Distribution Centric Approach to Correlated Multi-Armed Bandits

Neharika Jali, Ishank Juneja
160040101, 16D070012

Department of Electrical Engineering, IIT Bombay

July 01, 2020

Overview

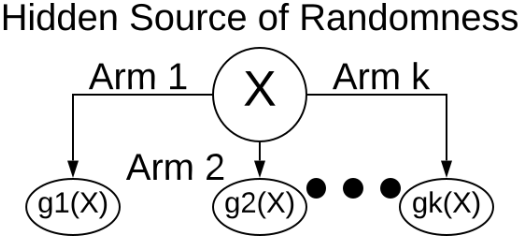
1. Introduction
2. Regret Minimization for Correlated Bandits
3. PAC Algorithm for Correlated Bandits

Introduction : What are Correlated Bandits?

- The usual MAB Setup has independent arms
- Independence assumption between arms is relaxed
- Correlation between arms can be exploited if present
- Skip pulling some arms based on correlation

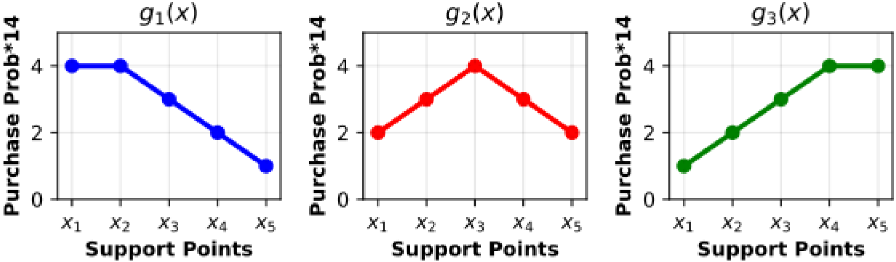
Introduction : What are correlated bandits?

- X is a discrete latent random variable
- Distribution of X is unknown
- $g_1(X), g_2(X), \ldots, g_k(X)$ are the dependent reward functions
- Here $g_1, g_2, \ldots, g_k$ are known functions



Introduction : An Example

An example of a situation where this model could be useful.
In general the arm functions are non-invertible.



Introduction : Overview of Reference Work

- Work of (Gupta et al. 2020)[2] describes a systematic method to exclude bad arms - CUCB
- Uses Distribution Agnostic Side information gathered for arm l using the pulls of arm k
- Skips sampling arms based on pairwise comparisons
- Excluded arms called *non competitive arms*
- To determine the arms to be excluded comparisons happen with certain reference arms
- Can we have a method that learns the underlying distribution and does more comparisons instead of just comparisons with a reference arm?

The UCUCB Algorithm

- Underlying distribution of X is learnt and used
- Called Pseudo distribution, an empirical estimate of the unknown distribution
- Computed as follows,

$$\tilde{p}_i(t) = \sum_{\tau=1}^{t} \frac{\beta_i(\tau)}{t} \tag{1}$$

Where, $\beta_i(\tau)$ is given by,

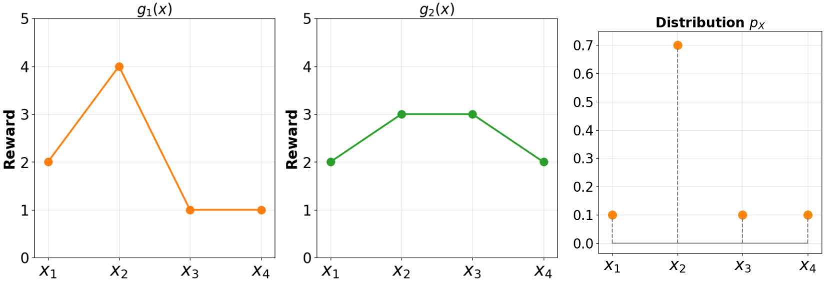
$$\beta_i(\tau) = \begin{cases} \frac{1}{|\text{inv}_k(r_\tau)|} & i \in \text{inv}_k(r_\tau) \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Here $\text{inv}_k(r_\tau)$ is the preimage of the reward r_τ under the function g_k and $|\text{inv}_k(r)|$ is the cardinality of the set $\text{inv}_k(r_\tau)$

