

Learning and Selecting Users for Achieving Reliability in Demand Response: A Multi-armed Bandit Approach

Yingying Li^a, Qinran Hu^b, Na Li^a

^aHarvard University, Cambridge, MA 02138, USA.

^bSoutheast University, Nanjing, China

Abstract

One challenge in the optimization and control of societal systems is to handle the unknown and uncertain user behavior. This paper focuses on residential demand response (DR) and proposes a closed-loop learning scheme to address these issues. In particular, we consider DR programs where an aggregator calls upon residential users to change their demand so that the total load adjustment is close to a target value. To learn and select the right customers, we formulate the DR problem as a combinatorial multi-armed bandit (CMAB) problem with a reliability objective. We propose a learning algorithm: CUCB-Avg (Combinatorial Upper Confidence Bound-Average), which utilizes both upper confidence bounds and sample averages to balance the tradeoff between exploration (learning) and exploitation (selecting). We consider both a fixed time-invariant target and time-varying targets and show that CUCB-Avg achieves $O(\log T)$ and $O(\sqrt{T \log(T)})$ regrets respectively. We further numerically test our algorithms using synthetic realistic DR data and demonstrate that our CUCB-Avg performs significantly better than the classic CUCB and better than Thompson Sampling.

Key words: learning theory; optimization under uncertainties; real time simulation and dispatching; multi-armed bandit; demand response; regret analysis.

1 Introduction

In many sequential decision-making problems of societal systems such as transportation, electricity grids, communication, crowd-sourcing, and resource allocation problems in general, one key challenge is to handle the unknown and uncertain user behavior. This paper focuses on residential demand response and aims to design a closed-loop learning scheme to address this challenge. Demand response (DR) has been playing an increasing role in electricity markets for improving grid resilience and sustainability [1–7]. However, most of the existing successful DR programs are for commercial and industrial customers. Despite a growing effort, residential DR remains underexploited [6]. In most residential DR programs, an aggregator such as a utility company requests load changes from users, for example, by changing the temperature set points of air conditioners. To encour-

age users' participation, most programs use incentive schemes such as prices, rewards, coupons, raffles, etc, under the assumption that customers are price responsive [2,6–8]. However, because the average monetary reward budget for a single household is usually small, it is reported that rewards play a limited role in affecting users' DR behaviors [6]. On the other side, there are many other factors that affect residential DR decisions, such as house sizes and types, household demographics and lifestyles, humidity and temperature, etc. However, it is unclear how these factors quantitatively affect DR actions. People with similar factors might react to a same DR signal in very different ways. Moreover, the DR aggregator usually has limited knowledge of these factors. These intrinsic, heterogeneous uncertainties associated with residential customers call for learning approaches to understand and interact with residential customers in a delicate way.

Multi-armed bandit (MAB) emerges as a natural framework to handle such uncertainties [9,10]. In a simple setting, MAB considers n independent arms, each providing a random contribution according to its own distribution at time steps $1 \leq t \leq T$. Without knowing these distributions, a decision maker picks one arm at each time

* The work was supported by NSF CAREER 1553407, NSF ECCS 1839632, AFOSR YIP, ONR YIP, and ARPA-E through the NODES program.

Email addresses: yingyingli@g.harvard.edu (Yingying Li), qhu@seu.edu.cn (Qinran Hu), nali@seas.harvard.edu (Na Li).

step, and tries to maximize the total expected contribution. The decision maker decides whether to *explore* arms to learn the unknown distributions, or to *exploit* the current knowledge by selecting the arm that has provided the highest mean contribution. When the decision maker can select multiple arms at each time, the problem is referred to as CMAB (Combinatorial MAB) in literature [11–15]. (C)MAB captures a fundamental trade-off in most learning problems: *exploration vs. exploitation*. A common metric to evaluate the performance of (C)MAB learning algorithms is regret, which captures the difference between the optimal expected value assuming the distributions are known and the achieved expected value of the online learning algorithm. A sub-linear regret implies good performance because it indicates that the learning algorithm eventually learns the optimal solution.

When applying CMAB framework to residential demand response, we can treat each customer as one arm. Then the aggregator follows CMAB methods to explore (learn) and exploit (select) the customers to achieve the goal of its DR program. There have been studies of DR via (C)MAB [16–18,7]. However, most literature sets the goal as maximizing the load reduction for peak hours rather than following a load reduction target for reliability issues. Besides, conducting rigorous regret analysis for achieving sub-linear regret is always challenging.

1.1 Our Contributions:

In this paper, we formulate the DR as a CMAB problem whose objective is to minimize the deviation between the total load adjustment and a target signal. The target might due to a sudden change of renewable energy or a peak load reduction request. for the sake of grid reliability. We consider a large number of residential customers, each of whom can commit one unit of load change (either reduction or increase) with an unknown probability. The task of the DR aggregator is to select a subset of the customers to approximate the target as close as possible. The size of the subset is not fixed, giving flexibility to the aggregator for achieving different levels of targets. Compared with the classic CMAB literature [11–15], a major difference of our formulation is that the reliability objective leads to a non-monotonic objective making the existing CMAB approaches and regret analysis inapplicable here [11].

In order to design our CMAB online learning algorithm, we first study the corresponding offline combinatorial optimization problem assuming the probabilities of the load change are known. Based on the structure of this offline algorithm, we propose an online algorithm CUCB-Avg (Combinatorial Upper Confidence Bound-Average) and provide a rigorous regret analysis. We show that, over T time steps, CUCB-Avg achieves $O(\log T)$ regret given a static target and $O(\sqrt{T \log(T)})$ regret given a time-varying target. The dependence of regret on the dimension n (the number of customers) is polynomial in

both cases. We also conduct numerical studies using synthetic DR data, showing that the performance of CUCB-Avg is much better than the classic algorithm CUCB [11,14,15] and slightly better than Thompson sampling, another popular CMAB method which uses a Bayesian learning approach and has good empirical performance [19–21].

Lastly, we would like to point out that despite a simplified DR model used in this paper, the model was motivated by real-world field studies of residential DR and the results in this paper have served as a guideline to design learning protocols for these studies [22]. Moreover, through modifying the detailed algorithms for other applications, the overall closed-loop learning scheme and the technical tools for analyzing its theoretical performance are useful for other societal systems especially with human-in-the-loop.

1.2 Related Work in CMAB.

Most literature in CMAB studies a classic formulation which aims to maximize the total (weighted) contribution of K arms with a fixed integer K (and known weights) [10,15,14,13,12,23].

As for more general settings, Chen et. al. design Combinatorial Upper Confidence Bound (CUCB) using the principle of *optimism in the face of uncertainty* when the objective function is *monotonically nondecreasing* with the parameters of arms given a fixed selected subset [11]. However, the reliability objective of our CMAB problem does not satisfy the monotone conditions. Another line of work follows the *Bayesian* approach and studies Thompson sampling [24,25]. However, the regret bounds of Thompson sampling depend exponentially on the number of the selected arms K and [25] shows that the exponential dependence is unavoidable. In the residential DR problems, K is usually large, thus Thompson sampling may generate poor performance. Finally, there is a lack of analysis on time-varying objective functions, but in many real-world applications the objectives change with time. For example, in residential DR, the load reduction targets would depend on the time-varying renewable generation. Therefore, either the learning algorithms or the theoretical analysis in literature do not directly apply to our CMAB problem, motivating the work of this paper.

Some of the preliminary work was presented in the conference paper [26]. This journal version strengthens the regret bounds, especially for the time-varying target case, conducts much more intensive numerical analysis using realistic data from ISOs, provides more complete proofs, and adds more physical intuitions and discussions on both theoretical and numerical results.

Notations. Let \bar{E} and $|E|$ be the complement and the cardinality of the set E respectively. For any positive integer n , let $[n] = \{1, \dots, n\}$. Let $I_E(x)$ be the indicator function: $I_E(x) = 1$ if $x \in E$ and $I_E(x) = 0$ if $x \notin E$.

For any two sets A, B , $A - B := \{x \mid x \in A, x \notin B\}$. When $k = 0$, let $\sum_{i=1}^k a_i = 0$ for any a_i , and the set $\{\sigma(1), \dots, \sigma(k)\} = \emptyset$ for any $\sigma(i)$. For $x \in \mathbb{R}^k$, we write $f(x) = O(g(x))$ as $x \rightarrow +\infty$ if there exists a constant M such that $|f(x)| \leq M|g(x)|$ for any x such that $x_i \geq M \forall i \in [k]$; and $f(x) = o(g(x))$ if $\lim_{x \rightarrow +\infty} f(x)/g(x) = 0$. We usually omit “as $x \rightarrow +\infty$ ” for simplicity. For the asymptotic behavior near zero, consider the inverse of x .

2 Problem Formulation

Motivated by the discussion in the introduction, we formulate the DR as a CMAB problem in this section. We focus on load reduction to illustrate the problem. The load increase can be treated in the same way. Consider a demand response (DR) program with an aggregator and n residential customers (arms) over T time steps where each time step corresponds to one DR event.¹ Each customer may respond to a DR event by reducing one unit of power consumption with probability $0 \leq p_i \leq 1$, or not respond with probability $1 - p_i$. The demand reduction by customer i at time step t is denoted by $X_{t,i}$, which is assumed to follow Bernoulli distribution: $X_{t,i} \sim \text{Bern}(p_i)$ and is independent across time.²

At each time $1 \leq t \leq T$, there is a DR event with a demand reduction target $D_t \geq 0$ determined by the power system. This reduction target might be due to a sudden drop of renewable energy generation or a peak load reduction request, etc. The aggregator aims to select a subset of customers $S_t \subseteq [n]$, such that the total demand reduction is as close to the target as possible. The loss/cost at time t can be captured by the *squared deviation* of the total reduction from the target D_t :

$$L_t(S_t) = \left(\sum_{i \in S_t} X_{t,i} - D_t \right)^2$$

Since demand reduction $X_{t,i}$ are random, the goal is to minimize the expected squared deviation,

$$\min_{S_t \subseteq [n]} \mathbb{E} L_t(S_t). \quad (1)$$

In this paper, we will first study the scenario where the target D is time-invariant (Section 3 and 4). Then we will extend the results to cope with time-varying targets to incorporate different DR signals resulted from the fluctuations of power supply and demand (Section 5).

¹ The specific definition of DR events and the duration of each event is up to the choice of the system designer. Our methods can accommodate different scenarios.

² For simplicity, we only consider that each customer has one unit to reduce. Our learning method can be extended to multi-unit setting and/or the setting where different users have different size of units. But the regret analysis will be more complicated which we leave as future work. As mentioned before, results in the paper have been used as a guideline for DR field studies[22].

When the response probability profile $p = (p_1, \dots, p_n)$ is known, the problem (1) is a combinatorial optimization, and an offline optimization algorithm is provided in Section 3. Let S_t^* denote an optimal solution.

In reality, the response probabilities are usually unknown. Thus, the aggregator should learn the probabilities from the feedback of previous demand response events, then make online decisions to minimize the difference between the total demand reduction and the target D_t . The learning performance is measured by $\text{Regret}(T)$, which compares the total expected cost of online decisions and the optimal total expected costs in T time steps:³

$$\text{Regret}(T) := \mathbb{E} \left[\sum_{t=1}^T R_t(S_t) \right] \quad (2)$$

where $R_t(S_t) := L_t(S_t) - L_t(S_t^*)$ and the expectation is taken with respect to $X_{t,i}$ and the possibly random S_t .

The feedback of previous demand response events includes the responses of every selected customer, i.e., $\{X_{t,i}\}_{i \in S_t}$. Such feedback structure is called *semi-bandit* in literature [11], and carries more information than bandit feedback which only includes the realized cost $L_t(S_t)$.

Lastly, we note that our problem formulation can be applied to other applications beyond demand response. One example is introduced below.

Example 1 Consider a crowd-sourcing related problem. Given budget D_t , a survey planner sends out surveys and offers one unit of reward for each participant. Each potential participant may participate with probability p_i . Let $X_{t,i} = 1$ if agent i participates; and $X_{t,i} = 0$, if agent i ignores the survey. The survey planner wants to maximize the total number of responses without exceeding the budget too much. One possible formulation is to select subset S_t such that the total number of responses is close to the budget D_t ,

$$\min_{S_t} \mathbb{E} \left(\sum_{i \in S_t} X_{t,i} - D_t \right)^2$$

Since the participation probabilities are unknown, the planner can learn the participation probabilities from the previous actions of its selected agents and then try to minimize the total costs during the learning process.

3 Algorithm Design

This section considers time-invariant target D . We will first provide an optimization algorithm for the offline

³ Strictly speaking, this is the definition of pseudo-regret, because its benchmark is the optimal expected cost: $\min_{S_t \subseteq [n]} \mathbb{E} L_t(S_t)$, instead of the optimal cost for each time, i.e. $\min_{S_t \subseteq [n]} L_t(S_t)$.

problem, then introduce the notations for online algorithms and discuss two simple algorithms: greedy algorithm and CUCB. Finally, we introduce our online algorithm CUCB-Avg.

3.1 Offline Optimization

When the probability profile p is known, the problem (1) becomes a combinatorial optimization problem:

$$\min_{S \subseteq [n]} \mathbb{E} L(S) \Leftrightarrow \min_{S \subseteq [n]} \left(\sum_{i \in S} p_i - D \right)^2 + \sum_{i \in S} p_i(1 - p_i) \quad (3)$$

Though combinatorial optimization is NP-hard and only has approximate algorithms in general, we are able to design a simple algorithm in Algorithm 1 to solve the problem (3) exactly. Roughly speaking, Algorithm 1 takes two steps: i) rank the arms according to p_i , ii) determine the number k according to the probability profile p and the target D and select the top k arms. The output of Algorithm 1 is denoted by $\phi(p, D)$ which is a subset of $[n]$. In the following theorem, we show that such algorithm finds an optimal solution to (3).

