UCBC: Variance-based Upper Confidence Bound Bandit Algorithm

Xiao Zhou

March 2021

1 Introduction

UCB1 is a famous multi-armed bandit algorithm that has a logarithmic total regret bound[2]. Employing the "Optimistic Under Uncertainty" technique, it plays the arm having the highest upper confidence bound at each time step. In this paper, a alternative UCB algorithm, called UCBC, is developed. UCBC uses an upper confidence bound that involves sample variances. This upper confidence bound is derived from Chebyshev's inequality, Central Limit Theorem, and Chi-square distribution (where the "C" is from). Compared with UCB1, UCBC has a better performance in all the tests deployed, especially when the rewards are beyond $[0, 1]$.

2 Problem Definition

The UCBC algorithm is design for the of the stochastic independent and identically distributed multi-armed bandit problem with the following parameters and annotations:

- N, the number of arms ($N \in \mathbb{N}$ and $N \geq 2$)
- T, total number of times played ($T \in \mathbb{N}$ and $T \geq 1$)
- i, the index of the arms ($i \in \mathbb{N}$ and $i \in [1, N]$)
- $*$, the index of the optimal arm
- μ_i the expected reward of arm i
- $\hat{\mu}_{i,t}$ the estimated reward of arm i from samples at time step t before t has been played
- σ_i^2 the variance of the reward of arm i
- Δ_i, the regret between the sub-optimal arm i and the optimal arm $*$, that is $\Delta_i = \mu_* - \mu_i$
- b, number of times that every arm got played at the initialization stage ($b \in \mathbb{N}$ and $T \geq b$). For example, for UCB1, $b = 1$.
- $n_{i,t}$, total number of times that arm i has been played ($n_{i,t} \in \mathbb{N}$ and $n_{i,t} \geq b$) before the time step t

3 UCBC Algorithm Annotations

For the UCBC algorithm, the following annotation is used:

- $U_{i,t}$ the upper confidence bound of arm i at time step t
- X_i, the random variable of arm i ($\mu_i = \mathbb{E}[X_i]$ and $\sigma_i^2 = Var(X_i)$)
- $\bar{X}_{i,t}$ the sample mean of arm i at time step t ($\bar{X}_{i,t} = \hat{\mu}_{i,t}$).
- $S_{n_{i,t}}^2$ the sample variance of arm i at time step t

4 UCBC Algorithm

Algorithm 1: UCBC

$\forall i \in [1, N]$ initialize count $n_{i,1} = 0$, sample mean $\bar{X}_{i,1} = 0$ and sample variance $S^2_{n_{i,1}} = 0$;
for $t = 1 \dots N$ **do**
  play arm $i = t$;
  $n_{i,t+1} \leftarrow n_{i,t} + 1$;
  observe reward $r_{i,t}$;
  use $r_{i,t}$ to update $\bar{X}_{i,t}$ and $S^2_{n_{i,t}}$ to get $\bar{X}_{i,t+1}$ and $S^2_{n_{i,t+1}}$;
end
for $t = N + 1 \dots T$ **do**
  for $i \in [1, N]$ **do**
    sample $Y_{i,t}$ from $\chi^2(n_{i,t})$;
    $V_{i,t} = \left(S^2_{n_{i,t}} + \frac{1}{n_{i,t}}\right) \tanh^{-1}(\log_{10}(n_{i,t} + 1))$;
  end
  play arm j=argmax$_{i\in[1,N]}\bar{X}_{i,t} + \sqrt{\frac{V_i \ln(t)}{Y_{i,t}}}$;
  $n_{j,t+1} \leftarrow n_{j,t} + 1$;
  observe reward $r_{j,t}$;
  use $r_{j,t}$ to update $\bar{X}_{j,t}$ and $S^2_{n_{j,t}}$ to get $\bar{X}_{j,t+1}$ and $S^2_{n_{j,t+1}}$;
end

The term $V_{i,t} = \left(S^2_{n_{i,t}} + \frac{1}{n_{i,t}}\right) \tanh^{-1}(\log_{10}(n_{i,t}+1))$ is to deal with the cold start problem of the UCBC algorithm as $n_{i,t}$ increases, $V_{i,t}$ approaches $S^2_{n_{i,t}}$. See Section 8.3 for detail.

5 UCBC Algorithm Derivation

This section derives the UCBC algorithm. While deriving, UCBC's sub-linear upper bound will be proven. The proof follows the same logic as in [3], though with modifications.

By Chebyshev's inequality,

$$\begin{aligned}
P(\bar{X}_{i,t} \geq \mathbb{E}[\bar{X}_{i,t}] + u_{i,t}) &= P(\hat{\mu}_{i,t} \geq \mu_i + u_{i,t}) && (1)\\
&\leq \frac{Var(\bar{X}_{i,t})}{u^2_{i,t}} && (2)\\
&\simeq \frac{Var(X_i)}{n_{i,t}u^2_{i,t}} \quad \text{when } n_{i,t} \text{ is large using Central Limit Theorem} && (3)\\
&= \frac{\sigma_i^2}{n_{i,t}u^2_{i,t}} && (4)\\
&= \delta && (5)
\end{aligned}$$

Let $\delta = \frac{1}{f(t)}$, then from (5)

$$u_{i,t} = \sqrt{\frac{\sigma_i^2 f(t)}{n_{i,t}}} \tag{6}$$

$$U_{i,t} = \hat{\mu}_{i,t} + u_{i,t} = \hat{\mu}_{i,t} + \sqrt{\frac{\sigma_i^2 f(t)}{n_{i,t}}} \tag{7}$$

If $\sigma_i = 0$, the $u_{i,t}$ will always be $\mu_{i,t}$, and arm i will always be judged and played correctly (to ensure performance, the UCBC algorithm uses a decaying sample variance if the real variance is 0). If $\sigma_i > 0$, for any arm i, the following two probability can be derived.

