2022 IEEE 61st Conference on Decision and Control (CDC)
December 6-9, 2022. Cancún, Mexico

A Safe Pricing Mechanism for Distributed Resource Allocation with Bandit Feedback

Spencer Hutchinson Berkay Turan Mahnoosh Alizadeh

2022 IEEE 61st Conference on Decision and Control (CDC) | 978-1-6654-6761-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/CDC51059.2022.9993064

***Abstract*— Pricing mechanisms are commonly advocated as a main tool to shape customers' demand in societal scale networked infrastructure such as power or transportation systems. Given that the price response function of each user is generally considered private and unknown, most existing algorithms rely on protocols that explicitly or implicitly solicit this information in order to design prices. However, approaches that rely solely on learning the price response through repeated interactions are more practical and gaining traction. In this paper, we model each customer's price response by an unknown parameter vector and we design a resource pricing mechanism to manage demand in order to maximize total welfare while ensuring that a set of linear constraints on the consumption are satisfied at all time steps with high probability. We propose an algorithm to address this problem that utilizes the well known principle of optimism in the face of uncertainty (OFU), while simultaneously being pessimistic with respect to constraint violation. Our analysis of this algorithm shows that, with high probability, it will not violate the constraints and will achieve $\mathcal{O}(\log(T)\sqrt{T})$ regret. Numerical experiments validate these results and demonstrate how our algorithm can be applied to demand response management in power distribution systems.**

I. Introduction

As optimization and learning expands in to safety-critical settings, there is a need for algorithms that can ensure that safety constraints are not violated, while maintaining good performance. Accordingly, various safe learning problems have been posed in recent years, including those that give the learner access to noisy observations (bandit feedback) of the optimization objective and/or the constraints (e.g. [1], [2], and [3]). In particular, online safe learning problems, such as those in [2] and [4], are especially relevant to many real-world settings as they guarantee good performance for the entire time horizon via regret analysis.

Although online safe learning algorithms are applicable to various real-world settings, here we are particularly motivated by applications in cyber-physical systems (CPS) with humans in the loop (e.g. power grid, transportation networks). In these systems, the automation algorithm often interacts with humans via pricing (e.g. electricity prices, road tolls). Through the choice of prices, these algorithms seek to achieve high social welfare while satisfying the physical constraints of the system. This fits well in to the online safe learning paradigm given that the *price response* of the users is initially unknown, but can be learned by posting prices and observing the response of the users.

S. Hutchinson, B. Turan and M. Alizadeh are with Dept. of ECE, UCSB, Santa Barbara, CA, USA. This work is supported by NSF grant #1847096. E-mails: shutchinson@ucsb.edu, bturan@ucsb.edu, alizadeh@ucsb.edu

One traditional approach for these types of CPS-related problems utilizes the network utility maximization (NUM) framework [5], [6], [7]. With utility maximization problems in general, the objective is to optimally allocate limited resources to n users in order maximize the total utility that these users derive from the resource. In many scenarios, however, it is not feasible to allocate the resources directly because the utility function of each user is private. Through dual decomposition [8], the NUM framework allows for the utility to be maximized by observing the resource consumption of each user and updating the dual variables (which equate to the resource prices) at each iteration. Although this approach has been applied to several types of CPS, including the coordination of generation and demand in power grids [5], and managing congestion in transportation networks [7], existing algorithms cannot ensure that the constraints will be satisfied at each time step prior to convergence. As such, in safety-critical CPS such as power systems, they are only used as a negotiation protocol to find the optimal operational point and cannot be implemented without prior coordination between the users and the system operator. This motivates the work here, where we aim to develop distributed resource allocation algorithms that operate online and satisfy safety constraints at each time step.

The contributions of this work are:

- We pose a novel distributed resource allocation problem where the price response of each users is paramaterized by an unknown vector. At each time step, the learner posts a price and then receives bandit feedback of the price response. The objective is to maximize the utility, which is a known function of the price response, while ensuring with high probability that a set of linear constraints are satisfied for all time steps.
- We propose a linear bandit-inspired algorithm for the stated problem that, with high probability, guarantees constraint satisfaction at each iteration. We then prove that the proposed algorithm achieves $\mathcal{O}(\log(T)\sqrt{T})$ regret with high probability.
- Our numerical experiments provide empirical validation of the theoretical results and demonstrate how this algorithm can be implemented in a power distribution network with price responsive users.

Related work: This work is related to literature in the fields of (1) safe learning, (2) NUM and (3) linear stochastic bandits.

1) *Safe learning:* In general, safe learning requires that the learner satisfy constraints at each time step. One class

of these problems are optimization problems in which the learner is only given access to bandit feedback of the objective and constraints, and is required to satisfy the constraints at each iteration [1], [4], [9], [10]. In another direction, there has been work that models the unknown objective as a Gaussian process and uses the corresponding confidence region to ensure safety [3], [11]. There is also existing work addressing safety in linear stochastic bandits. For example, [12] and [13] study settings where there are constraints on the costs or rewards received at each round. The work in [2] addresses a safe linear bandit problem that has some similarities to the one discussed here. In particular, we use multiple linear constraints that jointly apply to mulitple users (or equivalently, multiple bandits), while [2] considers one constraint for a single bandit that is bilinear with respect to the action and the unknown variable. This necessitates the distinct approach that we propose here.