Algorithm 1 Offline optimization algorithm

- 1: **Inputs:** $p_1, \dots, p_n \in [0, 1]$, $D > 0$.
- 2: Rank p_i in a non-increasing order:
 $p_{\sigma(1)} \geq \dots \geq p_{\sigma(n)}$.
- 3: Find the smallest $k \geq 0$ such that

$$\sum_{i=1}^k p_{\sigma(i)} > D - 1/2$$

Let $k = n$ if $\sum_{i=1}^n p_{\sigma(i)} \leq D - 1/2$. Ties are broken randomly.

- 4: **Outputs:** $\phi(p, D) = \{\sigma(1), \dots, \sigma(k)\}$
-

Theorem 1 For any $D > 0$, the output of Algorithm 1, $\phi(p, D)$, is an optimal solution to (3).

Proof Sketch. We defer the detailed proof to [27] and only introduce the intuition here. An optimal set S roughly has two properties: i) the total expected contribution of S , $\sum_{i \in S} p_i$, is closed to the target D , ii) the total variance of arms in S is minimized. i) is roughly guaranteed by Line 3 of Algorithm 1: it is easy to show that $|\sum_{i \in \phi(p, D)} p_i - D| \leq 1/2$. ii) is roughly guaranteed by only selecting arms with higher response probabilities, as indicated by Line 2 of Algorithm 1. The intuition is the following. Consider an arm with large parameter p_1 and two arms with smaller parameters p_2, p_3 . For simplicity, we let $p_1 = p_2 + p_3$. Thus replacing p_1 with p_2, p_3 will not affect the first term in (3). However, $p_1(1 - p_1) \leq p_2(1 - p_2) + p_3(1 - p_3)$ by $p_1^2 = (p_2 + p_3)^2 \geq p_2^2 + p_3^2$. Hence, replacing one arm with higher response probability by two arms with lower response probabilities will increase the variance. \square

Corollary 1 When $D < 1/2$, the empty set is optimal.

Remark 1 There might be more than one optimal subset. Algorithm 1 only outputs one of them.

3.2 Notations for Online Algorithms

Let $\bar{p}_i(t)$ denote the sample average of parameter p_i by time t (including time t), then

$$\bar{p}_i(t) = \frac{1}{T_i(t)} \sum_{\tau \in I_i(t)} X_{\tau, i}$$

where $I_i(t)$ denotes the set of time steps when arm i is selected before time t and $T_i(t) = |I_i(t)|$ denotes the number of times that arm i has been selected before time t . Let $\bar{p}(t) = (\bar{p}_1(t), \dots, \bar{p}_n(t))$. Notice that before making decisions at time t , only $\bar{p}(t-1)$ is available.

3.3 Two Simple Online Algorithms: Greedy Algorithm and CUCB

In this subsection, we introduce two simple algorithms: greedy algorithm and CUCB, and explain why they perform poorly in our problem to gain intuitions for our algorithm development.

Greedy algorithm uses the sample average of each parameter $\bar{p}_i(t-1)$ as an estimation of the unknown probability p_i and chooses a subset based on the offline oracle described in Algorithm 1, i.e. $S_t = \phi(\bar{p}(t-1), D)$. The greedy algorithm is known to perform poorly because it only exploits the current information, but fails to explore the unknown information, as demonstrated below.

Example 2 Consider two arms with parameters $p_1 > p_2$. The goal is to select the arm with the higher parameter. Suppose after some time steps, we have explored the suboptimal arm 2 for enough times such that $\bar{p}_2 \approx p_2$, but haven't explored the optimal arm 1 enough so that \bar{p}_1 is an underestimate and satisfies: $\bar{p}_1 < \bar{p}_2 < p_1$. In this case, the greedy algorithm will keep selecting the suboptimal arm 2 based on current estimation \bar{p}_1, \bar{p}_2 and fail to explore arm 1. As a result, the regret will be $O(T)$.

A well-known algorithm in CMAB literature that balances the exploration and exploitation is CUCB [11,15]. Instead of using sample average \bar{p} directly, CUCB considers an upper confidence bound:

$$U_i(t) = \min(\bar{p}_i(t-1) + \sqrt{\frac{\alpha \log t}{2T_i(t-1)}}, 1) \quad (4)$$

where $\alpha \geq 0$ is the parameter to balance the tradeoff between $\bar{p}_i(t-1)$ (exploitation) and $T_i(t-1)$ (exploration). The output of CUCB is $S_t = \phi(U(t), D)$. CUCB performs well in classic CMAB problems, such as maximizing the total contribution of K arms for a fixed K .

However, CUCB performs poorly in our problem, as shown in Section 6. The major problem of CUCB is the over-estimate of the arm parameter p . By choosing $S_t = \phi(U(t), D)$, CUCB selects less arms than needed,

which not only results in a large deviation from the target, but also discourages exploration.

Algorithm 2 CUCB-Avg

- 1: **Notations:** $T_i(t)$ is the number of times selecting
 - 2: arm i by time t , and $\bar{p}_i(t)$ is the sample average of
 - 3: arm i by time t (both including time t).
 - 4: **Inputs:** $\alpha > 2$, D
 - 5: **Initialization:** For $t = 1, \dots, \lceil \frac{n}{2D} \rceil$, select $2D$ arms each time until each arm has been selected for at least once. Let S_t be the set of arms selected at time t . Initialize $T_i(t)$, $\bar{p}_i(t)$ by the observation $\{X_{t,i}\}_{i \in S_t}$.⁴
 - 6: **for** $t = \lceil \frac{n}{2D} \rceil + 1, \dots, T$ **do**
 - 7: Compute the upper confidence bound for each i

$$U_i(t) = \min(\bar{p}_i(t-1) + \sqrt{\frac{\alpha \log t}{2T_i(t-1)}}, 1).$$
 - 8: Rank $U_i(t)$ by a non-increasing order:

$$U_{\sigma(t,1)}(t) \geq \dots \geq U_{\sigma(t,n)}(t).$$
 - 9: Find the smallest $k_t \geq 0$ such that

$$\sum_{i=1}^{k_t} \bar{p}_{\sigma(t,i)}(t-1) > D - 1/2$$
 or $k_t = n$ if $\sum_{i=1}^n \bar{p}_{\sigma(t,i)}(t-1) \leq D - 1/2$.
 - 10: Select $S_t = \{\sigma(t,1), \dots, \sigma(t, k_t)\}$ and update $T_i(t)$ and $\bar{p}_i(t)$ based on the observations $\{X_{t,i}\}_{i \in S_t}$
 - 11: **end for**
-

3.4 Our Proposed Online Algorithm: CUCB-Avg

Based on our discussion above, we propose a new algorithm, CUCB-Avg. The novelty of our algorithm is that it utilizes both sample averages and upper confidence bounds by exploiting the structure of the offline optimal algorithm.

We note that the offline Algorithm 1 selects the right subset of arms in two steps: i) rank (top) arms, ii) determine the number k of the top k arms to select. In CUCB-Avg, we use the upper confidence bound $U_i(t)$ to rank the arms in a non-increasing order. This is the same as CUCB. However, the difference is that our CUCB-Avg uses the sample average $\bar{p}_i(t-1)$ to decide the number of arms to select at time t . The details of the algorithm are given in Algorithm 2.

Now we explain why the ranking rule and the selection rule of CUCB-Avg would work for our problem.

The ranking rule is determined by $U_i(t)$. An arm with larger $U_i(t)$ is given a priority to be selected at time t . We note that $U_i(t)$ is the summation of two terms: the sample average $\bar{p}_i(t-1)$ and the confidence interval radius that is related to how many times the arm has been explored. Therefore, an arm with a large $U_i(t)$ may

⁴ The initialization method is not unique and can be any method that selects each customer for at least once.

either has a small $T_i(t-1)$, meaning that the arm has not been explored enough or has a large $\bar{p}_i(t-1)$, indicating that the arm frequently responds in the history. In this way, CUCB-Avg selects both the under-explored arms (*exploration*) and the arms with good performance in the past (*exploitation*).

When determining k , CUCB-Avg uses the sample averages and selects enough arms such that the total sample average is close to D . Compared with CUCB which uses upper confidence bounds to determine k , our algorithm selects more arms, which reduces the load reduction difference from the target and also encourages exploration.

4 Regret analysis

In this section, we will prove that our algorithm CUCB-Avg achieves $O(\log T)$ regret when D is time invariant.

4.1 The Main Result

Theorem 2 *There exists a constant $\epsilon_0 > 0$ determined by p and D , such that for any $\alpha > 2$, the regret of CUCB-Avg is upper bounded by*

$$\text{Regret}(T) \leq M \left(\lceil \frac{n}{2D} \rceil + \frac{2n}{\alpha - 2} \right) + \frac{\alpha M n \log T}{2\epsilon_0^2} \quad (5)$$

where $M = \max(D^2, (n - D)^2)$. □

We make a few comments before the proof.

Dependence on T and n . The dependence on the horizon T is $O(\log T)$, so the average regret diminishes to zero as T increases, demonstrating that our algorithm learns the customers' response probabilities effectively. The dependence on the number of arms n is polynomial $O(n^3)$ since $M \sim O(n^2)$, demonstrating that our algorithm can effectively learn a large number of customers.

Role of ϵ_0 . The bound depends on a constant term ϵ_0 determined by p and D and such a bound is referred to as a *distribution-dependent bound* in literature. We defer the explicit expression of ϵ_0 to [27] and only explain the intuition behind ϵ_0 here. Roughly speaking, ϵ_0 is a robustness measure of our offline optimal algorithm, in the sense that if the probability profile p is perturbed to be \bar{p} by ϵ_0 (i.e., $|\bar{p}_i - p_i| < \epsilon_0$ for all i), the output $\phi(\bar{p}, D)$ of Algorithm 1 would still be optimal for the true profile p . Intuitively, if ϵ_0 is large, the learning task is easy because we are able to find an optimal subset given a poor estimation, leading to a small regret. This explains why the upper bound in (5) decreases when ϵ_0 increases.

To discuss what factors will affect the robustness measure ϵ_0 , we provide an explicit expression of ϵ_0 under two assumptions in the following proposition.

Proposition 1 *If the following two assumptions hold: (A1): p_i are positive and distinct: $p_{\sigma(1)} > \dots > p_{\sigma(n)} > 0$; and (A2): There exists $k \geq 1$ such that $\sum_{i=1}^k p_{\sigma(i)} >$*

$D - 1/2$, and $\sum_{i=1}^{k-1} p_{\sigma(i)} < D - 1/2$, then the ϵ_0 in Theorem 2 can be determined by:

$$\epsilon_0 = \min\left(\frac{\delta_1}{k}, \frac{\delta_2}{k}, \frac{\Delta_k}{2}\right) \quad (6)$$

where $k = |\phi(p, D)|$, $\sum_{i=1}^k p_{\sigma(i)} = D - 1/2 + \delta_1$, $\sum_{i=1}^{k-1} p_{\sigma(i)} = D - 1/2 - \delta_2$, and $\Delta_i = p_{\sigma(i)} - p_{\sigma(i+1)}$, $\forall i = 1, \dots, n - 1$.

We defer the proof of the proposition to our online report [27] and only make two comments here. Firstly, it is easy to verify that Assumptions (A1) and (A2) imply $\epsilon_0 > 0$. Secondly, we explain why ϵ_0 defined in (6) is a robustness measure. That is, if $\forall i$, $|\bar{p}_i - p_i| < \epsilon_0$, we have $\phi(\bar{p}, D) = \phi(p, D)$. This can be proved in two steps. Step 1: when $\epsilon_0 \leq \frac{\Delta_k}{2}$, the k arms with higher \bar{p}_i are the same k arms with higher p_i because for any $1 \leq i \leq k$ and $k+1 \leq j \leq n$, we have $\bar{p}_{\sigma(i)} > p_{\sigma(k)} - \epsilon_0 \geq p_{\sigma(k+1)} + \epsilon_0 > \bar{p}_{\sigma(j)}$. Step 2: because $\epsilon_0 \leq \frac{\delta_1}{k}, \frac{\delta_2}{k}$, we have i) $\sum_{i=1}^k \bar{p}_{\sigma(i)} > \sum_{i=1}^k (p_{\sigma(i)} - \epsilon_0) = D - 1/2 + \delta_1 - k\epsilon_0 \geq D - 1/2$ and ii) $\sum_{i=1}^{k-1} \bar{p}_{\sigma(i)} < \sum_{i=1}^{k-1} (p_{\sigma(i)} + \epsilon_0) = D - 1/2 - \delta_2 + (k-1)\epsilon_0 \leq D - 1/2$. Therefore, by Algorithm 1, $\phi(\bar{p}, D) = \{\sigma(1), \dots, \sigma(k)\} = \phi(p, D)$.

Finally, we briefly discuss how to generalize the expression (6) of ϵ_0 to cases without (A1) and (A2). When (A1) does not hold, we only consider the gap between the arms that are not in a tie, i.e. $\{\Delta_i \mid \Delta_i > 0, 1 \leq i \leq n - 1\}$. When (A2) does not hold and $\sum_{i=1}^{k-1} p_{\sigma(i)} = D - 1/2$, we consider less than $k - 1$ arms to make the total expected contribution below $D - 1/2$. Again, an explicit expression of ϵ_0 is provided in our report [27].

