Karlsruhe Institute of Technology

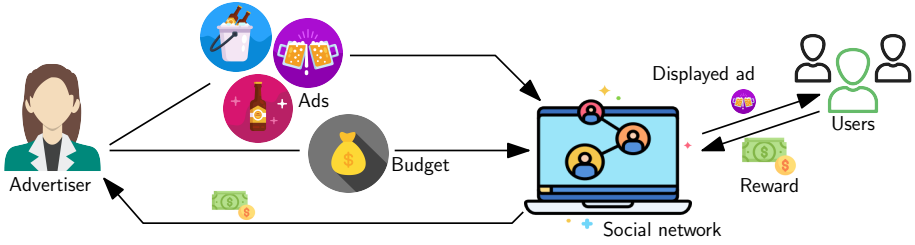
Budgeted Multi-Armed Bandits with Asymmetric Confidence Intervals

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

KDD '24 | 27. August 2024

Example
Social media advertising

create marketing campaign ! users interact with ads ! pay advertising cost ! receive reward



Budgeted Multi-armed Bandits

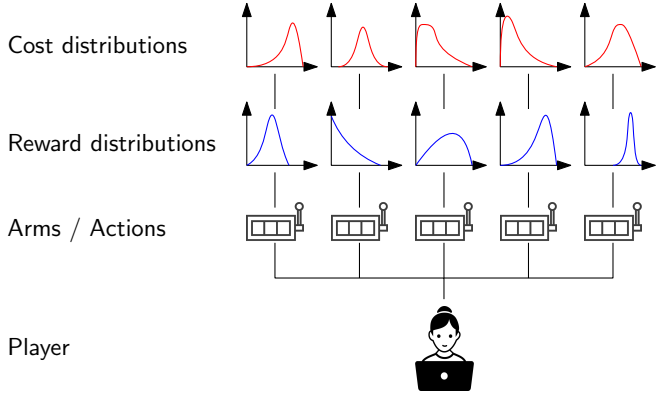
While budget B not empty:

- 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
- ! 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:

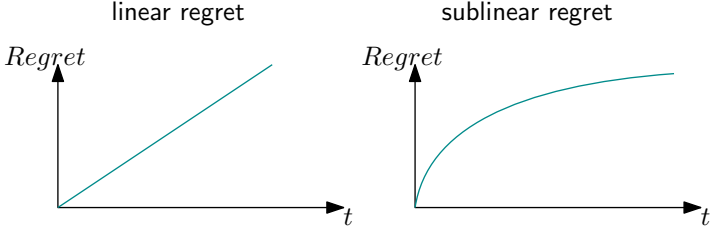
- $n_k(T)$: Number of plays of arm k until time step T
- $r_k \in [0;1]$; $c_k \in (0;1]$: Expected rewards and costs
- $r_k(T)$; $c_k(T)$: sample average of rewards and costs at time step T

- Let **arm 1** be the arm with the **highest reward-cost ratio** $\frac{r_1}{c_1} = \frac{r_k}{c_k}$, w.l.o.g.
- Maximize reward = **minimize regret** of playing arms $k > 1$:

$$\text{Regret} = \sum_{i=1}^{\cancel{K}} \frac{c_1}{c_k} \Delta_k \mathrm{E}[n_k(T_B)]; \qquad \text{where } \Delta_k = \frac{\frac{r_1}{c_1}}{} - \frac{\frac{r_k}{c_k}}{}$$

Regret
What is desirable?

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



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

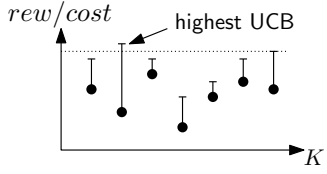
Approaches adopt ideas from traditional MAB policies

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

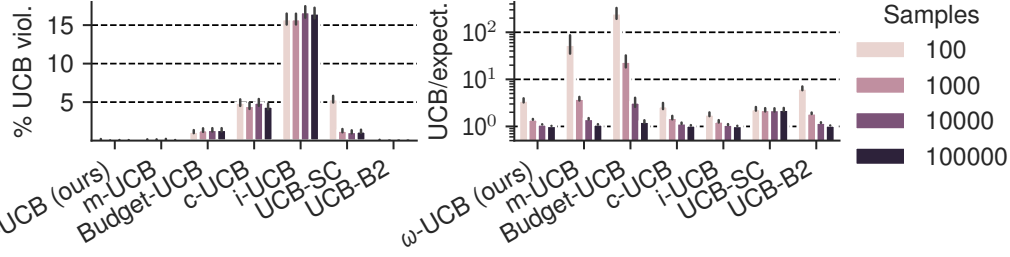


Existing UCB approaches have issues

The UCB for the reward-cost ratio should be

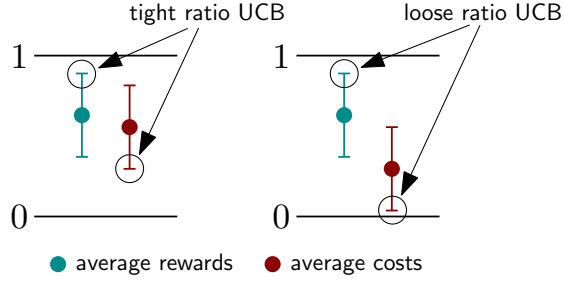
- as **accurate** as possible (UCB > expected value)
- as **tight** as possible

! but this is not the case.



