Designing Bounded min-knapsack Bandits algorithm for Sustainable demand response

Akansha Singh [1] **P Meghana Reddy** [1] **Shweta Jain** [1] **Sujit Gurjar** [2] **Zoltan Nagy** [3]

Abstract

Around 40% of global energy produced is consumed by buildings. By using renewable energy resources we can alleviate the dependence on electrical grids. Recent trends focus on incentivizing consumers to reduce their demand consumption during peak hours for sustainable demand response. To minimize the loss, the distributor companies should target the right set of consumers and demand the right amount of electricity reductions.This paper proposes a novel bounded integer min-knapsack algorithm and shows that the algorithm, while allowing for multiple unit reduction, also optimizes the loss to the distributor company within a factor of two (multiplicative) and a problem dependent additive constant. Existing CMAB algorithms fail to work in this setting due to non-monotonicity of reward function and time varying optimal sets. We propose a novel algorithm (Twin-MinKPDR-CB) to learn these compliance probabilities efficiently. Twin-MinKPDR-CB works for non-monotone reward functions, bounded min-knapsack constraints, and time-varying optimal sets. We find that Twin-MinKPDR-CB achieves sub-linear regret of $O(\log T)$ with T being the number of rounds demand response is run.

1. Introduction

The residential and industrial sector consumes more than 40% of global electricity produced give reference even some urls are fine. Renewable energy resources can be effectively used. But due to various uncertainties involved in renewable resources' use, instability can be caused. One of the major problems being high peak load. Towards this, one can make the smart grid learn human behavior intelligibly and use it to implement informed decisions about shifting the peak energy consumption over time via a *demand response program*. There are many ways in which a distributing company (DC) can implement a demand response program. The popular one being introducing dynamic pricing by DC based on the supply shortage. The anticipation is that the consumers will shift their electricity loads to lower-priced – non-peaked hours whenever possible, thus reducing the peak demand.

This paper considers a demand response program where a DC asks the consumers to voluntarily optimize their electricity consumption by offering certain incentives. To give these incentives, DC desires to select an optimal subset of consumers along with an allocation vector depicting the number of electricity unit reduction it is going to ask the selected consumers. This allocation vector also depends on the shortage of electricity DC faces. Every consumer has a certain value associated with every unit (KWh) of electricity at that time and expects a compensation equivalent to this valuation for reducing the load. Additionally, each consumer has a limit to the amount of electricity it can reduce. Due to external stochastic factors such as climate change, uncertainty in renewable energy resources at consumers' end, or a sudden increase in workload, there is a certain probability with which the consumer can reduce the electricity. We refer to such probability as *compliance probability* (CP). The DC's goal is thus to minimize (i) the expected loss, which is a function of the cost of buying the electricity from the market, which in turn depends upon CPs, and (ii) the cost incurred for compensating the consumers via the demand-response program. By exploiting the heterogeneity in the consumer base, multiple units reduction provides more flexibility to the DC and ensures cost effective allocation.

In this work, we introduce a novel transformation of the problem to the bounded min knapsack framework for demand response, MinKPDR and show its equivalence up to an additional problem dependent constant factor. Bounded min-knapsack problem is a well studied problem in theoretical computer science and there exists $2-$approximate algorithms to solve the problem. Thus, MinKPDR framework helps us in obtaining polynomial time algorithm with

[*]Equal contribution [1]Department of Computer Science, Indian Institute of Technology Ropar, Punjab, India [2]Department of Computer Science and Engineering, Indian Institute of Information Technology Hyderabad, Telangana, India [3]University of Texas, Austin, USA. Correspondence to: Akansha Singh <2017csb1065@iitrpr.ac.in>, P Meghana Reddy <2017csb1094@iitrpr.ac.in>.

Tackling Climate Change with Machine Learning Workshop at ICML 2021.

approximate guarantees.

When CPs of the consumers are not known, they can be learnt using *combinatorial multi-armed bandits* (CMAB) algorithm by selecting different subsets of consumers at different rounds (Jain et al., 2014; Li et al., 2018). The existing combinatorial MAB (CMAB) literature (Chen et al., 2016) heavily rely on two assumptions: (i) The reward function is monotone in terms of the stochastic rewards (compliance probabilities in our case), and (ii) The optimal set is fixed over a period of time. The first assumption does not hold even for a single unit reduction case and since the amount of shortage of electricity varies over time, the optimal set changes everytime thus violating the second assumption. Typically, if one has monotone reward functions, upper confidence bound (UCB) based algorithms work well in practice. Non-monotone reward function necessitates the design of a novel MAB algorithm. Towards this, we propose an ingenious combination of UCB and LCB (lower confidence bounds) to learn CPs in demand-response. Basically, we solve the problem twice, once with UCB in constraints and its twin problem – the same problem with LCB in constraints and opt for a solution better out of these two. We call the learning version of MinKPDR as Twin-MinKPDR-CB. We show that Twin-MinKPDR-CB achieves sub-linear regret of $O(\log T)$ to learn CPs, with T being the number of rounds for which demand response is run.