2) *NUM:* The framework of NUM allows for layering of an optimization problem through dual and primal decompositions [8]. This layering allows for a network-based problem to be modularized and solved in a distributed fashion [14]. It has been successfully applied to various network problems, notably congestion control for internet networks [15], [16], [17]. Most relevantly, NUM has also been applied to the control of CPS with humans in the loop [5], [6], [7].

3) *Linear Stochastic Bandits:* Bandit problems are a class of online learning problems where the learner chooses an action at each round from a decision set and observes a reward according to the action taken [18]. With linear stochastic bandits, the reward is an unknown linear function of the action plus some additive noise [19]. In the problem presented here, we use a similar parameterization and noise model for the price response. One prominent approach for solving linear stochastic bandit problems is *optimism in the face of uncertainty* (OFU), from which linear upper confidence bound (LUCB) algorithms have emerged [20]. For our algorithm, we adopt the OFU approach as well. We also utilize least squares confidence regions that have been developed for LUCB in [21]. Additionally, we draw on the regret analysis of LUCB, such as in [21], [22] to establish regret guarantees for our algorithm.

Paper Organization: The problem is stated in Section II. An algorithm designed to address this problem is presented in Section III, and the analysis of this algorithm is discussed in Section IV. Lastly, Section V describes a numerical experiment that demonstrates how the algorithm might be applied to demand response management in power systems.

Notations: We use $\|z\|$ to refer to the Euclidean norm of z and $\|z\|_A$ to refer to $\sqrt{z^T A z}$ for $z \in \mathbb{R}^m$ and positive definite $A \in \mathbb{R}^{m\times m}$. For vectors z_1 and z_2, $z_1 \leq z_2$ means that each element of z_1 is less than or equal to z_2. The set $\{1, 2, ..., n\}$ is referred to by $[n]$, where n is a positive integer. For a vector $z \in \mathbb{R}^m$ and $i \in [m]$, $[z]_i$ refers to the ith element of z. The zero vector is denoted with $\mathbf{0}$. The positive reals and non-negative reals are referred to as $\mathbb{R}_{++}$ and $\mathbb{R}_+$ respectively.

II. Problem Setup

We consider a distributed resource allocation problem where a central coordinator and n users work together to find a solution over T time steps, while satisfying all p resource constraints at each time step. Each user, time step and constraint are indexed by $i \in [n]$, $t \in [T]$ and $j \in [p]$ respectively. The central coordinator interacts with the users at each time step by broadcasting prices to each of them. After receiving the price, each user responds with a consumption level as specified by the *The Individual User Problem* in Section II-A. In turn, the central coordinator observes a noisy version of the consumption of each user and chooses the price for the next time step as discussed in the *Central Coordinator's Problem* in Section II-B.

A. The Individual User Problem

With this problem, we formalize the way users determine their resource consumption in response to prices. At time step t, user i receives the price γ_i^t from the central coordinator. User i's consumption, x_i^t, is defined according to the user's individual price response function:

$$x_i^t = x_i(\gamma_i^t; \theta_i^*) = h_i(\gamma_i^t)^T \theta_i^*. \tag{1}$$

The parameter vector $\theta_i^* \in \mathbb{R}_+^m$ is unknown to the central coordinator, while $h_i : \mathbb{R} \to \mathbb{R}_+^m$ is a known function that is continuous and non-increasing. It follows from the definitions of θ_i^* and h_i that $x_i(\cdot)$ is positive non-increasing, which is expected given that the consumption of a resource will typically not increase as the resource price increases. In the following assumption, we assume that h_i and θ_i^* are bounded, and that there exists a high enough price such that all elements of h_i are zero.

Assumption 1. *The following applies for all $i \in [n]$. There exists positive constants S and L such that $\|\theta_i^*\| \leq S$ and $\|h_i(\gamma_i)\| \leq L$ for all $\gamma_i \in \mathbb{R}$. Additionally, there exists a non-negative constant y such that $h_i(y) = \mathbf{0}$.*

Due to random variation in the user's behavior or corruption of the measurement and communication system, the central coordinator observes the noisy consumption:

$$\bar{x}_i^t = x_i^t + \mu_i^t, \tag{2}$$

where μ_i^t is a conditionally subgaussian random variable as follows.

Assumption 2. *Let $\mathcal{F}_i^t = \sigma(\gamma_i^1, \gamma_i^1, ..., \gamma_i^{t+1}, \mu_i^1, \mu_i^2, ..., \mu_i^t)$ be the history at round t for user i. For all t and i, μ_i^t is conditionally σ-subgaussian such that $\mathbb{E}[\mu_i^t | \mathcal{F}_i^{t-1}] = 0$ and $\mathbb{E}[e^{\lambda \mu_i^t} | \mathcal{F}_i^{t-1}] \leq \exp(\frac{\lambda^2 \sigma^2}{2}), \forall \lambda \in \mathbb{R}$.*

The noise model characterized by Assumption 2 is commonly used in the literature (e.g. [2], [4], [21]). Next, we define the central coordinator's problem.

B. The Central Coordinator's Problem

The central coordinator is assumed to be a welfare maximizing entity. Specifically, its goal is to choose the prices $\gamma^t = [\gamma_1^t \ \gamma_2^t \ ... \ \gamma_n^t], \forall t$, to maximize the users' welfare

from consuming resources within the physical capacity of the system. To mathematically define the central coordinator's objective, various goals could be considered. One approach which is particularly suitable for societal-scale CPS that serve basic needs of users is to ensure fairness in how resources are allocated. For example, underserved communities should not be charged high prices or, large commercial users should not block access to resources for smaller residential users that share the same network. To ensure this, the central coordinator can design appropriate utility functions for different classes of users. The properties of different utility functions has been studied in the literature (e.g. [15], [23], [24]) and is not a focus of this work.

