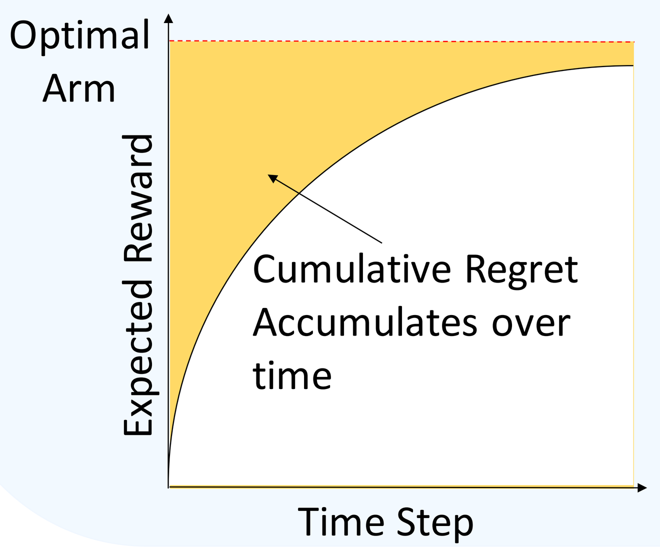
A New Approach to Correlated Multi-Armed Bandits

Juneja, I[1]., Gaharwar, D.S., Varshney, D., & Moharir, S.

Department Of Electrical Engineering, Indian Institute of Technology, Bombay, [1]Presenting Author

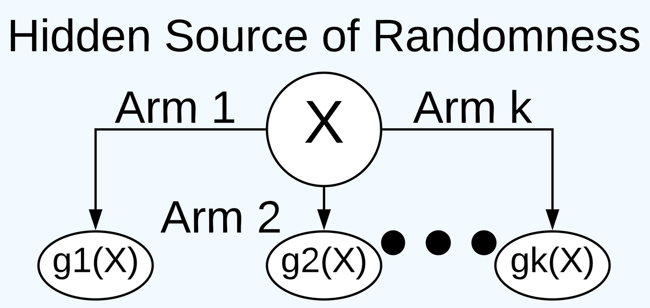
The Multi-Armed Bandit Problem

- In an MAB problem a player chooses one among many (k) choices
- Playing an arm returns a stochastic reward to the player
- The goal is to maximize cumulative rewards over time



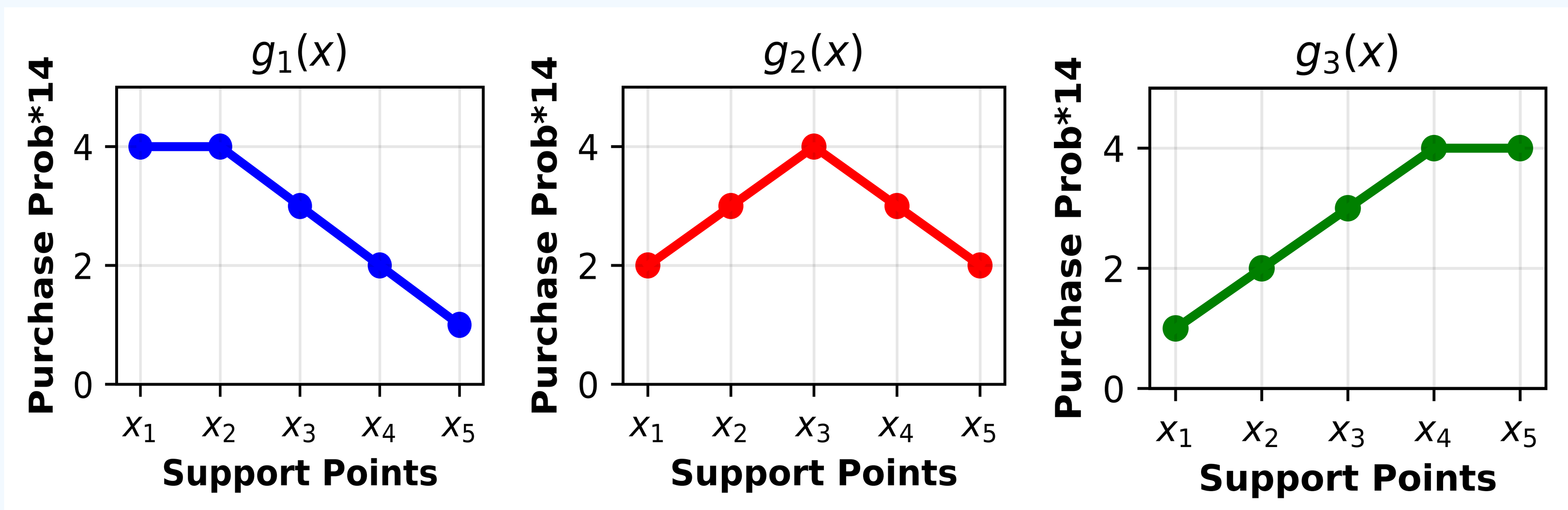
- Simple regret at time 't': Difference between expected reward obtained from sampling algorithm at time 't' and expected reward on sampling optimal arm
- Cumulative Regret at time 't': The sum of all simple regrets up to time 't'
- Under a good policy simple regret will decrease with time