- if $n_{i,t} \geq \frac{K^2\sigma_i^2 f(t)}{\Delta_i^2}$, (where $K \in \mathbb{R}$, $K \geq 2$ and K makes $\frac{K^2\sigma_i^2 f(t)}{\Delta_i^2}$ an integer $\geq b$)

$$P(\hat{\mu}_{i,t} \leq \mu_i + \frac{\Delta_i}{K}) \geq 1 - \frac{1}{f(t)} \tag{8}$$

- the upper confidence bound $U_{i,t}$,

$$P(U_{i,t} > \mu_i) \geq 1 - \frac{1}{f(t)} \tag{9}$$

At any time step t, if any sub-optimal arm i has been played for $\frac{K^2\sigma_i^2 f(t)}{\Delta_i^2}$ times, then the probability of arm i is played at time t is no greater than $\frac{2}{f(x)}$. That is,

$$P\left(I_{t+1} = i \middle| n_{i,t} \geq \frac{K\sigma_i^2 f(t)}{\Delta_i^2}\right) \leq \frac{2}{f(x)} \quad \text{when } n_{i,t} \geq \frac{K^2\sigma_i^2 f(t)}{\Delta_i^2} \tag{10}$$

Proof of (10):

$$\begin{align}
U_{i,t} &= \hat{\mu}_{i,t} + \sqrt{\frac{\sigma_i^2 f(t)}{n_{i,t}}} \tag{11}\\
&\leq \hat{\mu}_{i,t} + \frac{\Delta_i}{K} \quad \text{because } n_{i,t} \geq \frac{K^2\sigma_i^2 f(t)}{\Delta_i^2} \tag{12}\\
&\leq \mu_i + \frac{\Delta_i}{K} + \frac{\Delta_i}{K} \quad \text{by (8) this fails at probability} \leq \frac{1}{f(t)} \tag{13}\\
&\leq \mu_i + \Delta_i \quad K \geq 2 \tag{14}\\
&= \mu_* \quad \text{definition of } \Delta_i \tag{15}\\
&\leq U_{*,t} \quad \text{by (9) this fails at probability} \leq \frac{1}{f(t)} \tag{16}
\end{align}$$

Therefore, by the union bound $P(U_{i,t} > U_{*,t}) \leq \frac{2}{f(t)}$, and that is when an suboptimal arm is played, and hence the probability of arm i gets played is no greater than $\frac{2}{f(t)}$.

To get sub-linear regret can be achieved by getting sub-linear $n_{i,T}$.

$$\begin{align}
\mathbb{E}[n_{i,T}] &= b + \mathbb{E}\left[\sum_{t=bN+1}^{T} \mathbb{1}(I_t = i)\right] \tag{17}\\
&= \underbrace{b + \mathbb{E}\left[\sum_{t=bN+1}^{T} \mathbb{1}\left(I_t = i, n_{i,t} < \frac{K^2\sigma_i^2 f(t)}{\Delta_i^2}\right)\right]}_{\text{can be played at most } \frac{K^2\sigma_i^2 f(T)}{\Delta_i^2} times} + \mathbb{E}\left[\sum_{t=bN+1}^{T} \mathbb{1}\left(I_t = i, n_{i,t} \geq \frac{K^2\sigma_i^2 f(t)}{\Delta_i^2}\right)\right] \tag{18}\\
&\leq \frac{K^2\sigma_i^2 f(T)}{\Delta_i^2} + \mathbb{E}\left[\sum_{t=bN+1}^{T} \mathbb{1}\left(I_t = i, n_{i,t} \geq \frac{K^2\sigma_i^2 f(t)}{\Delta_i^2}\right)\right] \tag{19}\\
&= \frac{K^2\sigma_i^2 f(T)}{\Delta_i^2} + \sum_{t=bN+1}^{T} P\left(I_t = i, n_{i,t} \geq \frac{K^2\sigma_i^2 f(t)}{\Delta_i^2}\right) \tag{20}\\
&= \frac{K^2\sigma_i^2 f(T)}{\Delta_i^2} + \sum_{t=bN+1}^{T} P\left(I_t = i \middle| n_{i,t} \geq \frac{K^2\sigma_i^2 f(t)}{\Delta_i^2}\right) P\left(n_{i,t} \geq \frac{K^2\sigma_i^2 f(t)}{\Delta_i^2}\right) \tag{21}\\
&\leq \frac{K^2\sigma_i^2 f(T)}{\Delta_i^2} + \sum_{t=bN+1}^{T} \left[\frac{2}{f(t)}\right] \tag{22}
\end{align}$$

The goal then is to choose an $f(t)$ that make the $\mathbb{E}[n_{i,T}]$ upper bound sub-linear to T. For the UCBC algorithm, $f(t) = \ln(t)$ is chosen, because:

- it result in a sub-linear upper bound
- it match the formula of UCB1 for comparison
- both $u_{i,t} = \sqrt{\frac{\sigma_i^2 \ln(t)}{n_{i,t}}}$ and $\delta = \frac{1}{\ln(t)}$ monotonically decreasing as t grows, which follows the rules for δ ([1] p103): "the first is whether the width of the confidence interval at a given confidence level can be significantly decreased, and the second is whether the confidence level is chosen in a reasonable fashion."

