CSE599i: Online and Adaptive Machine Learning **Winter 2018**

Lecture 3: Stochastic Multi-Armed Bandits, Regret Minimization

Lecturer: Kevin Jamieson *Scribes: Walter Cai, Emisa Nategh, Jennifer Rogers*

Probability and Statistics Review

We begin with a review of the basics of probability and statistics, including independence, the law of large numbers, and the central limit theorem. This will lay the foundation for an introduction of tail bounds and their use in analyzing stochastic bandit problems.

Let X and Y be random variables. We say that X and Y are *independent* if, $\forall A, B$,

$$P(Y \in A | X \in B) = P(Y \in A)$$

We can use this definition of independence to show that the expectation of a product of functions of X and Y is the product of the expectations, as long as X and Y are independent:

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[\mathbb{E}[f(X)g(Y)|Y = y]] = \mathbb{E}[\mathbb{E}[f(X)]g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$$

We say $X_i \overset{iid}{\sim} P$ for $i = 1, \dots, n$ if each X_i is independent and identically distributed.

Lemma 1. *Suppose the true distribution P has mean $\mathbb{E}[X_i] = \mu$, and variance $\mathbb{E}[(X_i - \mu)^2] = \sigma^2$. If we define an estimator of the mean $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$, then $\mathbb{E}[(\hat{\mu}_n - \mu)^2] = \frac{\sigma^2}{n}$.*

Proof. We begin by substituting the definition of $\hat{\mu}_n$ and completing the square

$$\begin{aligned}
\mathbb{E}[(\hat{\mu}_n - \mu)^2] &= \mathbb{E}\left[\frac{1}{n}\sum_i (X_i - \mu)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_i (X_i - \mu)^2 + \frac{1}{n^2}\sum_{i \neq j = 1}(X_i - \mu)(X_j - \mu)\right]
\end{aligned}$$

Since the X_i and X_j are independent when $i \neq j$, the expectation of the second term is 0. This allows us to simplify the expression,

$$\begin{aligned}
\mathbb{E}[(\hat{\mu}_n - \mu)^2] &= \frac{1}{n^2}\sum_i \mathbb{E}\left[(X_i - \mu)^2\right] \\
&= \frac{1}{n^2}\sum_i \sigma^2 \\
&= \frac{\sigma^2}{n}
\end{aligned}$$

□

This bound on the squared variation of our estimator from the true mean will allow us to prove the Weak Law of Large Numbers. Before we begin that proof, we will need another result: Markov's Inequality.

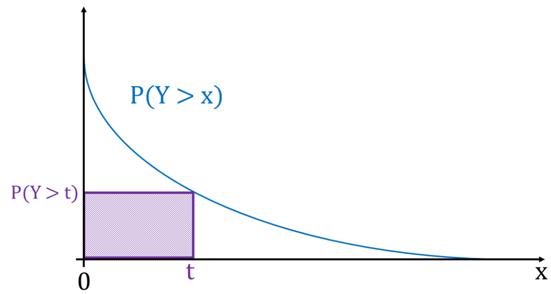

Figure 1: Since $P(Y > x)$ is a decreasing function supported only on the nonnegative numbers, the integral of $P(Y > x)$ is bounded below by $tP(Y > t)$

Lemma 2. *(Markov's Inequality): If Y is a nonnegative random variable, then $P(Y > t) \leq \frac{\mathbb{E}[Y]}{t}$*

We present two different proofs of Markov's Inequality.

Proof. We can write the expectation of Y as the integral

$$\mathbb{E}[Y] = \int_{x=0}^{\infty} P(Y > x)dx$$

Note that we can take this integral from 0 since Y is nonnegative, and that $P(Y > x)$ is a nonincreasing function. As Figure 1 illustrates, the nonincreasing nature of $P(Y > x)$ implies the following lower bound:

$$\mathbb{E}[Y] \geq tP(Y > t)$$

We have recovered Markov's Inequality, that $P(Y > t) \leq \frac{\mathbb{E}[Y]}{t}$. □

Proof. (Alternate proof of Markov's Inequality) Let Y be a positive random variable. Then

$$Y \geq t\mathbf{1}\{Y \geq t\}$$

To see why this is true, first consider that, when $Y < t$, the indicator function is zero, and by definition we know $Y \geq 0$. In the second case, when $Y \geq t$, we see that the indicator function is 1, simplifying this equation to our assumption, $Y \geq t$.

Next, we take the expectation of both sides to get

$$\begin{aligned}\mathbb{E}[Y] &\geq t\mathbb{E}[\mathbf{1}\{Y \geq t\}] \\ &= tP(Y \geq t)\end{aligned}$$

We have recovered Markov's Inequality, that $P(Y \geq t) \leq \frac{\mathbb{E}[Y]}{t}$. □

Theorem 1. *(Weak Law of Large Numbers): For all $\varepsilon > 0$, $\lim_{n\to\infty} P(|\hat{\mu}_n - \mu| > \varepsilon) = 0$*

Proof. Fix $\varepsilon > 0$. Then, $P(|\hat{\mu}_n - \mu| > \varepsilon) = P(|\hat{\mu}_n - \mu|^2 > \varepsilon^2)$. Now, the random variable in question, $|\hat{\mu}_n - \mu|^2$, is nonnegative. This means we can apply Markov's Inequality, yielding

$$\begin{aligned}P(|\hat{\mu}_n - \mu| > \epsilon) &= P(|\hat{\mu}_n - \mu|^2 > \epsilon^2) \\ &\leq \frac{\mathbb{E}[|\hat{\mu}_n - \mu|^2]}{\varepsilon^2} \\ &= \frac{\sigma^2}{n\varepsilon^2}\end{aligned}$$

where, in the last step, we applied lemma 1. Next, we take the limit,

$$\lim_{n\to\infty} P(|\hat{\mu}_n - \mu| > \varepsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2}$$
$$= 0$$