We use the function $f_i : \mathbb{R} \to \mathbb{R}$ to specify the utility function assigned to user i's consumption. Therefore, the utility at time step t is given by $\sum_{i=1}^{n} f_i(x_i^t)$. Furthermore, we assume that each of the utility functions are Lipschitz continuous.

Assumption 3. *The utility function f_i is M-Lipschitz continuous for all $i \in [n]$, such that $|f_i(y_1) - f_i(y_2)| \leq M|y_1 - y_2|$ $\forall y_1, y_2$.*

The physical limits of the system are specified by a set of linear constraints on the consumptions of the users

$$\sum_{i=1}^{n} a_{ji} x_i(\gamma_i^t; \theta_i^*) \leq c_j^t, \quad \forall j \in [p], \forall t \in [T], \tag{3}$$

where $a_{ji} \in \mathbb{R}_+$ and $c_j^t \in \mathbb{R}_+$ are known to the central coordinator. Since the a_{ji}s are non-negative, these constraints represent capacity limits on the network (e.g. in demand response this would be transformer capacities, wire capacities). Also note that c_j^t is allowed to vary with time. This could be used to capture varying network conditions on a daily basis (e.g., more renewable generation at certain nodes that must be consumed locally). Given the utility functions and the constraints, the central coordinator would choose the prices at time t as follows if the θ_i^*s were known:

$$\begin{aligned} \breve{\gamma}^t \in \arg\max_{\gamma \in \mathbb{R}^n} & \sum_{i=1}^{n} f_i(x_i(\gamma_i; \theta_i^*)) \\ \text{s.t. } & \sum_{i=1}^{n} a_{ji} x_i(\gamma_i; \theta_i^*) \leq c_j^t, \quad \forall j \in [p], \\ & \gamma_i \geq \Gamma_i, \ \forall i \in [n], \end{aligned} \tag{4}$$

where $\gamma = [\gamma_1 \ \gamma_2 \ ... \ \gamma_n]$, and Γ_i is a lower bound on the price for user i. This lower bound might be useful for restricting the price in some applications, but if it's not applicable, then it can be set low enough to not impact the solution of (4). We let $\breve{x}^t = [x_1(\breve{\gamma}_1^t; \theta_1^*) \ \dots \ x_n(\breve{\gamma}_n^t; \theta_n^*)]$ be the optimal resource consumption vector in response to the optimal prices.

Since θ_i^*'s are assumed to be private to the users and unknown by the central coordinator, it cannot solve (4) to find $\breve{\gamma}^t$ exactly. However, the central coordinator observes the noisy resource consumption vector $\bar{x}^\tau = [\bar{x}_1^\tau \ \bar{x}_2^\tau \ ... \ \bar{x}_n^\tau]$ at each time step τ, and therefore, at time t, it has access to the noisy consumption vectors of the users $\{\bar{x}^\tau\}_{\tau=1}^{t-1}$ in response to the prices $\{\gamma^\tau\}_{\tau=1}^{t-1}$. Since the price response model (1) is also known by the central coordinator, it can exploit this historical data to estimate θ_i^*'s in some way in order to choose γ^t. How well the central coordinator does in this task is measured by the difference in total utility between the optimal choice of prices in (4) and the actual choice of prices. This is referred to as the regret, which can be stated as:

$$R_T = \sum_{t=1}^{T} \sum_{i=1}^{n} [f_i(x_i(\breve{\gamma}_i^t)) - f_i(x_i^t)], \tag{5}$$

In addition to ensuring low regret, the central coordinator also needs to ensure that the constraints in (3) are satisfied with high probability for all time steps. The reason that the central coordinator is not required to ensure constraint satisfaction with probability one is because under the noise model in Assumption 2, the exact price response cannot be known in general. Therefore, the best that the central coordinator can do is to ensure that the constraints are satisfied with high probability. To this end, in the next section we propose a safe pricing algorithm that achieves sublinear regret and satisfies the constraints at each time step with high probability.

III. Safe Price Response Algorithm

The goal of the safe pricing algorithm given by Algorithm 1 is to encourage utility-maximizing demand while ensuring safety in terms of the constraints. Given that in (4), the θ_i^*'s are the only information the central coordinator lacks in order to determine the prices optimally and safely, an estimation of possible θ_i^*'s is crucial for the design of an efficient and safe algorithm. To do so, the algorithm relies on building confidence regions in which the users' unknown parameter vectors θ_i^* lie in with high probability. These confidence regions are essential for two reasons when we design the prices: 1) to implement the principle of OFU to efficiently trade-off exploration and exploitation; 2) to ensure safety with respect to the constraints (3) (in a pessimistic sense).

To define the confidence sets we use a modified version of the confidence region specified in [21]. We first give the least-squares estimator of θ_i^* with regularization paramater $\nu > 0$,

$$\hat{\theta}_i^t = [V_i^t]^{-1} \sum_{s=1}^{t} h_i(\gamma_i^s) \bar{x}_i^s, \ \forall i \in [n], \forall t \in [T], \tag{6}$$

where

$$V_i^t = \nu I + \sum_{s=1}^{t} h_i(\gamma_i^s) h_i(\gamma_i^s)^T, \ \forall i \in [n], \forall t \in [T]. \tag{7}$$

The confidence set then follows.