With $f(t) = \ln(t)$, the instance dependent bound can be derived:

$$\mathbb{E}[n_{i,T}] \leq \frac{K^2\sigma_i^2 \ln(T)}{\Delta_i^2} + \sum_{t=bN+1}^{T}\left[\frac{2}{\ln(t)}\right] \tag{23}$$

$$\leq \frac{K^2\sigma_i^2 \ln(T)}{\Delta_i^2} + \sum_{t=3}^{T}\left[\frac{2}{\ln(t)}\right] \tag{24}$$

$$\propto \frac{K^2\sigma_i^2 \ln(T)}{\Delta_i^2} + \frac{2T}{\ln(T)} \tag{25}$$

However, σ_i is unknown and still need to be taken care of. The idea is to use sampled variance $S^2_{n_{i,t}}$ to approximate.

$$u_{i,t} = \sqrt{\frac{\sigma_i^2 f(t)}{n_{i,t}}} \tag{26}$$

$$= \sqrt{\frac{\sigma_i^2 (n_{i,t}-1) S^2_{n_{i,t}} f(t)}{n_{i,t}(n_{i,t}-1)S^2_{n_{i,t}}}} \tag{27}$$

$$= \sqrt{\frac{(n_{i,t}-1) S^2_{n_{i,t}} f(t)}{n_{i,t} Y'}} \quad Y' \sim \chi^2(n_{i,t}-1) \tag{28}$$

$$\simeq \sqrt{\frac{S^2_{n_{i,t}} f(t)}{Y}} \quad Y \sim \chi^2(n_{i,t}) \tag{29}$$

Line (28) assumes Chi-square distribution, which is probably true when $n_{i,t}$ is large. Line (29), simplify line (28) by treating $n_{i,t} \simeq n_{i,t} - 1$. Let $f(t) = \ln(t)$, which gives the upper confidence bound of UCBC:

$$U_{i,t} = \hat{\mu}_{i,t} + \sqrt{\frac{S^2_{n_{i,t}} \ln(t)}{Y}} \tag{30}$$

Based on (25), the upper bound of the total regret of UCBC grows sub-linearly, but it is worse than logarithmic, so its performance should be worse than UCB1. However, it performs consistently better in all test cases (see Section 7), indicating there is a tighter upper bound. Next section gives an asymptotic analysis of this upper bound.

6 Asymptotic Analysis of Sub-linearity and the Choice of $f(t)$

Consider an UCB algorithm using (7) as the upper confidence bound, and $f(t)$ is monotonically increasing.

$$W_{i,t} \triangleq U_{i,t} - U_{*,t} \tag{31}$$

$$= \hat{\mu}_{i,t} - \hat{\mu}_{*,t} + \sqrt{\frac{\sigma_i^2 f(t)}{n_{i,t}}} - \sqrt{\frac{\sigma_*^2 f(t)}{n_{*,t}}} \tag{32}$$

At certain time step t, if both $n_{i,t}$ and $n_{*,t}$ are large enough so that, by the Central Limit Theorem, $\hat{\mu}_{i,t} \sim N(\mu_i, \frac{\sigma_i^2}{n_{i,t}})$ and $\hat{\mu}_{*,t} \sim N(\mu_*, \frac{\sigma_*^2}{n_{*,t}})$. From this,

$$W_{i,t} \sim N\left(\mu_i - \mu_* + \sqrt{\frac{\sigma_i^2 f(t)}{n_{i,t}}} - \sqrt{\frac{\sigma_*^2 f(t)}{n_{*,t}}}, \frac{\sigma_i^2}{n_{i,t}} + \frac{\sigma_*^2}{n_{*,t}}\right) \tag{33}$$

That is,

$$\frac{W_{i,t} - \left(\mu_i - \mu_* + \sqrt{\frac{\sigma_i^2 f(t)}{n_{i,t}}} - \sqrt{\frac{\sigma_*^2 f(t)}{n_{*,t}}}\right)}{\sqrt{\frac{\sigma_i^2}{n_{i,t}} + \frac{\sigma_*^2}{n_{*,t}}}} \sim N(0,1) \tag{34}$$

Choose $f(t) = \ln(t)$ and define

$$\begin{aligned} z &\triangleq -\frac{\mu_i - \mu_* + \sqrt{\frac{\sigma_i^2 f(t)}{n_{i,t}}} - \sqrt{\frac{\sigma_*^2 f(t)}{n_{*,t}}}}{\sqrt{\frac{\sigma_i^2}{n_{i,t}} + \frac{\sigma_*^2}{n_{*,t}}}} && (35)\\ &= \frac{\mu_* - \mu_i + \sqrt{\frac{\sigma_*^2 \ln(t)}{n_{*,t}}} - \sqrt{\frac{\sigma_i^2 \ln(t)}{n_{i,t}}}}{\sqrt{\frac{\sigma_*^2}{n_{*,t}} + \frac{\sigma_i^2}{n_{i,t}}}} && (36)\\ &= \frac{\Delta_i + \sqrt{\frac{\sigma_*^2 \ln(t)}{n_{*,t}}} - \sqrt{\frac{\sigma_i^2 \ln(t)}{n_{i,t}}}}{\sqrt{\frac{\sigma_*^2}{n_{*,t}} + \frac{\sigma_i^2}{n_{i,t}}}} && (37) \end{aligned}$$

Therefore,

$$P(W_{i,t} > 0) = \Phi^c(z) \tag{38}$$

Consider a 2-arm bandit. From (38) we can see that,

- if $z > 0$, $P(W_{i,t} > 0) < 0.5$, and the algorithm is more likely to play the optimal arm which in turn decreases z.
- if $z < 0$, $P(W_{i,t} > 0) > 0.5$, and the algorithm is more likely to play the sub-optimal arm which in turn increases z

Asymptotically, on average, the algorithm is balancing $n_{i,t}$ and $n_{*,t}$ as t grows. From (25), we know that $O(n_{*,t}) > O(\ln(t))$, and $O(n_{i,t}) \geq O(\ln(t))$. Therefore, the numerator of (37) converge to a constant C and $0 \leq C \leq \Delta_i$.

