Optimal betting for a multi-armed bandit

Kurt Ehlert
kehlert@math.wisc.edu

Abstract

If we are betting on a slot machine, then the natural question to ask is "how much should we bet"? If the probability of winning is p, it turns out that betting a fraction of our wealth equal to $2p - 1$ is an optimal strategy in many senses. However, if we are faced with many slot machines with unknown probabilities of winning, then how should we bet, and which slot machines should we bet on? We will present the Kelly-UCB algorithm, which addresses this more complicated problem. We will also provide analytic and numerical results regarding the performance of the Kelly-UCB algorithm.

1 Introduction

A slot machine is also known as a one-armed bandit, because early slot machines were operated by pulling a lever (arm) attached to their side [Glimne, 2015]. Suppose that we are pulling an arm with a probability p of winning, where $1/2 < p < 1$. Also, we start with a finite amount of money, and we can bet any fraction of our wealth each time we pull the arm. We win an amount equal to our bet with probability p, and conversely we lose our bet with probability $1 - p$. If $p > 1/2$, then we are likely to gain money by playing the game, but we need to be careful with our betting strategy. If we bet a fraction of our wealth equal to $2p - 1$, then this strategy is known as the "Kelly bet" [Kelly, 1956, Thorp, 2006]. In some senses it maximizes our wealth, and it also avoids bankruptcy. In many situations, we do not know the probability p. We will discuss a version of the Kelly bet for that type of problem.

Now consider a situation where we can choose between many arms. If we pull arm i, then we win with probability p_i and lose with probability $1 - p_i$. If we are only allowed to bet a fixed amount and we do not know the value of p_i for any i, then this is called the stochastic multi-armed bandit problem. There is a trade-off between exploitation and exploration: we want to exploit an arm with a seemingly high probability of winning, but we also want to explore and find the best arm. Here we consider a new version of the stochastic bandit problem. Instead of fixed-size bets, we can bet any fraction of our wealth. The KL-UCB algorithm in Cappé et al. [2013] does not address how much to bet, so we plan to extend the KL-UCB algorithm and its analysis to the variable-bet situation.

In Section 2, we review the Kelly bet for the one-armed bandit with known probabilities. Then we extend the Kelly bet to the situtation where we do not know the probabilities of the possible outcomes. Section 3 outlines our Kelly-UCB algorithm for playing a variable-bet multi-armed bandit. Then we will present analytic and numerical results on the algorithm's performance.

2 Optimal betting for a single-armed bandit

2.1 Kelly bet for a bandit with a known probability of winning

Suppose we are playing a slot machine that allows us to bet any proportion of our wealth. Let $X_n \in \{-1, 1\}$ be a random variable with $P(X_n = 1) = 1 - P(X_n = -1) = p$. When we pull the arm, we obtain of profit of X_n per unit bet on the n^{th} pull. If we bet \$100 and $X_1 = -1$, then we

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

lose our entire bet of \$100. If instead $X_1 = 1$, then we win \$100 (and get back our original bet of \$100). Assume that the X_n are i.i.d. for all n, and that $p > 1/2$.

If we want to maximize our expected profit, then we should bet our entire fortune on every pull. However, we would go bankrupt with probability 1. If instead we bet a fixed amount, then there is still a positive probability of going bankrupt. We will take a different approach and bet a fixed proportion of our wealth on each pull. Betting a proportion of our wealth avoids bankruptcy (although our wealth could become incredibly small), and if we choose the proportion carefully then our bets are optimal in many ways. Next we will derive the "optimal" proportional bet, which is called the "Kelly bet" or "Kelly criterion" [Kelly, 1956, Thorp, 2006]. Ethier [2010] analyzes the Kelly bet in much more detail and shows that it is optimal in many more ways than we consider here. Ethier [2010] also considers more general random variables X_n. Our later results rely on X_n being a Bernoulli random variable, so we do not consider the more general case in this section.