□

Example: Estimating the bias of a coin with Markov's Inequality

We have already shown in the proof of theorem 1 that, given fixed $\varepsilon > 0$, $P(|\hat{\mu}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$. Now, suppose we are trying to estimate the bias of a coin. We know the bias is bounded between $[0,1]$, and that the variance of a distribution with this support is bounded by $\sigma^2 \leq \frac{1}{4}$. This gives us

$$P(|\hat{\mu}_n - \mu| > \varepsilon) \leq \frac{1}{4n\varepsilon^2}$$

If we want the probability of such an event to be bounded by δ, then we set the right hand side equal to δ, and solve for ε. This yields

$$|\hat{\mu}_n - \mu| \leq \sqrt{\frac{1}{4n\delta}}$$

with probability at least $1-\delta$. Thus, if we desire that $|\hat{\mu}_n - \mu| \leq \epsilon$ with probability at least $1-\delta$ then we must have $n \geq \frac{\epsilon^{-2}}{4\delta}$. Later, we will see that the Central Limit Theorem suggests this to be very loose. Indeed, the CLT implies it suffices to take just $n = \epsilon^{-2}\log(2/\delta)/2$ which is substantially smaller.

Theorem 2. *(Central Limit Theorem (CLT))* $\lim_{n\to\infty} \sqrt{n}(\hat{\mu}_n - \mu) \sim \mathcal{N}(0, \sigma^2)$.

Proof. Consider the random variable

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}}$$

We will prove the central limit theorem by calculating the characteristic function of this random variable, and showing that, in the limit as $n \to \infty$, it is the same as the characteristic function for $\mathcal{N}(0,1)$. We begin by rewriting the random variable Z_n,

$$\begin{aligned} Z_n &= \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n\sigma^2}} \\ &= \sum_{j=1}^{n} \frac{X_j - \mu}{\sqrt{n\sigma^2}} \\ &= \sum_{j=1}^{n} \frac{1}{\sqrt{n}} Y_j \end{aligned}$$

where $Y_j = \frac{X_j - \mu}{\sigma}$. Note that these Y_j are i.i.d. with mean 0 and variance 1. We want to find a closed form for the characteristic function of Z_n, which is given by

$$\phi_{Z_n}(t) = \mathbb{E}\left[\exp\left(itZ_n\right)\right]$$

where $i = \sqrt{-1}$. We substitute in our definition of Z_n, yielding

$$\phi_{Z_n}(t) = \mathbb{E}\left[\exp\left(it \sum_j \frac{1}{\sqrt{n}} Y_j\right)\right]$$

By the properties of exponentials, we can change the sum in the exponent into a product of exponentials:

$$\phi_{Z_n}(t) = \mathbb{E}\left[\prod_j \exp\left(it\frac{1}{\sqrt{n}}Y_j\right)\right]$$

Since the Y_j are independent, the expectation commutes with the product, and we can write

$$\phi_{Z_n}(t) = \prod_j \mathbb{E}\left[\exp\left(it\frac{1}{\sqrt{n}}Y_j\right)\right]$$

We know the Y_j are identically distributed, so each expectation in the product must have the same value. This enables us to simplify the equation using Y_1, which is representative of all Y_j

$$\phi_{Z_n}(t) = \mathbb{E}\left[\exp\left(it\frac{1}{\sqrt{n}}Y_1\right)\right]^n$$

Using the definition of the characteristic function, we can write this as a power of the characteristic function of Y_1

$$\phi_{Z_n}(t) = \left(\phi_{Y_1}\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

Now, we can use Taylor's Theorem to approximate the characteristic function. For some (possibly complex) constant c, we have, as $\frac{t}{\sqrt{n}} \to 0$,

$$\phi_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + c\frac{t^3}{6n^{\frac{3}{2}}} + o\left(\frac{t^3}{n^{\frac{3}{2}}}\right)$$

Next, we recognize that, as $n \to \infty$, the characteristic function approaches $\phi_{Z_n}(t) \to \left(1 - \frac{t^2}{2n}\right)^n$. Using the identity $e^x = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n$, we conclude that

$$\lim_{n\to\infty} \phi_{Z_n}(t) = e^{-\frac{1}{2}t^2}$$

which is exactly the characteristic function of the standard normal distribution, $\mathcal{N}(0,1)$. Recalling that our definition of Z_n was a transformation of the random variables X_i, we see that the sum of the X_i's will converge to a normal distribution $\mathcal{N}(n\mu, n\sigma^2)$. □

Example: Estimating the bias of a coin using the Central Limit Theorem

Revisiting our coin flip example, we can use the central limit theorem to improve our asymptotic bound on $|\hat{\mu}_n - \mu|$. The central limit theorem tells us that our random variable $\sqrt{n}(\hat{\mu}_n - \mu) \sim \mathcal{N}(0, \sigma^2)$ as $n \to \infty$, so we begin with the definition of a normal distribution:

$$\begin{aligned} P(\hat{\mu}_n - \mu > \varepsilon) &\leq \int_{\varepsilon}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2/n}} e^{\frac{-nx^2}{2\sigma^2}}\, dx \\ &\leq e^{\frac{-n\varepsilon^2}{2\sigma^2}} \end{aligned}$$

We can set this bound equal to δ and solve for ε, as in our previous example. Doing this, we find that, with probability at least $1-\delta$, $|\hat{\mu}_n - \mu| \leq \sqrt{\frac{2\sigma^2\log(2/\delta)}{n}}$ as $n \to \infty$. This suggests that for "sufficiently large" n_0, we have with probability at least $1-\delta$ that $|\hat{\mu}_n - \mu| \leq \epsilon$ whenever $n \geq \max\{n_0, 2\sigma^2\epsilon^{-2}\log(2/\delta)\}$. Here, n_0 is encoding the fact that the CLT is an asymptotic statement. To make such a statement rigorous for all finite n, without knowledge of some sufficiently large n_0, we must appeal to different techniques. Note, however, that Markov's inequality holds for all n but is substantially looser than we would expect.