Comparison with the regret bound of classic CMAB. In classic CMAB literature when the goal is to select K arms with highest parameters given a fixed integer K , the regret bound usually depends on $\frac{\Delta_K}{2}$ [15]. We note that $\frac{\Delta_K}{2}$ is similar to ϵ_0 in our problem, as it is the robustness measure of the top- K -arm problem in the sense that given any estimation \bar{p} with estimation error at most $\frac{\Delta_K}{2}$: $\forall i$, $|\bar{p}_i - p_i| < \frac{\Delta_K}{2}$, the highest K arms with the profile \bar{p} are the same as that with the profile p . In addition, we would like to mention that the regret bound in literature is usually linear on $1/\Delta_K$, while our regret bound is $1/\epsilon_0^2$. This difference may be an artificial effect of the proof techniques because our CMAB problem is more complicated. We leave it as future work to strengthen the results.

Choice of α . When α increases, the term $\frac{2Mn}{\alpha-2}$ decreases while the term $\frac{\alpha M n \log T}{2\epsilon_0^2}$ increases. Since the second term is $O(\log T)$ while the first term is $O(1)$, α should be close to 2 when T is large.

4.2 Proof of Theorem 2

Proof outline: We divide the T time steps into four parts, and bound the regret in each part separately. The

partition of the time steps are based on event E_t and the event $B_t(\epsilon_0)$ defined below. Let E_t be the event when the sample average is outside the confidence interval considered in Algorithm 2:

$$E_t := \{\exists i \in [n], |\bar{p}_i(t-1) - p_i| \geq \sqrt{\frac{\alpha \log t}{2T_i(t-1)}}\}$$

For any $\epsilon > 0$, let $B_t(\epsilon)$ denote the event when Algorithm 2 selects an arm who has been explored for no more than $\frac{\alpha \log T}{2\epsilon^2}$ times:

$$B_t(\epsilon) := \{\exists i \in S_t, \text{ s.t. } T_i(t-1) \leq \frac{\alpha \log T}{2\epsilon^2}\} \quad (7)$$

Let $\epsilon_0 > 0$ be a small number such that Lemma 3 holds. Now, we will define the four parts of the T time steps, and briefly introduce the regret bound of each part.

- (1) *Initialization:* the regret bound does not depend on T (Inequality (8)).
- (2) *When E_t happens:* the regret bound does not depend on T because E_t happens rarely due to concentration properties in statistics (Lemma 1).
- (3) *When $\bar{E}_t, B_t(\epsilon_0)$ happen:* the regret is at most $O(\log T)$ because $B_t(\epsilon_0)$ happens for at most $O(\log T)$ times (Lemma 2).
- (4) *When $\bar{E}_t, \bar{B}_t(\epsilon_0)$ happen,* no regret due to enough exploration of the selected arms (Lemma 3).

Notice that the time steps are not divided sequentially here. For example, it is possible that $t = 10, 30$ belong to Part 2 while $t = 10$ belongs to Part 3.

Proof details: Firstly, it is without loss of generality to require $D \geq 1/2$ because when $D < 1/2$, the optimal set is empty and there is no regret by Corollary 1.

Secondly, we note that for all time steps $1 \leq t \leq T$ and any $S_t \subseteq [n]$, the regret at t is upper bounded by

$$R_t(S_t) \leq L_t(S_t) \leq \max(D^2, (n-D)^2) =: M \quad (8)$$

Thus, the regret of initialization (Part 1) at $t = 1, \dots, \lceil \frac{n}{\lceil 2D \rceil} \rceil$ is bounded by $M \lceil \frac{n}{\lceil 2D \rceil} \rceil$.

Next, we bound the regret of Part 2 by the Chernoff-Hoeffding's concentration inequality. The intuition behind the proof is that E_t happens rarely because the sample average $\bar{p}_i(t)$ concentrates around the true value p_i with a high probability.

Theorem 3 (Chernoff-Hoeffding's inequality)

X_1, \dots, X_m i.i.d in $[0, 1]$ with mean μ , then

$$\mathbb{P}\left(\sum_{i=1}^m X_i - m\mu \geq m\epsilon\right) \leq 2e^{-2m\epsilon^2} \quad (9)$$

Lemma 1 When $\alpha > 2$, $\mathbb{E} \sum_{t=1}^T I_{E_t} R_t(S_t) \leq \frac{2Mn}{\alpha-2}$.

Proof. The number of times E_t happens is bounded by

$$\begin{aligned}
\mathbb{E} \sum_{t=1}^T I_{E_t} &= \sum_{t=1}^T \mathbb{P}(E_t) \\
&\leq \sum_{t=1}^T \sum_{i=1}^n \mathbb{P} \left(|\bar{p}_i(t-1) - p_i| \geq \sqrt{\frac{\alpha \log t}{2T_i(t-1)}} \right) \\
&\leq \sum_{t=1}^T \sum_{i=1}^n \sum_{s=1}^{t-1} \mathbb{P}(|\bar{p}_i(t-1) - p_i| \geq \sqrt{\frac{\alpha \log t}{2s}}, T_i(t-1) = s) \\
&\leq \sum_{t=1}^T \sum_{i=1}^n \sum_{s=1}^{t-1} \frac{2}{t^\alpha} \leq \sum_{t=1}^T n \frac{2}{t^{\alpha-1}} \leq \frac{2n}{\alpha-2}
\end{aligned}$$

where the first inequality is by enumerating possible $i \in [n]$, the second inequality is by enumerating possible values of $T_i(t-1)$: $\{1, \dots, t-1\}$, the third inequality is by Chernoff-Hoeffding's inequality, and the last inequality is by $\sum_{t=1}^T \frac{1}{t^{\alpha-1}} \leq \int_1^{+\infty} \frac{1}{t^{\alpha-1}} \leq \frac{1}{\alpha-2}$. Then by inequality (8) the proof is completed. \square

Next, we show the regret of Part 3 is at most $O(\log T)$.

Lemma 2 *For any $\epsilon_0 > 0$, the regret in Part 3 is bounded by $\mathbb{E} \sum_{t=1}^T R_t(S_t) I_{\{\bar{E}_t, \bar{B}_t(\epsilon_0)\}} \leq \frac{\alpha \bar{M} n \log T}{2\epsilon_0^2}$*

Proof. By the definition of $B_t(\epsilon_0)$ in (7), whenever $B_t(\epsilon_0)$ happens, the algorithm selects an arm i that has not been selected for $\frac{\alpha \log T}{2\epsilon_0^2}$ times, increasing the selection time counter $T_i(t)$ by one. Hence, $B_t(\epsilon_0)$ can happen for at most $\frac{\alpha n \log T}{2\epsilon_0^2}$ times. Then, by inequality (8), the proof is completed. \square

When \bar{E}_t and $\bar{B}_t(\epsilon_0)$ happen (Part 4), every selected arm is fully explored and every arm's sample average is within the confidence interval. As a result, CUCB-Avg selects the right subset and hence contributes zero regret. This is formally stated in the following lemma.

Lemma 3 *There exists $\epsilon_0 > 0$, such that for each $1 \leq t \leq T$, if \bar{E}_t and $\bar{B}_t(\epsilon_0)$ happen, CUCB-Avg selects an optimal subset and $\mathbb{E} R_t(S_t) I_{\{\bar{E}_t, \bar{B}_t(\epsilon_0)\}} = 0$. Consequently, the regret in Part 4 is 0.*

Proof Sketch: We defer the proof to [27] and sketch the proof ideas here, which is based on two facts:

Fact 1: when \bar{E}_t and $\bar{B}_t(\epsilon_0)$ happen, the upper confidence bounds can be bounded by $U_i(t) > p_i$ for all $i \in [n]$, and the confidence bounds of the selected arm j satisfy

$$|\bar{p}_j(t-1) - p_j| < \epsilon_0, U_j(t) < p_j + 2\epsilon_0, \forall j \in S_t.$$

Fact 2: when ϵ_0 is small enough, CUCB-Avg selects an optimal subset.

To get the intuition for Fact 2, we consider the expression of ϵ_0 in (6) under Assumption (A1) (A2) in Proposition 1. Let $\phi(p, D) = \{\sigma(1), \dots, \sigma(k)\}$ denote the optimal

subset. In the following, we roughly explain why the selected subset S_t is optimal given ϵ_0 defined in (6):

i) By $\epsilon_0 \leq \frac{\Delta_{\min}}{2}$, we can show that the selected subset S_t is either a superset or a subset of the optimal subset $\{\sigma(1), \dots, \sigma(k)\}$.

ii) By $\epsilon_0 \leq \delta_1/k$, we can show that we will not select more than k arms, because, informally, even if we underestimate p_i , the sum of arms in $\{\sigma(1), \dots, \sigma(k)\}$ is still larger than $D - 1/2$.

iii) By $\epsilon_0 \leq \delta_2/k$, we can show that we will not select less than k arms, because, informally, even if we overestimate p_i , the sum of $k - 1$ arms in $\{\sigma(1), \dots, \sigma(k)\}$ is still smaller than $D - 1/2$. \square

The proof of Theorem 2 is completed by summing up the regret bounds of Part 1-4.

5 Time-varying target

In practice, the load reduction target is usually time-varying. We will study the performance of CUCB-Avg in the time-varying case below.

Notice that CUCB-Avg can be directly applied to the time-varying case by using D_t in Line 5 and Line 9 in Algorithm 2 at each time step t .

Next, we provide a regret bound for CUCB-Avg in the time-varying case. Notice that we impose no assumption on D_t except that it is bounded, which is almost always the case in practice.

Assumption 1 *There exists a finite $\bar{D} > 0$ such that $0 < D_t \leq \bar{D}$, $\forall 1 \leq t \leq T$.*

Theorem 4 *Suppose Assumption 1 holds. For any $\alpha > 2$, the regret of CUCB-Avg is upper bounded by*

$$\begin{aligned}
\text{Regret}(T) &\leq \bar{M}n + \frac{2\bar{M}n}{\alpha-2} + \frac{\alpha\bar{M}n \log T}{2\epsilon_1^2} \\
&\quad + n^2 \sqrt{2\alpha \log T} \sqrt{T + \frac{\alpha \log T}{2\epsilon_1^2}}
\end{aligned}$$

where $\bar{M} = \max(\bar{D}^2, n^2)$, $\epsilon_1 = \min(\frac{\Delta_{\min}}{2}, \frac{\beta_{\min}}{n})$, $\Delta_{\min} = \min\{\Delta_i \mid 1 \leq i \leq n-1, \Delta_i > 0\}$ and $\beta_{\min} = \min\{p_i \mid 1 \leq i \leq n, p_i > 0\}$.

Before the proof, we make a few comments below.

Dependence on T . The bound is $O(\sqrt{T \log T})$, still sublinear in T , meaning that our algorithm learns the customers' response probabilities well enough to yield diminishing average regret in the time-varying case.

The dependence on T is worse than the static case which is $O(\log T)$. We briefly discuss the intuition behind this difference. In the proof of Theorem 2, we show that there exists a threshold ϵ_0 depending on D such that when the estimation errors of parameter p_i for $i \in S_t$ are below ϵ_0 , our algorithm selects the optimal subset (Lemma 3).

Moreover, we also show that as t increases, with high probability the estimation error will decrease and eventually our algorithm will find the optimal subset and generate no more regret. However, in the time-varying case the argument above no longer holds because the threshold ϵ_0 will change with D_t , denoted as $\epsilon_0(D_t)$, and it is possible that the estimation error will always be larger than $\epsilon_0(D_t)$, as a result the algorithm will not find the optimal subset with high probability. This roughly explains why the bound of the time-varying case is worse than that of the static case.

Nevertheless, we are able to show that under some conditions the regret at time t is almost bounded by the estimation error at time t (Lemma 5), and the estimation error almost scales like $O(\sqrt{\log T/t})$. Therefore, even in the worst case, the total regret can be roughly bounded by $\sum_{t=1}^T O(\sqrt{\log T/t}) = O(\sqrt{T \log T})$.

Finally, we note that the regret bound is for the worst-case scenario and the regret in practice may be smaller.

Dependence on n . The bound is polynomial on the number of arms n : $O(n^3)$, because $M \sim O(n^2)$, demonstrating that our algorithm can learn a large number of arms effectively in the time-varying case.

Role of ϵ_1 . Notice that ϵ_1 only depends on p and does not depend on the target D_t . Roughly speaking, ϵ_1 captures how difficult it is to rank the arms correctly by the value of p_i , in the sense that as long as the estimation error of each p_i is smaller than ϵ_1 , the rank based on the estimation will be the correct rank based on the true parameter p_i .

5.1 Proof of Theorem 4

Most of the proof is similar to the static case. We also divide the time steps into four parts and complete the proof by combining the regret bound of each part. The first three parts can be bounded in similar ways as the static case. The major difference comes from Part 4.

(1) Initialization: the regret can be bounded by $\bar{M}n$ because the initialization at most lasts for n time steps and \bar{M} is an upper bound of the single-step regret.

(2) When E_t happens: notice that Lemma 1 still holds in the time-varying case if replacing M with \bar{M} , so the second part is bounded by $\mathbb{E} \sum_{t=1}^T I_{E_t} R_t(S_t) \leq \frac{2\bar{M}n}{\alpha-2}$.

