorem* applied to transform T_d where $T_d v = v$ is the *Fixed Point Equation* associated with transform T_d.

By the def. of conditional expectations and (1) we have

$$\begin{aligned}\mathbb{E}^{\pi_d}\big(r(X_{k+1},d(X_{k+1}))\big|X_1,X_0\big) &= \sum_{s'\in S} r(s',d(s'))\mathbb{P}^{\pi_d}(X_{k+1}=s'|X_1,X_0)\\ &= \sum_{s'\in S} r(s',d(s'))\mathbb{P}^{\pi_d}(X_{k+1}=s'|X_1)\\ &= \mathbb{E}^{\pi_d}\big(r(X_{k+1},d(X_{k+1}))\big|X_1\big)\end{aligned}$$

Hence

$$\begin{aligned}&\mathbb{E}^{\pi_d}\big(r(X_{k+1},d(X_{k+1}))\big|X_1,X_0\big) = \mathbb{E}^{\pi_d}\big(r(X_{k+1},d(X_{k+1}))\big|X_1\big)\\ \implies\ &\mathbb{E}^{\pi_d}\big(r(X_{k+1},d(X_{k+1}))\big|X_0=s\big) = \mathbb{E}^{\pi_d}\big(\mathbb{E}^{\pi_d}\big(r(X_{k+1},d(X_{k+1}))\big|X_1\big)\big|X_0=s\big)\end{aligned}$$

Using this, we can conclude a representation of $v^{\pi_d}(\cdot)$

$$\begin{aligned}v^{\pi_d}(s) &= r(s,d(s)) + \sum_{t=1}^{\infty}\alpha^t\mathbb{E}^{\pi_d}\big(r(X_t,d(X_t))\big|X_0=s\big)\\ &= r(s,d(s)) + \sum_{k=0}^{\infty}\alpha^{k+1}\mathbb{E}^{\pi_d}\big(\mathbb{E}^{\pi_d}(r(X_{k+1},d(X_{k+1}))|X_1)\big|X_0=s\big)\\ &= r(s,d(s)) + \alpha\mathbb{E}^{\pi_d}\left(\mathbb{E}^{\pi_d}\left(\sum_{k=0}^{\infty}\alpha^k r(X_{k+1},d(X_{k+1}))\middle|X_1\right)\middle|X_0=s\right)\\ \implies v^{\pi_d}(s) &= r(s,d(s)) + \alpha\mathbb{E}^{\pi_d}\left(\mathbb{E}^{\pi_d}\left(\sum_{k=0}^{\infty}\alpha^k r(X_{k+1},d(X_{k+1}))\middle|X_1\right)\middle|X_0=s\right)\end{aligned}$$

Let $R^{\pi_d}(s)$ denote the expected total reward when using policy π_d and $X_1 = s$

$$R^{\pi_d}(s) := \mathbb{E}^{\pi_d}\left(\sum_{k=0}^{\infty}\alpha^k r(X_{k+1},d(X_{k+1}))\middle|X_1=s\right)$$

Thus

$$v^{\pi_d}(s) = r(s,d(s)) + \alpha\mathbb{E}^{\pi_d}\left[R^{\pi_d}(X_1)|X_0=s\right]$$

Step 2 - Identification of R^{π_d}.

We want to find the link between $R^{\pi_d}(s)$ and $v^{\pi_d}(s)$. We can re-write R^{π_d} as

$$R^{\pi_d}(s) = \sum_{k=0}^{\infty}\alpha^k\mathbb{E}^{\pi_d}\left[r(X_{k+1},d(X_{k+1}))\big|X_1=s\right]$$

By the definition of conditional expectations

$$\begin{aligned}\mathbb{E}^{\pi_d}\left[r(X_{k+1},d(X_{k+1}))\big|X_1=s\right] &= \sum_{s'\in S} r(s',d(s'))\mathbb{P}^{\pi_d}(X_{k+1}=s'|X_1=s)\\ &= \sum_{s'\in S} r(s',d(s'))\mathbb{P}^{\pi_d}(X_k=s'|X_0=s)\\ &= \mathbb{E}^{\pi_d}\left[r(X_k,d(X_k))\big|X_0=s\right]\\ \implies \mathbb{E}^{\pi_d}\left[r(X_{k+1},d(X_{k+1}))\big|X_1=s\right] &= \mathbb{E}^{\pi_d}\left[r(X_k,d(X_k))\big|X_0=s\right]\end{aligned}$$

This gives the following re-expression of $R^{\pi_d}(s)$

$$\begin{aligned}R^{\pi_d}(s) &= \sum_{k=0}^{\infty}\alpha^k\mathbb{E}^{\pi_d}\left[r(X_{k+1},d(X_{k+1}))|X_1=s\right]\\ &= \sum_{k=0}^{\infty}\alpha^k\mathbb{E}^{\pi_d}\left[r(X_k,d(X_k))|X_0=s\right]\\ &= \mathbb{E}^{\pi_d}\left(\sum_{k=0}^{\infty}\alpha^k r(X_k,d(X_k))\middle|X_0=s\right)\\ &= v^{\pi_d}(s)\\ \implies R^{\pi_d}(s) &= v^{\pi_d}(s)\end{aligned}$$

Step 3 - Conclusion We want to show that v^{π_d} is a solution to $T_d v = v$.

By *Step 1* we have

$$v^{\pi_d}(s) = r(s, d(s)) + \alpha \mathbb{E}^{\pi_d} \left[R^{\pi_d}(X_1) | X_0 = s \right]$$

and, by *Step 2* we have

$$R^{\pi_d}(s) = v^{\pi_d}(s)$$

This means that

$$v^{\pi_d}(s) = r(s, d(s)) + \alpha \mathbb{E}^{\pi_d} \left[v^{\pi_d}(X_1) | X_0 = s \right]$$

By this result and the definition of conditional expectations, we have

$$\begin{aligned} v^{\pi_d}(s) &= r(s, d(s)) + \alpha \mathbb{E}^{\pi_d} \left[v^{\pi_d}(X_1) | X_0 = s \right] \\ &= r(s, d(s)) + \alpha \sum_{s' \in S} v^{\pi_d} \mathbb{P}^{\pi_d}(X_1 = s' | X_0 = s) \\ &= r(s, d(s)) + \alpha \sum_{s' \in S} v^{\pi_d}(s') p(s' | s, d(s)) \\ \implies v^{\pi_d}(s) &= r(s, d(s)) + \alpha \sum_{s' \in S} v^{\pi_d}(s') p(s' | s, d(s)) \end{aligned}$$

By this result and the definition of transform T_d we have

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

This means that v^{π_d} is a solution to $T_d v = v$.

In `Theorem 2.13 i)` $v_d(s)$ is the unique solution to $T_d v = v$, thus $v_d = v^{\pi_d}$. □

Theorem 2.14 - *Upper-Bound on* $\|v_N^* - v^*\|$

Here is a bound on the distance between the optimal value function in the infinite-time case and in the finite-time case (ie N epoch)

$$\|v_N^* - v^*\| \leq \frac{c\alpha^N}{1-\alpha} \ \forall \ N \geq 1 \text{ where } c := \max_{s \in S, a \in A(s)} |r(s, a)|$$

Proof 2.12 - *Theorem 2.14*

By the definition of c, we have $|r(X_t, Y_t)| \leq c$. Hence

$$\begin{aligned} \left| \sum_{t=0}^{\infty} \alpha^t r(X_t, Y_t) - \sum_{t=0}^{N_1} \alpha^t r(X_t, Y_t) \right| &= \left| \sum_{t=N}^{\infty} \alpha^t r(X_t, Y_t) \right| \\ &\leq \sum_{t=N}^{\infty} \alpha^t |r(X_t, Y_t)| \\ &\leq \sum_{t=N}^{\infty} \alpha^t c \\ &= \frac{\alpha^N}{1-\alpha} c \\ \implies \left| \sum_{t=0}^{\infty} \alpha^t r(X_t, Y_t) - \sum_{t=0}^{N_1} \alpha^t r(X_t, Y_t) \right| &= \frac{\alpha^N}{1-\alpha} c \\ \implies \sum_{t=0}^{N-1} \alpha^t r(X_t, Y_t) &\leq \sum_{t=0}^{\infty} \alpha^t r(X_t, Y_t) + \frac{\alpha^N}{1-\alpha} c \\ \text{and} \quad \sum_{t=0}^{N-1} \alpha^t r(X_t, Y_t) &\geq \sum_{t=0}^{\infty} \alpha^t r(X_t, Y_t) - \frac{\alpha^N}{1-\alpha} c \end{aligned}$$

Taking expectations of both sides of these last two expressions we get

$$\begin{aligned}\mathbb{E}^\pi\left(\sum_{t=0}^{N-1}\alpha^t r(X_t,Y_t)\middle|X_0=s\right) &\leq \mathbb{E}^\pi\left(\sum_{t=0}^{\infty}\alpha^t r(X_t,Y_t)\middle|X_0=s\right)+\frac{\alpha^n}{1-\alpha}c\\ \text{and } \mathbb{E}^\pi\left(\sum_{t=0}^{N-1}\alpha^t r(X_t,Y_t)\middle|X_0=s\right) &\geq \mathbb{E}^\pi\left(\sum_{t=0}^{\infty}\alpha^t r(X_t,Y_t)\middle|X_0=s\right)-\frac{\alpha^n}{1-\alpha}c\end{aligned}$$

Applying the definition of $v^\pi(s)$ to these results we get

$$\begin{aligned}\mathbb{E}^\pi\left(\sum_{t=0}^{N-1}\alpha^t r(X_t,Y_t)\middle|X_0=s\right) &\leq v^\pi(s)+\frac{\alpha^N}{1-\alpha}c\\ \text{and } \mathbb{E}^\pi\left(\sum_{t=0}^{N-1}\alpha^t r(X_t,Y_t)\middle|X_0=s\right) &\geq v^\pi(s)-\frac{\alpha^N}{1-\alpha}c\end{aligned}$$

Taking the maximum value when considering all policies $\max_\pi$ of both sides we get

$$\begin{aligned}\max_\pi \mathbb{E}^\pi\left(\sum_{t=0}^{N-1}\alpha^t r(X_t,Y_t)\middle|X_0=s\right) &\leq \max_\pi v^\pi(s)+\frac{\alpha^N}{1-\alpha}c\\ \text{and } \max_\pi \mathbb{E}^\pi\left(\sum_{t=0}^{N-1}\alpha^t r(X_t,Y_t)\middle|X_0=s\right) &\geq \max_\pi v^\pi(s)-\frac{\alpha^N}{1-\alpha}c\end{aligned}$$

From the definition of $v_N^*(s)$ we have

$$\begin{aligned}v_N^*(s) &\leq v^*(s)+\frac{\alpha^N}{1-\alpha}c\\ \text{and}\quad v_N^*(s) &\geq v^*(s)-\frac{\alpha^N}{1-\alpha}c\\ \implies -\frac{\alpha^N}{1-\alpha}c &\leq v_N^*(s)-v^*(s)\leq\frac{\alpha^N}{1-\alpha}c\\ \implies |v_N^*(s)-v^*(s)| &\leq \frac{\alpha^N}{1-\alpha}c\end{aligned}$$

As s is an arbitrary element of S, we can deduce that

$$\begin{aligned}\max_{s\in S}|v_N^*(s)-v^*(s)| &\leq \frac{\alpha^N}{1-\alpha}c\\ \implies \|v_N^*-v^*\| &\leq \frac{\alpha^N}{1-\alpha}c \text{ by def. } \|\cdot\|\end{aligned}$$

□

Theorem 2.15 - *Solution to Bellman Equation*
$v^*(s)$ is the unique solution to the *Bellman Equation.*

$$(Tv^*)(s)=v^*(s)\ \forall\ s\in S$$

Moreover, $v^*(s)=v^{\pi^*}(s)\ \forall\ s\in S$.