Theorem 1. *(Theorem 2 in [21] for multiple users) Let Assumptions 1 and 2 hold, fix $\delta \in (0, 1)$ and let*

$$\sqrt{\beta^t} = \sigma \sqrt{m \log\left(\frac{1 + tL^2/\nu}{\delta/n}\right)} + \sqrt{\nu} S, \tag{8}$$

Algorithm 1 Safe Price Response Algorithm

Input: Initialize basis functions h_i for all i, confidence level δ, confidence set $C_i^0 = \{\theta_i \in \mathbb{R}^m : \|\theta_i\| \leq S\}$ for all i, horizon T, noise bound σ, price limit Γ_i for all i, constraints $\{a_{ji}\}_{\forall j,i}, \{c_j^t\}_{\forall j,t}$, and bounds L and S.

1: **for** $t = 1$ to T **do**
2: Calculate $\tilde{x}^t$ with (10).
3: Find a γ_i^t that satisfies (11) for all i.
4: Broadcast γ^t to users.
5: Observe noisy consumption $\bar{x}^t$.
6: Update confidence set C_i^t for all i with (9).
7: **end for**

then θ_i^* is in

$$C_i^t = \{\theta_i \in \mathbb{R}^m : \|\theta_i - \hat{\theta}_i^t\|_{V_i^t} \leq \sqrt{\beta^t}, \|\theta_i\| \leq S, \theta_i \geq \mathbf{0}\} \tag{9}$$

for all $i \in [n]$ and $t \geq 1$ with probability at least $1 - \delta$.

Proof: Given Assumptions 1 and 2, we can use Theorem 2 in [21] for each user. We then apply union bound to substitute in δ/n. Using the fact that $\theta_i \geq \mathbf{0}$ and Assumption 1 to restrict the confidence set completes the proof. ∎

In order to design an effective pricing algorithm, we use this confidence region to both balance exploration and exploitation via the OFU principle, and maintain safety. In the field of online learning, the OFU principle states that the learner should behave as though the unknown (e.g. the unknown paramater) is as favorable as reasonably possible given what has been learned so far [18]. In problems where the unknown is a parameter, OFU-type algorithms use confidence regions to define what parameter values are reasonably possible. These algorithms then find the point in the confidence region that is most favorable and use it to choose the next action. The OFU principle has been successful for a variety of bandit settings (e.g. [2], [20], [21]). Our algorithm implements the OFU principle by finding an *optimistic* consumption vector $\tilde{x}^t$ at each time step; that is, the algorithm finds the consumption vector $x_i(\gamma_i; \theta_i)$ that maximizes the utility subject to the resource constraints (as in (4)), while θ_i is allowed to take any value in C_i^t. This will naturally select the $\theta_i \in C_i^t$ that is most favorable, hence accomplishing OFU. However, the algorithm can only choose prices—not consumption—and therefore cannot assign the consumption $\tilde{x}^t$ to the users directly. To maintain safety while still being optimistic, the algorithm chooses prices *pessimistically* with respect to the safety constraints such that, when $\theta_i \in C_i^t$ is as unfavorable as possible for constraint violation, the consumption will be $\tilde{x}^t$. In doing so, our algorithm utilizes the benefits of OFU, while maintaining safety.

Next, we explain specifically how the optimistic consumption vector $\tilde{x}^t$ is calculated. For time step t, the algorithm uses the confidence region from time step $t-1$ to find the optimistic consumption vector:

$$\begin{aligned} \tilde{x}^t = \arg\max_{x \in \mathbb{R}^n} & \sum_{i=1}^n f_i(x_i) \\ \text{s.t. } & \sum_{i=1}^n a_{ji} x_i \leq c_j^t, \ \forall j \in [p], \\ & 0 \leq x_i \leq \max_{\theta_i \in C_i^{t-1}} h_i(\Gamma_i)^T \theta_i, \ \forall i \in [n]. \end{aligned} \tag{10}$$

This equation finds an optimistic consumption vector because it maximizes utility subject to the safety constraints, while allowing x to take all possible values of $x_i(\gamma_i; \theta_i) = h_i(\gamma_i^t)^T \theta_i$ for $\theta_i \in C_i^{t-1}$ and $\gamma_i \geq \Gamma_i$. To see this, note that h_i is element-wise non-increasing so $h_i(\Gamma_i) \geq h_i(\gamma_i)$ for all $\gamma_i \geq \Gamma_i$. Also note from Assumption 1 that there exists a $y \in \mathbb{R}^m$ such that $h_i(y) = \mathbf{0}$. It follows that $h_i(\gamma_i)^T \theta_i$ can take any value in $[0, h_i(\Gamma_i)^T \theta_i]$ for any $\theta_i \in \mathbb{R}_+^m$. Therefore, restricting θ_i to C_i^{t-1} establishes the range for x_i given in the last constraint of (10). The optimization problem in (10) implements OFU because it maximizes utility while constraining x_i to the range of possible values of $x_i(\gamma_i; \theta_i)$ for $\theta_i \in C_i^{t-1}$; therefore, it will implicitly choose the most favorable $\theta_i \in C_i^{t-1}$.

