Exploration vs Exploitation in Stationary Multi-Armed Bandit Problems

Kun Joo Michael Ang

July 6, 2021

Abstract

This paper examines the multi-armed bandit problem in the case where the bandits' rewards are drawn from stationary but unknown distributions. Unlike the classical problem, players must factor in the informational value of each future sample to balance exploration against exploitation. Using no distributional assumptions, we derive some properties of the optimal strategy, using the notion of a bandit's fair-value. Then, by restricting the class of possible distributions, we cast the problem in a Bayesian framework and find complete solutions through dynamic programming and an updating prior. We find upper and lower bounds for each bandit's fair value and in the special case of a normal distribution, explicit formulae and numerical simulations are computed.

Keywords— multi-armed bandit problems, stationary unknown distribution, exploration, exploitation, dynamic programming, Bayesian inference, improper prior, one-step lookahead policy, fair-value

1 Introduction

Consider the problem of a player deciding between M one-armed bandits over T rounds. Each bandit yields a reward drawn from the distribution $f_i(x)$ for $i = 1, 2, ..., M$. The distributions f_i maybe be discrete or continuous and are hidden from the player. The player wishes to maximize his total expected reward at so over the T rounds, he must decide between **exploration** (choosing bandits to reduce the uncertainty of $\mathbb{E}[f_i]$) and **exploitation** (choosing the bandit with maximum expected reward conditional on current information). In this particular problem, the distributions f_i are stationary in time, though in general, they do not have to be.

This is a popular problem in reinforcement learning that is typically solved with approximate algorithms: ϵ-greedy algorithms, Gradient Bandit Algorithms, Upper Confidence Bound Action selection, etc) [1]. The first two are probabilistic algorithms where the decision taken at each step is random, and the third is deterministic. In several of these applications, a pre-defined learning rate needs to be specified, which controls how quickly the algorithm updates the estimated value of each bandit. Another common drawback of these algorithms is that they often do not take into account

the number of rounds remaining relative to the $\mathbb{E}[f_i]$. Indeed, for the ϵ-greedy algorithm, once we have low-variance estimates of $\mathbb{E}[f_i]$, we it makes sense to switch to exploitation by reducing ϵ, but what counts as sufficiently 'low-variance' should be higher if there are more rounds remaining. The UCB algorithm similarly does not take into account the value of exploration relative to the number of remaining rounds, and only computes a reasonable upper bound for $\mathbb{E}[f_i]$ relative to the current information.

We can use optimization and statistical theory to better quantify exactly when to switch from exploration to exploitation.

2 Methodology

2.1 Fair Value Bandit

Before we dive into an exact mathematical formulation of the problem, it is helpful to introduce the notion of a bandit's fair value. Imagine we are in a situation where we have exactly two bandits, A and B. A always gives a fixed constant reward λ, while B gives rewards drawn from a distribution $f(X)$. If f is known to the player, then $\lambda = \mathbb{E}[f]$ corresponds to the constant reward that would make the player indifferent between choosing either bandit. We say that bandit B has fair value λ at time $t < T$. It is easy to see that if we know the complete distribution of f, then if we are indifferent between A and B at some time t, then we remain indifferent at all future times $t' > t$. But if the distribution of f is unknown to us, then the result not longer holds. Future realizations of rewards from B might change our estimate of $\mathbb{E}[f]$, making one bandit strictly preferred over the other.

However, in such a scenario, there is still the notion of *weak indifference*, where we are indifferent between A and B at some time t. Furthermore, we can also show that for all times $t < t' \leq T$, it is never better to switch to B if your previous choice was A. To formalize this, let's consider a filtration $\mathcal{F}_t := \sigma(X_i \,|\, i < t)$ where X_i denotes the reward received from the bandit chosen at time i. Suppose now we have a strategy X which chooses A at time t, then switches to B at $t+1$. We also have an alternate strategy $\tilde{X}$ which follows X for $t' < t$, but chooses B and A at times t, $t+1$ respectively. Immediately after, $\tilde{X}$ will mirror the decision X under the filtration $\sigma(X_1, ..., X_{t-1}, \tilde{X}_t, \tilde{X}_{t+1})$. The stationarity of the distribution means that

$$\mathbb{E}\left[\mathbb{E}\left[\sum_{i=t+2}^{T} X_i \middle| \mathcal{F}_{t+2}\right] \middle| \mathcal{F}_t\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{i=t+2}^{T} \tilde{X}_i \middle| \tilde{\mathcal{F}}_{t+2}\right] \middle| \mathcal{F}_t\right]$$

as the decisions made at $t' \geq t+2$ have the same number of realizations of f, and the information is identical at $t' = t$. Also, since X and $\tilde{X}$ both have exactly one realization from A and B, their total expectations from t to $t+1$ are equal. This shows that if a strategy calls for switching from A to B, it is equally as good to reverse the order, selecting B then A, before returning to the original strategy.

In fact, it is better to select B first, since $\max(\lambda, \mathbb{E}[f|\tilde{\mathcal{F}}_{t+1}]) \geq \lambda$. At time $t+1$, we always have the option of choosing between A or B, which dominates a fixed strategy. If we choose B, we are also not worse off at future times $t' \geq t+2$, as we can always

ignore the information gained from the extra realization. This leads to the intuitive conclusion that the value of each state is a function of only the time remaining and the prior (unordered) samples of B. Furthermore, with the same time remaining, the total expected future reward increases with the number of prior samples of B. Note that this result holds true in expectation over the samples, and not for specific realizations of B.

