Supplementary Materials For "A Practical Method for Solving Contextual Bandit Problems Using Decision Trees"

1 Proof of Theorem 1

For the standard multi-armed bandit problem, let n_k be the number of total observed responses for action k and p_k be the observed success rate for action k (proportion of success responses out of total observed responses). For simplicity, we assume that we have observed at least one success and one failure for each action, i.e. $n_k \geq 2$ and $p_k \in (0, 1)$ for all k. Then, the following theorem holds:

Theorem 1. *Let a_t^{TS} be the action chosen by the Thompson sampling algorithm, and let a_t^B be the action chosen by the bootstrapping algorithm given data (n_k, p_k) for each action $k \in \{1, 2, \ldots, K\}$. Then,*

$$|P(a_t^{TS} = k) - P(a_t^B = k)| \leq C_k(p_1, ..., p_K) \sum_{j=1}^{K} \frac{1}{\sqrt{n_j}}$$

holds for every $k \in \{1, 2, \ldots, K\}$, for some function $C_k(p_1, ..., p_K)$ of $p_1, ..., p_K$.

To prove the theorem, we first provide a few lemmas. The proofs for all lemmas are given in Section 2 of this supplementary materials document. For notational convenience, we define $\alpha_k = n_k p_k$ and $\beta_k = n_k(1 - p_k)$, which indicate the number of success and failure responses observed so far for action k. For each action $k \in \{1, 2, \ldots, K\}$, the Thompson sampling algorithm first draws a random sample of the true (unknown) success probability according to a beta distribution with parameters α_k and β_k, and it then chooses the action with the highest success probability. The bootstrapping algorithm samples n_k observations with replacement from action k's observed rewards, and the generated success probability is then the proportion of successes observed in the bootstrapped dataset. This procedure is equivalent to generating a binomial random variable with n_k trials and p_k success rate, divided by n_k. The following lemma shows how close the distributions of these two random probability estimates are in terms of the number of available data points for each action.

Lemma 1. *Let X be a beta random variable with integer parameters $\alpha > 0$ and $\beta > 0$, and let Y be a binomial random variable with n trials and success rate p, where $\alpha + \beta = n$ and $p = \frac{\alpha}{n}$. Then,*

$$\max_{z \in [0,1]} \left| P(X \leq z) - P\left(\frac{Y}{n} \leq z\right) \right| \leq \frac{c(p)}{\sqrt{n}},$$

for some function $c(p)$ that is independent of n.

We now provide a second lemma which will prove useful in deriving Theorem 1. Recall that for each $k \in \{1, 2, \ldots, K\}$, n_k is the total number of observations and p_k is the observed success rate. Let p_k^{TS}

be the randomly drawn success probability of action k by the Thompson sampling algorithm and let p_k^B be the randomly drawn success probability of action k by the bootstrapping algorithm. Under each algorithm $h \in \{TS, B\}$, an action is chosen arbitrarily from the set $M^h := \{k : p_k^h \geq p_j^h \text{ for every } j \neq k\}$. In the Thompson sampling algorithm, p_k^{TS} is sampled from a continuous (beta) distribution, and thus M^{TS} will have cardinality one almost surely. Hence,

$$P(a_t^{TS} = k) = P(p_k^{TS} \geq p_j^{TS} \text{ for every } j \neq k)\,. \tag{1}$$

However, these events are not necessarily equivalent with respect to the bootstrapping algorithm. Since p_k^B is sampled using a discrete (binomial) distribution, it is possible that two actions will have the same sampled probabilities. Thus, it is possible for an action $k \in M^B$ to not be chosen if there exists another action $l \in M^B$. We provide a lemma which examines the difference in these two events:

Lemma 2. *For some function $b(p_k)$ that is independent of n_k,*

$$|P(a_t^B = k) - P(p_k^B \geq p_j^B \text{ for every } j \neq k)| \leq \frac{b(p_k)}{\sqrt{n_k}}(1 + O(1/n_k))\,.$$

We can now proceed with the derivation of the theorem. We define $err_k(z|X) := P(p_k^{TS} \leq z|X) - P(p_k^B \leq z|X)$ with respect to some random variable X. Assuming X is independent of p_k^{TS} and p_k^B (but not necessarily of z), then it follows from Lemma 1 that $|err_k(z|X)| \leq \frac{c(p_k)}{\sqrt{n_k}}$ for some function $c(p_k)$ that is independent of n_k. Note that this result also holds for the function $\widetilde{err}_k(z|X) := P(p_k^B \leq z|X) - P(p_k^{TS} \leq z|X)$.

