Budgeted Multi-Armed Bandits with Asymmetric Confidence Intervals

Marco Heyden
marco.heyden@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Vadim Arzamasov
vadim.arzamasov@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Edouard Fouché
edouard.fouche@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

Klemens Böhm
klemens.boehm@kit.edu
Karlsruhe Institute of Technology
Karlsruhe, Germany

ABSTRACT

We study the stochastic Budgeted Multi-Armed Bandit (MAB) problem, where a player chooses from K arms with unknown expected rewards and costs. The goal is to maximize the total reward under a budget constraint. A player thus seeks to choose the arm with the highest reward-cost ratio as often as possible. Current approaches for this problem have several issues, which we illustrate. To overcome them, we propose a new upper confidence bound (UCB) sampling policy, ω-UCB, that uses asymmetric confidence intervals. These intervals scale with the distance between the sample mean and the bounds of a random variable, yielding a more accurate and tight estimation of the reward-cost ratio compared to our competitors. We show that our approach has sublinear instance-dependent regret in general and logarithmic regret for parameter $\rho \geq 1$, and that it outperforms existing policies consistently in synthetic and real settings.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**; *Online learning settings*; • **Information systems** → Decision support systems.

KEYWORDS

Budgeted multi-armed bandits, Multi-armed bandits, Decision making under uncertainty, Online decision making

ACM Reference Format:
Marco Heyden, Vadim Arzamasov, Edouard Fouché, and Klemens Böhm. 2024. Budgeted Multi-Armed Bandits with Asymmetric Confidence Intervals. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24), August 25–29, 2024, Barcelona, Spain.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3637528.3671833

1 INTRODUCTION

In the stochastic Multi-Armed Bandit (MAB) problem, a player repeatedly plays one of K arms and receives a corresponding random reward. The goal is to maximize the cumulative reward by playing the arm with the highest expected reward as often as possible. The expected rewards are initially unknown, so the player must balance trying arms to learn their expected rewards (exploration) versus using the current information to play arms with known high expected rewards (exploitation).

In the stochastic Budgeted MAB problem [16], a player must consider not only the potential rewards but also the associated random costs for each arm. The player chooses arms until the available budget is exhausted. Budgeted MABs model real-world situations such as the selection of a cloud service provider [1], energy-efficient task selection for battery-powered embedded devices [17], bid optimization [6], or optimizing advertising on social media.

Example 1 (Social media advertising)**.** Consider a retail company that wants to advertise products on a social network platform. The retail company provides to the platform an advertisement campaign consisting of multiple ads, as well as an advertisement budget. Each time a user interacts with an ad (an arm), the platform charges the retail company (a cost). Within the given budget, the retailer wants to find the ads which maximize the likelihood of a subsequent purchase (a reward). Both the reward and the cost are random variables since they depend on the actions of users and the competition from other advertisers. A Budgeted MAB algorithm can help find the most promising ads in real time.

A variety of policies has been proposed to address the Budgeted MAB problem. Section 3 provides a summary. Some policies model the problem as Bandits with Knapsacks, which are able to take the budget limit into account [4, 17]. Others extend ideas from traditional multi-armed bandit algorithms (in which the costs of arms are assumed to be constant), including Thompson Sampling [23] and Upper Confidence Bound (UCB) sampling [24]. Several studies indicate that the latter – and in particular UCB-sampling – perform well in practice [18, 19, 22, 24, 25].

UCB-sampling policies continuously update an upper bound of the ratio of the expected rewards and costs of each arm, and play the arm with the highest upper bound. We distinguish between three types: Some policies [8, 18, 19, 24] compute the bound from the ratio of the sample average reward and average cost plus some uncertainty-related term (cf. Eq. (1), left). We call this type "united" (u). Other policies [4, 24, 25] divide the reward's upper confidence bound by the cost's lower confidence bound (LCB) (cf. Eq. (1), right).

KDD '24, August 25–29, 2024, Barcelona, Spain

ACM ISBN 979-8-4007-0490-1/24/08
https://doi.org/10.1145/3637528.3671833

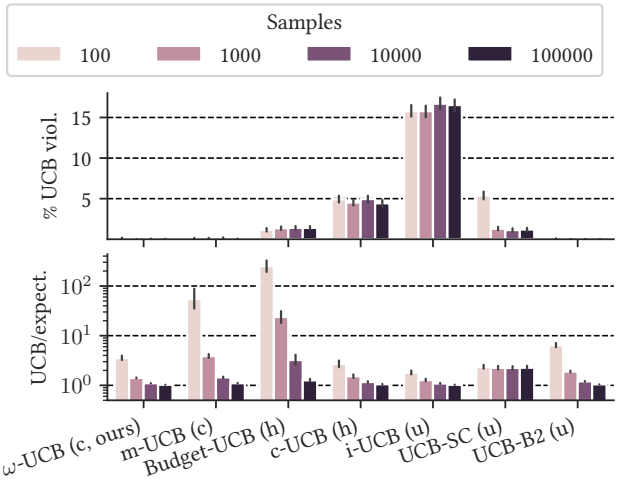


Figure 1: Issues of existing work

We refer to this type as "composite" (c). There also are "hybrid" (h) policies [22, 24] that combine the united and composite types.

$$UCB_u = \frac{average\ reward}{average\ cost} + uncertainty$$
$$UCB_c = \frac{average\ reward + uncertainty}{average\ cost - uncertainty} \quad (1)$$

However, most of the current policies have at least one of the following issues: (i1) **Over-optimism:** The policy often computes UCB that are too tight. (i2) **Over-pessimism:** The policy often computes UCB that are too loose. (i3) **Invalid values:** Negative or undefined UCB occur if the cost's lower confidence bound in Eq. (1) becomes negative or zero. The latter can happen, for instance, when computing the lower confidence bound of an arm's expected cost with Hoeffding's inequality [22, 24].

To illustrate (i1) and (i2), we randomly parameterized 10 000 Bernoulli reward and cost distributions and sampled from them. We used these samples to compute 99% confidence intervals of the reward-cost ratio using several state-of-the-art UCB-sampling policies. The upper plot of Figure 1 shows the share of cases when the expected reward-cost ratio exceeds (i.e., violates) its UCB. Values above 1% indicate overly tight bounds. The lower plot shows the UCB of the reward-cost ratio divided by its expectation, with higher values indicating looser bounds. Existing united policies (u) tend to suffer from issue (i1), hybrid approaches suffer from either of both issues, and issue (i2) mainly affects composite approaches. An exception is UCB-B2 [8], who's UCB is both tight and reliable (although not as tight as that of ω-UCB).

To address (i1) and (i2), some approaches provide a hyperparameter that allows to adjust the confidence interval manually [24]. However, setting such a hyperparameter is difficult since it depends on the unknown mean and variance of the reward and cost distributions. To address (i3), current approaches set the UCB of the ratio to infinity [24] or the cost LCB to a small positive value [22]. These heuristic solutions largely ignore the information already acquired about the cost distribution and tend to cause either (i1) or (i2).

Contributions. (1) We derive asymmetric confidence intervals for bounded random variables. These intervals have the same range as the random variable and scale with the distance between the sample mean and the boundaries. Our formula generalizes Wilson's score interval for binomial proportions [21] to arbitrary bounded random variables. (2) We introduce a policy called ω-UCB, which leverages these confidence intervals to address issues (i1)–(i3). We also propose an extension of ω-UCB, called ω^*-UCB, that uses the observed variances of the arms' rewards and cost to further tighten the UCB. (3) We prove that ω-UCB has sublinear regret in general and logarithmic regret for parameter $\rho \geq 1$. (4) We conduct experiments on typical settings found in the literature and real-world social network advertising data to compare the performance of ω-UCB and ω^*-UCB against state-of-the-art policies. Our results demonstrate that both policies have substantially lower regret than the competitors for both small and large budgets. (5) We share the code of our experiments.[1]

2 PROBLEM DEFINITION

We focus on a stochastic setting with K arms. Each arm k has continuous or discrete reward and cost distributions with unknown expected values $\mu_k^r \in [0, 1)$ and $\mu_k^c \in (0, 1]$, respectively. Assume without loss of generality that arm $k = 1$ has the highest ratio μ_k^r/μ_k^c among all arms. At time t a player chooses an arm $k_t \in \{1, \ldots, K\}$ and observes the reward $r_t \in [0, 1]$ and the cost $c_t \in [0, 1]$. We assume that the arms are independent and that rewards and costs observed at different time steps are independent and identically distributed (iid). This is consistent with previous work [18, 19, 22–24]. We do not make any assumptions about the correlation between rewards and costs of the same arm. The game ends after T_B plays that exhaust the available budget B.

Let $\mathbb{1}_k(k_t)$ be the indicator function: $\mathbb{1}_k(k_t) = 1$ iff $k_t = k$, else $\mathbb{1}_k(k_t) = 0$. The number of plays, and the sample average of rewards and costs of arm k at time T are:

$$n_k(T) = \sum_{t=1}^{T} \mathbb{1}_k(k_t)$$
$$\hat{\mu}_k^r(T) = \frac{1}{n_k(T)} \sum_{t=1}^{T} \mathbb{1}_k(k_t) r_t$$
$$\hat{\mu}_k^c(T) = \frac{1}{n_k(T)} \sum_{t=1}^{T} \mathbb{1}_k(k_t) c_t$$

The goal of the player is to minimize the pseudo-regret compared to the cumulative reward R^* of an optimal policy, given by $R^* - \mathbb{E}\sum_{t=1}^{T_B} r_t$. Finding the optimal policy in Budgeted MABs is known to be np-hard, due to the "knapsack problem" [16]. However, always choosing arm 1 leads to a suboptimality of at most $2\mu_1^r/\mu_1^c$, negligible for not too small budgets [23]. Thus, previous work [18, 19, 22–24], as well as our own approach, aim to minimize regret relative to a

[1] https://github.com/heymarco/OmegaUCB

policy that always selects arm 1:

$$\text{Regret} = \sum_{i=1}^{K} \mu_k^c \Delta_k \mathbb{E}[n_k(T_B)], \quad \text{where } \Delta_k = \frac{\mu_1^r}{\mu_1^c} - \frac{\mu_k^r}{\mu_k^c}$$

3 RELATED WORK

There exists many different MAB-related settings and policies. We refer to [11] for an overview and focus on algorithms developed for the Budgeted MAB setting in this section.

