COMP 400 Project Report

A Theoretical Analysis of Upper Confidence Bound applied to Trees

Submitted by

Yutong Yan

School of Computer Science
McGill University
Montreal, Quebec
December 2019

Contents

1 UCB1 with Markov's Inequality **2**

2 UCB1 with Chernoff-Hoeffding Inequality **6**

3 Upper Confidence Bounds applied to Trees (UCT) **9**

4 UCT with Laplace Bound **18**

Chapter 1

UCB1 with Markov's Inequality

Regret analysis guidelines:

1. Decompose the regret over the arms.
2. On a "good" event prove that the sub-optimal arms are not played too often.
3. Show that the "good" event occurs with high probability.

Suppose $X_1, ..., X_T$ is a sequence of independent Gaussian random variables, with mean μ and variance 1. The mean is $\hat{\mu} = \frac{1}{T}\sum_{t=1}^{T} X_t$. For any $\delta \in (0, 1)$,

$$\mathbb{P}(\hat{\mu} \geq \mu + \sqrt{\frac{2\log(1/\delta)}{T}}) \leq \delta \tag{1.1}$$

Similarly,

$$\mathbb{P}(\hat{\mu} \leq \mu - \sqrt{\frac{2\log(1/\delta)}{T}}) \leq \delta \tag{1.2}$$

Markov inequality:

$$\mathbb{P}(X \geq C) \leq \frac{\mathbb{E}[X]}{C} \tag{1.3}$$

If $X \sim \mathcal{N}(0,1)$,

$$\begin{aligned}
\mathbb{P}(\hat{\mu} \geq \mu + \epsilon) &= \mathbb{P}(\frac{1}{T}\textstyle\sum_{i=1}^{T} X_t \geq \mu + \epsilon) \\
&= \mathbb{P}(\textstyle\sum_{i=1}^{T}(X_t - \mu) \geq \epsilon T) \\
&= \mathbb{P}\big(\exp(\lambda \textstyle\sum_{i=1}^{T}(X_t - \mu)) \geq \exp(\lambda \epsilon T)\big) \\
&\text{Apply Markov Inequality} \\
&\leq \exp(-\lambda\epsilon T)\mathbb{E}[\exp(\lambda \textstyle\sum_{i=1}^{T}(X_t - \mu))] \\
&= \exp(-\lambda\epsilon T)\textstyle\prod_{i=1}^{T}\exp(\lambda^2/2) \\
&= \exp(-\epsilon T\lambda + \frac{\lambda^2 T}{2}) \\
&\text{To minimize the above, choose } \lambda = \epsilon \\
&= \exp(-\frac{\epsilon^2 T}{2}) \\
&\text{According to Equation 1.1 and 1.2} \\
&= \delta
\end{aligned} \tag{1.4}$$

Regret decomposition:
Define $\Delta_a = \mu^\star - \mu_a$ and $T_a(t) = \sum_{s=1}^{t} \mathbb{1}(A_s = a)$

$$\begin{aligned}
\mathcal{R}_n &= n\mu^\star - \mathbb{E}[\sum_{t=1}^{n} R_t] \\
&= \mathbb{E}[\sum_{t=1}^{n}(\mu^\star - R_t)] \\
&= \mathbb{E}[\sum_{t=1}^{n} \Delta_{A_t}] \\
&= \mathbb{E}[\sum_{t=1}^{n}\sum_{a\in\mathcal{A}} \mathbb{1}(A_t = a)\Delta_a] \\
&= \sum_{a\in\mathcal{A}} \Delta_a \mathbb{E}[T_a(n)]
\end{aligned} \tag{1.5}$$

Assumption 1. Assume that the estimated value of arm a is not too large. In other words, it is smaller than the true value plus the confidence interval.

$$\mu_a + \sqrt{\frac{2\log(1/\delta)}{T_a(t-1)}} \geq \hat{\mu}_a(t-1) \tag{1.6}$$

Assumption 2. Assume that the estimated value of the optimal arm $a^\star$ is not too small, so that it would not be underestimated and not selected.

$$\hat{\mu}_{a^\star}(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_{a^\star}(t-1)}} \geq \mu^\star \tag{1.7}$$

Now suppose $A_t = a$ in round t,

$$\begin{aligned}
&\text{According to Equation 1.6} \\
&\mu_a + 2\sqrt{\frac{2\log(1/\delta)}{T_a(t-1)}} \geq \hat{\mu}_a(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_a(t-1)}} \\
&\text{Optimal arm is not selected} \\
&\geq \hat{\mu}_{a^\star}(t-1) + \sqrt{\frac{2\log(1/\delta)}{T_{a^\star}(t-1)}} \\
&\geq \mu_{a^\star} \\
&= \mu_a + \Delta_a
\end{aligned} \tag{1.8}$$

Therefore, we can obtain

$$2\sqrt{\frac{2\log(1/\delta)}{T_a(t-1)}} \geq \Delta_a \tag{1.9}$$

Simultaneously,

$$T_a(t-1) \leq \frac{8\log\frac{1}{\delta}}{\Delta_a^2} \tag{1.10}$$

If **good** event happens (Assumption 1 and 2):

$$T_a(n) \leq 1 + \frac{8\log\frac{1}{\delta}}{\Delta_a^2} \tag{1.11}$$

Let $\hat{\mu}_{a,s}$ be the empirical mean of arm a after s plays. The concentration theorem shows that

$$\mathbb{P}(\hat{\mu}_{a,s} \geq \mu_a + \sqrt{\frac{2\log(1/\delta)}{s}}) \leq \delta \tag{1.12}$$

Combining with a union bound, defined as $\mathbb{P}(\bigcup_i B_i) \leq \sum_i \mathbb{P}(B_i)$

$$\mathbb{P}(\exists\, s \leq n : \hat{\mu}_{a,s} \geq \mu_a + \sqrt{\frac{2\log(1/\delta)}{s}}) \leq n\delta \tag{1.13}$$

Now we know when the **bad** event happens, a sub-optimal arm can be played n times with probability $2n\delta$, where 2 means we have to apply concentration bounds upper and lower. The regret

bound is then

$$
\begin{aligned}
\mathcal{R}_n &= \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)] \\
&\leq \sum_{a \in \mathcal{A}: \Delta_a > 0} \Delta_a \big(2\delta n^2 + 1 + \frac{8 \log(1/\delta)}{\Delta_a^2}\big) \\
&\quad \text{Choose } \delta = 1/n^2 \\
&\leq \sum_{a \in \mathcal{A}: \Delta_a > 0} 3\Delta_a + \frac{16 \log(n)}{\Delta_a}
\end{aligned}
\tag{1.14}
$$

Chapter 2

UCB1 with Chernoff-Hoeffding Inequality

Theorem 1. For all $K \geq 1$, if policy UCB1 is run on K machines having arbitrary reward distribution $P_1, ..., P_k$ with support in $[0, 1]$, then its expected regret after any number n of plays is at most

$$\Big[8 \sum_{i:\mu_i<\mu^\star} \Big(\frac{\ln n}{\Delta_i}\Big)\Big] + \Big(1 + \frac{\pi^2}{3}\Big)\Big(\sum_{j=1}^{K} \Delta_j\Big) \tag{2.1}$$

where $\mu_1, ..., \mu_k$ are the expected values of $P_1, ..., P_k$.