Chernoff Bounds

Central Limit Theorem guarantees are useful for large sample sizes, but if n is small, we would still like to bound the deviation of $\hat{\mu}_n$ from the true mean. Chernoff Bounds are a technique for bounding a random variable using its moment generating function. We wish to bound the quantity $P(\hat{\mu}_n - \mu > \varepsilon)$. For $\lambda > 0$, we can use the fact that e^x is monotonically increasing to transform our variable:

$$\begin{aligned} P(\hat{\mu}_n - \mu > \varepsilon) &= P(\lambda(\hat{\mu}_n - \mu) > \lambda\varepsilon) \\ &= P(e^{\lambda(\hat{\mu}_n - \mu)} > e^{\lambda\varepsilon}) \end{aligned}$$

Now, our random variable is nonnegative, and we can apply Markov's Inequality.

$$\begin{aligned} P(\hat{\mu}_n - \mu > \varepsilon) &\leq e^{-\lambda\varepsilon}\mathbb{E}\left[e^{\lambda(\hat{\mu}_n - \mu)}\right] \\ &= e^{-\lambda\varepsilon}\mathbb{E}\left[e^{\lambda(\frac{1}{n}\sum_i (X_i - \mu))}\right] \\ &= e^{-\lambda\varepsilon}\mathbb{E}\left[\prod_i e^{\frac{\lambda}{n}(X_i - \mu)}\right] \\ &= e^{-\lambda\varepsilon}\prod_i \mathbb{E}\left[e^{\frac{\lambda}{n}(X_i - \mu)}\right] && (1) \\ &= e^{-\lambda\varepsilon}\mathbb{E}\left[e^{\frac{\lambda}{n}(X_i - \mu)}\right]^n && (2) \end{aligned}$$

In equation 1, we have used the independence of X_i's, which means that the product commutes with the expectation. In equation 2, we have leveraged their identical distribution (so the expectations are identical). This sequence of steps, exponentiating and applying Markov's inequality with independence, is the technique known as the Chernoff bound.

Hoeffding's Inequality

Moving on from the Chernoff bound technique, we describe a more general bound: Hoeffding's Inequality. We prove the inequality by leveraging The Chernoff bound as well as Hoeffding's Lemma which we define and prove first.

Lemma 3. *(Hoeffding's Lemma) Let X be a random variable from domain $[a, b]$ almost surely and $\mathbb{E}[X] = 0$. Then for any real s*

$$\mathbb{E}\left[e^{sX}\right] \leq e^{\frac{s^2(b_i - a_i)^2}{8}}$$

Proof. We adapt the proof found from Duchi [7]. First note that e^{sx} is convex w.r.t. x. Thus we have:

$$e^{sX} \leq \frac{b - X}{b - a}e^{sa} + \frac{X - a}{b - a}e^{sb}$$

By linearity:

$$\mathbb{E}\left[e^{sX}\right] \leq \frac{b - \mathbb{E}[X]}{b - a}e^{sa} + \frac{\mathbb{E}[X] - a}{b - a}e^{sb} = \frac{b}{b - a}e^{sa} - \frac{a}{b - a}e^{sb}$$

For ease of reading, let $p = \frac{-a}{b-a}$ also noting that $a = -p(b - a)$. We isolate a factor e^{sa}:

$$\begin{aligned} \frac{b}{b - a}e^{sa} - \frac{a}{b - a}e^{sb} &= (1 - p)e^{sa} + pe^{sb} \\ &= \left((1 - p) + pe^{s(b-a)}\right)e^{sa} \\ &= \left(1 - p + pe^{s(b-a)}\right)e^{-sp(b-a)} \end{aligned}$$

Substitute $u = s(b-a)$:

$$\left(1 - p + pe^{s(b-a)}\right) e^{-sp(b-a)} = (1 - p + pe^{u})\, e^{pu}$$

Define function ϕ of u as the logarithm of the above expression:

$$\phi(u) = \log\left((1 - p + pe^{u})\, e^{pu}\right) = pu + \log\left(1 - p + pe^{u}\right)$$

We write $\mathbb{E}\left[e^{eX}\right] \leq e^{\phi(u)}$ and we proceed to bound $\phi(u)$. Our route is to apply Taylor's theorem; there must exist some $z \in [0, u]$ where

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{1}{2}u^2\phi''(z) \leq \phi(0) + u\phi'(0) + \sup_z \frac{1}{2}u^2\phi''(z) \tag{3}$$

We derive the first and second derivatives of $\phi(u)$:

$$\phi'(u) = p + \frac{pe^u}{1 - p + pe^u}$$
$$\phi''(u) = \frac{p(1-p)e^u}{(1 - p + pe^u)^2}$$

We have $\phi(0) = \phi'(0) = 0$ so we may rewrite Equation (3):

$$\phi(u) \leq \underbrace{\phi(0)}_{=0} + \underbrace{u\phi'(0)}_{=0} + \sup_z \frac{1}{2}u^2\phi''(z) = \sup_z \frac{1}{2}u^2\phi''(z)$$