The idea of a fair value bandit offering constant reward gives us a way of comparing bandits in the multi-armed problem. At any time t, each bandit i has an associated unique fair value $\lambda_i(t, \mathcal{F}_t)$. If a bandit offering constant reward λ_i for all future rounds was added at time t, we would be indifferent between i and this new bandit. The best bandit to choose at time t is therefore the one with the highest fair value. Note that these λ_i need to be reevaluated at each timestep, even if no new information is gained about i. Similar to an option in the financial markets, the time value of exploration is lost with each round. Indeed, if the information was exactly the same at an early round t' and a later round t, the player at t' always has the option of picking the constant reward up to time t. However, because he also has the option (but not the obligation) to pick i and potentially gain more information, a higher constant reward is needed to achieve fair value. We can therefore focus exclusively on the problem of finding a bandit's fair value and this will generalize to the solution of the stationary multi-armed case.

2.2 Distributional Assumptions

Any strategy X will depend on the underlying distributional assumptions placed on f_i, but there are a few general results that hold.

The Central Limit Theorem tells us that $X_1, X_2, ...X_N$ drawn from distribution f with finite first and second moments $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathbb{E}[(X - \mu)^2]$ satisfies

$$\frac{\hat{\mu} - \mu}{\sigma/\sqrt{N}} \xrightarrow{d} \mathcal{N}(0, 1)$$

where $\hat{\mu} = \frac{1}{N}\sum_i X_i$. We can also replace σ by its estimator $s = \sqrt{\frac{1}{N-1}\sum_i (X_i - \hat{\mu})^2}$

$$\frac{\hat{\mu} - \mu}{s/\sqrt{N}} \xrightarrow{d} t_N$$

where t_N denotes a t-distribution with N degrees of freedom. This allows us to create confidence intervals around μ. Although for any specific set of realizations we cannot infer anything about the distribution of μ (as it is some value with probability 1 and 0 everywhere else), for most practical intents, it serves as a good prior and is our best estimate if we had to assign a distribution to μ.

With this distribution as a prior, we can formulate a simple greedy strategy. In the fair value bandit problem, we pick bandit B if $\mathbb{E}_\mu[X] \geq \lambda$. Additionally, $\mathbb{E}_\mu[X] = \mathbb{E}[\mu] = \hat{\mu}$ by symmetry of the t-distribution. Translated to the multi-armed bandit problem, the greedy strategy is to always select the bandit with the highest running mean given the current information set. Without making any assumptions on the distribution of f, this simple strategy is always optimal at T, but not in earlier rounds.

But to do better than a greedy strategy, we need to be able to model how $\hat{\mu}$ changes and how we expect it to change in subsequent rounds based on subsequent draws. This is most easily done in a Bayesian setting, where we have a distributional assumption on X, say $X \sim \mathcal{N}(\mu, \sigma^2)$ with parameters in a domain, say $\Theta : \{\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+\}$. If we start with some prior over the distribution parameters θ, then Bayes rules allows us to update the density of θ, $h(\theta)$, conditional on the available information in each round. The key advantage here is that the prior allows us to simulate subsequent draws of X using the best information we have at each timestep, which in turn allows us to compute the expected value of future states. The prior density $h_0(\theta)$ updates as follows:

$$h(\theta \,|\, \mathcal{F}_k) = h_k(\theta) = \begin{cases} A f(X_k|\,\theta)\, h_{k-1}(\theta) & \text{if } X_k \in B \\ h_{k-1}(\theta) & \text{otherwise} \end{cases}$$

Here A is a constant of proportionality that ensures $h_k(\theta)$ remains a valid probability density. $h_k(\theta)$ is only updated when new samples are drawn from bandit B. However, this Bayesian formulation is not without its drawbacks. The initial prior $h_k(\theta)$ needs to be specified as well as integrable and this might not always be possible.

2.3 Improper Prior for Gaussian Distribution

Returning to the example of $X \sim \mathcal{N}(\mu, \sigma^2)$ with the domain of parameters $\Theta : \{\theta = (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+$. Assume that all samples X_i are drawn from B and we have observed samples $X_1, X_2, ..., X_k$. An improper uniform prior at $T = 0$ has $h_0(\theta) \equiv c$ and

$$\begin{aligned} h_k(\theta) &= A_k f(X_k|\,\theta)\, h_{k-1}(\theta) \\ &= A_k \prod_{i=1}^{k} f(X_i|\,\theta) \\ &= A_k \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(X_i-\mu)^2}{2\sigma^2}\right\} \\ &= A_k \left(2\pi\sigma^2\right)^{-\frac{k}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}(X_i-\mu)^2\right\} \end{aligned}$$

and we can verify that this is indeed a valid probability density.

$$\begin{aligned} A_k^{-1} &= \int_0^\infty \int_{-\infty}^\infty \left(2\pi\sigma^2\right)^{-\frac{k}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}(X_i-\mu)^2\right\} d\mu\, d\sigma \\ &= \int_0^\infty \left(2\pi\sigma^2\right)^{-\frac{k-1}{2}} \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{k}(X_i-\mu)^2\right\} d\mu\, d\sigma \\ &= \int_0^\infty \left(2\pi\sigma^2\right)^{-\frac{k-1}{2}} \exp\left\{\frac{\frac{1}{k}\left(\sum X_i\right)^2 - \sum X_i^2}{2\sigma^2}\right\} \cdot \\ &\qquad \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{k\left(\mu - \frac{1}{k}\sum X_i\right)^2}{2\sigma^2}\right\} d\mu\, d\sigma \\ &= \frac{1}{\sqrt{k}} \int_0^\infty \left(2\pi\sigma^2\right)^{-\frac{k-1}{2}} \exp\left\{\frac{\frac{1}{k}\left(\sum X_i\right)^2 - \sum X_i^2}{2\sigma^2}\right\} d\sigma \end{aligned}$$