2. Mathematical Model

There are $N = \{1, 2, \ldots, n\}$ consumers available for the demand response to whom a distributor company (DC) is distributing the electricity. Each consumer i has three quantities associated with them, k_i representing maximum units that the consumer i can reduce, c_i representing the compensation cost per unit reduction, and p_i denoting the probability of reducing one unit of electricity also known as compliance probability (CP). The DC asks the consumers to report their compensation cost c_i and maximum units of reduction k_i to participate in the demand response. If a consumer successfully reduces one electricity unit, he receives the offer of c_i per unit reduction. However, due to uncertainties at consumers' end such as failing to generate the expected electricity (renewable resources at consumers end), such uncertainty is depicted by the quantity p_i which denotes the probability of reducing one unit of electricity. We would like to design a demand response that subsumes these uncertainties in optimization problem itself. Therefore, our goal is not only to select the consumers who have lower cost of reducing the electricity but at the same time should also have higher probability of reducing the electricity once committed for the demand response. Thus, apart from minimizing the costs, the demand response would minimizes the variance to weed out the consumers with low CPs

At each round t, a distributor company encounters a shortage of $\mathcal{E}_t \neq 0$ and the goal is to select an allocation vector of reduction units $\mathbf{x}_t = (x_{1,t}, x_{2,t}, \ldots, x_{n,t})$ where $x_{i,t}$ represents the amount of electricity units asked from a consumer i at time t to reduce. Let S_t and $|S_t|$ be the set and number of consumers who are asked to reduce at least one unit of electricity i.e. $S_t = \{i | x_{i,t} > 0\}$. At the round t, whatever shortage the distributor company faces, it would have to buy from the market and the cost of buying the electricity from the market leads to quadratic loss (Li et al., 2018; Jain et al., 2018). Even if a consumer i is asked to reduce $x_{i,t}$ units of electricity at time t, due to uncertainties involved, the actual units of electricity reduced will be a random variable. Let $X_{i,t}$ denote the actual units of electricity that consumer i reduces at time t. If the allocation vector at time t is $\mathbf{x}_t$, then the cost of buying the electricity from the market is proportional to: $M_t(\mathbf{x}_t) = \left(\sum_{i \in S_t} X_{i,t} - \mathcal{E}_t\right)^2$.

Here, $X_{i,t}$ is a binomial random variable with parameters $(x_{i,t}, p_i)$ such that $0 \leq x_{i,t} \leq k_i$. We assume that if the consumer i is asked to reduce $x_{i,t}$ units than he/she reduces each unit independently with probability p_i. Thus, the final expected loss $\mathbf{E}L(\mathbf{x}_t)$ at round t is given as the sum of the loss incurred due to buying electricity from the market and the expected compensation to the agents, i.e.

$$\mathbf{E}\left[CM_t(\mathbf{x}_t) + \sum_{i \in S_t} X_{i,t} c_i\right] = C\mathbf{E}M_t(\mathbf{x}_t) + \sum_{i \in S_t} p_i x_{i,t} c_i$$

Here C represents the cost to buy the electricity from the market. Let $Y_{i,t} = X_{i,t} - \mathcal{E}_t/|S_t|$, then $C\mathbf{E}M_t(\mathbf{x}_t)$ is:

$$= C\mathbf{E}\left[(\sum_{i \in S_t} Y_{i,t})^2\right] = Cvar(\sum_{i \in S_t} Y_{i,t}) + C(\mathbf{E}[\sum_{i \in S_t} Y_{i,t}])^2$$

$$= C\sum_{i \in S_t} x_{i,t} p_i (1 - p_i) + C\left(\sum_{i \in S_t} x_{i,t} p_i - \mathcal{E}_t\right)^2$$