Fact 2. (Chernoff-Hoeffding bound) Let $X_1, ..., X_n$ be random variables with common range $[0, 1]$ and such that $\mathbb{E}[X_t|X_1, ..., X_{t-1}] = \mu$. Let $S_n = X_1 + \cdots + X_n$. Then for all $a \geq 0$

$$\mathbb{P}\big(S_n \geq n\mu + a\big) \leq \mathrm{e}^{-2a^2/n} \tag{2.2}$$

Similarly,

$$\mathbb{P}\big(S_n \leq n\mu - a\big) \leq \mathrm{e}^{-2a^2/n} \tag{2.3}$$

Proof. Let $c_{t,s} = \sqrt{\frac{2 \ln t}{s}}$. We need to upper bound $T_i(n)$ on any sequence of plays. More precisely, for each $t \geq 1$, we bound the indicator function of $I_t = i$ as follows. Let l be any positive inteter.

$$T_i(n) = 1 + \sum_{t=K+1}^{n} \{I_t = i\} \quad \text{Every arm visited once}$$

Sum over how many times an arm gets visited after l plays

$$\leq l + \sum_{t=K+1}^{n} \{I_t = i, T_i(t-1) \geq l\}$$

Sum over how many times the UCB of the selected arm larger than the optimal arm after l plays

$$\leq l + \sum_{t=K+1}^{n} \left\{ \overline{X}^{\star}_{T^{\star}(t-1)} + c_{t-1,T^{\star}(t-1)} \leq \overline{X}_{i,T_i(t-1)} + c_{t-1,T_i(t-1)}, T_i(t-1) \geq l \right\}$$

For any t, the optimal arm $i^{\star}$ can be played $s$times, where $s \in (0, t)$

The arm i can be played s_i times, where $s_i \in (l, t)$

Therefore, if the worst case happens,

the minimum value of the optimal arm is less than the maximum value of the arm in their time step range

$$\leq l + \sum_{t=K+1}^{n} \left\{ \min_{0<s<t} \overline{X}^{\star}_{s} + c_{t-1,s} \leq \max_{l \leq s_i < t} \overline{X}_{i,s_i} + c_{t-1,s_i} \right\}$$

Rearrange

$$\leq l + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=l}^{t-1} \left\{ \overline{X}^{\star}_{s} + c_{t,s} \leq \overline{X}_{i,s_i} + c_{t,s_i} \right\} \tag{2.4}$$

Now observe that $\overline{X}^{\star}_{s} + c_{t,s} \leq \overline{X}_{i,s_i} + c_{t,s_i}$ implies one of the following must hold

$$\overline{X}^{\star}_{s} \leq \mu^{\star} - c_{t,s} (\text{Underestimate optimal arm}) \tag{2.5}$$

$$\overline{X}_{i,s_i} \geq \mu_i + c_{t,s_i} \tag{2.6}$$

$$\mu^{\star} < \mu_i + 2c_{t,s_i} \tag{2.7}$$

First, bound the probability events of 2.5 and 2.6 using Fact 2

$$\mathbb{P}\left(\overline{X}^{\star}_{s} \leq \mu^{\star} - c_{t,s} \right) \leq \mathrm{e}^{-4 \ln t} = t^{-4} \tag{2.8}$$

$$\mathbb{P}\left(\overline{X}_{i,s_i} \geq \mu_i + c_{t,s_i} \right) \leq \mathrm{e}^{-4 \ln t} = t^{-4} \tag{2.9}$$

To make 2.7 false, $l = \lceil (8 \ln n)/\Delta_i^2 \rceil$

$$\begin{aligned} \mu^\star - \mu_i - 2c_{t,s_i} &= \mu^\star - \mu_i - 2\sqrt{2(\ln t)/s_i} \\ &\geq \mu^\star - \mu_i - \Delta_i \\ &= 0 \end{aligned} \tag{2.10}$$

Therefore, for $s_i \geq (8 \ln n)/\Delta_i^2$, we can get

$$\begin{aligned} \mathbb{E}[T_i(n)] &\leq \lceil \frac{8 \ln n}{\Delta_i^2} \rceil + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i = \lceil (8 \ln n)/\Delta_i^2 \rceil}^{t-1} \Big(\mathbb{P}(\overline{X}_s^\star \leq \mu^\star - c_{t,s}) + \mathbb{P}(\overline{X}_{i,s_i} \geq \mu_i + c_{t,s_i}) \Big) \\ &\leq \lceil \frac{8 \ln n}{\Delta_i^2} \rceil + \sum_{t=1}^{\infty} \sum_{s=1}^{t} \sum_{s_i=1}^{t} 2t^{-4} \\ &\leq \frac{8 \ln n}{\Delta_i^2} + 1 + \frac{\pi^2}{3} \end{aligned} \tag{2.11}$$

□

Chapter 3

Upper Confidence Bounds applied to Trees (UCT)

Problem definition:

Consider a bandit problem with K arms, defined by the sequence of random payoffs X_{it}, $i = 1, ..., K$, $t \geq 1$, where each i is the index of a gambling machine (the "arm" of a bandit). Successive plays of machine i yield the payoffs X_{i1}, $X_{i2}, \cdots$. For simplicity, we shall assume that X_{it} lies in the interval $[0, 1]$. An allocation policy is a mapping that selects the next arm to be played based on the sequence of past selections and the payoffs obtained. The expected regret of an allocation policy A after n plays is defined by Expected cumulative regret is defined as

$$\mathcal{R} = \max_i \mathbb{E}[\sum_{t=1}^{n} X_{it}] - \mathbb{E}[\sum_{i=1}^{K} \sum_{t=1}^{T_i(n)} X_{i,t}], \tag{3.1}$$

where $T_i(n) = \sum_{s=1}^{n} \mathbb{1}(\boldsymbol{I}_s = i)$ is the number of times arm i was played up to time n, $I_t \in \{1, ..., K\}$ is the index of the arm selected at time t.

Remark 1. There is no policy whose regret would grow slower than $\mathcal{O}(\ln n)$ for a large class of payoff distributions [1].