Problems with UCUCB

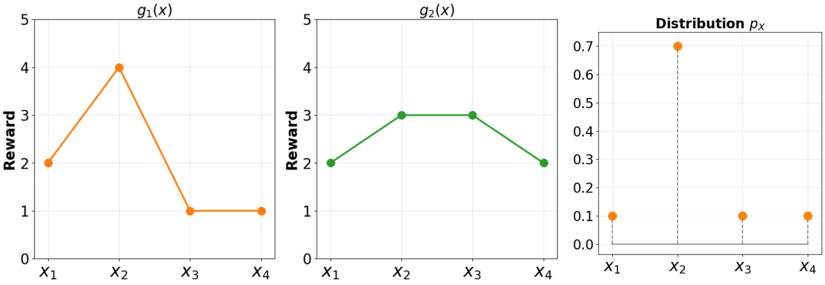
- UCUCB based on conjecture that persistent bias introduced into the distribution can only be due to indistinguishable points
- Uses a clause based on global pseudo distribution to remove non competitive arms
- Approach has limitations since it uses a global information estimate
- The idea of reference arms from CUCB is useful, makes comparison between performance of arms possible
- A global reference estimate such as the pseudo distribution cannot be relied upon for comparisons
- Pairwise distribution estimates do not make sense since they conceal all information

Problems: A Counter Example



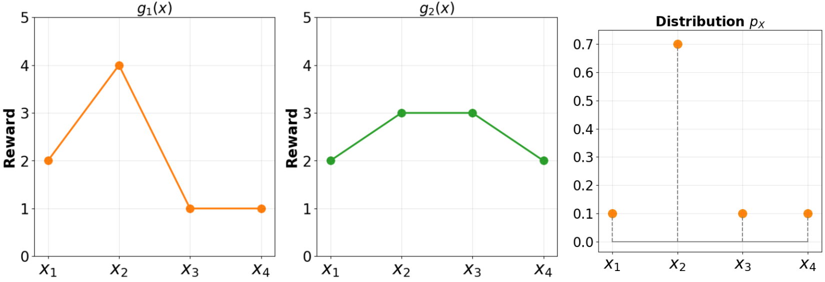
- Arm 2 pulled first, assume reward of 3 obtained
- As per the update rule we get $\tilde{P}_X(2) = [0, 0.5, 0.5, 0]$ for use in the next step
- This skewed distribution would make algorithm believe arm 2 is better

Problems: A Counter Example



- Now when arm 2 is sampled again and again the probability masses of positions 1 and 4 will remain in the neighbourhood of $\frac{1-(0.7+0.1)}{2} = 0.1$
- And through samples of arm 2, x_2 and x_3 will remain indistinguishable and their individual probability masses will never exceed 0.5
- $\tilde{P}(X)$ will always lie between $[0, 0.5, 0.5, 0]$ and $[0.1, 0.4, 0.4, 0.1]$. These distributions and everything in between will consider arm 2 to be superior to arm 1

Problems: A Counter Example



$E(t)$: The event that the algorithm breaks ties by pulling arm 2 first and obtains 3 as the reward

$$\sum_{t=1}^{T} \mathbb{E}[R(t)] = \sum_{t=1}^{T} \mathbb{E}[R(t)|E(t)]\mathbb{P}(E(t)) + \sum_{t=1}^{T} \mathbb{E}[R(t)|E^c(t)]\mathbb{P}(E^c(t)) \quad (3)$$

The first term on the RHS contributes linear regret. The second term is expected regret when $E(t)$ does not occur, which would still be non-negative.