Tran-Thanh et al. [16] introduced the Budgeted MAB problem and proposed an ε-first policy. Subsequent policies KUBE [17] and PD-BwK [4] propose to formulate the problem as *Bandits with Knapsacks*, where the size of the knapsack represents the available budget. Both policies require knowledge of B. This restricts their applicability when B is an unknown quantity. However, we argue that such a restriction is unnecessary, since the advantage of exploiting B becomes negligible for sufficiently large budgets (cf. Section 2). Another approach, UCB-BV1 [10], addresses the special case of discrete random costs. Later solutions [18, 19, 22–24] adapted concepts from traditional MABs, such as Upper Confidence Bound (UCB) [2] or Thompson sampling [14, 15] and can deal with continuous random costs and unknown budget. However, the one policy based on Thompson sampling, BTS [23], requires transforming continuous rewards and costs into Bernoulli-samples. As a result, the policy disregards information about the variance of rewards and costs, causing over-pessimism (i2) when the variance of the reward or cost distribution is small. MRCB [22] deals with the challenge of playing multiple arms in each time step; when playing only one arm at a time, the policy becomes m-UCB [24] that is similar to our policy. However, m-UCB relies on Hoeffding's inequality, which does not take the distance between a random variable's sample mean and boundaries into account. To see why this is problematic, consider the following example[2]:

Example 2 (m-UCB)**.** Assume two arms with $\mu_1^r = 0.8, \mu_1^c = 0.2$, $\mu_2^r = 0.1, \mu_2^c = 0.1$. Clearly, arm 1 should be preferred for a reasonably large budget. However, m-UCB is biased towards pulling arm 2. For instance, if $t = 10000$ and $n_1 = n_2 = 1000$, using m-UCB with $\alpha = 1$ would yield reward-cost UCB values of ≈ 2.95 for arm 1 and ≈ 48.6 for arm 2 due to the high influence of the denominator in Eq. (1) (rhs). m-UCB would hence pull arm 2. In comparison, ω-UCB would compute values of 5.5 and 2.1, and pull arm 1.

More recent algorithms are adaptations of exiting ones to specific scenarios. For example, Avadhanula et al. [3] develop a multi-platform algorithm for Bandits with Knapsacks, and Das et al. [9] extend BwK to the combinatorial setting in which the algorithm can pull one or more arms in each round. However, the authors assume that an arm's cost is a known, fixed quantity, while we address the challenge of dealing with cost distributions. [7] proposes a novel 'bandits with interruptions' framework in which a player can interrupt a taken action to limit the cost. [8] extends Budgeted MABs to handle unbounded cost and reward distributions and presents policies for various settings. Policy UCB-B2 is tailored for our setting of uncorrelated, bounded rewards and costs, and relies on an empirical version of Bernstein's inequality. However, using this bound does not resolve the bias illustrated in Example 2.

The above policies either have issues (i1)–(i3) [8, 18, 19, 22–24], are not designed for continuous random costs [10, 17, 23], require knowledge of B [4, 17], and/or have been shown to perform inferior to others [10, 16, 17]. Figure 2 provides a compilation of existing head-to-head empirical comparisons between various policies. Upwards pointing triangles indicate that the policy in the corresponding row outperformed the policy in the corresponding column in the respective paper, while downward pointing triangles indicate the opposite. Circles represent cases where both policies performed similarly, while horizontal lines indicate that the policies have not been compared. One sees that KUBE [17] outperforms ε-first [16], while UCB-BV1 [10] and BTS[22] outperform KUBE. UCB-BV1 is inferior to more recent policies [18, 19, 22, 23]. BTS [23], b-greedy [24], and {i, c, m}-UCB [24] outperform PD-BwK [4]. We will compare our policy to the best performing existing policies (BTS, Budget-UCB, {i, c, m}-UCB, b-greedy, and UCB-SC+) and to the most recent UCB-B2.

4 OUR POLICY

We now detail our policy ω-UCB and analyze it theoretically.

4.1 ω-UCB

Our policy ω-UCB, summarized in Alg. 1, starts by playing each arm once. At each subsequent time step t, the policy chooses the arm k_t with the highest upper confidence bound of the ratio of the expected reward μ_k^r to the expected cost μ_k^c. Let $\omega_{k+}^r(\alpha, t)$ denote the upper confidence bound of μ_k^r for a confidence level $1-\alpha$. Similarly, $\omega_{k-}^c(\alpha, t)$ is the lower confidence bound of μ_k^c. ω-UCB chooses k_t according to:

$$k_t = \arg\max_{k \in [K]} \Omega_k(\alpha, t), \text{ where } \Omega_k(\alpha, t) = \frac{\omega_{k+}^r(\alpha, t)}{\omega_{k-}^c(\alpha, t)} \tag{2}$$

Unlike other policies that rely on the same principle [4, 8, 22, 24, 25], ω-UCB calculates asymmetric confidence bounds that are shifted towards the center of the range of the random variable. This leads to tighter UCB for the reward-cost ratio especially when an arm's expected cost or the number of plays is low.

Theorem 1 (Asymmetric confidence interval for bounded random variables)**.** Let X be a random variable bounded in the interval $[m, M]$, with unknown expected value $\mu \in [m, M]$ and variance σ^2. Let z denote the number of standard deviations required to achieve $1-\alpha$ confidence in coverage of the standard normal distribution. Let $\hat{\mu}$ be the sample mean of n iid samples of X. Then,

$$\Pr[\mu \notin [\omega_-(\alpha), \omega_+(\alpha)]] \leq \alpha,$$

with

$$\omega_\pm(\alpha) = \frac{B}{2A} \pm \sqrt{\frac{B^2}{4A^2} - \frac{C}{A}}, \tag{3}$$

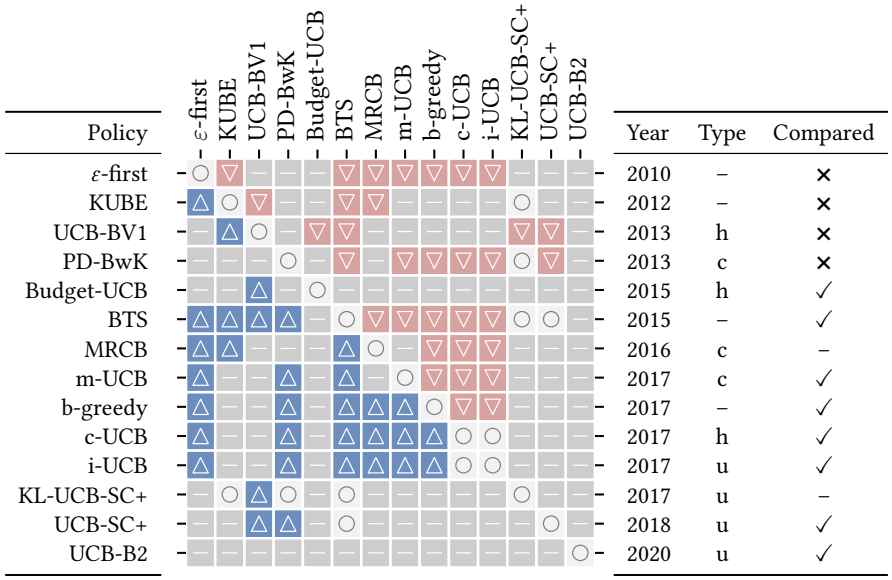
where

$$A = n + z^2\eta,\ B = 2n\hat{\mu} + z^2\eta(M+m),\ C = n\hat{\mu}^2 + z^2\eta Mm$$

and

$$\eta = \begin{cases} \frac{\sigma^2}{(M-\mu)(\mu-m)} & \text{if } \mu \in (m, M) \\ 1 & \text{if } \mu \in \{m, M\}. \end{cases}$$

[2] One can construct analogous examples for other symmetric bounds, such as Bernstein's inequality.



Policy	Year	Type	Compared
ε-first	2010	–	×
KUBE	2012	–	×
UCB-BV1	2013	h	×
PD-BwK	2013	c	×
Budget-UCB	2015	h	✓
BTS	2015	–	✓
MRCB	2016	c	–
m-UCB	2017	c	✓
b-greedy	2017	–	✓
c-UCB	2017	h	✓
i-UCB	2017	u	✓
KL-UCB-SC+	2017	u	–
UCB-SC+	2018	u	✓
UCB-B2	2020	u	✓

Figure 2: Empirical performance of different Budgeted MAB policies according to related work

Proof. Using the central limit theorem and Bhatia-Davis inequality, we follow similar steps as [21]. We first handle the case where $\mu \in (m, M)$. We then address the cases $\mu = m$ and $\mu = M$.