We therefore need only maximize $\phi''(z)$. We substitute y for e^u:

$$\frac{p(1-p)y}{(1-p+py)^2}$$

We note that the expression is a linear expression over a quadratic expression and therefore concave for $y > 0$. It therefore suffices to find the critical point for y:

$$\frac{d}{dy}\frac{p(1-p)y}{(1-p+py)^2} = \frac{p(1-p)(1-p-py)}{(1-p+py)^3}$$

We have two critical points to consider; $y = \frac{1-p}{p}$, and $y = \frac{p-1}{p}$. We note $\mathbb{E}[X] \geq 0 \implies a \leq 0 \implies p = \frac{-a}{b-a} \in [0, 1]$. Hence $\frac{1-p}{p} \geq 0$, and $y = \frac{p-1}{p} \leq 0$. We therefore select the candidate that falls inside the nonegative window: $y = \frac{1-p}{p}$. Note that if $\frac{p-1}{p} = 0 \implies p = 1 \implies \frac{p-1}{p} = 0$. That is, there is in fact only a single critical point in this situation. Substituting back in, we have:

$$\phi''(u) \leq \frac{p(1-p)\frac{1-p}{p}}{(1-p+p\frac{1-p}{p})^2} = \frac{1}{4}$$

We may conclude:

$$\mathbb{E}\left[e^{sX}\right] \leq e^{\phi(u)} \leq e^{\frac{u^2}{8}} = e^{\frac{s^2(b-a)^2}{8}}$$

$\square$

We now prove the primary implication of Lemma 3 and result of this subsection.

Theorem 3. *(Hoeffding's Inequality) Given independent random variables $\{X_1, \dots, X_m\}$ where $a_i \leq X_i \leq b_i$ almost surely (with probability 1) we have:*

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[X_i] \geq \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2 m^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right)$$

Proof. We adapt the proof found from Duchi [7]. As mentioned earlier, the result is a straightforward application of Lemma 3.

For all $1 \leq i \leq m$ define a new variable Z_i as the difference between X_i and its expectation.

$$Z_i = X_i - \mathbb{E}[X_i]$$

This implies that $\mathbb{E}[Z_i] = 0$. Moreover, we may bound the domain of Z_i inside $[a_i - \mathbb{E}[X_i], b_i - \mathbb{E}[X_i]]$. In particular, we note that the interval must still have length $b_i - a_i$ independent of the expectation of X_i.

Let s be some positive value. We have:

$$\mathbb{P}\left(\sum_{i=1}^{m} Z_i \geq t\right) = \mathbb{P}\left(\exp\left(s\sum_{i=1}^{m} Z_i\right) \geq e^{st}\right) \overset{\text{Chernoff}}{\leq} \frac{\mathbb{E}\left[\prod_{i=1}^{m} e^{sZ_i}\right]}{e^{st}}$$

By independence of the Z_i we may shift the expectation inside the product and continue. Recall that Z_i must still live in an interval of length $b_i - a_i$.

$$\frac{\mathbb{E}\left[\prod_{i=1}^{m} e^{sZ_i}\right]}{e^{st}} = \frac{\prod_{i=1}^{m}\mathbb{E}\left[e^{sZ_i}\right]}{e^{st}} \overset{\text{Hoeffding Lemma}}{\leq} e^{-st}\prod_{i=1}^{m} e^{\frac{s^2(b_i-a_i)^2}{8}} = \exp\left(-st + \frac{s^2}{8}\sum_{i=1}^{m}(b_i - a_i)^2\right)$$

We now substitute a conveniently engineered value of s to conclude our result. Note that $s > 0$ so the earlier restriction on s is satisfied.

$$s = \frac{4t}{\sum_{i=1}^{m}(b_i - a_i)^2}$$

substituting in s as well as $t = \epsilon m$, we have:

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{m} Z_i \geq \epsilon m\right) &= \mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}[X_i] \geq \epsilon\right) \\
&\leq \exp\left(-\frac{4\epsilon m}{\sum_{i=1}^{m}(b_i - a_i)^2}\epsilon m + \frac{1}{8}\left(\frac{4\epsilon m}{\sum_{i=1}^{m}(b_i - a_i)^2}\right)^2\sum_{i=1}^{m}(b_i - a_i)^2\right) \\
&= \exp\left(\frac{-2\epsilon^2 m^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right)
\end{aligned}$$

□

Stochastic Multi-Armed Bandits

In the stochastic multi-armed bandit problem, the player is presented with a collection of actions, or *arms*, to choose from in each round of play. Each arm distributes rewards according to some (unknown) subgaussian distribution over $[0, 1]$. Rewards are *i.i.d.*, with $\mathbb{E}[X_{i,t}] = \mu_i$ for all arms i and times t. The goal of the player is to minimize the cumulative *regret*, which is defined as the difference between the player's rewards after T time steps, and the best reward possible given the strategy of playing a single arm. If the player chooses arm I_t at time t, the regret can be written as

$$R(T) = \max_j \mathbb{E}\left[\sum_{t=1}^{T}(X_{j,t} - X_{I_t,t})\right]$$

For an alternative formulation of the regret, define each arm's gap from the best arm as $\Delta_i = \max_j \mu_j - \mu_i$. We can rewrite the regret by taking the expectation inside the summation,

$$\begin{aligned} R(T) &= \max_j \sum_t \mathbb{E}[X_{j,t}] - \mathbb{E}[X_{I_t,t}] \\ &= \max_j \sum_t \mu_j - \mathbb{E}[\mu_{I_t}] \\ &= \sum_{i=1}^n \Delta_i \mathbb{E}[T_i] \end{aligned}$$

where T_i is the total number of times we have played arm i. We see that the regret is the product of the number of times each suboptimal arm is played and that arm's gap with the optimal arm.

