TUNING CONFIDENCE BOUND FOR STOCHASTIC BANDITS WITH BANDIT DISTANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

We propose a novel modification of the standard upper confidence bound (UCB) method for the stochastic multi-armed bandit (MAB) problem which tunes the confidence bound of a given bandit based on its distance to others. Our UCB distance tuning (UCB-DT) formulation enables improved performance as measured by expected regret by preventing the MAB algorithm from focusing on non-optimal bandits which is a well-known deficiency of standard UCB. "Distance tuning" of the standard UCB is done using a proposed distance measure, which we call bandit distance, that is parameterizable and which therefore can be optimized to control the transition rate from exploration to exploitation based on problem requirements. We empirically demonstrate increased performance of UCB-DT versus many existing state-of-the-art methods which use the UCB formulation for the MAB problem. Our contribution also includes the development of a conceptual tool called the *Exploration Bargain Point* which gives insights into the tradeoffs between exploration and exploitation. We argue that the Exploration Bargain Point provides an intuitive perspective that is useful for comparatively analyzing the performance of UCB-based methods.

1 INTRODUCTION

Multi-armed bandit (MAB) (Slivkins, 2019) can model a broad range of applications, such as selecting the best website layout for users, or choosing the most profitable stocks among many. Stochastic bandits is an important setting in which MAB problems have been studied extensively. One of the most influential and widely used stochastic bandit policy is the upper confidence bound method (UCB) (Auer et al., 2002). UCB works by maintaining a mean estimation and confidence radius[1] for each bandit, and selects the bandit whose sum of mean and confidence bound is the maximum among all bandits at each step. The confidence bound can grow larger for less frequently used bandits which represents a higher degree of uncertainty to serve the exploration purpose.

However, the original UCB algorithm by its nature can lead to unsatisfactory results by being over-optimistic on non-optimal bandits. In this work, we propose UCB-DT (Upper Confidence Bound - Distance Tuning), a simple modification to the original UCB method which makes the confidence bound of a given bandit depend on its distance to others. Since the UCB-DT policy will select the largest bandit more often, exploration will naturally lean towards the neighbors of the largest bandit and prevent the algorithm from focusing on bandits that are farther away. Our proposed bandit distance is parmeterizable thereby offering the opportunity of customization over policies through different distances. Therefore, our formulation can represent a family of policies which inherently provide the flexibility of both pro-exploration policies, such as UCB, and pro-exploitation policies like ϵ-greedy.

[1] We will refer confidence radius as confidence bound in the rest of the paper.

Moreover, unlike previous UCB-based methods which focus on the $\log$ function in the confidence bound because of its analytical tractability, our method works differently by extending the denominator term. Using this enhancement to standard UCB, we propose a concept named *Exploration Bargain Point* to provide a novel perspective on analyzing performance of UCB-based methods. Using our new analysis tool, we intuitively and even graphically show that our method can always perform better than standard UCB.

We review existing work on the design of confidence bound for UCB in Sec. 3. While maintaining connections to some of these previous approaches, we make the following novel contributions in this paper.

- We propose UCB-DT policy, which tunes confidence bound by bandit distance. We conduct analysis and numerical experiments to show that our formulation is simple, extensible, and performant.
- We present a concept called *Exploration Bargain Point*, which provides a novel viewpoint on analyzing performance of upper confidence bound methods.

2 PRELIMINARIES

Let $k \in \mathbb{Z}^+$ be the number of bandits, T denote the time horizon, $\mu_i \in \mathbb{R}$ be the unknown mean for the subgaussian reward distribution of bandit i, $\mu_* = \max_i \mu_i$ be the mean reward of the optimal bandit, B_i be the shorthand for "ith bandit" where $i \in [k]$. In each round $t \in [T] = \{1, 2, ..., T\}$, the policy chooses a bandit $A_t \in [k]$ and whose reward is denoted by random variable X_t.

We use $\Delta_i = \mu_* - \mu_i$ to represent the suboptimality gap for B_i, and $N_i(t)$ to represent the number of times bandit i gets chosen till t. The regret over T rounds is

$$\mathcal{R}_T = \sum_{i=1}^{k} \Delta_i \mathbb{E}\left[N_i(T)\right], \text{ where } N_i(t) = \sum_{s=1}^{t} \mathbb{I}\left[A_s = i\right] \tag{1}$$

which serves as the main metric for stochastic bandit policies. A policy is called *asymptotically optimal* if

$$\lim_{T\to\infty} \frac{\mathcal{R}_T}{\log(T)} = \sum_{i:\Delta_i>0} \frac{2}{\Delta_i} \tag{2}$$

(Lai & Robbins, 1985; Burnetas & Katehakis, 1997) show that the above forms a regret upper bound for all consistent policies. A policy is called *sub-UCB* (Lattimore, 2018), which is a stricter requirement than *asymptotically optimality*, if there exists universal constants $C_1, C_2 > 0$, such that the regret can be finitely bounded as

$$\mathcal{R}_t \leq C_1 \sum_{i=1}^{k} \Delta_i + C_2 \sum_{i:\Delta_i(\mu)>0} \frac{\log(n)}{\Delta_i} \tag{3}$$

The UCB policy works by choosing bandit A_t such that

$$A_t = \arg\max_{i\in[k]} \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(t-1)}{N_i(t-1)}} \tag{4}$$

where the first term $\hat{\mu}_i(t) = (\sum_{c=1}^{t} \mathbb{I}(A_c = i)X_c)/N_i(t)$ represents the estimation of μ_i at time t, and the second term $\sqrt{2\log(t)/N_i(t)}$ denotes the confidence bound. UCB satisfies *sub-UCB* requirement (Auer et al., 2002) and is an *anytime* policy i.e. selection of the optimal bandit in Eq. 4 does not depend on T.