The solution to (10), $\tilde{x}^t$, is optimistic and satisfies the safety constraints. However, the central coordinator can not directly allocate $\tilde{x}^t$ to the users (which would ensure safety), but instead has to determine prices that are both safe and will result in low regret. Accordingly, we next explain how the algorithm chooses the prices pessimistically given the confidence region and $\tilde{x}^t$. A pessimistic choice implies that in the worst-case scenario, the constraints will not be violated. Since all of the a_{ji}s in (3) are non-negative and $\tilde{x}^t$ satisfies the constraints, constraint violation can be avoided by making sure that $x_i^t \leq \tilde{x}_i^t$ for all $i \in [n]$. Therefore, the algorithm ensures safety by choosing γ_i^t such that $x_i(\gamma_i^t; \theta_i) \leq \tilde{x}_i^t$ for all possible $\theta_i \in C_i^t$. That is, at time step t and for all $i \in [n]$, the algorithm chooses a γ_i^t that satisfies

$$\max_{\theta_i \in C_i^{t-1}} h_i(\gamma_i^t)^T \theta_i = \tilde{x}_i^t. \tag{11}$$

The above equation implies that the maximum possible resource consumption of user i in response to price γ_i^t should be equal to $\tilde{x}_i^t$. For later use, let $\tilde{\theta}_i^t$ be the maximizer of the left hand side of (11), and $\bar{\theta}_i = \arg\max_{\theta_i \in C_i^{t-1}} h_i(\Gamma_i)^T \theta_i$. It needs to be shown that a solution to (11) always exists.

Proposition 1. *Let Assumption 1 hold. Then, there exists a γ_i^t that satisfies* (11) *for all i.*

Proof: Let $\ell_i(\gamma_i) = \max_{\theta_i \in C_i^{t-1}} h_i(\gamma_i)^T \theta_i$. From (10), we know that $\tilde{x}_i^t$ can take any value in $D_i = [0, h_i(\Gamma_i)^T \bar{\theta}_i]$. We show that 1) ℓ_i can attain the maximum and minimum elements of D_i, and 2) ℓ_i is continuous:

1) From Assumption 1, we have that there exists y such that $[h_i(y)]_k = 0$ for all $k \in [m]$. This implies that there exists y such that $\ell_i(y) = 0$. Additionally, it is clear that $\ell_i(\Gamma_i) = h_i(\Gamma_i)^T \bar{\theta}_i$. Therefore, ℓ_i can attain the minimum and maximum elements of D_i.

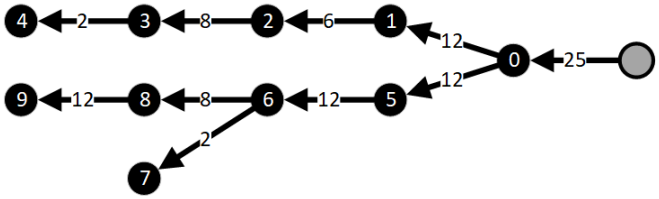

Fig. 1. Distribution network with nodes representing users and edge labels indicating capacity.

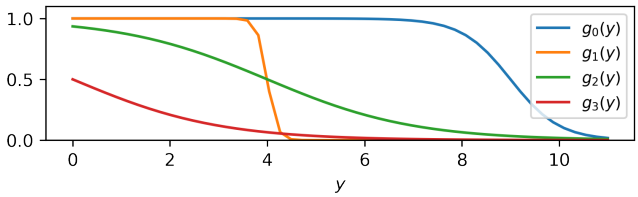

Fig. 2. Each element of basis function g.

2) First, consider the support function of C_i^{t-1}: $S_{C_i^{t-1}}(y) = \max_{z \in C_i^{t-1}}(y^T z)$ where $y \in \mathbb{R}^m$. Note that $S_{C_i^{t-1}}(y)$ is finite because C_i^{t-1} is bounded. Then, the function $S_{C_i^{t-1}}(y)$ is continuous because support functions are convex [25] and a convex function that is finite on all $\mathbb{R}^m$ is continuous [26]. We have that $\ell_i(\gamma_i) = S_{C_i^{t-1}}(h_i(\gamma_i))$. Since the composition of continuous functions is continuous, ℓ_i is continuous as well.

By the intermediate value theorem, it follows that ℓ_i can attain any value in D_i. The proof is completed by noting that everything stated previously applies to all i in $[n]$. ∎

We have now established the three main building blocks of Algorithm 1: 1) Building the confidence region C_i^t given by (9) that contains θ_i^* with high probability for all $i \in [n]$ (Step 6), 2) determining the optimistic resource consumption vector $\tilde{x}^t$ as the solution of (10) (Step 2), and 3) choosing the prices γ^t such that (11) holds for all $i \in [n]$ (Step 3). In the next section, we prove that the prices produced by Algorithm 1 induce safe resource consumption vectors that satisfy the constraints at all time steps and achieve a sublinear regret.

IV. Safety and Regret Analysis

In this section, we will prove that the prices set by Algorithm 1 induce resource consumption vectors that 1) satisfy the constraints for all $t \in [T]$, and 2) achieve $\mathcal{O}(\log(T)\sqrt{T})$ regret after T time steps. Since regret is a well-defined metric only when the solutions are feasible (otherwise an infeasible solution can have a higher objective value and negative regret), we will first prove the safety of the algorithm. Building on the intuition established in the previous section, we formally state the safety guarantee:

Theorem 2. *Let Assumptions 1 and 2 hold. Then Algorithm 1 will ensure that*

$$\sum_{i=1}^{n} a_{ji} x_i^t \leq c_j^t \quad \forall j \in [p], \quad \forall t \in [T], \tag{12}$$

with probability at least $1 - \delta$.