UCB (Upper Confidence Bounds)

Auer et al. (2002) [3] introduced simple and efficient allocation strategies based on upper confidence bounds for a bandit problem with any reward distribution with known bounded support. Their algorithms demonstrate logarithmic regret performance uniformly over time, not just asymptotically.

To implement the UCB1 algorithm, we need both $\hat{\mu}_{i,T_i}$, the empirical reward estimate of arm i after it has been pulled T_i many times, and an upper confidence bound on that estimate. To calculate our empirical reward estimate, we simply average the observed rewards over all rounds where we pull arm i. In the below expression, we assume $|\{t : I_t = i\}| = T_i$. That is, we have progressed through an appropriate number of rounds where arm i has been pulled T_i many times:

$$\hat{\mu}_{i,T_i} = \frac{\sum_{t:I_t=i} X_t}{T_i}$$

In addition to our empirical reward estimates, we need an upper confidence bound to describe the largest plausible mean of each arm. Using Hoeffding's Inequality and Chernoff Bounds, we can construct such a confidence interval. With probability at least $1 - t^{-\alpha}$, the empirical mean $\hat{\mu}_{i,T_{i,t}}$ will differ from the true mean by at most $\epsilon = \sqrt{\frac{\alpha \log t}{2T_i}}$. The UCB1 algorithm chooses the largest such upper bound:

$$\text{UCB}_{i,t} = I_t := \underset{i\in[n]}{\arg\max}\ \hat{\mu}_{i,T_{i,t}} + \sqrt{\frac{\alpha\ \log(t)}{2\ T_i}}$$

We see that our confidence bound, $\sqrt{\frac{\alpha \log(t)}{2T_i}}$, grows slowly as we play for more rounds (as t increases), ensuring that we never stop playing any given arm. The confidence bound for arm i shrinks quickly as we pull the arm (as T_i increases).

The pseudocode may be found in Algorithm 1.

Theorem 4. *Regret bound for the UCB algorithm*
For $T \geq 1$

$$R(T) \leq \sum_{i:\Delta_i>0} 4\alpha\Delta_i^{-1}\log(T) + \frac{2\alpha}{\alpha-1}\Delta_i$$

Proof. Suppose, without loss of generality, that arm 1 is optimal. Then, arm $i \neq 1$ will only be played in two cases: either arms 1 and i have been sampled insufficiently to distinguish between their means, or the upper confidence bound given by Hoeffding's inequality fails for either arm 1, or arm i. We begin by bounding the chance that we pull a suboptimal arm due to insufficient sampling.

Suppose that we have the following two events A_t, B_t.

Algorithm 1 UCB

procedure UCB($\{1, 2, \ldots, n\}, T$) ▷ Arms 1 through n, max steps T
  for $1 \leq t \leq n$ **do**
    $I_t \leftarrow t$ ▷ Play each arm once
  end for
  for $n + 1 \leq t \leq T$ **do**
    $I_t = \underset{i \in \{1, \cdots, n\}}{\arg\max} \mathrm{UCB}_{i,t-1}$
    Observe reward $X_{T_t,t}$
  end for
end procedure

A_t) $\hat{\mu}_{i,T_i} \leq \mu_i + \sqrt{\frac{\alpha \log t}{2T_i}}$

B_t) $\hat{\mu}_{1,T_1} \geq \mu_1 - \sqrt{\frac{\alpha \log t}{2\ T_1}}$

We wish to bound the probabilities of the complements of events A_t and B_t occuring. We will apply Hoeffding's inequality (Theorem 3). A_t fails when

$$\hat{\mu}_{i,T_i} - \mu_i > \sqrt{\frac{\alpha \log t}{2\ T_i}}$$

By Theorem 3 we have:

$$\begin{aligned}
\mathbb{P}(A_t^c) = \mathbb{P}\left(\hat{\mu}_{i,T_i} - \mu_i > \epsilon\right) &\leq \exp\left(\frac{-2\epsilon^2 t^2}{\sum_{i=1}^{t}(b_i - a_i)^2}\right) \\
&= \exp\left(\frac{-2\epsilon^2 t^2}{\sum_{i=1}^{t}(1-0)^2}\right) \\
&= \exp(-2\epsilon^2 t)
\end{aligned}$$

We plug in our bounding value $\epsilon = \sqrt{\frac{\alpha \log t}{2T_i}}$.

$$\begin{aligned}
\mathbb{P}\left(\hat{\mu}_{i,T_i} - \mu_i > \sqrt{\frac{\alpha \log t}{2T_i}}\right) &\leq \exp\left(\frac{-2t\alpha \log t}{2T_i}\right) \\
&= \exp\left(\frac{-t\alpha \log t}{T_i}\right) \\
&\leq \exp\left(\frac{-t\alpha \log t}{t}\right) \\
&= e^{\alpha \log t} \\
&= t^{-\alpha} \qquad (4)
\end{aligned}$$

The statement and justification is identical for the complement of event B_t and we return to the task of bounding the number of suboptimal arm pulls.

A suboptimal arm i is only played if its upper confidence bound exceeds that of arm 1, meaning that

$$\hat{\mu}_{i,T_i} + \sqrt{\frac{\alpha \log t}{2\ T_i}} > \hat{\mu}_{1,T_1} + \sqrt{\frac{\alpha\ \log t}{2\ T_1}}$$

Suppose that both A_t and B_t both hold. In this case, suboptimal arm i is pulled due to insufficient sampling up to this point.