3 RELATED WORK

Multi-armed bandits. MAB (Slivkins, 2019) is a simple yet powerful framework for decision making under uncertainty. There are many categories of MAB problems including stochastic bandits proposed by (Gittins,

1979; Lai & Robbins, 1985; Katehakis & Veinott Jr, 1987), as the most classical one, where it is assumed that $\forall t, r_{ti}$ are samples drawn from a stationary sub-gaussian distribution bound to bandit i. Here r_{ti} represents the reward at time t on arm i. Adversarial bandits (Auer et al., 1995) assumes that for $\forall t, r_{ti}$ do not have to belong to any stationary distribution and can be set by an adversary. It is common to model r_{ti} as a secret codebook set by an enemy who knows the policy before playing. Contextual bandits (Langford & Zhang, 2007) introduces an observable context variable and assumes r_{ti} is drawn a distribution parameterized by both bandit i and a context variable. There are many other variants under the MAB framework, whose details are beyond the scope of this paper. In this paper, we focus on the stochastic bandits problem.

UCB formulation for stochastic bandits. Since being first proposed, UCB (Auer et al., 2002) has received strong research interest and several variants of the original UCB policy have been proposed. For example, KL-UCB (Garivier & Cappé, 2011; Cappé et al., 2013) and KL-UCB++ (Ménard & Garivier, 2017) transform the UCB policy as a procedure that calculates the best possible arm using a Kullback-Leibler divergence bound at each time step. UCBV-Tune (Audibert et al., 2009) incorporates the estimated variance of reward distribution instead of assuming unit variance.

In the context of this paper, several previous authors have proposed formulations which redesign the confidence bound of standard UCB as shown in the term of Eq. 4. MOSS (Audibert & Bubeck, 2010) makes the confidence bound depend on the number of plays for each bandit by replacing $\log(t)$ with $\log(t/N_i(t))$ in Eq. 4, and policies similar to MOSS include OCUCB (Lattimore, 2016) and UCB* (Garivier et al., 2016). UCB† (Lattimore, 2018) improves upon the previous ones significantly by designing a more advanced log function component.

Compared with these previous approaches, our method works differently by extending the denominator term rather than the $\log$ function. It turns out to be intuitive in foresight and delivers strong performance. Moreover, instead of being non-parametric like most of the above methods, our policy has parameters which are tunable, which allows that the formulation of our method to encompass a family of policies.

4 METHOD

4.1 INTUITION

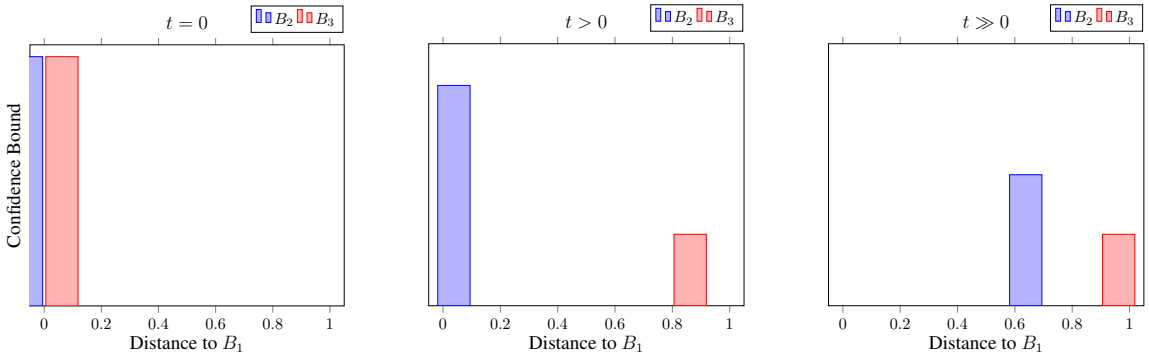
The core idea of UCB-DT is that when a bandit is selected, instead of increasing the confidence bounds of all other bandits uniformly, we increase them more for bandits which are similar to the current chosen bandit and vice versa. The intuition is that a similar bandit has a higher chance to be equally good as the current chosen one. As a result, this strategy will naturally lean towards the optimal bandit and spend exploration budget on similarly good bandits and save unnecessary trials with poor bandits.

