Stochastic Optimisation - Reviewed Notes

Dom Hutchinson

November 11, 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.

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

□

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

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

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

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

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

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

The result of the theorem. □

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

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

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

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

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

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. □

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 $n\in\mathbb{N},\ \alpha\in(0,1)$.

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

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 $p \in [0,1]$

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

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

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

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

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

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 $UCB(\alpha, i)$.

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

(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 $s + 1$. □

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

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

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

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

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 $u_{t,i} := \left\lceil \dfrac{2\alpha \ln(t)}{\Delta_i^2} \right\rceil$. 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 $u_{t,i}$, <u>if</u> $I(s+1) = i$ *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

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

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

Bernoulli Arms

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 Beta$(\alpha+1,\beta)$.
- Else: Set posterior to 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.

In the case of Bernoulli arms, Beta priors are conjugate. (See `Theorem2.5`)

Theorem 2.5 - *Beta 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. Then

$$\pi_1(\mu|x)\sim\begin{cases}\text{Beta}(\alpha+1,\beta) & \text{if } x=1\\ \text{Beta}(\alpha,\beta+1) & \text{if } x=0\end{cases}$$

Proof 2.4 - *Theorem 2.5*
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

$$\pi_1(\mu|x)\sim\begin{cases}\text{Beta}(\alpha+1,\beta) & \text{if } x=1\\ \text{Beta}(\alpha,\beta+1) & \text{if } x=0\end{cases}$$

□

TODO move this theorem somewhere relevant.
Theorem 2.6 - *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 2.5 - *Theorem 2.6*
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. □

Poisson Observations

Gaussian Observations

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 2.7 - *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 2.6 - *Theorem 2.7*
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

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

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 `Theorem 2.5`, 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 2.6`, 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.

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

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}.

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

Consider the following restatement of event B_t

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

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 $N_2(t)$

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

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 2.7` □

Proposition 2.8 - *Bound of Regret*
Using `Theorem 2.7` 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 2.8 - *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 2.9 - *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 2.7 - *Theorem 2.9*
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

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

the last step occurs because [1]

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

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

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

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

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

Next, we note that conditional on $S_2(t)$ and $N_2(t)$, by `Theorem 2.5` 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 `Theorem 2.6`, 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,

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

we can derive an explicit upper-bound on the probability.

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

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

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

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 2.9` □

0 Reference

0.1 Notation

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.
$N_j(t)$	$:=$	$\displaystyle\sum_{s=1}^{t} \mathbb{1}(I(s) = j)$	The number of times arm j has been played in the first t rounds.
$S_j(t)$	$:=$	$\displaystyle\sum_{s=1}^{t} X_j(s)\mathbb{1}(I(s) = j)$	The total reward from arm j in the first t rounds.
$\hat{\mu}_{j,n}$	$:=$	$\dfrac{S_j(t)}{N_j(t)}$	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.

0.2 Definition

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 $[0, t]$