An application of the Cauchy-Schwarz inequality gives

$$r_k^2 := \sum X_i^2 - \frac{1}{k}\left(\sum X_i\right)^2 \geq 0$$

and performing the substitution $y = \frac{r_k^2}{2\sigma^2}$ gives

$$\begin{aligned} A_k^{-1} &= \frac{1}{\sqrt{k}} \int_0^\infty \left(2\pi\sigma^2\right)^{-\frac{k-1}{2}} \exp\left\{-\frac{r_k^2}{2\sigma^2}\right\} d\sigma \\ &= \frac{r_k^{2-k}}{\sqrt{8k \cdot \pi^{k-1}}} \int_0^\infty y^{\left\{\frac{k-2}{2}-1\right\}} e^{-y} dy \\ &= \frac{r_k^{2-k}}{\sqrt{8k \cdot \pi^{k-1}}} \cdot \Gamma\left(\frac{k-2}{2}\right) \end{aligned}$$

which is finite for all $k > 2$. As an additional remark, the higher dimensional Bernstein-von Mises theorem tells us that the posterior distribution over θ will converge to a Gaussian distribution centered on the Maximum Likelihood Estimator $(\hat{\mu}, \hat{\sigma})$. Either by using this result with the symmetry of the Gaussian distribution, or directly integrating over $h_{T-1}(\theta)$, we can formulate our strategy for the final round.

$$\begin{aligned} \mathbb{E}\Big[X \,\Big|\, \mathcal{F}_T\Big] &= \mathbb{E}\Big[X \,\Big|\, h_{T-1}(\theta)\Big] \\ &= \int \mathbb{E}_{\mu,\theta}\Big[X\Big] h_{T-1}(\theta) d\theta \\ &= A_{T-1} \int\int \mu \prod_{i=1}^{T-1} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(X_i-\mu)^2}{2\sigma^2}\right\} d\mu d\sigma \\ &= A_{T-1} \frac{1}{\sqrt{T-1}} \int \left[\frac{1}{T-1} S_{T-1}\right] \left(2\pi\sigma^2\right)^{-\frac{T-2}{2}} \exp\left\{-\frac{r_{T-1}^2}{2\sigma^2}\right\} d\sigma \\ &= \frac{S_{T-1}}{T-1} \end{aligned}$$

where for notational convenience we define

$$S_k = \sum_{i=1}^k X_i \qquad\qquad S_k^2 = \sum_{i=1}^k X_i^2$$

Our strategy at T is therefore to pick bandit A whenever $\lambda \geq \mathbb{E}[X \,|\, \mathcal{F}_T]$ and bandit B otherwise. Using V_k to denote the value of the sum of expected future rewards (following an optimal strategy) at time k, we have

$$\begin{aligned} V_T(X_1, ..., X_{T-1}) &= \mathbb{E}\left[\lambda \cdot \mathbb{1}_{\{\lambda \geq \mathbb{E}[X|\mathcal{F}_T]\}} + X \cdot \mathbb{1}_{\{\lambda < \mathbb{E}[X|\mathcal{F}_T]\}} \,\middle|\, \mathcal{F}_T\right] \\ &= \max\Big(\lambda,\, \mathbb{E}\big[X \,|\, \mathcal{F}_T\big]\Big) \\ &= \lambda + \left(\frac{S_{T-1}}{T-1} - \lambda\right)_+ \end{aligned}$$

And we are now in good shape to extend our strategy to earlier time steps.

2.4 Strategy at $T-1$

Recall that it is never optimal to switch from bandit A to bandit B, so at time $T-1$, we either pick bandit A twice or pick bandit B and receive $(X+V_T)$.

$$\begin{aligned}
V_{T-1}(\mathcal{F}_{T-1}) &= 2\lambda \cdot \mathbb{1}_{\{2\lambda \geq \mathbb{E}[X+V_T|\mathcal{F}_{T-1}]\}} + \mathbb{E}\left[X+V_T|\mathcal{F}_{T-1}\right] \cdot \mathbb{1}_{\{2\lambda \geq \mathbb{E}[X+V_T|\mathcal{F}_{T-1}]\}} \\
&= \max\left(2\lambda,\ \mathbb{E}\Big[X+V_T\ \Big|\ \mathcal{F}_{T-1}\Big]\right) \\
&= \max\left(2\lambda,\ \ \frac{S_{T-2}}{T-2}+\lambda+\mathbb{E}\left[\left(\frac{S_{T-1}}{T-1}-\lambda\right)_+\ \middle|\ \mathcal{F}_{T-1}\right]\right) \\
&= 2\lambda+\left(\frac{S_{T-2}}{T-2}-\lambda+\mathbb{E}\left[\left(\frac{S_{T-2}+X}{T-1}-\lambda\right)_+\ \middle|\ \mathcal{F}_{T-1}\right]\right)_+ \\
&= 2\lambda+\left(\frac{S_{T-2}}{T-2}-\lambda+\frac{1}{T-1}\mathbb{E}\left[\Big(X-\big\{\lambda(T-1)-S_{T-2}\big\}\Big)_+\ \middle|\ \mathcal{F}_{T-1}\right]\right)_+
\end{aligned}$$

The value of each state can thus be interpreted as a stream of fixed rewards λ plus a sequence of nested call options on the values of future states. The fair value bandit is the smallest λ^* that makes the option value 0. At time T, λ^* is the sample mean, but at time $T-1$, because the value of the nested option is non-negative, λ^* must be equal to the sample mean or greater. In fact, λ^* will be strictly greater than the sample mean if the distribution of X is continuous with infinite support, as there will be some strictly positive value on the option. One such example is when X is normally distributed.

