Stochastic Optimisation - Reviewed Notes

Dom Hutchinson

December 21, 2020

Contents

1 Probability

1.1 Probabilistic Inequalities

Theorem 1.1 - *Markov's Inequality*
Let X be a non-negative random variable.
Markov's Inequality states

$$\forall\ c > 0 \quad \mathbb{P}(X \geq c) \leq \frac{\mathbb{E}(X)}{c}$$

Proof 1.1 - *Markov's Inequality*
Let X be a non-negative random variable and fix $c > 0$.
Consider partitioning the expectation of X around the value c.

$$\mathbb{E}(X) = \mathbb{P}(X < c) \cdot \mathbb{E}[X|X < c] + \mathbb{P}(X \geq c) \cdot \mathbb{E}[X|X \geq c]$$

Note that $\mathbb{E}[X|X < c] > 0$ since X is non-negative and $\mathbb{E}[X|X \geq c] \geq c$ since it only considers the cases where $X \geq c$. Thus

$$\mathbb{E}(X) \geq \mathbb{P}(X < c) \cdot 0 + \mathbb{P}(X \geq c) \cdot c$$

Rearranging we get the result of the theorem.

$$\mathbb{P}(X \geq c) \leq \frac{\mathbb{E}(X)}{c}$$

$\square$

Theorem 1.2 - *Chebyshev's Inequality*
Let X be a random-variable with finite mean μ and variance σ^2.
Chebyshev's Inequality states

$$\forall\ c > 0 \quad \mathbb{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$$

Further for $X_1, \dots, X_n$ IID RVs with finite mean μ and variance σ^2. We have

$$\forall\ c > 0 \quad \mathbb{P}\left(\left|\left(\sum_{i=1}^n X_i\right) - n\mu\right| \geq nc\right) \leq \frac{\sigma^2}{nc^2}$$

Proof 1.2 - *Chebyshev's Inequality - Single Random Variable*
Let X be a random-variable with finite mean μ and variance σ^2, and fix $c > 0$.
Define random variable $Y := (X-\mu)^2$, noting that Y is non-negative and $\mathbb{E}[Y] = \mathbb{E}\left[(X - \mu)^2\right] =: \text{Var}(X) = \sigma^2$.
By *Markov's Inequality* we have that

$$\mathbb{P}(Y \geq c^2) \leq \frac{\mathbb{E}(Y)}{c^2} = \frac{\text{Var}(X)}{c^2}$$

Note that the event $\{Y \geq c^2\} = \{(X - \mu)^2 \geq c^2\}$ is equivalent to the event $\{|X - \mu| \geq c\}$ since $c > 0$.
Substituting this result into the above expression gives the result of the theorem.

$$\mathbb{P}(|X - \mu| \geq c) \leq \frac{\mathbb{E}(Y)}{c^2} = \frac{\text{Var}(X)}{c^2}$$

□

Proof 1.3 - *Chebyshev's Inequality - Sum of IID Random Variables*
Let $X_1, \dots, X_n$ be IID RVs with finite mean μ and variance σ^2.
Define random variable $Y := \sum_{i=1}^n X_i$. Note that

$$\begin{array}{rcl} \mathbb{E}(Y) & = & \mathbb{E}\left(\sum_{i=1}^n X_i\right) \\ & = & \sum_{i=1}^n \mathbb{E}(X_i) \\ & = & n\mu \end{array} \qquad \begin{array}{rcll} \mathrm{Var}(Y) & = & \mathrm{Var}\left(\sum_{i=1}^n X_i\right) & \\ & = & \sum_{i=1}^n \mathrm{Var}(X_i) & \text{by independence} \\ & = & n\sigma^2 & \text{by identical distribution} \end{array}$$

By applying *Chebyshev's Inequality* to Y bounded by $(nc)^2$, we get

$$\begin{array}{rrcl} & \mathbb{P}\left(|Y - \mathbb{E}(Y)| \geq c\right) & \leq & \frac{\mathrm{Var}(Y)}{(nc)^2} \\ \implies & \mathbb{P}\left(\left|\left(\sum_{i=1}^n X_i\right) - n\mu\right| \geq c\right) & \leq & \frac{n\sigma^2}{(nc)^2} = \frac{\sigma^2}{nc^2} \end{array}$$

The result of the theorem for the sum of IID RVs. □

Theorem 1.3 - *Chernoff Bounds*
Let X be a random variable whose moment-generating function $\mathbb{E}[e^{\theta X}]$ is finite $\forall\ \theta$.
Chernoff Bounds state

$$\forall\ c \in \mathbb{R} \quad \mathbb{P}(X \geq c) \leq \inf_{\theta > 0} e^{-\theta c}\mathbb{E}[e^{\theta X}] \quad \text{and} \quad \mathbb{P}(X \leq c) \leq \inf_{\theta < 0} e^{-\theta c}\mathbb{E}[e^{\theta X})$$

Further for $X_1, \dots, X_n$ IID RVs with finite moment-generating functions $\forall\ \theta$. We have

$$\forall\ c \in \mathbb{R} \quad \mathbb{P}\left(\sum_{i=1}^n X_i \geq nc\right) \leq \inf_{\theta > 0} e^{-n\theta c}\mathbb{E}[e^{\theta X}]^n \quad \text{and} \quad \mathbb{P}\left(\sum_{i=1}^n X_i \leq c\right) \leq \inf_{\theta < 0} e^{-n\theta c}\mathbb{E}[e^{\theta X}]^n$$

Proof 1.4 - *Chernoff Bounds - Single Random Variable*
Let X be a random variable whose moment-generating function $\mathbb{E}[e^{\theta X}]$ is finite $\forall\ \theta$.
Note that $\forall\ \theta > 0$ the events $\left\{X \geq c\right\}$ and $\left\{e^{\theta X} \geq e^{\theta c}\right\}$ are equivalent. Giving

$$\mathbb{P}(X \geq c) = \mathbb{P}(e^{\theta X} \geq e^{\theta c})$$

By *Markov's Inequality* we have that

$$\mathbb{P}(e^{\theta X} \geq e^{\theta c}) \leq \frac{\mathbb{E}[e^{\theta X}]}{e^{\theta c}} = e^{-\theta c}\mathbb{E}[e^{\theta X}]$$

As θ is any positive real and we want the tightest bound, we take the infinum of the bound wrt θ. Giving

$$\begin{array}{rrcl} & \mathbb{P}(e^{\theta X} \geq e^{\theta c}) & \leq & \inf_{\theta > 0} e^{-\theta c}\mathbb{E}[e^{\theta X}] \\ \implies & \mathbb{P}(X \geq c) & \leq & \inf_{\theta > 0} e^{-\theta c}\mathbb{E}[e^{\theta X}] \end{array}$$

The result of the theorem. □
An equivalent proof is used for the event $\{X \leq c\}$.

Proof 1.5 - *Chernoff Bounds - Sum of IID Random Variables*

Let $X_1,\dots,X_n$ be IID RVs with finite moment-generating functions $\forall\ \theta$.
Note that $\forall\ \theta>0$ the events $\{\sum_{i=1}^n X_i \geq nc\}$ and $\left\{e^{\theta\sum_{i=1}^n X_i}\geq e^{nc\theta}\right\}$ are equivalent. Giving

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq nc\right)=\mathbb{P}\left(e^{\theta\sum_{i=1}^n X_i}\geq e^{nc\theta}\right)$$

By *Markov's Inequality* we have that

$$\mathbb{P}\left(e^{\theta\sum_{i=1}^n X_i}\geq e^{nc\theta}\right)\leq\frac{\mathbb{E}\left[e^{\theta\sum_{i=1}^n X_i}\right]}{e^{nc\theta}}=e^{-nc\theta}\mathbb{E}[e^{\theta X}]^n$$

As θ is any positive real and we want the tightest bound, we take the infinum of the bound wrt θ. Giving

$$\begin{array}{rrcl} & \mathbb{P}\left(e^{\theta\sum_{i=1}^n X_i}\geq e^{nc\theta}\right) & \leq & \inf_{\theta>0}\frac{\mathbb{E}\left[e^{\theta\sum_{i=1}^n X_i}\right]}{e^{nc\theta}}=e^{-nc\theta}\mathbb{E}[e^{\theta X}]^n \\ \implies & \mathbb{P}\left(\sum_{i=1}^n X\geq c\right) & \leq & \inf_{\theta>0}e^{-nc\theta}\mathbb{E}[e^{\theta X}]^n \end{array}$$

The result of the theorem. $\square$

An equivalent proof is used for the event $\left\{\sum_{i=1}^n X_i\leq c\right\}$.

Theorem 1.4 - *Jensen's Inequality*
Let f be a convex function and X be a random variable.
Jensen's Inequality states that

$$\mathbb{E}[f(X)]\geq f(E[X])$$

Theorem 1.5 - *Hoeffding's Inequality*
Let $X_1,\dots,X_n$ be IID random variables taking values in $[0,1]$ and finite mean μ.
Hoeffding's Inequality states

$$\begin{array}{rrrcl} & \forall\ c>0 & \mathbb{P}\left(\sum_{i=1}^n (X_i-\mu)>nc\right) & \leq & e^{-2nc^2} \\ \iff & \forall\ c>0 & \mathbb{P}\left(\hat{\mu}-\mu>c\right) & \leq & e^{-2nc^2} \end{array}$$

The value of the bound is the same for inequalities in the other direction

$$\begin{array}{rrrcl} & \forall\ c>0 & \mathbb{P}\left(\sum_{i=1}^n (X_i-\mu)<nc\right) & \leq & e^{-2nc^2} \\ \iff & \forall\ c>0 & \mathbb{P}\left(\hat{\mu}-\mu<c\right) & \leq & e^{-2nc^2} \end{array}$$

the n used in the expression involving sample mean is the size of the sample used to calculate the sample mean.

Theorem 1.6 - *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 1.6 - *Hoeffding's Theorem*
Let $X_1,\dots,X_n$ be IID random variables taking values in $[0,1]$ and finite mean μ. Fix $c>0$.
Chernoff Bounds on $\sum_{i=1}^n(X_i-\mu)$ bounded below by nc state

$$\forall\ \theta>0\quad \mathbb{P}\left(\sum_{i=1}^n(X_i-\mu)>nc\right)\leq e^{-\theta nc}\left(\mathbb{E}[e^{\theta(X-\mu)}]\right)^n$$

By `Theorem 1.6`

$$\forall\ \theta\in\mathbb{R}\quad \mathbb{E}[e^{\theta(X-\mu)}]^n\leq\left[e^{\frac{\theta^2}{8}}\right]^n=e^{n\frac{\theta^2}{8}}$$

Incorporating this bound into the above expression we get

$$\forall\ \theta>0\quad \mathbb{P}\left(\sum_{i=1}^n(X_i-\mu)>nt\right)\leq e^{-\theta nt}\cdot e^{n\frac{\theta^2}{8}}=e^{n\left(-\theta t+\frac{\theta^2}{8}\right)}$$

To get the tightest upper-bound we want to find the θ which minimises the expression on the RHS. This is equivalent to minimising the expression $-\theta t+\frac{\theta^2}{8}$ wrt θ.

$$\begin{array}{rrcl} & \frac{\partial}{\partial\theta}\left(-\theta t+\frac{\theta^2}{8}\right) &=& -t+\frac{\theta}{4}\\ & \frac{\partial^2}{\partial\theta^2}\left(-\theta t+\frac{\theta^2}{8}\right) &=& \frac{1}{4}>0\\ \text{Setting} & \frac{\partial}{\partial\theta}\left(-\theta t+\frac{\theta^2}{8}\right) &=& 0\\ \implies & -t+\frac{\theta}{4} &=& 0\\ \implies & \theta &=& 4t \end{array}$$

As the second derivative is strictly positive, the expression above is minimise for $\theta=4t$.
By substituting this value of θ into the expression we get

$$\forall\ \theta>0\quad \mathbb{P}\left(\sum_{i=1}^n(X_i-\mu)>nt\right)\leq e^{n\left(-4t\cdot t+\frac{(4t)^2}{8}\right)}=e^{n\left(-4t^2+\frac{16t^2}{8}\right)}=e^{-2nt^2}$$

The result of the theorem. $\square$

Theorem 1.7 - *Pinsker's Theorem*
For any distributions $p,q\in[0,1]$

$$K(q;p)\geq 2(p-q)^2$$

1.1.1 Special Cases

Theorem 1.8 - *Chernoff Bound - Binomial Random Variable*
Let $X\sim\text{Bin}(n,\alpha)$ with $n\in\mathbb{N},\ \alpha\in(0,1)$.

$$\begin{array}{l}\forall\ \beta>\alpha\quad \mathbb{P}(X\geq\beta n)\leq e^{-nK(\beta;\alpha)}\\ \forall\ \beta<\alpha\quad \mathbb{P}(X\leq\beta n)\leq e^{-nK(\beta;\alpha)}\end{array}$$

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 $x = 0$. Note that $K(\cdot;\cdot)$ is the *Kullback-Leibler Divergence* for two *Binomial Random Variables.*

Theorem 1.9 - *Heoffding's Inequality - Binomial Random Variables*
Let $X \sim \text{Bin}(n,p)$ with $n \in \mathbb{N}$ and $p \in [0,1]$

$$\begin{array}{ll}\forall\ \varepsilon > 0 & \mathbb{P}(X \leq (p-\varepsilon)n) \leq \exp\left(-2n\varepsilon^2\right)\\ \forall\ \varepsilon > 0 & \mathbb{P}(X \geq (p+\varepsilon)n) \leq \exp\left(-2n\varepsilon^2\right)\end{array}$$

1.2 Transformation of Random Variables

Theorem 1.10 - *Monotone Functions*
Let X be a random variable and g be a differentiable and strictly monotone function.
Define $Y := g(X)$. Then

$$f_Y(y) = f_X(g^{-1}(y))\frac{1}{|g'(g^{-1}(y))}$$

Theorem 1.11 - *Non-Monotone Functions*
Let X be a random variable and g be a differentiable and non-monotone function.
Define $Y := g(X)$.
Since g is not monotone, then for a fixed y there are multiple x which solve $y = g(x)$. (Think trig functions). In this case we have to sum the probability contribution from each of these xs

$$f_Y(y) = \sum_{x\in\{x:g(x)=y\}} f_X(x)\frac{1}{|g'(x)|}$$

Theorem 1.12 - *Joint Distributions*
Let $\mathbf{X} := \{X_1,\dots,X_n\}$ be random variables on the same sample space and $g : \mathbb{R}^n \to \mathbb{R}^n$ be a differentiable function.
Define $\mathbf{Y} = (Y_1,\dots,Y_n) := g(X_1,\dots,X_n)$. Then

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{\mathbf{x}\in\{\mathbf{x}:g(\mathbf{x})=\mathbf{y}\}} f_{\mathbf{X}}(\mathbf{x})\frac{1}{|\det(J_g(\mathbf{x}))}$$

where $\det(J_g(\mathbf{x}))$ denotes the determinant of the Jacobian of g wrt $\mathbf{x}$ (See `Definition 0.1`).