Algorithm UCB1, whose finite-time regret is studied in [2]. It chooses the arm with the best upper confidence bound:

$$\boldsymbol{I}_t = \underset{i \in \{1,...,K\}}{\arg\max} \{\overline{X}_{i,T_i(t-1)} + c_{t-1,T_i(t-1)}\}, \tag{3.2}$$

where $c_{t,s}$ is a bias sequence chosen to be

$$c_{t,s} = \sqrt{\frac{2 \ln t}{s}} \tag{3.3}$$

The bias sequence is such that if X_{it} were i.i.d. (or form a martingale difference process shifted by a constant) then the inequalities

$$\mathbb{P}(\overline{X}_{is} \geq \mu_i + c_{t,s}) \leq t^{-4} \tag{3.4}$$

and

$$\mathbb{P}(\overline{X}_{is} \leq \mu_i + c_{t,s}) \leq t^{-4} \tag{3.5}$$

This follows from Hoeffding's (or more generally, the Hoeffding Azuma) inequality (see Lemma 8). Unlike in [2], we we allow the mean-value of the payoffs Xi· to drift as a function of time.

Laplace bound from Chapter 2 in [3]:

$$\mathbb{P}\Big(\frac{1}{N}\sum_{i=1}^{N} Z_i \geq \sigma\sqrt{2(1+\frac{1}{N})\frac{\ln(\sqrt{N+1}/\delta)}{N}}\Big) \leq \delta \tag{3.6}$$

The expected values of the averages

$$\overline{X}_{in} = \frac{1}{n}\sum_{t=1}^{n} X_{it} \tag{3.7}$$

converge. We let $\mu_{in} = \mathbb{E}[\overline{X}_{in}]$ and

$$\mu_i = \lim_{n\to\infty} \mu_{in}. \tag{3.8}$$

Further, we define δ_{in} by (Non-stationarity)

$$\mu_{in} = \mu_i + \delta_{in} \tag{3.9}$$

We start by analyzing UCB1 for non-stationary bandit problems. Remember that by assumption $0 \leq X_{it} \leq 1$. Quantities related to the optimal arm shall be upper indexed by a star, e.g., $\mu^\star, T^\star(t), \overline{X}_t^\star$, etc. For the sake of easy referencing, we summarize the assumptions on the rewards here:

Assumption 2. Fix $1 \leq i \leq K$. Let $\{\mathcal{F}_{it}\}_t$ be a filtration such that X_{itt} is $\{\mathcal{F}_{it}\}_t$-adapted and $X_{i,t}$ is conditionally independent of $\mathcal{F}_{i,t+1}$, $\mathcal{F}_{i,t+2}$, ... given $\mathcal{F}_{i,t-1}$. Then $0 \leq X_{it} \leq 1$ and the limit of $\mu_{in} = \mathbb{E}[\overline{X}_{in}]$ exists. Further, we assume that there exists a constant $C_p > 0$ and an integer N_p such that for $n \geq N_p$, for any $\delta > 0$, $\Delta_n(\delta) = C_p\sqrt{n\ln(1/\delta)}$, the following bounds hold:

$$\mathbb{P}(n\overline{X}_{in} \geq n\mathbb{E}[\overline{X}_{in}] + \Delta_n(\delta)) \leq \delta \tag{3.10}$$

and

$$\mathbb{P}(n\overline{X}_{in} \leq n\mathbb{E}[\overline{X}_{in}] - \Delta_n(\delta)) \leq \delta \tag{3.11}$$

Set δ to t^{-4}, according to Hoeffding's inequality,

$$c_{t,s} = \Delta_s(t^{-4})/s = C_p\sqrt{s \cdot 4 \cdot \ln(t)}/s = 2C_p\sqrt{\frac{\ln(t)}{s}} \tag{3.12}$$

We let $\Delta_i = \mu^\star - \mu_i$.

Definition 3. Convergence of δ_{it} with ϵ

Since δ_{it} converges by assumption to zero, for all $\epsilon > 0$ there exists an index $N_0(\epsilon)$ such that if $t \geq N_0(\epsilon)$ then $|\delta_{it}| \leq \epsilon\Delta_i/2$ and $|\delta_{j^\star,t}| \leq \epsilon\Delta_i/2$, whenever i is the index of a sub-optimal arm and $j^\star$ is the index of an optimal arm.

In particular, it follows that for any optimal arm $j^\star$, $t \geq N_0(\epsilon)$, $|\delta_{j^\star,t}| \leq \epsilon/2 \min_{\{i|\Delta_i>0\}} \Delta_i$.

Theorem 4. Consider UCB1 applied to a non-stationary problem where the pay-off sequence satisfies Assumption 2 and where the bias sequence $c_{t,s}$, used by UCB1 is given by Equation 3.12. Fix $\epsilon > 0$. Let $T_i(n)$ denote the number of plays of arm i. Then if i is the index of a suboptimal arm then

$$\mathbb{E}[T_i(n)] \leq \frac{16C_p^2 \ln n}{(1-\epsilon)^2\Delta_i^2} + N_0(\epsilon) + N_p + 1 + \frac{\pi^2}{3} \tag{3.13}$$

Note that here $N_0(\epsilon)$ comes from Definition 3 and N_p comes from Assumption 2.

Proof. Fix the index i of a suboptimal arm. We follow the proof of Theorem 1 in [2]. Let

$$A_0(n,\epsilon) = \min\{s|c_{t,s} \leq (1-\epsilon)\Delta_i/2\} \tag{3.14}$$

Note here the reason why $(1-\epsilon)$ is related to bound bad events later.

By the definition of $c_{t,s}$, $A_0(n,\epsilon) = \left\lceil \frac{16C_p^2 \ln n}{(1-\epsilon)^2\Delta_i^2} \right\rceil$. We let

$$A(n,\epsilon) = \max(A_0(n,\epsilon), N_0(\epsilon), N_p) \tag{3.15}$$

By definition,

$$\begin{aligned} T_i(n) &= 1 + \sum_{t=K+1}^{n} \mathbb{I}(I_t = i) \\ &\leq A(n,\epsilon) + \sum_{t=K+1}^{n} \mathbb{I}(I_t = i, T_i(t-1) \geq A(n,\epsilon)) \\ &\leq A(n,\epsilon) + \sum_{t=1}^{n}\sum_{s=1}^{t-1}\sum_{s'=A(n,\epsilon)} \mathbb{I}(\overline{X}_s^\star + c_{t,s} \leq \overline{X}_{i,s'} + c_{t,s'}) \end{aligned} \tag{3.16}$$

- For $n \geq t \geq s' \geq A(n,\epsilon)$, we have $\mu_t^\star \geq \mu_{it} + 2c_{t,s'}$. (Note that we assume after n time steps, we will not play sub-optimal arms.)
- Indeed, since $n \geq t$ and $c_{t,s}$ increases in t, $c_{t,s'} \leq c_{n,s'}$.

- Since $c_{t,s}$ decreases in s, and $s' \geq A(n,\epsilon) \geq A_0(n,\epsilon)$, $c_{n,s'} \leq c_{n,A_0(n,\epsilon)}$.