Symmetric CIs lead to increased UCB for reward-cost ratio

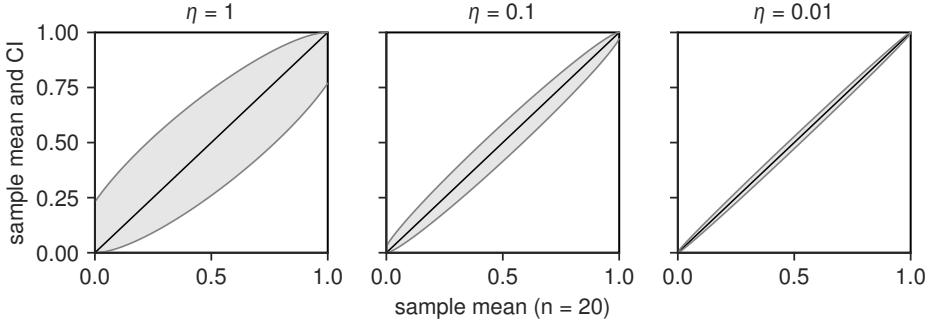
$$UCB = \frac{\textit{average reward} + \textit{uncertainty}}{\textit{average cost} \quad \textit{uncertainty}}$$



Our approach
Asymmetric confidence interval (illustration)

- Asymmetric CI) higher LCB of cost) tighter UCB of ratio
- : variance parameter (= 1 ! Bernoulli random variable)
- Generalization of Wilson Score Interval [Wil27]

$$= \frac{^{2}}{(1 \quad)}$$



Our approach
ω-UCB and ω -UCB

While budget B not empty:

- **Play arm k_t:**

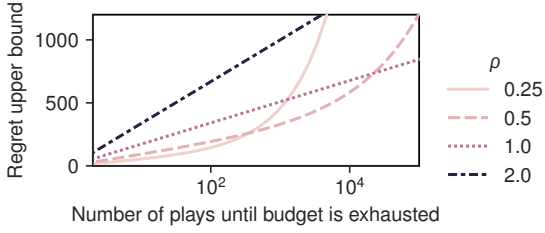
$$k_t = \arg\max_{k \in [K]} \quad _k(\ ; t); \qquad \text{where} \quad _k(\ ; t) = \frac{^r_k(t) + \text{asymmetric CI of rewards}}{^c_k(t) \quad \text{asymmetric CI of costs}}$$

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

Theoretical analysis

Proof structure (based on [Xia+17])

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



Asymptotic regret: The regret of $!$-UCB is

$$\text{Regret } 2\ O\ \ B^{1}\quad ;\ \text{for } 0 < \ < 1; \qquad \text{Regret } 2\ O(\log B);\ \text{for}\quad 1$$

Evaluation
Setup

Competitors

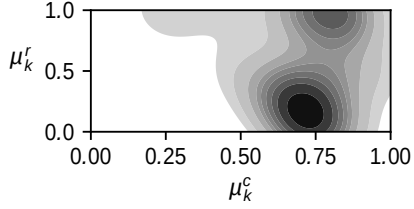
- 8 competitors (the strongest ones)
- 4 *!*-UCB variants
  - best versions theoretically
  - best versions empirically

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

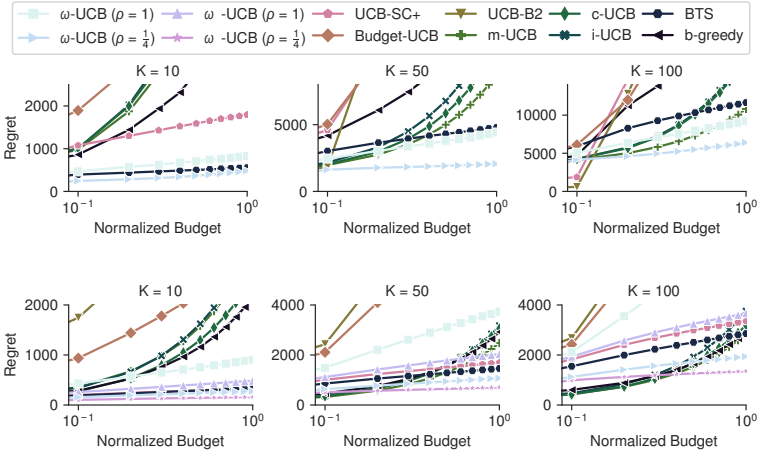


Evaluation
Results

Top: Bernoulli rewards / costs

Bottom: rewards / costs drawn from $f0; 0.25; : : : ; 1g$

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



Wrapping up

Summary

- We propose *!*-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

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.