2 The Multi-Armed Bandit Problem

2.1 The Problem

Definition

Definition 2.1 - *Multi-Armed Bandit Problem*
In the *Multi-Armed Bandit Problem* an agent is given the choice of K actions, with each action giving a different reward modelled by an unknown random variable X_i. The agent is allowed to

play a single at a time and the agent's aim is to maximise some measure of long-run reward (ie find the action with the greatest mean reward), typically whilst minimising loss during the learning stage.

Example 2.1 - *Motivating Example for Multi-Armed Bandit Problem*
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?

One approach is to assign a subset of the patients randomly to treatments, and then assign the rest to the best treatment. This leads to the questions around what is sufficient evidence for one treatment to be the best? And, how likely are you to choose a sub-optimal treatment?

Strategies

Definition 2.2 - *Strategy, $I(\cdot)$*
The agent's *Strategy* I is a function which determines which action the agent shall make at each time step. The only information a *Strategy* can utilise is which arms were played in the past and what reward was received each time.

As it is assumed that this knowledge is utilised, we simplify the notation to only take time as a parameter.

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

Definition 2.3 - *Policy*
A *Policy* $f(t)$ is a family of *Strategies* and the *Strategy* used at each time-step is chosen randomly from these *Strategies*, typically uniformly at random.

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

Measures of Success

Definition 2.4 - *Long-Run Average Reward Criterion, X_**
The *Long-Run Average Reward* X_* is the average reward a chosen *Strategy* $I(\cdot)$ produces. A *Strategy* is said to be *Optimal* if $X_* = \max_{k\in[1,K]} \mathbb{E}[X_k]$

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

The *Infinum* is taken as there is no guarantee the limit exists.

Definition 2.5 - *Regret $\mathcal{R}_n$*
Regret R_n is the total reward lost during the first n time-steps by using a strategy $I(\cdot)$, compared to if the optimal arm had been played every time.

$$\mathcal{R}_n := n\mu^* - \sum_{i=1}^{n} \mathbb{E}[X_{I(t)}(t)] \quad \text{where} \quad \mu^* := \max_{k\in[1,K]} \mathbb{E}[X_k]$$

Remark 2.1 - *Learning Regret*
Regret only involves expectations and thus can be learnt from observations.

Definition 2.6 - *Strongly Consistent*
A strategy for the multi-armed bandit problem is said to be *Strongly Consistent* if its regret satisfies $\mathcal{R}_T = o(T^\alpha)\ \forall\ \alpha > 0$. (i.e. its regret grows slower than any fractional power of T).

Theorem 2.1 - *Lai & Robbins Theorem*
Consider a K-armed bandit with Bernoulli arms.
Lai & Robbins Theorem states that, for any *Strongly Consistent* strategy, the number of times that a sub-optimal arm i is played up to time T ($N_i(T)$) satisfies

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_i(T)]}{\ln(T)} \geq \frac{1}{K(\mu_i;\mu^*)} \quad \text{where } \mu* := \max_{i=1}^{K} \mu_i$$

where $K(q;p)$ is the *KL-Divergence* of a Bern(q) distribution wrt a Bern(p) distribution (See `Theorem 1.7`).

Mathematical Setup

Proposition 2.1 - *Mathematical Setup for Multi-Armed Bandit Problem*
Consider a *Multi-Armed Bandit* with K arms and let $X_i(t)$ model the reward obtained by playing arm i at time set t, with $i \in [1, K]$ and $t \in \mathbb{N}$. We make two assumptions about the reward distributions

i). The reward distributions $X_1(\cdot), \cdots, X_K(\cdots)$ are mutually independent.

ii). The reward of each distribution is independent of time. ie $X_i(t)$ and $X_i(t+m)$ are independent $\forall\ i, t, m$

The agent is tasked with finding a *Strategy* which minimises *Regret* R_n over a time horizon T.

$$\text{Find } I(\cdot) \text{ which minimises } \mathcal{R}_T := T\mu^* - \sum_{t=1}^{T} \mathbb{E}[X_{I(t)}(t)] \text{ where } \mu^* := \max_{k\in[1,K]} \mathbb{E}[X_k]$$

There are strategies where *Regret* over time T grows sub-linearly (ie $\frac{1}{T}R_t \overset{T\to\infty}{\longrightarrow} 0$).

2.2 Naïve Approaches

Proposition 2.2 - *Naïve Heuristic - Single Test, Bernoulli*
Let X_1, X_2 be *Bernoulli* reward distributions for a 2-armed bandit and defined $\mu_i := \mathbb{E}[X_i]$. Assume WLOG that $\mu_1 > \mu_2$ consider the following heuristic

Play each arm once. Whichever arms returns the greatest reward, play it for all reamining rounds.

Since the reward distributions are *Beroulii* random variables, this heuristic picks the sub-optimal arm with probability $\mu_2(1-\mu_1)$. If the sub-optimal arm is chosen, then it is played a total of $T-1$ times over time T. Giving the following lower-bound on the regret $\mathcal{R}_T$

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

This regret grows linearly in T.

Proposition 2.3 - *Better Heuristic - N Tests, Bernoulli*
Let X_1, X_2 be *Bernoulli* reward distributions for a 2-armed bandit and defined $\mu_i := \mathbb{E}[X_i]$. Assume WLOG that $\mu_1 > \mu_2$ consider 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.

As X_1, X_2 are Bernoulli RVs, $S_i(n) \sim \text{Bin}(n, \mu_i)$ and S_1, S_2 are independent.
For $\beta \in (\mu_2, \mu_1)$

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

by `Theorem 1.7` 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 2.4 - *Optimal N for `Proposition 2.3`*
For the situation described in `Proposition 2.3` we want to find N which minimises regret, given a total time horizon of T.

With this heuristic it is guaranteed that $R_N = N(\mu_1-\mu_2)$ due to the learning phase. Regret only increases after the learning phase if the sub-optimal arm is chosen. This gives the following expression for regret over time horizon T.

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

$$\begin{array}{rcl}
\mathcal{R}_T &=& \underbrace{(T-2N)(\mu_1-\mu_2)\mathbb{P}\big(S_1(N)<S_2(N)\big)}_{\text{if wrong decision made}} + \underbrace{N(\mu_1-\mu_2)}_{\text{guaranteed regret}} \\
&\leq& (T-2N)(\mu_1-\mu_2)\underbrace{e^{-NJ(\mu_1,\mu_2)}}_{\texttt{Theorem 1.7}} + N(\mu_1-\mu_2) \\
&\simeq& (\mu_1-\mu_2)(N+Te^{-NJ(\mu_1,\mu_2)}) \quad \text{as } -2Ne^{-NJ(\mu_1,\mu_2)} \text{ is very small}
\end{array}$$

We want to minimise this expression wrt N

$$\begin{array}{rrcl}
& \frac{\partial}{\partial N}(\mu_1-\mu_2)(N+Te^{-NJ(\mu_1,\mu_2)}) &=& -TJ(\mu_1,\mu_2)e^{-NJ(\mu_1,\mu_2)} \\
& \frac{\partial^2}{\partial N^2}(\mu_1-\mu_2)(N+Te^{-NJ(\mu_1,\mu_2)}) &=& TJ(\mu_1,\mu_2)^2e^{-NJ(\mu_1,\mu_2)} > 0 \\
\text{Setting} & -TJ(\mu_1,\mu_2)e^{-NJ(\mu_1,\mu_2)} &=& 0 \\
\implies & TJ(\mu_1,\mu_2)e^{-NJ(\mu_1,\mu_2)} &=& 0 \\
\implies & \ln[TJ(\mu_1,\mu_2)] - NJ(\mu_1,\mu_2) &=& 0 \\
\implies & N &=& \frac{\ln[TJ(\mu_1,\mu_2)]}{J(\mu_1,\mu_2)} \\
\implies & N &=& \frac{\ln[T]}{J(\mu_1,\mu_2)} + O(1) \text{ as } \ln[J(\mu_1,\mu_2)] \text{ is very small}
\end{array}$$

As the second derivative is strictly positive, $N := \frac{\ln[T]}{J(\mu_1,\mu_2)} + O(1)$ is the optimal N used during training and gives the following expression for regret

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

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

2.3 UCB Algorithm

Remark 2.2 - *UCB Algorithm*
The *Upper Confidence Bound Algorithm* is a *frequentist* algorithm for solving the *Multi-Armed Bandit Problem* for a bandit with *Bernoulli*.

The premise of the algorithm is to play whichever arm has the greatest upper-bound on a confidence interval for the true value of the mean μ_i/

Remark 2.3 - *Motivation*
The heuristics in `Proposition 2.2, 2.3` treat the sample mean as if it is the true mean (*Certainty Equivalence*), which it is not. The *UCB Algorithm* considers a $1-\delta$ confidence interval for the value of μ_i.

Noting that *Hoeffding's Inequality* states

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

We can use this to find an upper-bound of a $1-\delta$ confidence interval for the value of μ_i. This can be done by setting $\delta = e^{-2nx^2}$, rearranging to get $x = \sqrt{\frac{1}{2n} \ln\left(\frac{1}{\delta}\right)}$, and substituting this value of x into *Hoeffding's Inequality* to get an upper bound on μ_i

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

Here δ is a value we choose from $[0,1]$ depending upon the setting.

2.3.1 Algorithm

Definition 2.7 - *UCB(α) Algorithm*
Consider the set up of a K-Armed bandit in `Proposition 2.1` with Bernoulli Arms and let $\alpha > 0$.
The *UCB Algorithm* over time horizon T is defined as

i). In rounds $t \in [1, K]$:

  (a) Play the t^{th} arm.

ii). Calculate the $UCB(\alpha, i)$ value for each arm.

$$UCB(\alpha, i) := \hat{\mu}_{i,N_i(t)} + \sqrt{\frac{1}{2N_i(t)} \alpha \ln(t)}$$

iii). In rounds $t \in (K, T]$:

(a) Play the arm i which maximises $UCB(\alpha, i)$.

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

(b) Update the $UCB(\alpha, i)$ value for the played arm.

2.3.2 Analysis

Remark 2.4 - *UCB is Strongly Consistent*
The *UCB(α)* algorithm is strongly consistent for all $\alpha > 1$ as its regret grows logarithmically with T.

Theorem 2.2 - *Upper Bound on Regret*
Consider the set up of a K-Armed bandit in `Proposition 2.1` with Bernoulli Arms, let $\alpha > 0$ and assume WLOG that arm 1 is the optimal arm (ie $\mu_1 > \mu_i\ \forall\ i \in [2, K]$).

If the *UCB(α)* algorithm is used, with $\alpha > 1$, then the regret in the first T rounds is bounded above by

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

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

This theorem is problem in `Proof 2.3`*.*

Remark 2.5 - *Setting α*
The result in `Theorem 2.1` grows fast if α is taken to be large. However, if α is small then the constant term dominates for smaller values of T. Thus we typically choose $\alpha = 2$.

Theorem 2.3 - *When a sub-optimal arm is played*
Consider the set up of a K-Armed bandit in `Proposition 2.1` with Bernoulli Arms, let $\alpha > 0$ and assume WLOG that arm 1 is the optimal arm (ie $\mu_1 > \mu_i\ \forall\ i \in [2, K]$).

Consider applying *UCB(α)* to this bandit and under what circumstances a sub-optimal arm is played in steps $t \geq K$ (ie $I(t) = i \neq 1$ for some $t > K$). One of the following statements is true:

i). The sample mean reward from the optimal arm is much smaller than the true mean.

$$\hat{\mu}_{1,N_1(s)} \leq \mu_1 - \sqrt{\frac{\alpha \ln(s)}{2N_1(s)}}$$

ii). The sample mean reward on arm i is much larger than its true mean.

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

iii). Arm i has been played very few times meaning its the confidence interval on its true mean μ_i is wide.

$$N_i(s) < \frac{2\alpha \ln(s)}{\Delta_i^2}$$

Proof 2.1 - *Theorem 2.2*
This is a proof by contradiction.
Consider the set up of a K-Armed bandit in `Proposition 2.1` with Bernoulli Arms, let $\alpha > 0$ and assume WLOG that arm 1 is the optimal arm (ie $\mu_1 > \mu_i \ \forall\ i \in [2, K]$).