To realize this idea, we start by looking at the confidence bound term $\sqrt{2\log(t)/N_i(t)}$ for B_i in Eq. 4. Suppose there exists a distance measure $d(i, j)$ ranges between $[0, 1]$ that compares the distance between B_i and B_j. Then we find that the above idea can be implemented by replacing $N_i(t)$ as

$$N_i(t) \Rightarrow N_i(t) + \sum_{j \in [k], j \neq i} d(i, j) N_j(t) \tag{5}$$

This modification can shrink the confidence bounds of bandits which are distant from others, while maintaining the confidence bounds of those ones that are closer for further exploration. Furthermore, this modification elegantly depicts the poles of exploration and exploitation as below:

Exploration When $d(i, j) \equiv 0$, Eq. 5 will degrade to the vanilla UCB case, which is an "optimistic" policy and encourages exploration.



(a) Given no prior knowledge, the policy assumes B_1, B_2, B_3 have no difference, $d_t(1,2) = d_t(1,3) = 0$, which renders large confidence bounds and promote exploration.

(b) Learning that B_3 is far from B_1, $d_t(1,3)$ will be close to 1, which inhibits its confidence bound to grow larger. Thus, the policy will focus on comparing B_1 and B_2.

(c) Given the expanding property, $d_t(1,2)$ will grow larger even if B_1, B_2 are close. So the confidence bound of B_2 will cease to dominate eventually, policy can focus on B_1.

Figure 1: A visual demonstration for our intuition with an example B_1, B_2, B_3 whose $\mu_1 > \mu_2 \gg \mu_3$ and $\mu_1 = \mu_2 + \epsilon$.

Exploitation When $d(i,j) \equiv 1$, Eq. 5 will degrade to greedy case because $N_i(t) = t$ and $\log(t)/t$ can rapidly converge to 0, which purely exploits.

These two special cases correspond exactly to the simple and commonly recognized truth: *In the stochastic bandit problem, you shall explore just enough to find the right bandit, then exploit that as long as you can.*

In our formulation, we can model this transition from exploration to exploitation by customizing d. It indicates that if we could find a transition from $d \equiv 0$ to $d \equiv 1$, like expanding d from the origin to a unit circle, then we can realize the above truth under the framework in Eq. 5.

Algorithm 1: UCB distance tuning (UCB-DT)

input : distance measure d_t

$$A_t = \arg\max_{i \in [k]} \hat{\mu}_i(t-1) + \sqrt{\frac{2\log(t-1)}{\widetilde{N}_i(t-1)}}$$

$$\text{where } \widetilde{N}_i(t) = N_i(t) + \sum_{j \in [k], j \neq i} d_t(i,j) N_j(t)$$

Therefore, we write d as time dependent as d_t and condense the above findings in Alg. 1. An example is provided to demonstrate our ideas in Fig. 1. A specific instance of d will be introduced in Sec. 4.2.

4.2 Bandit Distance

We design the following distance which composes UCB-DT(μ).

UCB-DT(μ)

$$d_t(i,j) = |\hat{\mu}_i(t) - \hat{\mu}_j(t)|^{1/\lfloor \gamma N_i(t) \rfloor} \tag{6}$$

First, it directly measures the distance between two bandits. Second, the more often a bandit gets pulled, the closer its distances from all other bandits will approach the maximum value of 1, which means that the policy

for this bandit transitions deeper into exploitation from exploration. Here γ is a speed parameter to control the transition rate. It could be pointed that the UCB-DT(μ) will not work properly for bandits whose $\mu_i \gg 1$ because the distance will saturate. But we argue that it is common practice to use normalization to bypass this constraint. Thus, we keep the above design for simplicity.

In Appendix B, we provide more designs for d and analyze their characteristics. In Fig. 2, we visualize the distance versus N_i using different γ and $|\hat{\mu}_i - \hat{\mu}_j|$.

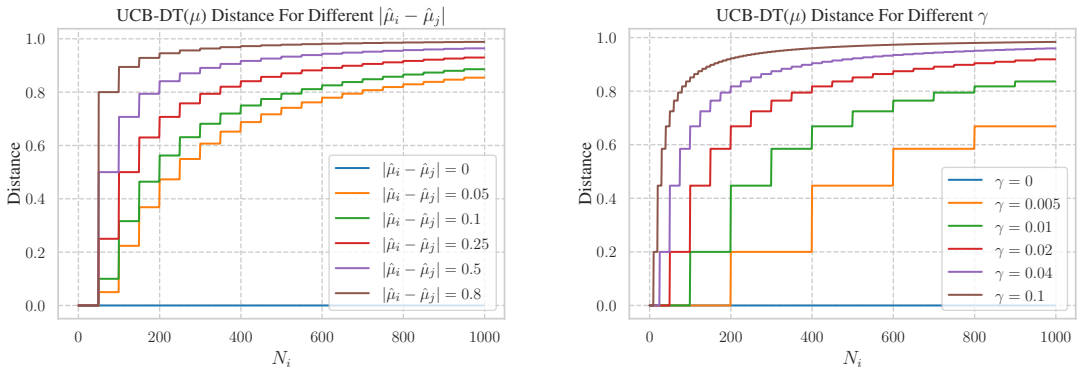


Figure 2: Visualization of the Distance in Eq. 6. γ is set to 0.02 in the left figure, and $|\hat{\mu}_i - \hat{\mu}_j|$ is set to 0.2 in the right figure. It can be clearly seen that the greater the difference between the mean rewards of two bandits is, the faster that the distance is expanded from 0 to 1. The rate of convergence of d to 1 can also be controlled by increasing γ. The curves are jagged because of the floor operation $\lfloor \gamma N_i \rfloor$.