Note that the events $\{p_k^h \geq p_j^h\}$ for $j \neq k$ are independent conditioned on p_k^h. Hence, using (1) we have

$$\begin{aligned}
P(a_t^{TS} = k) &= E_{p_k^{TS}}\left[P(p_k^{TS} \geq p_j^{TS} \text{ for every } j \neq k | p_k^{TS})\right] \\
&= E_{p_k^{TS}}\left[\prod_{j \neq k} P(p_k^{TS} \geq p_j^{TS} | p_k^{TS})\right] \\
&= E_{p_k^{TS}}\left[\prod_{j \neq k} \left(P(p_k^{TS} \geq p_j^{B} | p_k^{TS}) + err_j(p_k^{TS} | p_k^{TS})\right)\right]\,.
\end{aligned} \tag{2}$$

In the expansion of the product (2), only one term does not include $err_j(p_k^{TS}|p_k^{TS})$. This term is given as

$$\begin{aligned}
& E_{p_k^{TS}}\left[\prod_{j\neq k} P(p_k^{TS} \geq p_j^B|p_k^{TS})\right] \\
= \;& E_{p_k^{TS}}\left[P(p_k^{TS} \geq p_j^B \text{ for every } j\neq k|p_k^{TS})\right] \\
= \;& E_{p_j^B, j\neq k}\left[P(p_k^{TS} \geq \max_{j\neq k} p_j^B|p_j^B, j\neq k)\right] \\
= \;& E_{p_j^B, j\neq k}\left[P(p_k^B \geq \max_{j\neq k} p_j^B|p_j^B, j\neq k) + \widetilde{err}_k(\max_{j\neq k} p_j^B|p_j^B, j\neq k)\right] \\
= \;& P(p_k^B \geq \max_{j\neq k} p_j^B) + E_{p_j^B, j\neq k}[\widetilde{err}_k(\max_{j\neq k} p_j^B|p_j^B, j\neq k)] \\
= \;& P(a_t^B = k) + (P(p_k^B \geq \max_{j\neq k} p_j^B) - P(a_t^B = k)) + E_{p_j^B, j\neq k}[\widetilde{err}_k(\max_{j\neq k} p_j^B|p_j^B, j\neq k)]\,. \qquad (3)
\end{aligned}$$

The rest of (2) is the sum of multiplications of $K-1$ terms of $P(p_k^{TS} \geq p_j^B|p_k^{TS})$ and $err_j(p_k^{TS}|p_k^{TS})$. Because $P(p_k^{TS} \geq p_j^B|p_k^{TS}) \leq 1$ and $|err_j| \leq \frac{c(p_j)}{\sqrt{n_j}}$ from Lemma 1, the rest can be bounded by $\prod_{j\neq k}\left(1+\frac{c(p_j)}{\sqrt{n_j}}\right) - 1$. Hence, applying Lemma 1 and Lemma 2 to (3), we have

$$\begin{aligned}
& |P(a_t^{TS} = k) - P(a_t^B = k)| \\
\leq \;& |P(p_k^B \geq \max_{j\neq k} p_j^B) - P(a_t^B = k)| + |E_{p_j^B, j\neq k}[\widetilde{err}_k(\max_{j\neq k} p_j^B|p_j^B, j\neq k)]| + \prod_{j\neq k}\left(1+\frac{c(p_j)}{\sqrt{n_j}}\right) - 1 \\
\leq \;& \frac{b(p_k)}{\sqrt{n_k}}\left(1+O(1/n_k)\right) + \frac{c(p_k)}{\sqrt{n_k}} + \prod_{j\neq k}\left(1+\frac{c(p_j)}{\sqrt{n_j}}\right) - 1\,. \qquad (4)
\end{aligned}$$

Note that the error term from Sterling's approximation, $O(1/n_k)$, can be bounded from above by a constant. Furthermore, for any set S of actions,

$$\prod_{s\in S}\frac{c(p_s)}{\sqrt{n_s}} \leq \frac{1}{\sqrt{n_t}}\prod_{s\in S} c(p_s) \qquad \forall t \in S\,.$$