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

$$\begin{array}{rcll} \underbrace{\hat{\mu}_{1,N_1(s)} + \sqrt{\frac{\alpha \ln(s)}{2N_1(s)}}}_{UCB(\alpha,1)} & > & \mu_1 & \text{by not i)} \\ & = & \mu_i + \Delta_j & \text{by def. of } \Delta_i \\ & \geq & \mu_i + \sqrt{\frac{2\alpha \ln(s)}{N_i(s)}} & \text{by not iii)} \\ & \geq & \hat{\mu}_{i,N_i(s)} - \sqrt{\frac{\alpha \ln(s)}{2N_i(s)}} + \sqrt{\frac{2\alpha \ln(s)}{N_i(s)}} & \text{by not ii)} \\ & \geq & \hat{\mu}_{i,N_i(s)} + \left(\sqrt{2} - \frac{1}{\sqrt{2}}\right)\sqrt{\frac{\alpha \ln(s)}{N_i(s)}} & \text{by collecting terms} \\ & = & \underbrace{\hat{\mu}_{i,N_i(s)} + \sqrt{\frac{\alpha \ln(s)}{2N_i(s)}}}_{UCB(\alpha,i)} & \end{array}$$

But, this implies that the $UCB(\alpha, 1) > UCB(\alpha, i)$ at the end of round s. Hence arm i would not be played in time slot $s + 1$. $\square$

Theorem 2.4 - *Counting Lemma*
Let $\{I(t)\}_{t\in\mathbb{N}}$ be a $\{0, 1\}$-valued sequence and $N_i(t) := \sum_{s=1}^t \mathbb{1}I(s) = i$. Then

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

with an empty sum defined to be zero.

Note that $\big\{(N(s-1) \geq u)\ \&\ (I(s) = i)\big\}$ is the event where: arm i has been played at least u times so far <u>and</u> is played this turn.

Proof 2.2 - *Theorem 2.3*
Fix $t, u \in \mathbb{N}$. There are two cases

Case 1 $N_i(t) \leq u$. (ie Have not reached u yet). The result holds trivially here.

Case 2 $\exists\ s \in [1, t]$ st $N_i(s) > u$. (ie 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

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

The last step holds $u+1\leq s^*$ and thus the sum is done over more terms in the final expression than the one before. □

Proof 2.3 - *Upper Bound on Regret*
Consider the set up of a K-Armed bandit in `Proposition 2.1` with Bernoulli Arms, let $\alpha>0$ and assume WLOG that arm 1 is the optimal arm (ie $\mu_1>\mu_i\ \forall\ i\in[2,K]$).

Fix $t\in\mathbb{N}$ and define $u_{t,i}:=\left\lceil\frac{2\alpha\ln(t)}{\Delta_i^2}\right\rceil$. By `Theorem 2.3` we have that

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

Note that both sides involve random variables. By taking expectations of both sides we get

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

By `Theorem 2.2` and the definition of $u_{t,i}$, if $I(s+1)=i$ *and* $N_j(s)\geq u$ (ie `Theorem 2.2 iii)` does not hold) *then*

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

Thus

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

Consider trying to bound the two probabilities

$$\begin{aligned}
\mathbb{P}\left(\hat{\mu}_{i,N_i(s)}>\mu_i-\sqrt{\frac{\alpha\ln(s)}{2N_i(s)}}\right) &= \mathbb{P}\left(\hat{\mu}_{i,N_i(s)}-\mu_i>\sqrt{\frac{\alpha\ln(s)}{2N_i(s)}}\right) \\
&\leq e^{-2N_i(s)\cdot\frac{\alpha\ln(s)}{2N_i(s)}} \quad\text{by Hoeffding's Inequality} \\
&= e^{-\alpha\ln(s)} \\
&= s^{-\alpha}
\end{aligned}$$

The same bound can be applied to the other probability. Substituting these bounds into the previous expression gives

$$\begin{aligned}
\mathbb{E}[N_i(t)] &\leq u_{t,i} + \sum_{s=u}^{t-1} 2s^{-\alpha} \\
&\leq u_{t,i} + \int_{u-1}^{\infty} 2s^{-\alpha} ds \quad \text{assumption } \alpha > 1 \text{ required here} \\
&= u_{t,i} + \frac{2(u-1)^{-(\alpha-1)}}{\alpha - 1} \\
&\leq u_{t,i} + \frac{2}{\alpha - 1} \quad \text{since } u \geq 2 \implies (u-1)^{-(\alpha-1)} \leq 1 \\
&= \left\lceil \frac{2\alpha \ln(t)}{\Delta_i^2} \right\rceil + \frac{2}{\alpha - 1} \quad \text{by def. of } u_{t,i} \\
&\leq \frac{2\alpha \ln(t)}{\Delta_i^2} + 1 + \frac{2}{\alpha - 1} \quad \text{by def. of ceil} \\
&= \frac{2\alpha \ln(t)}{\Delta_i^2} + \frac{\alpha + 1}{\alpha - 1}
\end{aligned}$$

Due to the generality of i, this result holds $\forall\ i \in [2, K]$. Hence the total regret up to time T is bounded by

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

The result of the theorem. □

2.3.3 Can we Improve?

Remark 2.6 - *The regret for UCB is almost optimal.*
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.

Proposition 2.5 - *Lower Bound on Regret*
To show the regret of $UCB(\alpha)$ is almost optimal, we derive a lower bound for the regret of any strongly consistent strategy for the multi-armed bandit problem

$$\begin{aligned}
\liminf_{T\to\infty} \frac{\mathcal{R}_T}{\ln(T)} &= \liminf_{T\to\infty} \frac{1}{\ln(T)} \sum_{i\in\{i:\mu_i<\mu^*\}} \Delta_i \mathbb{E}[N_i(T)] \quad \text{by def } \mathcal{R}_T \\
&= \sum_{i\in\{i:\mu_i<\mu^*\}} \Delta_i \left[\liminf_{T\to\infty} \frac{\mathbb{E}[N_i(T)]}{\ln(T)} \right] \\
&\geq \sum_{i\in\{i:\mu_i<\mu^*\}} \frac{\Delta_i}{K(\mu_i;\mu^*)} \quad \text{by \textit{Lai \& Robbins Theorem}}
\end{aligned}$$

Proposition 2.6 - *Upper Bound on Regret from UCB*
To show the regret of $UCB(\alpha)$ is almost optimal, we derive an upper bound for the regret of

any strongly consistent strategy for the multi-armed bandit problem

$$\begin{aligned}
\limsup_{T\to\infty}\frac{\mathcal{R}_T}{\ln(T)} &\leq \limsup_{T\to\infty}\frac{1}{\ln(T)}\sum_{i=2}^{K}\left(\frac{2\alpha\ln(T)}{\Delta_i}+\Delta_i\frac{\alpha+1}{\alpha-1}\right) \quad \text{by } \texttt{Theorem 2.2}\\
&= \limsup_{T\to\infty}\sum_{i=2}^{K}\left(\frac{2\alpha}{\Delta_i}+\frac{\Delta_i}{\ln(T)}\cdot\frac{\alpha+1}{(\alpha-1)}\right)\\
&= \limsup_{T\to\infty}\sum_{i=2}^{K}\frac{2\alpha}{\Delta_i}\\
&\leq \sum_{i=2}^{K}\frac{2}{\Delta_i}\ \text{TODO check this}
\end{aligned}$$

Proposition 2.7 - *Comparing UCB & Minimum Lower Bound*
Consider `Proposition 2.5` and *Pinsker's Inequality*, when equality is reached

$$\liminf_{T\to\infty}\frac{\mathcal{R}_T}{\ln(T)} \geq \sum_{i\in\{i:\mu_i<\mu^*\}}\frac{\Delta_i}{K(\mu_i;\mu^*)} \geq \frac{1}{2\Delta_i}$$

Comparing this to the result in `Proposition 2.6`, we get that the regret $UCB(\alpha)$ is at most $\frac{\sum 2/\Delta_i}{\sum 1/2\Delta_i} = 4$ times worse that the absolute best.

2.4 Thompson Sampling

Remark 2.7 - *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.

2.4.1 Algorithm

Definition 2.8 - *Thompson Sampling Algorithm - Bernoulli Arms*
Consider the st up of a K-Armed bandit in `Proposition 2.1` with Bernoulli Arms.
The *Thompson Sampling Algorithm* over time horizon T is defined as

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

ii). For $t \in [1,T]$:

   (a) For $i \in [1,K]$ sample $\hat{\mu}_i(t)$ from the priors of each arm, breaking ties arbitrarily.
   (b) Play the arm with the greatest sample value.

$$I(t) = \text{argmax}_{i\in[1,K]}\hat{\mu}_i(t)$$

   (c) Use the observed reward to calculate the posterior of the played arm:
      - Given the arm for this prior at time t was $\text{Beta}(\alpha,\beta)$.
      - If (Reward Observed): Set posterior to $\text{Beta}(\alpha+1,\beta)$.
      - Else: Set posterior to $\text{Beta}(\alpha,\beta+1)$.

(d) For all un-played arms, assign their prior as their posterior.

(e) For the next round, use the posteriors from this round as the priors.

Remark 2.8 - *Choosing Priors for Thompson Sampling Algorithm*
In the *Thompson Sampling Algorithm* we choose priors which are *conjugate* with the distribution of the arms of the bandit so the priors and posteriors are from the same family.

For the *Multi-Armed Bandit Problem* we are only interested in estimated the mean reward of a random variable. Here I list some sets of conjugate priors which can be used in *Thompson Sampling* the means of specific distributions. See `Section 0.4` for a list of conjugate priors and their proofs.

2.4.2 Genie Analysis

Remark 2.9 - *Genie*
Analysing *Thompson Sampling* is hard as it is difficult to account for the scenario where there is an initial run of bad luck on the optimal arm.

In this section I analyse a simpler version of the Thompson Sampling algorithm for a 2-armed bandit. Consider the following scenario

> The value of μ_1 is known, but the value of μ_2 is unknown. Further, it is unknown whether μ_1 or μ_2 is greater (ie it is not known which is the optimal arm). We only define a prior & posterior for μ_2 and we play arm 2 if the value $\theta_2(t)$ sampled from its prior is greater than the true value of μ_1.

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

Theorem 2.5 - *Times Sub-Optimal arm is played*
Suppose WLOG that arm two is the suboptimal arm (ie $\mu_1 \geq \mu_2$) and consider a time horizon $T \in \mathbb{N}$. Define $L := \left\lceil \frac{2\ln(T)}{\Delta^2} \right\rceil$ & $\tau := \inf\{t \in [1, T] : N_2(t) \geq L\}$ (The round in which arm 2 is played for the L^{th} time). The probability arm two is played in any given round after round τ is bounded as

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

Futher, we can bound the expected number of times for arm two to be played after round τ

$$\begin{array}{rcl} \mathbb{E}[\#\text{ plays of arm two after round }\tau] & = & \underbrace{(T-\tau)}_{\#\ Rounds} \cdot\mathbb{P}(\theta_2(t) \geq \mu_1) \\ & \leq & (T-\tau)\frac{2}{T} \\ & \leq & 2 \end{array}$$

Proof 2.4 - *Theorem 2.5*
Consider a time horizon $T \in \mathbb{N}$ and define the quantities $L := \left\lceil \frac{2\ln(T)}{\Delta^2} \right\rceil$ & $\tau := \inf\{t \in [1, T] : N_2(t) \geq L\}$.
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 in round t is greater than μ_1 (ie arm two is played in round t). B_t is the event the average observed rewards from arm 2 up to round t is closer to μ_2 than μ_1. We can bound $\mathbb{P}(A_t)$ as follows

$$\begin{array}{rcll}\mathbb{P}(A_t) & = & \mathbb{P}(A_t \cap B_t) + \mathbb{P}(A_t \cap B_t^c) & \\ & = & \mathbb{P}(A_t|B_t)\mathbb{P}(B_t) + \mathbb{P}(A_t|B_t^c)\mathbb{P}(B_t^c) & \\ & \leq & \mathbb{P}(A_t|B_t) + \mathbb{P}(B_t^c) & (1)\end{array}$$

The inequality occurs since $\mathbb{P}(X) \geq \mathbb{P}(X)\mathbb{P}(Y)$ for all events X, Y.

We shall derive bounds, which are independent of the which round it is, for the two RH terms in the final expression separately. First I bound $\mathbb{P}(B_t^c)$.

