

Stochastic Optimisation - Notes

Dom Hutchinson

October 20, 2020

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

Contents

1	Multi-Armed Bandit	2
1.1	The Problem	2
1.2	Regret Minimisation	3
1.3	Best Arm Identification for Bernoulli Distribution	4
1.4	Heuristic	4
1.5	UCB Algorithm	6
1.5.1	Analysis	7
1.5.2	Improvements?	8
1.6	Thompson Sampling	10
2	Probability	12
2.1	Probability Inequalities	12
2.2	Markov Processes	15
2.2.1	Discrete Time Markov Chains	16
2.2.2	Continuous Time Markov Process	18
2.2.3	Poisson Process	19
2.3	Transformation of Random Variables	20
0	Reference	23
0.1	Notation	24
0.1.1	Asymptotic Notation	24

1 Multi-Armed Bandit

1.1 The Problem

Example 1.1 - Motivating Example

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

Definition 1.1 - Multi-Armed Bandit Problem

An agent is faced with a choice of K actions. Each (discrete) time step the agent plays action i they receive a reward from the random real-valued distribution ν_i . Each reward is independent of the past. The distributions ν_1, \dots, ν_K are unknown to the agent.

In the *Multi-Armed Bandit Problem* the agent seeks to maximise a measure of long-run reward.

Remark 1.1 - Informal Definition of Multi-Armed Bandit Problem

Given a finite set of actions and a random reward for each action, how best do we learn the reward distribution and maximise reward in the long-run.

Definition 1.2 - Formal Definition of Multi-Armed Bandit Problem

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

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

There are a number of ways to formalise this objective.

Definition 1.3 - Strategy, $I(\cdot)$

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

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

Definition 1.4 - Long-Run Average Reward Criterion, X_*

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

$$X_* = \liminf_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbb{E}(X_{I(t)})$$

The *Infinum* is taken as there is no guarantee the limit exists (depending on the strategy), typically we will only deal with strategies where this limit exists.

Most strategies as based only on realisations of $\{X_i(s)\}_{s \in [1, t]}$, thus $\mathbb{E}(X_{I(t)}) \leq \mu^*$ and thus $X_* \leq \mu^*$. A strategy $I(\cdot)$ is *Optimal* if $X_* = \mu^*$.

Remark 1.2 - It is not hard to find an Optimal Strategy in the (very) long run, so we are going to look at Regret Minimisation First.

Proposition 1.1 - *Mathematical Model & Assumptions for Multi-Armed Bandit Problem Model:*

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

Assumptions:

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

We define the following quantities to make analysis easier

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

Definition 1.5 - *Policy*

A *policy* is a family of functions f_t which specify what arm is to be played in round t . f_t should depend on the information available at time t $\{I(s), X_{I(s)}(s) : s \in [1, t-1]\}$.

Randomised policies are allowed. So, in addition to the history up to time t , f_t can depend upon a $U(t) \sim U[0, 1]$ random variable which is independent of $X_i(\cdot)$. Thus

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

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

1.2 Regret Minimisation

Definition 1.6 - *Regret, R_n*

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

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

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

Remark 1.3 - *Minimising the growth rate of R_T with T is quite hard.*

The best achievable regret scales as $R_T \sim c \log T$ (i.e. $R_T/c \log T \xrightarrow{T \rightarrow \infty} 1$) where c depends on the reward distributions $X_1(t), \dots, X_K(t)$.

Definition 1.7 - Pseudo-Regret, \tilde{R}_n

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

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

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

1.3 Best Arm Identification for Bernoulli Distribution**Example 1.2 - Best Arm Identification for Bernoulli Bandits**

Consider a bandit with two *Bernoulli* arms: $\{X_1(t)\}_{t \in \mathbb{N}}$ IID RVs with distribution $\text{Bern}(\mu_1)$; and, $\{X_2(t)\}_{t \in \mathbb{N}}$ IID RVs with distribution $\text{Bern}(\mu_2)$.

Suppose $\mu_1 > \mu_2$ (i.e. arm 1 is better). Let the player play each arm n times and declare the arm with the greatest empirical mean to be the better arm. *What is the probability of choosing the wrong arm (Arm 2)?*

An error occurs if $\sum_{t=1}^n X_2(t) \geq \sum_{t=1}^n X_1(t)$ and thus we want to calculate the probability of this event.

Define $\{Y(t)\}_{t \in \mathbb{N}}$ st $Y(t) := \{X_2(t) - X_1(t)\}$. This means $Y(t) \in \{-1, 0, 1\} \subset [-1, 1]$.

To use *Hoeffding's inequality* we need to scale Y to be in $[0, 1]$, so we define $Z(t) := \frac{1}{2}(Y(t) + 1)$. We have $\mathbb{E}(Z(t)) = \frac{1}{2}(1 + \mu_2 - \mu_1)$ and an error occurs if $\sum_{t=1}^n Y(t) > 0 \iff \sum_{t=1}^n Z(t) \geq \frac{n}{2}$. By *Hoeffding's Inequality*

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

1.4 Heuristic**Remark 1.4 - How many tests?**

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

Proposition 1.2 - Naïve Heuristic (Single Test)

Consider a 2-armed bandit & the following Heuristic

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

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

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

This regret grows linearly in T .