We can also go a step further to explicitly compute the distribution of X and hence the option value under the $\mathcal{F}_{T-1}$ filtration.

$$\begin{aligned}
f_{T-1}(X) &= \int\int \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{(X-\mu)^2}{2\sigma^2}\right\} h_{T-2}(\mu,\sigma)d\theta d\sigma \\
&= \int\int A_{T-2}\left(2\pi\sigma^2\right)^{-\frac{T-1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\cdot\left[(X-\mu)^2+\sum_{i=1}^{T-2}(X_i-\mu)^2\right]\right\} d\theta d\sigma \\
&= \frac{A_{T-2}}{A^*_{T-1}} \\
&= \frac{{r^*_{T-1}}^{3-T}}{r^{4-T}_{T-2}} \cdot \frac{\Gamma\left(\frac{T-3}{2}\right)}{\Gamma\left(\frac{T-4}{2}\right)} \cdot \sqrt{\frac{T-2}{T-1}} \cdot \sqrt{\pi} \\
&\propto \left(\frac{1}{\sqrt{S^2_{T-2}+X^2+\frac{1}{T-1}\left(S_{T-2}+X\right)^2}}\right)^{T-3} \\
&= \gamma\left(\frac{1}{\sqrt{S^2_{T-2}+\frac{1}{T}\left(S_{T-2}\right)^2+\frac{T}{T-1}\left(X+\frac{1}{T}S_{T-2}\right)^2}}\right)^{T-3}
\end{aligned}$$

where A_T^* and r_T^* are computed with X_T as X and γ is a normalizing constant. λ^* is therefore implicit solution to the following equation.

$$\begin{aligned}\lambda - \frac{S_{T-2}}{T-2} &= \frac{1}{T-1}\mathbb{E}\left[\Big(X - \big\{\lambda(T-1) - S_{T-2}\big\}\Big)_+ \,\middle|\, \mathcal{F}_{T-1}\right] \\ &= \frac{1}{T-1}\int f_{T-1}(X)\Big(X - \big\{\lambda(T-1) - S_{T-2}\big\}\Big)_+ dX \\ &= \frac{1}{T-1}\int_{\lambda(T-1)-S_{T-2}} f_{T-1}(X)\Big[X - \lambda(T-1) + S_{T-2}\Big]\, dX\end{aligned}$$

Rather fortunately, there exists a useful analytical result that makes calculating the integrals easier, Unfortunately, the result is a recurrence relation in T and does not have a simple closed form.

Consider the substitutions

* $A = \sqrt{S_{T-2}^2 + \frac{1}{T}\left(S_{T-2}\right)^2}$
* $B = \sqrt{\frac{T}{T-1}}$
* $C = \frac{1}{T}S_{T-2}$
* $N = T - 3$
* $Z = \lambda(T-1) - S_{T-2}$
* $\delta = \frac{1}{T-2}S_{T-2}$

Then we can solve

$$\tan\theta = \frac{B(u+C)}{A}$$

$$\begin{aligned}\int^X \left(\frac{1}{\sqrt{A^2 + B^2(u+C)^2}}\right)^N du &= \int^{\tan^{-1}\left(\frac{B(X+C)}{A}\right)} \frac{A}{B}\sec^2\theta\left(\frac{1}{A\sqrt{1+\tan^2\theta}}\right)^N d\theta \\ &= \frac{A^{N-1}}{B}\int^{\tan^{-1}\left(\frac{B(X+C)}{A}\right)} \cos^{(N-2)}\theta\, d\theta\end{aligned}$$

and by applying the recurrence relation

$$\int \cos^n\theta' d\theta' = \frac{1}{n}\cos^{n-1}\theta\sin\theta + \frac{n-1}{n}\int \cos^{n-2}\theta' d\theta'$$

this gives us

$$\gamma^{-1} = \begin{cases} \frac{A^{N-1}}{B}\cdot 2\left[\frac{N-3}{N-2}\cdot\frac{N-5}{N-4}\cdots\frac{2}{3}\right] & N \text{ odd} \\ \frac{A^{N-1}}{B}\cdot \pi\left[\frac{N-3}{N-2}\cdot\frac{N-5}{N-4}\cdots\frac{1}{2}\right] & N \text{ even}\end{cases}$$

And we resolve the remaining half of the call with the integral

$$\begin{aligned}\int^X u\left(\frac{1}{\sqrt{A^2 + B^2(u+C)^2}}\right)^N du &= \frac{1}{B^2\cdot(2-N)}\Big(A^2 + B^2(X+C)^2\Big)^{\frac{2-N}{2}} \\ &\quad - C\int^X \left(\frac{1}{\sqrt{A^2 + B^2(u+C)^2}}\right)^N du\end{aligned}$$

The final implicit solution looks like

$$\Big[Z-\delta\Big] = \frac{\gamma}{B^2(N-2)}\Big(A^2+B^2(Z+C)^2\Big)^{\frac{2-N}{2}} + \gamma\Big(Z+C\Big)\frac{A^{N-1}}{B}\left[\frac{1}{N-2}\cos^{N-3}\theta\sin\theta + \frac{N-3}{N-2}\frac{1}{N-4}\cos^{N-5}\theta\sin\theta+\cdots\right]_{\tan^{-1}\left(\frac{B(Z+C)}{A}\right)}^{\frac{\pi}{2}}$$