Restrictions to Bandit Framework

- From counter example, it is clear that any temporary bias in the distribution is problematic
- Restriction that all arms be invertible is required
- Problem is no longer a partial observability Bandit Problem
- Becomes similar to experts setting with the restriction of drawing from a distribution
- Under this setting, constant cumulative expected regret can be achieved

Regret Minimization with Distribution (RMD)

Algorithm 1 RMD Algorithm

Input: Alphabet $\{x_1, ..., x_n\}$, Functions $\{g_1, ..., g_K\}$, All Invertible
Initialize : $t = 0, \tilde{g}_k = \infty$ (like Vanilla UCB)
for Every round t **do**
  $\tilde{g}_k(t) \leftarrow \sum_{i=1}^{n} g_k(x_i)\tilde{p}_i(t)$
  $k_t = \arg\max_k \tilde{g}_k(t)$
  Receive reward r_t by sampling arm k_t
  Record the realization of x, $x \leftarrow g_k^{-1}(r_t)$
  $t \leftarrow t + 1$
  $\tilde{p}_i(t+1) \leftarrow (\tilde{p}_i(t) \times t + \mathbb{1}_{x=x_i})/t$
end for

Analysis: Regret Upper Bound

Lemma (Hoeffding Inequality)

For a random variable $X \in (a, b)$,

$$\mathbb{P}\left(\frac{\sum_{\tau=1}^{t} X_\tau}{t} - \mu \geq \epsilon\right) \leq \exp\left(\frac{-2t\epsilon^2}{(b-a)^2}\right) \tag{4}$$

Applying the Hoeffding Inequality and using the fact that rewards are bounded between 0 and B, we have

Lemma (Number of Sub Optimal pulls)

The expected number of sub optimal pulls are bounded by,

$$\mathbb{E}\left[\sum_{\tau=1}^{t} \mathbb{1}_{k_\tau \neq k^*}\right] \leq K \sum_{\tau=1}^{t} \exp\left(\frac{-\tau \Delta_{\min}^2}{2B^2}\right) \tag{5}$$

Analysis: Regret Upper Bound

Theorem

The expected cumulative regret is upper bounded by,

$$\sum_{t=1}^{T} \mathbb{E}[R(t)] \leq K\Delta_{\max} \sum_{\tau=1}^{t} \exp\left(\frac{-\tau\Delta_{\min}^2}{2B^2}\right) \tag{6}$$

Which is a constant

- This bound relies on the distribution being unbiased and well sampled
- Well sampled meaning the number of samples associated with each $\tilde{p}_i(t)$ should be large
- In the absence of the invertibility assumptions, even logarithmic regret is not guaranteed

Pure Exploration Framework

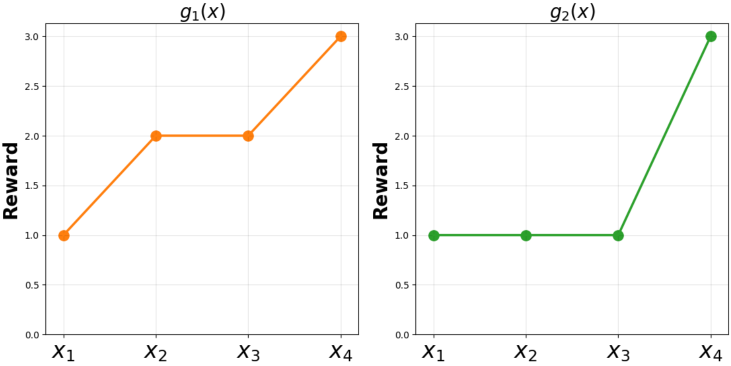
- Heavy restrictions required on allowed Bandit Instances to use method for regret minimization
- Regret minimization has exploration-exploitation trade-off
- Cannot actively learn distribution
- Look towards other frameworks - Pure Exploration Setting

PAC Algorithm for Correlated Bandits

- Given a set of arms with known reward functions $g_1, \ldots, g_K$, and an underlying latent random variable X with support points $X = \{x_1, ..., x_n\}$. The goal is to find the best arm
- We propose a $(0, \delta)$-PAC Algorithm for a correlated bandit based on RRPULL + PIEST algorithm by Gupta et al. (2018) [1] and the Successive Elimination algorithm (E Even-Dar et al. 2002 [4])
- The former learns distributions of rewards from indirect samples