By the definition of A_0, $c_{n,A_0(n,\epsilon)} \leq (1-\epsilon)\Delta_i/2$ (Equation 3.14)
Hence, $2c_{t,s'} \leq (1-\epsilon)\Delta_i$. $(c_{n,s'} \leq c_{n,A_0(n,\epsilon)})$
Since $t \geq A(n,\epsilon) \geq N_0(\epsilon)$, we have that $\delta_{it} \leq \epsilon\Delta_i$. (if $t \geq N_0(\epsilon)$ then $|\delta_{it}| \leq \epsilon\Delta_i/2$)
Hence, $\mu_t^\star - \mu_{it} - 2c_{t,s'} = \Delta_i - |\delta_t^\star| - \delta_{it} - 2c_{t,s'} \geq \Delta_i - \epsilon\Delta_i - (1-\epsilon)\Delta_i = 0$. $(\mu_{in} = \mu_i + \delta_{in})$
From [2], observing $\overline{X}_s^\star + c_{t,s} \leq \overline{X}_{i,s_i} + c_{t,s}$, one of the following must hold

$$\overline{X}_s^\star \leq \mu^\star - c_{t,s} \tag{3.17}$$

$$\overline{X}_{i,s_i} \geq \mu_i + c_{t,s_i} \tag{3.18}$$

$$\mu^\star < \mu_i + 2c_{t,s_i} \tag{3.19}$$

Hence, $\mathbb{I}(\overline{X}_s^\star + c_{t,s} \leq \overline{X}_{i,s_i} + c_{t,s}) \leq \mathbb{I}(\overline{X}_s^\star \leq \mu^\star - c_{t,s}) + \mathbb{I}(\overline{X}_{i,s_i} \geq \mu_i + c_{t,s_i})$.

Similar to chapter 1, now we can bound the two bad events using Equation 3.4 and 3.5, which is equivalent to $2t^{-4}$.

$$\begin{aligned}\mathbb{E}[T_i(n)] &\leq A(n,\epsilon) + 1 + \frac{\pi^2}{3} \\ &\leq \left\lceil \frac{16C_p^2 \ln n}{(1-\epsilon)^2\Delta_i^2} \right\rceil + N_0(\epsilon) + N_p + 1 + \frac{\pi^2}{3}\end{aligned} \tag{3.20}$$

$N_0(\epsilon)$ and N_p are included since Equation 3.15.

□

Theorem 5. Let

$$\overline{X}_n = \sum_{i=1}^{K} \frac{T_i(n)}{n} \overline{X}_{i,T_i(n)} \tag{3.21}$$

Under the assumption of Theorem 4,

$$|\mathbb{E}[\overline{X}_n] - \mu^\star| \leq |\delta_n^\star| + \mathcal{O}\left(\frac{K(C_p^2 \ln n + N_0)}{n}\right), \tag{3.22}$$

where $N_0 = N_0(1/2)$.

Proof. W.l.o.g, we assume there is a unique optimal arm, denoted by arm $i^\star$.
By triangle inequality, $|\mu^\star - \mathbb{E}[\overline{X}_n]| \leq |\overline{\mu}^\star - \overline{\mu}_n^\star| + |\overline{\mu}_n^\star - \mathbb{E}[\overline{X}_n]|$.

$$n|\overline{\mu}^{\star}_n - \mathbb{E}[\overline{X}_n| = \Big|\sum_{i=1}^{n}\mathbb{E}[X^{\star}_t] - \mathbb{E}\Big[\sum_{i=1}^{K}\overline{X}_{i,T_i(n)}\Big]\Big|$$

$$\text{Separate optimal arm and sub-optimal arms} \tag{3.23}$$

$$= \Big|\sum_{i=1}^{n}\mathbb{E}[X^{\star}_t] - \mathbb{E}\Big[T*(n)\overline{X}^{\star}_{T^{\star}(n)}\Big]\Big| + \mathbb{E}\Big[\sum_{i=1,i\neq i^{\star}}^{K}\overline{X}_{i,T_i(n)}\Big]$$

According to Theorem 4, and by the assumption that $0 \leq \overline{X}_{i,T_i(n)} \leq 1$, the second term is bounded by $\mathcal{O}\Big(K(C_p^2 \ln n + N_0)\Big)$.

According to Theorem 4, the parameters we can have control are ϵ, C_p, and n. Therefore, N_p disappears here.

In order to bound the first term in Equation 3.23, let us note that $T^{\star}(n)\overline{X}^{\star}_{T^{\star}(n)} = \sum_{t=1}^{T^{\star}(n)} X^{\star}_t$ and we have

$$\begin{aligned} D_n &\stackrel{\text{def}}{=} \sum_{t=1}^{n}\mathbb{E}[X^{\star}_t] - \mathbb{E}\left[\sum_{t=1}^{T^{\star}(n)} X^{\star}_t\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{n} X^{\star}_t - \sum_{t=1}^{T^{\star}(n)} X^{\star}_t\right] \\ &= \mathbb{E}\left[\sum_{t=T^{\star}(n)+1}^{n} X^{\star}_t\right] \end{aligned} \tag{3.24}$$

Since $0 \leq \overline{X}_{i,T_i(n)} \leq 1$, we can bound the term from above by $\mathbb{E}[n - T^{\star}(n)]$, which is just $\sum_{i\neq i^{\star}}\mathbb{E}[T_i(n)]$ and hence by Theorem 4, $D_n = \mathcal{O}\Big(K(C_p^2 \ln n + N_0)\Big)$ □

Theorem 6. (Lower Bound) Under the assumptions of Theorem 4, there exists some positive constant ρ such that for all arms i and n, $T_i(n) \geq \lceil \rho \log(n) \rceil$.

Proof. □

Theorem 7. Fix any arbitrary $\delta > 0$ and let $\Delta = 9\sqrt{2n\ln(2/\delta)}$. Let n_0 be such that

$$\sqrt{n_0} \geq \mathcal{O}(K(C_p^2 \ln n_0 + N_0(1/2))) \tag{3.25}$$

Then for any $n \geq n_0$, under the assumptions of Theorem 4 the following bounds hold true:

$$\mathbb{P}(n\overline{X}_n \geq n\mathbb{E}[\overline{X}_n] + \Delta_n) \leq \delta \tag{3.26}$$

$$\mathbb{P}(n\overline{X}_n \leq n\mathbb{E}[\overline{X}_n] - \Delta_n) \leq \delta \tag{3.27}$$

Proof. □

Theorem 8. (Convergence of Failure Probability) Under the assumptions of Theorem 4 it

holds that

$$\lim_{t\to\infty} P(I_t \neq i^\star) = 0 \tag{3.28}$$

Proof. □

Theorem 9. Consider algorithm UCT running on a game tree of Depth D, branching factor K with stochastic payoffs at the leaves. Assume that the payoffs lie in the interval $[0,1]$. Then the bias of the estimated expected payoff, $\overline{X}_n$, is $\mathcal{O}((KD\log(n)+K^D)/n)$. Further, the failure probability at the root converges to zero as the number of samples grows to infinity.