Let W_0 be our initial wealth, and let W_n is our wealth after pull n. Define f as the proportion of our wealth that we bet. Then

$$W_n = W_{n-1} + fW_{n-1}X_n \tag{2.1}$$
$$= W_{n-1}(1 + fX_n) \tag{2.2}$$

Using recursion, we can see that our wealth after pull n is

$$W_n = W_0 \prod_{i=1}^{n} (1 + fX_i) \tag{2.3}$$

Here is another way to write down W_n. Let $r_n(f) = n^{-1} \log(W_n/W_0)$, then we can rewrite our wealth after pull n as

$$W_n = W_0 e^{r_n(f)n} \tag{2.4}$$

We can interpret r_n as the average geometric growth rate of our wealth after pull n. Intuitively, we want to maximize the growth rate of our wealth. That intuition leads to the following lemma, which is based on Lemma 10.1.1 in Ethier [2010]

Lemma 2.1. *Let $\mu(f) = \mathbb{E}[\log(1 + fX_1)]$, which is defined for $f \in [0, 1)$. Then $\lim_{n\to\infty} r_n(f) = \mu(f)$. Also, $f^* = \arg\max_f \mu(f) = 2p - 1$ and $\mu(f^*) = \log(2) - H(p)$, where $H(p)$ is the entropy of the Bernoulli distribution with parameter p.*

Proof. The strong law of large numbers implies

$$\lim_{n\to\infty} r_n(f) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \log(1 + fX_i) = \mathbb{E}[\log(1 + fX_1)] \text{ a.s.} \tag{2.5}$$

$\mu(f)$ is strictly concave, because

$$\mu''(f) = -\mathbb{E}\left[\frac{X^2}{(1 + fX)^2}\right] < 0,\ 0 \le f < 1 \tag{2.6}$$

Therefore we can find the global maximum of $\mu(f)$ by just setting its derivative equal to 0. We find that $f^* = \arg\max_{0\le f\le 1} \mu(f) = 2p - 1$. Computing $\mu(2p - 1)$ almost immediately gives the last result. ☐

Above we showed that the Kelly bet is optimal in an asymptotic sense, so the natural question to ask is if it is also optimal for finite times. If we wish to maximize the logarithm of our wealth after n pulls, then once again the Kelly bet is optimal.

Lemma 2.2. *f^* is the unique maximizer of $\mathbb{E}[\log W_n(f)]$.*

Proof. From the definition of W_n and the linearity of expectations

$$\mathbb{E}[\log W_n(f)] = \log W_0 + \sum_{i=1}^{n} \mathbb{E}[\log(1 + fX_n)] \tag{2.7}$$
$$= \log W_0 + n\mu(f) \tag{2.8}$$

Therefore maximizing the log-wealth is equivalent to maximizing $\mu(f)$. In lemma 2.1 we showed that f^* is the unique maximizer of $\mu(f)$. ☐

The Kelly bet also approximately maximizes the median wealth. Additionally, in the long-run it will do better than any other "essentially different" strategy, which includes strategies that change betting proportions between pulls. Ethier [2010] provides the details.

2.2 Kelly bet for a bandit with an unknown probability of winning

A simple approach to estimating the Kelly bet would be to calculate f^* based on an estimated value of p. The obvious estimator is

$$\bar{p}_n = \frac{n_1}{n}$$

where $n_1 = \sum_{i=1}^{n} \mathbb{1}(X_i = 1)$. then we could let $\hat{f}_n = 2\bar{p}_n - 1$. This approach is somewhat reasonable, because our wealth would grow at a near-optimal rate if $\bar{p}$ is a good estimate. A major problem with the strategy is that it is very sensitive to the outcomes of the first few pulls. The probability of incurring major losses or even going bankrupt could be very high. Instead of using the above estimator of p, we will use the Krichevsky-Trofimov (KT) estimator. Essentially, it controls our losses while still converging to the Kelly bet as n goes to infinity. According to Cesa-Bianchi and Lugosi [2006], the KT estimator of p is