To propagate this strategy forward yet another time step, note that we can write $V_{T-1}(\mathcal{F}_{T-1}) = 2\lambda^*_{T-1}(X_1,...,X_{T-2})$ and similarly $V_{T-2} = 3\lambda^*_{T-2}$

$$\begin{aligned} V_{T-2}(\mathcal{F}_{T-2}) &= \max\left(3\lambda_{T-2},\, \mathbb{E}\Big[X+V_{T-1}\,\Big|\,\mathcal{F}_{T-2}\Big]\right) \\ &= \max\left(3\lambda_{T-2},\, \frac{S_{T-2}}{T-2}+2\,\mathbb{E}\Big[\lambda^*_{T-1}\,\Big|\,\mathcal{F}_{T-2}\Big]\right) \\ &= \max\left(3\lambda_{T-2},\, \frac{S_{T-2}}{T-2}+2\int f_{T-2}(X)\lambda^*_{T-1}dX\right) \\ \implies \lambda^*_{T-k} &= \frac{1}{k}\left[\frac{S_{T-k}}{T-k}+(k-1)\int \lambda^*_{T-k+1}f_{T-k}(X)dX\right] \end{aligned}$$

and this creates a recurrence relation of functions in λ^*, allowing us to solve for each bandit, a fair value $\lambda^*_t(\mathcal{F}_t)$ at time t. Of course, computing each λ^*_t is computationally very expensive, as integrals need to be performed over all combinations of $X_{t+1},...,X_T$.

3 Upper and Lower Bounds

3.1 Upper bounds λ^k_U

Explicit computation of λ^* is expensive, but we can extract an upper bound λ^1_U by adopting a sub-optimal strategy. Assume that $t<T$ and place the following restriction: Free choice is given between bandit A and B for one round, but immediately after a decision must be made between A or B and that will be the chosen bandit for all remaining rounds. By removing optionality, this strategy is inferior to the one before, and we require a higher constant reward $\lambda^1_U \geq \lambda^*$ to be indifferent between A and B initially.

λ^1_U is defined at time $t<T$ by

$$\begin{aligned} \lambda^1_U(T-t+1) &= \mathbb{E}[X|\mathcal{F}_t] + \mathbb{E}\left[(T-t)\cdot\max\Big(\lambda^1_U,\,\mathbb{E}[X|\mathcal{F}_{t+1}]\Big)\Big|\mathcal{F}_t\right] \\ &= \frac{1}{t-1}S_{t-1} + (T-t)\cdot\mathbb{E}\left[\max\Big(\lambda^1_U,\frac{1}{t}\Big[S_{t-1}+X\Big]\Big)\Big|\mathcal{F}_t\right] \\ \implies \lambda^1_U &= \frac{1}{t-1}S_{t-1} + \frac{T-t}{t}\cdot\int_{t\lambda^1_U-S_{t-1}} f_t(X)\Big[X-t\lambda^1_U+S_{t-1}\Big]dX \end{aligned}$$

and we have an implicit solution very similar to our $T-1$ strategy.

Using the following substitutions, with γ defined as before,

* $A = \sqrt{S_{t-1}^2 + \frac{1}{t+1}\left(S_{t-1}\right)^2}$
* $B = \sqrt{\frac{t+1}{t}}$
* $C = \frac{1}{t+1} S_{t-1}$
* $N = t - 2$
* $Z = t\lambda_U^1 - S_{t-1}$
* $\delta = \frac{1}{t-1} S_{t-1}$

$$\frac{1}{\gamma(T-t)}\Big[Z-\delta\Big] = \frac{1}{B^2(N-2)}\Big(A^2 + B^2(Z+C)^2\Big)^{\frac{2-N}{2}} +$$
$$\Big(Z+C\Big)\frac{A^{N-1}}{B}\left[\frac{1}{N-2}\cos^{N-3}\theta\sin\theta + \frac{N-3}{N-2}\frac{1}{N-4}\cos^{N-5}\theta\sin\theta + \cdots\right]_{\tan^{-1}\left(\frac{B(Z+C)}{A}\right)}^{\frac{\pi}{2}}$$

In fact, we can generalize this result. Let λ_U^k denote the strategy of choosing freely between A or B for the next k rounds, after which you must stick to a single bandit for the remaining rounds. The following inequalities must hold.

$$\lambda_U^1 \geq \lambda_U^2 \geq \cdots \geq \lambda_U^{T-t} = \lambda^*$$

3.2 Lower bounds λ_L^k

Similar to our upper bound which was computed from the fair value of a sub-optimal strategy, we compute the lower bound from the fair value of a super-optimal strategy. Assume that at time t, we are have to choose between bandit A and B as usual, but at the end of the round we will additionally be told the true value of μ. This scenario is optimal to before and will require a lower constant reward for fair-value indifference.

$$\begin{aligned}
\lambda_L^1(T-t+1) &= \mathbb{E}[X|\mathcal{F}_t] + \mathbb{E}\left[(T-t)\cdot\max\left(\lambda_L^1,\,\mu\right)\middle|\mathcal{F}_t\right] \\
\implies \lambda_L^1 &= \frac{1}{t-1}S_{t-1} + (T-t)\cdot\mathbb{E}\Big[\Big(\mu-\lambda_L^1\Big)_+\Big] \\
&= \frac{1}{t-1}S_{t-1} + (T-t)\cdot\int\int_{\lambda_L^1}\left(\mu-\lambda_L^1\right)h_{t-1}(\mu,\sigma)\,d\mu d\sigma
\end{aligned}$$

where in the third equality we once again consider the special case of the Gaussian distribution.

