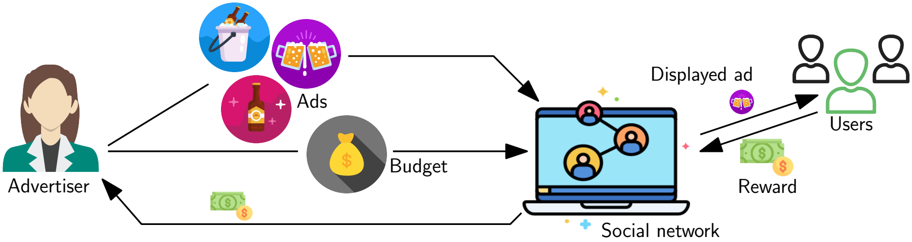
Budgeted Multi-Armed Bandits with Asymmetric Confidence Intervals

Marco Heyden*, Vadim Arzamasov, Edouard Fouché, Klemens Böhm

KDD ’24 | 27. August 2024

Example
Social media advertising

create marketing campaign $\rightarrow$ users interact with ads $\rightarrow$ pay advertising cost $\rightarrow$ receive reward



Budgeted Multi-armed Bandits

While budget B not empty:

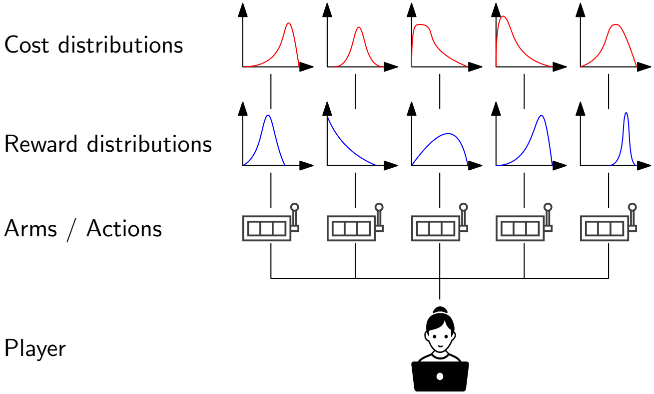
- play one of K arms
- observe reward and cost
- adjust arm selection strategy

Goal: maximize the total reward until the available budget runs out

Similar to traditional MABs, but:

- Budget B determines length of the game

$\rightarrow$ length of game is no longer deterministic



Notation

General:

- K: Number of arms
- B: Available budget
- T_B: Number of plays until budget is empty
- k: Some arm

For each arm:

- $n_k(T)$: Number of plays of arm k until time step T
- $\mu_k^r \in [0, 1), \mu_k^c \in (0, 1]$: Expected rewards and costs
- $\bar{\mu}_k^r(T), \bar{\mu}_k^c(T)$: sample average of rewards and costs at time step T

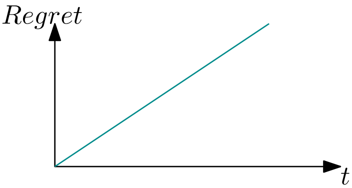
- Let **arm 1** be the arm with the **highest reward-cost ratio** μ_k^r / μ_k^c, w.l.o.g.
- Maximize reward = **minimize regret** of playing arms $k > 1$:

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

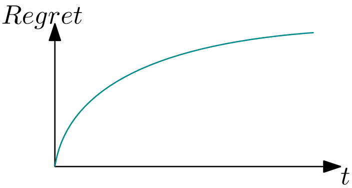
Regret
What is desirable?

- Playing any arm $k > 1$ leads to linear regret
- **Sublinear regret = the algorithm "learns" something** $\rightarrow$ this is what we want!

linear regret



sublinear regret



$$\text{Regret} = \sum_{i=1}^{K} \mu_k^c \Delta_k \mathbb{E}[n_k(T_B)]$$

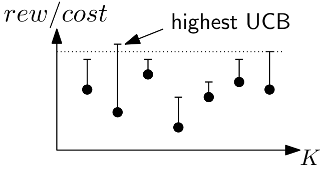
Related work
Approaches adopt ideas from traditional MAB policies

Thompson sampling (a.k.a. posterior sampling) [Xia+15b]