Proof: From (11), $\tilde{x}_i^t = \max_{\theta_i \in C_i^{t-1}} h_i(\gamma_i^t)^T \theta_i$. Then with probability at least $1 - \delta$, it follows that $\tilde{x}_i^t \geq h_i(\gamma_i^t)^T \theta_i^* = x_i^t$. Since all a_{ji}s are non-negative, we have that $\sum_{i=1}^n a_{ji} x_i^t \leq \sum_{i=1}^n a_{ji} \tilde{x}_i^t \leq c_j^t$. This completes the proof. ∎

According to Theorem 2, the prices set by Algorithm 1 guarantee that the users consume feasible amount of resources at all time steps, i.e., the algorithm produces safe prices. Therefore, regret is a valid metric of performance. We next prove that the regret incurred by the prices set by Algorithm 1 is sublinear in T.

Theorem 3. *Let Assumptions 1,2 and 3 hold. Then, the cumulative regret of Algorithm 1 satisfies*

$$\begin{aligned} R_T \leq & nM \max(LS, 1) \sqrt{8Tm \log\left(1 + \frac{TL^2}{m\nu}\right)} \\ & \times \left(\sigma \sqrt{m \log\left(\frac{n}{\delta}\left(1 + \frac{TL^2}{\nu}\right)\right)} + \sqrt{\nu} S\right) \end{aligned} \tag{13}$$

with probability at least $1 - \delta$.

Proof outline: We first use the Lipschitz assumption on f_i's to put the regret in terms of the resource consumption. Then the analysis proceeds similar to linear bandit analysis such as in [21]. However, unlike the bandit case, our algorithm is optimistic in the *total* utility of all users, i.e. $\sum_{i=1}^n f_i(\tilde{x}_i^t) \geq \sum_{i=1}^n f_i(\check{x}_i^t)$. As a result, our analysis requires careful handling of the regret due to each user.

The complete proof is given in Appendix A. According to Theorem 3, the regret incurred by Algorithm 1 is $\mathcal{O}(nm \log(Tn)\sqrt{T})$. Note that the factor of n in the regret bound is due to the fact that the definition of regret is the difference in *total* utility across all users. In fact, the *average* regret over the users (R_T/n) is $\mathcal{O}(m \log(Tn)\sqrt{T})$.

V. Numerical Experiment

To validate the algorithm and demonstrate how it can be applied, we simulate the algorithm choosing electricity prices for the users of a small power distribution system. The architecture of the distribution network is shown in Fig. 1.

We assume that the grid operator does not know enough about each user to assign different h_i's to each one. Therefore, we make the basis functions the same for all users, such that $g = h_i$ for all i. We use logistic-type functions for the elements of g:

$$g_k(y; t_k, d_k) = \frac{1}{1 + \exp((y - t_k)/d_k)}, \ \forall k \in [m]. \tag{14}$$

We use four different functions ($m = 4$) that represent the price response of different classes of appliances that users might use. These functions are shown in Fig. 2 which are

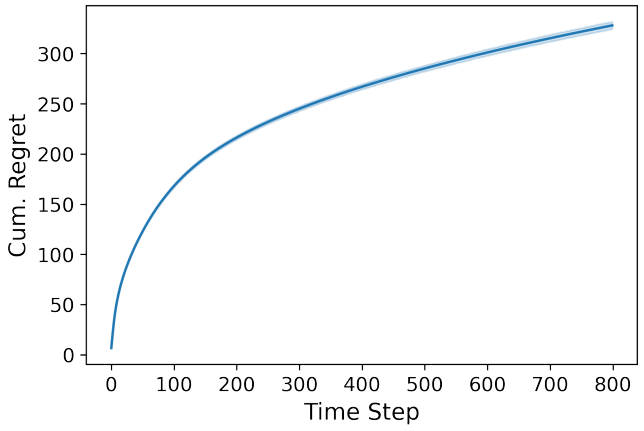

Fig. 3. The cumulative regret of the Safe Price Response Algorithm choosing electricity prices for a simulated power distribution system. The data is averaged over 100 trials with the 95% confidence interval shown.

defined with $t = \{9, 4, 4, 0\}$ and $d = \{0.5, 0.1, 1.5, 1.5\}$. The following list explains what the different elements of g represent.

- g_0: critical appliances (basic heating, lighting, cooking)
- g_1: non-critical appliances that will be on either entirely or not at all (electric vehicle charging)
- g_2: non-critical appliances that can be used in variable quantities (additional heating, cooling)
- g_3: luxury appliances (television, gaming)

Note that each of these functions don't represent the price response of each appliance in isolation, rather they represent the additional price response of that appliance given that the more fundamental appliances are in use.

The utility functions are chosen as shifted log functions:

$$f_i(x_i) = \alpha_i \log(x_i + 0.1) \quad \forall i \tag{15}$$

The shift is to ensure that the functions are Lipschitz continuous. The value of each α_i is sampled uniformly: $\alpha_i \sim U[0.5, 1]$ iid $\forall i$. The θ_is are chosen as $[\theta_i^*]_k \sim U[0.5, 1]$ iid $\forall i, k$. Also, $\Gamma_i = 0.1$ for all i. The noise variable $\mu_i^t \sim N(0, \sigma^2)$ iid $\forall i, t$, where $\sigma^2 = 0.2$.