- C can be greater than 0 when $\Delta_i >> \sigma_i$
- C is non-negative (approximately) because the above-mentioned balancing effect
- C is not greater than Δ_i because $\sqrt{\frac{\sigma_*^2 \ln(t)}{n_{*,t}}} \to 0$ as t grows

Hence,

$$\Delta_i + \sqrt{\frac{\sigma_*^2 \ln(t)}{n_{*,t}}} - \sqrt{\frac{\sigma_i^2 \ln(t)}{n_{i,t}}} = C \tag{39}$$

$$\Rightarrow \frac{\Delta_i - C}{\sqrt{\ln(t)}} = \frac{\sigma_i}{\sqrt{n_{i,t}}} - \frac{\sigma_*}{\sqrt{n_{*,t}}} \tag{40}$$

$$\Rightarrow n_{i,t} = \frac{\sigma_i^2 \ln(t) n_{*,t}}{\left((\Delta_i - C)\sqrt{n_{*,t}} + \sigma_* \sqrt{\ln(t)}\right)^2} \tag{41}$$

$$= \frac{\sigma_i^2 \ln(t)}{\left((\Delta_i - C) + \sigma_* \sqrt{\frac{\ln(t)}{n_{*,t}}}\right)^2} \tag{42}$$

Because $\lim_{t\to\infty} \frac{\ln(t)}{n_{*,t}} = 0$, on average, $n_{i,t}$ is $O(\frac{\sigma_i^2 \ln(t)}{(\Delta_i - C)^2})$. For multi-arm bandits, intuitively,the magnitude of $\mathbb{E}[n_{i,t}]$ should be smaller than this, because there are multiple sub-optimal arms to play.

Notice that unlike UCB1 whose $\mathbb{E}[n_{i,t}]$ has a proven upper bound at each time step, $O(\frac{\sigma_i^2 \ln(t)}{(\Delta_i - C)^2})$ is an asymptotic analysis for UCBC. The $\mathbb{E}[n_{i,t}]$ of UCBC at early time steps may be linear. The analysis shown in this section uses the real variance and mean. However, in reality, UCBC needs to estimate these values. Errors in estimation may lead to linear total regret. In the next section, we see that UCBC performs well in all the tests. Also notice that the derivation of UCBC does not require the [0,1]-bound of rewards, which indicates that it can be used in such conditions.

7 Empirical Comparison

In this section UCBC is compared with UCB1. For UCB1, the exploration constant is set to 1, unlike the original UCB1 algorithm which used 2 instead [2]. Concretely, The UCB1 upper confidence bound is

$$U_{i,t} = \hat{\mu}_{i,t} + \sqrt{\frac{c \ln(t)}{n_{i,t}}} \quad c \text{ is the exploration constant} \tag{43}$$

$$= \hat{\mu}_{i,t} + \sqrt{\frac{\ln(t)}{n_{i,t}}} \quad c = 1 \text{ is used in implementation} \tag{44}$$

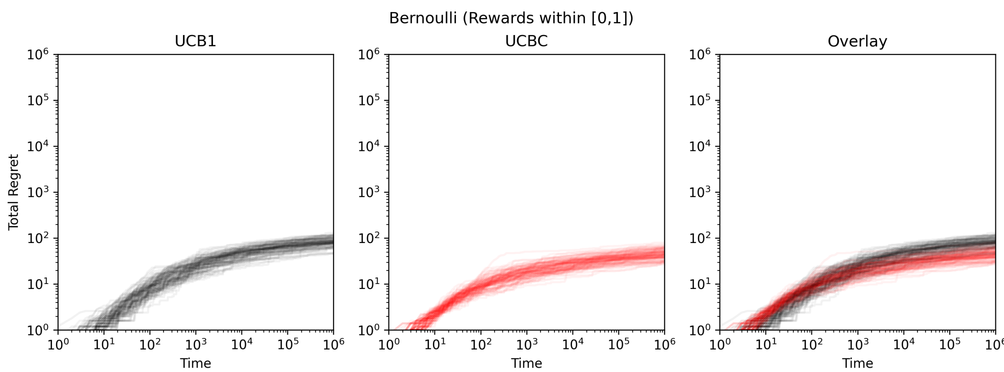
This is more comparable to UCBC's upper confidence bound (30), because $S^2_{n_{i,t}} \leq 1$ if rewards are bounded by $[0,1]$. The following bandits are used, all have three arms:

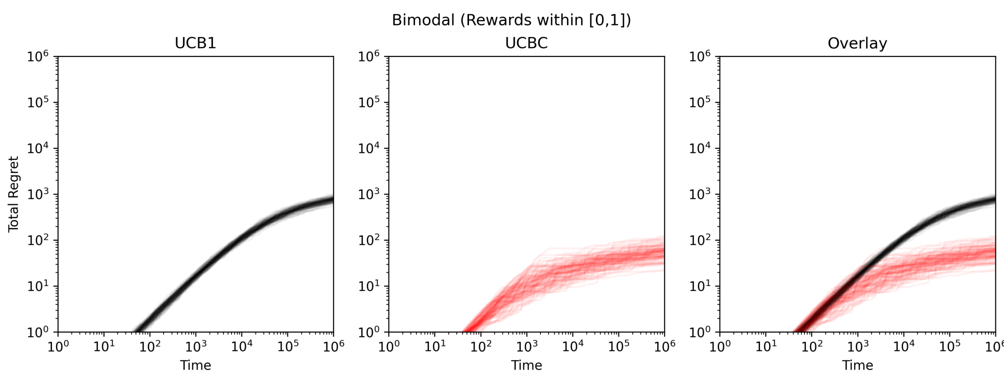
- Bernoulli distribution, rewards within [0,1]: $\theta_1 = 0.1$, $\theta_2 = 0.5$, $\theta_3 = 0.7$
- Bimodal distribution, rewards within [0,1]:
  $\theta_1 = N(0.4, 0.1^2), N(0.6, 0.3^2), 0.4$; $\theta_2 = N(0.5, 0.2^2), N(0.5, 0.2^2), 0.5$; $\theta_3 = N(0.6, 0.3^2), N(0.3, 0.1^2), 0.6$; (this is a mixture of two Normal distributions. $\theta = Normal1, Normal2, p$: with probability p use Normal1 else use Normal2).
- Bimodal distribution, rewards beyond [0,1]: $\theta_1 = N(4, 1^2), N(6, 3^3), 0.4$; $\theta_2 = N(5, 2^2), N(5, 2^2), 0.5$; $\theta_3 = N(6, 3^2), N(3, 1^2), 0.6$
- Normal distribution, rewards within [0,1]: $\theta_1 = (0.4, 0.1)$, $\theta_2 = (0.5, 0.2)$, $\theta_3 = (0.6, 0.3)$; first number is the μ, second is the σ
- Normal distribution, rewards beyond [0,1]: $\theta_1 = (4, 1)$, $\theta_2 = (5, 2)$, $\theta_3 = (6, 3)$;
- Uniform distribution, rewards within [0,1]: $\theta_1 = (0, 0.4)$, $\theta_2 = (0, .5)$, $\theta_3 = (0, 0.6)$; the two numbers are the range of the uniform distribution
- Uniform distribution, rewards beyond [0,1]: $\theta_1 = (0, 4)$, $\theta_2 = (0, 5)$, $\theta_3 = (0, 6)$; the two numbers are the range of the uniform distribution
- Bernoulli distribution, rewards within [0,1]: 10 arms with random parameters
- Bimodal distribution, rewards within [0,1]: 10 arms with random parameters
- Normal distribution, rewards within [0,1]: 10 arms with random parameters

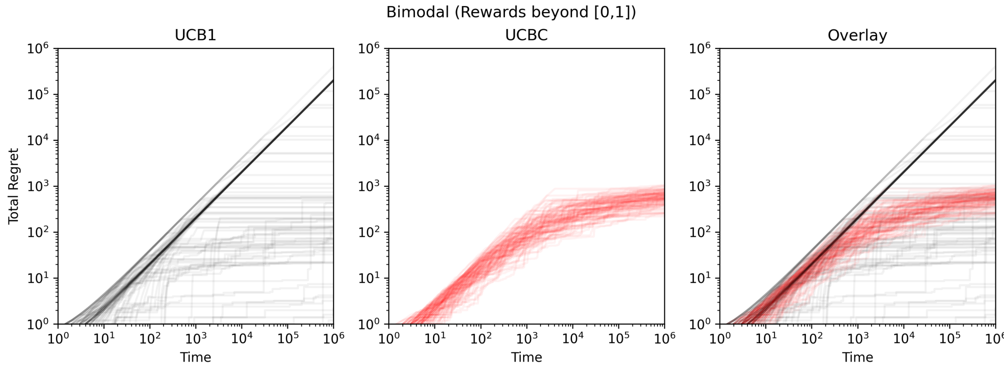
- Uniform distribution, rewards within [0,1]: 10 arms with random parameters
- Bernoulli distribution, rewards within [0,1]: 10 adversarial arms, $\theta s = 0.46, 0.47, 0.48, 0.49, 0.50, 0.51, 0.52, 0.53, 0.54, 0.55$
- Normal distribution, rewards within [0,1]: 10 adversarial arms, $\theta s = (0.46, 0.22), (0.47, 0.21), (0.48, 0.20), (0.49, 0.19), (0.50, 0.18), (0.51, 0.17), (0.52, 0.16), (0.53, 0.15), (0.54, 0.14), (0.55, 0.13)$. First number is μ, and second number is the σ.

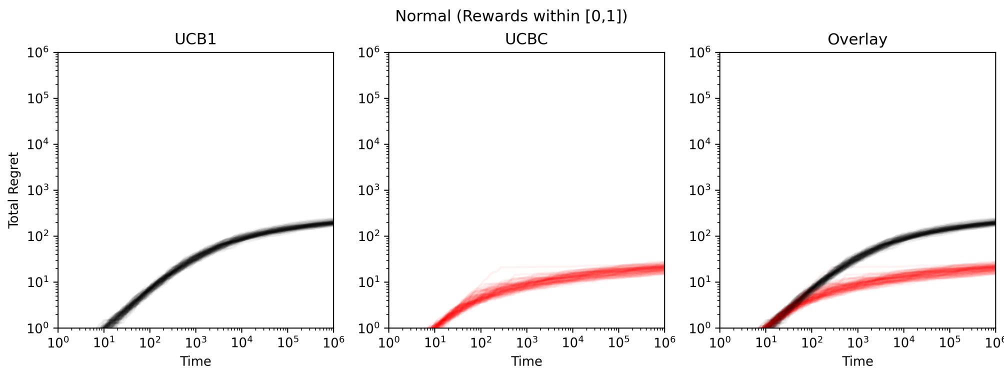
The rewards range of $[0, 1]$ for Bimodal distribution and Normal distribution is approximately true, but cannot be guaranteed due to the tails of the distribution. The adversarial arms are testing large-variance-small-regret arms. For each bandit, UCB1 (gray lines) and UCBC (red lines) were run for 1 million time steps and repeated 100 times, respectively. The color intensity represents the density of the lines in the plots: the darker the color the denser the lines. All the axes are in log scale.