The goal is to select an allocation vector $\mathbf{x}_t$ so as to minimize $\mathbf{E}L(\mathbf{x}_t)$ which is given as:

$$C\left(\sum_{i \in S_t} x_{i,t} p_i - \mathcal{E}_t\right)^2 + C\sum_{i \in S_t} x_{i,t} p_i (1 - p_i) + \sum_{i \in S_t} x_{i,t} p_i c_i \tag{1}$$

3. MinKPDR for Multi Unit Reduction

Let c_{max} denote the maximum cost that any consumer incurs for a single unit of electricity, i.e. $c_{max} = \max_i c_i$. We assume that the distributor company will always prefer to ask the consumers to reduce the electricity as opposed to buying from the electricity market i.e. $C \geq c_{max}$. We provide a novel framework by drawing an interesting relation from the min-knapsack problem for which a 2-approximate

greedy algorithm exists (Csirik, 1991). At any round t, if $\sum_{i=1}^{n} k_i p_i < \mathcal{E}_t$ then $\mathbf{x}_t = \{k_1, k_2, \ldots, k_n\}$ else solve the following:

$$\min_{\mathbf{x}_t} C \sum_{i \in S_t} x_{i,t} p_i (1 - p_i) + \sum_{i \in S_t} x_{i,t} p_i c_i$$
$$\text{s.t.} \sum_{i \in S_t} x_{i,t} p_i \geq \mathcal{E}_t \text{ and } 0 \leq x_{i,t} \leq k_i \; \forall i \tag{2}$$

This is the bounded min-knapsack problem where instead of one instance, k_i instances of the item i are available. Thus, any min-knapsack algorithm can be used to solve its bounded version with same approximation factor but maybe with an increased complexity. We now prove that solving min-knapsack problem will result in only constant factor approximation to the original problem. The proofs are given in the appendix.

Theorem 1. *Let $\tilde{\mathbf{x}}_t$ be the optimal allocation vector from solving Equation (2) and $\mathbf{x}_t^*$ be the allocation vector from solving Equation (1). Then $\mathbb{E}L(\tilde{\mathbf{x}}_t) \leq \mathbb{E}L(\mathbf{x}_t^*) + 4C + 1$*

4. Twin-MinKPDR-CB for Unknown CPs

When the compliance probabilities of the consumers are not known, these have to be learnt over a period of time to minimize the loss function. The problem can be formulated as the combinatorial multi-armed bandit problem, where at each round a subset of arms (consumers) need to be selected and the reward (amount of electricity reduced) from these arms are observed. The estimates of CPs are thus updated at each time. Under monotonicity assumption, existing algorithms (Chen et al., 2013; 2016) use Upper confidence bound (UCB) based algorithm that work on the principle of optimism in the face of uncertainty. Twin-MinKPDR-CB (Algorithm 1) uses both UCB and lower confidence bound (LCB) to intelligently select the allocation vector. However, our problem is not monotone.

Lemma 1. *The multi-unit loss function in Equation (1) is not monotone in terms of compliance probabilities.*

To solve this issue, we propose a novel algorithm Twin-MinKPDR-CB that simultaneously solve the optimization problem with UCB and LCB. The algorithm is provided in Algorithm 1. The performance of any learning algorithm is measured by regret.The regret of the algorithm is defined as the difference in the cost of the allocation vector $\mathbf{x}_t$ output by the algorithm at a round t with unknown CPs and the cost of the optimal allocation vector $\mathbf{x}_t^*$ with known CPs. Since finding the optimal allocation vector to Equation (2) is a hard problem, we define the regret at round t as the difference in the cost of the allocation vector $\mathbf{x}_t$ returned by our algorithm with unknown CPs and the cost of the allocation vector $\tilde{\mathbf{x}}_t$ obtained by MinKPDR with known CPs i.e. $\mathcal{R}_t(\mathbf{x}_t) = \mathbf{E}L(\mathbf{x}_t) - \mathbf{E}L(\tilde{\mathbf{x}}_t)$

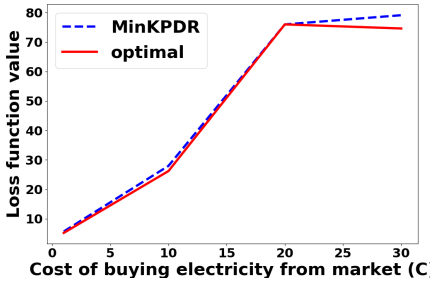

Figure 1. Difference between Loss value incurred by the distribution company in comparison to the cost of buying electricity from the market by using proposed minKPDR algorithm and optimal is always less than $4C + 1$