The *Optimal Markovian Decision Function* $d^*(s)$ is defined by

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

The *Optimal Policy* $\pi^* = \pi_{d^*}$ is the stationary policy in $HR(T)$ based on decision function $d^*(s)$.

Proof 2.13 - *Theorem 2.15*
By `Theorem 2.9` and the definition of T we get

$$\begin{aligned} v^*_{N+1}(s) &= \max_{a\in A(s)} \left(r(s,a) + \alpha \sum_{s'\in S} v^*_N(s') p(s'|s,a) \right) \\ &= (Tv^*_N)(s) \end{aligned}$$

Hence, the sequence $\{v^*_N\}_{N\geq 1}$ is generated through the recursion $v^*_{N+1}(s) = (Tv^*_N)(s)$ for $s \in S,\ N \geq 1$.

`Theorem 2.12` implies the following

i). $\lim_{N\to\infty} v^*_N(s)$ exists $\forall\ s \in S$.

ii). $\lim_{N\to\infty} v^*_N(s)$ is the unique solution to $Tv = v$.

By `Theorem 2.14`, we have

$$\begin{aligned} & & \|v^*_N - v^*\| &\leq \frac{\alpha^N}{1-\alpha} c \\ &\implies & \lim_{N\to\infty} \|v^*_N - v^*\| &\leq \lim_{N\to\infty} \frac{\alpha^N}{1-\alpha} c \\ & & &= 0 \\ &\implies & \lim_{N\to\infty} \|v^*_N - v^*\| &= 0 \\ &\implies & \lim_{N\to\infty} v^*_N(s) &= v^*(s) \end{aligned}$$

Since $\lim_{N\to\infty} v^*_N(s)$ is the unique solution to $Tv = v$, $v^*(s)$ is also the unique solution to the same equation.

$$(Tv^*)(s) = v^*(s)$$

We define the *Markov Decision Function* $d^*(s)$ as

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

Consequently, we get

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

By the definitions of T, T_{d^*} we get

$$\begin{aligned} (Tv^*)(s) &= \max_{a\in A(s)} \left(r(s,a) + \alpha \sum_{s'\in S} v^*(s') p(s'|s,a) \right) \\ &= r(s,d^*(s)) + \alpha \sum_{s'\in S} v^*(s') p(s'|s,d^*(s)) \\ &= (T_{d^*} v^*)(s) \\ \implies (Tv^*)(s) &= (T_{d^*} v^*)(s) \end{aligned}$$

Futher, we can deduce

$$\begin{aligned} v^*(s) &= (Tv^*)(s) \\ &= (T_{d^*}v^*)(s) \end{aligned}$$

Hence, $v^*(s)$ is a solution to $T_{d^*}v = v$. By `Theorem 2.13`, $v^{\pi_{d^*}}(s) = v^{\pi^*}(s)$ is the unique solution to $T_{d^*}(v) = v$ meaning $v^*(s) = v^{\pi^*}(s)$. □

Proof 2.14 - $\pi^* := \pi_{d^*}$ *is the Optimal Policy*
By `Theorem 2.15` we have that

$$v^*(s) = v^{\pi^*}(s) \ \forall \ s \in S$$

From the definition of *Optimal Value Functions* for *Discounted Reward MDPs*, we have

$$v^*(s) = \max_\pi v^\pi(s) \ \forall \ s \in S$$

Hence

$$v^\pi(s) \leq v^*(s) \ \forall \ s \in S$$

where this is an equality when $\pi = \pi^*$.

By the *Tower Property*, we can conclude

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

By the definition of *Value Functions* for *Discounted Reward MDPs*, we can deduce

$$\begin{aligned} & & v^\pi(X_0) &= \mathbb{E}^\pi\left[\sum_{t=0}^\infty \alpha^t r(X_t, Y_t)\middle| X_0\right] \\ \implies & \mathbb{E}^\pi\left[\sum_{t=0}^\infty \alpha^t r(X_t, Y_t)\middle| X_0\right] &= \mathbb{E}^\pi[v^\pi(X_0)] \\ & &\leq \mathbb{E}^\pi[v^*(X_0)] \\ \implies & \mathbb{E}^\pi\left[\sum_{t=0}^\infty \alpha^t r(X_t, Y_t)\middle| X_0\right] &\leq \mathbb{E}^\pi[v^*(X_0)] \end{aligned}$$

where the final expression is an equality when $\pi = \pi^*$.

Since X_0 is independent of π, we have that

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

Thus

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

which is an equality when $\pi = \pi^*$.

Hence, the maximum value of the expected discounted reward is $\mathbb{E}[v^*(X_0)]$ and this is attained when $\pi = \pi^*$. Therefore, π^* is an *Optimal Policy.* □

Policy Iteration Algorithm

Definition 2.30 - *Policy Iteration Algorithm*
The *Policy Iteration Algorithm* numerical solves any given *Discount Reward MDP* by calculating the *Optimal Value Function* and *Optimal Policy.*

Here are the stages of the *Policy Iteration Algorithm*

- *Initialisation* - Arbitrarily choose a *Markovian Decision Function* $d_0(s)$ and set $k = 0$.
- *Body* - For each $k \geq 0$, execute the following steps

  i). *Policy Evaluation* - Compute a solution $v_k(s)$ to the equation $v = T_{d_k} v$ where v is unknown.

  ii). *Policy Improvement* - Select a *Markovian Decision Function* $d_{k+1}(s)$ st

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

  iii). *Terminate?*
  - If $d_k(s) = d_{k+1}(s) \ \forall \ s \in S$: STOP. return $\hat{d}(s) := d_k(s)$ and $\hat{v}(s) = v_k(s)$.
  - Else: Increment k and repeat i)-iii).

Theorem 2.16 - *Properties of Policy Iteration Algorithm*
Here are some properties of the *Policy Iteration Algorithm*

i). The *Policy Iteration Algorithm* stops after a finite number of iterations.

ii). $\hat{d}(s) = d^*(s)$ and $\hat{v}(s) = v^*(s) \ \forall \ s \in S$.

Remark 2.17 - *Remarks about the Policy Iteration Algorithm*

- By `Theorem 2.13`, $v_k(s) = v^{\pi_k}(s) \ \forall \ s \in S$.
- $v^{\pi_k}(s)$ is the value function associated with π_k.
- π_k is the stationary policy in $HR(T)$ based on $d_k(s)$.

Equivalent Linear Programming Problem

Proposition 2.19 - *Linear Programming*
Linear Programming techniques can be used to numerically solve *Discounted Reward MDPs.*

Proposition 2.20 - *Equivalent Linear Program*
Here is a *Linear Program* which is equivalent to a *Discounted Reward MDP.*

- *Minimise* $\sum_{s \in S} \gamma(s) v(s)$. *Subject to* $r(s,a) + \sum_{s' \in S} \alpha p(s'|s,a) v(s') \leq v(s) \ \forall \ s \in S, a \in A(s)$.

The minimisation is in $v : S \to \mathbb{R}$, while $\gamma : S \to \mathbb{R}$ is any function which satisfies $\gamma(s) > 0 \ \forall \ s \in S$.

Proposition 2.21 - *Notation* - $\hat{v}(s), \hat{d}(s)$

- $\hat{v}(s)$ - Optimal solution to the equivalent linear program.
- $\hat{d}(s)$ - A *Markovian Decision Function* which satisfies

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

Theorem 2.17 -
$\hat{d}(s) = d^*(s)$ and $\hat{v}(s) = v^*(s) \ \forall \ s \in S$.

2.2.4 Average Reward Infinite-Horizon MDPs

Problem Formulation

Definition 2.31 - *Formulation of Average Reward Infinite-Horizon MDPs*
Here is the specification of the components of an *Average Reward Infinite-Horizon MDP*.

- *Number of Decision Epochs* - $N = \infty$.
- *Time-Horizon* - $T = \{0, 1, \dots\}$.
- *Transition Probabilities* - $p_t(s'|s, a) = p(s'|s, a)\ \forall\ t \in T$.
- *Rewards* $r_t(s, a) = r(s, a)\ \forall\ t \in T$.

,$\mathrm{Y}_{0:t} \sim X_{t+1}|X_t, Y_t$
$\sum p_t(\cdot|X_t, Y_t)$
$= p(\cdot|X_t, Y_t)$

Definition 2.32 - *Objective of an Average Reward Infinite-Horizon MDP*
Given the transition probabilities $p(s'|s, a)$ and the rewards $r(s, a)$, the agent is tasked to find a *History Dependent Randomised Policy* $\pi \in HT(T)$ which maximises the expected average reward per epoch over the time-horizon T.

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

Proposition 2.22 - *Derivation of Problem*
Consider the following two assumptions

i). The parameters of the stochastic system and the parameters of immediate rewards change slowly in time.

ii). $N \gg 1$.

By `i)` it is reasonable to approximate the transition probabilities and rewards as

$$p_t(s'|s, a) \approx p(s'|s, a) \quad r_t(s'|s, a) \approx r(s, a)$$

By `ii)`, we can neglect the terminal reward $r_N(S_N)$ as it is significantly less than the total reward to that point. Thus

$$\begin{aligned}\mathbb{E}^\pi \left[r_N(X_N)\sum_{t=0}^{N-1} r_t(X_t, Y_t)\right] &\approx \mathbb{E}^\pi \left[\sum_{t=0}^{N-1} r_t(X_t, Y_t)\right] \\ &\approx \mathbb{E}^\pi \left[\sum_{t=0}^{N-1} r(X_t, Y_t)\right]\end{aligned}$$

The problem of maximising the total reward and average reward are the same, as introducing a multiplicative constant $\frac{1}{N}$ does not affect the maximisation problem.

$$\begin{aligned}\mathrm{argmax}_\pi \mathbb{E}^\pi \left[\sum_{t=0}^{N-1} r(X_t, Y_t)\right] &= \mathrm{argmax}_\pi \frac{1}{N}\mathbb{E}^\pi \left[\sum_{t=0}^{N-1} r(X_t, Y_t)\right] \\ &= \mathrm{argmax}_\pi \mathbb{E}^\pi \left[\frac{1}{N}\sum_{t=0}^{N-1} r(X_t, Y_t)\right]\end{aligned}$$

By `ii)` it is reasonable to approximate the true expectation with its limit

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

As the limit is not guaranteed to exist, we take the limit of the infimum of the expectation.