Theorem 1.1 - Chernoff Bound of a Binomial Random Variable

Let $X \sim \text{Bin}(n, \alpha)$ with $n \in \mathbb{N}$, $\alpha \in (0, 1)$. Then

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

where

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

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

Similarly

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

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

Proposition 1.3 - Better Heuristic (N Tests)

Consider a 2-armed bandit problem & the following heuristic

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

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

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

where

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

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

Proposition 1.4 - Optimal N

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

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

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

Thus, the overall regret up to time T is

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

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

The corresponding regret is

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

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

1.5 UCB Algorithm

Remark 1.5 - UCB Algorithm

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

Remark 1.6 - Motivation

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

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

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

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

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

This suggests a heuristic:

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

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

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

Definition 1.8 - UCB(α) Algorithm

Consider the set up of the multi-armed bandit problem in **Proposition 1.1** and wlog that $\mu_1 > \mu_2 \geq \dots \geq \mu_K$.

Consider a k -armed bandit and let $\alpha > 0$.

- i). In the first K rounds, play each arm once.
- ii). At the end of each round $t \geq K$ compute the $UCB(\alpha)$ index of each arm i defined as

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

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

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

1.5.1 Analysis

Theorem 1.2 - Upper Bound on Regret

Consider a K -armed bandit and define $\Delta_i := \mu_1 - \mu_i$.

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

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

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

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

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

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

Theorem 1.3 - When a sub-optimal arm is played

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

- i). $\hat{\mu}_{1,N_1(s)} \leq \mu_1 - \sqrt{\frac{\alpha \ln(s)}{2N_1(s)}}$. The sample mean reward on the optimal arm is much smaller than the true mean.
- ii). $\hat{\mu}_{j,N_j(s)} \geq \mu_j + \sqrt{\frac{\alpha \ln(s)}{2N_j(s)}}$. The sample mean reward on arm j is much larger than its true mean.
- iii). $N_j(s) < \frac{2\alpha \ln(s)}{\Delta_j^2}$. Arm j has been played very few times.

Proof 1.1 - Theorem 1.3

This is a proof by contradiction.

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

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

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

Theorem 1.4 - Counting Lemma

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

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

with an empty sum defined to be zero.

Proof 1.2 - Theorem 1.4

Fix $t, u \in \mathbb{N}$. There are two possibilities

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

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

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

□

1.5.2 Improvements?

Remark 1.7 - *The regret of UCB grows logarithmically with T . No other algorithm can do better.*

Further, the constant factor of $\ln(T)$ used is almost optimal. This shall now be shown.

Definition 1.9 - Strongly Consistent

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

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

Theorem 1.5 - Lai and Robbins

Consider a K -armed bandit, where the rewards from arm i are iid $\text{Bern}(\mu_i)$ and rewards from distinct arms are mutually independent. Then, for any *strongly consistent* strategy, the number of times that a sub-optimal arm i is played up to time T , $N_i(T)$ satisfies

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

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

Proposition 1.5 - Lower bound on Regret

Here we derive a lower bound for the regret of any strongly consistent strategy from the multi-armed bandit problem.

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

Proposition 1.6 - Comparing to Lower bound of UCB(α)

We showed that the regret of the UCB(α) algorithm satisfies

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

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

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

Theorem 1.6 - Concentration Inequalities for Sample Means

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

Proof 1.3 - Theorem 1.5

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

Proof 1.4 - Theorem 1.2

Fix $t \in \mathbb{N}$ adn take $u_{t,j} := \left\lceil \frac{2\alpha \ln(t)}{\Delta_j^2} \right\rceil$.

By Theorem 1.4 we have that

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

Both sides involve random variables. Taking expectations we get

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

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

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

Thus

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

By Theorem 1.5

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

Thus

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

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

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

□

Remark 1.8 - *Is there an algorithm which achieves lower regret?*

No. There is no algorithm which has regret growing slower than $\ln(T)$.

1.6 Thompson Sampling

Remark 1.9 - *Thompson Sampling*

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

Theorem 1.7 - *Relationship between Beta & Gamma Distribution*

Let $X \sim \text{Gamma}(\alpha, \lambda)$ and $Y \sim \text{Gamma}(\beta, \lambda)$ (ie shared scale parameter but different shape parameters). Then

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

Proof 1.5 - *Theorem 1.7*

Consider the map $(X, Y) \mapsto \frac{X}{X+Y}$. This maps $\mathbb{R}^2 \rightarrow \mathbb{R}$ (reduces dimensions) and thus we cannot directly use the formula to compute the density of V .

We introduce an auxiliary random variable $W = X$. Now $(V, W) = g(X, Y) = \left(\frac{X}{X+Y}, X\right)$.

Note that X and Y are non-negative random variables. Hence, $V \in [0, 1]$ and $W \in \mathbb{R}^+$.

For $(v, w) \in [0, 1] \times \mathbb{R}^+$ we want to find all solutions of $g(x, y) = (v, w)$. Clearly $(x, y) = \left(w, w\frac{1-v}{v}\right)$ is a unique solution. The joint density of (V, W) is given by the formula

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

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

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

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

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

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

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

We are only interested in the marginal distribution of V

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

□

Proposition 1.7 - Thompson Sampling Algorithm

Consider a K -armed bandit with independent Bernoulli arms with parameters μ_1, \dots, μ_K . *Thompson Sampling* follows the following process.

- i). Define a prior distribution $\text{Beta}(1, 1)$ for the parameter of each arm.
- ii). At the start of round t , sample $\hat{\mu}_i(t)$ for $i \in [1, K]$, from the corresponding prior distributions.
- iii). Play the arm with the greatest sample value $I(t) \in \arg\max_{i \in [1, K]} \hat{\mu}_i(t)$ (breaking ties arbitrarily).
- iv). Compute a posterior distribution for that parameter, based on the observed reward. $\text{Beta}(\alpha + 1, \beta)$ if a reward is given and $\text{Beta}(\alpha, \beta + 1)$ if a reward is not given. (The priors for the other arms are the same as their priors as no result was observed).

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

The endpoint of the algorithm is when time runs out.