Correlated Bandits and their Application



- Independence assumption between arms relaxed
- Skip pulling some arms by exploiting correlation
- Model Used: Hidden Random Source X
- $g_1(X)$, $g_2(X)$, …, $g_k(X)$ are known reward functions dependent only on X

- An E-Retailer wants to expand to a new country
- Goal is to maximize sales of smartphones through advertising
- Functions g_1, g_2, …, g_k are purchase probabilities as a function of Income level – X, modeled as a discrete r.v. with unknown distribution
- Even when the distribution changes, reward functions remain unchanged



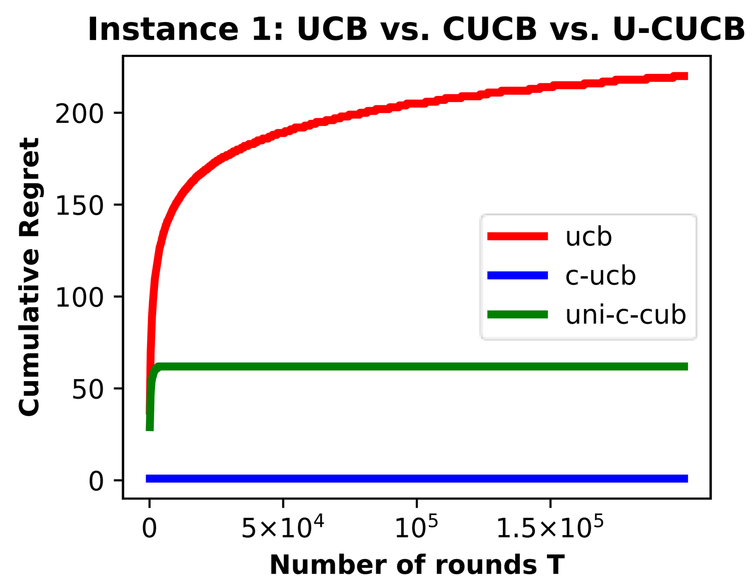
- For Instance consider above 3-arm Bandit Instance for three smart-phone models
- Horizontal Axis shows discretized income level support points
- Vertical Axis is the scaled product purchase probability

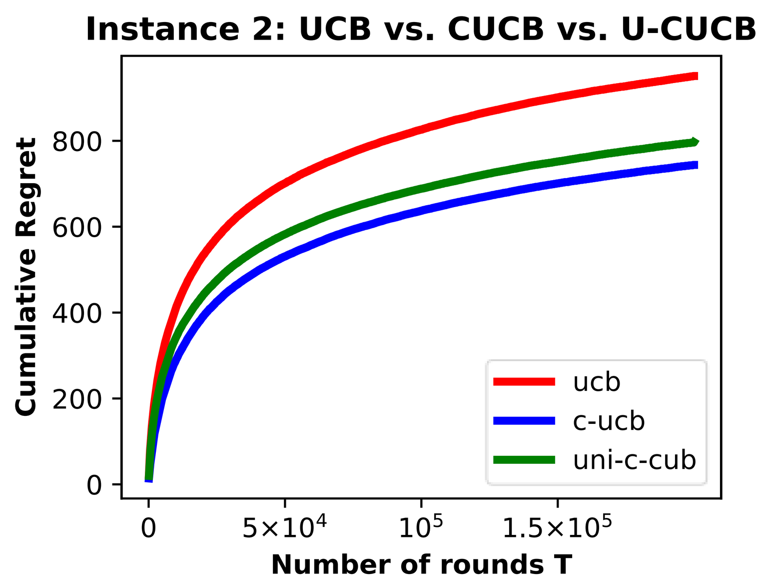
Sampling Algorithms - Correlated Bandits