$$\begin{aligned}
&\int_{\lambda_L^1}^{\infty}\int_0^{\infty}(\mu-\lambda_L^1)\,h_{t-1}(\mu,\theta)\,d\sigma d\mu \\
&= \int_{\lambda_L^1}^{\infty}(\mu-\lambda_L^*)\,A_{t-1}\int_0^{\infty}\left(2\pi\sigma^2\right)^{-\frac{t-1}{2}}\exp\Big\{-\frac{1}{2\sigma^2}\Big[S_{t-1}^2 - 2S_{t-1}\mu + (t-1)\mu^2\Big]\Big\}\,d\sigma d\mu \\
&= \int_{\lambda_L^1}^{\infty}(\mu-\lambda_L^1)\cdot A_{t-1}\cdot 2^{-\frac{3}{2}}\cdot\pi^{-\frac{t-1}{2}}\cdot\Big[S_{t-1}^2 - 2S_{t-1}\mu + (t-1)\mu^2\Big]^{-\frac{t-2}{2}}\Gamma\left(\frac{t-2}{2}\right)d\mu \\
&= r_{t-1}^{t-3}\cdot\sqrt{\frac{t-1}{\pi}}\cdot\frac{\Gamma\left(\frac{t-2}{2}\right)}{\Gamma\left(\frac{t-3}{2}\right)}\int_{\lambda_L^1}^{\infty}(\mu-\lambda_L^*)\cdot\Big[S_{t-1}^2 - 2S_{t-1}\mu + (t-1)\mu^2\Big]^{-\frac{t-2}{2}}\,d\mu
\end{aligned}$$

Once again make the following substitutions,

* $A = r_{t-1}$
* $B = \sqrt{t-1}$
* $C = \frac{1}{t-1} S_{t-1}$
* $N = t - 2$

and we get an implicit solution very similar in form to λ_U^1:

$$\frac{\lambda_L^1 - C}{r_{t-1}^{t-3} \cdot \sqrt{\frac{t-1}{\pi}} \cdot \frac{\Gamma\left(\frac{t-2}{2}\right)}{\Gamma\left(\frac{t-3}{2}\right)}(T-t)} = \frac{1}{B^2(N-2)}\left(A^2 + B^2(\lambda_L^1 + C)^2\right)^{\frac{2-N}{2}} - \left(\lambda_L^1 + C\right)\frac{A^{N-1}}{B}\left[\frac{1}{N-2}\cos^{N-3}\theta \sin\theta + \cdots\right]_{\tan^{-1}\left(\frac{B(\lambda_L^1 + C)}{A}\right)}^{\frac{\pi}{2}}$$

Similar to λ_U^k, we generalize this strategy by denoting it λ_L^1, where λ_L^k refers to the super-optimal strategy of playing optimal under the condition that μ will be revealed after k additional rounds. It is assumed here that μ is a fixed but unknown value, drawn from a distribution proportional to its likelihood estimate at time t. The following inequalities must hold.

$$\lambda^* \geq \lambda_L^{T-t+1} \geq \cdots \geq \lambda_L^2 \geq \lambda_L^1$$

3.3 Tight lower bound λ_L^*

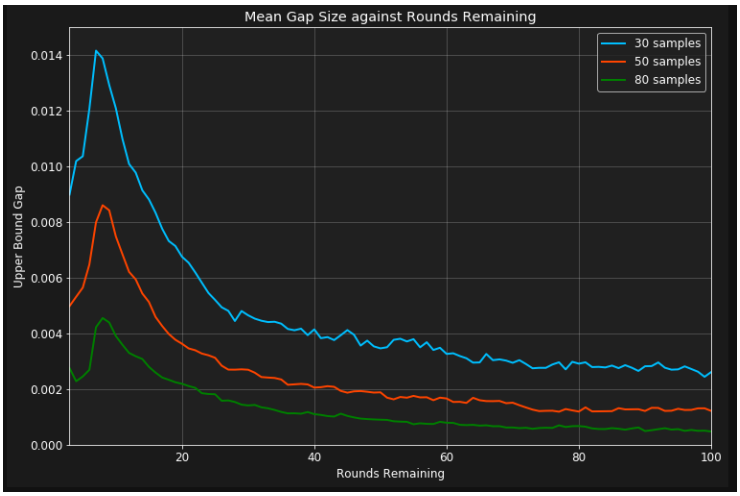
Although the formulation of λ_L^k as a sequence of increasing lower bounds is aesthetically appealing, the current sample mean $\lambda_L^* = \frac{1}{t-1} S_{t-1}$, is often a tighter lower bound in practice. Assuming there are sufficient samples to apply the Central Limit Theorem, the distribution of the true mean is Gaussian and centered about the sample mean. This means the expected value of all future draws from bandit B is equal to λ_L^* under the $\mathcal{F}_t$ filtration, and consequently $\lambda_L^k \leq \lambda_L^*$.