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

Proposition 2.23 - *Existence of Limits*
Since the state-space S and action-space A are finite sets, we have that

$$c = \max_{s\in S, a\in A(s)} |r(s,a)| < \infty$$

Consequently, the average reward is finite

$$\left|\frac{1}{N}\sum_{t=0}^{N-1} r(X_t, Y_t)\right| \leq \frac{1}{N}\sum_{t=0}^{N-1} |r(X_t, Y_t)| \leq c$$

Therefore, the limit of the infimum and the limit of the supremum exist and are finite[13]

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

If policy π is optimal, then the limit of the expected average reward exists and is finite.

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

Proposition 2.24 - *Approximating Finite-Horzion MDP with Average Reward MDP*
The *Average Reward MDP* accurately approximates the *Finite-Horizon MDP* under the assumptions in `Proposition 2.22`

$$\begin{aligned}
& \operatorname{argmax}_{\pi} \mathbb{E}^{\pi}\left[\sum_{t=0}^{N-1} r_t(X_t, Y_t) + r_N(X_N)\right] \\
\approx\ & \operatorname{argmax}_{\pi} \mathbb{E}^{\pi}\left[\sum_{t=0}^{N-1} r(X_t, Y_t)\right] && \text{as } N \gg 1 \\
=\ & \operatorname{argmax}_{\pi} \mathbb{E}^{\pi}\left[\frac{1}{N}\sum_{t=0}^{N-1} r(X_t, Y_t)\right] \\
\approx\ & \operatorname{argmax}_{\pi} \left(\lim_{N\to\infty} \inf \mathbb{E}^{\pi}\left[\sum_{t=0}^{N-1} r(X_t, Y_t)\right]\right) && \text{as } N \gg 1
\end{aligned}$$

Remark 2.18 - *Average Reward MDP vs Discounted Reward MDP*

[13] It is shown in `Proposition 2.??` that these limits have the same value then an optimal policy π is used.

- When the discounting factor α is close to 0, *Discounted Reward MDPs* place emphasis on the present/near-future over the far-future.
- The *Average Reward MDP* gives the same emphasis to all received rewards.
- When the discounting factor α is close to 1, *Discounted Reward MDPs* and *Average Reward MDP* behave similarly.

Remark 2.19 - *Average Reward MDPs are very similar to Irreducible Markov Chains*
See `0. Reference` for details on *Irreducible Markov Chains*.

Optimality Equations

Remark 2.20 - *For the rest of this section*
Assume that under any stationary *Marovian Determinisic Policy* $\{X_t\}_{t\geq 0}$ is an irreducible *Markov Chain*. This is equivalent to the condition that $p(s'|s, d(s))$ is an irreducible transition kernel for any *Markovian Decision Function* $d : S \to A$.

Definition 2.33 - *Bellman Equation*
The *Bellman Equation* is the *Optimality Equation* for the *Average Reward MDP* and is defined as

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

where $r^* \in \mathbb{R}, w^* : S \to \mathbb{R}$ are unknown.

Theorem 2.18 - *Solution to Bellman equation*
There exists $r^* \in \mathbb{R},\ w^* : S \to \mathbb{R}$ st that the *Bellman Equation* is satisfied.

Proof 2.15 - *Theorem 2.18*
Steo 1 - *Discounted Reward MDPs and Average Reward MDPs.*

Consider a *Discounted Reward MDP* $DRMDP(\alpha)$ with the following elements

- Time-horizon $T = \{0, 1, \dots\}$.
- Transition probabilities $p_t(s'|s, a) = p(s'|s, a)\ \forall\ t \in T$.
- Rewards $r_t(s, a) = \alpha^t r(s, a)\ \forall\ t \in T$.
- Discounting factor $\alpha \in (0, 1)$.

Step 2 - *Notation.*

Let $v^*_\alpha(s)$ be the *Optimal Value Function* for $DRMDP(\alpha)$, $d^*_\alpha(s)$ be the *Markovian Decision Function* and π^*_α be the stationary policy based on $d^*_\alpha(s)$.

$$\begin{aligned} v^*_\alpha &= \max_\pi \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \alpha^t r(X_t, Y_t) \big| X_0 = s \right] \\ d^*_\alpha(s) &\in \text{argmax}_{a\in A(s)} \left(r(s,a) + \alpha \sum_{s'\in S} v^*_\alpha(s')p(s'|s,a) \right) \end{aligned}$$

Step 3 - *Properties of* $DRMDP(\alpha)$.

By `Theorem 2.15` v^*_α is the unique solution to the *Bellman Equation* for $DRMDP(\alpha)$ and is the value function associated with policy π^*_α. Meaning π^*_α is the optimal policy for $DRMDP(\alpha)$.

Step 4 - *Sequence of Discounting Factors.*

Here I construct a sequence $\{\alpha_n\}_{n\geq 0}$ in $(0,1)$ with some special properties. Let $\{\beta_n\}_{n\geq 0}$ be any increasing sequence in $(0,1)$ which satisfies

$$\lim_{n\to\infty} \beta_n = 1$$

Since S, A are both finite, there any a finite number of distinct *Markovian Decision Functions* which can be constructed. Hence, the sequence $\{d^*_{\beta_n}(s)\}_{n\geq 0}$ has finitely many distinct elements (elements may be repeated). Consequently, there exists a *Markovian Decision Function* $d : S \to A$ and a subsequence $\{\alpha_n\}_{n\geq 0}$ of $\{\beta_n\}_{n\geq 0}$ st $d^*_{\alpha_n}(s) = d(s)\ \forall\ s \in S, n \geq 0$. Note that $\lim_{n\to\infty} \alpha_n = 1$.

Step 5 - *Notation.*

Let $r_d(s)$ be the reward function and $p_d(s'|s)$ be the transition kernel when decision function d is used and π_d be the stationary policy based on $d(s)$.

$$\begin{aligned} r_d(s) &:= r(s, d(s)) \\ p_d(s'|s) &:= p(s'|s, d(s)) \end{aligned}$$

Assume that $\{(X_t, Y_t)\}_{t\geq 0}$ is generated with policy π_d.

Step 6 - *Properties of $p_d(s'|s)$ and $\{X_t\}_{t\geq 0}$.*

Since $\{(X_t, Y_t)\}_{t\geq 0}$ is generated with policy π_d then `Theorem 2.2` implies the following hold

i). $Y_t = d(X_t)\ \forall\ t \geq 0$.

ii). $\{X_t\}_{t\geq 0}$ is a *Homogeneous Markov Chain.*

iii). $p_d(s'|s) := p(s'|s, d(s))$ is the *Transition Kernel* of $\{X_t\}_{t\geq 0}$, further we assume that $\{X_t\}_{t\geq 0}$ is an *Irreducible Markov Chain.* `Theorem 0.2` implies that $\{X_t\}_{t\geq 0}$ has a unique *Invariant PMF* $\mu_d(s)$.

Let $\bar{r}_d$ be the mean reward wrt μ_d

$$\bar{r}_d := \sum_{s\in S} r_d(s)\mu_d(s)$$

`Theorem 0.3` implies that the *Poisson Equation* associated with *Markov Chain* $\{X_t\}_{t\geq 0}$ and function $r_d(s)$ has a solution $w'_d(s)$ st

$$r_d(s) - \bar{r}(d) = w'_d(s) - \sum_{s'\in S} w'_d(s')p_d(s'|s) \quad \forall\ s \in S$$

Let $w_d(s)$ be the zero-mean solution to the *Poisson Equation* associated with *Markov Chain* $\{X_t\}_{t\geq 0}$ and function $r_d(s)$ has a solution $w'_d(s)$

$$w_d(s) := w'_d(s) - \sum_{s'\in S} w'_d(s')\mu_d(s')$$

Let $v_\alpha(s)$ be the value function and $\tilde{v}(a)_s$ be the α-resolvent associated with Markov Chain $\{X_t\}_{t\geq 0}$ and function $r_d(s)$

$$\begin{aligned} v_\alpha(s) &:= \mathbb{E}^{\pi_d}\left[\sum_{t=0}^{\infty} \alpha^t r_d(X_t)\middle| X_0 = s\right] \\ \tilde{v}_\alpha(s) &:= v_\alpha(s) - \left(\frac{\bar{r}_d}{1-\alpha} - w_d(s)\right) \\ \implies v_\alpha(s) &= \left(\frac{\bar{r}_d}{1-\alpha} - w_d(s)\right) + \tilde{v}_\alpha(s) \end{aligned}$$

The last equation is the *Laurent Expansion* of $v_\alpha(s)$ and $\tilde{v}_\alpha(s)$ is the residual in this expansion. `Theorem 0.4` implies that

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

Step 7 - *Relationship between $v_{\alpha_n}(s)$ and $v^*_{\alpha_n}(s)$.*

Since $d^*_{\alpha_n}(s) = d(s)\ \forall\ s \in S,\ n \geq 0$ we have that $\pi^*_{\alpha_n} = \pi_d$. As $v^*_{\alpha_n}(s)$ is identical to the value function associated with $\pi^*_{\alpha_n}$ we can deduce that

$$\begin{aligned}
v^*_{\alpha_n}(s) &= \mathbb{E}^{\pi^*_{\alpha_n}}\left[\sum_{t=0}^{\infty} \alpha_n^t r(X_t, Y_t) \middle| X_0 = s\right] \\
&= \mathbb{E}^{\pi_d}\left[\sum_{t=0}^{\infty} \alpha_n^t r(X_t, Y_t) \middle| X_0 = s\right] \\
&= \mathbb{E}^{\pi_d}\left[\sum_{t=0}^{\infty} \alpha_n^t r(X_t, d(X_t)) \middle| X_0 = s\right] \quad \text{since using } \pi \\
&= \mathbb{E}^{\pi_d}\left[\sum_{t=0}^{\infty} \alpha_n^t r_d(X_t) \middle| X_0 = s\right] \quad \text{by def. } r_d(\cdot) \\
&= v_{\alpha_n}(s)
\end{aligned}$$

This shows that the value function for $FHMDP(\alpha_n)$ $v^*_{\alpha_n}$ is the same as the α_n-resolvent $v_{\alpha_n}(s)$ associated with the *Markov Chain* $\{X_t\}_{t \geq 0}$ and function $r_d(s)$.