Proposition 1.8 - Choosing the Prior Distribution for Thompson Sampling

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

Remark 1.10 - Beta Distributions are conjugate priors for Bernoulli Random Variables

Theorem 1.8 - Posterior from a Beta Distribution Prior of a Bernoulli Random Variable

Let $X \sim \text{Bernoulli}(\mu)$ and assume a prior distribution $\mu \sim \text{Beta}(\alpha, \beta)$. Then, the posterior distribution of μ given $X = 1$ is $\text{Beta}(\alpha + 1, \beta)$ and given $X = 0$ is $\text{Beta}(\alpha, \beta + 1)$.

Proof 1.6 - Theorem 1.8

Let f_μ denote the density of μ . By our assumption of a $\text{Beta}(\alpha, \beta)$ prior we have

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

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

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

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

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

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

2 Probability

Definition 2.1 - Random Process

A *Random Process* is a collection of random variables indexed by time $\{X_t\}_{t \in T}$ (e.g. flipping a coin several times). Each of these random variables can take a value from a state space S . A random process is a *Discrete Time Process* if the index set T is discrete. A random process is a *Continuous Time Process* if the index set T is continuous.

2.1 Probability Inequalities

Remark 2.1 - We can use the moments of a random variable to determine bounds on the probability of it taking values in a certain set.

Theorem 2.1 - Markov's Inequality

Let X be a non-negative random variable. Then

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

Proof

Consider an event A and define its indicator $\mathbb{1}(A)(\omega) := \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$. Fix $c > 0$, then

$$\begin{aligned} \mathbb{E}(X) &\geq \mathbb{E}[X \mathbb{1}(X \geq c)] \\ &\geq \mathbb{E}[c \mathbb{1}(X \geq c)] \\ &= c \mathbb{P}(X \geq c) \\ \implies \mathbb{P}(X \geq c) &\leq \frac{1}{c} \mathbb{E}(X) \end{aligned}$$

Theorem 2.2 - Chebyshev's Inequality

Let X be a random-variable with finite mean and variance. Then

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

Proof

Note that the events $|X - \mathbb{E}(X)| \geq c$ and $(X - \mathbb{E}(X))^2 \geq c^2$ are equivalent. Note that $\text{Var}([X - \mathbb{E}(X)]^2) = \text{Var}(X)$. Then the result follows by *Markov's Inequality*.

Theorem 2.3 - Chebyshev's Inequality for Sum of IIDs

Let X_1, \dots, X_n be IID random variables with finite mean μ and finite variance σ^2 .

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

Proof

This is proved by extending the proof of **Theorem 2.2** and noting that the variance of a sum of IIDs is the sum of the individual variances.

Theorem 2.4 - Chernoff Bounds

Let X be a random variable whose moment-generating function $\mathbb{E}[e^{\theta X}]$ is finite $\forall \theta$. Then

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

Proof

Note that the events $X \geq c$ and $e^{\theta X} \geq e^{\theta c}$ are equivalent for all $\theta > 0$. The result follows by applying *Markov's Inequality* to $e^{\theta X}$ and taking the best bound over all possible θ .

$$\begin{aligned} \mathbb{P}(X \geq c) &= \mathbb{P}(e^{\theta X} \geq e^{\theta c}) \\ &\leq e^{-\theta c} \mathbb{E}(e^{\theta X}) \\ &\leq \inf_{\theta < 0} e^{-\theta c} \mathbb{E}(e^{\theta X}) \end{aligned}$$

Theorem 2.5 - Chernoff Bounds for Sum of IIDs

Let X_1, \dots, X_n be IID random variables. Then $\forall c \in \mathbb{R}$

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

Theorem 2.6 - Jensen's Inequality

Let f be a *Convex Function* and X be a random variable. Then

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

Theorem 2.7 - Bound on Moment Generating Function

Let X be a random variable taking values in $[0, 1]$ with finite expected value μ . Then we can bound the MGF of the centred random variable with

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

Proof (of weaker version)

Let X_1 be an independent copy of X , so both have mean μ . We can easily verify that $f(x) = e^{\theta x}$ is a convex function for all $\theta \in \mathbb{R}$. By *Jensen's Inequality* to $f(\cdot)$ and X_1

$$\mathbb{E}[e^{-\theta X_1}] \geq e^{-\theta \mathbb{E}[X_1]} = e^{-\theta \mu} \quad (1)$$

Consequently

$$\begin{aligned} \mathbb{E}[e^{\theta(X-X_1)}] &= \mathbb{E}[e^{\theta X}] \cdot \mathbb{E}[e^{-\theta X_1}] && \text{by independence} \\ &\geq \mathbb{E}[e^{\theta X}] \cdot e^{-\theta \mu} && \text{by (1)} \\ &= \mathbb{E}[e^{\theta(X-\mu)}] \\ \implies \mathbb{E}[e^{\theta(X-X_1)}] &\geq \mathbb{E}[e^{\theta(X-\mu)}] \end{aligned}$$

Since $X, X_1 \in [0, 1]$ then $(X - X_1) \in [-1, 1]$. As X, X_1 have the same distribution $\mathbb{E}(X - X_1) = 0$ and the distribution is symmetric around the mean.

Define random variable S which is independent of X, X_1 and takes values $\{-1, 1\}$, each with probability $p = \frac{1}{2}$. $S(X - X_1)$ has the same distribution as $(X - X_1)$ due to independence of S and symmetry of $(X - X_1)$. Hence

$$\begin{aligned} \mathbb{E}[e^{\theta(X-X_1)}] &= \mathbb{E}[e^{\theta S(X-X_1)}] && \text{by identical distribution} \\ &\leq \mathbb{E}[e^{\theta S}] && (2) \text{ since } (X - X_1) \in [-1, 1] \\ &= \frac{1}{2}(e^{\theta} + e^{-\theta}) && \text{by def. of expectation} \\ \implies \mathbb{E}[e^{\theta(X-X_1)}] &\leq \frac{1}{2}(e^{\theta} + e^{-\theta}) \end{aligned}$$

Note that $f(x) = e^x + e^{-x}$ is increasing for $x \in (0, \infty)$; decreasing for $x \in (-\infty, 0)$; and symmetric around 0.