$$\hat{p}_n = \frac{n_1 + 1/2}{n + 1} \tag{2.9}$$

The strong law of large numbers implies that $\lim_{n\to\infty} \hat{p}_n = p$. Define $\hat{f}_n = 2\hat{p}_n - 1$, which is the proportion that we will bet on next pull. Note that f_1 is not well-defined, so we will just assume that we get to pull the arm once for free and call the outcome X_0. Then $\hat{p}_1$ and f_1 are well-defined. Our wealth after pull n is

$$W_n = W_0 \prod_{i=1}^{n} (1 + \hat{f}_i X_i) \tag{2.10}$$

Let $Y_i = (1 + X_i)/2$, $f_i = 2p_i - 1$, and define the loss $\ell(p, Y)$ as

$$\ell(p, Y) = -Y \log p - (1 - Y) \log(1 - p) \tag{2.11}$$

Take the logarithm of both sides of 2.10 to get

$$\begin{aligned}
\log W_n &= \log W_0 + \sum_{i=1}^{n} \log(1 + f_i X_i) && (2.12)\\
&= \log W_0 + \sum_{i=1}^{n} \frac{1 + X_i}{2} \log(2p_i) + \frac{1 - X_i}{2} \log(2(1 - p_i)) && (2.13)\\
&= \log W_0 + n \log 2 + \sum_{i=1}^{n} \frac{1 + X_i}{2} \log(p_i) + \frac{1 - X_i}{2} \log(1 - p_i) && (2.14)\\
&= \log W_0 + n \log 2 - \sum_{i=1}^{n} \ell(p_i, Y_i) && (2.15)
\end{aligned}$$

The bove holds for any choice of p_i, in particular it holds for $p_i = p$ and $p_i = \hat{p}_i$. Let W_n^* be our wealth from using the Kelly bet, and let W_n° be our wealth using the bet $\hat{f}_n = \hat{p}_n$. Define the regret R_n as

$$R_n = \log W_n^* - \log W_n^\circ = \sum_{i=1}^{n} \ell(\hat{p}_i, Y_i) - \ell(p, Y_i) \tag{2.16}$$

According to Cesa-Bianchi and Lugosi [2006] the KT estimator satisfies

$$\sum_{i=1}^{n} \ell(\hat{p}_i, Y_i) = \log \left[\frac{1}{\pi} \text{Beta}(n_1 + 1/2, n - n_1 + 1/2) \right] \tag{2.17}$$

If $p_i = p$, then

$$\sum_{i=1}^{n} \ell(p, Y_i) = \log \left[p^{n_1} (1 - p)^{n - n_1} \right] \tag{2.18}$$

Therefore

$$R_n = \log \frac{\pi p^{n_1}(1-p)^{n-n_1}}{\mathrm{Beta}(n_1 + 1/2, n - n_1 + 1/2)} \tag{2.19}$$

The right-hand side of the above expression attains its maximum at $n_1 = np$. Therefore we have the following bound

$$R_n \leq -nH(p) - \log \mathrm{Beta}(np + 1/2, n(1-p) + 1/2) + \log(\pi) \tag{2.20}$$

$H(p)$ is the entropy of a Bernoulli distribution with parameter p. If np is an integer, then the bound is sharp. Suppose $0 < p < 1$, then we can use Stirling's formula to approximate the beta function as $n \to \infty$. Substitute $n_1 = np$ into equation 2.19 to get

$$\begin{aligned}
R_n &\leq \log \left[\frac{\pi p^{np}(1-p)^{n(1-p)}}{\mathrm{Beta}(np + 1/2, n(1-p) + 1/2)} \right] && (2.21)\\
&\sim \log \left[\frac{\pi p^{np}(1-p)^{n(1-p)}}{\sqrt{2\pi}(np+1/2)^{np}(n(1-p)+1/2)^{n(1-p)}} (n+1)^{n+1/2} \right] && (2.22)\\
&= \log \left[\sqrt{\pi/2}(n+1)^{1/2} \left(1 + \frac{1}{n}\right)^{n} \left(1 + \frac{1}{2np}\right)^{-np} \left(1 + \frac{1}{2n(1-p)}\right)^{-n(1-p)} \right] && (2.23)\\
&\sim \log \left[\sqrt{\pi/2}(n+1)^{1/2} e e^{-1/2} e^{-1/2} \right] && (2.24)\\
&= \frac{1}{2}\log(n+1) + \frac{1}{2}\log(\pi/2) && (2.25)
\end{aligned}$$