- Sample from posterior of arms' reward and cost distributions
- Play arm that maximizes ratio of the samples

UCB sampling (optimism in the face of uncertainty) [Xia+15a; Xia+16; Xia+17; Wat+17; Wat+18]

- Play arm with the highest UCB of reward-cost ratio
- Optimism encourages exploration

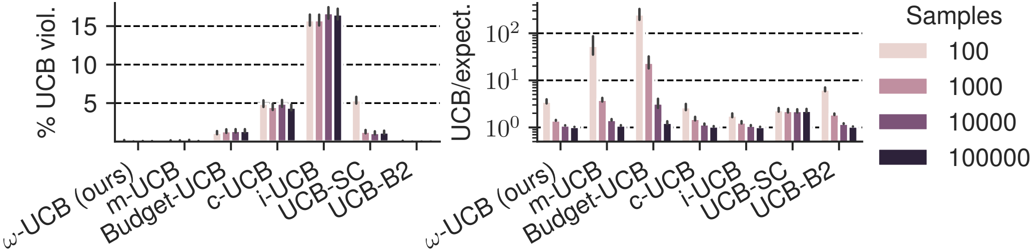


Related work
Existing UCB approaches have issues

The UCB for the reward-cost ratio should be

- as **accurate** as possible (UCB > expected value)
- as **tight** as possible

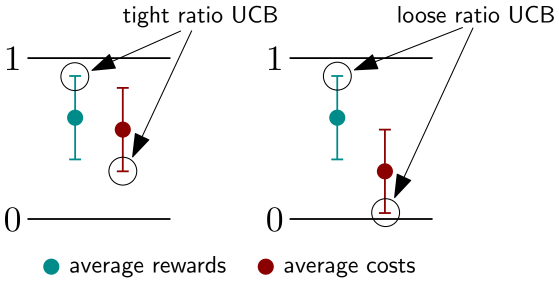
→ but this is not the case.



Foundation ○○○○○○●
The approach ○○○
Analysis & Evaluation ○○○
Summary ○

Our approach
Symmetric CIs lead to increased UCB for reward-cost ratio

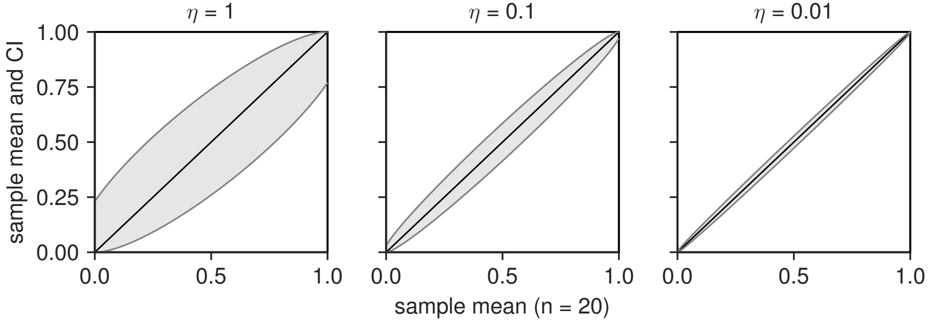
$$UCB = \frac{\textit{average reward} + \textit{uncertainty}}{\textit{average cost} - \textit{uncertainty}}$$



Our approach
Asymmetric confidence interval (illustration)

- Asymmetric CI $\Rightarrow$ higher LCB of cost $\Rightarrow$ tighter UCB of ratio
- η: variance parameter ($\eta = 1 \rightarrow$ Bernoulli random variable)
- Generalization of Wilson Score Interval [Wil27]

$$\eta = \frac{\sigma^2}{(1-\mu)\mu}$$



Our approach
ω-UCB and ω^*-UCB

While budget B not empty:

- **Play arm k_t:**

$$k_t = \underset{k \in [K]}{\arg\max}\, \Omega_k(\alpha, t), \qquad \text{where } \Omega_k(\alpha, t) = \frac{\bar{\mu}_k^r(t) + \text{asymmetric CI of rewards}}{\bar{\mu}_k^c(t) - \text{asymmetric CI of costs}}$$