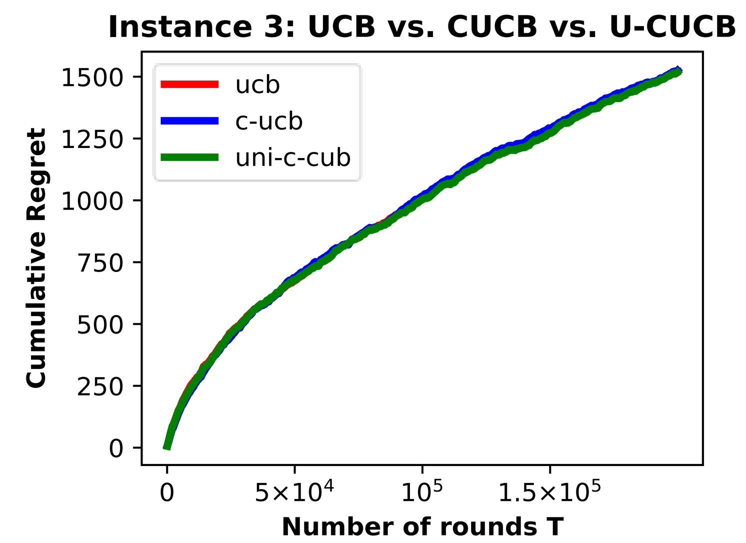
- UCB – Upper Confidence Bound, algorithm is a well-known solution to MAB arm selection problem
- The Algorithm chooses to sample an arm with the highest UCB index

$$I_k(t) = \hat{\mu}_k(t) + B\sqrt{\frac{2logt}{n_k(t)}}$$

- In general UCB achieves logarithmic regret scaling: R(T) = O(log T)
- If correlation between arms is exploited, order wise or constant factor improvement can be achieved by playing a modified strategy
- CUCB is a strategy for Correlated Bandits that was proposed in 2018







Probability Distributions Used

Instance 1: $\{0.8, 0.1, 0.04, 0.03, 0.03\}$

Instance 2: $\{0.03, 0.07, 0.41, 0.245, 0.245\}$

Instance 3: $\{0.13, 0.3, 0.24, 0.2, 0.13\}$

U-CUCB Terminology

- In our work, we propose the U-CUCB arm selection algorithm
- Key idea: Obtain an estimate of the distribution – "Pseudo Distribution"
- True distribution need not be learnable since reward functions are non-invertible
- If we observe reward r, at time t, the Pseudo-Distribution probability mass for the i^{th} support point is updated as follows:

$$p^{t+1}(x_i) = \frac{t \cdot p^t(x_i) + 1 \cdot \beta_i}{t+1}, \quad \beta_i = \begin{cases} \frac{1}{|\text{inv}_k(r)|} & i \in \text{inv}_k(r) \\ 0 & \text{otherwise,} \end{cases}$$

- Here, the inverse set is defined as $\text{inv}_k(r) := \{i : g_k(x_i) = r\}$
- Next, we define a confidence set C^* as per the following
  - Sort the empirical pseudo-distribution in descending order to obtain sorted indices q_1, q_2, …, q_n
  - Pick C* = $\{q_1, q_2, …, q_j\}$ where j is the smallest m s.t. $\sum_{i=1}^{m} p(x_i) > 1 - \epsilon$
- Here epsilon is a small number modelled as a Hyper-Parameter