Next, we discuss how the algorithm is implemented in the simulation. First, note that f_i in (15) is concave, so (10) is convex and can be solved efficiently with any convex solver (we use CVXPY [27], [28]). Additionally, note that the price update equation in (11) can be formulated as a scalar root-finding problem (i.e. it seeks the γ_i^t that yields $r(\gamma_i^t) = \max_{\theta_i \in C_i^{t-1}} h_i(\gamma_i^t)^T \theta_i - \tilde{x}_i^t = 0$). Since $\max_{\theta_i \in C_i^{t-1}} h_i(\gamma_i^t)^T \theta_i$ is a convex optimization problem, $r(\gamma_i^t)$ can be evaluated by solving a convex optimization problem. Therefore, we calculate (11) by using a scalar root-finding solver on $r(\gamma_i^t)$ (we use Scikit-learn [29]), and using a convex optimization solver (we use CVXPY [27], [28]) to evaluate $r(\gamma_i^t)$ each time it's called by the root-finding solver.

One hundred simulations were run for 800 time steps with different realizations of $\{\mu_i^t\}_{\forall i,t}$ for each simulation. From these trials, there were zero constraint violations. The average cumulative regret and a 95% confidence interval over all simulations is shown in Fig. 3.

VI. Conclusion

In this work, we posed a novel safe price design problem motivated by applications in cyber-physical systems with humans in the loop. To address this problem, we proposed an algorithm that first finds an optimistic consumption level and then finds a price that will achieve that consumption level in the worst case in terms of safety. Analysis shows that, with high probability, this algorithm maintains safety and achieves regret $\mathcal{O}(\log(T)\sqrt{T})$. Additionally, a numerical experiment demonstrates how this algorithm can be applied to a demand response management problem for a power distribution network. The numerical results from this experiment agree with the safety analysis and regret analysis.

References

[1] I. Usmanova, A. Krause, and M. Kamgarpour, "Safe convex learning under uncertain constraints," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 2106–2114.

[2] S. Amani, M. Alizadeh, and C. Thrampoulidis, "Linear stochastic bandits under safety constraints," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[3] Y. Sui, A. Gotovos, J. Burdick, and A. Krause, "Safe exploration for optimization with gaussian processes," in *International conference on machine learning*. PMLR, 2015, pp. 997–1005.

[4] S. Chaudhary and D. Kalathil, "Safe online convex optimization with unknown linear safety constraints," *arXiv preprint arXiv:2111.07430*, 2021.

[5] P. Samadi, A.-H. Mohsenian-Rad, R. Schober, V. W. Wong, and J. Jatskevich, "Optimal real-time pricing algorithm based on utility maximization for smart grid," in *2010 First IEEE International Conference on Smart Grid Communications*. IEEE, 2010, pp. 415–420.

[6] N. Li, L. Chen, and S. H. Low, "Optimal demand response based on utility maximization in power networks," in *2011 IEEE power and energy society general meeting*. IEEE, 2011, pp. 1–8.

[7] N. Mehr, J. Lioris, R. Horowitz, and R. Pedarsani, "Joint perimeter and signal control of urban traffic via network utility maximization," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2017, pp. 1–6.

[8] D. P. Palomar and M. Chiang, "A tutorial on decomposition methods for network utility maximization," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 8, pp. 1439–1451, 2006.

[9] I. Usmanova, A. Krause, and M. Kamgarpour, "Safe non-smooth black-box optimization with application to policy search," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 980–989.

[10] M. Fereydounian, Z. Shen, A. Mokhtari, A. Karbasi, and H. Hassani, "Safe learning under uncertain objectives and constraints," *arXiv preprint arXiv:2006.13326*, 2020.

[11] F. Berkenkamp, A. Krause, and A. P. Schoellig, "Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics," *Machine Learning*, pp. 1–35, 2021.

[12] A. Pacchiano, M. Ghavamzadeh, P. Bartlett, and H. Jiang, "Stochastic bandits with linear constraints," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 2827–2835.

[13] A. Moradipari, C. Thrampoulidis, and M. Alizadeh, "Stage-wise conservative linear bandits," *Advances in neural information processing systems*, vol. 33, pp. 11 191–11 201, 2020.

[14] M. Chiang, S. H. Low, A. R. Calderbank, and J. C. Doyle, "Layering as optimization decomposition: A mathematical theory of network architectures," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 255–312, 2007.

[15] F. P. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control for communication networks: shadow prices, proportional fairness and stability," *Journal of the Operational Research society*, vol. 49, no. 3, pp. 237–252, 1998.

[16] S. H. Low and D. E. Lapsley, "Optimization flow control. i. basic algorithm and convergence," *IEEE/ACM Transactions on networking*, vol. 7, no. 6, pp. 861–874, 1999.
[17] D. P. Palomar and M. Chiang, "Alternative distributed algorithms for network utility maximization: Framework and applications," *IEEE Transactions on Automatic Control*, vol. 52, no. 12, pp. 2254–2269, 2007.
[18] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.
[19] N. Abe and P. M. Long, "Associative reinforcement learning using linear probabilistic concepts," in *ICML*. Citeseer, 1999, pp. 3–11.
[20] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2, pp. 235–256, 2002.
[21] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," *Advances in neural information processing systems*, vol. 24, 2011.
[22] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," 2008.
[23] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Transactions on networking*, vol. 8, no. 5, pp. 556–567, 2000.
[24] T. Lan, D. Kao, M. Chiang, and A. Sabharwal, *An axiomatic theory of fairness in network resource allocation*. IEEE, 2010.
[25] S. Boyd, S. P. Boyd, and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.
[26] R. T. Rockafellar, *Convex analysis*. Princeton university press, 2015.
[27] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.
[28] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd, "A rewriting system for convex optimization problems," *Journal of Control and Decision*, vol. 5, no. 1, pp. 42–60, 2018.
[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