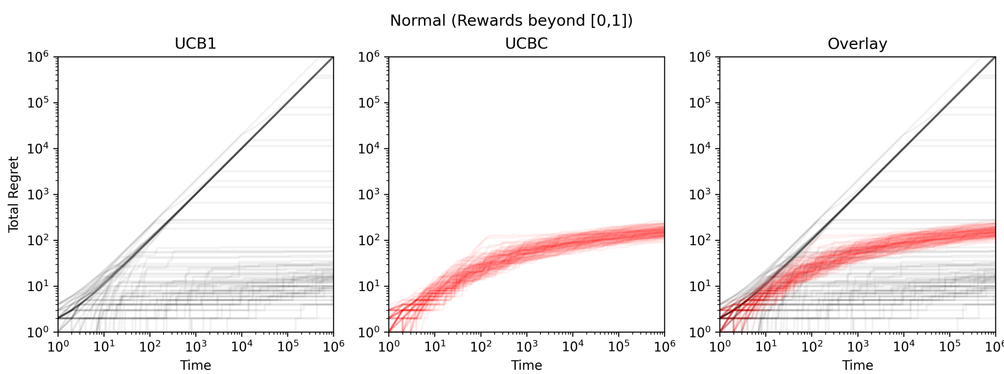
As shown in the following figures, when rewards are in $[0, 1]$, UCBC performs better than UCB1; when rewards are beyond $[0, 1]$, UCBC still has logarithmic total regret, while UCB1 suffers from linear total regret as expected.

