Safe Pricing Mechanisms for Distributed Resource Allocation with Bandit Feedback

Spencer Hutchinson, Berkay Turan, and Mahnoosh Alizadeh

***Abstract*—In societal-scale infrastructures, such as electric grids or transportation networks, pricing mechanisms are often used as a way to shape users' demand in order to lower operating costs and improve reliability. Existing approaches to pricing design for safety-critical networks often require that users are queried beforehand to negotiate prices, which has proven to be challenging to implement in the real-world. To offer a more practical alternative, we develop learning-based pricing mechanisms that require no input from the users. These pricing mechanisms aim to maximize the utility of the users' consumption by gradually estimating the users' price response over a span of T time steps (e.g., days) while ensuring that the infrastructure network's safety constraints that limit the users' demand are satisfied at all time steps. We propose two different algorithms for the two different scenarios when, 1) the utility function is chosen by the central coordinator to achieve a social objective and 2) the utility function is defined by the price response under the assumption that the users are self-interested agents. We prove that both algorithms enjoy $\tilde{\mathcal{O}}(T^{2/3})$ regret with high probability. We then apply these algorithms to demand response pricing for the smart grid and numerically demonstrate their effectiveness.**

I. Introduction

In safety-critical infrastructure systems, such as power and transportation networks, prices or tolls are often used to improve efficiency while ensuring safety constraints (e.g. power line or road capacities) are honored. Optimal design of such prices requires knowledge of self-interested users' preferences/utility functions, which are not often apriori available to any central coordinator. As such, one popular approach is to employ distributed resource allocation mechanisms such as network utility maximization (NUM), e.g., [1], [2]. These approaches are well suited for finding optimal shadow prices in such multi-agent network systems via prescribed interactions between agents with private preferences [3]–[5]. After the distributed optimization protocol converges, optimal prices may be posted, and the users will adjust their demand in response to the posted prices. However, in spite of their popularity in research papers, such resource allocation mechanisms have not been widely implemented in real-world safety-critical networks, such as power systems, due to several factors including: 1) the need for back and forth communications with users to negotiate over optimal prices; 2) they require fully automated personal demand management mechanisms to be adopted by each individual user in order to implement the distributed optimization protocol; 3) they require all users to take part in the protocol and cooperate with the central entity.

To circumvent these issues, in this paper, we adopt an alternative viewpoint wherein, instead of employing distributed mechanisms to find optimal prices given unknown user preferences, the central coordinator aims to *learn* the users' preferences over a span of T days through repeated interactions. Each day, the central coordinator posts a price and observes the users' noisy response through their resource consumption, and based on these observations, refines its knowledge of the users' preferences.

Adopting a learning-based pricing framework introduces several novel challenges that are not present in conventional approaches. The first challenge is ensuring the infrastructure's *safety constraints* when posting prices each day. As the price response of the users is being learnt and is hence not entirely known, the central coordinator needs to ensure that any posted price will not lead to user demand that will violate the network's constraints (e.g., power flow constraints in demand response applications). The second challenge is ensuring the efficiency of the posted prices. Despite the fact that the central coordinator lacks full knowledge of the users' preferences, it still needs to ensure that the aggregate utility of the users due to the resource consumption is high over the span of T days.

To model this problem, we study two different frameworks, *Safe Price Response (SPR)* and *Safe Utility Maximization (SUM)*. The SPR and the SUM problems differ in that, in the former, the utility functions for different user groups are chosen by the central coordinator and, in the latter, the utility functions are defined by the price response function under the assumption that the users are self-interested agents. SPR is appropriate when the optimization objective (i.e. the utility function) is a design choice of the central coordinator to achieve social objectives, while SUM is appropriate when the users are assumed to be self-interested agents and the goal is to maximize the total private benefit of these agents.

The contributions of this work are summarized as follows:

- We introduce two new learning-based frameworks for pricing design in safety-critical networks that are applicable for the two typical settings where 1) the utility functions are designed by a central entity and 2) the utility functions are defined by the user's response to prices under the assumption that the users behave as self-interested agents.
- Relative to prior works such as [3]–[5], our frameworks are more practical for safety-critical pricing applications because they *do not* require prices to be negotiated with users beforehand to ensure safety;

This work was supported by NSF grant #1847096.