Case $\mu \in (m, M)$. The central limit theorem states that for a large enough sample size, $\hat{\mu}$ approximately follows a normal distribution with mean μ and variance σ^2/n. I.e.,

$$\hat{\mu} \sim \mathcal{N}\left(\mu, \sqrt{\frac{\sigma^2}{n}}\right) \iff \frac{\hat{\mu} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim \mathcal{N}(0, 1).$$

Therefore, $\hat{\mu}$ likely falls into an interval that is centered around μ and scaled by σ. The value z is the number of standard deviations such that $\hat{\mu}$ falls out of the corresponding confidence interval with a probability of α.

$$\Pr\left[\hat{\mu} \notin \left[\mu - \frac{\sigma}{\sqrt{n}}z, \mu + \frac{\sigma}{\sqrt{n}}z\right]\right] = \alpha \tag{4}$$

Next, we apply the Bhatia-Davis inequality [5] to express σ as a function of μ. It states that $\sigma^2 \leq (M - \mu)(\mu - m)$. Hence, there exists a factor $\eta \in [0, 1]$ such that

$$\sigma^2 = \eta(M - \mu)(\mu - m). \tag{5}$$

This gives us an expression for the interval bounds in Eq. (4) that is quadratic w.r.t. μ:

$$(\hat{\mu} - \mu)^2 = \frac{\sigma^2}{n}z^2 = \frac{\eta(M - \mu)(\mu - m)}{n}z^2 \tag{6}$$

Solving Eq. (6) for μ yields the endpoints $\omega_-(\alpha)$ and $\omega_+(\alpha)$ of our confidence interval:

$$\Pr[\mu \notin [\omega_-(\alpha), \omega_+(\alpha)]] = \alpha$$

with

$$\omega_-(\alpha), \omega_+(\alpha) = \frac{B}{2A} \pm \sqrt{\frac{B^2}{4A^2} - \frac{C}{A}}$$

Algorithm 1 ω-UCB

Require: K ▷ Number of arms
Require: B ▷ Available budget
Require: ρ ▷ CI scaling parameter, defaults to $1/4$
Require: $\vec{\eta}^r$ ▷ Reward variance parameters, defaults to $[1]^K$
Require: $\vec{\eta}^c$ ▷ Cost variance parameters, defaults to $[1]^K$
  $t \leftarrow 0$
  $\vec{n} \leftarrow [0]^K$
  while $B > 0$ **do**
    if $t < k$ **then** ▷ Play each arm once
      Play arm t
      Observe r_t, c_t and update $\hat{\mu}^r_k(t), \hat{\mu}^c_k(t)$
      $\vec{n}[t] = 1$
    else
      $z_\rho(t) = \sqrt{2\rho \log t}$ ▷ Corresponds to $\alpha(t) < 1 - \sqrt{1 - t^{-\rho}}$
      for all arms $k \in 1 \dots K$ **do**
        Compute $\Omega_k(\alpha, t)$ from Eq. (2) and Eq. (3) with $z_\rho(t)$, $\vec{\eta}^r[k]$, $\vec{\eta}^c[k]$, $\vec{n}[k]$, $\hat{\mu}^r_k(t)$, and $\hat{\mu}^c_k(t)$
      Find $k_t = \arg\max_k \Omega_k(\alpha, t)$
      Play arm k_t
      Observe r_t, c_t and update $\hat{\mu}^r_k(t), \hat{\mu}^c_k(t)$
      $\vec{n}[k_t] = \vec{n}[k_t] + 1$
    $B = B - c_t$
    $t = t + 1$

and

$$A = n + z^2\eta, \quad B = 2n\hat{\mu} + z^2\eta(M + m), \quad C = n\hat{\mu}^2 + z^2\eta Mm.$$

Cases $\mu = m$ *and* $\mu = M$. For $\mu = m$ (the case for $\mu = M$ is analogous), the probability that μ is not in the confidence interval $[\omega_-(\alpha), \omega_+(\alpha)]$ is zero, which is less than α. Additionally, we have

$\sigma^2 = (M-\mu)(\mu-m) = 0$, which implies that Eq. (5) holds for any choice of η. However, since μ is an unknown quantity, we can never be certain that $\mu = m$ based on some sample from X. In the worst-case scenario, X is a random variable that takes only extreme values, i.e., $X \in \{m, M\}$, with μ greater than and approximately equal to m. In this case, $\eta = 1$ by definition. Hence, we define $\eta = 1$ for $\mu \in m, M$. Combining the special case that $\mu \in \{m, M\}$ with the result from the previous paragraph gives Theorem 1. □

Illustration of asymmetry. The center of the confidence interval is a weighted average of the sample mean and the center of the range of the random variable. This leads to confidence intervals that are shifted towards the center of the range of the random variable. To see this, we can compare the distance between $\hat{\mu}$ and the interval center $B/2A$ to half the width of the confidence interval:

$$\text{Asymmetry} = \frac{\hat{\mu} - \frac{B}{2A}}{\sqrt{\frac{B^2}{4A^2} - \frac{C}{A}}} \in [0, 1]$$

For Bernoulli random variables, after some derivations and inserting the definitions of A, B, C as specified in Theorem 1, we obtain

$$\text{Asymmetry} = \frac{(2\hat{\mu}-1)^2 z^2}{4n\hat{\mu}(1-\hat{\mu}) + z^2} \tag{7}$$

Figure 3 plots the asymmetry measure for different values of n over $\hat{\mu}$ and $z = 3$. (1) For $\hat{\mu} = 1$ and $\hat{\mu} = 0$, the asymmetry takes on a maximum value of 1, while for $\hat{\mu} = 0.5$, asymmetry is 0. (2) For a given value of $\hat{\mu} \in (0, 1)$, asymmetry decreases with increasing sample size. (3) Related to this, we see that asymmetry is maximal for a given $\hat{\mu}$ for $n = 1$.

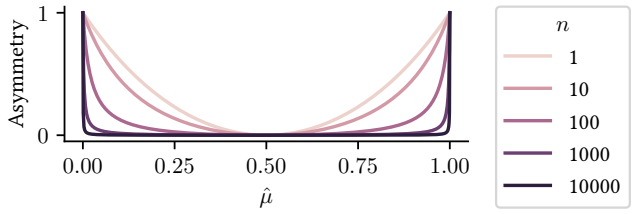


Figure 3: Asymmetry measure from Eq. (7) for Bernoulli rewards and costs for different n and $\hat{\mu}$ for $z = 3$.

Discussion. According to the Bhatia-Davis inequality [5], $0 \leq \sigma^2 \leq (M-\mu)(\mu-m)$, and hence $\eta \in [0, 1]$. For the special case of Bernoulli random variables, $\eta = 1$, $m = 0$, $M = 1$, and Theorem 1 recovers Wilson's original confidence interval for Binomial proportions [21]. However, this theorem is more flexible than Wilson's original method. It enables tighter confidence intervals for non-Bernoulli costs or rewards, by setting $\eta < 1$ when an estimate or an upper bound of the variance is available. Our experiments demonstrate that this flexibility leads to a significant performance improvement.

The following theorem defines an upper confidence bound $\Omega(\alpha)$ for the ratio of the expected values of two random variables. Combined with Theorem 1, it facilitates the computation of an arm's index according to Eq. (2).

Theorem 2 (UCB for ratio of expected values of two random variables)**.** Let R and C be two bounded random variables with expected values $\mu^r \geq 0$ and $\mu^c > 0$. Let $\omega^r_+(\alpha) \geq 0$ denote the upper confidence bound of R and $\omega^c_-(\alpha) > 0$ the lower confidence bound of C as given in Theorem 1. Let $\Omega(\alpha) = \omega^r_+(\alpha)/\omega^c_-(\alpha)$. Then,

$$\Pr\left[\frac{\mu^r}{\mu^c} > \Omega(\alpha)\right] \leq \alpha$$

Proof. Define events $E_1 = \mu^r > \omega^r_+(\alpha)$ and $E_2 = \mu^c < \omega^c_-(\alpha)$. A violation of the UCB of the reward-cost ratio requires that either E_1 or E_2 occurs, or that both events happen simultaneously. Therefore, by the union bound we have that $\Pr[\mu^r/\mu^c > \Omega_k] = \Pr\left[\mu^r/\mu^c > \omega^r_+(\alpha)/\omega^c_-(\alpha)\right] \leq \Pr[E_1] + \Pr[E_2]$. E_1 and E_2 both occur with probability $\leq \alpha/2$, hence $\Pr[\mu^r/\mu^c > \Omega_k] \leq \alpha$. □

A UCB-sampling policy that keeps parameter α constant leads to linear regret in the worst case. This is because such a policy will eventually stop exploring arms that may have high costs and low rewards in the beginning. To avoid this problem, ω-UCB decreases the value of α as the time t progresses, similarly to the UCB1-policy [2]. This adaptive approach helps to ensure continued exploration of arms and guarantees sub-linear regret. Theorem 3 introduces the scaling law and relates it to the confidence level.

Theorem 3 (Time-adaptive confidence interval)**.** For an arm k, let μ^r_k be its expected reward, μ^c_k its expected cost, and $\Omega_k(\alpha, t)$ the upper confidence bound for μ^r_k/μ^c_k, as in Eq. (2). For $\rho, t > 0$, and $\alpha(t) < 1 - \sqrt{1 - t^{-\rho}}$ it holds that

$$\Pr\left[\Omega_k(\alpha, t) \geq \frac{\mu^r_k}{\mu^c_k}\right] \geq 1 - \alpha(t),$$

that is, the upper confidence bound holds asymptotically almost surely.

Proof. We start from Theorem 1 and equate the confidence level $1-\alpha$ of the individual reward and cost distributions to the number of standard deviations z. This involves the cumulative density function (cdf) of the standard normal distribution. We then replace the cdf with an approximation that has a closed form solution for z. Our choice of z cancels out the exponential term in this approximation, similar to the UCB1-policy [2]. Last, we apply Theorem 2 to obtain the final result.