Step 8 - *Final Conclusions.*

$v_{\alpha_n}(s) = v^*_{\alpha_n}(s)$ is the unique solution to the *Bellman Equation* for a $FHMDP(\alpha_n)$. Further,

$$\begin{aligned}
& & v_{\alpha_n} &= \frac{\bar{r}_d}{1-\alpha_n} + w_d(s) + \tilde{v}_{\alpha_n}(s) \\
\implies & & \frac{\bar{r}_d}{1-\alpha_n} + w_d(s) + \tilde{v}_{\alpha_n}(s) &= \max_{a \in A(s)} \left(r(s,a) + \alpha_n \sum_{s' \in S} v_{\alpha_n}(s') p(s'|s,a) \right) \\
& & &= \max_{a \in A(s)} \left(r(s,a) + \alpha_n \sum_{s' \in S} \left(\frac{\bar{r}_d}{1-\alpha_n} + w_d(s') + \tilde{v}_{\alpha_n}(s') \right) p(s'|s \right. \\
& & &= \max_{a \in A(s)} \left(r(s,a) + \frac{\alpha_n \bar{r}_d}{1-\alpha_n} + \alpha_n \sum_{s' \in S} \left(w_d(s') + \tilde{v}_{\alpha_n}(s') \right) p(s'|s, \right. \\
& & &= \frac{\alpha_n \bar{r}_d}{1-\alpha_n} + \max_{a \in A(s)} \left(r(s,a) + \alpha_n \sum_{s' \in S} \left(w_d(s') + \tilde{v}_{\alpha_n}(s') \right) p(s'|s, \right. \\
\implies & \left(\frac{\bar{r}_d}{1-\alpha_n} + w_d(s) + \tilde{v}_{\alpha_n}(s) \right) - \frac{\alpha_n \bar{r}_d}{1-\alpha_n} &= \max_{a \in A(s)} \left(r(s,a) + \alpha_n \sum_{s' \in S} \left(w_d(s') + \tilde{v}_{\alpha_n}(s') \right) p(s'|s,a) \right) \\
\implies & \bar{r}_d + w_d(s) + \tilde{v}_{\alpha_n}(s) &= \max_{a \in A(s)} \left(r(s,a) + \alpha_n \sum_{s' \in S} \left(w_d(s') + \tilde{v}_{\alpha_n}(s') \right) p(s'|s,a) \right)
\end{aligned}$$

Since $\lim_{n\to\infty} \alpha_n = 1$ and $\lim_{n\to\infty} \tilde{v}_{\alpha_n}(s) = 0$, by taking limits, we can further deduce that

$$\begin{aligned}
\bar{r}_d + w_d(s) &= \lim_{n\to\infty} (\bar{r}_d + w_d(s) + \tilde{v}_{\alpha_n}(s)) \\
&= \lim_{n\to\infty} \max_{a\in A(s)} \left(r(s,a) + \alpha_n \sum_{s'\in S} (w_d(s') + \tilde{v}_{\alpha_n}(s'))p(s'|s,a) \right) \\
&= \max_{a\in A(s)} \lim_{n\to\infty} \left(r(s,a) + \alpha_n \sum_{s'\in S} (w_d(s') + \tilde{v}_{\alpha_n}(s'))p(s'|s,a) \right) \\
&= \max_{a\in A(s)} \left(r(s,a) + \lim_{n\to\infty} \alpha_n \sum_{s'\in S} (w_d(s') + \tilde{v}_{\alpha_n}(s'))p(s'|s,a) \right) \\
&= \max_{a\in A(s)} \left(r(s,a) + \sum_{s'\in S} w_d(s')p(s'|s,a) \right) \\
\implies \bar{r}_d + w_d(s) &= \max_{a\in A(s)} \left(r(s,a) + \sum_{s'\in S} w_d(s')p(s'|s,a) \right)
\end{aligned}$$

This shows that $(\bar{r}_d, w_d(s))$ is a solution to the *Bellman Equation* □

Theorem 2.19 - *Value of r^* and the policy*
Let $r^* \in \mathbb{R}$, $w^* : S \to \mathbb{R}$ be a solution to the *Bellman Equation.*
Let $d^*(s)$ be the *Markovian Decision Function* defined as

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

and π^* be the stationary policy based on $d^*(s)$. Then, for any $\pi \in HR(T)$, r^* is the maximum asymptotic value of the expected average reward.

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

Further, is an optimal policy

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

Proof 2.16 - *Theorem 2.19*
Let $(r^*, w^*(s))$ be a solution to the *Bellman Equation* and $d^*(s)$ be a *Markovian Decision Function.*

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

Thus

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

where this is an inequality when $a = d^*(s)$.

Setting $s = X_T, a = Y_t$ we get

$$r^* + w^*(X_t) \geq r(X_t, Y_t) + \sum_{s' \in S} w^*(s') p(s'|X_t, Y_t)$$

where this is an inequality when $Y_t = d^*(X_t)$.

From the definition of conditional expectations and transition probabilities, we have

$$\begin{aligned} \sum_{s' \in S} w^*(s') p(s'|X_t, Y_t) &= \sum_{s' \in S} w^*(s') \mathbb{P}^\pi(X_{t+1} = s'|X_t, Y_t) \\ &= \mathbb{E}^\pi[w^*(X_{t+1})|X_t, Y_t] \\ \implies \qquad r^* + w^*(X_t) &\geq r(X_t, Y_t) + \mathbb{E}^\pi[w^*(X_{t+1})|X_t, Y_t] \end{aligned}$$

Taking expectations of both sides, we get

$$\mathbb{E}^\pi[r^* + w^*(X_t)] \geq \mathbb{E}^\pi[r(X_t, Y_t) + \mathbb{E}^\pi(w^*(X_{t+1})|X_t, Y_t)]$$

since $Y_t := d^*(X_t)$ under policy π^*, this is an equality when $\pi = \pi^*$.

By the tower property we can deduce

$$\begin{aligned} \mathbb{E}^\pi[r^* + w^*(X_t)] &\geq \mathbb{E}^\pi[r(X_t, Y_t) + \mathbb{E}^\pi(w^*(X_{t+1})|X_t, Y_t)] \\ &= \mathbb{E}^\pi[r(X_t, Y_t)] + \mathbb{E}^\pi[\mathbb{E}^\pi[w^*(X_{t+1})|X_t, Y_t]] \\ &= \mathbb{E}^\pi[r(X_t, Y_t)] + \mathbb{E}^\pi[w^*(X_{t+1})] \\ \implies \mathbb{E}^\pi[r^* + w^*(X_t)] &\geq \mathbb{E}^\pi[r(X_t, Y_t)] + \mathbb{E}^\pi[w^*(X_{t+1})] \\ \implies \mathbb{E}^\pi[r^*] &\geq \mathbb{E}^\pi[r(X_t, Y_t)] + \mathbb{E}^\pi[w^*(X_{t+1})] - \mathbb{E}^\pi[w^*(X_t)] \\ \implies r^* &\geq \mathbb{E}^\pi[r(X_t, Y_t)] + \mathbb{E}^\pi[w^*(X_{t+1})] - \mathbb{E}^\pi[w^*(X_t)] \\ \implies r^* &\geq \frac{1}{N} \sum_{t=0}^{N-1} \Big(\mathbb{E}^\pi[r(X_t, Y_t)] + \mathbb{E}^\pi[w^*(X_{t+1})] - \mathbb{E}^\pi(w^*(X_t)) \Big) \\ &= \frac{1}{N} \sum_{t=0}^{N-1} \mathbb{E}^\pi[r(X_t, Y_t)] + \frac{1}{N} \sum_{t=0}^{N-1} \Big(\mathbb{E}^\pi[w^*(X_{t+1})] - \mathbb{E}^\pi[w^*(X_t)] \Big) \\ &= \mathbb{E}^\pi \left(\frac{1}{N} \sum_{t=0}^{N-1} r(X_t, Y_t) \right) + \frac{1}{N} \Big(\mathbb{E}^\pi[w^*(X_N)] - \mathbb{E}^\pi[w^*(X_0)] \Big) \end{aligned}$$

with equality when $\pi = \pi^*$.

Since S is a finite set, we have $\max_{s \in S} |w^*(s)| < \infty$. Thus

$$\lim_{N \to \infty} \left(\frac{1}{N} \big(\mathbb{E}^\pi[w^*(X_N)] - \mathbb{E}^\pi[w^*(X_0)] \big) \right) = 0$$

Thus

$$\begin{aligned} r^* &\geq \lim_{N \to \infty} \sup \mathbb{E}^\pi \left(\frac{1}{N} \sum_{t=0}^{N-1} r(X_t, Y_t) \right) + \lim_{N \to \infty} \frac{1}{N} \Big(\mathbb{E}^\pi[w^*(X_N)] - \mathbb{E}^\pi[w^*(X_0)] \Big) \\ &= \lim_{N \to \infty} \sup \mathbb{E}^\pi \left(\frac{1}{N} \sum_{t=0}^{N-1} r(X_t, Y_t) \right) \\ \implies r^* &= \lim_{N \to \infty} \sup \mathbb{E}^{\pi^*} \left(\frac{1}{N} \sum_{t=0}^{N-1} r(X_t, Y_t) \right) \end{aligned}$$

□

Theorem 2.20 - *Uniqueness of solutions to Bellman Equation*
Let $(r^*, w^*(s)))$ and $(\tilde{r}^*, \tilde{w}^*(s))$ be solutions to the *Bellman Equation*. Then the following hold

i). $r^* = \tilde{r}^*$. (The r-part is unique).

ii). $\exists\ c \in \mathbb{R}$ st $w^*(s) = w^*(s) + c\ \forall\ s \in S$. (The w-part is unique with an additive constant)

Policy Iteration Algorithm

Definition 2.34 - *Policy Iteration Algorithm*
Here is the *Policy Iteration Algorithm* for finding an optimal policy for an *Average Reward MDP*

Initialisation - Arbitrarily select a *Markovian Decision Function* $d_0(s)$ and set $k = 0$.

Body For each $k \geq 0$ iteration:

i). *Policy Evaluation*

- Compute a solution $\mu_k(s)$ to the equations (where $\mu(\cdot)$ is unknown)[14]

$$\begin{aligned} \textstyle\sum_{s' \in S} \mu(s') &= 1 \\ \mu(s) &= \textstyle\sum_{s' \in S} p(s|s', d_k(s'))\mu(s') \end{aligned}$$

- Compute a solution $w_k(s)$ to the equation (where $w(\cdot)$ is unknown) using the $\mu_k(\cdot)$ computed in the previous step.[15]

$$\begin{aligned} w(s) - \textstyle\sum_{s' \in S} w(s')p(s'|s, d_k(s)) &= r(s, d_k(s)) - r_k \\ \text{where } r_k &:= \textstyle\sum_{s \in S} r(s, d_k(s))\mu_k(s) \end{aligned}$$

ii). *Policy Improvement*

- Select a *Markovian Decision Function* $d_{k+1}(s)$ st for all $s \in S$

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

iii). *Termination?* - If $d_k(s) = d_{k+1}(s)\ \forall\ s \in S$, then stop the algorithm and return $\hat{d}(s) = d_k(s)$. Otherwise, increment k and repeat `i)-iii)`

Theorem 2.21 - *Optimality of Policy Iteration Algorithm*
The following are true for the *Policy Iteration Algorithm*

i). The *Policy Iteration Algorithm* stops after a finite number of iterations.

ii). $\hat{d}(s) = d^*(s)\ \forall\ s \in S$

Equivalent Linear Programming Problem

Proposition 2.25 - *Equivalent Linear Programming Problem for Average Reward MDP*
The equivalent *Linear Programming Problem* to the *Average Reward MDP* is to

Minimise $r \in \mathbb{R}$ subject to

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

Here, $r \in \mathbb{R}$ and $w : S \to \mathbb{R}$ are unknown.