References

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 $\quad 2\ [m; M]$ and variance $\quad^2$. Let z denote the number of standard deviations required to achieve $1\quad$ confidence in coverage of the standard normal distribution. Denote the sample mean of n iid samples of X as $\quad$. Then

$$\Pr[\quad 2 [\,!\ (\ );\ !_+(\ )]] \quad ; \quad with\ !\ (\ ) = \frac{B}{2A} \quad \sqrt{\frac{B^2}{4A^2} - \frac{C}{A}};$$

where

$$A = n + z^2\ ; \quad B = 2n \quad + z^2\ (M + m); \quad C = n\ ^2 + z^2\ Mm; \quad and$$

$$= \frac{^2}{(M \quad)(\quad m)}\ if\quad 2\ (m; M); \quad and \quad = 1\ if\quad 2\ fm; Mg:$$

References

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 r_k be its expected reward, $p\frac{c_k}{1 \quad t}$ its expected cost, and $_k(\ ; t)$ the upper confidence bound for $r_k = c_k$, as in Eq. 10. For $; t > 0$, and $(t) < 1$ it holds that

$$\Pr \quad _k(\ ; t) \quad \frac{r_k}{c_k} \quad 1 \quad (t);$$

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

References

Theoretical analysis
Proof idea for worst-case regret

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

- Playing a suboptimal arm k leads to expected "incremental" regret of $\ ^{c}_{k}\ _{k}$

Derive regret obtained until time step

- Sum incremental regret over arms and time horizon

$$\text{Regret} = \sum_{i=1}^{K} \ ^{c}_{k}\ _{k}\mathrm{E}[n_k(T_B)]$$

Find $\ _B$ that is larger than T_B with high probability

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

Evaluate asymptotic behavior of regret

- Behavior of regret for $\ _B \ ! \ 1$

References

Theorem (Number of suboptimal plays)

With ! -UCB, the expected number of plays of a suboptimal arm $k > 1$ before time step , $\mathrm{E}[n_k(\ )]$, is upper-bounded by:

$$\mathrm{E}[n_k(\ )] \quad 1 + n_k(\ ) + (\ ;\ );$$

where

$$(\ ;\ ) = (\quad K)\ 2\ \frac{p}{1}\ \sum_{t=K+1}\ \frac{p}{1\quad t}\ ;$$

$$n_k(\ ) = \frac{8\ \log}{\frac{2}{k}}\ \max\ \frac{\frac{r}{k}\ \frac{r}{k}}{1\quad \frac{r}{k}};\ \frac{\frac{c}{k}(1\quad \frac{c}{k})}{\frac{c}{k}}\ ;\quad {}_k = \frac{k}{k + \frac{1}{\frac{c}{k}}};$$

and K and ${}_k$ are defined as before.

References

Theoretical analysis
Results (II)

Theorem (Worst-case regret)

Define $_B = 2B=\min_{k2[K]}$ ^c_k *and* $_k; n_k(\ _B)$, *and* $(\ _B;\ )$ *as before. For any* > 0, *the regret of* $!$*-UCB is upper-bounded by*

$$\text{Regret} \quad \sum_{k=2} \ _k(1 + n_k(\ _B) + (\ _B;\ )) + X(B) \sum_{k=2} \ _k + \frac{2\ ^r_1}{^c_1};$$

where $X(B)$ *is in* $O\ \frac{B}{^c_{min}} e^{\ 0:5B\ ^c_{min}}$.

Theorem (Asymptotic regret)

The regret of $!$*-UCB is*

$$\text{Regret } 2\ O\ B^1 \quad ; \text{ for } 0 < \ < 1; \qquad \text{Regret } 2\ O(\log B); \text{ for } \quad 1$$

References

Evaluation
Competitors

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

Policy	Ref.	Evaluated
$''$-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]	✓
! -UCB	ours	✓

References

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**	$U(0; 1)$	10	[Xia+15b; Xia+17]
			50	[Xia+17]
			100	[Xia+15a; Xia+15b]
	Generalized Bernoulli	$U(0; 1)$	10	[Xia+15b; Xia+16]
			50	[Xia+16]
			100	[Xia+15b]
	Beta	$U(0; 5)$	10	[Xia+17; Xia+16]
			50	[Xia+17; Xia+16]
			100	[Xia+15a]
Facebook	Bernoulli	given	$[2; 97]$	–
	Beta	randomized	$[2; 97]$	–

References