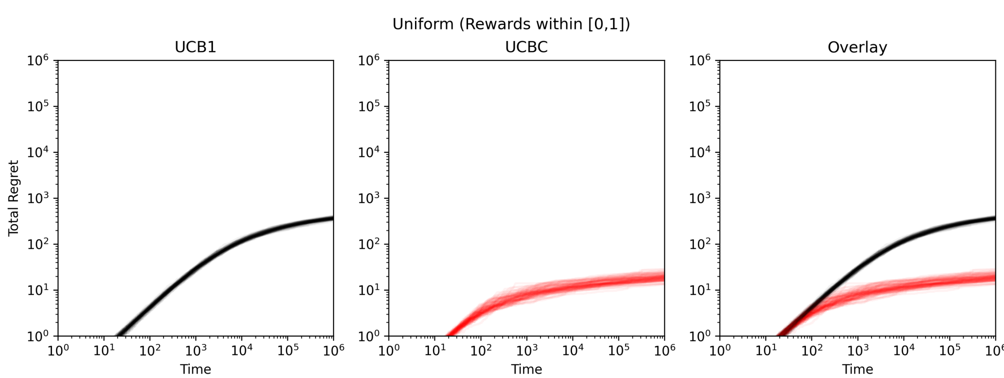


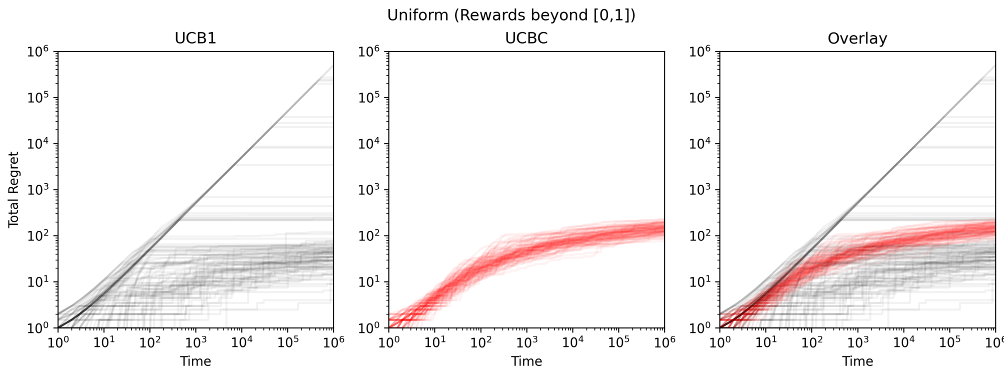


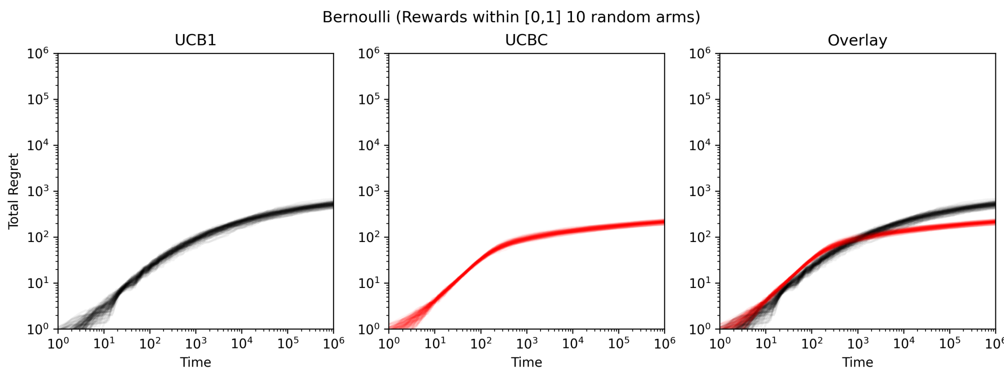

Bimodal (Rewards beyond [0,1])
UCB1
UCBC
Overlay
Total Regret
Time

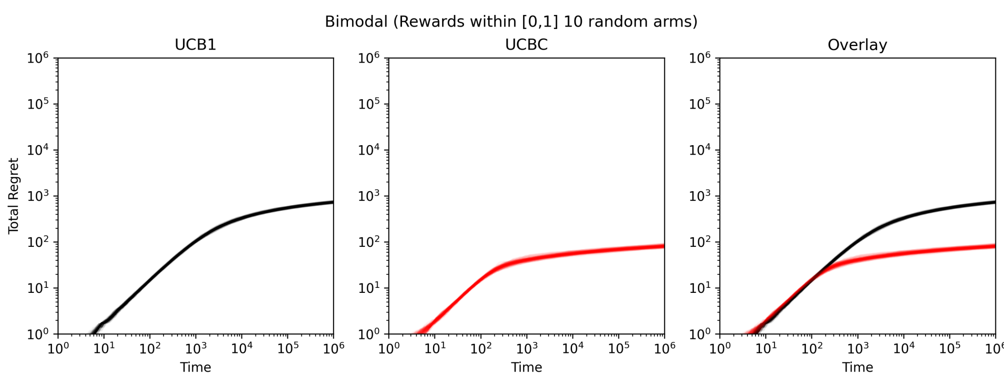

Normal (Rewards within [0,1])
UCB1
UCBC
Overlay
Total Regret
Time

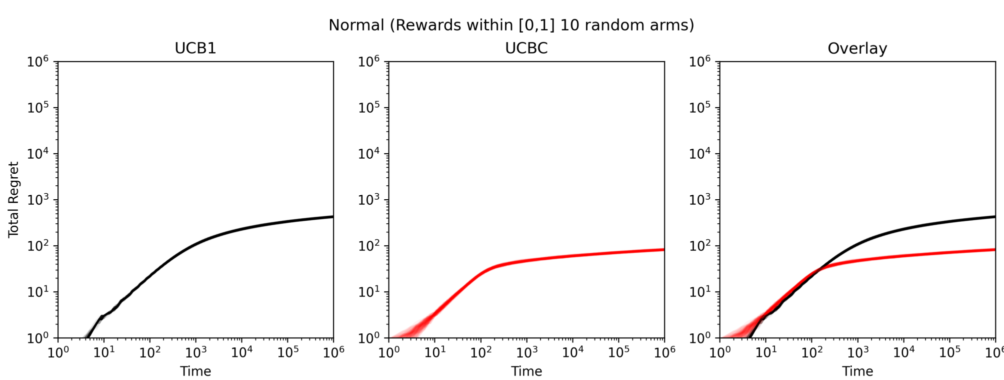

Normal (Rewards beyond [0,1])
UCB1
UCBC
Overlay
Total Regret
Time

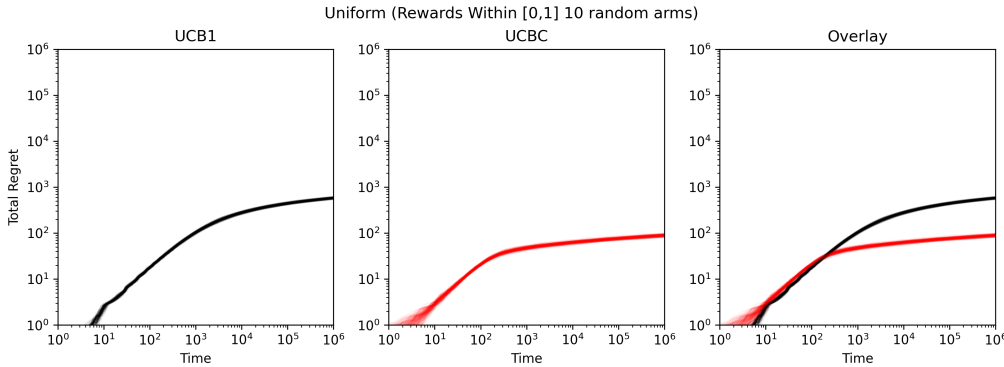

Uniform (Rewards within [0,1])
UCB1
UCBC
Overlay
Total Regret
Time

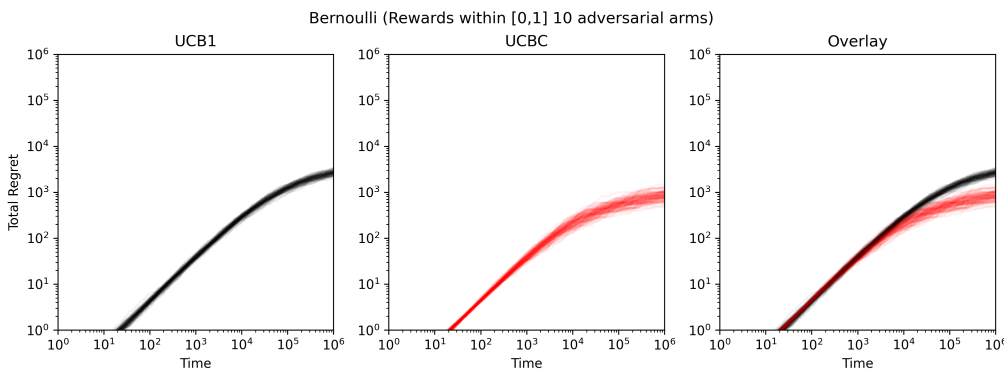

Uniform (Rewards beyond [0,1])
UCB1
UCBC
Overlay
Total Regret
Time

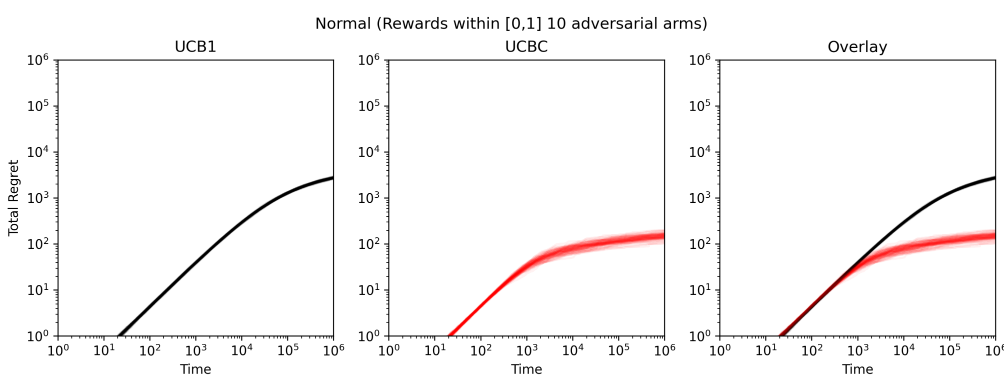

Bernoulli (Rewards within [0,1] 10 random arms)
UCB1
UCBC
Overlay
Total Regret
Time


Bimodal (Rewards within [0,1] 10 random arms)
UCB1
UCBC
Overlay
Total Regret
Time


Normal (Rewards within [0,1] 10 random arms)
UCB1
UCBC
Overlay
Total Regret
Time


Uniform (Rewards Within [0,1] 10 random arms)
UCB1
UCBC
Overlay
Total Regret
Time


Bernoulli (Rewards within [0,1] 10 adversarial arms)
UCB1
UCBC
Overlay
Total Regret
Time


Normal (Rewards within [0,1] 10 adversarial arms)
UCB1
UCBC
Overlay
Total Regret
Time

8 Discussion

8.1 Relationship between UCB1 and UCBC

From the upper confidence bound of UCB1 (45) and UCBC (30), it can be seen that UCB1 is using the upper bound of all arms' variances, while UCBC is actively estimating each arm's variance and use the estimated variance to calculated the upper bound.

Actually, if arm i's rewards are in range $[a_i, b_i]$, using the the general form of the Hoeffding's inequality and following the derivation of UCB1 [2], a more general form of upper confidence bound can be derived:

$$U_{i,t} = \hat{\mu}_{i,t} + \sqrt{\frac{(b_i - a_i)^2 \ln(t)}{n_{i,t}}} \tag{45}$$

$(b_i - a_i)^2$ is clearly larger than σ_i^2 which is estimated and used in UCBC.

8.2 Exploration Constant