Step 1. We relate the confidence level $1 - \alpha(t)$ at time t to the cumulative density function of the standard normal distribution (erf abbreviates the error function).

$$1 - \frac{\alpha(t)}{2} = \frac{1}{2}\left(1 + \text{erf}\left(\frac{z}{\sqrt{2}}\right)\right)$$

Solving for $\alpha(t)$ yields:

$$\alpha(t) = 1 - \text{erf}\left(\frac{z}{\sqrt{2}}\right)$$

Step 2. We now replace the error function $\text{erf}\left(\frac{z}{\sqrt{2}}\right)$ in the equation above with a series expansion based on Bürmann's theorem [13, 20]; we summarize all but the first addend in a remainder term

$\gamma\left(\frac{z}{\sqrt{2}}\right) > 0$. This term has a maximum of $\gamma(0.71) \approx 0.0554$ and approaches 0 for larger z:

$$\alpha(t) = 1 - \left(\sqrt{1 - \exp\left(-\frac{z^2}{2}\right)} + \gamma\left(\frac{z}{\sqrt{2}}\right)\right)$$

Omitting the γ-term gives an upper bound for $\alpha(t)$:

$$\alpha(t) < 1 - \sqrt{1 - \exp\left(-\frac{z^2}{2}\right)} \tag{8}$$

Step 3. Next, we choose z as a function of $\log t$ and $\rho > 0$, $z_\rho(t) = \sqrt{2\rho \log t}$. This results in a time-increasing confidence level $1-\alpha(t)$:

$$\Pr[\mu \notin [\omega_-(\alpha(t)), \omega_+(\alpha(t))]] \leq \alpha(t) \text{ with } \alpha(t) < 1 - \sqrt{1 - t^{-\rho}}$$

Step 4. Applying Theorem 2 gives

$$\Pr\left[\frac{\mu_k^r}{\mu_k^c} > \Omega_k(\alpha, t)\right] \leq \alpha(t) \text{ with } \alpha(t) < 1 - \sqrt{1 - t^{-\rho}}.$$

The complementary event, $\Omega_k(\alpha, t) \geq \mu_k^r/\mu_k^c$, holds with a probability of at least $1 - \alpha(t)$,

$$\Pr\left[\Omega_k(\alpha, t) \geq \frac{\mu_k^r}{\mu_k^c}\right] \geq 1 - \alpha(t) \text{ with } \alpha(t) < 1 - \sqrt{1 - t^{-\rho}},$$

which is the result given in Theorem 3.

□

With $\alpha(t) < 1 - \sqrt{1 - t^{-\rho}}$, the confidence level $1 - \alpha$ approaches 1 as time t goes to infinity. This encourages exploration of arms that are played less frequently. Moreover, it establishes a logarithmic dependence between z in Theorem 1 and t, i.e., $z_\rho(t) = \sqrt{2\rho \log t}$, which will be useful in our regret analysis. The next section analyzes the worst-case regret of ω-UCB. To simplify notation, we abbreviate $\Omega(\alpha, t)$ as $\Omega(t)$, $\omega_{k-}^c(\alpha, t)$ as $\omega_{k-}^c(t)$, and $\omega_{k+}^r(\alpha, t)$ as $\omega_{k+}^r(t)$ in our regret analysis.

4.2 Regret Analysis

We first bound the expected number of suboptimal plays $\mathbb{E}[n_k(\tau)]$ before some time step τ. We use the result to derive the regret bound of ω-UCB (cf. Theorem 5). Due to space restrictions, we omit longer proofs here but provide them in Appendix B.

Theorem 4 (Number of suboptimal plays)**.** For ω-UCB, the expected number of plays of a suboptimal arm $k > 1$ before time step τ, $\mathbb{E}[n_k(\tau)]$, is upper-bounded by

$$\mathbb{E}[n_k(\tau)] \leq 1 + n_k^*(\tau) + \xi(\tau, \rho), \tag{9}$$

where

$$\xi(\tau, \rho) = (\tau - K)\left(2 - \sqrt{1 - \tau^{-\rho}}\right) - \sum_{\text{t=K+1}}^{\tau} \sqrt{1 - t^{-\rho}},$$

$$n_k^*(\tau) = \frac{8\rho \log \tau}{\delta_k^2} \max\left\{\frac{\eta_k^r \mu_k^r}{1 - \mu_k^r}, \frac{\eta_k^c (1 - \mu_k^c)}{\mu_k^c}\right\},$$

$\delta_k = \Delta_k/(\Delta_k + \frac{1}{\mu_k^c})$, and K and Δ_k are defined as before, cf. Section 2.

The 1 in Eq. (9) represents the very first pull of arm k. We interpret the term $n_k^*(\tau)$ as a minimum number of plays that leads to a "sufficiently small" deviation between μ_k^r/μ_k^c and $\hat{\mu}_k^r(t)/\hat{\mu}_k^c(t)$. This number is logarithmic w.r.t. τ. The term $\xi(\tau, \rho)$ represents those plays that occur despite the arms' sufficiently small deviation between the true reward-cost ratio and the ratio of the sample means.

For each suboptimal play, a policy suffers a greater-than-zero regret in expectation. Hence, the number of suboptimal plays is closely linked to a policy's regret. Lemma 4 in [24] establishes this connection for Budgeted MABs. In combination with Theorem 4, this lemma thus yields the worst-case regret for ω-UCB.

Theorem 5 (Finite-budget instance-dependent regret)**.** Define $\Delta_k, n_k^*(\tau_B)$ and $\xi(\tau_B, \rho)$ as before, and $\tau_B = \left\lfloor 2B/\min_{k\in[K]} \mu_k^c \right\rfloor$. For any $\rho > 0$, ω-UCB suffers instance-dependent regret of

$$\begin{aligned}\text{Regret} \leq &\sum_{k=2}^{K} \Delta_k \left(1 + n_k^*(\tau_B) + \xi(\tau_B, \rho)\right) \\ &+ \mathcal{X}(B) \sum_{k=2}^{K} \Delta_k + \frac{2\mu_1^r}{\mu_1^c},\end{aligned} \tag{10}$$

where $\mathcal{X}(B)$ is in $O\left((B/\mu_{min}^c)e^{-0.5B\mu_{min}^c}\right)$.

Proof. Lemma 4 of [24] provides a policy-independent regret expression for Budgeted MAB policies:

$$\text{Regret} \leq \sum_{k=2}^{K} \Delta_k \mathbb{E}[n_k(\tau_B)] + \mathcal{X}(B) \sum_{k=2}^{K} \Delta_k + \frac{2\mu_1^r}{\mu_1^c} \tag{11}$$

where

$$\tau_B = \left\lfloor \frac{2B}{\min_{k\in[K]} \mu_k^c} \right\rfloor \text{ and } \mathcal{X}(B) \text{ is in } O\left((B/\mu_{min}^c)e^{-0.5B\mu_{min}^c}\right).$$

Substituting $\mathbb{E}[n_k(\tau_B)]$ in Eq. (11) with the result from Theorem 4 completes the proof. □

Theorem 5 and our definition of $n_k^*(\tau)$ show that the regret of ω-UCB decreases for small η_k^r and η_c^k. I.e., when the reward and cost distributions have small variance compared to a Bernoulli variable with the same expected value.

The term $\xi(\tau, \rho)$ decreases, while $n_k^*(\tau)$ increases with ρ. Further derivations show that for increasingly large budgets, $\xi(\tau, \rho)$ converges for $\rho > 1$, grows logarithmic for $\rho = 1$, and superlogarithmic (on the order of $O(B^{1-\rho})$) for $\rho < 1$; see Appendix B.4 for the details. This results in the following asymptotic behavior:

Theorem 6 (Asymptotic regret)**.** The regret of ω-UCB is in

$$O\left(B^{1-\rho}\right) \text{ for } 0 < \rho < 1, \quad \text{and in} \quad O(\log B) \text{ for } \rho \geq 1.$$

5 EXPERIMENTAL SETUP

This section presents the experimental setup used to evaluate the policies. We introduce the MAB settings, followed by the configurations of ω-UCB and its competitors. We conducted the experiments on an Ubuntu 20.04 server with x86-architecture, using 32 cores, each running at 2.0 GHz, and 128 GB of RAM.

5.1 Budgeted MAB Settings

We use MAB settings based on synthetic and real data, which we describe separately. Each setting comprises a specific combination of reward and cost distributions, and the number of arms K. See Table 1 for a summary.

Table 1: Evaluation settings

Type	Distr.	Params	K	Used in	Id
Synth.	Bernoulli	$\mathcal{U}(0,1)$	10	[23, 24]	S-Br-10
			50	[24]	S-Br-50
			100	[22, 23]	S-Br-100
	General. Bernoulli	$\mathcal{U}(0,1)$	10	[23, 25]	S-GBr-10
			50	[25]	S-GBr-50
			100	[23]	S-GBr-100
	Beta	$\mathcal{U}(0,5)$	10	[24, 25]	S-Bt-10
			50	[24, 25]	S-Bt-50
			100	[22]	S-Bt-100
Face-book	Bernoulli	given	[2, 97]	–	FB-Br
	Beta	random	[2, 97]	–	FB-Bt

Synthetic Data. Previous studies on Budgeted Multi-Armed Bandits (MABs) use synthetic settings with rewards and costs drawn from discrete (Bernoulli or Generalized Bernoulli with possible outcomes $\{0.0, 0.25, 0.5, 0.75, 1.0\}$) or continuous (Beta) distributions [18, 19, 22, 24]. These studies typically generate parameters randomly within a given range [22–25] and use 10 to 100 arms [19, 22–25]. We adopt this and set the parameter ranges to those used in related work.

Social-Media Advertisement Data. We also evaluate our policy in a social media advertisement scenario described in Example 1. We use real-world data from [12]. It contains information about different ads based on their target gender (female or male) and age category (30–34, 34–39, 40–44, 45–49), along with the number of displays and clicks, the total cost, and the number of purchases. We group the ads by target gender and age category, resulting in 19 "advertisement campaigns" (Budgeted MAB settings). Each campaign has between 2 and 93 ads (arms). We compute the expected rewards μ_k^r and costs μ_k^c of each ad as the average revenue per click and average cost per click, respectively. We model both discrete and continuous rewards and costs. For the discrete case, we sample rewards and costs from two Bernoulli distributions with expected values of μ_k^r and μ_k^c, respectively. In the continuous case, we use a Beta distribution and sample the distribution parameters from a uniform distribution with a range of (0, 5). We then adjust one of the parameters to ensure that the expected values of rewards and costs match μ_k^r and μ_k^c.