One can then use these facts to manipulate (4) to prove the desired result:

$$|P(a_t^{TS} = k) - P(a_t^B = k)| \leq C_k(p_1, ..., p_K)\sum_{j=1}^{K}\frac{1}{\sqrt{n_j}}\,,$$

where $C_k(p_1, ..., p_K)$ is a function of $p_1, ..., p_K$. □

2 Proof of Auxiliary Lemmas

Lemma 1. *Let X be a beta random variable with integer parameters $\alpha > 0$ and $\beta > 0$, and let Y be a binomial random variable with n trials and success rate p, where $\alpha + \beta = n$ and $p = \frac{\alpha}{n}$. Then,*

$$\max_{z\in[0,1]} \left| P(X \leq z) - P\left(\frac{Y}{n} \leq z\right) \right| \leq \frac{c(p)}{\sqrt{n}},$$

for some function $c(p)$ that is independent of n.

Proof. For notational convenience, we define $q = \frac{\beta}{n} = 1 - p$, and we denote the p.d.f., and the c.d.f. of a standard normal random variable by $\phi(\cdot)$ and $\Phi(\cdot)$, respectively. We will first show that, for each $z \in [0, 1]$, $P(X \leq z)$ and $P\left(\frac{Y}{n} \leq z\right)$ can be approximated using normal c.d.f.'s $\Phi(\frac{\sqrt{n}(z-p)}{\sqrt{pq+(z-p)^2}})$ and $\Phi(\frac{\sqrt{n}(z-p)}{\sqrt{pq}})$, respectively. Finally, we will bound the difference in these approximations and apply the triangle inequality:

$$\begin{aligned}
\left| P(X \leq z) - P\left(\frac{Y}{n} \leq z\right) \right| &\leq \left| P(X \leq z) - \Phi(\frac{\sqrt{n}(z-p)}{\sqrt{pq+(z-p)^2}}) \right| \\
&+ \left| \Phi(\frac{\sqrt{n}(z-p)}{\sqrt{pq}}) - \Phi(\frac{\sqrt{n}(z-p)}{\sqrt{pq+(z-p)^2}}) \right| \\
&+ \left| P(\frac{Y}{n} \leq z) - \Phi(\frac{\sqrt{n}(z-p)}{\sqrt{pq}}) \right| \qquad (5)
\end{aligned}$$

We first bound the normal approximation for $P\left(\frac{Y}{n} \leq z\right)$. From the Berry - Esseen theorem, we have

$$|P(\sqrt{n}\frac{(Y/n - p)}{\sqrt{pq}} \leq x) - \Phi(x)| \leq \frac{C(p^2 + q^2)}{\sqrt{npq}}$$

for every x, which implies that

$$|P(\frac{Y}{n} \leq z) - \Phi(\frac{\sqrt{n}(z-p)}{\sqrt{pq}})| \leq \frac{C(p^2 + q^2)}{\sqrt{npq}} \qquad (6)$$

holds for every $z \in [0, 1]$.

Next, we will show that $P(X \leq z)$ can be approximated by a similar (but not exactly the same) function, $\Phi(\frac{\sqrt{n}(z-p)}{\sqrt{pq+(z-p)^2}})$. Note that the beta random variable X has the same distribution as $\frac{A}{A+B}$, where A and B are independent Gamma random variables with shape parameters α and β, respectively. We first derive an approximation for the c.d.f. of Gamma random variables. Suppose that Γ is Gamma distributed with an integer shape parameter $m > 0$. Because Γ has the same distribution as the sum of m independent exponential random variables with parameter 1, from the Berry - Esseen theorem we have

$$|P(\sqrt{m}(\Gamma/m - 1) \leq x) - \Phi(x)| \leq \frac{C\rho}{\sqrt{m}}, \qquad (7)$$

where ρ is the third-order absolute moment of the unit exponential distribution.