4.3 Under Exploration Analysis

As compared to standard UCB, our proposed formulation UCB-DT performs less exploration. In this section, we conduct an analysis based on a novel concept called *Exploration Bargain Point* to show that our method can always give better performance than UCB. Based on our analysis, we also provide practical guidelines on how to set the parameter γ.

In the following discussion, we assume a scenario of two bandits where $\mu_1 > \mu_2$, and we also assume that $N_1 \geq N_2$. Since we only have two bandits, we can regard $N_2(T)$ as the exploration budget spent till time T, and $T = N_1 + N_2$. We conduct our analysis by hindsight[2] in the context of standard UCB.

4.3.1 Exploration Full Point

If we want to explore enough to ensure $P(A_T = 1) \geq 1 - \delta$, this implies that we recognize the optimal bandit as the dominant choice. Therefore, based on Eq. 17 which is discussed in the formulation of standard UCB in Appendix C, we have:

$$P\left(\sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)} \leq \Delta_2/2\right) \leq \delta \tag{7}$$

[2]"By hindsight" means we look back at a policy's decisions from time T.

where $\sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)}$ is the largest possible deviation of mean estimation. Since we have $\delta = 1/t$ in UCB, we can solve for $N_2(T)$ when equality holds for the argument of P in Eq. 7. To simplify notation, we skip T in the argument of N_2:

$$\begin{aligned} \sqrt{\frac{2}{N_2}\log(T)} &= \frac{\Delta_2}{2} \\ N_{full} = N_2 &= \frac{8\log(T)}{\Delta_2^2} \end{aligned} \tag{8}$$

Eq. 8 carries physical meaning, as it implies that if we explores for N_{full} times, then the confidence bound will shrink below to half of the suboptimality gap Δ_2. In this case we can safely choose B_1, and any more exploration is completely unnecessary. To put in a succinct manner of speaking, by exploring N_{full} times according to the above equation, we fulfill the confidence bound of UCB.

Therefore, we write N_2 in Eq. 8 as *Exploration Full Point* with N_{full}. Let G_{full} denotes the expected cumulative reward[3] at N_{full}:

$$G_{full} = (T - N_{full})\mu_1 + N_{full}\mu_2 \tag{9}$$

4.3.2 Exploration Bargain Point

The key question is whether we could stop exploring before N_{full} and still achieve better performance? The trade-off is that we have a smaller N_2 by exploring less, and the probability of recognizing the non-optimal bandit as the dominant choice $P(A_T = 2) = \delta$ will grow larger and become non-negligible. Therefore, with a slight abuse of notation, we can write a lower bound for expected cumulative reward as $G(N_2)$:

$$G(N_2) \geq \underbrace{\left((T-N_2)\,\mu_1 + N_2\mu_2\right)}_{\text{Correctly choose } B_1 \text{ as the best one}} \overbrace{(1-\delta)}^{\inf P(A_T=1)} + \underbrace{\left((T-N_2)\,\mu_2 + N_2\mu_1\right)}_{\text{Mistakenly choose } B_2 \text{ as the best one}} \overbrace{\delta}^{\sup P(A_T=2)} \tag{10}$$

where $\delta = e^{-N_2\Delta_2^2/8}$ is from Eq. 8

Next, to ensure we can achieve better or at least the same performance as N_{full}, we write

$$G(N_2) \geq G_{full} \tag{11}$$

Then, to find out the boundary condition where the same performance are achieved, we can let G_{full} in Eq. 9 to equal the right hand side of the Eq. 10.

$$G_{full} = \left((T-N_2)\,\mu_1 + N_2\mu_2\right)(1-\delta) + \left((T-N_2)\,\mu_2 + N_2\mu_1\right)\delta \tag{12}$$

We write the solution of Eq. 12 as *Exploration Bargain Point* with $N_{bargain} = \underset{N_2}{\text{solution}}\{\text{Equation 12}\}$ because we achieve at least the same expected reward by only exploring to the point of $N_{bargain}$. We continue the discussion of the form of the solution in Appendix D.

[3]For ease of analysis, we use reward instead of regret for under exploration analysis. The same conclusion holds if regret is used in this analysis.