Theorem 2.22 - *Optimality of Equivalent Linear Programming Problem*

[14] $\mu_k(s)$ is the invariant pmf of the transition kernel $p(s'|s, d_k(s))$
[15] This is the *Poisson Equation* associated with function $r(s, d_k(s))$ and with the transition kernel $p(s'|s, d_k(s))$

Let $(\hat{r}, \hat{w}(s))$ be an optimal solution to the equivalent *Linear Programming Problem* and $\hat{d}(s)$ be a Markovian Decision function which satisfies

$$\hat{d}(s) \in \text{argmax}_{a \in A(s)} \left(r(s,a) + \sum_{s' \in S} \hat{w}(s') p(s'|s,a) \right) \quad \forall\ s \in S$$

Then

$$\hat{d}(s) = d^*(s) \quad \forall\ s \in S$$

3 Probability

Definition 3.1 - *Random Process*
A *Random Process* is a collection of random variables indexed by time $\{X_t\}_{t \in T}$ (e.g. flipping a coin several times). Each of these random variables can take a value from a state space S. A random process a *Discrete Time Process* if the index set T is discrete. A random process a *Continuous Time Process* if the index set T is continuous.

3.1 Probability Inequalities

Remark 3.1 - *We can use the moments of a random variable to determine bounds on the probability of it taking values in a certain set.*
Theorem 3.1 - *Markov's Inequality*
Let X be a non-negative random variable. Then

$$\forall\ c > 0 \quad \mathbb{P}(X \geq c) \leq \frac{\mathbb{E}(X)}{c}$$

Proof

Consider an event A and define its indicator $\mathbb{1}(A)(\omega) := \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$. Fix $c > 0$, then

$$\begin{aligned} \mathbb{E}(X) &\geq \mathbb{E}[X \mathbb{1}(X \geq c)] \\ &\geq \mathbb{E}[c \mathbb{1}(X \geq c)] \\ &= c\mathbb{P}(X \geq c) \\ \implies \mathbb{P}(X \geq c) &\leq \tfrac{1}{c}\mathbb{E}(X) \end{aligned}$$

Theorem 3.2 - *Chebyshev's Inequality*
Let X be a random-variable with finite mean and variance. Then

$$\forall\ c > 0 \quad \mathbb{P}(|X - \mathbb{E}(X)| \geq c) \leq \frac{\text{Var}(X)}{c^2}$$

Proof
Note that the events $|X - \mathbb{E}(X)| \geq c$ and $(X - \mathbb{E}(X))^2 \geq c^2)$ are equivalent. Note that $\text{Var}([X - \mathbb{E}(X)]^2) = \text{Var}(X)$. Then the result follows by *Markov's Inequality.*

Theorem 3.3 - *Chebyshev's Inequality for Sum of IIDs*

Let $X_1, \dots, X_n$ be IID random variables with finite mean μ and finite variance σ^2.

$$\forall\ c > 0 \quad \mathbb{P}\left(\left|\left(\sum_{i=1}^{n} X_i\right) - n\mu\right| \geq nc\right) \leq \frac{\sigma^2}{nc^2}$$

Proof
This is proved by extending the proof of `Theorem 2.2` and noting that the variance of a sum of IIDs is the sum of the individual variances.

Theorem 3.4 - *Chernoff Bounds*
Let X be a random variable whose moment-generating function $\mathbb{E}[e^{\theta X}]$ is finite $\forall \theta$. Then

$$\forall\ c \in \mathbb{R} \quad \mathbb{P}(X \geq c) \leq \inf_{\theta > 0} e^{-\theta c}\mathbb{E}(e^{\theta X}) \quad \text{and} \quad \mathbb{P}(X \leq c) \leq \inf_{\theta < 0} e^{-\theta c}\mathbb{E}(e^{\theta X})$$

Proof
Note that the events $X \geq c$ and $e^{\theta X} \geq e^{\theta c}$ are equivalent for all $\theta > 0$. The result follows by applying *Markov's Inequality* to $r^{\theta X}$ and taking the best bound over all possible θ.

$$\begin{aligned} \mathbb{P}(X \geq c) &= \mathbb{P}(e^{\theta X} \geq e^{\theta c}) \\ &\leq e^{-\theta c}\mathbb{E}(e^{\theta X}) \\ &\leq \inf_{\theta < 0} e^{-\theta c}\mathbb{E}(e^{\theta X}) \end{aligned}$$

Theorem 3.5 - *Chernoff Bounds for Sum of IIDs*
Let $X_1, \dots, X_n$ be IID random variables. Then $\forall\ c \in \mathbb{R}$

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^{n} X_i \geq nc\right) &\leq \inf_{\theta > 0} e^{-n\theta c}\left(\mathbb{E}\left[e^{\theta X}\right]\right)^n \\ \mathbb{P}\left(\sum_{i=1}^{n} X_i \leq nc\right) &\leq \inf_{\theta < 0} e^{-n\theta c}\left(\mathbb{E}\left[e^{\theta X}\right]\right)^n \end{aligned}$$

Theorem 3.6 - *Jensen's Inequality*
Let f be a *Convex Function* and X be a random variable. Then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

Theorem 3.7 - *Bound on Moment Generating Function*
Let X be a random variable taking values in $[0, 1]$ with finite expected value μ. Then we can bound the MGF of the centred random variable with

$$\forall\ \theta \in \mathbb{R} \quad \mathbb{E}\left[e^{\theta(X-\mu)}\right] \leq e^{\theta^2/8}$$

Proof (of weaker version)
Let X_1 be an independent copy of X, so both have mean μ. We can easily verify that $f(x) = e^{\theta x}$ is a convex function for all $\theta \in \mathbb{R}$. By *Jensen's Inequality* to $f(\cdot)$ and X_1

$$\mathbb{E}[e^{-\theta X_1}] \geq e^{-\theta\mathbb{E}[X_1]} = e^{-\theta\mu} \quad (1)$$

Consequently

$$\begin{aligned} & \mathbb{E}[e^{\theta(X-X_1)}] &=& \mathbb{E}[e^{\theta X}] \cdot \mathbb{E}[e^{-\theta X_1}] && \text{by independence} \\ & &\geq& \mathbb{E}[e^{\theta X}] \cdot e^{-\theta\mu} && \text{by (1)} \\ & &=& \mathbb{E}[e^{\theta(X-\mu)}] \\ \implies & \mathbb{E}[e^{\theta(X-X_1)}] &\geq& \mathbb{E}[e^{\theta(X-\mu)}] \end{aligned}$$

Since $X, X_1 \in [0,1]$ then $(X - X_1) \in [-1,1]$. As X, X_1 have the same distribution $\mathbb{E}(X - X_1) = 0$ and the distribution is symmetric around the mean.
Define random variable S which is independent of X, X_1 and takes values $\{-1, 1\}$, each with probability $p = \frac{1}{2}$. $S(X - X_1)$ has the same distribution as $(X - X_1)$ due to independence of S and symmetry of $(X - X_1)$. Hence

$$\begin{aligned} \mathbb{E}[e^{\theta(X-X_1)}] &= \mathbb{E}[e^{\theta S(X-X_1)}] && \text{by identical distribution} \\ &\leq \mathbb{E}[e^{\theta S}] && (2) \text{ since } (X - X_1) \in [-1,1] \\ &= \tfrac{1}{2}(e^{\theta} + e^{-\theta}) && \text{by def. of expectation} \\ \implies \mathbb{E}[e^{\theta(X-X_1)}] &\leq \tfrac{1}{2}(e^{\theta} + e^{-\theta}) \end{aligned}$$

Note that $f(x) = e^x + e^{-x}$ is increasing for $x \in (0, \infty)$; decreasing for $x \in (-\infty, 0)$; and symmetric around 0.
Using a *Taylor Series* we an observed that

$$\begin{aligned} \tfrac{1}{2}(e^{\theta} - e^{-\theta}) &= \sum_{n=0}^{\infty} \tfrac{\theta^{2n}}{(2n)!} && \text{by Taylor expansion of} e^x \\ &\leq \sum_{n=0}^{\infty} \frac{(\theta^2/2)^n}{n!} \\ &\overset{\text{def.}}{=} e^{\theta^2/2} \\ \implies \tfrac{1}{2}(e^{\theta} + e^{-\theta}) &\leq e^{\theta^2/2} \end{aligned}$$

Combining all these results we get

$$\begin{aligned} & \mathbb{E}[e^{\theta(X-\mu)}] \leq \mathbb{E}[e^{\theta(X-X_1)}] \leq \tfrac{1}{2}(e^{\theta} + e^{-\theta}) \leq e^{\theta^2/2} \\ \implies \quad & \mathbb{E}[e^{\theta(X-\mu)}] \leq e^{\theta^2/2} \end{aligned}$$

□

Theorem 3.8 - *Hoeffding's Theorem*
Let $X_1, \ldots, X_n$ be IID random variables taking values in $[0,1]$ and with finite expected value μ. Then

$$\forall\, t > 0 \quad \mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu) > nt\right) \leq e^{-2nt^2}$$

Proof
From *Chernoff's Bound* we have that

$$\forall\, \theta > 0 \quad \mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu) > nt\right) \leq e^{-\theta nt}\left(\mathbb{E}[e^{\theta(X-\mu)}]\right)^n$$

Using `Theorem 2.7` to bound the moment generating function, we get

$$\forall\, \theta > 0 \quad \mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu) > nt\right) \leq e^{-\theta nt} \cdot e^{n\frac{\theta^2}{8}} = e^{n\left(-\theta t + \frac{1}{8}\theta^2\right)}$$

Thus, by taking logs and rearranging, we get

$$\forall\ \theta > 0 \quad \frac{1}{n}\log \mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu) > nt\right) \leq -\theta t + \frac{\theta^2}{8}$$

We have that $-\theta t + \frac{\theta^2}{8}$ is minimised at $\theta = 4t$ which is positive if t is positive. Thus, by applying this bound and substituting $\theta = 4t$ we get

$$\forall\ \theta > 0 \quad \mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mu) > nt\right) \leq e^{n\left(-4t^2 + \frac{1}{8}(16t^2)\right)} = e^{n(-4t^2+2t^2)} = e^{-2nt^2}$$

□

3.2 Markov Processes

Definition 3.2 - *Markov Property*
A random process has the *Markov Property* if the conditional probability of a future state only depends on the current state.

$$\mathbb{P}(X_{t+1} = y | X_t = x_t, X_{t-1} = x_{t-1}) = \mathbb{P}(X_{t+1} = y | X_t = x_t)$$

A random process with the *Markov Property* is called a *Markov Process/Chain.*

Remark 3.2 - *On this course we only deal with discrete time markov chains*
Definition 3.3 - *Transience*
A state $x \in S$ is *Transient* if $\mathbb{P}(\exists\ t > 0 : X_t = x | X_0 = x) < 1$. The number of times the markov chain returns to a transient state is finite, with probability 1.

Definition 3.4 - *Recurrent*
A state $x \in S$ is *Recuurent* if $\mathbb{P}(\exists\ t > 0 : X_t = x | X_0 = x) = 1$. The number of times the markov chain returns to a recurrent state is infinite, with probability 1.

Every markov chain, with a finite state space S, has a recurrent communicating class.

Definition 3.5 - *Communication Class*
We say $y \in S$ is *Accessible* from $x \in S$ if $\exists\ t \geq 0$ st $[P^t]_{xy} > 0$.