5.2 Budgeted MAB Policies

We test the performance of two variants of our policy: ω-UCB and ω^*-UCB. The ω-UCB variant uses a fixed value of $\eta = 1$ for rewards and costs. The ω^*-UCB uses $\eta_k^r = \eta_k^c = 1$ as default but estimates their values using sample variance and mean once arm k has been played sufficiently many times ($n_k(T) \geq 30$), $\hat{\eta}_k = \hat{\sigma}_k^2/((M - \hat{\mu}_k)(\hat{\mu}_k - m))$, where bars refer to sample estimates as before. We experiment with two values of the hyperparameter ρ: $\rho = 1$ and $\rho = 1/4$. The former is the minimum value for which we have proven logarithmic regret. The latter has performed well in our sensitivity study, as we will demonstrate in Section 6.2.

We compare the performance of our policy to several other state-of-the-art Budgeted MAB policies, including BTS, Budget-UCB, i-UCB, c-UCB, m-UCB, b-greedy, UCB-SC+, and UCB-B2. We set the hyperparameters for each competitor to the values recommended in their respective papers. Appendix A features details about the policies and their hyperparameters.

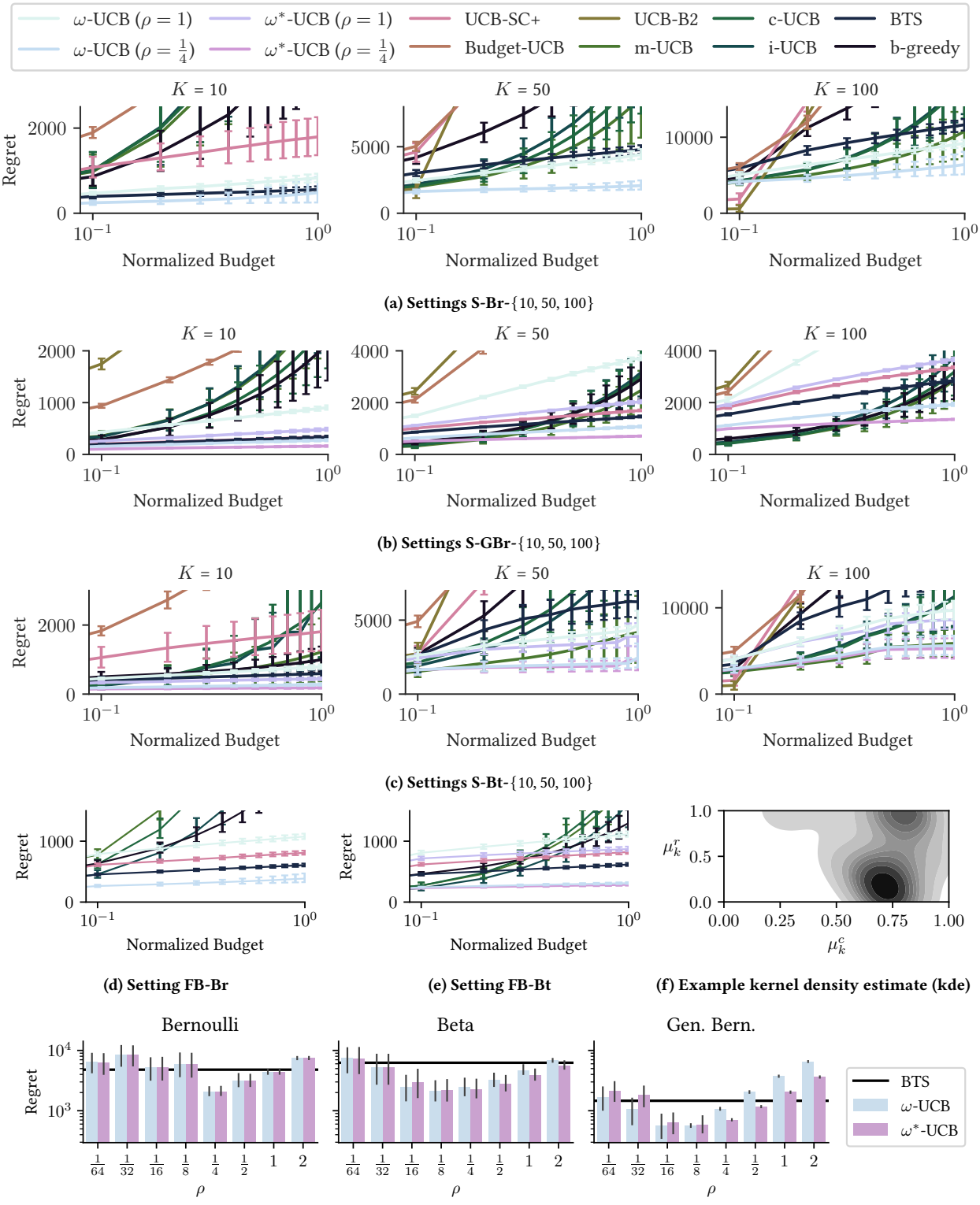
6 RESULTS

To observe the asymptotic behavior of the policies, we set the budget B to $1.5 \cdot 10^5$ times the minimum expected cost. We run each policy until the available budget is depleted and report the average results over 100 independent repetitions with the repetition index as the seed for the random number generator. Since we draw the expected cost μ_k^c and expected reward μ_k^r randomly in each repetition of the experiment, we normalize the budget in our graphs. Note that, in some instances, an approach may not appear in a graph if its regret exceeded the graph's upper y-axis limit.

6.1 Performance of Budgeted MAB Policies

6.1.1 Synthetic Bernoulli. Figure 4a shows the regret of the policies for Bernoulli-distributed rewards and costs with 95% confidence intervals. We do not present ω^*-UCB in this experiment since its results are almost identical to ω-UCB. ω-UCB and BTS achieve logarithmic regret and demonstrate better asymptotic behavior than other methods. Although {i,c,m}-UCB may outperform ω-UCB with $\rho = 1$ for small budgets, their regret grows rapidly as the budget increases, indicating poor asymptotic behavior. One surprising observation is that UCB-B2 is not competitive in our experiments despite its tight and accurate UCB (cf. Fig. 1). Our assumption is that this approach decreases the confidence level too fast, which we found can result in over-exploration. For $K = 50$ and $K = 100$, our policy has lower regret than BTS. BTS outperforms ω-UCB with $\rho = 1$ only on the 10-armed bandit. ω-UCB with $\rho = 1/4$ outperforms all other policies on small and large budgets and regardless of K. Comparing $\rho = 1/4$ with $\rho = 1$, one can see that the curve for $\rho = 1$ is linear (the x-axis is logarithmic), while for $\rho = 1/4$ it is convex. We conclude that $\rho = 1/4$ leads to smaller regret than $\rho = 1$ for not too large budgets but that $\rho = 1$ performs better asymptotically.

6.1.2 Synthetic Generalized Bernoulli and Synthetic Beta. Figure 4b shows the policies' regret for rewards and costs drawn from Generalized Bernoulli distributions and Figure 4c for Beta distributed rewards and costs. Besides ω-UCB, we also present the results for ω^*-UCB which estimates η_k^r and η_k^c from the sample variance of rewards and costs. ω-UCB and ω^*-UCB with $\rho = 1/4$ outperform their competitors, except for $K = 100$ in the Beta bandit where m-UCB performs comparable. We also notice that BTS is not competitive in this evaluation setting, likely because the policy cannot account for the (often very small) variance of the sampled Beta distributions. Our ω^*-UCB policy is advantageous in such cases.



(a) Settings S-Br-{10, 50, 100}

(b) Settings S-GBr-{10, 50, 100}

(c) Settings S-Bt-{10, 50, 100}

(d) Setting FB-Br

(e) Setting FB-Bt

(f) Example kernel density estimate (kde)

(g) Sensitivity study showing the regret for different choices of ρ

Figure 4: Evaluation results

6.1.3 Social-Media Advertisement Data. Figure 4d and Figure 4e show the results of our study on social-media advertisement data[3] [12]. Here, the default choice $\rho = 1/4$ outperforms all other competitors. $\rho = 1/4$ also outperforms $\rho = 1$ significantly, although both choices show good asymptotic behavior. The advertisement data settings seem to be easier for some competitors (BTS, KL-UCB-SC+) and harder for others (i,c,m-UCB). The likely cause is that the distribution of expected rewards and costs between arms is non-uniform: costs are biased towards 1, and rewards are biased towards the boundaries of $[0, 1]$, as the kernel density estimate (kde) plot for the MAB with $K = 33$ in Figure 4f illustrates exemplary. Last, we observe that ω^*-UCB has lower regret than ω-UCB, although the effect is not as strong as in the synthetic settings.

6.2 Sensitivity Study

We investigate the performance difference between ω-UCB and ω^*-UCB, as well as the sensitivity of our policy with respect to the hyperparameter ρ. The results based on our synthetic settings (cf. Table 1) for $K = 50$ are shown in Figure 4g. We omit the results for $K = 10$ and $K = 100$ as they look similar to $K = 50$. ω-UCB and ω^*-UCB perform best with $\rho = 1/4$ when rewards and costs follow Bernoulli distributions. Both policies achieve comparable performance in this case. It appears that estimating η (which is known to be 1 in Bernoulli bandits) does not result in a performance decrease of ω^*-UCB compared to ω-UCB. Also, even though $\rho = 1/8$ works well for ω-UCB when rewards and costs follow a generalized Bernoulli or Beta distribution, $\rho = 1/4$ remains a near-optimal choice for ω^*-UCB. Based on these results, we recommend using ω^*-UCB with $\rho = 1/4$ as a default.

7 CONCLUSIONS

We presented a new approach for Budgeted MABs called ω-UCB. It combines UCB sampling with asymmetric confidence intervals to address issues of existing approaches. Our interval generalizes Wilson's score interval to arbitrary bounded random variables. An extension of our approach, ω^*-UCB, tracks the variance of the reward and cost distributions on the fly to tighten the confidence intervals. This leads to even better performance when rewards or costs are continuous. Our analysis shows that ω-UCB achieves logarithmic regret for $\rho \geq 1$, while $\rho = 1/4$ performed best in our experiments. In the future, one could extend our approach to the successive elimination framework which repeatedly eliminates bad arms, or derive an instance-independent regret bound. One could also extend our approach to the non-stationary setting where the reward and cost distributions change over time. This is particularly relevant in scenarios like online advertising, where companies want to promote their products and services continuously.