Since A_t has been assumed to be true, we generate the following bound:

$$\mu_i + 2\sqrt{\frac{\alpha\log(t)}{T_i}} \geq \hat{\mu}_{i,T_i} + \sqrt{\frac{\alpha\log(t)}{2\ T_i}} \tag{5}$$

Next, we use our assumption of B_t being true to upper-bound the right hand side of Line (5):

$$\hat{\mu}_{1,T_1} + \sqrt{\frac{\alpha\log t}{2\ T_1}} \geq \mu_1 \tag{6}$$

Chaining equations (5) and (6) we have

$$\mu_i + 2\sqrt{\frac{\alpha\log(t)}{T_i}} \geq \mu_1$$

Rearranging we have:

$$\sqrt{\frac{\alpha\log(t)}{T_i}} \geq \frac{\mu_1 - \mu_i}{2}$$

Now, recall our definition of the optimality gap of an arm, $\Delta_i = \max_j \mu_j - \mu_i$. Since we know arm 1 is optimal, this becomes $\Delta_i = \mu_1 - \mu_i$. Our inequality becomes

$$\sqrt{\frac{\alpha\log(t)}{T_i}} \geq \frac{\Delta_i}{2}$$

Solving for the number of times T_i that an arm has been played, we arrive at

$$\begin{aligned} T_i &\leq 4\Delta_i^{-2}\alpha\log(t) \\ &\leq 4\Delta_i^{-2}\alpha\log(T) \end{aligned} \tag{7}$$

Thus when A_t and B_t hold, we only play suboptimal arm i at most $4\Delta_i^{-2}\alpha\log(T)$ times.

Recall that I_t can only be equal to i if either it has been sampled insufficiently (fewer than $4\Delta_i^{-2}\alpha\log(T)$ times) or either event A_t or B_t fails. For any arm i, the expected number of times it is played up to round T under UCB is:

$$\begin{aligned} \mathbb{E}[T_i] &= \sum_{t=1}^{T} \ \mathbb{E}[\mathbf{1}(I_t = i)] \\ &\leq 4\ \alpha\Delta_i^{-2}\log(T) + \sum_{t=1}^{T}\mathbb{E}[\mathbf{1}\{A_t^c \cup B_t^c\}] \\ &\leq 4\ \alpha\Delta_i^{-2}\log(T) + \sum_{t=1}^{T}\Big(\mathbb{E}[\mathbf{1}\{A_t^c\}] + \mathbb{E}[\mathbf{1}\{B_t^c\}]\Big) && (8) \\ &\leq 4\ \alpha\Delta_i^{-2}\log(T) + \sum_{t=1}^{T}\Big(t^{-\alpha} + t^{-\alpha}\Big) && (9) \\ &= 4\ \alpha\Delta_i^{-2}\log(T) + 2\sum_{t=1}^{T} t^{-\alpha} \end{aligned}$$

The inequality in line (8) comes from the union bound on events A_t^c and B_t^c. The inequality in line (9) comes from Equation (4). In order to bound the second term on the right hand side, we note:

$$\sum_{t=1}^{T} t^{-\alpha} \leq 1 + \int_{1}^{\infty} x^{-\alpha}\, dx = 1 + \frac{-1}{1-\alpha} = \frac{-\alpha}{1-\alpha}$$

Therefore we have:

$$\mathbb{E}[T_i] \leq 4\alpha\Delta_i^{-2}\log(T) + \frac{2\alpha}{\alpha-1} \tag{10}$$

The desired result follows from summing over all suboptimal arms:

$$\begin{aligned} R(T) &= \sum_{i\neq 1} \Delta_i \mathbb{E}[T_i] \\ &= \sum_{i\neq 1} 4\alpha\Delta_i^{-1}\log(T) + \frac{2\alpha}{\alpha-1}\Delta_i \end{aligned}$$

□

Theorem 5. *(Lai-Robin's (1985) [13]) Lai and Robbins, provided an optimal asymptotic lower bound on the expected regret of any bandit algorithm.*

If $\forall \beta > 0$, $R(T) \leq o(T^\beta)$, *then*

$$\mathbb{E}[T_i] \geq \frac{\log(T)}{\Delta_i^2}$$

We refer the reader to Kaufmann et al. (2012) [12] for a proof of the above theorem. For an overview of the UCB family of algorithms refer to Bubeck and Cesa-Bianchi (2012, chap. 2) [5].

If the gap between the best and second-best arm is very small, then the Δ_i^{-1} penalty in the regret bound becomes very large. However, as the gap becomes small, we would imagine that playing the second-best arm becomes a decent strategy. To this end, we seek a "worst case" regret bound for the UCB algorithm. This bound, which is independent of Δ_i, is shown in the following theorem.

Theorem 6. *For all* $T \geq n$, *a gap-agnostic bound achieved by the UCB algorithm in round* T *is*

$$\mathbb{E}[R(T)] \leq (1+4\alpha)\sqrt{nT\log(T)} + n\frac{2\alpha}{\alpha-1}$$

Proof. Divide the arms into two groups:

- Group G_1 contains "almost optimal" arms with $\Delta_i < \sqrt{\frac{n}{T}\log(T)}$.
- Group G_2 contains arms with $\Delta_i \geq \sqrt{\frac{n}{T}\log(T)}$.