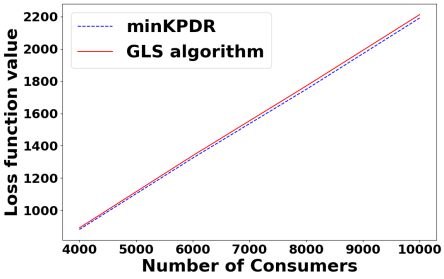

Figure 2. Loss value incurred by the distribution company in comparison to the number of consumers by minKPDR algorithm is lesser than GLS algorithm (benchmark algorithm from Shweta & Sujit (2020)), significant difference is seen as the number of consumers increase

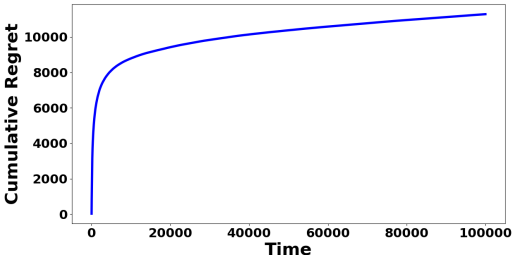

Figure 3. Regret is observed to be logarithmic with respect to time

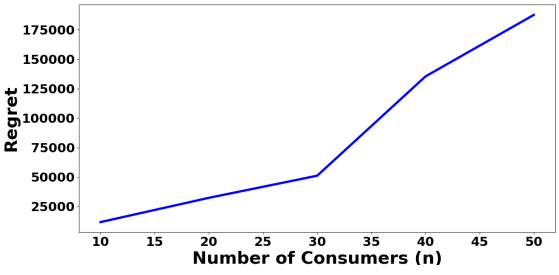

Figure 4. Regret is observed to be quadratic with respect to the number of consumers.

Algorithm 1 Twin-MinKPDR-CB

Input: $\{c_1, c_2, ...c_n\}, \{k_1, k_2, ...k_n\}$, Number of Rounds T.
Output: Allocations in each round $\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_T$

1. $\mathbf{x}_1 = \{k_1, k_2, ...k_n\}$ Make offer of full amount of electricity they can reduce to get initial estimate of CPs. i.e $n_i(1) = k_i \; \forall i$.
2. **for** $t \leftarrow 2 : T$ **do**

Observe the values of D_t and $X_{i,t-1}$ i.e actual amount reduced by i at t-1.

Update $\hat{p_i} = \frac{\sum_{t'=1}^{t-1} X_{i,t'}}{n_i(t-1)}$, $\hat{p}_i^+ = \hat{p}_i + \sqrt{\frac{2 \ln t}{n_i(t-1)}}$ and $\hat{p}_i^- = \hat{p}_i - \sqrt{\frac{2 \ln t}{n_i(t-1)}}$.

Solve for $\mathbf{x}_t^+, \mathbf{x}_t^-$ from

$$\min_{\mathbf{x}_t^+} C \sum_{i \in S_t^+} x_{i,t}^+ \hat{p}_{i,t}^- (1 - \hat{p}_{i,t}^+) + \sum_{i \in S_t^+} x_{i,t}^+ \hat{p}_{i,t}^- c_i \quad \text{s.t.} \sum_{i \in S_t^+} x_{i,t}^+ \hat{p}_{i,t}^+ \geq \mathcal{E}_t \tag{3}$$

$$\min_{\mathbf{x}_t^-} C \sum_{i \in S_t^-} x_{i,t}^- \hat{p}_{i,t}^- (1 - \hat{p}_{i,t}^+) + \sum_{i \in S_t^-} x_{i,t}^- \hat{p}_{i,t}^- c_i \quad \text{s.t.} \sum_{i \in S_t^-} x_{i,t}^- \hat{p}_{i,t}^- \geq \mathcal{E}_t \tag{4}$$

Obtain $\mathbb{E}L_{\hat{p}_t^+}(x_t^+), \mathbb{E}L_{\hat{p}_t^-}(x_t^-)$ as

$$\mathbb{E}L_{\hat{p}_t^+}(x_t^+) = C \left(\sum_{i \in S_t^+} x_{i,t}^+ \hat{p}_i^+ - \mathcal{E}_t \right)^2 + C \sum_{i \in S_t^+} x_{i,t}^+ \hat{p}_i^- (1 - \hat{p}_i^+) + \sum_{i \in S_t^+} x_{i,t}^+ \hat{p}_i^- c_i$$