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

$$\begin{array}{rcll}\mathbb{P}(B_t^c) & \equiv & \mathbb{P}\left(\dfrac{S_2(t)}{N_2(t)} > \mu_2 + \dfrac{\Delta}{2}\right) & \\ & \equiv & \mathbb{P}\left(\hat{\mu}_2(t) > \mu_2 + \dfrac{\Delta}{2}\right) & \\ & \leq & \exp\left(-2N_t\dfrac{\Delta^2}{4}\right) & \text{by Hoeffding's Ineq.} \\ & \leq & \exp\left(-L\dfrac{\Delta^2}{2}\right) & \text{since } N_2(t) \geq L \\ & \leq & \exp\left(-\dfrac{2\ln(T)}{\Delta^2}\cdot\dfrac{\Delta^2}{2}\right) = e^{-\ln(T)} & \text{by def. } L \\ & = & \dfrac{1}{T} & (2)\end{array}$$

Now I bound $\mathbb{P}(A_t|B_t)$. Let $\theta_2(t+1)$ is the value sampled from the posterior distribution of μ_2 after t rounds, thus, by `Proof 0.2`, it has the following distribution

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

Hence, by `Theorem 0.3`, the following events are equivalent

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

By applying the result in `Theorem 1.9`, for Hoeffding's Inequality on a binomial random variable, we can derive an explicit upper-bound on the probability of the RH event occurring.

$$\begin{array}{lrcl} & \mathbb{P}\left(\text{Bin}\left(N_2(t)+1, \mu_1\right) \leq S_2(t)\right) & \leq & \exp\left(-2(N_2(t)+1)\varepsilon^2\right) \text{ by } \texttt{Theorem 1.9} \\ \text{where} & (N_2(t)+1)(\mu_1 - \varepsilon) & = & S_2(t) \\ \implies & \varepsilon & = & \mu_1 - \dfrac{S_2(t)}{N_2(t)+1} \text{ since } N_2(t), S_2(t) \in \mathbb{N} \\ & & \geq & \mu_1 - \dfrac{S_2(t)}{N_2(t)} \text{ noting } \mu_1 < \dfrac{S_2(t)}{N_2(t)} \\ \implies & \exp(-\varepsilon^2) & \leq & \exp\left(-\left(\mu_1 - \dfrac{S_2(t)}{N_2(t)}\right)^2\right)\end{array}$$

Note that $\left(\mu_1 - \frac{S_2(t)}{N_2(t)}\right) \in [0,1]$ by definition of the terms and $\forall\, x \in [0,1]$, $\left(e^{-x}\right)^n \geq \left(e^{-x}\right)^{n+1}$. Using these results we derive an upper-bound on the binomial random variable and the equivalent event A_{t+1}.

$$\begin{array}{lrcl} & \mathbb{P}\big(\text{Bin}(N_2(t)+1, \mu_1) \leq S_2(t)\big) & \leq & \exp\left(-2N_2(t)\left(\mu_1 - \dfrac{S_2(t)}{N_2(t)}\right)^2\right) \\ \implies & \mathbb{P}(A_{t+1}|S_2(t), N_2(t)) & \leq & \exp\left(-2N_2(t)\left(\mu_1 - \dfrac{S_2(t)}{N_2(t)}\right)^2\right)\end{array}$$

Consider the following restatement of event B_t

$$\begin{aligned} & \left\{\frac{S_2(t)}{N_2(t)} \leq \mu_2 + \frac{\Delta}{2}\right\} \\ \iff & \left\{\frac{S_2(t)}{N_2(t)} \leq \mu_1 - \frac{\Delta}{2}\right\} \text{ by def. } \Delta \\ \iff & \left\{\frac{\Delta}{2} \leq \mu_1 - \frac{S_2(t)}{N_2(t)}\right\} \end{aligned}$$

Hence, we can state a bound for A_t given B_t and $N_2(t)$

$$\mathbb{P}\big(A_t|B_t, N_2(t)\big) \leq \exp\left(-2N_2(t)\left(\frac{\Delta}{2}\right)^2\right) = \exp\left(-2N_2(t)\frac{\Delta^2}{4}\right)$$

By the definition of τ, $\forall\ t \geq \tau,\ N_2(t) \geq L$. Hence we can derive a bound for A_t given B_t which is independent of $N_2(t)$

$$\begin{aligned} \forall\ t \geq \tau \quad \mathbb{P}(A_t|B_t) &\leq \exp\left(-2N_2(t)\frac{\Delta^2}{4}\right) \\ &\leq \exp\left(-L\frac{\Delta^2}{2}\right) \\ &= \exp\left(-\frac{2\ln(T)}{\Delta^2}\cdot\frac{\Delta^2}{2}\right) \text{ by def. } L \\ &\leq \exp(-\ln T) \\ &= \frac{1}{T} \end{aligned} \tag{3}$$

By substituting the bounds (2) and (3) into expression (1) we get the following bound for event A_t

$$\forall\ t \geq \tau \quad \mathbb{P}(A_t) \leq \mathbb{P}(A_t|B_t) + \mathbb{P}(B_t^c) \leq \frac{1}{T} + \frac{1}{T} = \frac{2}{T}$$

This is the stated result of `Theorem 2.5` □

Proposition 2.8 - *Bound of Regret*
Using `Theorem 2.5` we can bound the regret of Genie-Thompson Sampling 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.

2.4.3 Analysis

Remark 2.10 - *Analysis of Thompson Sampling is Hard*
Analysing *Thompson Sampling* is hard as it is difficult 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 worry about plays of the sub-optimal arm when they are played too often. However, in this scenario the posterior for the sub-optimal arm will be concentrated around the true parameter value and thus the samples arm truer representations.

Theorem 2.6 - *Upper Bound on Regret*

Consider a two-armed bandit with Bernoulli arms.
The regret of *Thompson Sampling* over time horizon T is bounded as

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

where c is an arbitrary constant which is independent of T.

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

Theorem 2.7 - *Number of times wrong arm is played*
Consider the set up of a 2-Armed bandit in `Proposition 2.1` with Bernoulli Arms and assume WLOG that arm 1 is the optimal arm (ie $\mu_1 > \mu_2$).

Consider using *Thompson Sampling* over time horizon T. Define $L = \left\lceil \frac{24\ln(T)}{\Delta^2} \right\rceil$ and $\tau = \inf\{t \in [0,T] : N_2(t) \geq L\}$ (The time at which arm 2 is played for the L^{th} time).
Then

$$\text{For } t \in [\tau, T] \quad \mathbb{P}\left(\theta_2(t) \geq \mu_2 + \frac{\Delta}{2}\right) \leq \frac{2}{T^3}$$

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

Proof 2.5 - *Theorem 2.7*
Consider using *Thompson Sampling* over time horizon T. Define $L = \left\lceil \frac{24\ln(T)}{\Delta^2} \right\rceil$ and $\tau = \inf\{t \in [0,T] : N_2(t) \geq L\}$ (The time at which arm 2 is played for the L^{th} time).
By the definition of τ, if $t \geq \tau$ then $N_2(t) \geq L$. Thus

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

the last step occurs because [1]

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

We now bound both terms in (1) seperately.

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* and noting that $\mathbb{E}\left(\frac{S_2(t)}{N_2(t)}\right) = \mu_2$, we have

$$\mathbb{P}\left(\frac{S_2(t)}{N_2(t)} > \mu_2 + \frac{\Delta}{4} \middle| N_2(t)\right) \leq \exp\left(-2N_2(t)\left(\frac{\Delta}{4}\right)^2\right) = \exp\left(-N_2(t)\frac{\Delta^2}{8}\right)$$

[1]For all random variables X, Y $\mathbb{P}(X|Y) \geq \mathbb{P}(X,Y)$ and $\mathbb{P}(X) \geq \mathbb{P}(X,Y)$

As we have assumed that $N_2(t) \geq L \geq \frac{1}{\Delta^2}(24\ln(T))$ meaning $-N_2(t) \leq \frac{1}{\Delta^2}(24\ln(T))$. Thus

$$\begin{aligned} -N_2(t)\frac{\Delta^2}{8} &\geq -\frac{24}{8}\ln(T) \\ &= -3\ln(T) \\ \implies \mathbb{P}\left(\frac{S_2(t)}{N_2(t)} > \mu_2 + \frac{\Delta}{4}\right) &\leq \exp\left(-3\ln(T)\right) \\ &= \frac{1}{T^3} \qquad (2) \end{aligned}$$

Next, we note that conditional on $S_2(t)$ and $N_2(t)$, by `Proof 0.2` the distribution of $\theta_2(t)$ is $\text{Beta}\big(\underbrace{S_2(t)+1}_{\alpha}, \underbrace{N_2(t)-S_2(t)+1}_{\beta}\big)$. Consequently, by `Proof 0.3`, we have that

$$\mathbb{P}\Big(\theta_2(t) \geq \underbrace{\mu_2 + \frac{\Delta}{2}}_{p}\Big) = \mathbb{P}\Big(\text{Bin}\Big(\underbrace{N_2(t)+1}_{\alpha+\beta-1}, \underbrace{\mu_2 + \frac{\Delta}{2}}_{p}\Big) \leq \underbrace{S_2(t)}_{\alpha-1}\Big)$$

By applying the result in `Theorem 1.9`, for Hoeffding's Inequality on a binomial random variable, we can derive an explicit upper-bound on the probability.

$$\begin{aligned} \mathbb{P}\left(\text{Bin}\left(N_2(t)+1, \mu_2 + \frac{\Delta}{2}\right) \leq S_2(t)\right) &\leq \exp\left(-2(N_2(t)+1)\varepsilon^2\right) \text{ by } \texttt{Theorem 1.9} \\ \text{where} \qquad (N_2(t)+1)\left(\mu_2 + \frac{\Delta}{2} - \varepsilon\right) &= S_2(t) \\ \implies \qquad \mu_2 + \frac{\Delta}{2} - \varepsilon &= \frac{S_2(t)}{N_2(t)+1} \\ \implies \qquad \varepsilon &= \mu_2 + \frac{\Delta}{2} - \frac{S_2(t)}{N_2(t)+1} \\ &\leq \mu_2 + \frac{\Delta}{2} - \left(\mu_2 + \frac{\Delta}{4}\right) \quad \text{assuming } \frac{S_2(t)}{N_2(t)} \leq \mu_2 + \frac{\Delta}{4} \\ &= \frac{\Delta}{4} \\ \implies \qquad \exp(-\varepsilon^2) &\leq \exp\left(-\left(\frac{\Delta}{4}\right)^2\right) = \exp\left(-\frac{\Delta^2}{16}\right) \end{aligned}$$

This gives us the following bound

$$\mathbb{P}\left(\text{Bin}\left(N_2(t)+1, \mu_2 + \frac{\Delta}{2}\right) \leq S_2(t) \middle| \frac{S_2(t)}{N_2(t)} \leq \mu_2 + \frac{\Delta}{4}\right) \leq \exp\left(-2(N_2(t)+1)\frac{\Delta^2}{16}\right)$$

Substituting this result into the original expression involving the binomial we get

$$\begin{aligned} \mathbb{P}\left(\theta_2(t) \geq \mu_2 + \frac{\Delta}{2} \middle| \frac{S_2(t)}{N_2(t)} \leq \mu_2 + \frac{\Delta}{4}\right) &\leq \exp\left(-2\big(N_2(t)+1\big)\frac{\Delta^2}{16}\right) \\ &\leq \exp\left(-\frac{L\Delta^2}{8}\right) \text{ since } N_2(t) \geq L \\ &\leq \exp\left(-\frac{24\ln(T)}{\Delta^2}\cdot\frac{\Delta^2}{8}\right) \text{ by def. of } L \\ &= \exp\left(-3\ln(T)\right) \\ &= \frac{1}{T^3} \qquad (3) \end{aligned}$$

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

$$\mathbb{P}\left(\theta_2(t) \geq \mu_2 + \frac{\Delta}{2}\right) \leq \frac{1}{T^3} + \frac{1}{T^3} = \frac{2}{T^3}$$

This is the stated result of `Theorem 2.7` □

3 Stochastic Dynamic Optimisation Problems

3.1 General

Definition 3.1 - *Stochastic System*
A *Stochastic System* is a dynamic system where at least one part of the system relies on a random process, modelled by random variables.

Definition 3.2 - *Stochastic Dynamic Optimisation*
Stochastic Dynamic Optimisation is the study of problems where an agent is tasked with making optimal or near-optimal decision in a *Stochastic System.*

Definition 3.3 - *Sequential Decision Process*
In a *Sequential Decision Process* an agent is tasked with choosing a sequence of actions such that a *Stochastic System* performs optimally wrt some pre-specified *Performance Criterion.* The agent is able to observe the current system-state before taking each action.

A *Sequential Decision Process* has the following components which need to be defined

- *Time-Horizon*, T - Time epochs in which actions are taken and their effect realised.
- *State-Space*, S - A mathematical encoding of available system information.
- *Action-Space*, A - Set of actions an agent is able to take, which affect the system. Available actions may depend on the current system state.
- *Transition Probabilities*, $p_t(\cdot|\cdot,\cdot)$ - A mathematical description of the underlying stocahstic system, relating agent actions and system states.
- *Immediate Rewards/Costs*, $r_t(\cdot,\cdot)$ - The reward/cost an agent recieves/incurs after taking an action.

Definition 3.4 - *Time-Horizon, T*
The *Time-Horizon* T is the set of all *Decision Epochs.* There are three types of *Time-Horizon*

i). *Continuous-Time* - $T=[t_0,t_1]$.[2]

ii). *Finite Discrete-Time* - $T=\{t_0,\dots,t_N\}$.

iii). *Infinite Discrete-Time* - $T=\{t_0,t_1,\dots\}$.

Definition 3.5 - *State-Space, S*
The *State-Space* S is the set of all states the *Stochastic System* can take. There are three types of *State-Space*[3]

i). *Continuous-State* - *State-Space* is uncountable.

ii). *Finite Discrete-State* - *State-Space* is countaly finite $S=\{s_1,\dots,s_n\}$.

iii). *Infinite Discrete-State* - *State-Space* is countably infinite.

[2]Continuous-time time-horizons are out of the scope of this module.
[3]Only *Finite Discrete-State-Spaces* are in scope of this module.

Definition 3.6 - *Action-Space, A*
The *Action-Space* A is the set of actions the agent can take. There are three types of *Action-Space*[4]

i). *Continuous-Action* - *Action-Space* is uncountable.

ii). *Finite Discrete-Action* - *Action-Space* is countaly finite $A = \{s_1, \dots, s_n\}$.

iii). *Infinite Discrete-Action* - *Action-Space* is countably infinite.

The *Admissible Action-Space* $A(s) \subseteq A$ is the set of actions the agent can take if the system is in state s.

Definition 3.7 - *Transition Probabilities, $p_t(\cdot|\cdot,\cdot)$*
Transition Probabilities $p_t(\cdot|\cdot,\cdot)$ are parametric-probability mass functions which define the probability of the

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

Definition 3.8 - *Decision Rules $q_t(\cdot)$,$d_t(\cdot)$*
A *Decision Rule* $q_t(\cdot)$ is a procedure the agent uses to decide what action to take, given available information (Current state X_t, previous states $X_{0:t-1}$, previous actions $Y_{0:t-1}$).