Using a *Taylor Series* we can observe that

$$\begin{aligned} \frac{1}{2}(e^{\theta} + e^{-\theta}) &= \sum_{n=0}^{\infty} \frac{\theta^{2n}}{(2n)!} && \text{by Taylor expansion of } e^x \\ &\leq \sum_{n=0}^{\infty} \frac{(\theta^2/2)^n}{n!} \\ &\stackrel{\text{def.}}{=} e^{\theta^2/2} \\ \implies \frac{1}{2}(e^{\theta} + e^{-\theta}) &\leq e^{\theta^2/2} \end{aligned}$$

Combining all these results we get

$$\begin{aligned} \mathbb{E}[e^{\theta(X-\mu)}] &\leq \mathbb{E}[e^{\theta(X-X_1)}] \leq \frac{1}{2}(e^{\theta} + e^{-\theta}) \leq e^{\theta^2/2} \\ \implies \mathbb{E}[e^{\theta(X-\mu)}] &\leq e^{\theta^2/2} \end{aligned}$$

□

Theorem 2.8 - Hoeffding's Theorem

Let X_1, \dots, X_n be IID random variables taking values in $[0, 1]$ and with finite expected value μ . Then

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

Proof

From *Chernoff's Bound* we have that

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

Using **Theorem 2.7** to bound the moment generating function, we get

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

Thus, by taking logs and rearranging, we get

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

We have that $-\theta t + \frac{\theta^2}{8}$ is minimised at $\theta = 4t$ which is positive if t is positive. Thus, by applying this bound and substituting $\theta = 4t$ we get

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

□

2.2 Markov Processes

Definition 2.2 - Markov Property

A random process has the *Markov Property* if the conditional probability of a future state only depends on the current state.

$$\mathbb{P}(X_{t+1} = y | X_t = x_t, X_{t-1} = x_{t-1}) = \mathbb{P}(X_{t+1} = y | X_t = x_t)$$

A random process with the *Markov Property* is called a *Markov Process/Chain*.

Remark 2.2 - On this course we only deal with discrete time markov chains

Definition 2.3 - Transience

A state $x \in S$ is *Transient* if $\mathbb{P}(\exists t > 0 : X_t = x | X_0 = x) < 1$. The number of times the markov chain returns to a transient state is finite, with probability 1.

Definition 2.4 - Recurrent

A state $x \in S$ is *Recurrent* if $\mathbb{P}(\exists t > 0 : X_t = x | X_0 = x) = 1$. The number of times the markov chain returns to a recurrent state is infinite, with probability 1.

Every markov chain, with a finite state space S , has a recurrent communicating class.

Definition 2.5 - Communication Class

We say $y \in S$ is *Accessible* from $x \in S$ if $\exists t \geq 0$ st $[P^t]_{xy} > 0$.

We say x and y *communicate* (denoted xCy) if: x is *accessible* from y and y is *accessible* from x .

Communication is an *equivalence relation* on the state space S . Hence, *communication* partitions S into equivalence classes called *Communication Classes*. All elements of a *Communication Class* communicate with all other elements in the class, it is possible for elements to be accessible from another class but not for those elements to *communicate*.

If one state in a *Communicating Class* is *Transient/Recurrent* then all states are in that class.

If a *Markov Chain* has only one communicating class it is called *Irreducible*.

2.2.1 Discrete Time Markov Chains

Proposition 2.1 - Characterising a Discrete Time Markov Process

A *Discrete Time Markov Process* can be characterised by the set of all 1-step conditional probabilities

$$\mathbb{P}(X_{t+1} = y | X_t = x) \quad \forall x, y \in S$$

A markov chain is *time-homogeneous* if the 1-step conditionals only depend on x, y and not on t ($\mathbb{P}(X_{t+1} = y | X_t = x) = \mathbb{P}(X_1 = t | X_0 = x)$). The 1-step conditional probabilities of a *time-homogeneous markov process* can be specified in an $|S| \times |S|$ matrix P where

$$p_{x,y} = \mathbb{P}(X_{t+1} = y | X_t = x)$$

P is a *Stochastic Matrix*.

Proposition 2.2 - n -Step Transition Probabilities from 1-Step Transition Matrix

Let P be the 1-step transition matrix for a *time-homogeneous*.

The 2-step transition probabilities (ie $\mathbb{P}(X_{t+2} = z | X_t = x)$) can be found as

$$\begin{aligned} \mathbb{P}(X_{t+2} = z | X_t = x) &= \mathbb{P}(X_2 = z | X_0 = x) && \text{by time-homogeneity} \\ &= \sum_{y \in S} \mathbb{P}(X_2 = z, X_1 = y | X_0 = x) \\ &= \sum_{y \in S} \mathbb{P}(X_1 = y | X_0 = x) \mathbb{P}(X_2 = z | X_1 = y, X_0 = x) \\ &= \sum_{y \in S} \mathbb{P}(X_1 = y | X_0 = x) \mathbb{P}(X_2 = z | X_1 = y) \\ &= \sum_{y \in S} p_{xy} p_{yz} \\ &\equiv [P^2]_{xz} \end{aligned}$$