$$\mathbb{E}L_{\hat{p}_t^-}(x_t^-) = C \left(\sum_{i \in S_t^-} x_{i,t}^- \hat{p}_i^- - \mathcal{E}_t \right)^2 + C \sum_{i \in S_t^+} x_{i,t}^- \hat{p}_i^- (1 - \hat{p}_i^+) + \sum_{i \in S_t^+} x_{i,t}^- \hat{p}_i^- c_i$$

if $\mathbb{E}L_{\hat{p}_t^+}(\mathbf{x}_t^+) < \mathbb{E}L_{\hat{p}_t^-}(\mathbf{x}_t^-)$ **then**
$\mathbf{x}_t = \mathbf{x}_t^+, \tilde{p} = p_i^+$
else
$\mathbf{x}_t = \mathbf{x}_t^-, \tilde{p} = p_i^-$

Theorem 2. *The regret of the algorithm is bounded by* $\left(\frac{8 \ln T}{(f^{-1}(\Delta))^2} + \frac{\pi^2}{3} + 1 \right) nC\mathcal{E}_{max}^2$

5. Simulation Results

We now present our simulation results to demonstrate the efficacy of the proposed MinKPDR and Twin-MinKPDR-CB algorithms. We take two benchmark algorithms to compare the offline MinKPDR algorithm. First is the optimal algorithm where the solution is computed via the brute force technique by considering all possible subsets (hence takes exponential time) and second, the GLS algorithm proposed by Shweta & Sujit (2020) having time complexity of $O(n \log n)$, with n being the number of consumers. We have used greedy algorithm (Csirik, 1991) to obtain the solution of minknapsack problem for both MinKPDR and Twin-MinKPDR-CB algorithms. The time complexity of MinKPDR using this greedy approach is also $O(n \log n)$.

Setting: For each consumer i, CPs p_i and compensation costs $c_i$both $\sim U[0, 1]$. The value of C is kept as 3 (except in figure (1)) and for figures (1, 3) the value of n is fixed at 10. The maximum units of reduction k_i by any consumer i is generated randomly from 1 to 5. The demand shortage $\mathcal{E}_t \sim U[1, \frac{K}{4}]$ with K being sum of maximum reductions from all the consumers.

Figure (1) compares the worst-case loss function of MinKPDR and the optimal algorithm over 500 samples. As can be seen from the figure that the loss differences between the optimal one and MinKPDR are very close and always less than $4C + 1$. Further, MinKPDR algorithm performed 20 times faster as compared to the optimal algorithm which is implemented using mixed-integer linear programming solver Gurobi(Gurobi Optimization, 2021). Figure (2) compare the worst-case loss over 500 samples for the GLS algorithm and MinKPDR algorithm. Since GLS works only for single unit, the figure is generated by implementing MinKPDR algorithm for single unit reduction case. It clearly shows the MinKPDR algorithm outperforming the GLS algorithm. Figures (3) and (4) represent the average cumulative regret over 100 runs obtained by Twin-MinKPDR-CB versus number of rounds and number of consumers respectively. Once the allocation vector is generated by Twin-MinKPDR-CB, the actual amount of electricity reduced by customer is generated as binomial random variable for every round. For Figure (4), the cumulative regret is computed with respect to the solution obtained by solving bounded min-knapsack problem for $T = 10^4$ rounds. As can be seen from the graph we get logarithmic regret in terms of T and quadratic regret in terms of n.

6. Conclusion

For the demand response problem, this paper presented a novel min-knapsack framework that can be used to shave off peak electricity consumption. Most of the work in this area considered only single unit reduction, which does not fully optimize to shave off the peak electricity consumption. The proposed novel transformation to min-knapsack can be used to solve multi-unit reduction. We designed a novel combinatorial multi-armed bandit algorithm that works for non-monotone reward function, non-convex con-

straints, and time-varying optimal set when the uncertainties are involved. Our proposed Twin-MinKPDR-CB algorithm achieves sub-linear regret of $O(\log T)$ where T is the number of rounds for which demand response is run. A combinatorial MAB algorithm for general non-monotone reward function is strongly required as these functions exist in many other settings such as resource allocation, influence maximization, etc. We believe that our novel Twin technique of combining UCB and LCB will be extremely beneficial for any setting involving such non-monotone reward functions.

References

Chen, W., Wang, Y., and Yuan, Y. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pp. 151–159, 2013.

Chen, W., Hu, W., Li, F., Li, J., Liu, Y., and Lu, P. Combinatorial multi-armed bandit with general reward functions. *Advances in Neural Information Processing Systems*, 29: 1659–1667, 2016.