There are four classes 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, using all available information.

$$\begin{array}{rcl} \mathbb{P}(Y_0 = a_0|X_0 = s_0) & = & q_0(a_0|s_0) \\ (Y_0|X_0) & \sim & q_0(\cdot|X_0) \\ \mathbb{P}(Y_t = a_t|X_{0:t} = s_{0:t}, Y_{0:t-1} = a_{0:t-1}) & = & q_t(a_t|s_{0:t}, a_{0:t-1}) \\ (Y_t|X_{0:t}, Y_{0:t-1}) & \sim & q_t(\cdot|X_{0:t}, Y_{0:t-1}) \end{array}$$

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

$$Y_t := d_t(X_{0:t}, Y_{0:t-1})$$

iii). *Markovian Randomised*, MR - The *Decision Rule* $q_t(\cdot)$ is a conditional mass function on the action-space A, using only the current system-state.

$$\begin{array}{rcl} \mathbb{P}(Y_t = a_t|X_{0:t} = s_{0:t}, Y_{0:t-1} = a_{0:t-1}) & = & \mathbb{P}(Y_T = a_t|X_t = s_t) \\ & = & q_t(a_t|s_t) \\ (Y_t|X_{0:t}, Y_{0:t-1}) & \sim & q_t(\cdot|X_t) \end{array}$$

iv). *Markovian Deterministic*, MD - The *Decision Rule* $d_t(\cdot)$ is a deterministic function of the current system-state

$$Y_t := d_t(X_t)$$

[4] Only *Finite Discrete-Action-Spaces* are in scope of this module.

Remark 3.1 - *Memoryless*
Markovian Decision Rules are memoryless.

Definition 3.9 - *Decision Policy* π
A *Decision Policy* π is a sequence of *Decision Rules*, specifying which *Decision Rule* $q_t(\cdot)$ to use in each epoch.

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

There are two types of *Decision Policy*

i). *Stationary Decision-Policy* - The same decision rule is applied in each epoch.

$$\exists\ q(\cdot) \text{ st } q_t(\cdot) = q(\cdot)\ \forall\ t \in T$$

ii). *Non-Stationary Decision-Policy* - A variety of *Decision Rules* are used. Which specific one is used depends on the current epoch t.

Remark 3.2 - *Static vs Dynamic Approach*
There are two approaches to a *Sequential Decision Process.*

Static The agent decides what actions they take before the first decision epoch. This means agent actions are independent of the system state.

Dynamic The agent decides their action each epoch, taking the current system state into account.

As there is no penalty for delaying choosing a move until the epoch in which you make it, there is little reason not to take the *dynamic* approach.

Definition 3.10 - *Induced Stochastic Process* $\{(X_t, Y_t)\}_{t\geq 0}$
The *Induced Stochastic Process* $\{(X_t, Y_t)\}_{t\geq 0}$ is the time-evolution of the agent actions X_t and system states Y_t in the stochastic system.

The *Induced Stochastic Process* can be fully defined by

i). The probability mass function of X_0.

ii). The system's transition probabilities $\{p_t(\cdot|\cdot,\cdot)\}_{t\in T}$.[5]

iii). The agent's decision policy $\pi := \{q_t(\cdot|\cdot)\}_{t\in T}$.[6]

Proposition 3.1 - *Distributions of Induced Stochastic System*
In an *Induced Stochastic Process* the following distributions exist

$$\begin{aligned}
\mathbb{P}(X_{0:t} = s_{0:t}, Y_{0:t-1} = a_{0:t-1}) &= \mathbb{P}(X_0 = s_0)\prod_{k=0}^{t-1}\underbrace{p_k(s_{k+1}|s_k, a_k)}_{\text{Transition}}\underbrace{q_k(a_k|s_{0:k}, a_{0:k-1})}_{\text{Decision}}\\
\mathbb{P}(X_{t+1} = s'|X_t = s) &= \sum_{a\in A(s)} p_t(s'|s, a)q_t(a, s)
\end{aligned}$$

Theorem 3.1 - *Markov Chains in SDPs*

[5]Specifies the probability the system is in state s', given the agent took action a which the system was in state s.

[6]Specifies the probability of an agent taken a given action, given the current state of the system.

When using a *Markovian Decision Policy* in a *Stochastic Dynamic Process*, the sequence of states $\{X_t\}_{t\in T}$ and the sequence of state-action pairs $\{(X_t, Y_t)\}_{t\in T}$ are *Markov Chains*.

Moreover, if the transition and decision probabilities are stationary (ie independent of t), then they are *Homogeneous Markov Chains*.

3.2 Markov Decision Processes

Definition 3.11 - *Markov Decision Process, MDP*
A *Markov Decision Process*, MDP, is a *Sequential Decision Problem* where the underlying *Stochastic System* has the *Markov Property*. This is realised by the state of the system in epoch $t+1$ only depending upon the system state and agent action in epoch t.

In each decision epoch, *Markov Decision Process* follows the following steps

i). The agent observes the system state X_t.

ii). Based on this observation, the agent chooses an action Y_t to take.

iii). The agent recieves an immediate reward $r(X_t, Y_t)$ and the system evolves X_{t+1}.

Definition 3.12 - *Markov Decision Problem*
In a *Markov Decision Problem*, the agent is tasked with finding a *Decision Policy* π which maximise the expected total reward received[7] in a given time-horizon T.

$$\max_{\pi} \mathbb{E}^{\pi}\left[\sum_{t\in T} r(X_t, Y_t)\right]$$

A *Markov Decision Problem* is defined by the same components as a *Sequential Decision Problem* (See `Definition 3.3`). The *Transition Probabilities* $p_t(\cdot|\cdot,\cdot)$ are required to have the *Markov Property*, meaning we have the following stochastic system

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

Remark 3.3 - *Initial State X_0*
The initital state X_0 of the system is independent of the agent's actions and thus the chosen policy π.

3.3 General Finite-Horizon MDPs

3.3.1 Problem Formulation

Definition 3.13 - *General Finite-Horizon MDP*
In a *Finite-Horizon Markov Decision Problem* the agent has a finite-number of epochs in which to take actions in and seeks to maximise the total expected reward received.

All *Finite-Horizon MDPs* have the following features[8]

[7] As the reward received in each epoch $r_t(\cdot,\cdot)$ depends upon random quantities (namely system states), we cannot maximise total reward directly and instead maximise its expectation wrt the chosen policy.

[8] The number of epochs N, state-space S, action-space A, transition probabilities $p_t(\cdot)$ and rewards $r_t(\cdot)$ are all specified on a problem-by-problem basis.

- *Number of Epochs* - $N \in \mathbb{N}$.
- *Time-Horizon* - $T = \{0, \dots, N-1\}$.
- *Transition Probabilities* - $p_0(s'|s,a), \dots, p_{N-1}(s'|s,a)$.
- *Immediate Rewards* - $r_0(s,a), \dots, r_{N-1}(s,a), r_N(s)$.[9]
- *Objective* - Given the transition probabilities $\{p_t(\cdot|\cdot,\cdot)\}_{t\in T}$, immediate rewards $\{r_t(\cdot,\cdot)\}_{t\in T}$ and terminal reward $r_N(\cdot)$, the agent is tasked to find a *History Dependent Randomised* policy $\pi \in HR(T)$ over time-horizon T which maximises the exepcted total reward

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

Proposition 3.2 - *Stochastic System of a Finite-Horizon MDP*
In epoch t *Finite-Horizon MDPs* have the following *Stochastic System*, given all available information

$$\begin{array}{rcl}(X_{t+1}|X_{0:t},Y_{0:t}) & \sim & (X_{t+1}|X_t,Y_t)\\ & \sim & p_t(X_{t+1}|X_t,Y_t)\end{array}$$

3.3.2 Optimisation

Remark 3.4 - *Computational Cost of Optimisation*
Calculating optimal strategies for *Finite-Horizon MDPs* is computationally expensive, especially for large N. Hence approximating *Finite-Horizon MDPs* are other problems is desirable. This is explored in `Section 3.4.2` and `Section 3.5.2` /

Remark 3.5 - *Dynamic Programming Algorithm*
The *Dynamic Programming Algorithm* is a system equation for determining the optimal *Decision Policy* π^* for a *Finite-Horizon MDP*. These equations are defined as a *Backwards Recursion*[10]

$$\begin{array}{rcl}u_N^*(s) & = & r_N(s)\\ u_t^*(s) & = & \max\limits_{a\in A(s)}\left\{r_t(s,a)+\sum\limits_{s'\in S}u_{t+1}^*(s')p_t(s'|s,a)\right\}\quad t\in T=[0,N-1]\\ d_t^*(s) & = & \text{argmax}_{a\in A(s)}\left\{r_t(s,a)+\sum\limits_{s'\in S}u_{t+1}^*(s')p_t(s'|s,a)\right\}\quad t\in t=[0,N-1]\end{array}$$

$u_t^*(s)$ is the *Optimality Equation* and takes the value of the maximum expected reward which can be earned in the last $N-t$ steps of the problem, due to its recursive definition.

$d_t^*(s)$ is the *Optimal Decision Rule* and takes the value of the action which produces the greatest expected reward from the last $N-t$ steps of the problem.

Definition 3.14 - *Optimal Policy* π^*
An *Optimal Policy* π^* is any policy which produces the maximum expected total reward when applied to the defined *Finite-Horizon MDP*. In this case it is

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

[9] $r_N(s)$ is the *Terminal Reward* and depends on the final state of the system.
[10] Iterate from $N-1$ to 0.

Definition 3.15 - *Value Function $v^\pi(\cdot)$ and Optimal Value Function $v^*(\cdot)$*
The *Value Function* $v^pi(\cdot)$ is the expected total reward, given the initial state of the system $X_0 = s$ and the policy $\pi \in HR(T)$ which is being used.

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

The *Optimal Value Function* $v^*(\cdot)$ is the maximum expected total reward, given the initial state of the system $X_0 = s$

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

Theorem 3.2 - *Optimal Value Function and Dynamic Programming Algorithm*
Here are two equivalent expressions of the *Optimal Value Function* $v^*(\cdot)$

$$\begin{array}{rclr} v^*(s) & = & u_0^*(s) & \forall\ s \in S \\ v^*(s) & = & v^{\pi^*}(s) & \forall\ s \in S \end{array}$$

3.3.3 Optimality Principle

Definition 3.16 - *Tail Subproblem*
Consider a *Finite-Horizon MDP* over N epochs.

The *Tail Subproblem of Length* L[11] of this *Finite-Horizon MDP* is a subproblem which is concerned with the last L epochs of the full problem. It has the following features

- *Number of Epochs* - $L \in [1, N]$.
- *Time-Horizon* - $T_L = \{N - L, \dots, N - 1\}$.
- *Transition Probabilities* - $p_{N-L}(s'|s, a), \dots, p_{N-1}(s'|s, a)$.
- *Immediate Rewards* - $r_{N-L}(s, a), \dots, r_{N-1}(s, a), r_N(s)$.
- *Objective*[12] - Given the transition probabilities $\{p_t(\cdot|\cdot, \cdot)\}_{t\in T}$, immediate rewards $\{r_t(\cdot, \cdot)\}_{t\in T}$ and terminal reward $r_N(\cdot)$, the agent is tasked to find a *History Dependent Randomised* policy $\pi \in HR(T)$ over time-horizon T which maximises the exepcted total reward
$$\text{argmax}_{\pi \in HR(T_L)} \mathbb{E}^\pi\left[\left(\sum_{t=N-L}^{N-1} r_t(X_t, Y_t)\right) + r_N(X_N)\right]$$

[11] $L \in [1, N]$
[12] Same as the full problem, except over the reduced *Time-Horizon*

Remark 3.6 - *Equivalence of MDP and Tail Subproblem*
The *Tail Subproblem of Length* L has *Time-Horizon* $T = \{N-L, \dots, N-1\}$ and thus is equivalent to the full *Finite-Horizon MDP* with time-horizon $T = \{0, \dots, L\}$[13]

This means optimising the *Tail Subproblem* only requires re-indexing the *Optimality Equations* of the full problem.

Proposition 3.3 - *Optimising Tail Subproblem*
The *Optimality Equations* for a *Tail Subproblem of Length* L are defined sub-recursively as

$$\begin{array}{rcl} u^*_{L,N}(s) &=& r_N(s) \\ u^*_{L,t}(s) &=& \max_{a\in A(s)}\left\{r_t(s,a) + \sum_{s'\in S} u^*_{L,t+1}(s')p_t(s'|s,a)\right\} \quad t \in T_L = [N-L, N-1] \\ d^*_{L,t}(s) &=& \text{argmax}_{a\in A(s)}\left\{r_t(s,a) + \sum_{s'\in S} u^*_{L,t+1}(s')p_t(s'|s,a)\right\} \quad t \in t = [N-L, N-1] \end{array}$$

and the *Optimal Policy* π^*_L is

$$\pi^*_L := \{d^*_{L,t}(s)\}_{t\in T_L}$$

Remark 3.7 - *Optimising Tail Subproblem vs Optimising MDP*
The initial condition of the *Optimality Equations* for both the tail subproblem $u^*_{L,N}(\cdot)$ and the full problem $u^*_N(\cdot)$ have the same definition

$$u^*_{L,N}(s) := r_N(s) =: u^*_N(s)$$

In the equivalent epoch t, the *Optimality Equation* of the tail subproblem $u^*_{L,t}(\cdot)$ is a sub-recursion of the *Optimality Equation* for the full problem $u^*_t(\cdot)$.

Given these two properties, the *Optimality Equations* for the subproblem are all identical to that of the full problem for the equivalent epoch.

$$\begin{array}{rcl} u^*_{L,t}(s) &=& u^*_t(s) \\ d^*_{L,t}(s) &=& d^*_t(s) \end{array}$$

Definition 3.17 - *Optimal Value Function of Tail-Subproblem* $v^*_L(\cdot)$
The *Optimal value Function* for the *Tail-Subproblem* of length L is defined as

$$v^*_L(\cdot) := \max_{\pi\in HR(T_L)} \mathbb{E}^\pi\left[\left(\sum_{t=N-L}^{N-1} r_t(X_t,Y_t)\right) + r_N(X_N)\middle| X_{N-t} = s\right]$$

This value can be interpreted as the maximum expected total reward received from the last L epochs, given the system is in state s at the start of epoch $t = N-L$.

$$v^*_L(s) = y^*_{L,N-L}(s)$$

Theorem 3.3 - *Optimality Principle*

[13] *Finite-Horizon MDP* over L epochs.