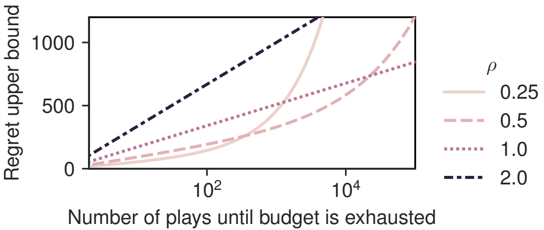
- **Observe reward r_t and cost c_t**
- **Update parameters for CI calculation**
  - Variant ω-UCB assumes maximum variance (e.g. Bernoulli random variable)
  - ω^*-UCB uses observed variance to tighten CI
- **Scale CI of all arms s. th. $\alpha(t) < 1 - \sqrt{1 - t^{-\rho}}$**
  - Ensures sufficient exploration of all arms
  - ρ: exploration parameter (high $\rho \rightarrow$ more exploration)

Theoretical analysis

Karlsruhe Institute of Technology

Proof structure (based on [Xia+17])

- Bound number of suboptimal plays $\mathbb{E}(n_k(\tau))$
  - up to time step τ
- Derive regret obtained until time step τ
- Choose τ_B that is larger than T_B with high probability
- Bound regret for "extra long" games where $T_B > \tau_B$
- Evaluate asymptotic behavior



Asymptotic regret: The regret of ω-UCB is

$$\text{Regret} \in \mathcal{O}\left(B^{1-\rho}\right), \text{ for } 0 < \rho < 1; \qquad \text{Regret} \in \mathcal{O}(\log B), \text{ for } \rho \geq 1$$

Evaluation
Setup

Competitors

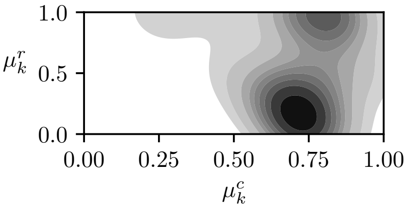
- 8 competitors (the strongest ones)
- 4 ω-UCB variants
  - best versions theoretically
  - best versions empirically

Synthetic MAB environments

- Bernoulli: rewards and costs follow Bernoulli distributions
- Generalized Bernoulli: rewards and costs sampled from $\{0, 0.25, 0.5, 0.75, 1\}$
- Beta: rewards and costs sampled from Beta distributions

Social media advertising

- Expected rewards and costs derived from real-world social media advertising campaigns [Lem17]
- Bernoulli and Beta distributed rewards and costs
- Below: KDE plot of a marketing campaign

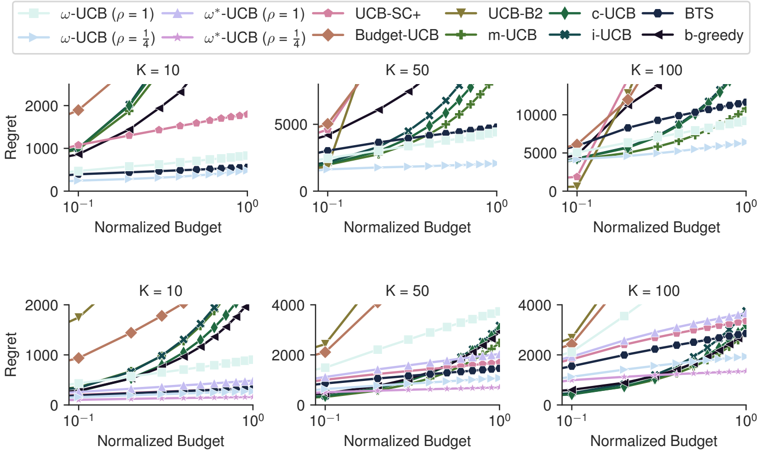


Evaluation
Results

Top: Bernoulli rewards / costs

Bottom: rewards / costs drawn from $\{0, 0.25, \ldots, 1\}$

- ω-UCB has lower regret than competitors
- ω^*-UCB performs even better on Beta bandits
- Straight line = logarithmic growth (x-axis is log-scaled)