If $p = 0$, then

$$\begin{aligned}
R_n &\leq \log \left[\frac{\pi}{\mathrm{Beta}(1/2, n + 1/2)} \right] && (2.26)\\
&\sim \frac{1}{2}\log(n + 1/2) + \frac{1}{2}\log(\pi) && (2.27)
\end{aligned}$$

If $p = 1$, we get the same answer.

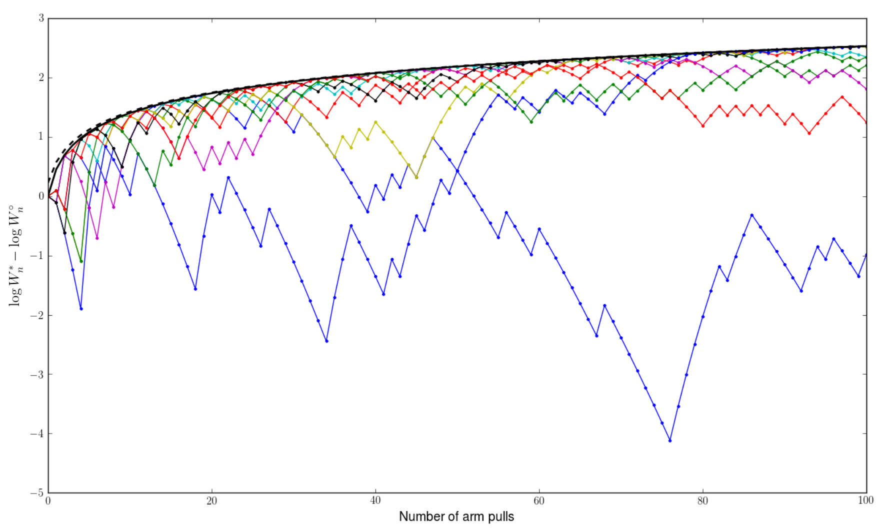

Figure 1: Comparison of regret paths to the KT regret bound for $p = 0.55$. The bold black curve on top is the regret bound $-nH(p) - \log \mathrm{Beta}(np + 1/2, n(1-p) + 1/2) + \log(\pi)$ from (2.20). The dashed black curve is the asymptotic bound (2.25). The others plots are the regrets of ten different paths.

In the above analysis, we left out a subtle point. If X_n is often equal to -1, then we could have $\hat{f}_n < 0$. The analysis still holds, but this only makes sense if we are allowed to "short" the arms. In other words, we can bet that we will see $X_n = -1$. If this is not possible due to the circumstances of the game, then it seems reasonable that we should either not bet at all or place a minimum bet.

3 Kelly-UCB algorithm for a multi-armed bandit

In this section, we consider a situation where we have a choice between many different arms. If we pull arm i, then we win with probability p_i and lose with probability $1 - p_i$. If we are only allowed to bet a fixed amount and we do not know the value of p_i for any i, then this is called the stochastic multi-armed bandit problem. There is a trade-off between exploitation and exploration: we want to exploit an arm with a seemingly high probability of winning, but we also want to explore and find the best arm.

There are many variations of the multi-armed bandit problem described in Bubeck et al. [2012]. Here we consider a new version of the stochastic bandit problem. Instead of fixed-size bets, we can bet any fraction of our wealth. The KL-UCB algorithm in Cappé et al. [2013] decides which arms to pull, but it does not address how much to bet. We also want to allow "shorting", which means that we gain our bet if a pull is a loss (i.e. bet negative amounts). We modified the KL-UCB algorithm to allow for variable bets and shorting.