Consider a *Finite-Horizon MDP* over N epochs and a *Tail Subproblem of Length L*.

The *Optimality Principle* states

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

Remark 3.8 - *Optimality Principle*
The *Optimality Principle* shows that the *Dynamic Programming Algorithm* can be solved by solving all the *Tail Subproblems of Length L* for all $L \in [1, N]$.

This can be used to restate the *Dynamic Programming Algorithm* as a forwards-recursion

$$\begin{array}{rcl} v_0^*(s) &:=& r_N(s) \\ v_t^*(s) &:=& \max_{a\in A(s)} \left\{r_{N-k}(s,a) + \sum_{s'\in S} v_{k-1}^*(s')p_{N-k}(s'|s,a)\right\} \quad t \in [1, N] \\ d_t^*(s) &:=& \text{argmax}_{a\in A(s)} \left\{r_{N-k}(s,a) + \sum_{s'\in S} v_{k-1}^*(s')p_{N-k}(s'|s,a)\right\} \quad t \in [1, N] \end{array}$$

3.4 Discounted Reward Infinite-Horizon MDPs

3.4.1 Problem Formulation

Definition 3.18 - *Discounted Reward Infinite-Horizon MDPs*
In a *Discounted Reward Infinite-Horizon MDP* the agent is tasked to find a policy which maximises the total expected discounted reward received.[14] All *Discounted Reward MDPs* have the following features

- *Number of Epochs* - $N = \infty$.
- *Time-Horizon* - $T_L = \{0, 1, \dots\}$.
- *Transition Probabilities* - $p_t(s'|s,a) = p(s'|s,a)\ \forall\ t \in T$.[15]
- *Immediate Rewards* - $r_t(s,a) = \alpha^t r(s,a)\ \forall\ t \in T$.[16]
- *Objective* - Given the transition probabilities $p(s'|s,a)$, immediate rewards $r(s,a)$ and discounting factor α, the agent is tasked to find a *History Dependent Randomised Policy* $\pi \in HR(T)$ over time-horizon T which maximises the expected total reward

$$\text{argmax}_{\pi\in HR(T)}\mathbb{E}^\pi\left[\sum_{t=0}^\infty r_t(X_t,Y_t)\right] = \text{argmax}_{\pi\in HR(T)}\mathbb{E}^\pi\left[\sum_{t=0}^\infty \alpha^t r(X_t,Y_t)\right]$$

Proposition 3.4 - *Stochastic System of a Discounted Reward MDP*
In epoch t, *Discounted Reward MDPs* have the following *Stochastic System*, given all available information

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

[14]The reward being "discounted" means that rewards received further into the further are weighted less. This is done by multiplying the expected reward in epoch $t \in T$ by α^t where $\alpha \in (0, 1)$.

[15]These are *Stationary Transition Probabilities*.

[16]These are *Stationary Rewards*.

Remark 3.9 - *Time-Importance of Rewards*
The value of α^t characterises the importance of the reward received in epoch t. The closer the value of α is to 0, the quicker the importance of rewards diminishes.

Theorem 3.4 - *Discounted Reward Converges*
Discounted Reward converges, thus it is reasonable to expect the *Value Function* to converge.

$$\sum_{t=0}^{\infty} \alpha^t |r(X_t, Y_t)| < \infty \quad \text{where } \alpha \in (0,1)$$

3.4.2 Using for Approximation

Proposition 3.5 - *Approximating Finite-Horizon MDPs as Discounted Reward MDPs*
Consider a *Finite-Horizon MDP* as defined in `Definition 3.13` and assume the following

i). The parameters of the *Stochastic System* and the parameters of the *Immediate Rewards* change slowly wrt time.

ii). $N \gg 1$.

Using these assumptions we can derive the following approximations of the *Transition probabilities* $p_t(s'|s,a)$, *Immediate Rewards* $r_t(s,a)$ and *Objective Function* of this *Finite-Horizon MDP* as

$$\begin{array}{rcll}
p_t(s'|s,a) & \approx & p(s'|s,a) & \text{By i)} \\
r_t(s,a) & \approx & r(s,a) & \text{By i)} \\
\left|\sum_{t=0}^{N-1} r_t(X_t,Y_t)\right| & \gg & |r_N(X_N)| & \text{By ii)} \\
\implies \mathbb{E}^\pi\left[\left(\sum_{t=0}^{N-1} r_t(X_t,Y_t)\right) + r_N(X_N)\right] & \approx & \mathbb{E}^\pi\left[\sum_{t=0}^{N_1} r_t(X_t,Y_t)\right] & \\
& \approx & \mathbb{E}^\pi\left[\sum_{t=0}^{N_1} r(X_t,Y_t)\right] & \\
& \approx & \lim_{N\to\infty} \mathbb{E}^\pi\left[\sum_{t=0}^{N_1} r(X_t,Y_t)\right] & \\
& = & \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} r(X_t,Y_t)\right] & \\
& \approx & \mathbb{E}^\pi\left[\sum_{t=0}^{\infty} \alpha^t r(X_t,Y_t)\right] & \text{for } \alpha \approx 1^{[17]}
\end{array}$$

These approximations can form the definition of a *Discounted Reward Infinite-Horizon MDP*.

[17] See `Remark 3.9`

Remark 3.10 - *The approximation is well-defined*
It is possible that the penultimate expression in `Proposition 3.5` is not well-defined[18]. To overcome this, we multiple the reward by α^t for $\alpha \in (0,1)$.

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

Since the *State-Space* S and the *Action-Space* A are both finite-sets, there is a finite upper-bound to the reward recieved

$$\begin{array}{rcl} \max_{s\in S, a\in A(s)} |r(s,a)| & = & c<\infty \\ \implies \sum_{t=0}^\infty \alpha^t |r(X_t,Y_t)| & \leq & \sum_{t=0}^\infty \alpha^t c \\ & = & \frac{c}{1-\alpha}<\infty \end{array}$$

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

Remark 3.11 - *Quality of Approximation*
We have that

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

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

i). $\alpha \in (0,1)$ and $\alpha \approx 1$.

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

iii). $N \gg 1$.

3.4.3 Optimisation

Definition 3.19 - *Value Function $v^\pi(\cdot)$ and Optimal Value Function $v^*(\cdot)$*
The *Value Function* $v^\pi(\cdot)$ of a policy $\pi \in HR(T)$ is the expected total discounted reward when using that policy, given the initial state of system $X_0 = s$.

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

The *Optimal Value Function* $v^*(\cdot)$ is the maximum expected total reward, given the initial state of the system is $X_0 = s$.

$$\begin{array}{rcl} v^*(s) & := & \max_{\pi\in HR(T)} v^\pi(s) \\ & := & \max_{\pi\in HR(T)} \mathbb{E}^\pi\left[\sum_{t=0}^\infty \alpha^t r(X_t,Y_t)\middle| X_0=s\right] \end{array}$$

[18] i.e. It may not have a finite value.

Theorem 3.5 - *Optimality Principle for $FHMDP(T_{N+1})$*
Let $u_N^*(s)$ be the maximum expected discounted reward for an *Approximated Finite-Horizon MDP* over N epochs, given the system starts in state $X_0 = s$.

$$u_N^*(s) := \max_{\pi \in HR(T_N)} \mathbb{E}^\pi \left[\sum_{t=0}^{N} \alpha^t r(X_t, Y_t) \middle| X_0 = s \right] \quad \text{where } T_n := \{0, \dots, N-1\}$$

The *Optimality Principle* for an *Approximated Finite-Horizon MDP* over $N+1$ epochs gives a recursive definition for $u_{N+1}^*(s)$

$$u_{N+1}^*(s) = \max_{a \in A(s)} \left(r(s,a) + \alpha \sum_{s' \in S} u_N^* p(s'|s,a) \right)$$

Definition 3.20 - *Bellman Equation*
The *Bellman Equation* is the *Optimality Equation* for a *Discounted Reward MDP*, stated as

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

where $v^*(\cdot)$ is unknown[19]

Proposition 3.6 - *Derivation of Bellman Equation*
We want to establish the relationship between $u_N^*(\cdot)$ and $v^*(\cdot)$.[20]
Since *Discounted Reward* converges[21]

$$\begin{array}{rrcll}
 & \lim_{N\to\infty} u_N^*(s) & = & \lim_{N\to\infty} \left(\max_{\pi \in HR(T)} \mathbb{E}^\pi \left[\sum_{t=0}^{N} \alpha^t r(X_t, Y_t) \middle| X_0 = s \right] \right) & \\
 & & = & \max_{\pi \in HR(T)} \mathbb{E}^\pi \left[\lim_{N\to\infty} \left(\sum_{t=0}^{N} \alpha^t r(X_t, Y_t) \right) \middle| X_0 = s \right] & \\
 & & = & \max_{\pi \in HR(T)} \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \alpha^t r(X_t, Y_t) \middle| X_0 = s \right] & \\
 & & = & v^*(s) \text{ by def.} & \\
\implies & v^*(s) & = & \lim_{N\to\infty} u_{N+1}^*(s) & \text{by above} \\
 & & = & \lim_{N\to\infty} \left(\max_{a \in A(s)} \left(r(s,a) + \alpha \sum_{s' \in S} u_N^*(s') p(s'|s,a) \right) \right) & \text{by Optimality Principle} \\
 & & = & \max_{a \in A(s)} \left(r(s,a) + \alpha \sum_{s' \in S} \left(\lim_{N\to\infty} u_N^*(s') \right) p(s'|s,a) \right) & \\
 & & = & \max_{a \in A(s)} \left(r(s,a) + \alpha \sum_{s' \in S} v^*(s') p(s'|s,a) \right) & \text{by above} \\
\implies & v^*(s) & = & \max_{a \in A(s)} \left(r(s,a) + \alpha \sum_{s' \in S} v^*(s') p(s'|s,a) \right) \text{[22]} &
\end{array}$$

[19] And thus the function we wish to find.
[20] i.e. relate the optimality equation for a *Finite-Horizon MDP* $u_N^*(\cdot)$ to the optimality equation for a *Discounted Reward MDP* $v^*(\cdot)$
[21] See `Theorem 3.4`.

Remark 3.12 - *Compact Bellman Equation*
The *Bellman Equation* can be written as[23]

$$(Tv)(s) = v(s)$$

where $v(\dot{})$ is unknown and $T : (S \to \mathbb{R}) \to (S \to \mathbb{R})$ is the following transform

$$T(v(s)) := \max_{a \in A(s)} \left(r(s,a) + \alpha \sum_{s \in S'} v(s')p(s'|s,a) \right)$$

Theorem 3.6 - *Transform T is a Contractive Mapping*
Let $v', v'' : S \to \mathbb{R}$ be value functions[24] and $\alpha \in (0,1)$ be the *Discount Factor*

The transform T, defined in `Remark 3.12`, is a *Contractive Mapping* since $\alpha \in (0,1)$

$$\|(Tv')(s) - (Tv'')(s)\| \leq \alpha \|v'(s) - v''(s)\| \ \forall \ s \in S$$

Definition 3.21 - *Transform T_d*
Let $d : S \to A$ be a *Markovian Decision Function.*

We define the transform $T_d : (S \to \mathbb{R}) \to (S \to \mathbb{R})$ as

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

This is the expected reward from the single epoch when the system in state s and decision function d is used.

Theorem 3.7 - *Bounding Distance between Transformations*
Let $v', v'' : S \to \mathbb{R}$ be value functions and $\alpha \in (0,1)$ be the *Discount Factor*

The transform T_d, defined in `Definition 3.21`, is a *Contractive Mapping* since $\alpha \in (0,1)$

$$\|(T_d v')(s) - (T_d v'')(s)\| \leq \alpha \|v'(s) - v''(s)\| \ \forall \ s \in S$$

Theorem 3.8 - *Banach Fixed-Point Theorem Applied to Transform T*
The following is the *Banach Fixed-Point Theorem* applied to transform T.

i). Let $v_0 : S \to \mathbb{R}$ be an arbitrary value function and $\{v_k\}_{k \geq 0}$ be recusively defined as

$$v_{k+1}(s) := (Tv_k)(s)^{[25]} \implies v_{k+1}(s) = (T^{k+1}v_0)(s)$$

[22]This is the *Bellman Equation.*
[23]This equation is known as the *Fixed-Point Equation* and is used in the *Banach Fixed Point Theorems* (`Theorem 3.8,3.9`)
[24]ie Value functions for different policies
[25]This is known as the *Fixed-Point Recursion*

Then, after applying transform T sufficiently many times to v_0 we get a solution to the *Compact Bellman Equation*

$$\begin{array}{rrcl} & \lim_{k\to\infty} v_k(s) = \lim_{k\to\infty}(T^k v)(s) & = & v(s) \quad \forall\ s\in S \\ \text{and} & (Tv)(s) & = & v(s) \quad \forall\ s\in S \end{array}$$

Moreover, we can bound how close we are to a solution

$$\begin{array}{rrcll} & \|v_k(s)-v(s)\| & \leq & \dfrac{\alpha^k\|v_1(s)-v_0(s)\|}{1-\alpha} & \forall\ s\in S, k\geq 1 \\ \iff & \|(T^k v_0)(s)-v(s)\| & \leq & \dfrac{\alpha^k\|(Tv_0)(s)-v_0(s)\|}{1-\alpha} & \forall\ s\in S, k\geq 1 \end{array}$$

ii).

$$\text{If } \exists\ v' : S\to\mathbb{R} \text{ st } \forall\ s\in S,\ (Tv')(s)=v'(s) \implies v'(s)=v(s)\ \forall\ s\in S$$

Theorem 3.9 - *Banach Fixed-Point Theorem Applied to Transform* T_d
The following is the *Banach Fixed-Point Theorem* applied to transform T.