Let N_1 and N_2 be independent standard normal random variables, and define

$$g(z) := P((\sqrt{\alpha}N_1 + \alpha)(1-z) \leq (\sqrt{\beta}N_2 + \beta)z)\,.$$

Then, from the triangle inequality we have that for all $z \in (0,1)$:

$$\begin{aligned} & |P(X \leq z) - g(z)| = |P(A(1-z) \leq Bz) - g(z)| \\ \leq\ & |P(A(1-z) \leq Bz) - P((\sqrt{\alpha}N_1 + \alpha)(1-z) \leq Bz)| + |P((\sqrt{\alpha}N_1 + \alpha)(1-z) \leq Bz) - g(z)|\,. \end{aligned}$$

From (7), we have

$$\begin{aligned} & |P(A(1-z) \leq Bz) - P((\sqrt{\alpha}N_1 + \alpha)(1-z) \leq Bz)| \\ =\ & \left|E_B\left[P(A(1-z) \leq Bz|B) - P((\sqrt{\alpha}N_1 + \alpha)(1-z) \leq Bz|B)\right]\right| \\ \leq\ & E_B\left[\left|P(A(1-z) \leq Bz|B) - P((\sqrt{\alpha}N_1 + \alpha)(1-z) \leq Bz|B)\right|\right] \\ \leq\ & \frac{C\rho}{\sqrt{\alpha}}\,. \end{aligned}$$

Similarly, again from (7), we have

$$\begin{aligned} & |P((\sqrt{\alpha}N_1 + \alpha)(1-z) \leq Bz) - g(z)| \\ =\ & \left|E_{N_1}\left[P((\sqrt{\alpha}N_1 + \alpha)(1-z) \leq Bz|N_1) - P(\sqrt{\alpha}N_1(1-z) - \sqrt{\beta}N_2 z \leq nz - \alpha|N_1)\right]\right| \\ \leq\ & E_{N_1}\left[\left|P((\sqrt{\alpha}N_1 + \alpha)(1-z) \leq Bz|N_1) - P(\sqrt{\alpha}N_1(1-z) - \sqrt{\beta}N_2 z \leq nz - \alpha|N_1)\right|\right] \\ \leq\ & \frac{C\rho}{\sqrt{\beta}}\,. \end{aligned}$$

Finally, because $-N_2$ is a standard normal random variable and the sum of two independent normal random variables is a standard normal random variable, we have

$$\begin{aligned} & g(z) = P(\sqrt{\alpha}N_1(1-z) - \sqrt{\beta}N_2 z \leq nz - \alpha) \\ =\ & \Phi\left(\frac{nz-\alpha}{\sqrt{\alpha(1-z)^2 + \beta z^2}}\right) = \Phi\left(\frac{\sqrt{n}(z-p)}{\sqrt{pq + (z-p)^2}}\right), \end{aligned}$$

which concludes that

$$|P(X \leq z) - \Phi\left(\frac{\sqrt{n}(z-p)}{\sqrt{pq + (z-p)^2}}\right)| \leq \frac{C\rho}{\sqrt{np}} + \frac{C\rho}{\sqrt{nq}} \tag{8}$$

holds for every $z \in [0,1]$.

Finally, we provide a bound for $|\Phi(\frac{\sqrt{n}(z-p)}{\sqrt{pq}}) - \Phi(\frac{\sqrt{n}(z-p)}{\sqrt{pq+(z-p)^2}})|$. For notational convenience, we define $d := \frac{z-p}{\sqrt{pq}}$. Then, the difference is given as $|\Phi(\sqrt{n}d) - \Phi(\frac{\sqrt{n}d}{\sqrt{1+d^2}})|$. Because $\Phi(y)$ is concave in y for $y \geq 0$,

for every $d \geq 0$ we have

$$\begin{aligned}
|\Phi(\sqrt{n}d) - \Phi(\frac{\sqrt{n}d}{\sqrt{1+d^2}})| &= \Phi(\sqrt{n}d) - \Phi(\frac{\sqrt{n}d}{\sqrt{1+d^2}}) \\
&\leq \left(\sqrt{n}d - \frac{\sqrt{n}d}{\sqrt{1+d^2}}\right)\phi(\frac{\sqrt{n}d}{\sqrt{1+d^2}}) \\
&= (\sqrt{1+d^2}-1)\frac{\sqrt{n}d}{\sqrt{1+d^2}}\phi(\frac{\sqrt{n}d}{\sqrt{1+d^2}}) \\
&\leq (\sqrt{1+d^2}-1)\phi(1) \\
&= \left(\frac{d^2}{\sqrt{1+d^2}+1}\right)\phi(1) \\
&\leq \frac{d^2}{2}\phi(1)\,,
\end{aligned}$$

where the first inequality is from the concavity of $\Phi(\cdot)$ and the second inequality is from the fact that $x\phi(x)$ is maximized at $x = 1$ for $x \geq 0$. We define $f(d) := \Phi(\sqrt{n}d) - \Phi(\frac{\sqrt{n}d}{\sqrt{1+d^2}})$ and $d^*(n) := \arg\max_{d\geq 0} f(d)$. Then, for every $d \geq 0$, $\Phi(\sqrt{n}d) - \Phi(\frac{\sqrt{n}d}{\sqrt{1+d^2}}) \leq \Phi(\sqrt{n}d^*(n)) - \Phi(\frac{\sqrt{n}d^*(n)}{\sqrt{1+d^*(n)^2}}) \leq \frac{d^*(n)^2}{2}\phi(1)$.