(3) When $\bar{E}_t, B_t(\epsilon_1)$ happen: notice that Lemma 2 still holds so $\mathbb{E} \sum_{t=1}^T R_t(S_t) I_{\{\bar{E}_t, B_t(\epsilon_1)\}} \leq \frac{\alpha \bar{M} n \log T}{2\epsilon_1^2}$.

(4) When $\bar{E}_t, \bar{B}_t(\epsilon_1)$ happen, we can show that the regret is $O(\sqrt{T \log T})$ as stated in the lemma below.

Lemma 4

$$\mathbb{E} \sum_{t=1}^T R_t(S_t) I_{\{\bar{E}_t, \bar{B}_t(\epsilon_1)\}} \leq 2n^2 \sqrt{\frac{\alpha \log T}{2}} \sqrt{T + \frac{\alpha \log T}{2\epsilon_1^2}}$$

Proof. Without loss of generality, we consider $D_t \geq 1/2$ for all t , because if $D_t < 1/2$ for some t , the algorithm will select the empty set and produce no regret at t .

Below is the intuition behind the proof. First, Lemma 5 shows that the regret at time t can roughly be bounded by the estimation error ϵ at t when $\epsilon \leq \epsilon_1$. Then, we show that the estimation error, roughly captured by $\sqrt{\log T/T_i(t-1)}$, almost decays as $O(\sqrt{\log T/t})$. Thus, the total regret is bounded by $\sum_t \sqrt{\log T/t} = \sqrt{T \log T}$.

Lemma 5 Consider any t and D_t . For any $0 < \epsilon \leq \epsilon_1$ such that $\mathbb{P}(\bar{E}_t, \bar{B}_t(\epsilon)) > 0$, let \mathbb{S}_t denote the set of all possible selections of CUCB-Avg at time t if event \bar{E}_t and $\bar{B}_t(\epsilon)$ happen, then $\forall s \in \mathbb{S}_t$, we have $\mathbb{E}[R_t(s)] \leq 2n\epsilon$.

Proof sketch. Due to the space limit, we defer the proof to [27] and only sketch the proof here. Firstly, we are able to show that under $\bar{E}_t, \bar{B}_t(\epsilon)$, the selected subset differs from the optimal subset for at most one arm. This is mainly due to $\epsilon \leq \epsilon_1$. Secondly, we can bound the regret of the suboptimal selections by $O(\epsilon)$, which is mainly due to our quadratic loss function. \square

Next, we will finish the proof by bounding the estimation errors. We introduce event H_t^q to represent that each selected arm i at time t has been selected for more than $\frac{\alpha \log T}{2\epsilon_1^2} + q$ times for $q = 0, 1, 2, \dots$:

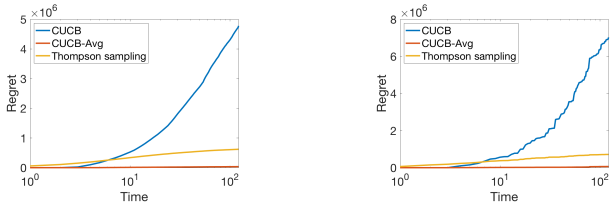
$$H_t^q := \{\forall i \in S_t, T_i(t-1) > \frac{\alpha \log T}{2\epsilon_1^2} + q\} \cap \bar{E}_t \cap \bar{B}_t(\epsilon_1)$$

In addition, we define the estimation error η_q by the confidence interval radius when an arm has been explored for $\frac{\alpha \log T}{2\epsilon_1^2} + q - 1$ times: $\frac{\alpha \log T}{2\eta_q^2} = \frac{\alpha \log T}{2\epsilon_1^2} + q - 1$, that is,

$$\eta_q = \sqrt{\frac{\frac{\alpha \log T}{2}}{q-1 + \frac{\alpha \log T}{2\epsilon_1^2}}}. \text{ The proof is completed by:}$$

$$\begin{aligned} \mathbb{E} \sum_{t=1}^T R_t(S_t) I_{\bar{E}_t \cap \bar{B}_t(\epsilon_1)} &= \sum_{q=1}^T \sum_{t=1}^T \mathbb{E} R_t(S_t) I_{(H_t^{q-1} - H_t^q)} \\ &\leq \sum_{q=1}^T \sum_{t=1}^T 2n\eta_q \mathbb{E} I_{(H_t^{q-1} - H_t^q)} \\ &\leq \sum_{q=1}^T 2n^2 \eta_q = 2n^2 \sum_{q=1}^T \sqrt{\frac{\frac{\alpha \log T}{2}}{q-1 + \frac{\alpha \log T}{2\epsilon_1^2}}} \\ &\leq 2n^2 \sqrt{\frac{\alpha \log T}{2}} \left(\sqrt{T + \frac{\alpha \log T}{2\epsilon_1^2}} - \sqrt{\frac{\alpha \log T}{2\epsilon_1^2}} \right) \end{aligned}$$

where the first equality is by $\bar{E}_t \cap \bar{B}_t(\epsilon_1) = \cup_{q=1}^T (H_t^{q-1} - H_t^q)$; the first inequality is by Lemma 5, $(H_t^{q-1} - H_t^q) \subseteq \bar{E}_t \cap \bar{B}_t(\eta_q)$ and $\eta_q \leq \epsilon_1$; the second one is because



(a) Average peak.

(b) Daily peak.

Fig. 1. The regret of three algorithms.

$$H_t^{q-1} - H_t^q \subseteq \bigcup_{i=1}^n \{i \in S_t, T_i(t-1) = \frac{\alpha \log T}{2\epsilon_1^2} + q\} \text{ and } \{i \in S_t, T_i(t-1) = \frac{\alpha \log T}{2\epsilon_1^2} + q\} \text{ occurs at most once. } \square$$

6 Numerical Experiments

In this section, we conduct numerical experiments to complement the theoretical analysis above.

6.1 Algorithms comparison

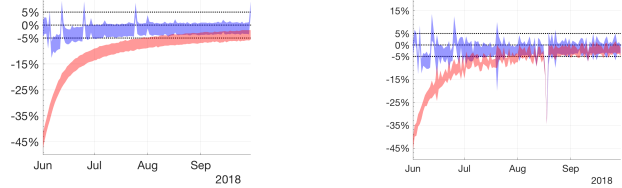
We will compare our algorithm with Thompson sampling (TS), an algorithm with good empirical performance in classic MAB problems. We briefly introduce the Thompson sampling below. In TS, the unknown parameter p is viewed as a random vector with a prior distribution. The algorithm selects a subset $S_t = \phi(\hat{p}_t, D)$ based on a sample \hat{p}_t from the prior distribution of p at $t = 1$ (or the posterior distribution at $t \geq 2$), then updates the posterior distribution of p by observations $\{X_{t,i}\}_{i \in S_t}$. For more details, we refer the reader to [19].

In our experiment, we consider a residential demand response program with 3000 customers. Each customer can either participate in the DR event by reducing 1kW or not. The probabilities of participation are i.i.d. $\text{Unif}[0, 1]$. The demand response events last for one hour on each day from June to September in 2018, with a goal of shaving the peak loads in Rhode Island. The *hourly* demand profile is from New England ISO [28]. We consider two schemes to determine the peak-load-shaving target D_t :

- i) Average peak: Consider the average daily load profile in these four months. The constant target D is the 5% of the difference between the peak average load and the average load at one hour before the peak hour.
- ii) Daily peak: On each day t , the target D_t is 5% of the difference between the peak load and the load at one hour before the peak hour of the daily demand.

In CUCB-Avg and CUCB, we set $\alpha = 2.5$. In Thompson sampling, the prior distribution of p is $\text{Unif}[0, 1]^n$.

Figure 1 plots the regret of CUCB, CUCB-Avg and TS under two schemes of peak shaving. The x-axis is in log scale. Both figures show that CUCB-Avg performs better than CUCB and TS. In addition, the regret of CUCB-Avg in Figure 1(a) is linear with respect to $\log(T)$, consistent with our theoretical result in Theorem 2. Moreover, the regret of CUCB-Avg in Figure 1(b) is almost



(a) Average peak.

(b) Daily peak.

Fig. 2. 90% confidence intervals of load reduction's relative errors of CUCB-Avg (blue) and Thompson sampling (red).

linear with $\log(T)$, demonstrating that in practice the regret can be much better than our worst case regret bound in Theorem 4.

Figure 2 plots the 90% confidence interval of the relative reduction error, $\frac{\sum_{i \in S_t} X_{t,i} - D_t}{D_t}$, of CUCB-Avg and TS by 1000 simulations. It is observed that the relative error of CUCB-Avg roughly stays within $\pm 5\%$, much better than Thompson sampling. This again demonstrates the reliability of CUCB-Avg. Interestingly, the figure shows that TS tends to reduce less load than the target, which is possibly because TS overestimates the customers' load reduction when selecting customers. Finally, on August 18th both algorithms cannot fulfill the daily peak target because it is very hot and the target is very large.

6.2 Effect of α and n

Next, we discuss the effect of the tradeoff parameter α by comparing the relative deviation of the load reduction, $\sqrt{\mathbb{E} L(S_t)} / D_t$, of CUCB-Avg with different α in Figure 3(a). The target is determined by scheme (i). It is observed that when T is small, a large α provides smaller relative deviation. This is because when T is small, the information of customers is limited. A larger α encourages exploration of the information, yielding better performance. When T is large, a smaller α leads to a better performance. This is because when T is large, the information of customers is sufficient, and a small α encourages the exploitation of the current information, thus generating better decisions. In addition, Figure 3(a) shows that for a wide range of α 's values, CUCB-Avg reduces the deviation to below 5% after a few days, showing that CUCB-Avg is pretty robust to the choice of α .

Finally, we discuss the effect of the total number of the customers n . The target is determined by scheme (i). Figure 3(b) shows that even with a large number of customers, CUCB-Avg reduces the relative deviation to below 5% very quickly, demonstrating that our algorithm can handle large n effectively. In addition, when T is small, a small n provides smaller relative deviation, because a small number of customers is easier to learn in a short time period. When T is large, a large n provides better performance, because there are more reliable customers to choose from a larger customer pool.

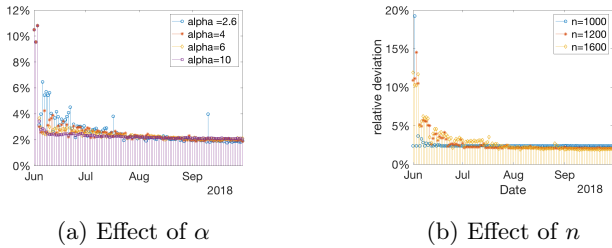


Fig. 3. The standard deviation of the actual load reduction of CUCB-Avg with different α or different number of arms n .

7 Conclusion

This paper studies a CMAB problem motivated by residential demand response with the goal of minimizing the difference between the total load adjustment and the target value. We propose CUCB-Avg, and show that CUCB-Avg achieves sublinear regrets in both static and time-varying cases. Future work includes considering dynamic population and general customer behavior models, e.g. continuous distributions, Markov processes, etc.

References

- [1] Pierluigi Siano. Demand response and smart grid – a survey. *Renewable and Sustainable Energy Reviews*, 30(C):461–478, 2014.
- [2] Na Li, Lijun Chen, and Steven H Low. Optimal demand response based on utility maximization in power networks. In *2011 IEEE power and energy society general meeting*, pages 1–8. IEEE, 2011.
- [3] PJM. PJM: Demand response. <http://www.pjm.com/markets-and-operations/demand-response.aspx>, 2018.
- [4] NYISO. New york ISO demand response program. http://www.nyiso.com/public/markets_operations/market_data/demand_response/index.jsp.
- [5] Farrokh Rahimi and Ali Ipekchi. Demand response as a market resource under the smart grid paradigm. *IEEE Transactions on Smart Grid*, 1(1):82–88, 2010.
- [6] FERC. Reports on Demand Response and Advanced Metering . Technical report, Federal Energy Regulatory Commission, 12 2017.
- [7] Daniel O’Neill, Marco Levorato, Andrea Goldsmith, and Urbashi Mitra. Residential demand response using reinforcement learning. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 409–414. IEEE, 2010.
- [8] Haiwang Zhong, Le Xie, and Qing Xia. Coupon incentive-based demand response: Theory and case study. *IEEE Transactions on Power Systems*, 28(2):1266–1276, 2013.
- [9] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.
- [10] Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.
- [11] Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, 17(1):1746–1778, 2016.
- [12] Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, et al. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, pages 2116–2124, 2015.
- [13] Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking (TON)*, 20(5):1466–1478, 2012.
- [14] Branislav Kveton, Zheng Wen, Azin Ashkan, Hoda Eydgahi, and Brian Eriksson. Matroid bandits: Fast combinatorial optimization with learning. *arXiv preprint arXiv:1403.5045*, 2014.
- [15] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543, 2015.
- [16] Qingsi Wang, Mingyan Liu, and Johanna L Mathieu. Adaptive demand response: Online learning of restless and controlled bandits. In *Smart Grid Communications (SmartGridComm), 2014 IEEE International Conference on*, pages 752–757. IEEE, 2014.
- [17] Antoine Lesage-Landry and Joshua A Taylor. The multi-armed bandit with stochastic plays. *IEEE Transactions on Automatic Control*, 2017.
- [18] Shweta Jain, Balakrishnan Narayanaswamy, and Y Narahari. A multiarmed bandit incentive mechanism for crowdsourcing demand response in smart grids. In *AAAI*, pages 721–727, 2014.
- [19] Daniel Russo, Benjamin Van Roy, Abbas Kazerouni, and Ian Osband. A tutorial on thompson sampling. *arXiv preprint arXiv:1707.02038*, 2017.
- [20] Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.
- [21] Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.
- [22] Con Edison. Consolidated edison smart ac program. <https://conedsmartac.com>.
- [23] Jean-Yves Audibert, Sébastien Bubeck, and Gábor Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2013.
- [24] Aditya Gopalan, Shie Mannor, and Yishay Mansour. Thompson sampling for complex online problems. In *International Conference on Machine Learning*, pages 100–108, 2014.
- [25] Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 5101–5109, 2018.
- [26] Yingying Li, Qinran Hu, and Na Li. Learning and selecting the right customers for reliability: A multi-armed bandit approach. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 4869–4874. IEEE, 2018.
- [27] Yingying Li, Qinran Hu, and Na Li. Learning and selecting users for achieving reliability: A multi-armed bandit approach. 2019. online report, <https://nali.seas.harvard.edu/files/nali/files/automaticadr.pdf>.
- [28] NEISO. New england iso express load and demand data. <https://www.iso-ne.com/isoexpress/web/reports/load-and-demand/-/tree/demand-by-zone>.