APPENDIX

A. Proof of Theorem 3

Given that (10) is optimistic and Assumption 3, we have with probability at least $1-\delta$ that

$$r_t = \sum_{i=1}^{n}[f_i(x_i(\check{\gamma}_i^t)) - f_i(x_i^t)] \tag{16}$$
$$\leq \sum_{i=1}^{n}[f_i(\tilde{x}_i) - f_i(x_i^t)] \tag{17}$$
$$\leq \sum_{i=1}^{n}|f_i(\tilde{x}_i) - f_i(x_i^t)| \tag{18}$$
$$\leq M\sum_{i=1}^{n}|\tilde{x}_i - x_i^t| \tag{19}$$
$$= M\sum_{i=1}^{n}|h_i(\gamma_i^t)^T\tilde{\theta}_i - h_i(\gamma_i^t)^T\theta_i^*| \tag{20}$$

Let $r'_{t,i} = |h_i(\gamma_i^t)^T(\tilde{\theta}_i^t - \theta_i^*)|$ and $R'_{T,i} = \sum_{t=1}^{T} r'_{t,i}$. Note that

$$R_T \leq M\sum_{t=1}^{T}\sum_{i=1}^{n} r'_{t,i} \tag{21}$$
$$= M\sum_{i=1}^{n} R'_{T,i} \tag{22}$$

We can then bound $R'_{T,i}$ using the standard linear stochastic bandit analysis.

$$r'_{t,i} = |h_i(\gamma_i^t)(\tilde{\theta}_i^t - \theta_i^*)| \tag{23}$$
$$= |h_i(\gamma_i^t)(\tilde{\theta}_i^t - \hat{\theta}_i^{t-1} + \hat{\theta}_i^{t-1} - \theta_i^*)| \tag{24}$$
$$\leq \|h_i(\gamma_i^t)\|_{[V_i^{t-1}]^{-1}}\|\tilde{\theta}_i^t - \hat{\theta}_i^{t-1} + \hat{\theta}_i^{t-1} - \theta_i^*\|_{V_i^{t-1}} \tag{25}$$
$$\leq \|h_i(\gamma_i^t)\|_{[V_i^{t-1}]^{-1}} \times (\|\tilde{\theta}_i^t - \hat{\theta}_i^{t-1}\|_{V_i^{t-1}} + \|\hat{\theta}_i^{t-1} - \theta_i^*\|_{V_i^{t-1}}) \tag{26}$$
$$\leq 2\|h_i(\gamma_i^t)\|_{[V_i^{t-1}]^{-1}}\sqrt{\beta^{t-1}} \tag{27}$$

To use the bandit analysis, we need a trivial bound on $r'_{t,i}$. We know that $\|h_i(\gamma_i^t)\| \leq L$, which implies

$$0 \leq h_i(\gamma_i^t)^T\theta_i^* \leq \|h_i(\gamma_i^t)\|\|\theta_i^*\| \leq LS, \tag{28}$$
$$0 \leq h_i(\gamma_i^t)^T\tilde{\theta}_i^t \leq \|h_i(\gamma_i^t)\|\|\tilde{\theta}_i^t\| \leq LS, \tag{29}$$

so we have the trivial bound on $r'_{t,i}$:

$$r'_{t,i} = |h_i(\gamma_i^t)^T\tilde{\theta}_i^t - h_i(\gamma_i^t)^T\theta_i^*| \leq LS \tag{30}$$

Therefore, assuming that T is large enough such that $\beta^T \geq 1$ (for simplicity), we have that

$$r'_{t,i} \leq \min(2\|h_i(\gamma_i^t)\|_{[V_i^{t-1}]^{-1}}\sqrt{\beta^{t-1}}, LS) \tag{31}$$
$$\leq 2\max(LS,1)\min(\|h_i(\gamma_i^t)\|_{[V_i^{t-1}]^{-1}}\sqrt{\beta^{t-1}}, 1) \tag{32}$$
$$\leq 2\max(LS,1)\sqrt{\beta^T}\min(\|h_i(\gamma_i^t)\|_{[V_i^{t-1}]^{-1}}, 1) \tag{33}$$
$$r'^2_{t,i} \leq 4\max(L^2S^2,1)\beta^T\min(\|h_i(\gamma_i^t)\|^2_{[V_i^{t-1}]^{-1}}, 1) \tag{34}$$

We can use the so-called elliptical potential lemma to bound this.

Lemma 1. *(Lemma 11 from [21]) Let Assumption 1 hold. Then,*

$$\sum_{t=1}^{T}\min(\|h_i(\gamma_i^t)\|^2_{[V_i^{t-1}]^{-1}}, 1) \leq 2m\log(1 + TL^2/(m\nu)) \tag{35}$$

Then, we can apply Cauchy-Schwarz on $R'_{T,i}$ to get the bound for R_T:

$$R_T \leq M\sum_{i=1}^{n} R'_{T,i} \leq M\sum_{i=1}^{n}\sqrt{T\sum_{t=1}^{T} r'^2_{t,i}} \tag{36}$$

Plugging in (34) and (8) completes the proof.