Csirik, J. Heuristics for the 0-1 min-knapsack problem. *Acta Cybernetica*, 10(1-2):15–20, 1991.

Gurobi Optimization, L. Gurobi optimizer reference manual, 2021. URL http://www.gurobi.com.

Jain, S., Narayanaswamy, B., and Narahari, Y. A multiarmed bandit incentive mechanism for crowdsourcing demand response in smart grids. 2014.

Jain, S., Gujar, S., Bhat, S., Zoeter, O., and Narahari, Y. A quality assuring, cost optimal multi-armed bandit mechanism for expertsourcing. *Artificial Intelligence*, 254:44 – 63, 2018. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2017.10.001. URL http://www.sciencedirect.com/science/article/pii/S000437021730125X.

Li, Y., Hu, Q., and Li, N. Learning and selecting the right customers for reliability: A multi-armed bandit approach, 2018. URL https://nali.seas.harvard.edu/files/nali/files/2018cdcmab.pdf.

Shweta, J. and Sujit, G. A multiarmed bandit based incentive mechanism for a subset selection of customers for demand response in smart grids. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 2046–2053, 2020.

7. Appendix

7.1. Proof of Theorem 1

Lemma 2. *If $\tilde{S}_t \neq N$ then $-1 < \varepsilon_t < 0$.*

Proof. If $\varepsilon_t < -1$, then MinKPDR algorithm can drop at least one consumer from $\tilde{S}_t$ and can strictly reduce the objective function in Equation (2). □

Lemma 3. *If $\varepsilon_t^* > 0$ then either $\varepsilon_t^* < 1$ or $S_t^* = N$.*

Proof. $S_t^* \neq N \implies \exists k \notin S_t^*$. If $\varepsilon_t^* > 1$, then:

$$\begin{aligned}
&\mathbb{E}L(S_t^*) - \mathbb{E}L(S_t^* \cup \{k\}) \\
&= C\varepsilon_t^{*2} - C(p_k - \varepsilon_t^*)^2 - Cp_k(1-p_k) - cp_k \\
&= -Cp_k^2 + 2C\varepsilon_t^* p_k - Cp_k + Cp_k^2 - c_k p_k > 0 \\
&\qquad\qquad (c \leq C, \varepsilon_t^* > 1)
\end{aligned}$$

Leading to the contradiction that S_t^* is the optimal set. □

Now, the proof of Theorem 1 is as follows:

Proof. Let $g(\tilde{S}_t)$ represent the objective function value of Equation (2). If $\varepsilon_t^* \leq 0$ then $g(\tilde{S}_t) \leq g(S_t^*)$. When $\varepsilon_t^* > 0$ and $S_t^* \neq N$ then let $S_{new} = S_t^* \cup S_{ext}$ be the set such that $\sum_{i \in S_{new}} p_i \geq \mathcal{E}_t$ and S_{ext} includes minimum number of such consumers. If such a set is not possible, $S_{new} = N$. From Lemma 3, $\varepsilon_t^* < 1$ and thus $\sum_{i \in S_{ext}} p_i \leq 2$. The reason is we are at max one unit short and we cannot overshoot much since $p_i < 1 \ \forall i$. $g(S_{new}) - g(S_t^*) = C\sum_{i \in S_{ext}} p_i(1-p_i) + \sum_{i \in S_{ext}} p_i c_i \leq 4C$. Further, if $\sum_{i \in S_t^*} p_i < \mathcal{E}_t$ and $S_t^* = N$, then $g(S_t^*) = g(\tilde{S}_t)$. Thus, $g(\tilde{S}_t) \leq g(S_t^*) + 4C$. We now have following two cases:

Case 1: $\tilde{S}_t \neq N$: From Lemma 2, $\mathbb{E}L(\tilde{S}_t) = g(\tilde{S}_t) + \varepsilon_t^2 \leq g(S_t^*) + 4C + 1 \leq \mathbb{E}L(S_t^*) + 4C + 1$.

Case 2: $\tilde{S}_t = N$: In this case, $\varepsilon_t < \varepsilon_t^*$. Thus, $\mathbb{E}L(\tilde{S}_t) = g(N) + \varepsilon_t^2 \leq g(S_t^*) + 4C + \varepsilon_t^{*2} \leq \mathbb{E}L(S_t^*) + 4C$. □

Note: If the selected set S_t is not optimal to Equation (2) but is $\alpha-$approx solution such that $g(S_t) \leq \alpha g(\tilde{S}_t)$, then $\mathbb{E}L(S_t) \leq g(S_t) + 1 \leq \alpha g(\tilde{S}_t) + 1 \leq \alpha(g(S_t^*) + 4C + 1) + 1$.