Proof. □

Let $\mathcal{F}_t$ denote a filtration over some probability space, Y_t be an $\mathcal{F}_t$-adapted real valued martingale-difference sequence. Define the partial sum martingale $S_n = \sum_{t=1}^n Y_t$, $n \geq 1$. Use Hoeffding-Azuma inequality:

Lemma 10. (Hoeffding-Azuma inequality) If Y_n is a martingale difference with $|Y_i| \leq C$, a.s., $i = 1, 2, ...$, where C is a positive real number, then

$$\mathbb{P}(S_n \geq \epsilon n) \leq \exp(-\frac{2n\epsilon^2}{C^2}) \tag{3.29}$$

Similarly,

$$\mathbb{P}(S_n \leq -\epsilon n) \leq \exp(-\frac{2n\epsilon^2}{C^2}) \tag{3.30}$$

Tail inequality for stopped martingales?

Lemma 11. Let N be an integer-valued random variable and let S_t be an $\mathcal{F}_t$-adapted real-valued process (not necessarily a martingale) $(t = 0, 1, 2, ...)$, which is centered: $\mathbb{E}[S_t] = 0$. Pick any integer $0 \leq a < b$ and $\epsilon > 0$. Then

$$\mathbb{P}(S_N \geq \epsilon N) \leq (b-a+1)\max_{a\leq t\leq b} \mathbb{P}(S_t \geq \epsilon t) + \mathbb{P}(N \notin [a,b]), \tag{3.31}$$

$$\mathbb{P}(S_N \leq -\epsilon N) \leq (b-a+1)\max_{a\leq t\leq b} \mathbb{P}(S_t \leq -\epsilon t) + \mathbb{P}(N \notin [a,b]), \tag{3.32}$$

Proof.

$$\begin{aligned}\mathbb{P}(S_N \geq \epsilon N) &\leq \mathbb{P}(S_N \geq \epsilon N, a \leq N \leq b) + \mathbb{P}(N \notin [a,b]) \text{(According to if } N \notin [a,b], \text{ then } a \leq N \leq b) \\ &= \mathbb{P}(S_N \geq \epsilon N | a \leq N \leq b)\mathbb{P}(a \leq N \leq b) + \mathbb{P}(N \notin [a,b])\end{aligned} \tag{3.33}$$

We also have

$$\begin{aligned}\mathbb{P}(S_N \geq \epsilon N, a \leq N \leq b) &= \mathbb{E}[\mathbb{I}(S_N \geq \epsilon N)|a \leq N \leq b] \\ &\leq \mathbb{E}\Big[\sum_{i=a}^{b} \mathbb{I}(S_i \geq \epsilon i)|a \leq N \leq b\Big] \\ &= \sum_{i=a}^{b} \mathbb{P}(S_i \geq \epsilon i|a \leq N \leq b).\end{aligned} \tag{3.34}$$

The second step follows that if the integer N can make $S_N \geq \epsilon N$, then at $[a, b]$, there must be more than or equal to one integers (including N) that can meet this requirement.

Therefore, $\mathbb{P}(S_N \geq \epsilon N|a \leq N \leq b)\mathbb{P}(a \leq N \leq b)$ can be bounded by $\sum_{i=a}^{b} \mathbb{P}(S_i \geq \epsilon i)$ (Because the summation from a from b).

Bounding the term by the maxima of its terms and multiplied by the number of terms, we get the desired inequality. □

Lemma 12. (Hoeffding-Azuma inequality for Stopped Martingales) Assume that S_t is a centered matingale such that the corresponding martingale difference process is uniformly bounded by C. Then, for any fixed $\epsilon > 0$, integers $0 \leq a < b$, the following inequalities hold:

$$\mathbb{P}(S_N > \epsilon N) \leq (b - a + 1)\exp(-2a^2\epsilon^2/C^2) + \mathbb{P}(N \notin [a, b]), \tag{3.35}$$

$$\mathbb{P}(S_N < -\epsilon N) \leq (b - a + 1)\exp(-2a^2\epsilon^2/C^2) + \mathbb{P}(N \notin [a, b]), \tag{3.36}$$

Proof. The result follows Lemma 10 and 11. □

Lemma 13. Let (Z_i), $i = 1, ..., n$ be a sequence of random variables such that Z_i is conditionally independent of Z_{i+1}, ..., Z_n given Z_1,..., Z_{i-1}. Then the Doob martingale $X_i = \mathbb{E}[f(Z_1, ..., Z_n)|Z_1, ..., Z_i]$ has bounded differences, in particular

$$|X_{i+1} - X_i| \leq 2C \tag{3.37}$$

Proof. Proof omitted. □

Now let $N = \sum_{i=1}^{n} Z_i$ where Z_i are $0-1$-valued random variables. We assume that Z_i is adapted to the filtration $\{\mathcal{F}_i\}_t$ and that Z_{i+1} is conditionally independent of $Z_{i+2}, Z_{i+3}, ..., Z_n$ given $\mathcal{F}_i$.

Note: Our aim is to obtain upper and lower tail bounds for the counting process N.

Lemma 14. We have

$$\mathbb{P}(N - \mathbb{E}[N] > u) \leq \exp(-u^2/(2n)). \tag{3.38}$$

Similarly,

$$\mathbb{P}(N - \mathbb{E}[N] < -u) \leq \exp(-u^2/(2n)). \tag{3.39}$$

Proof. Note that the function $f(z_1, ..., z_n) = z_1 + ... + z_n$ is 1-Lipschitz. Hence, the Doob martingale $X_i = \mathbb{E}[N|Z_1, ..., Z_i]$ is a bounded difference martingale with bound 2 (According to Lemma 13). The H-A inequality applied to the centered martingale $X_i - \mathbb{E}[N]$ yields the desired result. □

The next result gives an upper tail bound on N when $\mathbb{E}[\sum_{i=1}^n Z_i]$ is slowly growing:

Lemma 15. Let Z_i be as in Lemma 14, $N_n = \sum_{i=1}^n Z_i$. Assume that a_n is an upper bound on $\mathbb{E}[N_n]$. Then for all $\Delta > 0$, if n is such that $a_n \leq \Delta/2$ then

$$\mathbb{P}(N_n \geq \Delta) \leq \exp(-\Delta^2/(8n)). \tag{3.40}$$

Proof. We have

$$\mathbb{P}(N_n \geq \Delta) = \mathbb{P}(N_n > \mathbb{E}[N_n] + \Delta - \mathbb{E}[N_n]) \leq \mathbb{P}(N_n > \mathbb{E}[N_n] + \Delta/2) \tag{3.41}$$

since by the assumption $\mathbb{E}[N_n] \leq a_n \leq \Delta/2$.