Appendix.

A Proof of Theorem 1

Only in this subsection, we assume $p_1 \geq \dots \geq p_n$ to simplify the notation. This will not cause any loss of generality in the offline analysis. In other parts of the document, the order of p_1, \dots, p_n is unknown, and we will use $p_{\sigma(1)} \geq \dots \geq p_{\sigma(n)}$ to denote the non-increasing order of parameters.

Since adding or removing an arm i with $p_i = 0$ will not affect the regret, in the following we will assume $p_i > 0$ without loss of generality⁵.

The proof of Theorem 1 takes two steps.

- (1) In Lemma 6, we establish a local optimality condition, which is a necessary condition to the global optimality.
- (2) In Lemma 7, we show that there exists an optimal selection S^* which includes first k arms with an unknown k . Then we can easily show that the Algorithm 1 selects the optimal k .

We first state and prove Lemma 6.

Lemma 6 *Suppose S^* is optimal and $p_i > 0$ for $i \in S^*$. Then we must have*

$$\sum_{i \in S^* - \{j\}} p_i \leq D - 1/2, \quad \forall j \in S^*$$

If $S^* \neq [n]$, then we will also have

$$\sum_{i \in S^*} p_i \geq D - 1/2$$

Proof. Since S^* is optimal, removing an element will not reduce the expected loss, i.e.

$$\mathbb{E} L(S^*) \leq \mathbb{E} L(S^* - \{j\})$$

which is equivalent with

$$\begin{aligned} & \mathbb{E} L(S^*) - \mathbb{E} L(S^* - \{j\}) \\ &= (2 \sum_{i \in S^*} p_i - 2D - p_j)p_j + p_j(1 - p_j) \\ &= (2 \sum_{i \in S^*} p_i - 2D + 1 - 2p_j)p_j \leq 0 \end{aligned}$$

Since $p_j > 0$, we must have

$$\sum_{i \in S^* - \{j\}} p_i \leq D - 1/2$$

⁵ One way to think about this is that we only consider subset S such that $p_i > 0$ for $i \in S$.

If $S^* \neq [n]$, then adding an element will not reduce the cost. So we must have

$$\mathbb{E} L(S^*) \leq \mathbb{E} L(S^* \cup \{j\}), \quad \forall j \notin S^*$$

which is equivalent with

$$\begin{aligned} & \mathbb{E} L(S^*) - \mathbb{E} L(S^* \cup \{j\}) \\ &= -(2 \sum_{i \in S^*} p_i - 2D + p_j)p_j - p_j(1 - p_j) \\ &= -(2 \sum_{i \in S^*} p_i - 2D + 1)p_j \leq 0 \end{aligned}$$

Since $p_j > 0$, we must have

$$\sum_{i \in S^*} p_i \geq D - 1/2$$

□

Corollary 2 *When $D < 1/2$, the empty set is optimal.*

Proof. Suppose there exists a non-empty optimal subset $S \neq \emptyset$. Then

$$\sum_{i \in S^* - \{j\}} p_i \geq 0 > D - 1/2, \quad \forall j \in S^*$$

which results in a contradiction by Lemma 6. □

Corollary 3 *When $\sum_{i=1}^n p_i \leq D - 1/2$, the optimal subset is $[n]$.*

Proof. Suppose there exists an optimal subset $S \neq [n]$. Then

$$\sum_{i \in S^*} p_i < \sum_{i=1}^n p_i \leq D - 1/2$$

which results in a contradiction by Lemma 6. □

Next, we are going to show that there must exist an optimal subset containing all elements with highest mean values. This is done by contradiction.

Lemma 7 *When $D \geq 1/2$ and $\sum_{i=1}^n p_i > D - 1/2$, there must exist an optimal subset S^* whose elements' mean values are $q_1 \geq \dots \geq q_m$, such that for any $p_i > q_m$, we have $i \in S^*$.*

Proof. Let's prove by construction and contradiction. Consider S^* , assume there exists $p_i > q_m$ but $i \notin S^*$. In the following, we will ignore other random variables outside $S^* \cup \{i\}$ because they are irrelevant. Now, we rank the mean value $\{q_1, \dots, p_i, \dots, q_m\}$ and assume that p_i is the j th largest element here. To simplify the notation, we will call the newly ranked mean value set as

$$p_1 \geq \dots \geq p_{j-1} \geq p_j \geq p_{j+1} \geq \dots \geq p_{m+1}$$

The mean values of random variables in S^* are $\{p_1, \dots, p_{j-1}, p_{j+1}, \dots, p_{m+1}\}$ (used to be called as $\{q_1, \dots, q_m\}$) and the injected element (used to be denoted as p_i) now is called p_j . Under this simpler notation, we proceed to construct a subset of top k arms with some k whose expected loss is no more than the optimal expected loss.

Construct a subset A in the following way. Pick the smallest k such that

$$\begin{aligned} \sum_{i=1}^k p_i &> D - 1/2 \\ \sum_{i=1}^{k-1} p_i &\leq D - 1/2 \end{aligned}$$

Then let $A = \{1, \dots, k\}$. It is easy to see that $k \geq j$. (Since S^* is optimal, by Lemma 6, excluding any element will go below $D - 1/2$, so k must include the newcomer j to be beyond $D - 1/2$.)

We claim that $\mathbb{E}L(A) \leq \mathbb{E}L(S^*)$. Since $\mathbb{E}L(S^*)$ is optimal, we must have $\mathbb{E}L(A) = \mathbb{E}L(S^*)$ and A is also optimal. Then, we can construct a new subset A_1 with the same rule above. Since there are only finite elements, we can always end up with an optimal set \hat{A} which includes variables with the highest mean values. Then the proof is done.

The only remaining part is to prove $\mathbb{E}L(A) \leq \mathbb{E}L(S^*)$. Denote

$$\begin{aligned} \sum_{i=1}^k p_i &= D - 1/2 + \delta_x \\ \sum_{i \neq j, i=1}^{m+1} p_i &= D - 1/2 + \delta_y \end{aligned}$$

By construction of k , $\delta_x \in (0, p_k]$. By Lemma 6, $\delta_y \in [0, p_{m+1}]$. So we have

$$\sum_{i=1}^k p_i - \sum_{i \neq j, i=1}^{m+1} p_i = p_j - \sum_{i=k+1}^{m+1} p_i = \delta_x - \delta_y$$

Now let's do some simple algebra and try to prove $\mathbb{E}L(A) - \mathbb{E}L(S^*) \leq 0$. Basically, we are trying to write $\mathbb{E}L(A) - \mathbb{E}L(S^*)$ with δ_x, δ_y defined above and then bound it using bounds of δ_x, δ_y .

$$\begin{aligned} \mathbb{E}L(A) - \mathbb{E}L(S^*) &= \left(\sum_{i=1}^k p_i - D \right)^2 + \\ &\sum_{i=1}^k p_i(1-p_i) - \left(\sum_{i \neq j, i=1}^{m+1} p_i - D \right)^2 - \sum_{i \neq j, i=1}^{m+1} p_i(1-p_i) \end{aligned}$$

$$\begin{aligned} &= \left(\sum_{i=1}^k p_i - D + \sum_{i \neq j, i=1}^{m+1} p_i - D \right) \\ &\left(\sum_{i=1}^k p_i - D - \sum_{i \neq j, i=1}^{m+1} p_i + D \right) \\ &+ p_j(1-p_j) - \sum_{i=k+1}^{m+1} p_i(1-p_i) \\ &= (\delta_x + \delta_y - 1)(\delta_x - \delta_y) + p_j - \sum_{i=k+1}^{m+1} p_i + \sum_{i=k+1}^{m+1} p_i^2 - p_j^2 \\ &= (\delta_x + \delta_y - 1)(\delta_x - \delta_y) + \delta_x - \delta_y + \sum_{i=k+1}^{m+1} p_i^2 - p_j^2 \\ &= (\delta_x + \delta_y)(\delta_x - \delta_y) + \sum_{i=k+1}^{m+1} p_i^2 - p_j^2 \\ &= (\delta_x + \delta_y)(\delta_x - \delta_y) + \sum_{i=k+1}^{m+1} p_i^2 - \left(\sum_{i=k+1}^{m+1} p_i + \delta_x - \delta_y \right)^2 \\ &= (\delta_x + \delta_y)(\delta_x - \delta_y) + \sum_{i=k+1}^{m+1} p_i^2 \\ &- \left[\left(\sum_{i=k+1}^{m+1} p_i \right)^2 + (\delta_x - \delta_y)^2 + 2(\delta_x - \delta_y) \sum_{i=k+1}^{m+1} p_i \right] \\ &= (\delta_x - \delta_y)(\delta_x + \delta_y - \delta_x + \delta_y - 2 \sum_{i=k+1}^{m+1} p_i) \\ &+ \sum_{i=k+1}^{m+1} p_i^2 - \left(\sum_{i=k+1}^{m+1} p_i \right)^2 \\ &= (\delta_x - \delta_y)(2\delta_y - 2 \sum_{i=k+1}^{m+1} p_i) + \sum_{i=k+1}^{m+1} p_i^2 - \left(\sum_{i=k+1}^{m+1} p_i \right)^2 \end{aligned} \tag{A.1}$$

Now, we first notice that $\delta_y \leq p_{m+1} \leq \sum_{i=k+1}^{m+1} p_i$, so $(2\delta_y - 2 \sum_{i=k+1}^{m+1} p_i) \leq 0$.

Also notice that

$$\sum_{i=k+1}^{m+1} p_i^2 - \left(\sum_{i=k+1}^{m+1} p_i \right)^2 \leq 0$$

since $p_i \geq 0$.

Case 1: $\delta_x \geq \delta_y$. In this case, (A.1) ≤ 0 is straightforward.

Case 2: $\delta_x < \delta_y$. In this case, $p_{m+1} < p_j < \sum_{i=k+1}^{m+1} p_i$. So we must have $m - k \geq 1$. Since $(2\delta_y - 2 \sum_{i=k+1}^{m+1} p_i) \leq$

0, we can decrease δ_x to 0

$$(A.1) \leq -\delta_y(2\delta_y - 2 \sum_{i=k+1}^{m+1} p_i) + \sum_{i=k+1}^{m+1} p_i^2 - \left(\sum_{i=k+1}^{m+1} p_i \right)^2$$

$$= RHS$$

RHS is a quadratic function with respect to δ_y and it is increasing in the region $[0, \frac{\sum_{i=k+1}^{m+1} p_i}{2}]$. Since $m - k \geq 1$, we have

$$\frac{\sum_{i=k+1}^{m+1} p_i}{2} \geq (p_{m+1} + p_m)/2 \geq p_{m+1} \geq \delta_y$$

So the highest possible value is reached when $\delta_y = p_{m+1}$. Plugging this in RHS, we have

RHS

$$\leq -p_{m+1}(2p_{m+1} - 2 \sum_{i=k+1}^{m+1} p_i) + \sum_{i=k+1}^{m+1} p_i^2 - \left(\sum_{i=k+1}^{m+1} p_i \right)^2$$

$$= -2p_{m+1}^2 + 2 \left(\sum_{i=k+1}^{m+1} p_i \right) p_{m+1} + \sum_{i=k+1}^{m+1} p_i^2 - \left(\sum_{i=k+1}^{m+1} p_i \right)^2$$

$$= \left(\sum_{i=k+1}^{m+1} p_i \right) (p_{m+1} - \sum_{i=k+1}^m p_i) - p_{m+1}^2 + \sum_{i=k}^m p_i^2$$

$$= (p_{m+1} + \sum_{i=k}^m p_i) (p_{m+1} - \sum_{i=k+1}^m p_i) - p_{m+1}^2 + \sum_{i=k}^m p_i^2$$

$$= p_{m+1}^2 - \left(\sum_{i=k}^m p_i \right)^2 - p_{m+1}^2 + \sum_{i=k}^m p_i^2$$

$$= \sum_{i=k}^m p_i^2 - \left(\sum_{i=k}^m p_i \right)^2 \leq 0 \quad (\text{Since } p_i \geq 0)$$

Thus we have shown that

$$\mathbb{E}L(A) - \mathbb{E}L(S^*) \leq 0$$

□

Now, given Lemma 6 and Lemma 7, we can prove Theorem 1.

Proof of theorem 1: When $D < 1/2$ or $\sum_{i=1}^n p_i \leq D - 1/2$, see Corollary 2 and Corollary 3.