Acknowledgements. This work was supported by the German Research Foundation (DFG) as part of the Research Training Group GRK 2153: Energy Status Data – Informatics Methods for its Collection, Analysis, and Exploitation and by the Baden-Württemberg Foundation via the Elite Program for Postdoctoral Researchers.

[3]Available at https://www.kaggle.com/datasets/madislemsalu/facebook-ad-campaign under a PDDL license.

REFERENCES

[1] Danilo Ardagna, Barbara Panicucci, and Mauro Passacantando. 2011. A game theoretic formulation of the service provisioning problem in cloud systems. In *WWW*. ACM, 177–186.
[2] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. *Mach. Learn.* 47, 2-3 (2002), 235–256. https://doi.org/10.1023/A:1013689704352
[3] Vashist Avadhanula, Riccardo Colini-Baldeschi, Stefano Leonardi, Karthik Abinav Sankararaman, and Okke Schrijvers. 2021. Stochastic bandits for multi-platform budget optimization in online advertising. In *WWW '21*, Jure Leskovec, Marko Grobelnik, Marc Najork, Jie Tang, and Leila Zia (Eds.). ACM / IW3C2, 2805–2817. https://doi.org/10.1145/3442381.3450074
[4] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. 2013. Bandits with Knapsacks. In *FOCS*. IEEE Computer Society, 207–216.
[5] Rajendra Bhatia and Chandler Davis. 2000. A better bound on the variance. *Amer. Math. Monthly* 107, 4 (2000), 353–357.
[6] Christian Borgs, Jennifer T. Chayes, Nicole Immorlica, Kamal Jain, Omid Etesami, and Mohammad Mahdian. 2007. Dynamics of bid optimization in online advertisement auctions. In *WWW*. ACM, 531–540.
[7] Semih Cayci, Atilla Eryilmaz, and R. Srikant. 2019. Learning to control renewal processes with bandit feedback. *Proc. ACM Meas. Anal. Comput. Syst.* 3, 2 (2019), 43:1–43:32. https://doi.org/10.1145/3341617.3326158
[8] Semih Cayci, Atilla Eryilmaz, and R. Srikant. 2020. Budget-constrained bandits over general cost and reward distributions. In *AISTATS (PMLR, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, 4388–4398. http://proceedings.mlr.press/v108/cayci20a.html
[9] Debojit Das, Shweta Jain, and Sujit Gujar. 2022. Budgeted Combinatorial Multi-Armed Bandits. In *AAMAS*. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 345–353.
[10] Wenkui Ding, Tao Qin, Xu-Dong Zhang, and Tie-Yan Liu. 2013. Multi-Armed Bandit with Budget Constraint and Variable Costs. In *AAAI*, Vol. 27. AAAI Press, 232–238. https://doi.org/10.1609/aaai.v27i1.8637
[11] Tor Lattimore and Csaba Szepesvári. 2020. *Bandit algorithms*. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781108571401
[12] Madis Lemsalu. 2017. Facebook ad campaign. https://www.kaggle.com/madislemsalu/facebook-ad-campaign howpublished: Kaggle (https://www.kaggle.com/madislemsalu/facebook-ad-campaign).
[13] HM Schöpf and PH Supancic. 2014. On Bürmann's theorem and its application to problems of linear and nonlinear heat transfer and diffusion. *The Mathematica Journal* 16 (2014), 1–44.
[14] William R. Thompson. 1933. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* 25, 3/4 (1933), 285–294. https://doi.org/10.2307/2332286
[15] William Robin Thompson. 1935. On the Theory of Apportionment. *American Journal of Mathematics* 57 (1935), 450.
[16] Long Tran-Thanh, Archie C. Chapman, Enrique Munoz de Cote, Alex Rogers, and Nicholas R. Jennings. 2010. Epsilon-First Policies for Budget-Limited Multi-Armed Bandits. In *AAAI*, Vol. 24. AAAI Press. https://doi.org/10.1609/aaai.v24i1.7758
[17] Long Tran-Thanh, Archie C. Chapman, Alex Rogers, and Nicholas R. Jennings. 2012. Knapsack Based Optimal Policies for Budget-Limited Multi-Armed Bandits. In *AAAI*, Vol. 26. AAAI Press, 1134–1140. https://doi.org/10.1609/aaai.v26i1.8279
[18] Ryo Watanabe, Junpei Komiyama, Atsuyoshi Nakamura, and Mineichi Kudo. 2017. KL-UCB-Based Policy for Budgeted Multi-Armed Bandits with Stochastic Action Costs. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* 100-A, 11 (2017), 2470–2486.
[19] Ryo Watanabe, Junpei Komiyama, Atsuyoshi Nakamura, and Mineichi Kudo. 2018. UCB-SC: A Fast Variant of KL-UCB-SC for Budgeted Multi-Armed Bandit Problem. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* 101-A, 3 (2018), 662–667.
[20] E.T. Whittaker and G.N. Watson. 2020. *A course of modern analysis; an introduction to the general theory of infinite processes and of analytic functions* (4 ed.). Cambridge University Press, Cambridge. pages: 208 section: 7.3.
[21] Edwin B. Wilson. 1927. Probable Inference, the Law of Succession, and Statistical Inference. *J. Amer. Statist. Assoc.* 22, 158 (1927), 209–212. https://doi.org/10.2307/2276774
[22] Yingce Xia, Wenkui Ding, Xu-Dong Zhang, Nenghai Yu, and Tao Qin. 2015. Budgeted Bandit Problems with Continuous Random Costs. In *ACML (JMLR Workshop and Conference Proceedings, Vol. 45)*. JMLR.org, 317–332.
[23] Yingce Xia, Haifang Li, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2015. Thompson Sampling for Budgeted Multi-Armed Bandits. In *IJCAI*. AAAI Press, 3960–3966.
[24] Yingce Xia, Tao Qin, Wenkui Ding, Haifang Li, Xudong Zhang, Nenghai Yu, and Tie-Yan Liu. 2017. Finite budget analysis of multi-armed bandit problems. *Neurocomputing* 258 (2017), 13–29. https://doi.org/10.1016/j.neucom.2016.12.079
[25] Yingce Xia, Tao Qin, Weidong Ma, Nenghai Yu, and Tie-Yan Liu. 2016. Budgeted Multi-Armed Bandits with Multiple Plays. In *IJCAI*. IJCAI/AAAI Press, 2210–2216.

A RELATED UCB APPROACHES

Table 2 summarizes our direct competitors and how they compute the arm selection indexes $\Omega_k(t)$.

B PROOFS

B.1 Proof of Theorem 4

The proof starts with a general expression for the number of plays of a suboptimal arm $k > 1$, where $\mathbb{1}\{\cdot\}$ denotes the indicator function.

$$\begin{aligned} n_k(\tau) &\leq 1 + \sum_{t=K+1}^{\tau} \mathbb{1}\left\{\Omega_k(t) \geq \Omega_j(t), \forall j \neq i\right\} \\ &\leq 1 + \sum_{t=K+1}^{\tau} \mathbb{1}\left\{\Omega_k(t) \geq \Omega_1(t)\right\} \end{aligned}$$

This is upper-bounded by

$$\begin{aligned} n_k(\tau) &\leq 1 + \sum_{t=K+1}^{\tau} \left[\mathbb{1}\left\{\Omega_k(t) \geq \Omega_1(t), \Omega_1(t) < \frac{\mu_1^r}{\mu_1^c}\right\} \right. \\ &\quad \left. + \mathbb{1}\left\{\Omega_k(t) \geq \Omega_1(t), \Omega_1(t) \geq \frac{\mu_1^r}{\mu_1^c}\right\} \right] \\ &\leq 1 + \sum_{t=K+1}^{\tau} \left[\mathbb{1}\left\{\Omega_1(t) < \frac{\mu_1^r}{\mu_1^c}\right\} + \mathbb{1}\left\{\Omega_k(t) \geq \frac{\mu_1^r}{\mu_1^c}\right\} \right] \end{aligned}$$

The expected number of plays $\mathbb{E}[n_k(\tau)]$ is given by the probabilities of the individual events:

$$\mathbb{E}[n_k(\tau)] \leq 1 + \sum_{t=K+1}^{\tau} \left[\underbrace{\Pr\left\{\Omega_1(t) < \frac{\mu_1^r}{\mu_1^c}\right\}}_{\Pr[A]} + \underbrace{\Pr\left\{\Omega_k(t) \geq \frac{\mu_1^r}{\mu_1^c}\right\}}_{\Pr[B]} \right] \tag{12}$$

We now evaluate the sum in the equation above.

Sum of $\Pr[A]$. We apply Theorem 2:

$$\sum_{t=K+1}^{\tau} \Pr[A] < \sum_{t=K+1}^{\tau} \left[1 - \sqrt{1 - t^{-\rho}}\right] = (\tau - K) - \sum_{t=K+1}^{\tau} \sqrt{1 - t^{-\rho}} \tag{13}$$

Sum of $\Pr[B]$. For this step, let us first introduce a helpful lemma: Lemma 1 bounds $\Pr\left[\Omega_k(t) \geq \mu_1^r/\mu_1^c\right]$ after a minimum number of plays $n_k^*(\tau)$, which grows logarithmic with τ.

Lemma 1. *Define* $\delta_k = \frac{\Delta_k}{\Delta_k + 1/\mu_k^c}$ *and* $n_k^*(\tau)$ *as follows:*

$$n_k^*(\tau) = \frac{8\rho \log \tau}{\delta_k^2} \max\left\{\frac{\eta_k^r \mu_k^r}{1 - \mu_k^r}, \frac{\eta_k^c (1 - \mu_k^c)}{\mu_k^c}\right\}$$

The following inequality holds whenever $n_k(t) \geq n_k^*(\tau)$*:*

$$\Pr\left[\Omega_k(t) \geq \frac{\mu_1^r}{\mu_1^c}\right] < 1 - \sqrt{1 - \tau^{-\rho}}$$

Appendix B.3 contains the proof of Lemma 1. See Appendix B.2 for the derivation of δ_k.