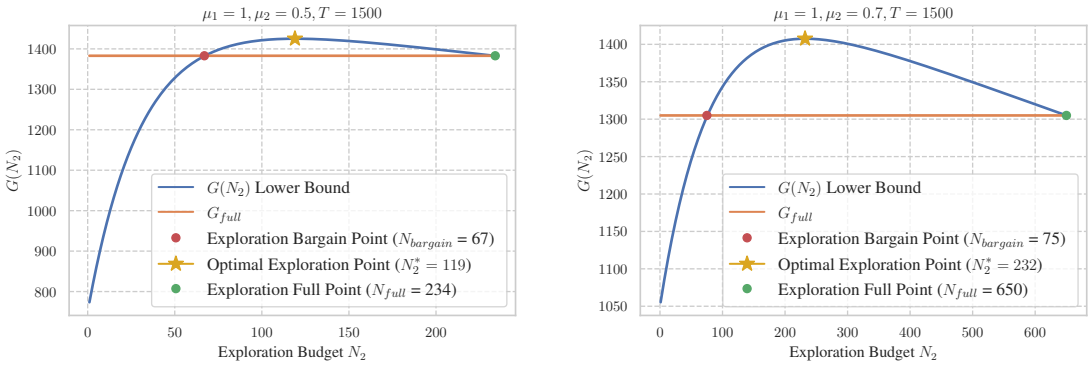


Figure 3: Examples of the relationship between expected cumulative reward $G(N_2)$ and exploration budget N_2

Determining γ with $\mathbf{N_{bargain}}$. In Fig. 3, we visualize the relationship among N_{full}, $N_{bargain}$, and $G(N_2)$ by examples. It is fascinating to see that the optimal exploration point is located between $N_{bargain}$ and N_{full} because of the concavity of the subgaussian reward distribution. Therefore, our method can always perform better than UCB as long as γ is set to $1/N_{bargain}$ in Eq. 6. It ensures that after exploring beyond $N_{bargain}$ (red dot), the distance in Eq. 6 will become larger than 0, which makes the confidence bounds of all bandits in Alg. 1 smaller than that of UCB. By having smaller confidence bounds, our policy will under-explore compared to UCB and stop before the N_{full} (green dot)[4]. Thus, the final N_2 will lie between $N_{bargain}$ and N_{full}, where $G(N_2) > G_{full}$. Moreover, as shown by Eq. 19 in Appendix D, $N_{bargain}$ only depends on suboptimality gap Δ_2 and time horizon T, and does not rely on the actual values of μ_1, μ_2. We argue that domain knowledge could be used to estimate the difference between optimal bandit and other ones in practice and thereby estimate γ.

Existence of $\mathbf{N_{bargain}}$. $N_{bargain}$ always exists and is less than N_{full} as long as $\mu_1 > \mu_2$, and this can be proven by contradiction. Because of the concavity of subgaussian distribution, the only possible case where this condition is not satisfied is when $N_{bargain} = N_{full}$, in which the G_{full} (orange curve) becomes the tangent line of $G(N_2)$ (blue curve). From Eq. 12, we get

$$((T - N_{full})\,\mu_1 + N_{full}\mu_2) = ((T - N_{full})\,\mu_1 + N_{full}\mu_2)\,(1 - \delta) + ((T - N_{full})\,\mu_2 + N_{full}\mu_1)\,\delta \tag{13}$$

It follows directly that $\delta = 0$. Since $\delta = e^{-N_2\Delta_2^2/8}$ from Eq. 10, we get $N_2 = \inf$. Then, from Eq. 8, $N_2 = \inf \rightarrow \Delta_2 = 0$. However, we assume $\Delta_2 = \mu_1 - \mu_2 > 0$. So $N_{bargain}$ must always exist.

Implications. Exploration Bargain Point describes exactly the over-exploring nature of UCB. We could use $N_{bargain}$ as an anchor to optimize the UCB method:

- It allows UCB practitioners to early stop and prevent over-exploration. We summarize this policy as UCB-then-Commit in Appendix B. In Table 2, we see that UCB-then-Commit outperforms UCB in most experiments but is never as good as UCB-DT(μ).
- $N_{bargain}$ helps us understand when UCB-DT(μ) crosses into the optimal territory to get better rewards than UCB.

This insight is not available using traditional regret bound analysis. We therefore believe that our analysis tool provides a novel and more intuitive perspective on analyzing UCB-based methods.

[4]A policy could stop before $N_{bargain}$ if γ is too high, while a γ too low causes the policy to approach N_{full}.

4.4 REGRET ANALYSIS

In addition to our analysis in Sec. 4.3, we offer a finite time regret bound analysis for UCB-DT following standard practice. For the sake of clarity of notation, we write $N_i(t)$ as $N_i(n)$. Suppose $\mu_1 > \mu_2 > .. > \mu_k$, we can infer from Eq. 1 that the task of bounding R_t can be translated as bounding $\mathbb{E}\left[N_i(t)\right], i \in [k], i \neq 1$. The main idea is to divide the analysis into two cases when we choose the suboptimal bandit over the optimal one:

(A) μ_1 is under estimated, so the optimal bandit B_1 appears worse.