This can be generalise for the n -step transition probabilities with

$$\mathbb{P}(X_{t+n} = z | X_t = x) = [P^n]_{xz}$$

Proposition 2.3 - Any Joint Probability from 1-Step Transition Matrix

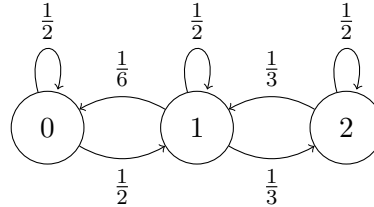
For a time homogeneous markov process the joint distribution of any transition can be computed by considering the individual steps of the transition.

$$\mathbb{P}(X_{n_0} = x_0, X_{n_1} = x_1, X_{n_2} = x_2, \dots) = \mathbb{P}(X_{n_0} = x_0) \cdot [P^{n_0 - n_1}]_{x_0 x_1} \cdot [P^{n_1 - n_2}]_{x_1 x_2} \dots$$

Proposition 2.4 - State Diagram Representation

A graph/automata can be drawn to represent the transition probability matrix P . A node is assigned for each member of the state space S and an arrow is drawn between each pair of nodes (x, y) where $P_{xy} \neq 0$. Generally the value of P_{xy} is denoted on the arrow.

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{6} & \frac{1}{2} & \frac{1}{3} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$


Definition 2.6 - Invariant Distribution

Let $\boldsymbol{\mu}(t)$ denote the probability distribution of random variable X (i.e, $\mu_x(t) = \mathbb{P}(X_t = x)$). Then

$$\begin{aligned} \boldsymbol{\mu}(t+1) &= \mathbb{P}(X_{t+1} = y) = \sum_{x \in S} \mathbb{P}(X_t = x, X_{t+1} = y) = \sum_{x \in S} \mu_x(t) p_{xy} = \boldsymbol{\mu}(t)P \\ \Rightarrow \boldsymbol{\mu}(t+1) &= \boldsymbol{\mu}(t)P \end{aligned}$$

A distribution $\boldsymbol{\pi}$ on the state space is called an *Invariant Distribution* if $\boldsymbol{\pi} = \boldsymbol{\pi}P$. If X_t has distribution $\boldsymbol{\pi}$ so will X_{t+1}, \dots . Every markov chain with a *finite* state space S has an *invariant distribution*. (Not necessarily true if S is infinite).

Proposition 2.5 - Finding an Invariant Distribution

If an *Invariant Distribution* it is easy to find by solving $\boldsymbol{\pi}P = \boldsymbol{\pi}$ and using normalising constant $\sum_{x \in S} \pi_x = 1$.

Remark 2.3 - If a Markov chain is irreducible, its invariant distribution (if one exists) is unique

If a *Markov Chain* is irreducible and has a finite state space, then it has a unique invariant distribution.

Example 2.1 - Markov Chains

- The *Asymmetric Simple Random Walk* on \mathbb{Z} is irreducible, transient and has no invariant distribution.
obvious not obvious
- The *Symmetric Simple Random Walk* on \mathbb{Z} is irreducible, recurrent and has no invariant distribution.
obvious not obvious

Theorem 2.9 - Ergodic Theorem for Markov Chains

Let $\{X_t\}_{t \in \mathbb{N}}$ be an irreducible markov chain on state space S (not necessarily finite) with unique invariant distribution $\boldsymbol{\pi}$. Then

$$\forall x \in S \quad \lim_{t \rightarrow \infty} \frac{1}{t} \sum_{s=1}^t \mathbb{1}(X_s = x) = \pi_x$$

i.e. The fraction of time spend in state $x \in S$ tends to π_x in the long run.

Definition 2.7 - Period

The *Period* of a state $x \in S$ is the greatest common divisor of all possible return times to x

$$\text{Period}(x) := \gcd(\{t > 0 : \mathbb{P}(X_t = x | X_0 = x) > 0\})$$

A state $x \in S$ is *Aperiodic* if $\text{Period}(x) = 1$. An irreducible markov chain is *aperiodic* if all its states are aperiodic.

All states in a *communicating class* have the same period.

Proposition 2.6 - Marginal Distribution of Irreducible, Aperiodic Markov Chain

If an irreducible, aperiodic Markov Chain has an invariant distribution π , then

$$\forall x \in S \quad \mu_x(t) \xrightarrow{t \rightarrow \infty} \pi_x$$

Definition 2.8 - Reversibility

A markov chain $\{X_t\}_{t \in \mathbb{Z}}$ is *Reversible* if all joint distributions are the same forwards and backwards in time. (i.e. the distribution of the chain is the same if it was reversed).

An irreducible markov chain $\{X_t\}_{t \in \mathbb{Z}}$ with transition matrix P is reversible iff

$$\exists \pi \quad \text{st} \quad \pi_x p_{xy} = \pi_y p_{yx} \quad \forall x, y \in S$$

This is the *Local/Detailed Balance Equation*. Note that this is a system of $\binom{|S|}{2}$ equations which need to be consistent for reversibility to exist.

2.2.2 Continuous Time Markov Process

Definition 2.9 - Continuous Time Markov Process

A stochastic process $\{X_t\}_{t \in \mathbb{R}}$ is a *Continuous Time Markov Process* on state space s if

$$\forall s < t \ \& \ x, y \in S \quad \mathbb{P}(X_t = y | X_s = x, X_u, u \leq s) = \mathbb{P}(X_t = y | X_s = x)$$

ie future values only depend on the present value and not past.

If $\forall t, s, x, y \ \mathbb{P}(X_t = y | X_s = x)$ depends only on $x, y, t - s$ (observed values & change in time) then the process is *Time-Homogeneous*.

For *Time-Homogenous Markov Processes* we let $P(t)$ denote the stochastic matrix with the probability of each possible transition after t time $[P(t)]_{xy} = \mathbb{P}(X_t = y | X_0 = x)$.