3.1 Description of the Kelly-UCB algorithm

Suppose there are K levers that are indexed by $i = 1, \dots, K$. At each time $t = 1, 2, \dots$, we can choose which lever to pull. Let $\{X_{i,t}\}$ be independent random variables indicating the payoff of lever i on pull t, where $P(X_{i,t} = 1) = 1 - P(X_{i,t} = -1) = p_i$. We also need to decide how much to bet. We start with a finite amount of wealth, and we stop betting if we go bankrupt.

Let W_0 be our initial wealth, and let $f_{i,t}$ be the fraction of our wealth we bet on lever i at time t. Also let $I_t \in \{1, \dots, K\}$ be the index of the lever we pull at time t. Then

$$W_T = W_0 \prod_{t=1}^{T} (1 + f_{I_t,t} X_{I_t,t}) \tag{3.1}$$

Below is the Kelly-UCB algorithm. Note that our lever choice is given by a modified version of the KL-UCB algorithm in Cappé et al. [2013]. Let $\mathrm{kl}(p, q)$ denote the Kullback-Leibler divergence between two Bernoulli distributions with parameters p and q. We do not explicity address the possibility of $f = 0$ in the algorithm, but an easy fix is to just bet some extremely small minimal amount instead.

3.2 Analysis of the Kelly-UCB algorithm

Similar to Cappé et al. [2013] and Bubeck et al. [2012], we will bound the regret at time T, which we denote as R_T. Let $W_T(\Phi)$ be our wealth at time T when using strategy Φ, where Φ is any permissible strategy (it does not look into the future and does not bet more than we have). Let $\hat{\Phi}$ be the strategy given in algorithm 1. Define

$$\mathbb{E}[R_T] = \max_{\Phi} E[\log W_T(\Phi) - \log W_T(\hat{\Phi})] \tag{3.2}$$

Let $H(p)$ denote the entropy of a Bernoulli distribution with parameter p, and define $f_i = 2p_i - 1$ (f_i is the Kelly bet for arm i).

Theorem 3.1. *Let $i^* \in \arg\max_{i=1,\dots,K} |p_i - 1/2|$, and let Φ^* be the strategy where we always pull lever i^* and bet $f = 2p_{i^*} - 1$. Then $\Phi^* = \arg\max_{\Phi} \mathbb{E}[\log W_T(\Phi)]$ for any $T \geq 0$.*

Algorithm 1 Kelly-UCB

Require: $\epsilon > 0$ where $\epsilon << 1$, initial wealth $W > 0$

for $i = 1$ to K **do**
    $X \leftarrow \pm 1$ where $P(X = 1) = 1 - P(X = -1) = p_i$
    $W \leftarrow W(1 + \epsilon X)$
    $S_i \leftarrow (X + 1)/2$
    $N_i \leftarrow 1$
    $f_i \leftarrow X/2$
end for
for $t = K + 1$ to T **do**
    for $i = 1$ to K **do**
        $q_{i,\text{long}} \leftarrow \max\{q \in [0, 1] \mid N_i \text{kl}(S_i/N_i, q) \leq \log t\}$
        $q_{i,\text{short}} \leftarrow 1 - \min\{q \in [0, 1] \mid N_i \text{kl}(S_i/N_i, q) \leq \log t\}$
        $q_i \leftarrow \max(q_{i,\text{long}}, q_{i,\text{short}})$
    end for
    choose $I \in \arg\max_{i=1,\ldots,K} q_i$
    $X \leftarrow \pm 1$ where $P(X = 1) = 1 - P(X = -1) = p_I$
    $W \leftarrow W(1 + f_I X)$
    $S_I \leftarrow S_I + X + 1)/2$
    $N_I \leftarrow N_I + 1$
    $f_I \leftarrow (N_I + 1)^{-1}(N_I f_I + X)$
end for