7.2. Proof of Theorem 2

Lemma 4. *Let $\mathbb{E}L_p(\mathbf{x}_t)$ be the expected loss function of $\mathbf{x}_t$ with true CP vector p. Then, $\mathbb{E}L_{\tilde{p}}(\mathbf{x}_t) \leq \mathbb{E}L_p(\mathbf{x}_t)$.*

Proof. Following inequalities holds with high probability for all consumers i and for all allocation vectors $\mathbf{x}_t$:

$$\sum_{i \in S_t} x_{i,t}\hat{p}_{i,t}^- - \mathcal{E}_t \leq \sum_{i \in S_t} x_{i,t}p_i - \mathcal{E}_t \leq \sum_i x_{i,t}\hat{p}_{i,t}^+ - \mathcal{E}_t$$

$$\sum_i x_{i,t}\hat{p}_{i,t}^-(1-\hat{p}_{i,t}^+) \leq \sum_i x_{i,t}p_i(1-p_i)$$

$$\sum_i x_{i,t}\hat{p}_{i,t}^- c_i \leq \sum_i x_{i,t}p_i c_i$$

All the above inequalities along with the definition of $\mathbb{E}L_{\tilde{p}}(x_t)$, we have $\mathbb{E}L_{\tilde{p}}(x_t) \leq \mathbb{E}L_p(x_t)$. □

Lemma 5. ***Bounded Smoothness Property:*** *Consider any two compliance probability vectors $p = \{p_1, p_2, \ldots, p_n\}$ and $p' = \{p_1', p_2', \ldots, p_n'\}$. Let $\mathbb{E}L_p(\mathbf{x}_t)$ denote the loss function of allocation vector $\mathbf{x}_t$ with compliance probability vector p, then $|\mathbb{E}L_p(\mathbf{x}_t) - \mathbb{E}L_{p'}(\mathbf{x}_t)| \leq f(\delta)$ if $|p_i - p_i'| \leq \delta \ \forall i$ where f is a strictly increasing and invertible function.*

Proof. Let $K = \sum_i k_i \geq \sum_{i \in S_t} k_i \geq \sum_{i \in S_t} x_{i,t}$ We have:

$$|\sum_{i \in S_t} x_{i,t}p_i c_i - \sum_{i \in S_t} x_{i,t}p_i' c_i| \leq \delta \sum_{i \in S_t} x_{i,t}c_i \leq KC\delta$$

$$\begin{aligned}
&|\sum_{i \in S_t} x_{i,t}p_i(1-p_i) - \sum_{i \in S_t} x_{i,t}p_i'(1-p_i')| \\
&\leq \sum_{i \in S_t} x_{i,t}|p_i - p_i'|(1 + |p_i + p_i'|) \leq 3K\delta
\end{aligned}$$

$$\begin{aligned}
&\left|\left(\sum_{i \in S_t} x_{i,t}p_i - \mathcal{E}_t\right)^2 - \left(\sum_{i \in S_t} x_{i,t}p_i' - \mathcal{E}_t\right)^2\right| \\
&\leq \left|\sum_{i \in S_t} x_{i,t}p_i + \sum_{i \in S_t} x_{i,t}p_i'\right| \left|\sum_{i \in S_t} x_{i,t}p_i - \sum_{i \in S_t} x_{i,t}p_i'\right| \leq 2K^2\delta
\end{aligned}$$

Thus, $|\mathbb{E}L_p(x) - \mathbb{E}L_{p'}(x)| \leq f(\delta) = (4CK + 2CK^2)\delta$ □

Once we have the bounded smoothness property from Lemma 5 and the monotonicity property from Lemma 4, the regret bound proof can follow the similar proof as in (Chen et al., 2013). However, the optimal set in our setting varies each time as opposed to a fixed optimal set.