The total regret is the sum of the regret of each group. The maximum total regret incurred due to pulling arms in G_1 is given by

$$\sum_{i\in G_1} T_i\Delta_i$$

By definition, the regret on any arm $i \in G_1$ is bounded by $\Delta_i < \sqrt{\frac{n}{T}\log(T)}$. We may therefore bound the total regret on arms in G_1 as follows:

$$\begin{aligned} \sum_{i\in G_1} T_i\Delta_i &\leq \sqrt{\frac{n}{T}\log(T)} \sum_{i\in G_1} T_i \\ &\leq T \cdot \sqrt{\frac{n}{T}\log(T)} \\ &= \sqrt{nT\log(T)} \end{aligned}$$

We may now shift our focus to group G_2. Recall by definition for all arms $i \in G_2$ we have $\Delta_i \geq \sqrt{\frac{n}{T}\log(T)}$. Rearranging we have:

$$\Delta_i^{-1} \leq \sqrt{\frac{T}{n\log(T)}} \tag{11}$$

We begin by building on Equation (10) and summing over all arms in G_2:

$$\begin{aligned}
\sum_{i \in G_2} \mathbb{E}[T_i]\Delta_i &\leq \sum_{i \in G_2} \left(4\alpha\Delta_i^{-2}\log(T) + \frac{2\alpha}{\alpha - 1}\right)\Delta_i \qquad (12)\\
&= \sum_{i \in G_2} \left(4\alpha\Delta_i^{-1}\log(T) + \Delta_i\frac{2\alpha}{\alpha - 1}\right)\\
&\leq \sum_{i \in G_2} \left(4\alpha\sqrt{\frac{T}{n\log(T)}}\log(T) + 1\frac{2\alpha}{\alpha - 1}\right) \qquad (13)\\
&= \sum_{i \in G_2} \left(4\alpha\sqrt{\frac{T\log(T)}{n}} + 1\frac{2\alpha}{\alpha - 1}\right)\\
&\leq n \cdot \left(4\alpha\sqrt{\frac{T\log(T)}{n}} + 1\frac{2\alpha}{\alpha - 1}\right) \qquad (14)\\
&\leq 4\alpha\sqrt{nT\log(T)} + n\frac{2\alpha}{\alpha - 1}
\end{aligned}$$

Line (12) follows from multiplying Δ_i (a necessarily nonnegative value) onto either side of equation (10) and summing over all arms in G_2. Line (13) follows from Equation (11), and the fact that μ_i lives in $[0, 1]$ implying Δ_i may not exceed 1. For line (14), note that the interior of the summation is independent of which arm i we are iterating over and $|G_2| \leq n$.

We sum the expected regret over all arms in groups G_1 and G_2 to arrive at the total expected regret:

$$\begin{aligned}
\mathbb{E}[R(T)] &= \sum_{1 \leq i \leq n} \mathbb{E}[T_i]\Delta_i\\
&= \sum_{i \in G_1} \mathbb{E}[T_i]\Delta_i + \sum_{i \in G_2} \mathbb{E}[T_i]\Delta_i\\
&\leq \sqrt{nT\log(T)} + 4\alpha\sqrt{nT\log(T)} + n\frac{2\alpha}{\alpha - 1}\\
&= (1 + 4\alpha)\sqrt{nT\log(T)} + n\frac{2\alpha}{\alpha - 1}
\end{aligned}$$

□

Putting Theorems 4 and 6 together, we see that the UCB algorithm operates under two distinct regimes. During the initial "burn-in" period, the algorithm experiences $O(\sqrt{nT\log T})$ regret to learn the arm payouts. As the game continues, and the gap between the arms becomes easier to distinguish, the algorithm moves into the second regime, where its performance is $O(\sum_i \Delta_i^{-1}\log T)$.

UCB algorithms are an active research field in machine learning, especially for the contextual bandit problem [5, 8, 15]. For an overview of the UCB family of algorithms refer to Bubeck and Cesa-Bianchi (2012, chap. 2) [5] and [4].

Thompson Sampling(Posterior Sampling or Probability Matching)

Thompson sampling (TS) is one of the oldest heuristics for multi-armed bandit problems [20]. Thompson sampling takes a Bayesian approach to find the optimal arm while balancing the trade off between exploration and exploitation of non-optimal arms. In TS, the reward of each arm is distributed Bernoulli and the expected reward is unknown. The objective is to find the optimal arm that gives maximum expected cumulative reward. The TS algorithm initially assumes arm i to have prior Beta $(1,1)$ on μ_i, which is natural because Beta$(1,1)$ is the uniform distribution on $(0,1)$. At time t, having observed $S_i(t)$ successes (reward=1) and $F_i(t)$ failures (reward=0) in $T_i(t) = S_i(t) + F_i(t)$ plays of arm i, the algorithm updates the distribution on μ_i as Beta$(S_i(t)+1, F_i(t)+1)$. The algorithm then samples from these posterior distributions of the μ_i's, and plays an arm according to the probability of its mean being the largest. The Thompson sampling algorithm is given in Algorithm 2.

Algorithm 2 Thompson Sampling

```
procedure THOMPSON({1, 2, ..., n}, T)        ▷ Arms 1 through n, max steps T
    S_{I_t}, F_{I_t}, T_i ← 0
    for 1 ≤ t ≤ T do
        for 1 ≤ i ≤ n do
            μ̂_i ~ Beta(S_{I_t} + 1, F_{I_t} + 1)   ▷ Draw each μ̂_i according to the posterior distribution
        end for
        I_t ← argmax_{i∈[n]} μ̂_i
        T_{I_t} ← T_{I_t} + 1                ▷ Increment the total counter for arm I_t.
        X_{I_t,t} ~ Bernoulli(μ_i)           ▷ Observe reward X_{I_t,t}
        S_{I_t} ← S_{I_t} + X_{I_t,t}        ▷ Update success counter appropriately
        F_{I_t} ← T_i I_t − S_{I_t}          ▷ Update failure counter appropriately
    end for
end procedure
```

Thompson Sampling has received considerable attention in industry as well (e.g. Scott (2010) [18], Graepel et al. (2010) [11], and Tang et al. (2013) [19]). For more details please see [1, 2, 6, 17].