Proof. We can define strategy Φ as the sequence $\{I_t, f_{I_t,t}\}$, so (3.1) implies

$$\begin{aligned}
\mathbb{E}[\log W_T(\Phi)] &= \log W_0 + \mathbb{E}\left[\sum_{i=1}^{T} \log(1 + f_{I_t,t} X_{I_t,t})\right] && (3.3)\\
&= \log W_0 + \sum_{i=1}^{T} \mathbb{E}\left[\mathbb{E}\left[\log(1 + f_{I_t,t} X_{I_t,t}) \mid I_t\right]\right] && (3.4)\\
&= \log W_0 + \sum_{i=1}^{T} \mathbb{E}\left[p_{I_t} \log(1 + f_{I_t,t}) + (1 - p_{I_t}) \log(1 - f_{I_t,t})\right] && (3.5)\\
&\leq \log W_0 + \sum_{i=1}^{T} \mathbb{E}\left[p_{I_t} \log(2p_{I_t}) + (1 - p_{I_t}) \log(2(1 - p_{I_t}))\right] && (3.6)\\
&= \log W_0 + T \log 2 - \sum_{i=1}^{T} H(p_{I_t}) && (3.7)\\
&\leq \log W_0 + T \log 2 - T H(p_{i^*}) && (3.8)\\
&= \mathbb{E}[\log W_T(\Phi^*)] && (3.9)
\end{aligned}$$

The third equality follows from the fact that conditioned on I_t, $f_{I_t,t}$ and $X_{I_t,t}$ are independent. The first inequality follows from theorem 2.1. $\square$

Theorem 3.2. *Let R_T be the regret defined in* (3.2)*. Define*

$$kl_i = kl(\max(p_i, 1 - p_i), \max(p_{i^*}, 1 - p_{i^*})) \tag{3.10}$$

Then for all $T \geq 0$

$$\mathbb{E}[R_T] \leq \frac{1}{2}\log(T + 1) + \sum_{i=1}^{K}\left[\frac{\Delta_i \log T}{kl_i}(1 + o(1)) + \frac{1}{2}\log\left(\frac{\log T}{kl_i} + 1\right)\right] + \frac{K}{2}\log(\pi/2) \tag{3.11}$$

Proof. Continuing from (3.2) and using theorem 3.1 leads to

$$\mathbb{E}[R_T] = \mathbb{E}\left[\log W_T(\Phi^*) - \log W_T(\hat{\Phi})\right] \tag{3.12}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} \log \frac{1 + f_{i^*} X_{i^*,t}}{1 + \hat{f}_{I_t,t} X_{I_t,t}}\right] \tag{3.13}$$

$$= \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \log \frac{1 + f_{i^*} X_{i^*,t}}{1 + f_{I_t} X_{I_t,t}}\right]}_{:=A} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{T} \log \frac{1 + f_{I_t} X_{I_t,t}}{1 + \hat{f}_{I_t,t} X_{I_t,t}}\right]}_{:=B} \tag{3.14}$$

We will simplify A and B separately. We can think of A as the regret caused by choosing the wrong arm, and B is the regret caused by using an approximation of the Kelly bet.

$$A = \sum_{t=1}^{T} \mathbb{E}[\log(1 + f_{i^*} X_{i^*,t})] - \sum_{t=1}^{T} \mathbb{E}[\log(1 + f_{I_t} X_{I_t,t})] \tag{3.15}$$

$$= \sum_{t=1}^{T} [\log(2) - H(p_{i^*})] - \sum_{t=1}^{T} \mathbb{E}[\log(2) - H(p_{I_t})] \tag{3.16}$$

$$= \sum_{t=1}^{T} \mathbb{E}[-H(p_{i^*}) - H(p_{I_t})] \tag{3.17}$$

$$= \sum_{i=1}^{K} \mathbb{E}\left[\sum_{j=1}^{N_i(t)} -H(p_{i^*}) - H(p_i)\right] \tag{3.18}$$