Use Lemma 14, we obtain the result. □

The following technical lemma is at the core of our results for propagating confidence bounds "upward in the tree":

Lemma 16. Let Z_i, $\mathcal{F}_i$, a_i be as in Lemma 13. Let $\{X_i\}$ be an i.i.d sequence with mean μ, and Y_i and $\mathcal{F}_i$-adapted process. We assume that both X_i and Y_i lie in the $[0, 1]$ interval. Consider the partial sums

$$S_n = \sum_{i=1}^n (1 - Z_i)X_i + Z_i Y_i. \tag{3.42}$$

Fix an arbitrary $\delta > 0$, let $\Delta = 9\sqrt{2n \ln(2/\delta)}$ and let

$$R_n = \mathbb{E}\Big[\sum_i X_i\Big] - \mathbb{E}[S_n] \tag{3.43}$$

Then for n such that $a_n \leq (1/9)\Delta$ and $R_n \leq (4/9)\Delta/2$,

$$\mathbb{P}(S_n \geq \mathbb{E}[S_n] + \Delta) \leq \delta \tag{3.44}$$

and

$$\mathbb{P}(S_n \leq \mathbb{E}[S_n] - \Delta) \leq \delta. \tag{3.45}$$

Proof. Let $p = \mathbb{P}(S_n \geq \mathbb{E}[S_n] + \Delta)$. We have $S_n = \sum_{i=1}^n X_i + \sum_{i=1}^n Z_i(Y_i - X_i) \leq \sum_{i=1}^n X_i + 2\sum_{i=1}^n Z_i$. Therefore,

$$p \leq \mathbb{P}\Big(\sum_{i=1}^n X_i + 2\sum_{i=1}^n Z_i \geq \mathbb{E}\Big[\sum_{i=1}^n X_i\Big] - R_n + \Delta\Big) \tag{3.46}$$

Using the elementary inequality $\mathbb{I}(A+B \geq \Delta) \leq \mathbb{I}(A \geq \alpha\Delta) + \mathbb{I}(B \geq (1-\alpha)\Delta)$ that holds for any $A, B \geq 0$, $0 \leq \alpha \leq 1$, we get

$$p \leq \mathbb{P}\Big(\sum_{i=1}^{n} X_i \geq \mathbb{E}\Big[\sum_{i=1}^{n} X_i\Big] + \Delta/9\Big) + \mathbb{P}\Big(2\sum_{i=1}^{n} Z_i \geq 8/9\Delta - R_n\Big) \tag{3.47}$$

□

Chapter 4

UCT with Laplace Bound

Problem definition:

Consider a bandit problem with K arms, defined by the sequence of random payoffs X_{it}, $i = 1, ..., K$, $t \geq 1$, where each i is the index of a gambling machine (the "arm" of a bandit). Successive plays of machine i yield the payoffs X_{i1}, $X_{i2}, \cdots$. For simplicity, we shall assume that X_{it} lies in the interval $[0, 1]$. An allocation policy is a mapping that selects the next arm to be played based on the sequence of past selections and the payoffs obtained. The expected regret of an allocation policy A after n plays is defined by Expected cumulative regret is defined as

$$\mathcal{R} = \max_i \mathbb{E}[\sum_{t=1}^{n} X_{it}] - \mathbb{E}[\sum_{i=1}^{K} \sum_{t=1}^{T_i(n)} X_{i,t}], \tag{4.1}$$

where $T_i(n) = \sum_{s=1}^{n} \mathbb{1}(\boldsymbol{I}_s = i)$ is the number of times arm i was played up to time n, $I_t \in \{1, ..., K\}$ is the index of the arm selected at time t.

The expected values of the averages

$$\overline{X}_{in} = \frac{1}{n} \sum_{t=1}^{n} X_{it} \tag{4.2}$$

converge. We let $\mu_{in} = \mathbb{E}[\overline{X}_{in}]$ and

$$\mu_i = \lim_{n \to \infty} \mu_{in}. \tag{4.3}$$

Further, we define δ_{in} by (Non-stationarity)

$$\mu_{in} = \mu_i + \delta_{in} \tag{4.4}$$

We start by analyzing UCB1 for non-stationary bandit problems. Remember that by assumption $0 \leq X_{it} \leq 1$. Quantities related to the optimal arm shall be upper indexed by a star, e.g., $\mu^\star, T^\star(t), \overline{X}_t^\star$, etc. For the sake of easy referencing, we summarize the assumptions on the rewards here:

Assumption 1. Fix $1 \leq i \leq K$. Let $\{\mathcal{F}_{it}\}_t$ be a filtration such that X_{itt} is $\{\mathcal{F}_{it}\}_t$-adapted and $X_{i,t}$ is conditionally independent of $\mathcal{F}_{i,t+1}$, $\mathcal{F}_{i,t+2}$, ... given $\mathcal{F}_{i,t-1}$. Then $0 \leq X_{it} \leq 1$ and the limit of $\mu_{in} = \mathbb{E}[\overline{X}_{in}]$ exists. Further, we assume that there exists a constant $C_p > 0$ and for all $n > 0$, for any $\delta > 0$, $\Delta_n(\delta) = 2C_p\sqrt{(n+1)\ln(K \cdot \sqrt{n+1}/\delta)}$ [1], the following bounds hold:

$$\mathbb{P}(n\overline{X}_{in} \geq n\mathbb{E}[\overline{X}_{in}] + \Delta_n(\delta)) \leq \delta \tag{4.5}$$

and

$$\mathbb{P}(n\overline{X}_{in} \leq n\mathbb{E}[\overline{X}_{in}] - \Delta_n(\delta)) \leq \delta \tag{4.6}$$

Set δ to t^{-1}, according to Laplace bound[2],

$$c_{t,s} = \Delta_s(t^{-1})/s = 2C_p\sqrt{(s+1)\ln(K \cdot \sqrt{s+1} \cdot t)}/s = 2C_p\sqrt{(1+\frac{1}{s}) \cdot \frac{\ln(K \cdot t \cdot \sqrt{s+1})}{s}} \tag{4.7}$$

We let $\Delta_i = \mu^\star - \mu_i$.

Definition 2. Convergence of δ_{it} with ϵ

Note that we have this definition because of the non-stationary assumption on UCB1 (or UCB-Laplace). So this part is consistent as original UCT.

Since δ_{it} converges by assumption to zero, for all $\epsilon > 0$ there exists an index $N_0(\epsilon)$ such that if $t \geq N_0(\epsilon)$ then $|\delta_{it}| \leq \epsilon\Delta_i/2$ and $|\delta_{j^\star,t}| \leq \epsilon\Delta_i/2$, whenever i is the index of a sub-optimal arm and $j^\star$ is the index of an optimal arm.

In particular, it follows that for any optimal arm $j^\star$, $t \geq N_0(\epsilon)$, $|\delta_{j^\star,t}| \leq \epsilon/2 \min_{\{i|\Delta_i>0\}} \Delta_i$.