Let $d : S\to A$ be a *Markovian Decision Function.* Then there is a unique solution $v_d : S\to\mathbb{R}$ to the *Copact Bellman Equation* for transform T_d

$$\exists!\ v_d : S\to\mathbb{R} \text{ st } \forall\ s\in S,\ (T_d v_d)(s)=v_d(s)$$

Moreover, this solution $v_d(\cdot)$ is equivalent to the value function $v^{\pi_d}(\cdot)$ for π_d[26]

$$v_d(s)=v^{\pi_d}(s)\quad\forall\ s\in S$$

Theorem 3.10 - *Bounding Distance between Optimal Value Function over Infinite* v^* *and Finite* v_N^* *Horizons*
Let v^* be the *Optimal Value Function* over an infinite time-horizon and v_N^* be the *Optimal Value Function* over a finite time-horizon with N epochs.

Then

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

Theorem 3.11 - *Solution to Bellman Equation*
The *Optimal Value Function* v^* is the unique solution to the *Bellman Equation*

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

Since $v^*(s)=v^{\pi^*}(s)\ \forall\ s\in S$ then we can deduce the *Optimal Decision Rule* is

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

The *Optimal Policy* is $\pi^*=\pi_{d^*}$.

Remark 3.13 - π_{d^*} *is an Optimal Policy*

[26] π_d is the stationary policy based on decision policy d.

3.4.4 Policy Iteration Algorithm

Definition 3.22 - *Policy Iteration Algorithm*
The *Policy Iteration Algorithm* is an algorithm for finding an optimal decision policy π^* for an *Discounted Reward MDP*. 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 $k \geq 0$ perform the following

  i). *Policy Evaluation* - Compute a solution $v_k(\cdot)$ to the *Compact Bellman Equation*

  $$(T_{d_k} v) = v$$

  where v is unknown and T_{d_k} is the *Transform* defined in `Definition 3.21`.

  ii). *Policy Improvement* - Use the function v_k which has just been computed, to select a *Markovian Decision Function* $d_{k+1}(s)$ which, in each states $s \in S$, maximises the *Bellman Equation.*

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

  iii). *Terminatiom?* -

  If $\forall\ s \in S\ d_k(s) = d_{k+1}(s)$:[27] Stop the algorithm and return the last calculated *Markovian Decision Function* $d_{k+1}(\cdot)$.

  Else : Increment k and repeat i)-iii).

Remark 3.14 - *Policy Iteration Algorithm*
Here are some properties of the *Policy Iteration Algorithm*

i). The algorithm terminates after a finite number of iterations.

ii). The returned decision function is optimal $\forall\ s \in S,\ d_k(s) = d^*(s)$.

3.4.5 Equivalent Linear Program

Remark 3.15 - *Linear Programming*
Linear Programming methods can be used to solve *Discounted Reward MDPs.*

Proposition 3.7 - *Equivalent Linear Programming Problem for Discounted Reward MDP*
The following *Linear Program* is equivalent to a *Discounted Reward MDP*

Find $v : S \to \mathbb{R}$ which minimises $\sum_{s \in S} \gamma(s) v(s)$ under the restrictions that

i). $r(s,a) + \sum_{s' \in S} \alpha p(s'|s,a) v(s') \leq v(s) \quad \forall\ s \in S, a \in A(s)$.

ii). $\gamma(s) > 0\ \forall\ s \in S$

[27] If the decision function is unchanged.

3.5 Average Reward Infinite-Horizon MDPs

3.5.1 Problem Formulation

Definition 3.23 - *Average Reward Infinite-Horizon MDPs*
In an *Average Reward Infinite-Horizon MDP* the agent is tasked to find a policy which maximises the average reward in the long-run. All *Average Reward MDPs* have the following features

- *Number of Epochs* - $N = \infty$.
- *Time-Horizon* - $T = \{0, 1, \dots\}$.
- *Transition Probabilities* - $p_t(s'|s,a) = p(s'|s,a)\ \forall\ t \in T$.[28]
- *Immediate Rewards* - $r_t(s,a) = r(s,a)\ \forall\ t \in T$.[29]
- *Objective* - Given the transition probabilities $p(s'|s,a)$ and immediate rewards $r(s,a)$, the agent is tasked to find a *History Dependent Randomised Policy* $\pi \in HR(T)$ which maximises the expected average reward-per-epoch over the infinite time-horizon

$$\begin{array}{rl} & \text{argmax}_{\pi\in HR(T)}\left\{\lim_{N\to\infty}\inf\mathbb{E}^\pi\left[\frac{1}{N}\sum_{t=0}^{N-1}r_t(X_t,Y_t)\right]\right\} \\ = & \text{argmax}_{\pi\in HR(T)}\left\{\lim_{N\to\infty}\inf\mathbb{E}^\pi\left[\frac{1}{N}\sum_{t=0}^{N-1}r(X_t,Y_t)\right]\right\}\end{array}$$

Proposition 3.8 - *Stochastic System of an Average Reward MDP*
In epoch t, *Average Reward MDPs* have the following *Stochastic System*, given all available information

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

Remark 3.16 - *The problem is well-defined*
Since the *State-Space* S and *Action-Space* A are finite-sets, the maximum reward is finite and thus the average reward over finite-time is finite.

$$\begin{array}{rrcl} & c := \max_{s\in S, a\in A(s)}|r(s,a)| & < & \infty \\ \implies & \left|\frac{1}{N}\sum_{t=0}^{N-1}r(X_t,Y_t)\right| & \leq & \frac{1}{N}\sum_{t=0}^{N-1}|r(X_t,Y_t)| \leq c\end{array}$$

Therefore the limit of the infimum and supremum exist and are finite

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

Remark 3.17 - *Average Reward MDP vs Discounted Reward MDP*

[28]The *Transitions Probabilities* are stationary.
[29]The *Rewards* are stationary.

An *Average Reward MDP* places the same emphasis on all received rewards. A *Discounted Reward MDP* only does this if the *Discount Factor* α is close to 1.

If $\alpha \approx 0$ then a *Discounted Reward MDP* places significantly more emphasis on near-future rewards than far-future.

Remark 3.18 - *Average Reward MDPs are similar to Irreducible Markov Chains*
See `Subsection 0.5` for details on *Irreducible Morkov Chains.*

3.5.2 Using for Approximation

Proposition 3.9 - *Approximating Finite-Horizon MDP as Average Reward MDP*
Consider a *Finite-Horizon MDP* as defined in `Definition 3.13` and assume the following

i). The parameters of the *Stochastic System* and the parameters of the *Immediate Rewards* change slowlt wrt time.

ii). $N \gg 1$.

Using thses assumptions we can derive the following approximation of the *Transition Probabilities* $p_t(s'|s,a)$, *Immediate Rewards* $r_t(s,a)$ of this *Finite-Horizon MDP*

$$\begin{aligned}
p_t(s'|s,a) &\approx p(s'|s,a) && \text{By i)}\\
r_t(s,a) &\approx r(s,a) && \text{By i)}\\
\left|\sum_{t=0}^{N-1} r_t(X_t,Y_t)\right| &\gg |r_N(X_N)| && \text{By ii)}\\
\implies \mathbb{E}^\pi\left[\left(\sum_{t=0}^{N-1} r_t(X_t,Y_t)\right)+r_N(X_N)\right] &\approx \mathbb{E}^\pi\left[\sum_{t=0}^{N-1} r_t(X_t,Y_t)\right]\\
&\approx \mathbb{E}^\pi\left[\sum_{t=0}^{N-1} r(X_t,Y_t)\right] && \text{By i)}
\end{aligned}$$

Since introducing a multiplicative constant does affect the argmax of an expression, we have that the following expressions are all equivalent. Thus optimising the total reward and average reward are equivalent objectives.

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

We can approximate this *Objective Function* using a limit

$$\begin{aligned}
\implies \mathbb{E}^\pi\left[\frac{1}{N}\sum_{t=0}^{N-1} r_t(X_t,Y_t)\right] &\approx \lim_{N\to\infty} \mathbb{E}^\pi\left[\frac{1}{N}\sum_{t=0}^{N-1} r(X_t,Y_t)\right] && \text{By ii)}\\
&= \lim_{N\to\infty} \inf \mathbb{E}^\pi\left[\frac{1}{N}\sum_{t=0}^{N-1} r(X_t,Y_t)\right] && \text{[30]}
\end{aligned}$$

These approximations can form the definition of an *Average Reward Infnite-Horizon MDP*.

Proposition 3.10 - *Quality of Approximation*
We have that

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

3.5.3 Optimisation

Definition 3.24 - *Bellman Equation*
The *Bellman Optimality Equation* for an *Average Reward MDP*, stated as

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

where $w^* : S \to \mathbb{R},\ r^* \in \mathbb{R}$ are unkowns to be found.[31]

Definition 3.25 - *Optimal Markovian Decision Function $d^*(\cdot)$*
The *Optimal Markovian Decision Function $d^* : S \to A$* for an *Average Reward MDP* is on which chooses the action $a \in A(s)$ which maximises the RHS of the *Bellman Equation*

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

Theorem 3.12 - *Solutions to the Bellman Equation Exist*
There exist $w^* : S \to \mathbb{R},\ r^* \in \mathbb{R}$ which staisfy the *Bellman Equation* for an *Average Reward MDP* (`Definition 3.25`).

Theorem 3.13 - *r^* is the Maximum Asymptotic Expected Average Reward*
Let $(w^*(s), r^*)$ be a solution-pair for the *Bellman Equation* for an *Average Reward MDP*.
Then

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

[30]We take the infimum to ensure the limit exists.
[31]w^* represents an invariant distribution in the stochastic system. r^* represents the reward value.

Further, let $d^*:S\to A$ be the *Optimal Markovian Decision Function* and π^* be the *Decision Policy* based on $d^*(s)$. Then

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

Theorem 3.14 - *Uniqueness of Solutions to the Bellman Equation*
Consider any two solutions-pairs $(w^*(s),r^*)$ and $(\tilde{w}^*(s),\tilde{r}^*(s))$ for the *Bellman Equation* for an *Average Reward MDP*. Then

i). The r-part will be the same in both solutions.

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

ii). The w-part will only differ by an additive constant

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

3.5.4 Policy Iteration Algorithm

Definition 3.26 - *Policy Iteration Algorithm*
The *Policy Iteration Algorithm* is an algorithm for finding an optimal decision policy π^* for an *Average Reward MDP*. Here are the stages of the *Policy Iteration Algorithm*.

- *Initialisation* - Arbitratily choose a *Markovian Decision Function* $d_0(s)$ and set $k=0$.
- *Body* - For $k\geq 0$ perform the following:

  i). *Policy Evaluation*
  - Compute a solution $\mu_k(\cdot)$ to the following equations[32]

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

    where $\mu(\cdot)$ is the unknown to be found.
  - Using this $\mu_k(\cdot)$, compute a solution $w_k(\cdot)$ to the following set of equations[33]

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

    where $w(\cdot)$ is the unkown to be found. Note that r_k is defined explicitly given we known $\mu_k(\cdot)$.

  ii). *Policy Improvement* - Select a *Markovian Decision Policy* $d_{k+1}(\cdot)$ which, in each state $s\in S$, chooses an action which maximises the *Bellman Equation*.

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

[32]This is the *Invariant Mass Function* for the current decision function d_k.
[33]This is the *Poisson Equation* associated with function $r(s,d_k(s))$ and the transition kernel $p(\cdot|s,d_k(s))$.

iii). *Termination?*

If $\forall\ s\in S\ d_k(s)=d_{k+1}(s)$:[34] Stop the algorithm and return $d_{k+1}(\cdot)$.

Else : Increment k and repeat i)-iii).

Theorem 3.15 - *Optimality of the Policy Iteration Algorithm*
The following are properties of the the *Policy Iteration Algorithm* for *Average Reward MDPs*

i). The algorithm terminates after a finite number of iterations.

ii). The returned decision function is optimal $\forall\ s\in S,\ d_k(s)=d^*(s)$.

3.5.5 Equivalent Linear Program

Remark 3.19 - *Linear Programming*
Linear Programming methods can be used to solve *Discounted Reward MDPs.*

Proposition 3.11 - *Equivalent Linear Programming Problem for Average Reward MDP*
The following *Linear Program* is equivalent to an *Average Reward MDP*.

Minimise $r\in\mathbb{R}$ under the restrictions that

- $\displaystyle r(s,a)+\sum_{s'\in S}p(s'|s,a)w(s')\leq r+w(s)\quad\forall\ s\in S,\forall\ a\in A(s)$.[35]

Theorem 3.16 - *Optimality of Equivalent Linear Programming Problem*
Let $(\hat{r},\hat{w}(s))$ be an optimal solution to the *Linear Program* defined in `Proposition 3.11` and $\hat{d}(\cdot)$ be a *Markovian Decision Function* which chooses actions which maximise the RHS of the *Bellman Equation* using the invariant distribution $\hat{w}$

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

Then the decision function $\hat{d}(\cdot)$ will be optimal

$$\forall\ s\in S,\ \hat{d}(s)=d^*(s)$$

[34]If the decision function is unchanged.

[35]Both $r\in\mathbb{R}$ and $w:S\to\mathbb{R}$ are unknown. IDK what to do about w.

0 Reference

0.1 Notation

0.1.1 Problem Specific

Proposition 0.1 - *Notation for Multi-Armed Bandit Problem*
The following notation is used to simplifier analysis of the *Multi-Armed Bandit Problem*

$I(t)$	$\in$	$[1, K]$	The arm out strategy I plays at 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.
$\hat{\mu}_{j,n}$	$:=$	$\frac{S_j(t)}{N_j(t)}$	The sample mean reward from arm j in the first n plays of arm j.
Δ_i	$:=$	$(\mu^* - \mu_i)$	The reward lost from playing arm i rather than the optimal arm.