$$= \sum_{i=1}^{K} \mathbb{E}[N_i(t)]\Delta_i \tag{3.19}$$

where $\Delta_i = -H(p_{i^*}) + H(p_i) \geq 0$. As for B

$$B = \mathbb{E}\left[\sum_{t=1}^{T} \log \frac{1 + f_{I_t} X_{I_t,t}}{1 + \hat{f}_{I_t,t} X_{I_t,t}}\right] \tag{3.20}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} \frac{1 + X_{I_t,t}}{2} \log\left(\frac{1 + f_{I_t}}{1 + \hat{f}_{I_t,t}}\right) + \frac{1 - X_{I_t,t}}{2} \log\left(\frac{1 - f_{I_t}}{1 - \hat{f}_{I_t,t}}\right)\right] \tag{3.21}$$

$$\tag{3.22}$$

Define $\hat{p}_{I_t,t}$ such that $\hat{f}_{I_t,t} = 2p_{I_t,t} - 1$, and let $Z_i \sim \text{Bernoulli}(p_i)$. Then continuing from above we get

$$B = \mathbb{E}\left[\sum_{t=1}^{T} Z_i \log\left(\frac{p_{I_t}}{\hat{p}_{I_t,t}}\right) + (1 - Z_i) \log\left(\frac{1 - p_{I_t}}{1 - \hat{p}_{I_t,t}}\right)\right] \tag{3.23}$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} -\ell(p_{I_t}, Z_{I_t}) + \ell(\hat{p}_{I_t,t}, Z_{I_t})\right] \tag{3.24}$$

$$= \sum_{i=1}^{K} \mathbb{E}\left[\sum_{j=1}^{N_i(t)} -\ell(p_i, Z_i) + \ell(\hat{p}_{i,t}, Z_i)\right] \tag{3.25}$$

Note that $\ell(p_i, y_i)$ was defined in (2.11). Combining the results for A and B leads to

$$\mathbb{E}[R_T] = \sum_{i=1}^{K} \mathbb{E}[N_i(t)]\Delta_i + \sum_{i=1}^{K} \mathbb{E}\left[\sum_{j=1}^{N_i(t)} -\ell(p_i, Z_i) + \ell(\hat{p}_{i,t}, Z_i)\right] \tag{3.26}$$

The rightmost sum appears because we are using an estimate of the Kelly bet. If $f_{i,t}$ is the exact Kelly bet f_i, then that sum is zero. Applying (2.20) to (3.26) leads to

$$\mathbb{E}[R_T] \leq \sum_{i=1}^{K} \mathbb{E}[N_i(T)]\Delta_i + \sum_{i=1}^{K} \mathbb{E}\left[\frac{1}{2}\log(N_i(t)+1) + \frac{1}{2}\log(\pi/2) + o(1)\right] \tag{3.27}$$

$$\leq \sum_{i=1}^{K}\left[\mathbb{E}[N_i(T)]\Delta_i + \frac{1}{2}\log(\mathbb{E}[N_i(T)]+1)\right] + \frac{K}{2}\log(\pi/2) + o(1) \tag{3.28}$$

The second inequality is Jensen's inequality. Cappé et al. [2013] showed that if $f_{I_t,t} \geq 0$, and $i \neq i^*$

$$\mathbb{E}[N_i(t)] \leq \frac{\log T}{\mathrm{kl}(p_i, p_{i^*})}(1 + o(1)) \tag{3.29}$$

Since we allow $f_{I_t,t} < 0$, we need to adjust the inequality. Let

$$\mathrm{kl}_i = \mathrm{kl}(\max(p_i, 1-p_i), \max(p_{i^*}, 1-p_{i^*})) \tag{3.30}$$