Based on this observation, if UCBC's $\mathbb{E}[n_{i,t}]$ is logarithmic in every time step, the exploration constant of UCB1 can be smaller in some cases. For example, for Bernoulli multi-armed bandits, the exploration constant of UCB1 can be as low as 0.25 which is the largest possible variance of a Bernoulli distribution. Another example is $[0, 1]$-uniformly distributed arms, the variance is $\frac{1}{12}$ which can be set as the exploration constant of UCB1. We do see UCBC out perform UCB1 in both examples in Section 7.

Another question is can a small exploration constant, for instance $\frac{1}{4}$, be added to $\ln(t)$ in UCBC's upper bound? First, if this constant is too small, Chebyshev's inequality will have no meaning for early time steps in (5). Second, the asymptotic analysis still holds, but it may take millions of time steps before UCBC reaches the logarithmic total regret stage.

8.3 Jump-start UCBC

Although UCBC has a better performance and does not required the knowledge of the rewards before hand, it has a cold start problem. UCBC will not perform well until the variances are sufficiently estimated. This is especially hard at the beginning. The UCBC algorithm shown in Algorithm 1 uses the following formula to replace the sample variances to jump-start UCBC:

$$V_{i,t} = \left(S^2_{n_{i,t}} + \frac{1}{n_{i,t}} \right) \tanh^{-1}(\log_{10}(n_{i,t} + 1)) \tag{46}$$

- $\frac{1}{n_{i,t}}$ is a decaying addition: in case $S_{n_{i,t}}$ is small in the beginning, it add 1 at $n_{i,t} = 1$ and decaying to 0 as $n_{i,t}$ increases
- $\tanh^{-1}(\log_{10}(n_{i,t} + 1))$ is a decaying exploration encourage factor: it ≈ 3.42 at $n_{i,t} = 1$ and decaying to 1 as $n_{i,t}$ increases

These two factors encourage exploration at early stages to get a good estimate of each arm's variance and fairly quickly phase out as $n_{i,t}$ grows large.

9 Source Code

The source code used in this paper can be found at:
https://github.com/XiaoMutt/ucbc

References

[1] Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

[2] Nicolò Cesa-Bianchi Peter Auer and Paul Fischer. "Finite-time Analysis of the Multiarmed Bandit Problem". In: *Machine learning* 47.2 (2002), pp. 235–256.

[3] Karl Stratos and Jang Sun Lee. *Multi-armed bandits and reinforcement learning Lecture 3: UCB Algorithm*. URL: `https://ieor8100.github.io/mab/Lecture%203.pdf`.