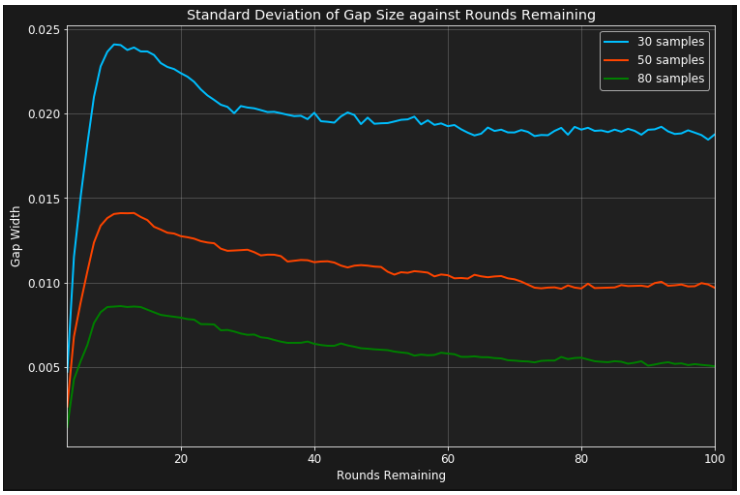
4 Simulated Results

We are interested in estimating λ^* as well as the uncertainty of our estimate. From the above results, two quantities that are easy to compute: λ_L^*, the sample mean, and λ_U^1, the upper bound given we must make a permanent choice after the next step. Given $\lambda^* \in [\lambda_L^*, \lambda_U^1]$, we define the gap size to be $\lambda_U^1 - \lambda_L^*$. A smaller gap is desirable corresponds to greater precision in estimating λ^*.

The following experiments are conducted where samples are independently drawn from an $N(0, 1)$ distribution. The two graphs below display the gap size for 30, 50 and 80 observed samples as a function of the number of remaining rounds. Taken over 1000 realizations, the first graph displays the mean gap size and the second displays the standard deviation of the gap size. At the bottom of the section, we also include a

third graph of the same experiment conducted over a wider range of numbers of samples.





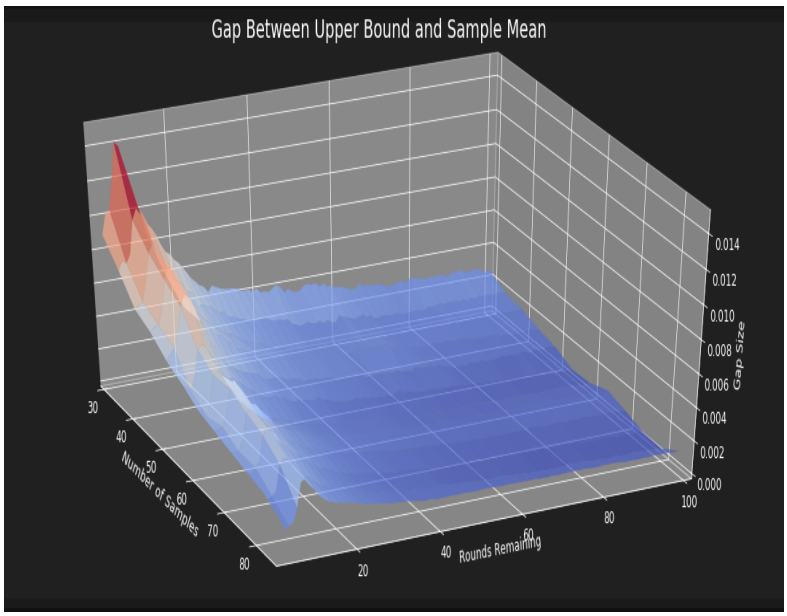
As we expect, having more samples results in a smaller mean gap size, as there is less uncertainty around the value of the bandit's true mean. But perhaps not so intuitively, we find that as the number of remaining rounds increases, the gap size increases sharply to a peak before decreasing quickly and then decaying exponentially. This can be explained from the implicit form of λ_U^1.

$$\lambda_U^1 = \frac{1}{t-1} S_{t-1} + \frac{T-t}{t} \cdot \int_{t\lambda_U^1 - S_{t-1}} f_t(X) \Big[X - t\lambda_U^1 + S_{t-1} \Big] dX$$

Holding everything constant, increasing $T - t$ on the right side increases λ_U^1 on the left side. But this will in turn decrease the value of the integral by increasing its lower

limit. Because we have set up X to follow a normal distribution with parameters drawn from another normal distribution, $f_t(X)$ has fast decaying tails. If we approximate the distribution of $f_t(X)$ as normal, then we see that at small values of $T-t$, the relative decrease of the integral is overcompensated by the linear increase in $T-t$. But at larger values, the relationship inverts.

These results tell us that as long as the time remaining or the number of observed samples is very large, the sample mean is a good approximation to λ^*. In the special case of the multi-armed bandit with infinite rounds, for each round we simply select the bandit with the highest sample mean. We also notice that the variance of the gap size follows a similar pattern, but the tails do not decay to zero for large values of $T-t$. This can be room for further investigation.



5 Summary and Further Work

To summarize the results of this paper, we have formulated a class of solutions to the stationary multi-armed bandit problem. Unlike the classical problem, the bandit distributions here are fixed but unknown. Using a popular trick in bandit literature, we solve for the case of a single bandit by introducing the idea of a fair-value bandit always giving a fixed reward λ. By calibrating λ conditional on the current information, each variable-reward bandit has an equivalent λ such that we would be indifferent between the fixed and the variable reward for the next round. The solution would then be to pick the bandit with the highest λ. We also show that much like the classical problem, once we choose the fixed-value bandit, it is never optimal to switch back to the variable bandit.