When $D \geq 1/2$ and $\sum_{i=1}^n p_i > D - 1/2$, by Lemma 7, there exists an optimal subset $S^* = \{1, \dots, k\}$ for some k , i.e. containing the first several arms with largest mean values. Since S^* is optimal, we must have, by Lemma 6, that

$$\sum_{i=1}^k p_i \geq D - 1/2$$

$$\sum_{i=1}^{k-1} p_i \leq D - 1/2$$

If $\sum_{i=1}^k p_i > D - 1/2$, then S^* is the output of Algorithm 1, so the output of Algorithm 1 is optimal.

If $\sum_{i=1}^k p_i = D - 1/2$, then it is easy to show that $\mathbb{E}L(\{1, \dots, k\}) = \mathbb{E}L(\{1, \dots, k+1\})$. So $\{1, \dots, k+1\}$ is also optimal. The output of Algorithm 1 is $\{1, \dots, k+1\}$. So the output of Algorithm 1 is still optimal. □

Corollary 4 If $\sum_{i=1}^{k-1} p_i = D - 1/2$, then the subset $\{1, \dots, k-1\}$ and $\{1, \dots, k\}$ are both optimal.

B Proof of Proposition 1 and Lemma 3 with Assumption (A1) and (A2) and proof of Proposition 1

We will prove Lemma 3 by using ϵ_0 's expression in Proposition 1 under Assumption (A1) and (A2). The proof of Proposition 1 follows naturally.

Before the proof, we provide some properties of the subset S_t selected by CUCB-Avg.

Lemma 8 (Properties of CUCB-Avg's Selection) Conditioning on any realization of the history by time $t-1$: $\mathcal{F}_{t-1} := \{S_\tau, \{X_{\tau,i}\}_{i \in S_\tau}\}_{\tau=1}^{t-1}$, if $S_t = s$ for some deterministic subset $s \subseteq [n]$, we have the following:

- i) $\sum_{i \in s} \bar{p}_i > D - 1/2$ or $s = [n]$
- ii) Define $i_{\min} \in \arg \min_{i \in s} U_i(t)$, then $\sum_{i \in s - \{i_{\min}\}} \bar{p}_i \leq D - 1/2$
- iii) For any $j \in s$ and any $i \in [n]$ such that $U_i(t) > U_j(t)$, $i \in s$.

Proof. The proof is straightforward by Algorithm 2. □

Proof of Lemma 3 given Assumption (A1) and (A2):

Let $S^* = \phi(p, D) = \{\sigma(1), \dots, \sigma(k)\}$. We will prove $S_t = S^*$ in Step 1-3 then prove zero regret in Step 4. The proof relies on ϵ_0 's expression in Proposition 1.

Step 1: Given $\bar{E}_t, \bar{B}_t(\epsilon_0)$, either $S_t \subseteq S^*$ or $S^* \subseteq S_t$: We prove this by contradiction. Suppose there is a realization of \mathcal{F}_{t-1} such that \bar{E}_t and $\bar{B}_t(\epsilon_0)$ hold, but $S_t \not\subseteq S^*$ and $S^* \not\subseteq S_t$, then there exists $j \in S_t - S^*$, $i \in S^* - S_t$ conditioning on this \mathcal{F}_{t-1} . We will show $U_i(t) > U_j(t)$ in the following:

$$U_i(t) = \min(\bar{p}_i(t-1) + \sqrt{\frac{\alpha \log t}{2T_i(t-1)}}, 1)$$

$$\geq p_i \quad (\text{by } \bar{E}_t)$$

$$\geq p_{\sigma(k)} \quad (\text{by } i \in S^*, \sigma(k) \text{ is the smallest one})$$

$$\begin{aligned}
U_j(t) &\leq \bar{p}_j(t-1) + \sqrt{\frac{\alpha \log t}{2T_j(t-1)}} && \text{(by def.)} \\
&< p_j + 2\sqrt{\frac{\alpha \log t}{2T_j(t-1)}} && \text{(by definition of } \bar{E}_t) \\
&< p_j + 2\epsilon_0 && \text{(by } j \in S_t, \text{ and definition of } \bar{B}_t(\epsilon_0)) \\
&\leq p_{\sigma(k+1)} + 2\epsilon_0 && \text{(by } j \notin S^*, \sigma(k+1) \text{ is the largest)} \\
&\leq p_{\sigma(k+1)} + 2\frac{p_{\sigma(k)} - p_{\sigma(k+1)}}{2} && \\
&\leq p_{\sigma(k)} \leq U_i(t) && \text{(by definition } \epsilon_0 \leq \frac{\Delta_k}{2})
\end{aligned}$$

Therefore, $U_i(t) > U_j(t)$. By Lemma 8 (iii), we have $i \in S_t$, which leads to a contradiction.

Step 2: Given $\bar{E}_t, \bar{B}_t(\epsilon_0), S_t \subsetneq S^*$ is impossible: We prove this by contradiction. Suppose $S_t \subsetneq S^*$, then $S_t \neq [n]$. We can show that

$$\begin{aligned}
&\sum_{i \in S_t} \bar{p}_i(t-1) \\
&< \sum_{i \in S_t} \left(p_i + \sqrt{\frac{\alpha \log t}{2T_i(t-1)}} \right) && \text{(by } \bar{E}_t) \\
&< \sum_{i \in S_t} (p_i + \epsilon_0) && \text{(by } \bar{B}_t(\epsilon_0)) \\
&\leq \sum_{i=1}^{k-1} (p_{\sigma(i)} + \epsilon_0) && \text{(by } S_t \subsetneq S^*) \\
&\leq D - 1/2 - \delta_2 + (k-1)\epsilon_0 && \text{(by definition of } \delta_2) \\
&\leq D - 1/2 && \text{(by definition of } \epsilon_0, \epsilon_0 \leq \frac{\delta_2}{k})
\end{aligned}$$

However, by Lemma 8 (i), $\sum_{i \in S_t} \bar{p}_i(t-1) > D - 1/2$, which leads to a contradiction.

Step 3: Given $\bar{E}_t, \bar{B}_t(\epsilon_0), S^* \subsetneq S_t$ is impossible: We prove this by contradiction. Suppose $S^* \subsetneq S_t$, so $S^* \neq [n]$, so $\sum_{i=1}^n p_i > D - 1/2$. We will show that $i_{\min} \in \arg \min_{i \in S_t} U_i(t) \subseteq S_t - S^*$, and $\sum_{i \in S_t - \{i_{\min}\}} \bar{p}_i(t-1) > D - 1/2$. Then by Lemma 8 (ii), we have a contradiction.

Now first we show that $i_{\min} \in \arg \min_{i \in S_t} U_i(t) \subseteq S_t - S^*$. It suffices to show that for any $i \in S^*$, and $j \in S_t - S^*$, $U_i(t) > U_j(t)$. This is proved by the following.

$$\begin{aligned}
U_i(t) &\geq p_i \geq p_{\sigma(k)} && \text{(by } \bar{E}_t \text{ and } i \in S_t) \\
U_j(t) &< p_j + 2\epsilon_0 && \text{(by } \bar{E}_t \text{ and } \bar{B}_t(\epsilon_0)) \\
&\leq p_{\sigma(k+1)} + 2\frac{\Delta_k}{2} && \text{(by } j \notin S^* \text{ and def. of } \epsilon_0) \\
&= p_{\sigma(k)}
\end{aligned}$$

Then we show $\sum_{i \in S_t - \{i_{\min}\}} \bar{p}_i(t-1) > D - 1/2$ by

$$\begin{aligned}
&\sum_{i \in S_t - \{i_{\min}\}} \bar{p}_i(t-1) \\
&\geq \sum_{i \in S^*} \bar{p}_i(t-1) && ((i_{\min} \in S_t - S^*, S_t \supsetneq S^*)) \\
&> \sum_{i \in S^*} (p_i - \epsilon_0) && \text{(by } \bar{E}_t) \\
&= D - 1/2 + \delta_1 - k\epsilon_0 \\
&\geq D - 1/2 && \text{(def of } \epsilon_0)
\end{aligned}$$

Step 4: Prove $\mathbb{E} R_t(S_t) I_{\{\bar{E}_t, \bar{B}_t(\epsilon_0)\}} = 0$. By the three steps above, we have $S_t = S^*$ under $\bar{E}_t, \bar{B}_t(\epsilon_0)$, then it is straightforward that

$$\begin{aligned}
\mathbb{E} R_t(S_t) I_{\{\bar{E}_t, \bar{B}_t(\epsilon_0)\}} &= \mathbb{E} R_t(S_t) I_{\{\bar{E}_t, \bar{B}_t(\epsilon_0), S_t = S^*\}} \\
&= \mathbb{E} R_t(S^*) I_{\{\bar{E}_t, \bar{B}_t(\epsilon_0), S_t = S^*\}} = 0
\end{aligned}$$

□

C A general expression of ϵ_0 and Proof of Lemma 3 without Assumption (A1) and (A2)

Without Assumption (A1) and (A2), we need additional technical discussion because there might be multiple optimal subsets due to ties in the p_i and Corollary 4. But the main idea behind the proof is the same.

We will first give an explicit expression of ϵ_0 , then we will show $\mathbb{E} R_t(S_t) I_{\{\bar{E}_t, \bar{B}_t(\epsilon_0)\}} = 0$ given the new ϵ_0 .

Without loss of generality, we will consider $D > 1/2$ due to Corollary 2 and 4.

Now we present the expression of ϵ_0 in the general case.

Definition 1 Let $p_{\sigma(1)} \geq \dots \geq p_{\sigma(n)}$. The ϵ_0 in Lemma 3 can be determined by

$$\epsilon_0 = \min\left(\frac{\delta_1}{l_1}, \frac{\delta_2}{l_2}, \frac{\Delta_{k_1}}{2}, \frac{\Delta_{k_2}}{2}\right)$$

In the following, we define each parameter respectively.

Definition of l_1 :

$$l_1 = \begin{cases} \min G_1 & \text{if } \sum_{i=1}^n p_i > D - 1/2 \\ n & \text{if } \sum_{i=1}^n p_i \leq D - 1/2 \end{cases}$$

where

$$G_1 = \{1 \leq k \leq n \mid \sum_{i=1}^k p_{\sigma(i)} > D - 1/2\}$$

The offline algorithm output is $\phi(p, D) = \{\sigma(1), \dots, \sigma(l_1)\}$. CUCB-Avg's output can be $\{\sigma(1), \dots, \sigma(l_1)\}$ (there

might be other outputs due to the random tie-breaking rule).

Definition of δ_1 :

$$\delta_1 = \begin{cases} \sum_{i=1}^{l_1} p_{\sigma(i)} - (D - 1/2) & \text{if } l_1 < n \\ n & \text{if } l_1 = n \end{cases}$$

Note that when $l_1 = n$, we let $\delta_1 = n$, so that $\frac{\delta_1}{l_1} = 1$ is large enough to not affect the definition of ϵ_0 .

Definition of l_2 :

$$l_2 = \max\{0 \leq k < l_1 \mid \sum_{i=1}^k p_{\sigma(i)} < D - 1/2\}$$

When $D - 1/2 = 0$, define $l_2 = 0$. Note that $0 \leq l_2 \leq n - 1$, and $l_1 > l_2$.

Definition of δ_2 :

$$\delta_2 = \begin{cases} (D - 1/2) - \sum_{i=1}^{l_2} p_{\sigma(i)} & \text{if } l_2 \geq 1 \\ 1 & \text{if } l_2 = 0 \end{cases}$$

Note that when $l_2 = 0$, let $\delta_2 = 1$, then $\frac{\delta_2}{l_2} = +\infty$, which is large enough to not affect the definition of ϵ_0

Definition of Δ_i :

$$\Delta_i = \begin{cases} p_{\sigma(i)} - p_{\sigma(i+1)} & \text{if } 1 \leq i \leq n - 1 \\ 2 & \text{if } i = 0, n \end{cases}$$

Note that when $i = 0, n$, we let $\Delta_i = 2$ which is large enough to keep the ϵ_0 unaffected.

Definition of k_1 :

$$k_1 = \max\{0 \leq i \leq l_1 - 1 \mid \Delta_i > 0\}$$

Intuitively, k_1 is the largest index under the non-increasing order such that the parameter $p_{\sigma(k_1)}$ is the second smallest among $p_{\sigma(1)}, \dots, p_{\sigma(l_1)}$.

Definition of k_2 :

$$k_2 = \min\{l_1 \leq i \leq n \mid \Delta_i > 0\}$$

Intuitively, k_2 is the largest index under the non-increasing order that can be possibly selected by the offline optimization algorithm.

In the following, we will first prove some supportive lemmas, then provide a proof of Lemma 3.

Lemma 9 For any $\epsilon > 0$ satisfying $\epsilon \leq \Delta_k/2$ for some k with $0 \leq k \leq n$, given \bar{E}_t and $\bar{B}_t(\epsilon)$, then we have either $S_t \subseteq \{\sigma(1), \dots, \sigma(k)\}$ or $\{\sigma(1), \dots, \sigma(k)\} \subseteq S_t$.

Proof. When $k = 0, n$, it is trivially true.