Lemma 3.1 shows that for $i \neq i^*$

$$\mathbb{E}[N_i(t)] \leq \frac{\log T}{\mathrm{kl}_i}(1 + o(1)) \tag{3.31}$$

And since $N_{i^*}(T) \leq T$, we have

$$\mathbb{E}[R_T] \leq \frac{1}{2}\log(T+1) + \sum_{i=1}^{K}\left[\frac{\Delta_i \log T}{\mathrm{kl}_i}(1+o(1)) + \frac{1}{2}\log\left(\frac{\log T}{\mathrm{kl}_i} + 1\right)\right] + \frac{K}{2}\log\left(\frac{\pi}{2}\right) \tag{3.32}$$

□

The $\frac{1}{2}\log(T+1)$ term is generally dominated by the other $\log T$ term inside the sum. In other words, the error from choosing the wrong arm generally dominates the error from betting incorrectly. Numerical results for $\vec{p} = (0.4, 0.5, 0.8)$ are shown in figure 2.

Lemma 3.1. *Define* $kl_i = kl(\max(p_i, 1-p_i), \max(p_{i^*}, 1-p_{i^*}))$. *Then*

$$\mathbb{E}[N_i(t)] \leq \frac{\log T}{kl_i}(1 + o(1))$$

Proof. adapt the Cappe proof □

We would also like to find a lower bound for the regret. Cappé et al. [2013] gives a lower bound for the number times each suboptimal arm is pulled. We do not find a lower bound here. However, we could get a lower bound for the Kelly-UCB regret by letting $\hat{f}_{i,t} = 2p_i - 1$ for all t, and then we use the aforementioned lower bound in Cappé et al. [2013].

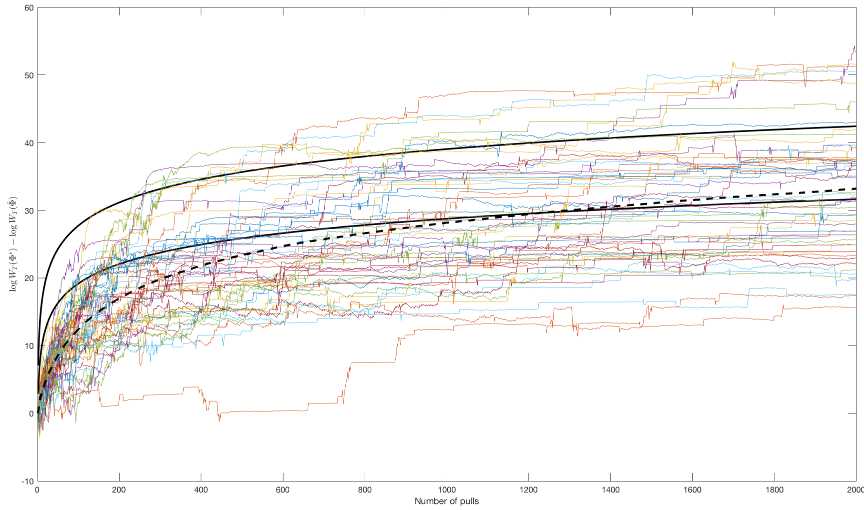

Figure 2: $\vec{p} = (0.4, 0.5, 0.6, 0.8)$. The top solid black curve is the expected regret bound(3.32). The bottom solid black curve is the lower bound we get if we always use the Kelly bet (i.e. the regret only accumulates from choosing the wrong arm). The dashed black line is the mean regret of 2000 paths. The other plots are 50 randomly chosen regret paths.

References

S. Bubeck, N. Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, G. Stoltz, et al. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

S. Ethier. *The doctrine of chances: probabilistic aspects of gambling*. Springer Science & Business Media, 2010.

D. Glimne. Slot machine. `https://www.britannica.com/topic/slot-machine`, 2015. [Online; accessed 2017-10-28].

J. L. Kelly. A new interpretation of information rate. *Bell Labs Technical Journal*, 35(4):917–926, 1956.

E. O. Thorp. The kelly criterion in blackjack, sports betting and the stock market. *Handbook of asset and liability management*, 1:385–428, 2006.