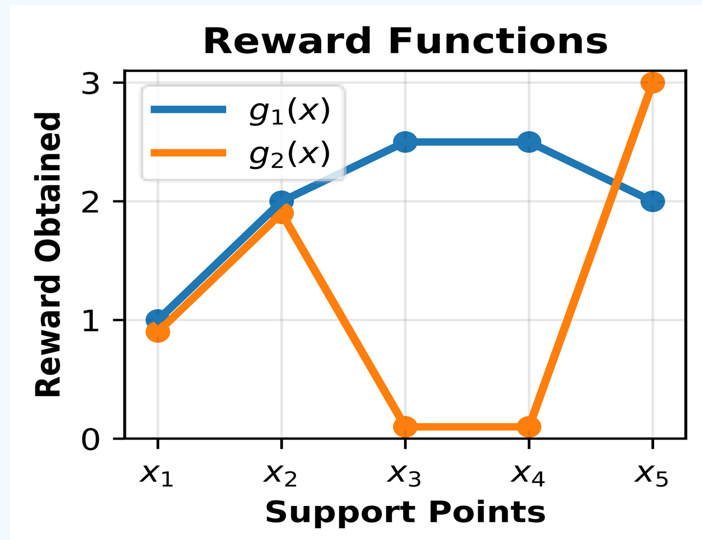
U-CUCB Sampling Algorithm

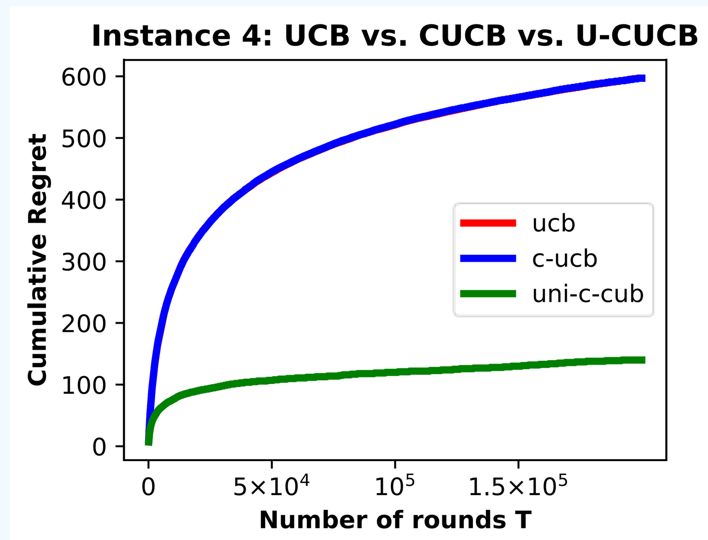
- To achieve an order wise or constant factor improvement over UCB we must avoid sampling certain 'non-competitive' arms
- An arm can be called 'non-competitive' if it can be determined to be sub-optimal through indirect sampling
- Under U-CUCB (U: Uniform and C: Correlated) we say arm k is non-competitive if there exists an arm j s.t.,

$$g_k(x) < g_j(x) \ \forall \ x \in C^* \text{ and } \ \hat{g}_k(X) < \hat{g}_j(X)$$

- Hatted variables represent empirically expected rewards from the respective arms
- Thus we identify an arm as non-competitive if its reward function lies entirely below some other reward function for all support points lying in C^*
- U-CUCB involves performing UCB arm selection over a reduced set consisting of only competitive arms.
- The regret scaling of U-CUCB will be at least as good as UCB with a constant factor improvement in most cases
- An order wise improvement to O(1) is achieved in cases where only the optimal arm is competitive – Instances 1 and 4 on this poster

Order Wise Improvement over CUCB





Distribution: $\{0.45, 0.45, 0.045, 0.045, 0.01\}$

- For the above Bandit Instance consisting of arms $g_1(X)$ and $g_2(X)$, CUCB is unable to identify arm 2 as non-competitive but U-CUCB is successful in doing so
- In general, Bandit Instances with arms having high rewards at support points outside of the confidence set C^* will perform better under U-CUCB
- Ongoing work includes finite time regret analysis of U-CUCB and finding sufficient conditions for Pseudo-Distribution to be reliable for arm classification

References

- Code Repository: https://github.com/ishank-juneja/Correlated-Multi-Armed-Bandits
- Gupta, S., Joshi, G., & Yağan, O. "Correlated Multi-Armed Bandits with Latent Random Source" International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2020.

Please see our extended paper for a more in depth explanation of U-CUCB, CUCB and the Correlated Bandit Framework