We say x and y *communicate* (denoted xCy) if: x is *accessible* from y and y is *accessible* from x.

Communication is an *equivalence relation* on the state space S. Hence, *communication* partitions S into equivalence classes called *Communication Classes.* All elements of a *Communication Class* communicate with all other elements in the class, it is possible for elements to be accessible from another class but not for those elements to *communicate.*

If one state in a *Communicating Class* is *Transient/Recurrent* then all states are in that class.

If a *Markov Chain* has only one communicating class it is called *Irreducible.*

3.2.1 Discrete Time Markov Chains

Proposition 3.1 - *Characterising a Discrete Time Markov Process*

A *Discrete Time Markov Process* can be characterised by the set of all 1-step conditional probabilities

$$\mathbb{P}(X_{t+1} = y | X_t = x) \ \forall \ x, y \in S$$

A markov chain is *time-homogeneous* if the 1-step conditionals only depend on x, y and not on t ($\mathbb{P}(X_{t+1} = y|X_t = x) = \mathbb{P}(X_1 = t|X_0 = x)$). The 1-step conditional probabilities of a *time-homogeneous markov process* can be specified in an $|S| \times |S|$ matrix P where

$$p_{x,y} = \mathbb{P}(X_{t+1} = y|X_t = x)$$

P is a *Stochastic Matrix*.

Proposition 3.2 - *n-Step Transition Probabilities from* 1*-Step Transition Matrix*
Let P be the 1-step transition matrix for a *time-homogeneous*.

The 2-step transition probabilities (ie $\mathbb{P}(X_{t+2} = z|X_t = x)$) can be found as

$$\begin{aligned}
\mathbb{P}(X_{t+2} = z|X_t = x) &= \mathbb{P}(X_2 = z|X_0 = x) && \text{by time-homogeneity} \\
&= \textstyle\sum_{y\in S} \mathbb{P}(X_2 = z, X_1 = y|X_t = x) \\
&= \textstyle\sum_{y\in S} \mathbb{P}(X_1 = y|X_0 = x)\mathbb{P}(X_2 = z|X_1 = t, X_0 = x) \\
&= \textstyle\sum_{y\in S} \mathbb{P}(X_1 = y|X_0 = x)\mathbb{P}(X_2 = z|X_1 = y) \\
&= \textstyle\sum_{y\in S} p_{xy}p_{yz} \\
&\equiv [P^2]_{xz}
\end{aligned}$$

This can be generalise for the n-step transition probabilities with

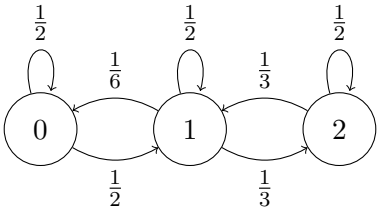
$$\mathbb{P}(X_{t+n} = z|X_t = x) = [P^n]_{xz}$$

Proposition 3.3 - *Any Joint Probability from* 1*-Step Transition Matrix*
For a time homogeneous markov process the joint distribution of any transition can be computed by considering the individual steps of the transition.

$$\mathbb{P}(X_{n_0} = x_0, X_{n_1} = x_1, X_{n_2} = x_2, \dots) = \mathbb{P}(X_{n_0} = x_0) \cdot [P^{n_0-n_1}]_{x_0x_1} \cdot [P^{n_1-n_2}]_{x_1x_2} \dots$$

Proposition 3.4 - *State Diagram Representation*
A graph/automata can be drawn to represent the transition probability matrix P. A node is assigned for each member of the state space S and an arrow is drawn between each pair of nodes (x, y) where $P_{xy} \neq 0$. Generally the value of P_{xy} is denoted on the arrow.

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$



Definition 3.6 - *Invariant Distribution*
Let $\boldsymbol{\mu}(t)$ denote the probability distribution of random variable X (i,e, $\mu_x(t) = \mathbb{P}(X_t = x)$). Then

$$\begin{aligned} & \boldsymbol{\mu}(t+1) &=& \mathbb{P}(X_{t+1} = y) = \textstyle\sum_{x\in S} \mathbb{P}(X_t = x, X_{t+1} = y) = \sum_{x\in S} \mu_x(t) p_{xy} = \boldsymbol{\mu}(t)P \\ \implies & \boldsymbol{\mu}(t+1) &=& \boldsymbol{\mu}(t)P \end{aligned}$$

A distribution $\boldsymbol{\pi}$ on the state space is called an *Invariant Distribution* if $\boldsymbol{\pi} = \boldsymbol{\pi}P$. If X_t has distribution $\boldsymbol{\pi}$ so will $X_{t+1}, \dots$. Every markov chain with a *finite* state space S has an *invariant distribution.* (Not necessarily true if S is infinite).

Proposition 3.5 - *Finding an Invariant Distribution*
If an *Invariant Distribution* it is easy to find by solving $\boldsymbol{\pi}P = \boldsymbol{\pi}$ and using normalising constant $\sum_{x\in S} \pi_x = 1$.

Remark 3.3 - *If a Markov chain is irreducible, its invariant distribution (if one exists) is unique*
If a *Markov Chain* is irreducible and has a finite state space, then it has a unique invariant distribution.

Example 3.1 - *Markov Chains*

- The *Asymmetric Simple Random Walk* on $\mathbb{Z}$ is $\underbrace{\text{irreducible}}_{\text{obvious}}$, $\underbrace{\text{transient}}_{\text{not obvious}}$ and has no invariant distribution.
- The *Symmetric Simple Random Walk* on $\mathbb{Z}$ is $\underbrace{\text{irreducible}}_{\text{obvious}}$, $\underbrace{\text{recurrent}}_{\text{not obvious}}$ and has no invariant distribution.

Theorem 3.9 - *Ergodic Theorem for Markov Chains*
Let $\{X_t\}_{t\in\mathbb{N}}$ be an irreducible markov chain on state space S (not necessarily finite) with unique invariant distribution $\boldsymbol{\pi}$. Then

$$\forall\ x \in S \quad \lim_{t\to\infty} \frac{1}{t} \sum_{s=1}^{t} \mathbb{1}(X_s = x) = \pi_x$$

i.e. The fraction of time spend in state $x \in S$ tends to π_x in the long run.

Definition 3.7 - *Period*
The *Period* of a state $x \in S$ is the greatest common divisor of all possible return times to x

$$\text{Period}(x) := \gcd\big(\{t > 0 : \mathbb{P}(X_t = x | X_0 = x) > 0\}\big)$$

A state $x \in S$ is *Aperiodic* if $\text{Period}(x) = 1$. An irreducible markov chain is *aperiodic* if all its states are aperiodic.

All states in a *communicating class* have the same period.

Proposition 3.6 - *Marginal Distribution of Irreducible, Aperiodic Markov Chain*
If an irreducible, aperiodic Markov Chain has an invariant distribution $\boldsymbol{\pi}$, then

$$\forall\ x \in S \quad \mu_x(t) \overset{t\to\infty}{\longrightarrow} \pi_x$$

Definition 3.8 - *Reversibility*
A markov chain $\{X_t\}_{t\in\mathbb{Z}}$ is *Reversible* if all joint distributions are the same forwards and backwards in time. (i.e. the distribution of the chain is the same if it was reversed).

An irreducible markov chain $\{X_t\}_{t\in\mathbb{Z}}$ with transition matrix P is reversible iff

$$\exists\ \boldsymbol{\pi} \quad \text{st} \quad \pi_x p_{xy} = \pi_y p_{yx}\ \forall\ x, y \in S$$

This is the *Local/Detailed Balance Equation*. Note that this is a system of $\binom{|S|}{2}$ equations which need to be consistent for reversibility to exist.

3.2.2 Continuous Time Markov Process

Definition 3.9 - *Continuous Time Markov Process*
A stochastic process $\{X_t\}_{t\in\mathbb{R}}$ is a *Continuous Time Markov Process* on state space s if

$$\forall\ s < t\ \&\ x, y \in S \quad \mathbb{P}(X_t = y | X_s = x, X_u, u \leq s) = \mathbb{P}(X_t = y | X_s = x)$$

ie future values only depend on the present value and not past.

If $\forall\ t, s, x, y\ \mathbb{P}(X_t = y | X_s = x)$ depends only on $x, y, t - s$ (observed values & change in time) then the process is *Time-Homogeneous*.

For *Time-Homogenous Markov Processes* we let $P(t)$ denote the stochastic matrix with the probability of each possible transition after t time $[P(t)]_{xy} = \mathbb{P}(X_t = y | X_0 = x)$.

Remark 3.4 - *A Time-Homogenous Markov Process is completely described by its initial condition and the family of transition probability matrices* $\{P(t) : t \geq 0$
This set of matrices $\{P(t) : t \geq 0$ is uncountably large.

Definition 3.10 - *Chapman-Kolmogorov Equations*
For a *Time-Homogeneous Markov Process* the family of stochastic matrices $\{P(t) : t \geq 0$ satisfy the following:

i). $P(0) = I$;

ii). $P(t + s) = P(t)P(s) = P(s)P(t)$

Hence

$$\frac{d}{dt}P(t) := \lim_{\delta\to 0} \frac{P(t+\delta) - P(t)}{\delta} = \lim_{\delta\to 0} \frac{\overbrace{P(t)P(\delta)}^{\text{ii)}} - P(t)}{\delta} = P(t) \lim_{\delta\to 0} \frac{(P(\delta) - I)}{\delta}$$

Suppose that $Q := \lim_{\delta\to 0} \frac{P(\delta) - P(0)}{\delta} = \lim_{\delta\to 0} \frac{P(\delta) - \overbrace{I}^{\text{i)}}}{\delta}$ exists.
Then $P(t)$ solve the following differential equations, known as the *Chapman-Kolmogorov Equations*

$$\frac{d}{dt}P(t) = \underbrace{P(t)Q}_{\text{forward eqn.}} = \underbrace{QP(t)}_{\text{backward eqn.}}$$

The solution to these equations is

$$P(t) = P(0)e^{Qt} = e^{Qt}P(0) = e^{Qt}I = e^{Qt}$$

N.B. Q is called the *Rate Matrix* or *Infinesimal Generator* of the markov process.

Proposition 3.7 - *Properties of the Rate Matrix, Q*
Let Q be the rate matrix of a *continuous-time markov process*. Q has the following properties

- If $n \neq y$ then $q_{xy} := \lim_{\delta \to 0} \frac{[P(\delta)]_{xy} - 0}{\delta} \geq 0$. (The off-diagonal elements are non-negative).
- $\forall\ x \in S$, $\sum_{y \in S} q_{xy} = \lim_{\delta \to 0} \frac{1-1}{\delta} = 0$. The rows of Q sum to 0.
- Thus, the diagonal entries q_{xx} are negative. (We denote $-q_{xx}$ by q_x)

Proposition 3.8 - *Interpreting the Rate Matrix Q*
Let Q be the rate matrix of a *continuous-time markov process*.

If the markov process enters state x at time t, it will remain in x for a random time which is distributed $\text{Exp}(q_x)$. (Note that $q_x := -q_x x$).