Remark 2.4 - A Time-Homogenous Markov Process is completely described by its initial condition and the family of transition probability matrices $\{P(t) : t \geq 0$

This set of matrices $\{P(t) : t \geq 0$ is uncountably large.

Definition 2.10 - Chapman-Kolmogorov Equations

For a *Time-Homogeneous Markov Process* the family of stochastic matrices $\{P(t) : t \geq 0$ satisfy the following:

- i). $P(0) = I$;
- ii). $P(t + s) = P(t)P(s) = P(s)P(t)$

Hence

$$\frac{d}{dt}P(t) := \lim_{\delta \rightarrow 0} \frac{P(t + \delta) - P(t)}{\delta} = \lim_{\delta \rightarrow 0} \overbrace{\frac{P(t)P(\delta) - P(t)}{\delta}}^{\text{ii)}} = P(t) \lim_{\delta \rightarrow 0} \frac{(P(\delta) - I)}{\delta}$$

Suppose that $Q := \lim_{\delta \rightarrow 0} \frac{P(\delta) - P(0)}{\delta} = \lim_{\delta \rightarrow 0} \frac{P(\delta) - \overset{i)}{I}}{\delta}$ exists.

Then $P(t)$ solve the following differential equations, known as the *Chapman-Kolmogorov Equations*

$$\frac{d}{dt}P(t) = \underbrace{P(t)Q}_{\text{forward eqn.}} = \underbrace{QP(t)}_{\text{backward eqn.}}$$

The solution to these equations is

$$P(t) = P(0)e^{Qt} = e^{Qt}P(0) = e^{Qt}I = e^{Qt}$$

N.B. Q is called the *Rate Matrix* or *Infinitesimal Generator* of the markov process.

Proposition 2.7 - Properties of the Rate Matrix, Q

Let Q be the rate matrix of a *continuous-time markov process*. Q has the following properties

- If $n \neq y$ then $q_{xy} := \lim_{\delta \rightarrow 0} \frac{[P(\delta)]_{xy} - 0}{\delta} \geq 0$. (The off-diagonal elements are non-negative).
- $\forall x \in S, \sum_{y \in S} q_{xy} = \lim_{\delta \rightarrow 0} \frac{1 - 1}{\delta} = 0$. The rows of Q sum to 0.
- Thus, the diagonal entries q_{xx} are negative. (We denote $-q_{xx}$ by q_x)

Proposition 2.8 - Interpreting the Rate Matrix Q

Let Q be the rate matrix of a *continuous-time markov process*.

If the markov process enters state x at time t , it will remain in x for a random time which is distributed $\text{Exp}(q_x)$. (Note that $q_x := -q_{xx}$).

It the jumps to state y with probability $\frac{q_{xy}}{q_x}$, independent of the past.

Definition 2.11 - Invariant Distributions

Suppose a *Continuous-Time Markov Process* starts with distribution $\mu(0)$ on state space S (i.e. $\mathbb{P}(X_0 = x) = [\mu(0)]_x$). Then, the distribution of X_t is $\mu(t) := \mu(0)P(t) = \mu(0) \underbrace{e^{Qt}}_{\text{CK Eqns}}$.

If there exists a distribution π on the state space S st $\forall t \geq 0 \pi = \pi P(t) = \pi \underbrace{e^{Qt}}_{\text{CK Eqns}}$, then π is

an *Invariant Distribution*. This distribution is invariant wrt time.

If a markov process has a finite state space then it definitely has an invariant distribution.

Invariant Distributions are not guaranteed to be unique.

Proposition 2.9 - Finding an Invariant Distribution

Starting with $\pi = \pi e^{Qt}$ we find that differentiating wrt t and then evaluating at time $t = 0$ we get $0 = \pi Q$ (The Global Balance Equations). This system of equations can be solved to find an *Invariant Distribution*.

A markov process is *reversible* iff there exists a distribution π on S which satisfies $\pi_x q_{xy} = \pi_y q_{yx} \forall x, y \in S$ (Local balance equations). Solving this system of equations will also find an *Invariant Distribution* but it is not guaranteed to have a solution.

Theorem 2.10 - Ergodic Theorem

Let $[X_t]_{t \in \mathbb{R}^+}$ is an *Irreducible Markov Process* on a state space S and has an invariant distribution π . Then

$$\forall x \in S \quad \pi_x = \lim_{t \rightarrow \infty} \frac{1}{t} \int_0^t \mathbb{1}(X_s = x) ds$$

Moreover, for an arbitrary initial distribution $\mu(0)$, $\mu(t)$ converges to π pointwise (i.e. $\mu_x(t) \xrightarrow{t \rightarrow \infty} \pi_x$)

2.2.3 Poisson Process

Definition 2.12 - Counting Process

A *Counting Process* is a stochastic process $\{N(t)\}_{t \in \mathbb{R}}$ st

- i). $N(0) = 0$ and $N(t) \in \mathbb{Z}$ for all $t \geq 0$; and,
- ii). $N(t)$ is a non-decreasing function of t .

Definition 2.13 - Independent Increments

A Process $\{N(t)\}_{t \in \mathbb{R}^+}$ is said to have *Independent Increments* if $\forall s \in (0, t)$, $(N(t) - N(s))$ is independent of $\{N(u) : u \in [0, s]\}$.

Definition 2.14 - Poisson Process

A *Poisson Process* is a *counting process* $\{N(t)\}_{t \in \mathbb{R}^+}$ which has independent increments and at least one of the following equivalence statements are true

- $\forall t \in [0, t] \quad (N(t) - N(s)) \sim \text{Po}(\lambda(t - s))$.
- $\mathbb{P}(N(t + \delta) - N(t) = 1) = \lambda\delta + o(\delta)$ and
 $\mathbb{P}(N(t + \delta) - N(t) = 0) = 1 - \lambda\delta + o(\delta)$ and
 $\mathbb{P}(N(t + \delta) - N(t) \geq 2) = o(\delta)$.
- The times between successive increments of the process $N(\cdot)$ are iid $\text{Exp}(\lambda)$ random variables.