Preprocessing

- Merge into 'superpoints'. Here x_2 and x_3 can be merged
- Normalise all rewards to [0,1], for ease of analysis



Notation

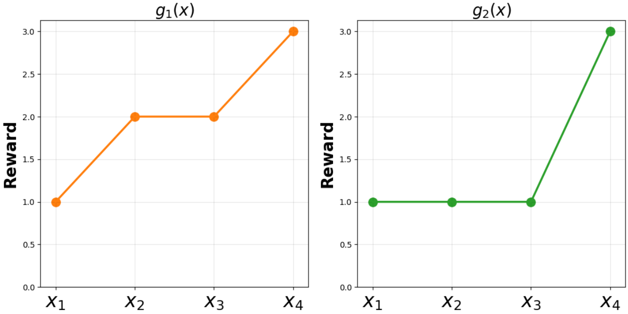
Borrowing notation from Gupta et al. (2018) [1], we define the following.

- As before,
  The true distribution of X is $P_X = [p_1, p_2, ..., p_n]^T$
  Our best estimate of P_X is $\tilde{P}_X = [\tilde{p}_1(t), \tilde{p}_2(t), ..., \tilde{p}_n(t)]^T$
- $\{z_{k,1}, ..., z_{k,m_k}\}$ - set of possible outcomes of the function g_k where m_k is the number of distinct outputs of g_k
- Sample Generation Matrix A_k of size $m_k \times n$

$$A_k(i,j) = \begin{cases} 1 & g_k(x_j) = z_{k,i} \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

- $A = [A_1^T, A_2^T, ..., A_K^T]$ of size $m \times n$ where $m = m_1 + m_2 + ... + m_K$
- A is the combined matrix and it captures information about the entire instance

Notation



$$A = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \hline 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right] \xrightarrow{\text{becomes}} A = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hline 1 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right] \tag{8}$$

Notation

- $q_{k,i}$ is the probability of observing $z_{k,i}$ each time arm k is pulled; $Q = [q_{1,1}, ..., q_{1,m_1}, ..., q_{K,m_K}]^T$, $\hat{Q}$ is the empirical estimate of Q
- Matrix A relates the entries of P_X and Q

$$q_{k,i} = \sum_{j=1}^{n} A_k(i,j) p_j \tag{9}$$

$$AP_X = Q \implies P_X = A^+ Q \tag{10}$$

- Lastly, as before, we have

$$\tilde{g}_s^t = \sum_{j=1}^{n} g_s(x_j) \tilde{p}_j(t) \tag{11}$$

Algorithm

Algorithm 2 Successive Elimination for Correlated Bandits

Input: Alphabet $\{x_1, ..., x_n\}$, Functions $\{g_1, ..., g_K\}$, Set of arms S
Initialize: $t = 0, t_k = 0 \; \forall \; k, \; t_{k,i} = 0 \; \forall \; i, k, \;\; \tilde{p}_j(0) = \frac{1}{n} \; \forall \; j$
while $|S| > 1$ **do**
  Preprocess by merging into superpoints
  Update the matrix A based on S and reduced alphabet
  Pull arm $s_t = mod(t, |S|) + 1$, observe output y_t
  $t_{s_t} = t_{s_t} + 1, \; t = t + 1$
  if $y_t = z_{s_t,i}$ **then**
    $t_{s_t,i} = t_{s_t,i} + 1$
  end if
  $\hat{q}_{s,i} = \frac{t_{s,i}}{t_s} \; \forall \; i, s$
  Obtain estimates $\tilde{p}_j(t)$ as $\tilde{P}_X = A^+ \hat{Q}$
  Let $\tilde{g}^t_{max} = \max_{s \in S} \tilde{g}^t_s, \; \alpha_t = \sqrt{\frac{\log(cKt^2/\delta)}{t}}$
  For every arm $s \in S$ s.t $\tilde{g}^t_{max} - \tilde{g}^t_s \geq 2\alpha_t$. Set $S = S \backslash s$
end while

Results Needed for Analysis

Theorem (Gupta et al., 2018 [1], Theorem 1)

It is possible to achieve asymptotically consistent estimation of probability distribution of X if and only if $rank(A) = n$.