It the jumps to state y with probability $\frac{q_{xy}}{q_x}$, independent of the past.

Definition 3.11 - *Invariant Distributions*
Suppose a *Continuous-Time Markov Process* starts with distribution $\boldsymbol{\mu}(0)$ on state space S (i.e. $\mathbb{P}(X_0 = x) = [\boldsymbol{\mu}(0)]_x$). Then, the distribution of X_t is $\boldsymbol{\mu}(t) := \boldsymbol{\mu}(0)P(t) = \boldsymbol{\mu}(0) \underbrace{e^{Qt}}_{\text{CK Eqns}}$.

If there exists a distribution $\boldsymbol{\pi}$ on the state space S st $\forall\ t \geq 0\ \boldsymbol{\pi} = \boldsymbol{\pi}P(t) = \boldsymbol{\pi} \underbrace{e^{Qt}}_{\text{CK Eqns}}$, then $\boldsymbol{\pi}$ is an *Invariant Distribution*. This distribution is invariant wrt time.

If a markov process has a finite state space then it definitely has an invariant distribution.

Invariant Distributions are not guaranteed to be unique.

Proposition 3.9 - *Finding an Invariant Distribution*
Starting with $\boldsymbol{\pi} = \boldsymbol{\pi}e^{Qt}$ we find that differentiating wrt t and then evaluating at time $t = 0$ we get $\mathbf{0} = \boldsymbol{\pi}Q$ (The Global Balance Equations). This system of equations can be solved to find an *Invariant Distribution*.

A markov process is *reversible* iff there exists a distribution $\boldsymbol{\pi}$ on S which satisfies $\boldsymbol{\pi}_x q_{xy} = \boldsymbol{\pi}_y q_{yx}\ \forall\ x, y \in S$ (Local balance equations). Solving this system of equations will also find an *Invariant Distribution* but it is not guaranteed to have a solution.

Theorem 3.10 - *Ergodic Theorem*
Let $[X_t]_{t \in \mathbb{R}^+}$ is an *Irreducible Markov Process* on a state space S and has an invariant distribution $\boldsymbol{\pi}$. Then

$$\forall\ x \in S \quad \pi_x = \lim_{t \to \infty} \frac{1}{t} \int_0^t \mathbb{1}(X_s = x) ds$$

Moreover, for an arbitrary initial distribution $\boldsymbol{\mu}(0)$, $\boldsymbol{\mu}(t)$ converges to $\boldsymbol{\pi}$ pointwise (i.e. $\mu_x(t) \overset{t \to \infty}{\longrightarrow} \pi_x$)

3.2.3 Poisson Process

Definition 3.12 - *Counting Process*
A *Counting Process* is a stochastic process $\{N(t)\}_{t\in\mathbb{R}}$ st

i). $N(0) = 0$ and $N(t) \in \mathbb{Z}$ for all $t \geq 0$; and,

ii). $N(t)$ is a non-decreasing function of t.

Definition 3.13 - *Independent Increments*
A *Process* $\{N(t)\}_{t\in\mathbb{R}^+}$ is said to have *Independent Increments* if $\forall\ s \in (0,t), \quad (N(t) - N(s))$ is independent of $\{N(u) : u \in [0,s]\}$.

Definition 3.14 - *Poisson Process*
A *Poisson Process* is a *counting process* $\{N(t)\}_{t\in\mathbb{R}^+}$ which has independent increments and at least one of the following equivalence statements are true

- $\forall\ t \in [0,t] \quad (N(t) - N(s)) \sim \text{Po}(\lambda(t-s))$.
- $\mathbb{P}(N(t+\delta) - N(t) = 1) = \lambda\delta + o(\delta)$ and
  $\mathbb{P}(N(t+\delta) - N(t) = 0) = 1 - \lambda\delta + o(\delta)$ and
  $\mathbb{P}(N(t+\delta) - N(t) \geq 2) = o(\delta)$.
- The times between successive increments of the process $N(\cdot)$ are iid $\text{Exp}(\lambda)$ random variables.

The parameter $\lambda \in \mathbb{R}^{>0}$ is called the *rate* of the poisson process. *Poisson Processes* are continuous time markov chains.

Example 3.2 - *Poisson Process*
Counting the number of cars which have passed a given point over time.

Proposition 3.10 - *Properties of Poisson Processes*
Define $\{N(t)\}_{t\in\mathbb{R}^+}$ to be a *Poisson Process* with rate λ. Then the following properties hold

i). The counting process $\{N(\beta t)\}_{t\in\mathbb{R}^+}$, with $\beta > 0$, is a Poisson Process with rate $\beta\lambda$.

ii). If $\{N_1(t)\}_{t\in\mathbb{R}^+}$ and $\{N_2(t)\}_{t\in\mathbb{R}^+}$ are independent poisson processes with rates λ_1 and λ_2, respectively, then $\{N(t) := N_1(t) + N_2(t)\}_{t\in\mathbb{R}^+}$ is a poisson process with rate $\lambda := \lambda_1 + \lambda_2$.

iii). Let $X_1, X_2, \ldots$ be a sequence of iid $\text{Bern}(p)$ random vairables, independent of $N(\cdot)$. Define $N_1(t) := \sum_{i=1}^{N(t)} X_i$ and $N_2(t) := \sum_{i=1}^{N(t)} (1 - X_i)$ (These are called *Bernoulli Thinnings*). These assign increments in $N(\cdot)$ randomly to either N_1 or N_2 (with probability p).
Then, $N_1(\cdot)$ and $N_2(\cdot)$ are independent poisson processes with rates λp and $\lambda(1-p)$, respectively.

3.3 Transformation of Random Variables

Example 3.3 - *Discrete Case*
Consider rolling a fair die where $\Omega := \{1,2,3,4,5,6\}$ and $\forall\ \omega \in \Omega \quad \mathbb{P}(\omega) = \frac{1}{6}$.
Let $X(\omega) = \omega\ \forall\ \omega \in \Omega$ so the pmf of X is given by

$$p_X(i) = \frac{1}{6}\ \forall\ i \in \{1, \ldots, 6\}$$

Consider $Y := X^2$. The pmf of Y is straightforward to work out as each value of X maps to a unique value of Y

$$P_Y(i) = \frac{1}{6} \text{ for } \sqrt{i} \in \{1, \dots, 6\}$$

Consider $Z := (X-2)^2$. This is not quite so simple as multiple values of X map to the same value of Z.

$$p_Z(1) = \frac{2}{6}; \quad p_Z(i) = \frac{1}{6} \text{ for } i \in \{0, 4, 9, 16\}$$

Example 3.4 - *Continuous Case*
Let $X \sim \text{Uniform}[0,1]$ and define $Y := 2X$. We have

$$f_X(x) = \begin{cases} 1 & x \in [0,1] \\ 0 & \text{otherwise} \end{cases} \qquad F_X(x) = \begin{cases} 0 & x < 0 \\ x & x \in [0,1] \\ 1 & x > 1 \end{cases}$$

Now consider Y the cdf is

$$F_Y(y) := \mathbb{P}(Y \leq y) = \mathbb{P}(2X \leq y) = \mathbb{P}\left(X \leq \frac{y}{2}\right) = F_X\left(\frac{y}{2}\right)$$

We then obtain the pdf for Y by differentiation and the chain rule.

$$f_Y(y) = F_Y'(y) = \underbrace{\frac{1}{2} F_X'\left(\frac{y}{2}\right)}_{\text{chain rule}} = \frac{1}{2} f_X\left(\frac{y}{2}\right) = \begin{cases} \frac{1}{2} & \frac{y}{2} \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

Proposition 3.11 - *Increasing Functions*
Let X be a random variable and $Y := g(X)$ where $g : \mathbb{R} \to \mathbb{R}$ is a strictly increasing function. Then, g is invertible on its range (denoted g^{-1}). Thus the cdfs for X and Y are related as

$$F_Y(y) = \underbrace{\mathbb{P}(g(X) \leq y) = \mathbb{P}(X \leq g^{-1}(y))}_{\text{as } g \text{ is increasing}} = F_X(g^{-1}(y))$$

Differentiating this gives us the pdfs

$$f_Y(y) = f_X\left(g^{-1}(y)\right) \frac{dg^{-1}(y)}{dy} = \left.\frac{f_X(x)}{g'(x)}\right|_{x = g^{-1}(y)}$$

The probability mass of X in the interval $(x, x+dx)$ gets mapped to the interval $\big(g(x), g(x+dx)\big)$. By *Taylor Expansion* of g we get $g(x+dx) \simeq g(x) + g'(x)dx$

Proposition 3.12 - *General Mappings - Single Random Variable*
Let X be a (discrete) random variable and define $Y := g(x)$ for any $g : \mathbb{R} \to \mathbb{R}$ which is differentiable.

There is a contribution $\frac{f_X(x)}{|g'(x)|}$ from each x such that $g(x) = y$. Where $g'(x)$ is positive or negative does not matter, as it only determines where the pre-image of $(y, y+dy)$ is of the form $(x, x+dx)$ or $(x-dx, x)$. Only the relative widths of the intervals matters for the contribution.

By summing the contributions from all solutions of $g(x) = y$ we get

$$p_Y(y) = \sum_{x : g(x) = y} \frac{f_X(x)}{|g'(x)|}$$

This formula is valid so long as the set $\{x : g(x) = y\}$ is countable. If it is not countable, then Y is a continuous RV (or is a mixed random variable)

Proposition 3.13 - *General Mappings - Random Vectors*
Consider the random vector $\mathbf{X} := (X_1, \dots, X_d)$ with joint density $f_{\mathbf{X}}$ and define $\mathbf{Y} := g(\mathbf{X})$ for any differentiable $g : \mathbb{R}^d \to \mathbb{R}^d$.

Let $J_g(\mathbf{x})$ denote the *Jacobian* of g at $\mathbf{x}$. Then

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{\mathbf{x} \in \mathbb{R}^d : g(\mathbf{x}) = \mathbf{y}} \frac{f_{\mathbf{X}}(\mathbf{x})}{|\det(J_g(\mathbf{x}))}$$

Let $\mathbf{x}$ solve $g(\mathbf{x}) = \mathbf{y}$. In a neighbourhood of $\mathbf{x}$, g is approximately a linear function. By *Taylor Expansion*

$$g(\mathbf{x}') \simeq f(\mathbf{x}) + J_g(\mathbf{x})(\mathbf{x}' - \mathbf{x}) = \mathbf{y} + J_g(\mathbf{x})(\mathbf{x}' - \mathbf{x})$$

for $\mathbf{x}'$ in a small enough neighbourhood of $\mathbf{x}$.

0 Reference

Definition 0.1 - *Stochastic Matrix*
A matrix is called a *Stochastic matrix* if:

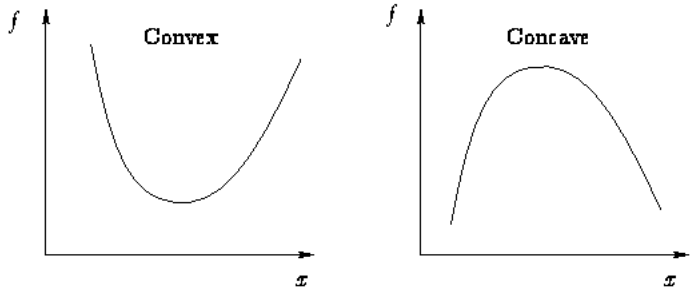
i). All elements are non-negative.

ii). All rows sum to 1.