When $1 \leq k \leq n - 1$. Suppose there is a realization of \mathcal{F}_{t-1} such that \bar{E}_t and $\bar{B}_t(\epsilon)$ hold, but $S_t \not\subseteq \{\sigma(1), \dots, \sigma(k)\}$ and $\{\sigma(1), \dots, \sigma(k)\} \not\subseteq S_t$. Fix this realization of \mathcal{F}_{t-1} , then there exists $j \in S_t - \{\sigma(1), \dots, \sigma(k)\}$, $i \in \{\sigma(1), \dots, \sigma(k)\} - S_t$. We will show that $U_i(t) > U_j(t)$, then by Lemma 8 (iii), we have $i \in S_t$ which leads to contradiction.

$$U_i(t) = \min(\bar{p}_i(t-1) + \sqrt{\frac{\alpha \log t}{2T_i(t-1)}}, 1) \geq p_i \geq p_{\sigma(k)}$$

$$\begin{aligned} U_j(t) &\leq \bar{p}_j(t-1) + \sqrt{\frac{\alpha \log t}{2T_j(t-1)}} \\ &< p_j + 2\sqrt{\frac{\alpha \log t}{2T_j(t-1)}} < p_j + 2\epsilon \\ &\leq p_{\sigma(k+1)} + 2\epsilon \leq p_{\sigma(k+1)} + 2\frac{p_{\sigma(k)} - p_{\sigma(k+1)}}{2} \\ &= p_{\sigma(k)} \leq U_i(t) \end{aligned}$$

□

Corollary 5 Given \bar{E}_t and $\bar{B}_t(\epsilon_0)$, then we have either $S_t \not\subseteq \{\sigma(1), \dots, \sigma(k_1)\}$ or $\{\sigma(1), \dots, \sigma(k_1)\} \subseteq S_t \subseteq \{\sigma(1), \dots, \sigma(k_2)\}$, or $\{\sigma(1), \dots, \sigma(k_2)\} \not\subseteq S_t$.

Proof. It is easy to see that $k_1 < k_2$ by definition, and by $\epsilon_0 \leq \Delta_{k_1}/2$, $\epsilon_0 \leq \Delta_{k_2}/2$, then by Lemma 9, it is straightforward. □

Finally, we are ready to prove Lemma 3.

Proof of Lemma 3: The major part of the proof is to show that if $\bar{E}_t, \bar{B}_t(\epsilon_0)$ happen, then S_t must be optimal. Then, given that S_t is optimal, it is easy to prove zero regret at time t .

Now, let's first prove $S_t \in \mathbb{S}^*$ given $\bar{E}_t, \bar{B}_t(\epsilon_0)$, where \mathbb{S}^* denotes the set of all possible optimal subsets.

Step 1: prove \bar{E}_t and $\bar{B}_t(\epsilon_0) \Rightarrow S_t \in \mathbb{S}^*$. We will list all possible scenarios, and prove that in each scenario, when \bar{E}_t and $\bar{B}_t(\epsilon_0)$ hold, we have $S_t \in \mathbb{S}^*$.

Scenario 1: When $l_1 > l_2 + 2$, we will first show that \mathbb{S}^* contains any set S that satisfies $S \supseteq \{\sigma(1), \dots, \sigma(l_2 + 1)\}$. To prove this, notice that by the definition of l_1, l_2 , when $l_1 > l_2 + 2$, we have $\sum_{i=1}^{l_2+1} p_{\sigma(i)} = \sum_{i=1}^{l_2+2} p_{\sigma(i)} = D - 1/2$, so $p_{\sigma(l_2+2)} = 0$. Accordingly, $p_{\sigma(i)} = 0$ for all $i \geq l_2 + 2$. This leads to $l_1 = n$. Moreover, by Corollary 4, $\{\sigma(1), \dots, \sigma(l_2 + 1)\}$ is optimal. Since the union of set $\{\sigma(1), \dots, \sigma(l_2 + 1)\}$ and some arms with zero probabilities is also optimal, any set with a subset $\{\sigma(1), \dots, \sigma(l_2 + 1)\}$ is also optimal.

Next, we will show S_t is optimal. It is easy to verify that $k_1 = l_2 + 1$ by the definition of k_1, l_2 . By Corollary 5, we have either $S_t \subsetneq \{\sigma(1), \dots, \sigma(l_2 + 1)\}$ or $\{\sigma(1), \dots, \sigma(l_2 + 1)\} \subseteq S_t \subseteq \{\sigma(1), \dots, \sigma(n)\}$. Since the second possible case guarantees $S_t \in \mathbb{S}^*$, we only need to show that $S_t \subsetneq \{\sigma(1), \dots, \sigma(l_2 + 1)\}$ is impossible. This is done by contradiction. Suppose $S_t \subsetneq \{\sigma(1), \dots, \sigma(l_2 + 1)\}$, then

$$\begin{aligned} \sum_{i \in S_t} \bar{p}_i(t-1) &\leq \sum_{i=1}^{l_2} \bar{p}_{\sigma(i)}(t-1) \\ &\leq \sum_{i=1}^{l_2} p_{\sigma(i)} + \epsilon_0 \quad (\text{by } \bar{E}_t, \bar{B}_t(\epsilon_0)) \\ &= \sum_{i=1}^{l_2} p_{\sigma(i)} + l_2 \epsilon_0 < D - 1/2 \quad (\text{by def. of } l_2, \epsilon_0) \end{aligned}$$

By Lemma 8 (i), this leads to a contradiction. Therefore, S_t is optimal in this scenario.

Scenario 2: When $l_1 = l_2 + 1$, we will first show that \mathbb{S}^* contains any set that contains and only contains $\sigma(1), \dots, \sigma(k_1)$ together with $l_1 - k_1$ arms from subset $\{\sigma(k_1 + 1), \dots, \sigma(k_2)\}$. To prove this, notice that by definition of l_1 and Theorem 1, $\phi(p, D) = \{\sigma(1), \dots, \sigma(l_1)\}$ is optimal. Then, by definition of k_1, k_2 , we have $k_1 + 1 \leq l_1 \leq k_2$ and the arms $\{\sigma(k_1 + 1), \dots, \sigma(k_2)\}$ are in a tie with probability $p_{\sigma(l_1)}$. Moreover, there are $l_1 - k_1$ arms in $\phi(p, D)$ whose value is $p_{\sigma(l_1)}$. Therefore, replacing these $l_1 - k_1$ arms with any $l_1 - k_1$ arms with the same value will still yield an optimal subset. This proves that any set is optimal if it contains $\sigma(1), \dots, \sigma(k_1)$ together with $l_1 - k_1$ arms from subset $\{\sigma(k_1 + 1), \dots, \sigma(k_2)\}$.

Next, we will show S_t is optimal. This is done by proving the following three statements: (a) $S_t \supsetneq \{\sigma(1), \dots, \sigma(k_1)\}$, (b) $S_t \subseteq \{\sigma(1), \dots, \sigma(k_2)\}$, and (c) S_t has exactly $l_1 - k_1$ arms with parameter $p_{\sigma(k_1 + 1)}$.

Firstly, we prove statement (a) by contradiction. Suppose $S_t \subseteq \{\sigma(1), \dots, \sigma(k_1)\}$,

$$\begin{aligned} \sum_{i \in S_t} \bar{p}_i(t-1) &\leq \sum_{i=1}^{k_1} \bar{p}_{\sigma(i)}(t-1) \\ &\leq \sum_{i=1}^{k_1} p_{\sigma(i)} + k_1 \epsilon_0 \quad (\text{by } \bar{E}_t, \bar{B}_t(\epsilon_0)) \\ &\leq \sum_{i=1}^{l_2} p_{\sigma(i)} + l_2 \epsilon_0 \quad (\text{by def. } k_1 \leq l_1 - 1 = l_2) \\ &= D - 1/2 - \delta_1 + l_2 \epsilon_0 \leq D - 1/2 \\ &\quad (\text{by def. of } l_2, \delta_2, \epsilon_0) \end{aligned}$$

Secondly, we prove statement (b) by contradiction. Suppose $S_t \supsetneq \{\sigma(1), \dots, \sigma(k_2)\}$, then $i_{min} \in$

$\arg \min_{i \in S_t} U_i(t) \subseteq S_t - \{\sigma(1), \dots, \sigma(k_2)\}$. This is because for any $i \in \{\sigma(1), \dots, \sigma(k_2)\}$, and $j \in S_t - \{\sigma(1), \dots, \sigma(k_2)\}$,

$$\begin{aligned} U_i(t) &\geq p_i \geq p_{\sigma(k_2)} && (\text{by } \bar{E}_t) \\ U_j(t) &< p_j + 2\epsilon_0 && (\text{by } \bar{E}_t \text{ and } \bar{B}_t(\epsilon_0)) \\ &\leq p_{\sigma(k_2+1)} + 2\frac{\Delta_{k_2}}{2} \\ &= p_{\sigma(k_2)} \leq U_i(t) \end{aligned}$$

so $U_i(t) > U_j(t)$. As a result, $i_{min} \in \arg \min_{i \in S_t} U_i(t) \notin \{\sigma(1), \dots, \sigma(k_2)\}$. Moreover, we have

$$\begin{aligned} \sum_{i \in S_t - \{i_{min}\}} \bar{p}_i(t-1) &\geq \sum_{i=1}^{k_2} \bar{p}_{\sigma(i)}(t-1) \\ &\geq \sum_{i=1}^{l_1} \bar{p}_{\sigma(i)}(t-1) > \sum_{i=1}^{l_1} p_{\sigma(i)} - l_1 \epsilon_0 \\ &= D - 1/2 + \delta_1 - l_1 \epsilon_0 \geq D - 1/2 \end{aligned}$$

This leads to a contradiction by Lemma 8 (ii). Therefore, statement (b) is true.

Thirdly, we prove statement (c). By using similar proof techniques, it is straightforward to show that if we select more or less than $l_1 - k_1$ arms, the sum of parameters is either more than $D - 1/2$ excluding i_{min} or less than $D - 1/2$, which leads to contradiction by Lemma 8 (ii) and (1). Therefore, statement (c) is true.

In conclusion, S_t is optimal in this scenario.

Scenario 3: When $l_1 = l_2 + 2$, it can be shown that \mathbb{S}^* contains subsets that contains and only contains $\sigma(1), \dots, \sigma(k_1)$ together with either $l_1 - k_1$ or $l_1 - 1 - k_1$ arms with parameter $p_{\sigma(l_1)}$. To prove this, notice that by $l_1 = l_2 + 2$, $\sum_{i=1}^{l_1-1} p_{\sigma(i)} = D - 1/2$. Hence by Corollary 4, both $\{\sigma(1), \dots, \sigma(l_1)\}$ and $\{\sigma(1), \dots, \sigma(l_1 - 1)\}$ are optimal. In addition, by definition of k_1, k_2 , we have $k_1 + 1 \leq l_1 \leq k_2$ and the arms $\{\sigma(k_1 + 1), \dots, \sigma(k_2)\}$ are in a tie with probability $p_{\sigma(l_1)}$. Moreover, there are $l_1 - k_1$ and $l_1 - k_1 - 1$ arms with probability $p_{\sigma(l_1)}$ in $\{\sigma(1), \dots, \sigma(l_1)\}$ and $\{\sigma(1), \dots, \sigma(l_1 - 1)\}$ respectively. Therefore, replacing these arms by arms with the same probability will still yield an optimal subset. This proves that any set is optimal if it contains $\sigma(1), \dots, \sigma(k_1)$ together with $l_1 - k_1$ or $l_1 - k_1 - 1$ arms from subset $\{\sigma(k_1 + 1), \dots, \sigma(k_2)\}$.

Next, we will prove S_t is optimal by showing a) $S_t \supsetneq \{\sigma(1), \dots, \sigma(k_1)\}$, b) $S_t \subseteq \{\sigma(1), \dots, \sigma(k_2)\}$, and c) S_t has either $l_1 - k_1$ or $l_1 - 1 - k_1$ arms with parameter $p_{\sigma(l_1)}$.

Firstly, we prove statement (a) by contradiction. The proof is very similar to that for Scenario 2. Suppose $S_t \subsetneq$

$\{\sigma(1), \dots, \sigma(k_1)\}$. By $k_1 \leq l_1 - 1 = l_2 + 1$,

$$\begin{aligned} \sum_{i \in S_t} p_i &\leq \sum_{i=1}^{k_1-1} \bar{p}_{\sigma(i)} \leq \sum_{i=1}^{k_1-1} p_{\sigma(i)} \\ &\leq \sum_{i=1}^{l_2} p_{\sigma(i)} + l_2 \epsilon_0 \leq D - 1/2 \end{aligned}$$

This contradicts Lemma 8 (i).

We can also prove statement (b) and (c) using the same proof in Scenario 2. In conclusion, S_t is optimal in this scenario.

Step 2: prove $\mathbb{E} R_t(S_t) I_{\{\bar{E}_t, \bar{B}_t(\epsilon_0)\}} = 0$. Notice that \bar{E}_t and $\bar{B}_t(\epsilon_0)$ are determined by \mathcal{F}_{t-1} . Consider \mathbb{S}^* to be the set of all optimal subsets,

$$\begin{aligned} \mathbb{E} I_{\{\bar{E}_t, \bar{B}_t(\epsilon_0)\}} R_t(S_t) &= \mathbb{E} I_{\{\bar{E}_t, \bar{B}_t(\epsilon_0), S_t \in \mathbb{S}^*\}} R_t(S_t) \\ &= \sum_{s \in \mathbb{S}^*} \mathbb{E} I_{\{\bar{E}_t, \bar{B}_t(\epsilon_0), S_t = s\}} R_t(S_t) \\ &= \sum_{s \in \mathbb{S}^*} \mathbb{P}(\bar{E}_t, \bar{B}_t(\epsilon_0), S_t = s) \mathbb{E} R_t(s) = 0 \end{aligned}$$

where the second inequality is because $R_t(s)$ and $I_{\{\bar{E}_t, \bar{B}_t(\epsilon_0), S_t = s\}}$ are independent. \square

D Proof of Lemma 5

To illustrate the intuition, we first provide the proof under the assumption that p and D_t satisfies Assumption (A1) and (A2) in this appendix. Then, we will provide a proof without these assumptions based on the same idea in Appendix D.1.