Wrapping up

Summary

- We propose ω-UCB, an **upper confidence bound sampling** policy that uses **asymmetric confidence intervals**
- Asymmetric confidence intervals lead to tighter estimation of UCB for reward-cost ratio
- Desirable theoretical properties and empirical performance

In the paper

- Definition and derivation of asymmetric intervals
- In-depth analysis (finite budget) and proofs
- Pseudocode
- Additional experiments

Paper and code:

- doi.org/10.1145/3637528.3671833
- github.com/heymarco/OmegaUCB

Paper

GitHub

References I

[1] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. "Finite-time Analysis of the Multiarmed Bandit Problem". In: *Mach. Learn.* 47.2-3 (2002), pp. 235–256. DOI: https://doi.org/10.1023/A:1013689704352.

[2] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. "Bandits with Knapsacks". In: *FOCS*. IEEE Computer Society, 2013, pp. 207–216.

[3] Wenkui Ding et al. "Multi-Armed Bandit with Budget Constraint and Variable Costs". In: *AAAI*. Vol. 27. AAAI Press, 2013, pp. 232–238. DOI: 10.1609/aaai.v27i1.8637.

[4] Madis Lemsalu. *Facebook ad campaign*. howpublished: Kaggle (https://www.kaggle.com/madislemsalu/facebook-ad-campaign). 2017. URL: https://www.kaggle.com/madislemsalu/facebook-ad-campaign (visited on 01/05/2023).

[5] Long Tran-Thanh et al. "Epsilon-First Policies for Budget-Limited Multi-Armed Bandits". In: *AAAI*. Vol. 24. AAAI Press, 2010. DOI: 10.1609/aaai.v24i1.7758.

References II

[6] Long Tran-Thanh et al. “Knapsack Based Optimal Policies for Budget-Limited Multi-Armed Bandits”. In: *AAAI*. Vol. 26. AAAI Press, 2012, pp. 1134–1140. DOI: `https://doi.org/10.1609/aaai.v26i1.8279`.

[7] Ryo Watanabe et al. “KL-UCB-Based Policy for Budgeted Multi-Armed Bandits with Stochastic Action Costs”. In: *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* 100-A.11 (2017), pp. 2470–2486.

[8] Ryo Watanabe et al. “UCB-SC: A Fast Variant of KL-UCB-SC for Budgeted Multi-Armed Bandit Problem”. In: *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* 101-A.3 (2018), pp. 662–667.

[9] Edwin B. Wilson. “Probable Inference, the Law of Succession, and Statistical Inference”. In: *Journal of the American Statistical Association* 22.158 (1927), pp. 209–212. DOI: `https://doi.org/10.2307/2276774`.

[10] Yingce Xia et al. “Budgeted Bandit Problems with Continuous Random Costs”. In: *ACML*. Vol. 45. JMLR Workshop and Conference Proceedings. JMLR.org, 2015, pp. 317–332.

[11] Yingce Xia et al. “Budgeted Multi-Armed Bandits with Multiple Plays”. In: *IJCAI*. IJCAI/AAAI Press, 2016, pp. 2210–2216.

References III

[12] Yingce Xia et al. “Finite budget analysis of multi-armed bandit problems”. In: *Neurocomputing* 258 (2017), pp. 13–29. ISSN: 0925-2312. DOI: `https://doi.org/10.1016/j.neucom.2016.12.079`.

[13] Yingce Xia et al. “Thompson Sampling for Budgeted Multi-Armed Bandits”. In: *IJCAI*. AAAI Press, 2015, pp. 3960–3966.

Our approach
Asymmetric confidence interval (definition)

Theorem (Asymmetric confidence interval for bounded random variables)

Let X be a random variable that falls in the interval $[m, M]$ and has an 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. Denote the sample mean of n iid samples of X as $\bar{\mu}$. Then

$$\Pr[\mu \notin [\omega_-(\alpha), \omega_+(\alpha)]] \leq \alpha, \quad \textit{with}\ \omega_\pm(\alpha) = \frac{B}{2A} \pm \sqrt{\frac{B^2}{4A^2} - \frac{C}{A}},$$

where

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

$$\eta = \frac{\sigma^2}{(M - \mu)(\mu - m)}\ \textit{if}\ \mu \in (m, M), \quad \textit{and}\ \eta = 1\ \textit{if}\ \mu \in \{m, M\}.$$