Definition 0.2 - *Convex Function*
A function $f : \mathbb{R} \to (\mathbb{R} \cup \{+\infty\})$ is *Convex* if, $\forall\ x, y \in \mathbb{R},\ \alpha \in [0, 1]$, we have

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$$

A smooth function f is convex iff f is twice differentiable and $f''(x) \geq 0\ \forall\ x \in \mathbb{R}$.
Visually, a function is convex if you can draw a line between any two points on the function and the function lies below the line.



Definition 0.3 - *Equivalence Relation*
A relation is an *Equivalence Relation* if it is

i). Reflexive: $i \leftrightarrow i$.

ii). Symmetric: If $i \to j$ then $j \to i$.

iii). Transitive: If $i \to j$ and $j \to k$ then $i \to k$.

Definition 0.4 - *Simple Random Walk*
A *Simple Random Walk* is a random walk which moves only one step at a time. (i.e. $X_{t+1} = X_t \pm 1$). A probability p is defined for $\mathbb{P}(X_{t+1} = X_t + 1)$, this means $1 - p$ is the probability of stepping in the other direction. A *Simple Random Walk* is *Assymetric* if $p \neq p - 1$.

Definition 0.5 - *Matrix Exponential, e^X*
Let X be a matrix then we define the *Matrix Exponential* as $e^X := I + X + \frac{X^2}{2!} + \dots$

Theorem 0.1 - *Pinsker's Inequality*
For two distributions Bern(p) and Bern(q)

$$K(q; p) \geq 2(q - p)^2$$

Definition 0.6 - *Jacobian Matrix*
The *Jacobian Matrix* is the first-order partial-derivatives of a multidimensional function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ wrt each parameter.

$$J_f := \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

0.1 Notation

Multi-Armed Bandit Problem

$p_{x,y}$	$\mathbb{P}(X_{t+1} = x \mid X_t = y)$ for a *time homogenous markov process.*
$I(t)$	Arm played in round t.
$N_i(t)$	$\sum_{s=1}^{t} \{I(s) = i\}$ Number of times arm i was played in first t rounds.
$S_i(t)$	$\sum_{s=1}^{t} X_i(s)\{I(s) = i\}$ Total reward from arm i in first t rounds.
$\hat{\mu}_{i,N_i(t)}$	$\frac{S_i(t)}{N_i(t)}$ sample mean reward from arm i in first t rounds.
Δ_i	$\mu^* - \mu_i$ the arm gaps from a K-armed bandit.
$X_i(t)$	RV modelling the result if arm i was played on round t.

Markov Decision Problems

T	Time-Horizon.
S	State-Space.
A	Action-Space.
$A(s)$	Available actions when in system state s.
Y_t	Random variable modelling action taken by the agent at epoch t.
$r_t(s, a)$	The value of the reward received at epoch t if the system is in state s and the agent takes actio
$p_t(s'\|s, a)$	The probability that the system is in state s' at time $t+1$ given at time t the system was in sta
$d_t(\cdot),\ q_t(\cdot)$	The decision function at epoch t.
π	A decision policy.
$HR(T)$	The set of policies on time-horizon T in which the decision rules are HRs.
$MD(T)$	The set of policies on time-horizon T in which the decision rules are MDs.
$\mathbb{P}^\pi,\ \mathbb{E}^\pi$	Probability and expectation given policy π is being used.
T^π	The expected reward for a given policy.
$u^*(s)$	The terminal reward for a given MDP.
$v_N^*(s)$	The optimal value function for the expected discount reward over N epochs.

General

$z_{m:n}$	The m^{th} to the n^{th} elements of a sequence $\{z_t\}$. height

0.1.1 Asymptotic Notation

Definition 0.7 - *Oh Notation*
Let $f, g : \mathbb{R}^+ \to \mathbb{R}$.

We say $f = o(g)$ (little oh of g) at 0 if $\frac{f(x)}{g(x)} \xrightarrow{x \to 0} 0$.

We say $f = O(g)$ (big oh of g) at 0 if $\exists\ c > 0$ st $|f(x)| \leq c|g(x)|$ in a neighbourhood of 0.

$f = o(g)$ at infinity and $f = O(g)$ at infinity are defined analogously.

Definition 0.8 - *Omega Notation*

Let $f, g : \mathbb{R}^+ \to \mathbb{R}$.
We say $g = \omega(g)$ if $o(f)$ and we say $f = \Omega(g)$ if $g = O(f)$.

Example 0.1 - *Oh & Omega Notation*
Define $f(x) = x,\ g(x) = \sin(x)$ and $h(x) = x^2$.
Then, $g = O(f)$ at 0 and $g = o(f)$ at infinity. $h = o(f)$ at 0 and $h = \omega(f)$ at infinity.

0.2 Irreducible Markov Chains

Definition 0.9 - *Markov Chain*
A *Stochastic Process* $\{X_t\}_{t\geq}$ taking values in S is a *Markov Chain* if it has the *Markov Property*

$$\mathbb{P}(X_{t+1} = s_{t+1}|X_t = s_t, \dots, X_0 = s_0) = \mathbb{P}(X_{t+1} = s_{t=1}|X_t = s_t)\ \forall\ t \in T$$

A *Markov Chain* $\{X_t\}_{t\geq 0}$ is *Homogeneous* if the transitions probabilities are the same in all time-periods

$$\mathbb{P}(X_{t+1} = s'|X_t = s) = \mathbb{P}(X_1 = s'|X_0 = s)\ \forall\ t \in T$$

The *Transition Kernel* of a *Homogeneous Markov Chain* $\{X_t\}_{t\geq 0}$ is the transition probabilities

$$p(s'|s) := \mathbb{P}(X_1 = s'|X_0 = s)$$

A *Homogeneous Markov Chain* $\{X_t\}_{t\geq 0}$ is *Irreducible* if $\forall\ s, s' \in S$ there exists $t \geq 1$ st

$$p^t(s'|s) > 0$$

Definition 0.10 - *Invariant Probability Mass Function*
A function $\mu(s)$ is an *Invariant Probability Mass Function* of a *Homogeneous Markov Chain* $\{X_t\}_{t\geq 0}$ if

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

Theorem 0.2 - *Invariant PMF exists for all Irreducible Markov Chain*
Let $\{X_t\}_{t\geq 0}$ be an *Irreducible Markov Chain.* Then the follow hold

i). $\{X_t\}_{t\geq 0}$ has a unique invariant probability mass function $\mu(s)$.

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

Theorem 0.3 - *Weak Law of Large Numbers*
Let $\{X_t\}_{t\geq 0}$ be an *Irreducible Markov Chain* with invariant pmf $\mu(\cdot)$ and let $f : S \to \mathbb{R}$ by any function. The *Weak Law of Large Numbers* state

$$\lim_{N\to\infty} \mathbb{E}\left[\frac{1}{N}\sum_{t=0}^{N-1} f(X_t)\right] = \sum_{s\in S} f(s)\mu(s)$$

Note that the RHS is the expected value of $f(s)$ wrt $\mu(s)$.

Theorem 0.4 - *Poisson Equation*
Let $\{X_t\}_{t\geq 0}$ be an *Irreducible Markov Chain* with *transition kernel* $p(s'|s)$ and *Invariant PMF* (s). Let $f : S \to \mathbb{R}$ be any function and $\bar{f}$ the expected value of f wrt μ

$$\bar{f} := \sum_{s\in S} f(s)\mu(s)$$

Then the following hold

i). There exists a function $\check{f} : S \to \mathbb{R}$ st

$$f(s) - \bar{f} = \check{f}(s) - \sum_{s'\in S} \check{f}(s')p(s'|s) \quad \forall\ s \in S$$

ii). Further, if there exists another function $\check{f}' : S \to \mathbb{R}$ st

$$f(s) - \bar{f} = \check{f}(s) - \sum_{s'\in S} \check{f}'(s')p(s'|s) \quad \forall\ s \in S$$

then $\exists\ c \in \mathbb{R}$ st

$$\check{f}'(s) = \check{f}(s) + c \quad \forall\ s \in S$$

This $\check{f}$ is known as the *Poisson Equation* for $\{X_t\}_{t\geq 0}$ and $f(s)$.

Theorem 0.5 - *Laurent Expansion of Resolvent*
Let $\{X_t\}_{t\geq 0}$ be an *Irreducible Markov Chain* with *transition kernel* $p(s'|s)$ and *Invariant PMF* (s). Let $f : S \to \mathbb{R}$ be any function, $\bar{f}$ the expected value of f wrt μ and $\check{f}(s)$ be a solution to the *Poisson Equation*

$$\begin{aligned} \bar{f} &:= \sum_{s\in S} f(s)\mu(s) \\ f(s) - \bar{f} &= \check{f} - \sum_{s'\in S} \check{f}(s')p(s'|s) \end{aligned}$$

$$\bar{f} := \sum_{s\in S} f(s)\mu(s)$$

Consider the following function

$$\begin{aligned} \check{f}'(s) &:= \check{f}(s) - \sum_{s'\in S} \check{f}(s') \\ f_\alpha(s) &:= \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \alpha^t f(X_t)\big|X_0 = s\right] \quad \alpha \in (0,1) \\ \tilde{f}_\alpha &:= f_\alpha(s) - \left(\frac{\bar{f}}{1-\alpha} + \check{f}')s_\prime\right) \end{aligned}$$

$\tilde{f}_\alpha(s)$ is known as the *Residual* in the *Laurent Expansion of* $f_\alpha(s)$ Then

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

Remark 0.1 - *Poisson Equation*
The function $\check{f}'(s)'$ is a solution to the *Poisson Equation* associated with *Markov Chain* $\{X_t\}_{t\geq 0}$ and function $f(s)$

$$f(s) - \bar{f} = \check{f}'(s) - \sum_{s'\in S} \check{f}'(s')p(s'|s)$$

Remark 0.2 - *Expectation of $\check{f}''(s)$ wrt $\mu(s)$*
The expectation of function $\check{f}(s)$ wrt $\mu(s)$ is zero

$$\sum_{s\in S} \check{f}'(s)\mu(s) = 0$$

Remark 0.3 - *$f_\alpha(s)$ and $f\tilde{f}_\alpha(s)$*
$f_\alpha(s)$ is known as the *alpha-resolvent* associated with *Markov Chain* $\{X_t\}_{t\geq 0}$ and function $f(s)$.
The defining equation for $\tilde{f}_\alpha$ can be rewritten as

$$f_\alpha(s) = \frac{\bar{f}}{(1-\alpha)} + \check{f}(s) + \tilde{f}_\alpha(s)$$

This equation is known as the *Laurent Expansion* of $f_\alpha(s)$ at $\alpha = 1$.
By the *Laurent Expansion of Resolvent*,

$$f_\alpha(s) \approx \frac{\bar{f}}{1-\alpha} + \check{f}(s) \quad \text{when } \alpha \approx 1$$