(B) $\mu_i, i \in [k], i \neq 1$ is over estimated, so the non-optimal bandit B_i appears better.

$$
\begin{aligned}
N_i(n) &= \sum_{t=1}^{n} \mathbb{I}\left\{A_t = i\right\} \\
&\leq \underbrace{\sum_{t=1}^{n} \mathbb{I}\left\{\hat{\mu}_1(t) + \sqrt{\frac{2\log(t)}{\widetilde{N}_1(t)}} \leq \mu_1 - \varepsilon\right\}}_{\text{(A)}} + \underbrace{\sum_{t=1}^{n} \mathbb{I}\left\{\hat{\mu}_i(t) + \sqrt{\frac{2\log t}{\widetilde{N}_i(t)}} \geq \mu_1 - \varepsilon \text{ and } A_t = i\right\}}_{\text{(B)}}
\end{aligned} \tag{14}
$$

Based on Eq. 5 we have $\widetilde{N}_i(t) \geq N_i(t)$, we can therefore relax (B) by replacing $\widetilde{N}_i(t)$ with $N_i(t)$, which makes it identical to the case of UCB and (B) can be bounded by $1 + \frac{2(\log t + \sqrt{\pi \log t} + 1)}{(\Delta_i - \varepsilon)^2}$ according to (Lattimore & Szepesvári, 2020). At the same time, while $\widetilde{N}_i(t) \leq t$, we cannot apply a similar procedure to (A) because directly replacing $\widetilde{N}_i(t)$ with t will make (A) grow much faster and violate the *sub-UCB* condition.

Assumption 1 *There exists a time step τ, which ensures the policy chooses A_1 often enough:*

$$
\tau = \min\left\{t \leq T : \sup_{s \geq t} |\hat{\mu}_1(s) - \mu_1| < \varepsilon\right\}
$$

If we can leverage assumption 1, then a simple observation is that:

$$
\sum_{t=1}^{T} \mathbb{I}\left\{\hat{\mu}_1(t) + \sqrt{\frac{2\log(t)}{\widetilde{N}_1(t)}} \leq \mu_1 - \varepsilon\right\} \leq \tau + \sum_{t=\tau+1}^{T} \mathbb{I}\{\mu_1 - \hat{\mu}_1(t) \geq \varepsilon + \delta\} \tag{15}
$$

It is straightforward to see that $P(\mu_1 - \hat{\mu}_1(t) \geq \varepsilon + \delta \mid t > \tau) = 0$, where $\sqrt{2\log(t)/t}$ is replaced with $\delta > 0$. Therefore, (A) can be bounded by τ. Furthermore, if we let bandit reward distribution to be gaussian, then by the concentration Lemma 1 in (Lattimore, 2018), $\mathbb{E}[\tau] \leq 1 + 9/\varepsilon^2$. Thus the regret of UCB-DT satisfies the *sub-UCB* requirement.

In the case of UCB-DT(μ), we can satisfy the assumption 1 as long as γ is small enough, such as $1/N_{bargain}$ according to Sec. 4.3.2. In fact, UCB-DT can represent a family of policies through parameterization of d, and the regret bound here only describes the boundary behavior as it approaches the standard UCB. Based on our previous analysis, there exists a parameter space of d which can be used to customize the policy to achieve better exploration-exploitation trade-off. Therefore, instead of diving into a detailed regret expression of any specific form, we provide a general analysis without explicitly including this degree of freedom.

5 NUMERICAL EXPERIMENTS

We compare UCB-DT to UCB (Auer et al., 2002), UCB† (Lattimore, 2018), UCBV-Tune (Audibert et al., 2009), KL-UCB (Garivier & Cappé, 2011; Cappé et al., 2013) and KL-UCB++ (Ménard & Garivier, 2017).

Among these methods, our main comparisons are done for UCB, UCB† and UCBV-Tune. UCB-DT can also be extended to support variance adaption. The remaining two methods, KL-UCB and KL-UCB++, which deliver excellent performance, are *non-anytime* policies. Because of these structural differences, we do not regard them as our main comparisons, however we still keep them as important references.

Experiment	KL-UCB	KL-UCB++	UCBV-Tune	UCB†	UCB	UCB-DT(μ)
B5 *	41.73	**38.03**	98.6	129.01	251.28	70.95
B20 *	168.54	**129.5**	415.23	410.5	939.99	383.82
B(0.02, 0.01) *	28.34	22.39	64.99	131.99	249.17	**21.6**
B(0.9, 0.88) *	38.44	33.78	49.6	73.9	119.91	**19.19**
N5 *	194.82	109.75	119.97	88.3	142.21	**83.65**
N20	539.91	**410.25**	840.26	560.47	1014.27	640.52

Table 1: Expected regret of different policies at $T = 20000$ in each experiment. For each experiment we mark bold text for the best result and "*" if UCB-DT outperforms UCB, UCB† and UCBV-Tune.

For thorough evaluation, we design 6 experiments as below

B5 Bernoulli reward, 5 bandits with expected rewards 0.9, 0.8, 0.7, 0.2, 0.5. This experiment is modified from (Garivier & Cappé, 2011) by adding more bandits.

B20 Bernoulli reward with many bandits, 20 bandits with expected rewards 0.9, 0.85, 0.8, 0.8, 0.7, 0.65, 0.6, 0.6, 0.55, 0.5, 0.4, 0.4, 0.35, 0.3, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05.

B(0.02, 0.01) Bernoulli reward with low means, 3 bandits with expected rewards 0.05, 0.02, 0.01. This experiment is borrowed identically from (Garivier & Cappé, 2011).