It is at this point where we cannot progress much further beyond classical results like the Glivenko-Cantelli or Central Limit Theorems on the bandits' distributions and their means. Absent any distributional assumptions on the bandits' rewards, the best approximation we have is their discrete sample distributions. This simple strategy would have an equivalent solution in the classical problem where the distributions are known. An alternative is to cast the problem in a Bayesian setting, where we assume each bandit has a distribution from a family of distributions parameterized by θ. θ follows a prior distribution which is updated after each sample.

The Bayesian formalism gives us the framework to formulate strategies based on how we expect our model to change in the future, expressed through the posterior distribution of θ. λ is computed by expressing the value of each state as $V_t(\mathcal{F}_t)$, the sum of future expected rewards given we are at round t, where $\mathcal{F}_t$ denotes the information we have on the distribution thus far. Each V_t is solved iteratively backwards from T, where V_T is the sample mean. Each $\lambda(\mathcal{F}_t) = (T-t)V_t$ will have an implicit solution expressed as a sequence of nested call options. But unfortunately, solving for λ is computationally expensive exponentially in $T-t$.

We get around this by employing a strictly inferior strategy, one where we have free choice for 1 round before having to permanently stick with either the fixed or variable reward. The indifference value of the fixed reward is $\lambda_U^1 > \lambda$ as more reward is needed for fair value to an inferior strategy. Similarly, we can generalize this to λ_U^k where we have free choice for only k rounds. This generates a nested sequence of upper bounds converging to λ. Similarly, we artificially manufacture a superior strategy by assuming we are given the true value of θ after k rounds, which generates a nested sequence of lower bounds λ_L^k. Finally, the sample mean itself, is yet another lower bound, which is often tighter than λ_L^k in practice. Both λ_U^1 and the sample mean are easy to compute and we will use them as bounds for the true value of λ. The difference between the two values is referred to as the gap size.

Running experiments on simulated data drawn from a normal distribution, we observed qualitative results about the gap size as a function of rounds remaining and rounds elapsed that are in line with what our models predict. What is also interesting is that the largest gap size for 30 samples (rough limit for application of CLT) is only $\sim 1.4\%$ of the standard deviation of the samples, making the sample mean a fairly good approximation to λ. This result will vary based on the sample distribution family.

The normal distribution was chosen not only for its ubiquity but also for analytical convenience. There is room for further work in finding closed-form solutions in other families of distributions, and comparing the gap sizes to those of the normal distribution. Alternatively, work can be put towards finding a computationally cheap solution to λ, or tighter bounds. Finally, we could extend the framework beyond maximizing the expected payoff, introducing a penalty term on the variance as well. This would require significant overhaul of most, if not all the previous results, but should be of strong academic interest.

A Python Code for Experiments

```
import numpy as np
import scipy.factorial2 as fac2
import scipy.gamma as gm
from scipy.optimize import minimize

def antidercos(n, theta):
    if n==0:
        return theta + np.pi/2
    elif n==1:
        return 1 + np.sin(theta)
    else:
        return np.cos(theta)**(n-1)*np.sin(theta)/n +\
        (n-1)/n*antidercos(n-2, theta)

def compute_gamma(A, B, N):
    gamma_inv = A**(N-1)/B*fac2(N-3)/fac2(N-2)
    if N%2==1:
        gamma_inv *= 2
    else:
        gamma_inv *= np.pi
    return 1/gamma_inv

def upper_error(Z, A, B, C, N, delta, time_remaining):
    gamma = compute_gamma(A, B, N)
    theta = np.arctan(B*(Z+C)/A)
    intgd = antidercos(N-2, np.pi/2)-antidercos(N-2, theta)
    lhs = 1/time_remaining/gamma*(Z-delta)
    rhs = B**(-2)/(N-2)*(A**2+B**2*(Z+C)**2)**((2-N)/2) +\
    (Z+C)*A**(N-1)/B*intgd
    return (lhs-rhs)**2

def lower_error(lambda_L, A, B, C, N, time_remaining):
    theta = np.arctan(B*(lambda_L+C)/A)
    intgd = antidercos(N-2, np.pi/2)-antidercos(N-2, theta)
    lhs = (lambda_L-C)/(A**(N-1))/np.sqrt((N+1)/np.pi)\
    /gm(N/2.)*gm((N-1)/2.)/time_remaining
    rhs = B**(-2)/(N-2)*(A**2+B**2*(lambda_L+C)**2)**((2-N)/2\
    -(lambda_L+C)*A**(N-1)/B*intgd
    return (lhs-rhs)**2

def find_lambda_U(T, samples):
    t = len(samples)+1
    S_tm1 = np.sum(samples)
    S2_tm1 = np.sum([x**2 for x in samples])

    A = np.sqrt(S2_tm1+S_tm1**2/(t+1))
    B = np.sqrt((t+1)/t)
```

```
    C = S_tm1/(t+1)
    N = t-2
    delta = S_tm1/(t-1)
    time_remaining = T-t
    opt_args = (A, B, C, N, delta, time_remaining)
    opt_val = minimize(upper_error, delta, args=opt_args).x[0]

    return (opt_val+S_tm1)/t

def find_lambda_L(T, samples):
    t = len(samples)+1
    S_tm1 = np.sum(samples)
    S2_tm1 = np.sum([x**2 for x in samples])

    A = np.sqrt(S2_tm1-S_tm1**2/(t-1))
    B = np.sqrt(t-1)
    C = S_tm1/(t-1)
    N = t-2
    time_remaining = T-t
    opt_args = (A, B, C, N, time_remaining)
    opt_val = minimize(lower_error, C, args=opt_args).x[0]

    return opt_val
```

References

[1] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.