We will later show that $d^*(n)^4 < \frac{30\ln(10)}{n}$. Hence,

$$|\Phi(\frac{\sqrt{n}(z-p)}{\sqrt{pq}}) - \Phi(\frac{\sqrt{n}(z-p)}{\sqrt{pq+(z-p)^2}})| \leq \frac{\phi(1)\sqrt{30\ln(10)}}{2\sqrt{n}} \tag{9}$$

holds for every $z \geq p$. The case of $z < p$ can be shown via symmetry.

Finally, from applying the bounds (6), (8), and (9) to equation (5), we have that

$$\left|P(X \leq z) - P\left(\frac{Y}{n} \leq z\right)\right| \leq \frac{C(p^2+q^2)}{\sqrt{npq}} + \frac{C\rho}{\sqrt{np}} + \frac{C\rho}{\sqrt{nq}} + \frac{\phi(1)\sqrt{30\ln(10)}}{2\sqrt{n}}\,,$$

which proves the lemma.

It remains to show that $d^*(n)^4 < \frac{30\ln(10)}{n}$. The first-order condition for $d^*(n)$ is given as

$$f'(d) = \sqrt{n}\phi(\sqrt{n}d) - \sqrt{n}(1+d^2)^{-\frac{3}{2}}\phi(\frac{\sqrt{n}d}{\sqrt{1+d^2}}) = 0\,.$$

Note that $f(d)$ is nonnegative and differentiable on $d \geq 0$. Since $f(0) = 0$, then if $\lim_{d\to\infty} f'(d) \leq 0$ we can conclude that there exists a global maximizer of $f(d)$ over $d \geq 0$ which satisfies the first-order condition. $f'(d) \leq 0$ can be simplified as

$$\begin{aligned}
\exp\left(-\frac{nd^2}{2} + \frac{nd^2}{2(1+d^2)}\right) &\leq (1+d^2)^{-\frac{3}{2}} \\
-\frac{nd^4}{2(1+d^2)} &\leq -\frac{3}{2}\ln(1+d^2) \\
n &\geq \frac{3(1+d^2)\ln(1+d^2)}{d^4}\,.
\end{aligned}$$

Note that the right hand side approaches zero as $d \to \infty$, proving that $\lim_{d\to\infty} f'(d) \leq 0$. Thus, $d^*(n)$ satisfies the first-order condition, given in a simplified form below:

$$n = \frac{3(1+d^2)\ln(1+d^2)}{d^4}\,.$$

Note that

$$\frac{d}{dx}\left(\frac{(1+x)\ln(1+x)}{x^2}\right) = \frac{\ln(1+x)}{x^2} + \frac{1}{x^2} - 2\frac{(1+x)\ln(1+x)}{x^3}$$
$$= \frac{1}{x^3}\left(x - (x+2)\ln(x+1)\right) \leq 0\,,$$

because $(x-(x+2)\ln(x+1))$ is decreasing in x and is zero when $x=0$. Hence, $d^*(n)$ is decreasing in n.

From the first-order condition and the decreasing property of $d^*(n)$, we have

$$d^*(n)^4 = \frac{3(1+d^*(n)^2)\ln(1+d^*(n)^2)}{n} \leq \frac{3(1+d^*(1)^2)\ln(1+d^*(1)^2)}{n}\,.$$

From the decreasing property and the first-order condition, we can show that $d^*(1) < 3$, which implies that $\frac{3(1+d^*(1)^2)\ln(1+d^*(1)^2)}{n} < \frac{30\ln(10)}{n}$. □

Lemma 2. *For some function $b(p_k)$ that is independent of n_k,*

$$|P(a_t^B = k) - P(p_k^B \geq p_j^B \text{ for every } j \neq k)| \leq \frac{b(p_k)}{\sqrt{n_k}}(1 + O(1/n_k))\,.$$