Our approach
Increasing the upper confidence bound over time

Intuition:

- Confidence intervals increase over time
- Guarantees that initially “unlucky” arms will be explored again at some point in the future
- Inspired by the UCB1-policy for “traditional” MABs [ACF02]

Theorem (Time-adaptive confidence interval)

For an arm k, let μ_k^r be its expected reward, μ_k^c its expected cost, and $\Omega_k(\alpha, t)$ the upper confidence bound for μ_k^r/μ_k^c, as in Eq. 10. For $\rho, t > 0$, and $\alpha(t) < 1 - \sqrt{1 - t^{-\rho}}$ it holds that

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

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

References

Theoretical analysis
Proof idea for worst-case regret

Bound number of suboptimal plays $\mathbb{E}(n_k(\tau))$ up to time step τ

- Playing a suboptimal arm k leads to expected “incremental” regret of $\mu_k^c \Delta_k$

Derive regret obtained until time step τ

- Sum incremental regret over arms and time horizon

Find τ_B that is larger than T_B with high probability

- Bound regret for “extra long” games where $T_B > \tau_B$
  - Already done by [Xia+17]

Evaluate asymptotic behavior of regret

- Behavior of regret for $\tau_B \to \infty$

$$\text{Regret} = \sum_{i=1}^{K} \mu_k^c \Delta_k \mathbb{E}[n_k(T_B)]$$

Theoretical analysis
Results (I)

Karlsruhe Institute of Technology

Theorem (Number of suboptimal plays)

With 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:

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

where

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

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

and K and Δ_k are defined as before.

References
○○○●○○○

Theoretical analysis
Results (II)

Theorem (Worst-case regret)

Define $\tau_B = \left\lfloor 2B/\min_{k\in[K]} \mu_k^c \right\rfloor$ and Δ_k, $n_k^*(\tau_B)$, and $\xi(\tau_B, \rho)$ as before. For any $\rho > 0$, the regret of ω-UCB is upper-bounded by

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

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

Theorem (Asymptotic regret)

The regret of ω-UCB is

$$\text{Regret} \in \mathcal{O}\left(B^{1-\rho}\right), \ \text{for } 0 < \rho < 1; \qquad \text{Regret} \in \mathcal{O}(\log B), \ \text{for } \rho \geq 1$$

Evaluation
Competitors

- We compare our approach against existing approaches
- We exclude:
  - Poorly performing baselines
  - “Older” versions of more recent approaches

Policy	Ref.	Evaluated
ε-first	[Tra+10]	×
KUBE	[Tra+12]	×
UCB-BV1	[Din+13]	×
PD-BwK	[BKS13]	×
Budget-UCB	[Xia+15a]	✓
BTS	[Xia+15b]	✓
MRCB	[Xia+16]	(✓)
m-UCB	[Xia+17]	✓
b-greedy	[Xia+17]	✓
c-UCB	[Xia+17]	✓
i-UCB	[Xia+17]	✓
KL-UCB-SC+	[Wat+17]	(✓)
UCB-SC+	[Wat+18]	✓
ω-UCB	ours	✓

Evaluation
Budgeted MAB settings

Synthetic and real world Budgeted MAB settings

- Adopt synthetic evaluation settings from related work
- Use openly available social media advertising data [Lem17]

Type	Distribution	Parameters	K	Used in
Synthetic	**Bernoulli**	$\mathcal{U}(0,1)$	10	[Xia+15b; Xia+17]
			50	[Xia+17]
			100	[Xia+15a; Xia+15b]
	Generalized Bernoulli	$\mathcal{U}(0,1)$	10	[Xia+15b; Xia+16]
			50	[Xia+16]
			100	[Xia+15b]
	Beta	$\mathcal{U}(0,5)$	10	[Xia+17; Xia+16]
			50	[Xia+17; Xia+16]
			100	[Xia+15a]
Facebook	Bernoulli	given	$[2,97]$	–
	Beta	randomized	$[2,97]$	–