Theorem 3. Consider UCB-Laplace applied to a non-stationary problem where the pay-off sequence satisfies Assumption 1 and where the bias sequence $c_{t,s}$, used by UCB-Laplace is given by Equation 4.7. Fix $\epsilon > 0$. Let $T_i(n)$ denote the number of plays of arm i. Then if i is the index of a suboptimal arm then

$$\mathbb{E}[T_i(n)] \leq \frac{32C_p^2}{(1-\epsilon)^2\Delta_i^2}\log\frac{2\sqrt{2}dC_p}{(1-\epsilon)\Delta_i\delta} + N_0(\epsilon) + 3 + \log n \tag{4.8}$$

Proof. Fix the index i of a suboptimal arm. Let

$$A_0(n,\epsilon) = \min\{s|c_{t,s} \leq (1-\epsilon)\Delta_i/2\} \tag{4.9}$$

[1]Note that here C_p is chosen to be $\frac{1}{\sqrt{2}}$ for both UCT and UCT-L.

[2]We will explain why we set δ to be t^{-1}, instead of t^{-4} in original UCT paper in the next

Substituting $c_{t,s}$ and squaring gives

$$\frac{T_i(n)^2}{T_i(n)+1} \leq \frac{4}{\Delta_i^2(1-\epsilon)^2} \cdot 4 \cdot C_p^2 \log\left(\frac{d}{\delta}(1+T_i(n))^{1/2}\right) \tag{4.10}$$

Rearranging terms

$$T_i(n)+1 < \frac{16C_p^2}{\Delta_i^2(1-\epsilon)^2}\log\frac{d}{\delta} + \frac{8C_p^2}{\Delta_i^2(1-\epsilon)^2}\log(1+T_i(n)) \tag{4.11}$$

By using Lemma 8 of [4], we get that for all $T_i(n) \geq 0$

$$T_i(n)+1 < \frac{16C_p^2}{\Delta_i^2(1-\epsilon)^2}\left(\log\frac{8C_p^2}{\Delta_i^2(1-\epsilon)^2} + 2\log\frac{d}{\delta}\right), \tag{4.12}$$

with $a = \frac{\Delta_i^2(1-\epsilon)^2}{8C_p^2}(T_i(n)+1)$ and $b = -2\log\frac{d}{\delta}$.

Rearranging terms

$$T_i(n) \leq 3 + \frac{32C_p^2}{(1-\epsilon)^2\Delta_i^2}\log\frac{2\sqrt{2}dC_p}{(1-\epsilon)\Delta_i\delta} \tag{4.13}$$

Note that $c_{t,s}$ holds with confidence $1-\delta$.

Therefore, suppose at time step t, the failure probability is $1/t$. We can bound the number of plays of arm i for n rounds. With failure probability δ,

$$T_i(n) \leq \log n \tag{4.14}$$

By the definition of $c_{t,s}$, $A_0(n,\epsilon) = 3 + \frac{32C_p^2}{(1-\epsilon)^2\Delta_i^2}\log\frac{2\sqrt{2}dC_p}{(1-\epsilon)\Delta_i\delta}$. We let

$$A(n,\epsilon) = \max\{A_0(n,\epsilon), N_0(\epsilon)\} \tag{4.15}$$

(There is no N_p. The reason why see Assumption 1.)

Therefore we can derive our conclusion

$$\mathbb{E}[T_i(n)] \leq \frac{32C_p^2}{(1-\epsilon)^2\Delta_i^2}\log\frac{2\sqrt{2}dC_p}{(1-\epsilon)\Delta_i\delta} + N_0(\epsilon) + 3 + \log n \tag{4.16}$$

□

Theorem 4. Let

$$\overline{X}_n = \sum_{i=1}^{K}\frac{T_i(n)}{n}\overline{X}_{i,T_i(n)} \tag{4.17}$$

Under the assumption of Theorem 3,

$$|\mathbb{E}[\overline{X}_n] - \mu^\star| \leq |\delta_n^\star| + \mathcal{O}\left(\frac{K(C_p^2\ln(C_pn) + N_0)}{n}\right), \tag{4.18}$$

where $N_0 = N_0(1/2)$.

Proof. W.l.o.g, we assume there is a unique optimal arm, denoted by arm $i^\star$.
By triangle inequality, $|\mu^\star - \mathbb{E}[\overline{X}_n]| \leq |\overline{\mu}^\star - \overline{\mu}_n^\star| + |\overline{\mu}_n^\star - \mathbb{E}[\overline{X}_n]|$.

$$
\begin{aligned}
n|\overline{\mu}_n^\star - \mathbb{E}[\overline{X}_n| &= \Big|\sum_{i=1}^{n}\mathbb{E}[X_t^\star] - \mathbb{E}\Big[\sum_{i=1}^{K}\overline{X}_{i,T_i(n)}\Big]\Big| \\
&\qquad \text{Separate optimal arm and sub-optimal arms} \qquad (4.19)\\
&= \Big|\sum_{i=1}^{n}\mathbb{E}[X_t^\star] - \mathbb{E}\Big[T*(n)\overline{X}^\star_{T^\star(n)}\Big]\Big| + \mathbb{E}\Big[\sum_{i=1, i\neq i^\star}^{K}\overline{X}_{i,T_i(n)}\Big]
\end{aligned}
$$

- According to Theorem 3, and by the assumption that $0 \leq \overline{X}_{i,T_i(n)} \leq 1$, the second term is bounded by $\mathcal{O}\Big(K(C_p^2\ln(C_pn) + N_0)\Big)$.

  Note: According to Theorem 3, the parameters we can have control are ϵ, C_p, and n.

  In order to bound the first term in Equation 4.19, let us note that $T^\star(n)\overline{X}^\star_{T^\star(n)} = \sum_{t=1}^{T^\star(n)} X_t^\star$ and we have

$$
\begin{aligned}
D_n &\overset{\text{def}}{=} \sum_{t=1}^{n}\mathbb{E}[X_t^\star] - \mathbb{E}\left[\sum_{t=1}^{T^\star(n)} X_t^\star\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{n} X_t^\star - \sum_{t=1}^{T^\star(n)} X_t^\star\right] \qquad (4.20)\\
&= \mathbb{E}\left[\sum_{t=T^\star(n)+1}^{n} X_t^\star\right]
\end{aligned}
$$

Since $0 \leq \overline{X}_{i,T_i(n)} \leq 1$, we can bound the term from above by $\mathbb{E}[n - T^\star(n)]$, which is just $\sum_{i\neq i^\star}\mathbb{E}[T_i(n)]$ and hence by Theorem 3, $D_n = \mathcal{O}\Big(K(C_p^2\ln(C_pn) + N_0)\Big)$. □

Theorem 5. (Lower Bound) Under the assumptions of Theorem 3, there exists some positive constant ρ such that for all arms i and n, $T_i(n) \geq \lceil \rho\log(n)\rceil$.