Proof. Note that

$$\begin{aligned} P(p_k^B \geq p_j^B \text{ for every } j \neq k) &= P(p_k^B \geq \max_{j\neq k} p_j^B) \\ &= P(p_k^B \geq \max_{j\neq k} p_j^B, a_t^B = k) + P(p_k^B \geq \max_{j\neq k} p_j^B, a_t^B \neq k) \\ &= P(a_t^B = k) + P(p_k^B \geq \max_{j\neq k} p_j^B, a_t^B \neq k)\,. \end{aligned}$$

Let Y_k denote the binomial variable associated with p_k^B for each action k, i.e. $Y_k := n_k p_k^B$. Then,

$$\begin{aligned} |P(a_t^B = k) - P(p_k^B \geq \max_{j\neq k} p_j^B)| &= P(p_k^B \geq \max_{j\neq k} p_j^B, a_t^B \neq k) \\ &\leq P(p_k^B = \max_{j\neq k} p_j^B) \\ &= P(Y_k = n_k \max_{j\neq k} p_j^B) \\ &\leq \max_i P(Y_k = i) \\ &= \max_i \binom{n_k}{i} p_k^i (1-p_k)^{n_k - i}\,. \end{aligned}$$

One can show that the binomial p.d.f. with parameters (n, p) is maximized when $i = \lfloor (n+1)p \rfloor$. Using the fact that $n_k p_k$ is an integer and that $p_k \in (0,1)$, $\lfloor (n_k+1)p_k \rfloor = \lfloor n_k p_k + p_k \rfloor = n_k p_k$. Further, through applying Stirling's formula, $n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n (1 + O(1/n))$, one obtains

$$\begin{aligned}
\binom{n_k}{n_k p_k} &= \frac{n_k!}{(n_k p_k)!(n_k - n_k p_k)!} \\
&= \sqrt{\frac{n_k}{2\pi n_k p_k (n_k - n_k p_k)}} \left(\frac{n_k}{n_k p_k}\right)^{n_k p_k} \left(\frac{n_k}{n_k - n_k p_k}\right)^{n_k - n_k p_k} (1 + O(1/n_k)) \\
&= \frac{1}{\sqrt{2\pi n_k p_k (1-p_k)}} \left(\frac{1}{p_k}\right)^{n_k p_k} \left(\frac{1}{1-p_k}\right)^{n_k - n_k p_k} (1 + O(1/n_k)) \,.
\end{aligned}$$

Thus,

$$\begin{aligned}
& \max_i \binom{n_k}{i} p_k^i (1-p_k)^{n_k - i} \\
= \; & \binom{n_k}{n_k p_k} p_k^{n_k p_k} (1-p_k)^{n_k - n_k p_k} \\
= \; & \frac{1}{\sqrt{2\pi n_k p_k (1-p_k)}} \left(\frac{1}{p_k}\right)^{n_k p_k} \left(\frac{1}{1-p_k}\right)^{n_k - n_k p_k} p_k^{n_k p_k} (1-p_k)^{n_k - n_k p_k} (1 + O(1/n_k)) \\
= \; & \frac{1}{\sqrt{2\pi n_k p_k (1-p_k)}} (1 + O(1/n_k)) \,,
\end{aligned}$$

which proves the lemma. □

3 Pseudocode for Logistic UCB

Algorithm 0: LogisticUCB()

Define $\mathcal{L}(y) = \frac{e^y}{1+e^y}$

Define $\rho(s) = \sqrt{M \log s \log(sT/\delta)}$

Initialize $t_a = 0, X_{t,a} = I_M \; \forall a = 1, ..., K$

for $t = 1, ..., T$ **do**

  Observe context vector x_t

  for $a = 1, ..., K$ **do**

    $UCB_{t,a} = \mathcal{L}(x_t^T \hat{\theta}_{t,a_t}) + \rho(t_a)||x_t||_{X_{t,a}^{-1}}$

  end

  Choose action $a_t = \arg\max_a \; UCB_{t,a}$

  Update $t_{a_t} = t_{a_t} + 1, X_{t+1,a_t} = X_{t,a_t} + x_t x_t^T$

  Update $\hat{\theta}_{t,a_t}$ with (x_t, r_{t,a_t})

end