Suppose p and D_t satisfies Assumption (A1) and (A2), the definition of ϵ_1 and $\xi(D_t)$ can be written as

$$\begin{aligned} \epsilon_1 &= \min\left(\frac{\Delta_1}{2}, \dots, \frac{\Delta_T}{2}, \frac{\beta}{n}\right) \\ \xi(D_t) &= \min\left(\frac{\delta_1^t}{l_1^t}, \frac{\delta_2^t}{l_1^t}, \epsilon_1\right) \end{aligned} \quad (\text{D.1})$$

where

$$\begin{aligned} l_1^t &= \lfloor \phi(p, D_t) \rfloor \\ \sum_{i=1}^{l_1^t} p_{\sigma(i)} &= D_t - 1/2 + \delta_1^t, \\ \sum_{i=1}^{l_1^t-1} p_{\sigma(i)} &= D_t - 1/2 - \delta_2^t \end{aligned}$$

when $D_t > 1/2$. When $D_t \leq 1/2$, $\xi(D_t) = \epsilon_1$.

Proof. When $\xi(D_t) \geq \epsilon$, show zero regret. This is because $\epsilon \leq \xi(D_t) \leq \epsilon_0(D_t)$, where $\epsilon_0(D_t)$ denotes the ϵ_0 defined in Definition 1 with respect to D_t . As a result, $E_t \cap \bar{B}_t(\epsilon) \subseteq \bar{E}_t \cap \bar{B}_t(\epsilon_0(D_t))$, so by Lemma 3, there is no regret at t .

When $\xi(D_t) < \epsilon$, show S_t at most differs from the optimal set by one arm. Formally, we will show $S_t \subseteq \{\{\sigma(1), \dots, \sigma(l_1^t - 1)\}, \{\sigma(1), \dots, \sigma(l_1^t)\}, \{\sigma(1), \dots, \sigma(l_1^t + 1)\}\}$ by $\epsilon \leq \epsilon_1$. The proof takes three steps.

Step 1: S_t must satisfy one of the five conditions: i) $S_t \subsetneq \{\sigma(1), \dots, \sigma(l_1^t - 1)\}$, ii) $S_t = \{\sigma(1), \dots, \sigma(l_1^t - 1)\}$, iii) $S_t = \{\sigma(1), \dots, \sigma(l_1^t)\}$, iv) $S_t = \{\sigma(1), \dots, \sigma(l_1^t + 1)\}$, v) $S_t \supsetneq \{\sigma(1), \dots, \sigma(l_1^t + 1)\}$. This is proved by $\epsilon_1 < \frac{\Delta_1^{l_1^t-1}}{2}, \frac{\Delta_1^{l_1^t}}{2}, \frac{\Delta_1^{l_1^t+1}}{2}$ and Lemma 9 in Appendix C.

Step 2: i) is not possible. This is because for any subset $s \subsetneq \{\sigma(1), \dots, \sigma(l_1^t - 1)\}$, the total estimated mean is less than $D_t - 1/2$, so by Lemma 8 (i), $S_t \neq s$.

$$\begin{aligned} &\sum_{i \in s} \bar{p}_i(t-1) \\ &< \sum_{i \in s} \left(p_i + \sqrt{\frac{\alpha \log t}{2T_i(t-1)}} \right) \quad (\text{by } \bar{E}_t) \\ &< \sum_{i \in s} (p_i + \epsilon_1) \quad (\text{by } \bar{B}_t(\epsilon_1)) \\ &\leq \sum_{i=1}^{l_1^t-2} (p_{\sigma(i)} + \epsilon_1) \quad (\text{by } s \subsetneq \{\sigma(1), \dots, \sigma(l_1^t - 1)\}) \\ &= D_t - 1/2 - \delta_2^t - p_{\sigma(l_1^t-1)} + (l_1^t - 2)\epsilon_1 \quad (\text{by def of } \delta_2^t) \\ &\leq D_t - 1/2 \quad (\text{by } \epsilon_1 \leq \frac{\beta}{n} \leq \frac{p_{\sigma(l_1^t-2)}}{l_1^t-2}) \end{aligned}$$

Step 3: v) is not possible. This is proved by contradiction. Suppose (v) is true, then $i_{\min} = \arg \min_{i \in S_t} U_i(t)$ must be in $S_t - \{\sigma(1), \dots, \sigma(l_1^t + 1)\}$ because by the definition of \bar{E}_t and $\bar{B}_t(\epsilon_1)$, CUCB-Avg must rank the arms in S_t in the correct order. We can show that $\sum_{i \in S_t - \{i_{\min}\}} p_i > D_t/2$ below:

$$\begin{aligned} &\sum_{i=1}^{l_1^t} \bar{p}_{\sigma(i)}(t-1) \\ &> \sum_{i=1}^{l_1^t} \left(p_{\sigma(i)} - \sqrt{\frac{\alpha \log t}{2T_i(t-1)}} \right) \quad (\text{by } \bar{E}_t) \\ &> \sum_{i=1}^{l_1^t} (p_{\sigma(i)} - \epsilon_1) \quad (\text{by } \bar{B}_t(\epsilon_1)) \\ &= D_t - 1/2 + \delta_1^t + p_{\sigma(l_1^t+1)} - (l_1^t + 1)\epsilon_1 \quad (\text{by def of } \delta_1^t) \\ &\geq D_t - 1/2 \quad (\text{by } p_i \geq \beta \geq n\epsilon_1) \end{aligned}$$

By Lemma 8 (ii), there is a contradiction. Therefore, v) is not true.

When $\xi(D_t) < \epsilon$, show $\mathbb{E} R_t(s) \leq 2n\epsilon$. We only need to discuss condition ii), iii) and iv). Notice that in condition iii), the regret is 0.

In condition ii),

$$\begin{aligned}
\mathbb{E} R_t(s) &= \left(\sum_{i=1}^{l_1^t-1} p_{\sigma(i)} - D_t \right)^2 \\
&\quad - \left(\sum_{i=1}^{l_1^t} p_{\sigma(i)} - D_t \right)^2 - p_{\sigma(l_1^t)}(1 - p_{\sigma(l_1^t)}) \\
&= -p_{\sigma(l_1^t)} \left(2 \sum_{i=1}^{l_1^t-1} p_{\sigma(i)} - 2D_t + p_{\sigma(l_1^t)} \right) \\
&\quad - p_{\sigma(l_1^t)}(1 - p_{\sigma(l_1^t)}) \\
&= -p_{\sigma(l_1^t)} \left(2 \sum_{i=1}^{l_1^t-1} p_{\sigma(i)} - 2D_t + 1 \right) \\
&= 2p_{\sigma(l_1^t)} \delta_2^t \leq 2\delta_2^t \\
&\leq 2n\epsilon
\end{aligned}$$

where the last inequality is by $\delta_t^2 < (l_1^t - 1)\epsilon$ which can be proved by contradiction. Suppose $\delta_t^2 \geq (l_1^t - 1)\epsilon$, then $s = \{\sigma(1), \dots, \sigma(l_1^t - 1)\} \in \mathbb{S}_t$ is not possible because the total estimated value is no more than $D_t - 1/2$.

$$\begin{aligned}
&\sum_{i=1}^{l_1^t-1} \bar{p}_{\sigma(i)}(t-1) \\
&\leq \sum_{i=1}^{l_1^t-1} (p_{\sigma(i)} + \epsilon) \quad (\text{by def of } \bar{E}_t \text{ and } \bar{B}_t(\epsilon)) \\
&= D_t - 1/2 - \delta_2^t + (l_1^t - 1)\epsilon \quad (\text{by def of } \delta_2^t) \\
&\leq D_t - 1/2 \quad (\text{by } \delta_t^2 \geq (l_1^t - 1)\epsilon)
\end{aligned}$$

In condition iv),

$$\begin{aligned}
\mathbb{E} R_t(s) &= \left(\sum_{i=1}^{l_1^t+1} p_{\sigma(i)} - D_t \right)^2 \\
&\quad - \left(\sum_{i=1}^{l_1^t} p_{\sigma(i)} - D_t \right)^2 + p_{\sigma(l_1^t+1)}(1 - p_{\sigma(l_1^t+1)}) \\
&= p_{\sigma(l_1^t+1)} \left(2 \sum_{i=1}^{l_1^t} p_{\sigma(i)} - 2D_t + p_{\sigma(l_1^t+1)} \right) \\
&\quad + p_{\sigma(l_1^t+1)}(1 - p_{\sigma(l_1^t+1)}) \\
&= p_{\sigma(l_1^t+1)} \left(2 \sum_{i=1}^{l_1^t} p_{\sigma(i)} - 2D_t + 1 \right)
\end{aligned}$$

$$= 2p_{\sigma(l_1^t)} \delta_1^t \leq 2\delta_1^t \leq 2n\epsilon$$

where the last inequality is by $\delta_t^1 < l_1^t \epsilon$ which can be proved by contradiction. Suppose $\delta_t^1 \geq l_1^t \epsilon$, then iv) is not possible because CUCB-Avg will terminate at l_1^t :

$$\begin{aligned}
&\sum_{i=1}^{l_1^t} \bar{p}_{\sigma(i)}(t-1) \\
&> \sum_{i=1}^{l_1^t} \left(p_{\sigma(i)} - \sqrt{\frac{\alpha \log t}{2T_i(t-1)}} \right) \quad (\text{by } \bar{E}_t) \\
&> \sum_{i=1}^{l_1^t} (p_{\sigma(i)} - \epsilon) \quad (\text{by } \bar{B}_t(\epsilon)) \\
&= D_t - 1/2 + \delta_1^t - l_1^t \epsilon \quad (\text{by def of } \delta_1^t) \\
&\geq D_t - 1/2 \quad (\text{by } \delta_t^1 \geq l_1^t \epsilon)
\end{aligned}$$

□

D.1 Proof of Lemma 5 without additional assumptions

In this appendix, we provide a proof of Lemma 5 without additional assumption (A1) (A2).

Before the proof, we note that adding or deleting a zero-valued arm from the selected subset s will not affect the regret $\mathbb{E} R_t(s)$. Therefore, without loss of generality, we will focus on S_t without zero-valued arms, and assume all subsets in \mathbb{S}_t do not consist of any zero-valued arm.

In addition, we note that zero regret under $\xi(D_t) > \epsilon_1$ can be proved in the same way as Lemma 3. Therefore, we only need to focus on $\xi(D_t) \leq \epsilon_1$.

Scenario 1: target D_t is reachable. In this scenario, there exists $1 \leq l_1^t \leq n$ such that

$$\sum_{i=1}^{l_1^t} p_{\sigma(i)} > D_t - 1/2$$

$$\sum_{i=1}^{l_2^t} p_{\sigma(i)} < D_t - 1/2$$

This also suggests that $p_{\sigma(l_1^t)} > 0$.

We will characterize \mathbb{S}_t in two ways. First, we will show that S_t must follow the right ordering. Second, we will show that S_t will at most select one more or one less arm than the optimal selection from the oracle.

Firstly, by Lemma 9 and $\epsilon_1 \leq \Delta_{k_1^t}, \Delta_{k_2^t}$, we know the subset S_t selected by CUCB-Avg must satisfy one of the following:

- i) $S_t \subseteq \{\sigma(1), \dots, \sigma(k_1^t)\}$,
- ii) $\{\sigma(1), \dots, \sigma(k_1^t)\} \subsetneq S_t \subseteq \{\sigma(1), \dots, \sigma(k_2^t)\}$, and S_t has no more than $l_2^t - k_1^t$ arms whose parameter is equal to $p_{\sigma(k_1^t)}$.

- iii) $\{\sigma(1), \dots, \sigma(k_1^t)\} \subsetneq S_t \subseteq \{\sigma(1), \dots, \sigma(k_2^t)\}$, and S_t has $l_2^t - k_1^t + 1 \leq u \leq l_1^t - k_1^t$ arms whose parameter is equal to $p_{\sigma(k_1^t)}$.
- iv) $\{\sigma(1), \dots, \sigma(k_1^t)\} \subsetneq S_t \subseteq \{\sigma(1), \dots, \sigma(k_2^t)\}$, and S_t has more than $l_1^t - k_1^t$ arms whose parameter is equal to $p_{\sigma(k_1^t)}$.
- v) $\{\sigma(1), \dots, \sigma(k_2^t)\} \subsetneq S_t$.

By the proof of Lemma 3, iii) generates no regret. Similarly to the proof in Appendix C, we can rule out i) and v) and show that ii) and iv) only happens under some restrictions on δ_1^t and δ_2^t . Then following the same argument, we can bound the regret by $2n\epsilon$.

Scenario 2: target D_t too large to reach. In this scenario, $\sum_{i=1}^n p_i \leq D_t - 1/2$. Similarly, we can show that S_t satisfy $|S_t| \geq n - 1$ and S_t must include the top $n - 1$ arms. Only when $\delta_2^t < (n - 1)\epsilon$, the regret is not 0, and the regret bound will also hold by the same argument.