The lemma allows to decompose $\sum_{t=K+1}^{\tau} \Pr[B]$ into "initial plays" ($n_k(t) < n_k^*(\tau)$) and "later plays" ($n_k(t) \geq n_k^*(\tau)$):

$$\begin{aligned} \sum_{t=K+1}^{\tau} \Pr[B] &= \sum_{t=K+1}^{\tau} \Pr\left\{\Omega_k(t) \geq \frac{\mu_1^r}{\mu_1^c}\right\} \\ &= n_k^*(\tau) + \sum_{t=K+1}^{\tau} \Pr\left\{\Omega_k(t) \geq \frac{\mu_1^r}{\mu_1^c}, n_k(t) \geq n_k^*(\tau)\right\} \end{aligned}$$

We now apply Lemma 1 to evaluate the sum in above equation,

$$\begin{aligned} \sum_{t=K+1}^{\tau} \Pr[B] &\leq n_k^*(\tau) + \sum_{t=K+1}^{\tau} \left[1 - \sqrt{1 - \tau^{-\rho}}\right] \\ &= n_k^*(\tau) + (\tau - K)\left(1 - \sqrt{1 - \tau^{-\rho}}\right). \end{aligned} \tag{14}$$

Inserting the results of Eq. (13) and Eq. (14) in Eq. (12) yields a bound on $\mathbb{E}[n_k(\tau)]$:

$$\mathbb{E}[n_k(\tau)] \leq 1 + n_k^*(\tau) + \underbrace{(\tau - K)\left(2 - \sqrt{1 - \tau^{-\rho}}\right) - \sum_{t=K+1}^{\tau} \sqrt{1 - t^{-\rho}}}_{\xi(\tau,\rho)}$$

B.2 Derivation of Proportional δ-gap

Let δ_k be the proportional δ-gap of arm k. It measures how much one must increase μ_k^r and decrease μ_k^c in order to make arm k have the same reward-cost ratio as arm 1, or in other words, to bridge the suboptimality gap Δ_k between arm k and arm 1. The δ-gap is proportional to the possible increase in rewards (decrease in costs) without violating their range $[0, 1]$. We start our derivation with Eq. (15), which states this mathematically.

$$\frac{\mu_k^r + \delta_k(1 - \mu_k^r)}{\mu_k^c - \delta_k \mu_k^c} = \frac{\mu_1^r}{\mu_1^c} \tag{15}$$

Rearranging the equation yields

$$\frac{\mu_1^r}{\mu_1^c} - \frac{\mu_k^r}{\mu_k^c} = \delta_k \left(\frac{\mu_1^r}{\mu_1^c} - \frac{\mu_k^r}{\mu_k^c} + \frac{1}{\mu_k^c}\right).$$

Now, recall the definition of an arm's suboptimality, $\Delta_k = \mu_1^r/\mu_1^c - \mu_k^r/\mu_k^c$, and solve for δ_k:

$$\delta_k = \frac{\Delta_k}{\Delta_k + \frac{1}{\mu_k^c}}$$

B.3 Proof of Lemma 1

Recall Lemma 1:

Lemma 1. *Define* $\delta_k = \frac{\Delta_k}{\Delta_k + 1/\mu_k^c}$ *and* $n_k^*(\tau)$ *as follows:*

$$n_k^*(\tau) = \frac{8\rho \log \tau}{\delta_k^2} \max\left\{\frac{\eta_k^r \mu_k^r}{1 - \mu_k^r}, \frac{\eta_k^c (1 - \mu_k^c)}{\mu_k^c}\right\}$$

The following inequality holds whenever $n_k(t) \geq n_k^*(\tau)$*:*

$$\Pr\left[\Omega_k(t) \geq \frac{\mu_1^r}{\mu_1^c}\right] < 1 - \sqrt{1 - \tau^{-\rho}}$$

Table 2: Overview of our competitors

Policy	$\Omega_k(t)$	Comment
BTS [23]	$\frac{\text{sample from Beta}(\alpha_r(t),\beta_r(t))}{\text{sample from Beta}(\alpha_c(t),\beta_c(t))}$	$\alpha_r(t) = n_k(t)\hat{\mu}_k^r(t) + 1$ $\beta_r(t) = n_k(t) + 2 - \alpha_r(t)$ $\alpha_c(t) = n_k(t)\hat{\mu}_k^c(t) + 1$ $\beta_c(t) = n_k(t) + 2 - \alpha_c(t)$ Discretization of continuous values
m-UCB [24]	$\frac{\min\{\hat{\mu}_k^r(t)+\varepsilon_k(t),1\}}{\max\{\hat{\mu}_k^c(t)-\varepsilon_k(t),0\}}$	$\varepsilon_k(t) = \alpha\sqrt{\frac{\log(t-1)}{n_k(t)}}$ Recommendation: $\alpha = 2^{-4}$
c-UCB [24]	$\frac{\hat{\mu}_k^r(t)}{\hat{\mu}_k^c(t)} + \frac{\varepsilon_k(t)}{\hat{\mu}_k^c(t)}$	$\varepsilon_k(t) = \alpha\sqrt{\frac{\log(t-1)}{n_k(t)}}$ Recommendation: $\alpha = 2^{-3}$
i-UCB [24]	$\frac{\hat{\mu}_k^r(t)}{\hat{\mu}_k^c(t)} + \varepsilon_k(t)$	$\varepsilon_k(t) = \alpha\sqrt{\frac{\log(t-1)}{n_k(t)}}$ Recommendation: $\alpha = 2^{-2}$
Budget UCB [22]	$\frac{\hat{\mu}_k^r(t)}{\hat{\mu}_k^c(t)} + \frac{\varepsilon_k(t)}{\hat{\mu}_k^c(t)}\left(1 + \frac{\min\{\hat{\mu}_k^r(t)+\varepsilon_k(t),1\}}{\max\{\hat{\mu}_k^c(t)-\varepsilon_k(t),\lambda\}}\right)$	$\varepsilon_k(t) = \sqrt{\frac{\log(t-1)}{n_k(t)}}$ $\lambda > 0$: minimum cost
UCB-SC+ [19]	$\frac{\hat{\mu}_k^r(t)+\alpha_k(t)\hat{\mu}_k^c(t)}{\hat{\mu}_k^c(t)-\alpha_k(t)\hat{\mu}_k^r(t)}$, if $\hat{\mu}_k^c(t)^2 > \frac{\log\frac{t}{n_k(t)}}{2n_k(t)}$ ∞, else	$\alpha_k(t) = \sqrt{\frac{\log\frac{t}{n_k(t)}}{2\kappa n_k(t) - \log\frac{t}{n_k(t)}}}$ with $\kappa = \hat{\mu}_k^r(t)^2 + \hat{\mu}_k^c(t)^2$
UCB-B2 [8]	$1.4\frac{\epsilon_{k,t}+\hat{r}_{k,t}}{\hat{\mu}_k^c(t)} + \hat{r}_{k,t}$ if condition 7 in [8] for $\lambda = 1.28$ ∞, else	$\hat{r}_{k,t} = \frac{\max\{0,\hat{\mu}_k^r(t)\}}{\max\{\text{minimum cost},\hat{\mu}_k^c(t)\}}$ $\epsilon_{k,t} = \sqrt{\frac{2\hat{V}_{k,t}^r \log t^\alpha}{n_k(t)}} + \frac{3\log t^\alpha}{n_k(t)}$ $\eta_{k,t} = \sqrt{\frac{2\hat{V}_{k,t}^c \log t^\alpha}{n_k(t)}} + \frac{3\log t^\alpha}{n_k(t)}$ $\hat{V}_{k,t}^r, \hat{V}_{k,t}^c$: variance estimates $\alpha > 2$; small choices preferable

The proof is structured in four steps. First, we derive an expression for the maximum deviation between the observed sample mean and the unknown expected value of an arm's rewards and costs that we use later on. Second, we decompose the probability $\Pr\left[\Omega_k(t) \geq \mu_1^r/\mu_1^c\right]$. Third, we evaluate the decomposed probabilities for cases where $n_k(t)$ is sufficiently large, that is, $n_k(t) \geq n_k^*(\tau)$. Finally, we recombine the decomposed probabilities to obtain the final result.

Deviation between sample mean and expected value. Let $\hat{\mu}_k(t)$ be the sample mean and μ_k the expected value of arm k's rewards or costs at time t. To quantify the deviation between mean $\hat{\mu}_k(t)$ and μ_k we start from Eq. (6) (central limit theorem) and set $[m, M] = [0, 1]$ and $n = n_k(t)$):

$$(\hat{\mu}_k(t) - \mu_k)^2 \leq \frac{\eta_k \mu_k (1-\mu_k)}{n_k(t)} z^2 \tag{16}$$

The above inequality holds with the same probability as our confidence interval since it is the basis of the confidence interval derivation. This allows us to bound the deviation between sample mean and expected value, denoted $\varepsilon_k(t)$, for our choice of $z_\rho(t) = \sqrt{2\rho \log t}$:

$$\Pr[|\hat{\mu}_k(t) - \mu_k| > \varepsilon_k(t)] \leq \alpha(t) \tag{17}$$

with

$$\varepsilon_k(t) = \sqrt{\frac{2\eta_k \mu_k (1-\mu_k)\rho \log t}{n_k(t)}} \text{ and } \alpha(t) < 1 - \sqrt{1 - t^{-\rho}}$$

Whenever we refer to $\varepsilon_k(t)$ w.r.t. rewards or costs we use the notations $\varepsilon_k^r(t)$ and $\varepsilon_k^c(t)$.