The parameter $\lambda \in \mathbb{R}^{>0}$ is called the *rate* of the poisson process. *Poisson Processes* are continuous time markov chains.

Example 2.2 - Poisson Process

Counting the number of cars which have passed a given point over time.

Proposition 2.10 - Properties of Poisson Processes

Define $\{N(t)\}_{t \in \mathbb{R}^+}$ to be a *Poisson Process* with rate λ . Then the following properties hold

- i). The counting process $\{N(\beta t)\}_{t \in \mathbb{R}^+}$, with $\beta > 0$, is a Poisson Process with rate $\beta\lambda$.
- ii). If $\{N_1(t)\}_{t \in \mathbb{R}^+}$ and $\{N_2(t)\}_{t \in \mathbb{R}^+}$ are independent poisson processes with rates λ_1 and λ_2 , respectively, then $\{N(t) := N_1(t) + N_2(t)\}_{t \in \mathbb{R}^+}$ is a poisson process with rate $\lambda := \lambda_1 + \lambda_2$.
- iii). Let X_1, X_2, \dots be a sequence of iid $\text{Bern}(p)$ random variables, independent of $N(\cdot)$. Define $N_1(t) := \sum_{i=1}^{N(t)} X_i$ and $N_2(t) := \sum_{i=1}^{N(t)} (1 - X_i)$ (These are called *Bernoulli Thinnings*). These assign increments in $N(\cdot)$ randomly to either N_1 or N_2 (with probability p). Then, $N_1(\cdot)$ and $N_2(\cdot)$ are independent poisson processes with rates λp and $\lambda(1 - p)$, respectively.

2.3 Transformation of Random Variables

Example 2.3 - Discrete Case

Consider rolling a fair die where $\Omega := \{1, 2, 3, 4, 5, 6\}$ and $\forall \omega \in \Omega \quad \mathbb{P}(\omega) = \frac{1}{6}$.

Let $X(\omega) = \omega \forall \omega \in \Omega$ so the pmf of X is given by

$$p_X(i) = \frac{1}{6} \forall i \in \{1, \dots, 6\}$$

Consider $Y := X^2$. The pmf of Y is straightforward to work out as each value of X maps to a unique value of Y

$$P_Y(i) = \frac{1}{6} \text{ for } \sqrt{i} \in \{1, \dots, 6\}$$

Consider $Z := (X - 2)^2$. This is not quite so simple as multiple values of X map to the same value of Z .

$$p_Z(1) = \frac{2}{6}; \quad p_Z(i) = \frac{1}{6} \text{ for } i \in \{0, 4, 9, 16\}$$

Example 2.4 - Continuous Case

Let $X \sim \text{Uniform}[0, 1]$ and define $Y := 2X$. We have

$$f_X(x) = \begin{cases} 1 & x \in [0, 1] \\ 0 & \text{otherwise} \end{cases} \quad F_X(x) = \begin{cases} 0 & x < 0 \\ x & x \in [0, 1] \\ 1 & x > 1 \end{cases}$$

Now consider Y the cdf is

$$F_Y(y) := \mathbb{P}(Y \leq y) = \mathbb{P}(2X \leq y) = \mathbb{P}\left(X \leq \frac{y}{2}\right) = F_X\left(\frac{y}{2}\right)$$

We then obtain the pdf for Y by differentiation and the chain rule.

$$f_Y(y) = F'_Y(y) = \underbrace{\frac{1}{2} F'_X\left(\frac{y}{2}\right)}_{\text{chain rule}} = \frac{1}{2} f_X\left(\frac{y}{2}\right) = \begin{cases} \frac{1}{2} & \frac{y}{2} \in [0, 1] \\ 0 & \text{otherwise} \end{cases}$$

Proposition 2.11 - Increasing Functions

Let X be a random variable and $Y := g(X)$ where $g : \mathbb{R} \rightarrow \mathbb{R}$ is a strictly increasing function. Then, g is invertible on its range (denoted g^{-1}). Thus the cdfs for X and Y are related as

$$F_Y(y) = \underbrace{\mathbb{P}(g(X) \leq y)}_{\text{as } g \text{ is increasing}} = \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

Differentiating this gives us the pdfs

$$f_Y(y) = f_X(g^{-1}(y)) \frac{dg^{-1}(y)}{dy} = \frac{f_X(x)}{g'(x)} \Big|_{x=g^{-1}(y)}$$

The probability mass of X in the interval $(x, x + dx)$ gets mapped to the interval $(g(x), g(x + dx))$. By *Taylor Expansion* of g we get $g(x + dx) \simeq g(x) + g'(x)dx$

Proposition 2.12 - General Mappings - Single Random Variable

Let X be a (discrete) random variable and define $Y := g(x)$ for any $g : \mathbb{R} \rightarrow \mathbb{R}$ which is differentiable.

There is a contribution $\frac{f_X(x)}{|g'(x)|}$ from each x such that $g(x) = y$. Where $g'(x)$ is positive or negative does not matter, as it only determines where the pre-image of $(y, y + dy)$ is of the form $(x, x + dx)$ or $(x - dx, x)$. Only the relative widths of the intervals matters for the contribution.