Theorem (Gupta et al., 2018 [1], Theorem 6)

It is possible to achieve estimation error of probability distribution of X of $O(\frac{1}{t})$ if $rank(A) = n$.

The preprocessing step ensures that all instances have $rank(A) = n$.

Results Needed for Analysis

Theorem

The empirical estimate of the distribution P_X, $\tilde{P}_X = A^+\hat{Q}$ is unbiased

Proof:
By the definition of $\hat{q}_{s,i}$, $\hat{Q}$ is unbiased. Therefore, $\mathbb{E}[\hat{Q}] = Q$. Hence,

$$\mathbb{E}[\tilde{P}_X] = \mathbb{E}[A^+\hat{Q}] = A^+\mathbb{E}[\hat{Q}] = A^+Q = P_X \tag{12}$$

Theorem

The Successive Elimination for Correlated Bandits is a $(0, \delta)$-PAC algorithm, and with probability $(1 - \delta)$ its arm complexity is bounded by $O\left(\frac{\log(K/\delta\Delta_{min})}{\Delta_{min}^2}\right)$

Proof of $(0, \delta)$-PAC

For any time t and action $s \in S_t$, we have,

$$Pr[|\tilde{g}_s^t - \mu_s| \geq \alpha_t] \leq \exp^{-\alpha_t^2 t} \leq \frac{\delta}{cKt^2} \tag{13}$$

because the Hoeffding inequality can be applied to $\tilde{g}_s^t$, being an unbiased estimate of μ_s from theorem [6].

With probability at least $(1 - \frac{\delta}{K})$ for any time t and action $s \in S_t$, $|\tilde{g}_s^t - \mu_s| \leq \alpha_t$.

Hence, with probability $(1 - \delta)$, best arm is never eliminated since as $\alpha_t \to 0$ as t increases, eventually every non-best arm is eliminated.

Sample Complexity

To eliminate a non-best arm s_i, we need to reach a time t_i such that,

$$\hat{\Delta}_{t_i} = \tilde{g}_{s^*}^{t_i} - \tilde{g}_{s_i}^{t_i} \geq 2\alpha_{t_i} \tag{14}$$

Definition of α_t combined with the assumption that $|\tilde{g}_s^t - \mu_s| \leq \alpha_t$ yields,

$$\Delta_i - 2\alpha_t = (\mu_{s^*}(X) - \alpha_t) - (\mu_{s_i}(X) + \alpha_t) \geq \tilde{g}_{s^*}(X) - \tilde{g}_{s_i}(X) \geq 2\alpha_t \tag{15}$$

which holds with probability atleast $(1 - \frac{\delta}{K})$ for

$$t_i = O\left(\frac{\log(K/\delta\Delta_i)}{\Delta_i^2}\right) \tag{16}$$

The last non-best arm will thus be eliminated and the best arm output at

$$t = O\left(\frac{\log(K/\delta\Delta_{min})}{\Delta_{min}^2}\right) \tag{17}$$

Questions?

References

- S. Gupta, G. Joshi, O. Yagan. "Active Distribution Learning from Indirect Samples." *Allerton Conference on Communication, Control and Computing, 2018*
- S. Gupta, G. Joshi and O. Yağan, "Correlated Multi-Armed Bandits with A Latent Random Source," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
- I. Juneja, D. S. Gaharwar, D. Varshney and S. Moharir, "A New Approach to Correlated Multi Armed Bandits," 2020 COMSNETS
- E Even-Dar, S Mannor, Y Mansour. "PAC bounds for multi-armed bandit and Markov decision processes". International Conference on Computational Learning Theory 2002, 255-270

Thank you!