B(0.9, 0.88) Bernoulli reward with close means, 2 bandits with expected rewards 0.9, 0.88.

N5 Gaussian reward, 5 bandits with unit variance and expected rewards 1, 0.8, 0.5, 0.3, -0.2.

N20 Gaussian reward with many bandits, 20 bandits with unit variance and expected rewards 0, -0.03, -0.03, -0.07, -0.07, -0.07, -0.15, -0.15, -0.15, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -0.5, -1, -1. This experiment is identical to the one used in (Lattimore, 2018).

We set T to 20000 and run 2000 simulations for each method on every experiment, and we set γ to 0.02 in all experiments. For other methods, we use implementations and default parameters from (Besson, 2018). We summarize cumulative regrets in table 1 and Fig. 4.

From our simulation results, we can see that UCB-DT is always better than UCB and outperforms UCB†, UCBV-Tune in first 5 experiments. The underlying reason for the under performance of UCB-DT as compared to UCB† in N20 can be attributed to the high variance of the bandit reward distribution, which makes the estimation of distance unstable. This performance degradation can be mitigated by using a different distance as discussed in Appendix B.

6 Conclusion

By leveraging bandit distance, we create a policy called UCB-DT, which is simple, extensible and performant. Using our proposed framework, we propose the concept of *Exploration Bargain Point* to provide a new perspective on analyzing performance of UCB-based methods. Admittedly, this work bears its own limitations. We do not dive deeper in our regret bound analysis to describe the relationship between convergence behavior and possible properties of d. We only study the *Exploration Bargain Point* in the context of our method and standard UCB and have not applied it to other UCB-based policies. However, we believe that these issues could be addressed in future work, and it may be promising to adopt the idea in this paper to more general settings like adversarial bandit and reinforcement learning.

REFERENCES

Jean-Yves Audibert and Sébastien Bubeck. Regret bounds and minimax policies under partial monitoring. *The Journal of Machine Learning Research*, 11:2785–2836, 2010.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pp. 322–331. IEEE, 1995.

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

Lilian Besson. SMPyBandits: an Open-Source Research Framework for Single and Multi-Players Multi-Arms Bandits (MAB) Algorithms in Python. Online at: `github.com/SMPyBandits/SMPyBandits`, 2018. URL `https://github.com/SMPyBandits/SMPyBandits/`. Code at https://github.com/SMPyBandits/SMPyBandits/, documentation at https://smpybandits.github.io/.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback-leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, pp. 1516–1541, 2013.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th annual conference on learning theory*, pp. 359–376. JMLR Workshop and Conference Proceedings, 2011.

Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. *Advances in Neural Information Processing Systems*, 29:784–792, 2016.

John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):148–164, 1979.

Wolfram Research, Inc. Mathematica, Version 12.3.1. URL `https://www.wolfram.com/mathematica`. Champaign, IL, 2021.

Michael N Katehakis and Arthur F Veinott Jr. The multi-armed bandit problem: decomposition and computation. *Mathematics of Operations Research*, 12(2):262–268, 1987.

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

John Langford and Tong Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. *Advances in neural information processing systems*, 20(1):96–1, 2007.

Tor Lattimore. Regret analysis of the anytime optimally confident ucb algorithm. *arXiv preprint arXiv:1603.08661*, 2016.

Tor Lattimore. Refining the confidence level for optimistic bandit strategies. *The Journal of Machine Learning Research*, 19(1):765–796, 2018.

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

Pierre Ménard and Aurélien Garivier. A minimax and asymptotically optimal algorithm for stochastic bandits. In *International Conference on Algorithmic Learning Theory*, pp. 223–237. PMLR, 2017.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *CoRR*, abs/1904.07272, 2019. URL http://arxiv.org/abs/1904.07272.

Eric W Weisstein. Lambert w-function. *https://mathworld. wolfram. com/*, 2002.

A REGRET CURVES

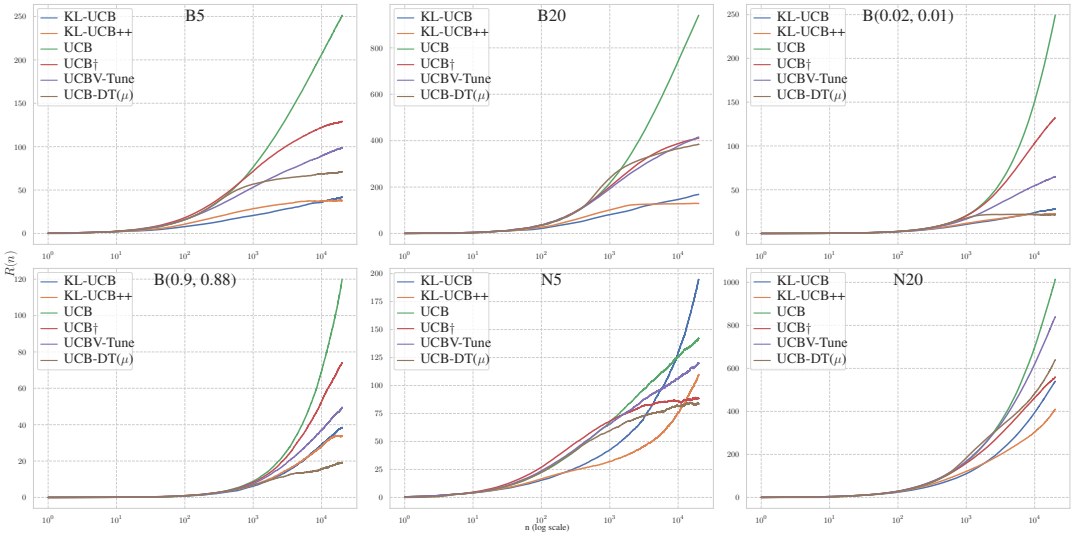


Figure 4: Regret of different policies as a function of time (log scale) in experiments from Sec. 5.