KL-UCB

The Kullback-Leibler UCB algorithm (KL-UCB) presents a modern approach to UCB for the standard stochastic bandits problem. KL-UCB improves the regret bounds from earlier UCB algorithms by considering the distance between the estimated distributions of each arm. The algorithms differ only at the arm selection step. Recall that UCB uses the following rule for arm selection:

$$\arg\max_{i\in[n]} \hat{\mu}_{i,T_{i,t}} + \sqrt{\frac{\alpha\ \log(t)}{2\ T_i}}$$

In contrast, KL-UCB uses:

$$I_t := \arg\max_{i\in[n]} \left(\max\left(q \in [0,1] : T_i \cdot d\left(\hat{\mu}_{i,T_{i,t}}, q\right) \le \log(t) + c\log(\log(t))\right) \right) \tag{15}$$

where $d(\cdot,\cdot)$ is the Bernoulli Kullback-Leibler divergence:

$$d(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

and c is a tuning parameter. The inner expression on the right side of Equation (15) is the stronger upper confidence bound. For each arm $i \in [n]$, the maximal q in the inner statement may be efficiently approximated using Newton's method. The psuedo-code may be found in Algorithm 3.

KL-UCB is optimal for Bernoulli distributions and strictly dominates UCB for any bounded reward distributions. For more details please see [9, 14].

Algorithm 3 KL-UCB

procedure KL-UCB($\{1, 2, \ldots, n\}, T$) ▷ Arms 1 through n, max steps T
  for $1 \le t \le n$ **do**
    $I_t \leftarrow t$ ▷ Play each arm once
  end for
  for $n + 1 \le t \le T$ **do**
    $I_t = \arg\max_{i \in [n]} \left(\max \left(q \in [0,1] : T_i \cdot d \left(\hat{\mu}_{i,T_{i,t}}, q \right) \le \log(t) + c \log(\log(t)) \right) \right)$
    Observe reward $X_{T_t,t}$
  end for
end procedure

Examples/Applications for TS and UCB

One of the early motivations for studying the Multi-Armed Bandit problem was clinical trials. Suppose that we have N different treatments of unknown efficacy for a certain disease. Patients arrive sequentially, and we must decide on a treatment to administer for each arriving patient. To make this decision, we could learn from how the previous choices of treatments fared for the previous patients. After a sufficient number of trials, we may have a reasonable idea of which treatment is most effective, and from then on, we could administer that treatment for all the patients. In applications like display advertising, product assortment, recommendation system (e.g. news article recommendation, cascading recommendation, recommending courses to learners), reinforcement learning in Markov decision processes, and active learning with neural networks, Thompson sampling is competitive to or better than popular methods such as UCB. Web advertising, job scheduling (or exercise scheduling), and routing (shortest path problem) examples could be another motivations in MAB problems. For more details please see [10].

Thompson Sampling and offers significant advantages over the UCB approach, and can be applied to problems with finite or infinite action spaces and complicated relationships among action rewards, refer to Russo and Van Roy (2014) [16].

UCB algorithms have been proposed for a variety of problems, including bandit problems with independent arms, bandit problems with linearly parameterized arms, bandits with continuous action spaces and smooth reward functions, and exploration in reinforcement learning. UCB1 is the building block for tree search algorithms (e.g. Upper Confidence bound applied to Trees (UCT)) used to, e.g., play games. There are some limitations for using Thompson sampling. For example, it is certainly a poor fit for sequential learning problems that do not require much active exploration. It may also perform poorly in time-sensitive learning problems where it is better to exploit a high performing suboptimal action than to invest resources exploring arms that might offer slightly improved performance.

References

[1] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. *CoRR*, abs/1111.1797, 2011.

[2] Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. *CoRR*, abs/1209.3353, 2012.

[3] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, May 2002.

[4] Peter Auer and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem.

[5] Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *CoRR*, abs/1204.5721, 2012.

[6] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pages 2249–2257, USA, 2011. Curran Associates Inc.

[7] John C. Duchi. Probability bounds. 2009.

[8] Sarah Filippi, Olivier Cappe, Aurélien Garivier, and Csaba Szepesvári. Parametric bandits: The generalized linear case. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 586–594. Curran Associates, Inc., 2010.

[9] Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *COLT*, 2011.

[10] Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages I–100–I–108. JMLR.org, 2014.

[11] Thore Graepel, Joaquin Quiñonero Candela, Thomas Borchert, and Ralf Herbrich. Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 13–20, USA, 2010. Omnipress.

[12] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *J. Mach. Learn. Res.*, 17(1):1–42, January 2016.

[13] T.L Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Adv. Appl. Math.*, 6(1):4–22, March 1985.

[14] Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. A Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences. In Sham Kakade & Ulrike von Luxburg, editor, *24th Annual Conference on Learning Theory : COLT'11*, page 18, Budapest, Hungary, July 2011.

[15] Paat Rusmevichientong and John N. Tsitsiklis. Linearly parameterized bandits. *CoRR*, abs/0812.3465, 2008.

[16] Daniel Russo and Benjamin Van Roy. Learning to optimize via posterior sampling. *CoRR*, abs/1301.2609, 2013.

[17] Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, and Ian Osband. A tutorial on thompson sampling. *CoRR*, abs/1707.02038, 2017.

[18] Steven L. Scott. A modern bayesian look at the multi-armed bandit. *Appl. Stoch. Model. Bus. Ind.*, 26(6):639–658, November 2010.

[19] Liang Tang, Romer Rosales, Ajit Singh, and Deepak Agarwal. Automatic ad format selection via contextual bandits. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*, CIKM '13, pages 1587–1594, New York, NY, USA, 2013. ACM.

[20] WILLIAM R THOMPSON. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933.