Proof. Proof is obvious. □

Theorem 6. For an arbitraty $\delta > 0$, and let $\Delta_n = \sqrt{2t(1+\frac{1}{t})\ln(\sqrt{t+1}/\delta)}$. Let n_0 be such that

$$
\sqrt{n_0} \geq \mathcal{O}\Big(K\big(C_p^2\ln(C_pn) + N_0(1/2)\big)\Big). \qquad (4.21)
$$

Then for any $n \geq n_0$, under the assumptions of Theorem 3 the following bounds hold true:

$$
\mathbb{P}(n\overline{X}_n \geq n\mathbb{E}[\overline{X}_n] + \Delta_n) \leq \delta \qquad (4.22)
$$

and

$$
\mathbb{P}(n\overline{X}_n \leq n\mathbb{E}[\overline{X}_n] - \Delta_n) \leq \delta \qquad (4.23)
$$

Proof. In original UCT paper, this theorem is proved using lemmas of Hoeffding-Azuma inequality for Stopped Martingales. Let Z_t be the indicator of the event that a suboptimal arm is chosen at time step t. Then by Theorem 3, $\mathbb{E}[\sum_{t=1}^{n} Z_t] \leq \mathcal{O}(K \ln(n))$. Hence, a_t can be chosen to be $\mathcal{O}\Big(K\big(C_p^2 \ln(C_p t) + N_0(1/2)\big)\Big)$. Further, we denote X_t as the payoff sequence of the best arm. We let Y_t be the payoff received at time step t. By assumption, X_t, Y_t lie in the $[0,1]$ interval and $n\overline{X}_n = \sum_{t=1}^{n}(1-Z_t)X_t + Z_t Y_t$. $R_n = \mathcal{O}\Big(K\big(C_p^2 \ln(C_p t) + N_0(1/2)\big)\Big)$. Let n_0 be an index such that if $n \geq n_0$ and X_t and Y_t are 1-sub-Gaussian. Such an index exists since $\Delta_n = \mathcal{O}(\sqrt{n})$ and $R_n = \mathcal{O}(\ln n)$Hence, for $n \geq n_0$, the conditions of Lemma 2.7 in [3] are satisfied and the desired tail-inequalities hold for $\overline{X}_n$. Since for $\delta < 1$, $\Delta_n = \sqrt{2t(1+\frac{1}{t})\ln(\sqrt{t+1}/\delta)} > \sqrt{2t(1+\frac{1}{t})\ln(\sqrt{t+1})}$, it follows that n_0 can be selected independently of δ. In fact, for a suitable constant of c, n_0 is the first integer such that $\sqrt{n_0} \geq c\Big(K\big(C_p^2 \ln(C_p n) + N_0(1/2)\big)\Big)$. This finishes the proof of the theorem. □

Theorem 7. (Convergence of Failure Probability) Under the assumptions of Theorem 3 it holds that

$$\lim_{t\to\infty} \mathbb{P}(I_t \neq i^\star) = 0 \tag{4.24}$$

The reason why we need this theorem is that we want the probability of suboptimal choices would converge to zero.

Proof. Unlike in [5], using Laplace bound, we can have control of failure probability, therefore, by selecting δ to be $1/t$ at every time step t, we can decrease $\mathbb{P}(I_t \neq i^\star)$ to 0 with $t \to \infty$. This is also one of the advantages of Laplace bound. □

Theorem 8. Consider algorithm UCT running on a game tree of depth D, branching factor K with stochastic payoffs at the leaves. Assume that the payoffs lie in the interval $[0,1]$. Then the bias of the estimated expected payoff, $\overline{X}_n$, is $\mathcal{O}((KD\log(n) + K^D)/n)$. Further, the failure probability at the root converges to zero as the number of samples grows to infinity.

Proof. The proof is done by induction on D. Consider first the case $D = 1$ (in this case, actually, UCT just corresponds to UCB1). Our assumptions on the payoffs hold, thanks to Laplace bound. Now the result on the bias follows directly from Theorem 4 and consistency follows from Theorem 7.

Now, assume that the result holds for all trees of up to depth $D-1$ and consider a tree of depth D. Let us only concentrate on the root node. We claim that from the point of the root node, running UCT is equivalent to running UCB1 with non-stationary, correlated payoffs for the various moves (arms). Fix a move i. In fact, the payoff for move i of the root at time t will depend on all previous "entries" into the subtree originating at the successor node of move i. For simplicity we shall denote this node by i, as well. We claim that the payoff process experienced at node i will satisfy the conditions required by Theorems 3-7. First, the payoffs lie in the interval $[0,1]$. Now, since the tree starting at node i has depth $D-1$, by the induction hypothesis we may apply Theorem 4 to show that the expected average payoff converges. That the conditions on the

exponential concentration of the payoffs are satisfied follows from Theorem 6. Since this holds for any i, it follows by Theorem 4 that the bias at the root converges at the rate of

$$|\delta_n^\star| + \mathcal{O}(K(\ln n + N_0)/n), \tag{4.25}$$

where $\delta_n^\star$ is the rate of convergence of the bias for the best move and

$$N_0 = \min\left\{n \middle| |\delta_{in}| \leq \frac{1}{2}\Delta_i, i \neq i^\star\right\}. \tag{4.26}$$

Now, by the induction hypothesis,

$$|\delta_{in}| = \mathcal{O}((K(D-1)\ln n + K^{D-1})/n),\ , i = 1, ..., K \tag{4.27}$$

Hence, $N_0 = \mathcal{O}(K^{D-1})$, yielding the desired result for the bias at the root. The proof is finished by noting that the failure probability converges to zero thanks to Theorem 7.

Note that it follows from the proof that when the payoffs are deterministic then the bias terms prescribe too much exploration at the nodes immediately preceding the leaves. Here, no exploration would be needed at all. This can be achieved gradually by making the bias more uniform. From this, one might conjecture that more uniform bias terms are desirable in the vicinity of the leafs. Indeed, it is reasonable to use stronger exploration bonus close to the root: at the beginning of runs the large unexplored parts of the tree can be expected to behave "randomly".

In fact, for deterministic problems, convergence can be shown for a larger class of bias terms: the role of the bias term can be viewed as taking care of the shifts in the payoff in the subtrees as time goes by. However, we do not pursue this direction further in this paper. □

Lemma 9. (Lemma 2.6 in [3] Time-uniform concentration inequalities) Let $X_{i1}, ..., X_{in}$ be a sequence of n i.i.d. real valued random variables bounded in $[0, 1]$, and the limit $\mu_{in} = \mathbb{E}[\overline{X}_{in}]$ exists. Then, for all $\delta \in (0, 1)$, it holds

$$\mathbb{P}\left(\exists n \in \mathbb{N}, n\overline{X}_{in} - n\mathbb{E}[\overline{X}_{in}] \geq 2C_p\sqrt{(n+1)\ln(K\sqrt{n+1}/\delta)}\right) \leq \delta \tag{4.28}$$

Bibliography

[1] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

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

[3] Odalric-Ambrym Maillard. *Mathematics of Statistical Sequential Decision Making*. PhD thesis, 2019.

[4] András Antos, Varun Grover, and Csaba Szepesvári. Active learning in heteroscedastic noise. *Theoretical Computer Science*, 411(29-30):2712–2728, 2010.

[5] Levente Kocsis, Csaba Szepesvári, and Jan Willemson. Improved monte-carlo search. *Univ. Tartu, Estonia, Tech. Rep*, 1, 2006.