B MORE BANDIT DISTANCES

We experiment the following two additional distances:

UCB-then-Commit

$$d_t(i,j) = \begin{cases} 0 & N_i(t) \leq \lfloor 1/\gamma \rfloor \\ 1 & N_i(t) > \lfloor 1/\gamma \rfloor \end{cases}$$

UCB-DT(μ margin) $d_t(i,j) = (|\hat{\mu}_i(t) - \hat{\mu}_j(t)| - m)^{1/\lfloor \gamma N_i(t) \rfloor}$

We name the first strategy as UCB-then-Commit, which enables the transition from exploration to exploitation occur based on $N_{bargain}$, thereby allowing UCB practitioners to early stop and prevent over exploration. We also introduce a second strategy called UCB-DT(μ margin), which reduces the distance $|\hat{\mu}_i(t) - \hat{\mu}_j(t)|$ by a margin. This distance reduction encourages the policy to explore more among similar bandits and allows us to better handle noisy environments with high variance and low mean. We summarize their expected regrets in Table 2.

Experiment	UCB	UCB-DT(μ)	UCB-then-Commit	UCB-DT(μ margin)
B5	251.28	70.95	76.11	73.49
B20	939.99	383.82	436.36	415.68
B(0.02, 0.01)	249.17	21.6	58	243.04
B(0.9, 0.88)	119.91	19.19	105.7	110.01
N5	142.21	83.65	185.44	108.11
N20	1014.27	640.52	778.72	628.14

Table 2: Expected regret of UCB and UCB-DT on more distances in each experiment. γ is set to 0.02 in all experiments, m is set to 0.05.

It is interesting to see that UCB-then-Commit generally outperforms UCB by a large margin, which indicates the benefit of under exploration as in Sec. 4.3. UCB-DT(μ margin) appears to be more robust than UCB-DT(μ) in noisy environments by performing slightly better in N20, where the mean is much smaller than variance compared to other experiments.

C Formulation of Standard UCB

According to (Lattimore & Szepesvári, 2020, Chapter 7), UCB is derived from Hoeffding's inequality on sum of subgaussian variables. Let $X_1, X_2, \ldots, X_n$ be independent and L1-subgaussian random variables with zero mean and $\hat{\mu} = \sum_{t=1}^{n} X_t/n$, then

$$\mathbb{P}(\hat{\mu} \geq \varepsilon) \leq \exp\left(-n\varepsilon^2/2\right) \tag{16}$$

Replacing $\exp\left(-n\varepsilon^2/2\right)$ with δ then we get

$$\mathbb{P}\left(|\hat{\mu}| \geq \sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)}\right) \leq \delta \tag{17}$$

If we use $1/t$ for δ, we arrive at the formulation of UCB.

D Optimal Exploration and Exploration Bargain Point

Using our analysis in Sec. 4.3, we gain a deeper understanding as to why UCB is an "optimistic" policy. Furthermore, we can set the derivative of the lower bound in Eq. 10 on N_2 to 0 and solve for the optimal bound, which represents the optimal exploration-exploitation trade-off point. We calculate this result in Eq. 18 with (Inc.).

$$\left\{ N_2^* \to \frac{\mu_1^2 T - 2\mu_2\mu_1 T + \mu_2^2 T - 16 W_{c_1}\left(\frac{1}{2}e^{\frac{\mu_1^2 T}{16} - \frac{1}{8}\mu_2\mu_1 T + \frac{\mu_2^2 T}{16} + 1}\right) + 16}{2\left(\mu_2 - \mu_1\right)^2} \middle| c_1 \in \mathbb{Z} \right\} \tag{18}$$

where W denotes the Lambert W function (Weisstein, 2002)

In the case of $N_{bargain}$, the exact solution cannot be found analytically and we can only solve this numerically in Fig. 3. The difficulty lies in Eq. 19, which is transcendental and has no closed form expression for N_2 in this case.

$$0 = e^{-\frac{1}{16}\Delta_2^2 N_2}\left(2N_2 - T\right) - N_2 + \frac{8\log(T)}{\Delta_2^2} \tag{19}$$

the solution of this equation for N_2 is *Exploration Bargain Point*

There are interesting insights which can be derived from these equations. For example, based on Eq. 18, the optimal exploration point N_2^* may not be unique. It will be interesting to study the relationship between the solutions and determine which ones have practical relevance in future work.