Decomposition of $\Pr\left[\Omega_k(t) \geq \mu_1^r/\mu_1^c\right]$. Next, we decompose the probability that $\Omega_k(t) \geq \mu_1^r/\mu_1^c$. The decomposition is analogous to the one in the proof of Theorem 2 and thus omitted here:

$$\Pr\left[\Omega_k(t) \geq \frac{\mu_1^r}{\mu_1^c}\right] = \Pr\left[\frac{\omega_{k+}^r(t)}{\omega_{k-}^c(t)} \geq \frac{\mu_1^r}{\mu_1^c}\right]$$
$$= \Pr\left[\frac{\omega_{k+}^r(t)}{\omega_{k-}^c(t)} \geq \frac{\mu_k^r + (1-\mu_k^r)\delta_k}{\mu_k^c - \mu_k^c\delta_k}\right] \tag{18}$$

$$\leq \Pr\left[\omega_{k+}^r(t) \geq \mu_k^r + (1-\mu_k^r)\delta_k\right]$$
$$+ \Pr\left[\omega_{k-}^c(t) \leq \mu_k^c - \mu_k^c\delta_k\right] \tag{19}$$

For the next step, note that if $\mu_k^r \leq \omega_{k+}^r(t)$ (and $\mu_k^c \geq \omega_{k-}^c(t)$), we have that $\mu_k^r - \hat{\mu}_k^r(t) \leq \varepsilon_k^r(t)$ (and $\hat{\mu}_k^c(t) - \mu_k^c \leq \varepsilon_k^c(t)$). This allows us to rewrite the terms in Eq. (19) as follows:

$$\begin{aligned}\Pr\left[\omega_{k+}^r(t) \geq \mu_k^r + (1-\mu_k^r)\delta_k\right] &= \Pr\left[\mu_k^r + (1-\mu_k^r)\delta_k - \hat{\mu}_k^r(t) \leq \varepsilon_k^r(t)\right] \\ &= \Pr\left[\hat{\mu}_k^r(t) - \mu_k^r \geq (1-\mu_k^r)\delta_k - \varepsilon_k^r(t)\right] \quad (20)\end{aligned}$$

$$\begin{aligned}\Pr\left[\omega_{k-}^c(t) \leq \mu_k^c - \mu_k^c\delta_k\right] &= \Pr\left[\hat{\mu}_k^c(t) - (\mu_k^c - \mu_k^c\delta_k) \leq \varepsilon_k^c(t)\right] \\ &= \Pr\left[\mu_k^c - \hat{\mu}_k^c(t) \geq \mu_k^c\delta_k - \varepsilon_k^c(t)\right] \quad (21)\end{aligned}$$

In the next two paragraphs, we evaluate Eq. (20) (confidence bound of rewards) and Eq. (21) (confidence bound of costs). We combine both results afterwards.

Evaluation of Eq. (20) for $n_k(t) \geq n_k^{*,r}(\tau)$. We now consider the cases in which the number of times arm k was played is at least logarithmic w.r.t. τ, i.e., $n_k(t) \geq n_k^{*,r}(\tau)$ with

$$n_k^{*,r}(\tau) = \frac{2\rho\log\tau}{\delta_k^2(1-\kappa)^2}\frac{\eta_k^r\mu_k^r}{1-\mu_k^r}, \quad \text{for any } \kappa \in (0,1).$$

In those cases, $\varepsilon_k^r(t) \leq (1-\kappa)\delta_k(1-\mu_k^r)$. One can verify this by inserting $n_k^{*,r}(\tau)$ in the definition of $\varepsilon_k^r(t)$, cf. Eq. (17). This gives the following inequality for the right side of Eq. (20):

$$\Pr\left[\omega_{k+}^r(t) \geq \mu_k^r + (1-\mu_k^r)\delta_k\right] \leq \Pr\left[\hat{\mu}_k^r(t) - \mu_k^r \geq \kappa(1-\mu_k^r)\delta_k\right]$$

Last, we compute the number of standard deviations z^* that corresponds to a deviation between $\hat{\mu}_k^r(t)$ and μ_k^r of at maximum $\kappa(1-\mu_k^r)\delta_k$ based on Eq. (16). In particular, we solve the right-most inequality in the expression below:

$$\begin{aligned}(\hat{\mu}_k^r(t) - \mu_k^r)^2 \leq \frac{\eta_k^r\mu_k^r(1-\mu_k^r)}{n_k(t)}z^{*2} \\ \leq \frac{\eta_k^r\mu_k^r(1-\mu_k^r)}{n_k^{*,r}(\tau)}z^{*2} \leq (\kappa(1-\mu_k^r)\delta_k)^2\end{aligned}$$

With our choice of $n_k^{*,r}(\tau)$, this yields $z^* = \left(2\rho\log\tau\kappa^2(1-\kappa)^{-2}\right)^{\frac{1}{2}}$. Inserting z^* in Eq. (8) (upper bound for $\alpha(t)$) results in the following bound:

$$\Pr\left[\omega_{k+}^r(t) \geq \mu_k^r + (1-\mu_k^r)\delta_k\right] < \frac{1}{2}\left(1 - \sqrt{1 - \tau^{-\frac{\kappa^2\rho}{(1-\kappa)^2}}}\right) \quad (22)$$

Evaluation of Eq. (21) for $n_k(t) > n_k^{*,c}(\tau)$. Again, we consider the cases in which the number of times arm k was played is at least logarithmic in τ, i.e., $n_k(t) \geq n_k^{*,c}(\tau)$ with

$$n_k^{*,c}(\tau) = \frac{2\rho\log\tau}{\delta_k^2(1-\kappa)^2}\frac{\eta_k^c(1-\mu_k^c)}{\mu_k^c}, \quad \kappa \in (0,1).$$

In those cases, $\varepsilon_k^c(t) \leq (1-\kappa)\delta_k\mu_k^c$. Following analogous steps as in the previous paragraph yields Eq. (23):

$$\Pr\left[\omega_{k-}^c(t) \leq \mu_k^c - \mu_k^c\delta_k\right] < \frac{1}{2}\left(1 - \sqrt{1 - \tau^{-\frac{(1-\kappa)^2\rho}{\kappa^2}}}\right) \quad (23)$$

Obtaining the final result. With the results in Eq. (22) and Eq. (23) we can finally evaluate Eq. (18). A choice of $\kappa = 0.5$ and

$$n_k^*(\tau) = \frac{8\rho\log\tau}{\delta_k^2}\max\left\{\frac{\eta_k^r\mu_k^r}{1-\mu_k^r}, \frac{\eta_k^c(1-\mu_k^c)}{\mu_k^c}\right\}$$

yields the bound given in Lemma 1:

$$\Pr\left[\Omega_k(t) \geq \frac{\mu_1^r}{\mu_1^c}\right] < 1 - \sqrt{1-\tau^{-\rho}}, \quad n_k(t) \geq n_k^*(\tau)$$

B.4 Proof of Theorem 6

First note that the latter two terms and $n_k^*(\tau_B)$ in Eq. (10) are in $O(\log B)$ [24]. Next we show that $\xi(\tau_B, \rho)$ is in $O(\log B)$ for $\rho \geq 1$ and in $O(B^{1-\rho})$ for $0 < \rho < 1$.

We exploit two inequalities in our proof; the latter is based on the integral test for convergence and holds for continuous, positive, decreasing functions.

$$\sqrt{1-t^{-\rho}} \geq 1 - t^{-\rho}, \quad t \geq 1, \rho > 0 \quad (24)$$

$$\sum_{t=K+1}^{\tau_B} t^{-\rho} \leq (K+1)^{-\rho} + \int_{t=K+1}^{\tau_B} t^{-\rho}\,dt \quad (25)$$

We use Eq. (24) to obtain an integrable expression for the sum in $\xi(\tau_B, \rho)$. We replace the sum with an integral-based upper bound as in Eq. (25):

$$\begin{aligned}\xi(\tau_B,\rho) &= (\tau_B - K)\left(2 - \sqrt{1-\tau_B^{-\rho}}\right) - \sum_{t=K+1}^{\tau_B}\sqrt{1-t^{-\rho}} \\ &\leq (\tau_B - K)\left(2 - \sqrt{1-\tau_B^{-\rho}}\right) - \sum_{t=K+1}^{\tau_B} 1 - t^{-\rho} \\ &= (\tau_B - K)\left(1 - \sqrt{1-\tau_B^{-\rho}}\right) + \sum_{t=K+1}^{\tau_B} t^{-\rho} \\ &\leq (\tau_B - K)\left(1 - \sqrt{1-\tau_B^{-\rho}}\right) + (K+1)^{-\rho} + \int_{t=K+1}^{\tau_B} t^{-\rho}\,dt\end{aligned}$$

Next, we evaluate the integral for the cases $\rho = 1$ and $\rho \neq 1$.

Case 1: $\rho = 1$.

$$\begin{aligned}\xi(\tau_B,\rho) \leq (\tau_B - K)\left(1 - \sqrt{1-\tau_B^{-1}}\right) \\ + (K+1)^{-1} + \log\tau_B - \log(K+1)\end{aligned}$$

For $\rho = 1$, the first term in above equation converges, so $\xi(\tau_B, \rho = 1)$ is in $O(\log B)$. This implies that the overall regret of ω-UCB is in $O(\log B)$ for $\rho = 1$.

Case 2: $\rho \neq 1$.

$$\begin{aligned}\xi(\tau_B,\rho) \leq (\tau_B - K)\left(1 - \sqrt{1-\tau_B^{-\rho}}\right) \\ + (K+1)^{-\rho} + \frac{1}{1-\rho}\left(\tau_B^{1-\rho} - (K+1)^{1-\rho}\right)\end{aligned}$$

For $\rho > 1$, $\xi(\tau_B, \rho)$ converges and thus the overall regret is in $O(\log B)$. For $0 < \rho < 1$, one can show that $\xi(\tau_B, \rho)$ is in $O(B^{1-\rho})$. Hence, the regret of ω-UCB is in $O(B^{1-\rho})$ for $0 < \rho < 1$.

To summarize, the regret of our policy is in $O(B^{1-\rho})$ for $0 < \rho < 1$ and in $O(\log B)$ for $\rho \geq 1$.