Lemma 6. *If $\mathbf{x}_{i,t} = 0 \ \forall i$ then $\tilde{\mathbf{x}}_{i,t} = 0 \ \forall i$*

Proof. $\mathbf{x}_{i,t} = 0 \ \forall i$ then $\mathbb{E}L_{\tilde{p}}(x_t) = C\mathcal{E}_t^2 \leq \mathbb{E}L_{\tilde{p}}(\tilde{\mathbf{x}}_t) \leq \mathbb{E}L_p(\tilde{\mathbf{x}}_t) \implies \tilde{\mathbf{x}}_{i,t} = 0 \ \forall i$. Here first inequality is due to the optimization problem solved by MinKPDR and second inequality is due to Lemma 4. □

The above result is an important result as it implies that whenever we are selecting a sub-optimal allocation (including no allocation), we are incrementing the counter for exactly one consumer. Let us define $\Delta = \min\{\mathbb{E}L_p(\mathbf{x}_t) - \mathbb{E}L_p(\tilde{\mathbf{x}}_t) | \mathbb{E}L_p(\mathbf{x}_t) > \mathbb{E}L_p(\tilde{\mathbf{x}}_t)\}$. Further, define $l_t = \frac{8 \ln t}{(f^{-1}(\Delta))^2}$. Let $\mathcal{E}_{max} = \max_t\{\mathcal{E}_t\}$, then the maximum regret at any round t is upperbounded by $\mathcal{E}_{max}$

and the expected regret of the algorithm is bounded by: $\mathbb{E}[\sum_{i=1}^{n} N_{i,T}]C\mathcal{E}_{max}^2$. The proof of theorem 2 is as follows:

Proof. The following steps are similar to (Chen et al., 2013):

$$\sum_{i=1}^{n} N_{i,T} - n(l_T + 1) = \sum_{n+1}^{T} \mathbb{I}(\mathbf{x}_t \neq \tilde{\mathbf{x}}_t) - nl_T$$
$$\leq \sum_{t=n+1}^{T} \sum_{i=1}^{n} \mathbb{I}(\mathbf{x}_t \neq \tilde{\mathbf{x}}_t, N_{i,t} > N_{i,t-1}, N_{i,t-1} > l_T)$$
$$\leq \sum_{t=n+1}^{T} \sum_{i=1}^{n} \mathbb{I}(x_t \neq x_t^*, N_{i,t} > N_{i,t-1}, N_{i,t-1} > l_t)$$
$$= \sum_{t=n+1}^{T} \mathbb{I}(\mathbf{x}_t \neq \tilde{\mathbf{x}}_t, \forall i : x_{i,t} > 0, T_{i,t-1} > l_t)$$

When $T_{i,t-1} > l_t \ \forall i$, from Hoeffding's bound we have:

$$\mathbb{P}\left(|\hat{p}_{i,t}^{+} - p_i| > f^{-1}(\Delta)\right) \leq \mathbb{P}\left(|\hat{p}_{i,t}^{+} - p_i| \geq 2\sqrt{\frac{2\ln t}{T_{i,t-1}}}\right)$$
$$\leq 2t^{-2}$$
$$\mathbb{P}\left(|\hat{p}_{i,t}^{-} - p_i| > f^{-1}(\Delta)\right) \leq \mathbb{P}\left(|\hat{p}_{i,t}^{-} - p_i| \geq 2\sqrt{\frac{2\ln t}{T_{i,t-1}}}\right)$$
$$\leq 2t^{-2}$$

Thus with probability $1 - 2nt^{-2}$,

$$\mathbb{E}L_p(\mathbf{x}_t) < \mathbb{E}L_{\tilde{p}}(\mathbf{x}_t) + \Delta \leq \mathbb{E}L_{\tilde{p}}(\tilde{\mathbf{x}}_t) + \Delta \leq \mathbb{E}L_p(\tilde{\mathbf{x}}_t) + \Delta$$

Here, first inequality comes from Bounded smoothness property, second from definition of $\mathbf{x}_t$, and third from Lemma 4. Thus, leading to the contradiction to the definition of Δ. Thus, the expected regret is bounded as:

$$\mathbb{E}[\sum_{i=1}^{n} N_{i,T}]C\mathcal{E}_{max}^2 \leq \left(n(l_T + 1) + \sum_{t=1}^{T} \frac{2n}{t^2}\right) C\mathcal{E}_{max}^2$$
$$\leq \left(\frac{8\ln T}{(f^{-1}(\Delta))^2} + \frac{\pi^2}{3} + 1\right) nC\mathcal{E}_{max}^2$$

□

From Lemma 5, $f^{-1}(\Delta) \propto K^2$, thus leading $O(n^5)$ regret. This upper bound is attributed to the fact that although we are pulling several instances of arm i at one instance, we are incrementing the counter $N_{i,t}$ only once. However, we can see from the simulation section, that in practice, the regret turns out to be quadratic in n.