Proposition 0.2 - *Notation for Stochastic Optimisation Processes*
The following notation is used to simplifier analysis of the *Stochastic Optimisation Processes*

X_t	System state at the start of epoch t.
Y_t	Agent action in epoch t.
T	The time horizon.
A	Action-space.
$A(s)$	Admissible action-space.
S	State-space.
$p_t(s'\|s, a)$	Transition probabilities.
$q_t(a\|s)$	Policy decision probabilities.

0.2 Definitions

Definition 0.1 - *Jacobian* $J(\cdot)$
Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathbf{x} \in \mathbb{R}^n$.

$$J_f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \dots & \frac{\partial f_m}{\partial x_1}(\mathbf{x}) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_n}(\mathbf{x}) & \dots & \frac{\partial f_m}{\partial x_n}(\mathbf{x}) \end{pmatrix}$$

0.3 Theorems

Theorem 0.1 - *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

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

A proof for this is given in the full notes.

Theorem 0.2 - *Result for Poisson Processes*

Let $\{N_s\}_{s\in\mathbb{N}}$ be a poisson process with intensity $\lambda > 0$ and fix $n, t \in \mathbb{N}$. Then, given that $N_t = n$, the random times at which the process sees an increment in time $[0, t]$ are mutually independent and uniformly distributed on $[0, t]$

Theorem 0.3 - *Relationship between Beta and Bernoulli Random Variables*
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 0.1 - *Theorem 0.3*
Let $\{N_t\}_{t\in\mathbb{N}}$ be a poisson process with unit-intensity and T_n be the time of the n^{th} increment.

Let $X \sim \text{Beta}(\alpha, \beta)$ then by `Theorem 0.1` we can write $X = \frac{V}{V+W}$ where V, W are independent with distributions $V \sim \text{Gamma}(\alpha, 1)$, $W \sim \text{Gamma}(\beta, 1)$. If α, β are integers then we can interpret $T_\alpha \sim V$ and $(T_{\alpha+\beta} - T_\alpha) \sim W$.

Hence, the following events are equivalent

$$\{X > p\} \iff \left\{\frac{T_\alpha}{T_{\alpha+\beta}} > p\right\} \iff \{T_\alpha > pT_{\alpha+\beta}\}$$

N_t increments $\alpha + \beta - 1$ times in $(0, T_{\alpha+\beta})$. By `Theorem 0.2`, these increments are uniformly and independently distributed in $[0, T_{\alpha+\beta}]$.

Hence the number of increments in time $[0, pT_{\alpha+\beta}]$ has a $\text{Bin}(\alpha + \beta - 1, p)$ distribution. This the same distribution as Y from the stated theorem.

The event $\{T_\alpha > pT_{\alpha+\beta}\}$ is the event that the number of increments of N_t in $[0, T_{\alpha+\beta}]$ is at most $\alpha - 1$. Meaning the following events are equivalent

$$\{T_\alpha > pT_{\alpha+\beta}\} \iff \{Y \leq \alpha - 1\}$$

. Thus, we have a full chain of equivalent events

$$\begin{array}{rl} & \{X > p\} \iff \left\{\frac{T_\alpha}{T_{\alpha+\beta}} > p\right\} \iff \{T_\alpha > pT_{\alpha+\beta}\} \iff \{Y \leq \alpha - 1\} \\ \implies & \{X > p\} \iff \{Y \leq \alpha - 1\} \\ \implies & \mathbb{P}(X > p) \iff \mathbb{P}(Y \leq \alpha - 1) \end{array}$$

The result of the theorem. □

0.4 Conjugate Priors

Reward Distribution X	**Prior** π_0	**Posterior** $\pi_1(\cdot\|x)$	**Proof**
Bernoulli(p) with p unknown	$\pi_0(p) \sim \text{Beta}(\alpha, \beta)$	$\pi_1(p\|x) \sim \begin{cases} \text{Beta}(\alpha+1, \beta) & \text{if } x = 1 \\ \text{Beta}(\alpha, \beta+1) & \text{if } x = 0 \end{cases}$	`Proof 0.2`
Poisson(λ) with λ unknown	$\pi_0(\lambda) \sim \text{Gamma}(\alpha, \beta)$	$\pi_1(\lambda\|n) \sim \text{Gamma}(\alpha + n, \beta + 1)$	`Proof 0.3`
Normal(μ, 1) with μ	$\pi_0 \sim \text{Normal}(\mu_0, \sigma_0^2)$	$\pi_1(\mu\|x) \sim \text{Normal}\left(\frac{\mu_0 + x\sigma_0^2}{1+\sigma_0^2}, \frac{\sigma_0^2}{1+\sigma_0^2}\right)$	`Proof 0.4`

Proof 0.2 - *Beta Distributions are Conjugate Priors for Bernoulli Observations*
Let $X \sim \text{Bern}(\mu)$, let $\pi_0 \sim \text{Beta}(\alpha, \beta)$ be the prior for μ and $\pi_1(\cdot|X)$ be the posterior distribution for μ given X was observed. This means

$$\pi_1(\mu|x) \propto \pi_0(\mu)p_X(x)$$

Note that
$$\pi_0(\mu)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\mu^{\alpha-1}(1-\mu)^{\beta-1}\quad\text{and}\quad p_X(x)=\begin{cases}\mu & \text{x=1}\\ 1-\mu & \text{x=0}\end{cases}$$
First, consider the case when $X=1$
$$\begin{array}{rcl}\pi_1(\mu|X=1) & \propto & \pi_0(\mu)p_X(1)\\ & \propto & [\mu^{\alpha-1}(1-\mu)^{\beta-1}]\cdot\mu \text{ (only terms involving } \mu)\\ & = & \mu^{\alpha}(1-\mu)^{\beta-1}\\ & \sim & \text{Beta}(\alpha+1,\beta)\end{array}$$
Now, consider the case when $X=0$
$$\begin{array}{rcl}\pi_1(\mu|X=0) & \propto & \pi_0(\mu)p_X(0)\\ & \propto & [\mu^{\alpha-1}(1-\mu)^{\beta-1}]\cdot(1-\mu) \text{ (only terms involving } \mu)\\ & = & \mu^{\alpha-1}(1-\mu)^{\beta}\\ & \sim & \text{Beta}(\alpha,\beta+1)\end{array}$$
Combining these two cases we get the result of the theorem
$$\pi_1(\mu|x)\sim\begin{cases}\text{Beta}(\alpha+1,\beta) & \text{if } x=1\\ \text{Beta}(\alpha,\beta+1) & \text{if } x=0\end{cases}$$
□

Proof 0.3 - *Gamma Distributions are Conjugate Priors for Poisson Observations*
Let $X\sim\text{Poisson}(\lambda)$ where λ is unknown, π_0 be the prior distribution for λ and $\pi_1(\cdot|n)$ be the posterior distribution for λ, given the value n was sampled from X. This means
$$\pi_1(\lambda|n)\propto\pi_0(\lambda)p_\lambda(n)$$
where $p_\lambda(n):=\mathbb{P}(X=n)$ given $X\sim\text{Poisson}(\lambda)$.

Suppose $\pi_0\sim\text{Gamma}(\alpha,\beta)$ and note that
$$p_\lambda(n)=\frac{\lambda^n e^{-\lambda}}{n!}\quad\text{and}\quad\pi_0(\lambda)=\frac{\lambda^{\alpha-1}e^{-\lambda\beta}}{\Gamma(\alpha)\beta^\alpha}$$
As we are considering proportionality wrt λ, we can ignore terms which do not involve λ. Giving
$$p_\lambda(n)\propto\lambda^n e^{-\lambda}\quad\text{and}\quad\pi_0(\lambda)\propto\lambda^{\alpha-1}e^{-\lambda\beta}$$
Using these results we can build an expression for the posterior $\pi_1(\cdot|n)$
$$\begin{array}{rcl}\pi_1(\lambda|n) & \propto & \pi_0(\lambda)p_\lambda(n)\\ & \propto & \left(\lambda^{\alpha-1}e^{-\lambda\beta}\right)\cdot\left(\lambda^n e^{-\lambda}\right)\\ & = & \lambda^{n+\alpha-1}e^{-\lambda(\beta+1)}\end{array}$$
By comparing this expression to that of a Gamma distribution we have that
$$\pi_1(\lambda|n)\sim\text{Gamma}(\alpha+n,\beta+1)$$
□

Proof 0.4 - *Normal Distributions are Conjugate Priors for Normal Distributions with Unit Variance*

Let $X \sim \text{Normal}(\theta, 1)$ with θ unknown and fix $\mu_0 \in \mathbb{R}, \sigma_0^2 > 0$.
Let $\pi_0 \sim \text{Normal}(\mu_0, \sigma_0^2)$ be the prior for θ and $\pi_1(\cdot|x)$ be the posterior for θ given x is observed from X. This means

$$\pi_1(\theta|x) \propto \pi_0(\theta) f_\theta(x) \quad \text{where } f_\theta(x) := \mathbb{P}(X = x|\theta)$$

Note that

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2} \quad \text{and} \quad \pi_0(\theta) = \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-\frac{1}{2}\frac{(\theta-\mu_0)^2}{\sigma_0^2}}$$

By considering only the terms involving θ we have

$$f_\theta(x) \propto e^{-\frac{1}{2}(x-\theta)^2} \quad \text{and} \quad \pi_0(\theta) \propto e^{-\frac{1}{2}\frac{(\theta-\mu_0)^2}{\sigma_0^2}}$$

This means

$$\pi_1(\theta|x) \propto e^{-\frac{1}{2}(x-\theta)^2} \cdot e^{-\frac{1}{2}\frac{(\theta-\mu_0)^2}{\sigma_0^2}} = \exp\left(-\frac{1}{2}\left((x-\theta)^2 + \frac{(\mu_0-\theta)^2}{\sigma_0^2}\right)\right)$$

Consider just the term of the exponent involving θ

$$\begin{array}{rl}
& (x-\theta)^2 + \frac{(\mu_0-\theta)^2}{\sigma_0^2} \\
= & x^2 - 2x\theta + \theta^2 + \frac{1}{\sigma_0^2}\left(\mu_0^2 - 2\mu_0\theta + \theta^2\right) \\
\propto & -2x\theta + \theta^2 + \frac{1}{\sigma_0^2}\left(-2\mu_0\theta + \theta^2\right) \\
= & \frac{1}{\sigma_0^2}\left[-2\sigma_0^2 x\theta + \sigma_0^2\theta^2 - 2\mu_0\theta + \theta^2\right] \\
= & \frac{1}{\sigma_0^2}\left[\theta^2(1+\sigma_0^2) - 2\theta(\mu_0 + x\sigma_0^2)\right] \\
= & \frac{1+\sigma^2}{\sigma_0^2}\left[\theta^2 - 2\theta\left(\frac{\mu_0+x\sigma_0^2}{1+\sigma_0^2}\right)\right] \\
\propto & \frac{1+\sigma^2}{\sigma_0^2}\left(\theta - \left(\frac{\mu_0+x\sigma_0^2}{1+\sigma_0^2}\right)\right)^2 \text{ by completing the square}
\end{array}$$

Substituting this result back into the expression for the posterior gives

$$\begin{array}{rcl}
\pi_1(\theta|x) & \propto & \exp\left(-\frac{1}{2} \cdot \frac{\left(\theta - \left(\frac{\mu_0+x\sigma_0^2}{1+\sigma_0^2}\right)\right)^2}{\sigma_0^2/(1+\sigma_0^2)}\right) \\
& \sim & \text{Normal}\left(\frac{\mu_0+x\sigma_0^2}{1+\sigma_0^2}, \frac{\sigma_0^2}{1+\sigma_0^2}\right)
\end{array}$$

Thus $\pi_1 \sim \text{Normal}(\mu_1, \sigma_1^2)$ where

$$\mu_1 := \frac{\mu_0 + x\sigma_0^2}{1+\sigma_0^2} \quad \text{and} \quad \sigma_1^2 := \frac{\sigma_0^2}{1+\sigma_0^2}$$

□

0.5 Irreducible Markov Chains

Definition 0.2 - *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 0.3 - *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 0.4 - *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 0.5 - *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 0.6 - *Poisson Equation*
Let $\{X_t\}_{t\geq 0}$ be an *Irreducible Markov Chain* with *transition kernel* $p(s'|s)$ and *Invariant PMF* $\mu(s)$. 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 0.7 - *Laurent Expansion of Resolvent*
Let $\{X_t\}_{t\geq 0}$ be an *Irreducible Markov Chain* with *transition kernel* $p(s'|s)$ and *Invariant PMF* $\mu(s)$. 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*

$$\begin{array}{rcl}\bar{f} & := & \sum_{s\in S} f(s)\mu(s) \\ f(s)-\bar{f} & = & \check{f}-\sum_{s'\in S}\check{f}(s')p(s'|s)\end{array}$$

$$\bar{f} := \sum_{s\in S} f(s)\mu(s)$$

Consider the following function

$$\begin{array}{rcl}\check{f}'(s) & := & \check{f}(s)-\sum_{s'\in S}\check{f}(s') \\ f_\alpha(s) & := & \mathbb{E}^\pi\left[\sum_{t=0}^\infty \alpha^t f(X_t)\big|X_0=s\right] \quad \alpha\in(0,1) \\ \tilde{f}_\alpha & := & f_\alpha(s)-\left(\frac{\bar{f}}{1-\alpha}+\check{f}'(s)\right)\end{array}$$

$\tilde{f}_\alpha(s)$ 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*,

$$f_\alpha(s)\approx\frac{\bar{f}}{1-\alpha}+\check{f}(s) \quad \text{when } \alpha\approx 1$$