By summing the contributions from all solutions of $g(x) = y$ we get

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

This formula is valid so long as the set $\{x : g(x) = y\}$ is countable. If it is not countable, then Y is a continuous RV (or is a mixed random variable)

Proposition 2.13 - General Mappings - Random Vectors

Consider the random vector $\mathbf{X} := (X_1, \dots, X_d)$ with joint density $f_{\mathbf{X}}$ and define $\mathbf{Y} := g(\mathbf{X})$ for any differentiable $g : \mathbb{R}^d \rightarrow \mathbb{R}^d$.

Let $J_g(\mathbf{x})$ denote the *Jacobian* of g at \mathbf{x} . Then

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

Let \mathbf{x} solve $g(\mathbf{x}) = \mathbf{y}$. In a neighbourhood of \mathbf{x} , g is approximately a linear function. By *Taylor Expansion*

$$g(\mathbf{x}') \simeq f(\mathbf{x}) + J_g(\mathbf{x})(\mathbf{x}' - \mathbf{x}) = \mathbf{y} + J_g(\mathbf{x})(\mathbf{x}' - \mathbf{x})$$

for \mathbf{x}' in a small enough neighbourhood of \mathbf{x} .

0 Reference

Definition 0.1 - Stochastic Matrix

A matrix is called a *Stochastic matrix* if:

- i). All elements are non-negative.
- ii). All rows sum to 1.

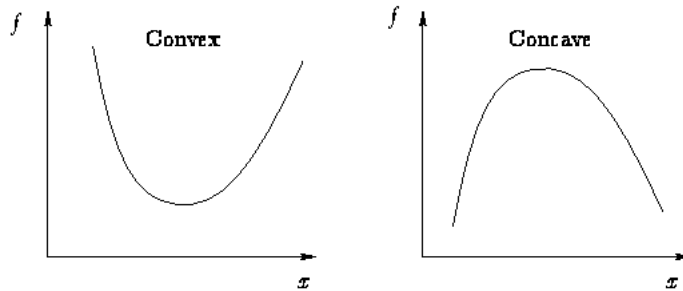
Definition 0.2 - Convex Function

A function $f : \mathbb{R} \rightarrow (\mathbb{R} \cup \{+\infty\})$ is *Convex* if, $\forall x, y \in \mathbb{R}, \alpha \in [0, 1]$, we have

$$f(\alpha x + (1 - \alpha)y) \leq \alpha f(x) + (1 - \alpha)f(y)$$

A smooth function f is convex iff f is twice differentiable and $f''(x) \geq 0 \forall x \in \mathbb{R}$.

Visually, a function is convex if you can draw a line between any two points on the function and the function lies below the line.



Definition 0.3 - Equivalence Relation

A relation is an *Equivalence Relation* if it is

- i). Reflexive: $i \leftrightarrow i$.
- ii). Symmetric: If $i \rightarrow j$ then $j \rightarrow i$.
- iii). Transitive: If $i \rightarrow j$ and $j \rightarrow k$ then $i \rightarrow k$.

Definition 0.4 - Simple Random Walk

A *Simple Random Walk* is a random walk which moves only one step at a time. (i.e. $X_{t+1} = X_t \pm 1$). A probability p is defined for $\mathbb{P}(X_{t+1} = X_t + 1)$, this means $1 - p$ is the probability of stepping in the other direction. A *Simple Random Walk* is *Assymmetric* if $p \neq 1 - p$.

Definition 0.5 - Matrix Exponential, e^X

Let X be a matrix then we define the *Matrix Exponential* as $e^X := I + X + \frac{X^2}{2!} + \dots$

Theorem 0.1 - Pinsker's Inequality

For two distributions $\text{Bern}(p)$ and $\text{Bern}(q)$

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

Definition 0.6 - Jacobian Matrix

The *Jacobian Matrix* is the first-order partial-derivatives of a multidimensional function $\mathbf{f} : \mathbb{R}^n \rightarrow \mathbb{R}^m$ wrt each parameter.

$$J_f := \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{pmatrix}$$

0.1 Notation

$p_{x,y}$	$\mathbb{P}(X_{t+1} = x X_t = y)$ for a <i>time homogenous markov process</i> .
$I(t)$	Arm played in round t .
$N_i(t)$	$\sum_{s=1}^t \{I(s) = i\}$ Number of times arm i was played in first t rounds.
$S_i(t)$	$\sum_{s=1}^t X_i(s) \{I(s) = i\}$ Total reward from arm i in first t rounds.
$\hat{\mu}_{i,N_i(t)}$	$\frac{S_i(t)}{N_i(t)}$ sample mean reward from arm i in first t rounds.
Δ_i	$\mu^* - \mu_i$ the arm gaps from a K -armed bandit.
$X_i(t)$	RV modelling the result if arm i was played on round t .

0.1.1 Asymptotic Notation

Definition 0.7 - Oh Notation

Let $f, g : \mathbb{R}^+ \rightarrow \mathbb{R}$.

We say $f = o(g)$ (little oh of g) at 0 if $\frac{f(x)}{g(x)} \xrightarrow{x \rightarrow 0} 0$.

We say $f = O(g)$ (big oh of g) at 0 if $\exists c > 0$ st $|f(x)| \leq c|g(x)|$ in a neighbourhood of 0.
 $f = o(g)$ at infinity and $f = O(g)$ at infinity are defined analogously.

Definition 0.8 - Omega Notation

Let $f, g : \mathbb{R}^+ \rightarrow \mathbb{R}$.

We say $g = \omega(f)$ if $o(f)$ and we say $f = \Omega(g)$ if $g = O(f)$.

Example 0.1 - Oh & Omega Notation

Define $f(x) = x$, $g(x) = \sin(x)$ and $h(x) = x^2$.

Then, $g = O(f)$ at 0 and $g = o(f)$ at infinity. $h = o(f)$ at 0 and $h = \omega(f)$ at infinity.