S. Hutchinson, B. Turan and M. Alizadeh are with Department of Electrical and Computer Engineering, University of California-Santa Barbara, Santa Barbara, CA 93106 USA (email: shutchinson@ucsb.edu; bturan@ucsb.edu; alizadeh@ucsb.edu).

- We propose two bandit algorithms for these frameworks and prove that they enjoy sublinear regret and satisfy the safety constraints at all rounds with high probability;
- We apply these algorithms to demand response in the smart grid and demonstrate their effectiveness through simulation of a real distribution network.

Related Work: Evidently, this work is related to existing approaches for demand management in safety-critical infrastructure. Several works [6]–[8] take a mechanism design perspective, where the system users are modeled as strategic agents in their interactions with the central coordinator (e.g. a user may submit untruthful estimates of future demand to reduce their own costs). In this paper, we take a different perspective in that we model the users' demand to be a private function of the price that does not change in response to the central coordinator's pricing policy. This is more closely related to distributed resource allocation approaches, which are useful for finding optimal shadow prices in multi-agent systems with private utility functions. The most relevant distributed resource allocation framework is Network Utility Maximization (NUM), which allows for a resource allocation problem with private utility functions to be decomposed such that it can be solved in a distributed fashion where a central coordinator communicates with each user [9], [10]. NUM has been applied to congestion control for internet networks [4], [5] as well as the control of power and transportation systems [1], [2], [11]. Recently, a NUM algorithm that respects stage-wise constraints was presented in [12]. Another class of distributed resource allocation approaches consider a fully distributed system with cooperative agents and limited information sharing [3], [13]. Our problem formulations are different than existing distributed resource allocation approaches in that there is noisy bandit feedback from users, a parametric form for the price response, and stage-wise safety constraints that must always be respected in spite of uncertainty about the users' response.

Given that this work is focused on learning in safety-critical applications, learning-based control techniques (surveyed in [14]) are particularly relevant because they use previous data to improve performance while ensuring safe operation. This includes learning-based adaptive control [15], [16], learning-based robust control [17], learning-based robust MPC [18] and model predictive safety certification [19]. Although we use similar techniques to ensure safety, our problem fundamentally differs from the aforementioned approaches because the algorithm in our problem interacts with the same static environment at every time step, albeit with progressively more information (i.e. the environment does not evolve as a dynamical system).

In addition to learning-based control, there is also relevant literature on safe optimization. This includes constrained optimization algorithms with unknown constraints and feasible iterates where the constraints are either linear [20] or nonlinear [21], as well as online convex optimization with unknown constraints that need to be satisfied in the long term [22] or constraints that need to be satisfied stage-wise [23]. Several works have also considered the problem of safe learning under a Gaussian process prior [24], [25]. However, none of these consider a multi-agent optimization problem with stage-wise constraints as we do here.

Most relevant to our work, prior work has also studied safety in linear stochastic bandits, where the reward is an unknown linear function of the action and the learner receives noisy bandit feedback of this action. Different types of safety constraints have been considered, including constraints on the objective [26], constraints on another linearly parameterized function with bandit feedback [27] and constraints that are linear with respect to the decision variable and the unknown parameter [28]. Although our algorithm and analysis are inspired by [28], the key difference is that we have multiple constraints that jointly apply to multiple users (or equivalently, bandits) which necessitates different analysis techniques. Refer to Section II-C for a more detailed comparison of the analysis.

This work studies a similar, but more general problem as the conference paper [29]. In particular, [29] only considers the SPR problem formulation with $a_{ji} \geq 0$ for all $i \in [n], j \in [p]$, which allows for a simpler algorithm and analysis than what is presented here. The SUM formulation was not considered in [29].

Organization: Our study of the SPR and SUM problems are located in Section II and III respectively. Additionally, the algorithms developed for these problems are applied to demand response in smart grid in Section IV.

Notation: For a positive integer n, we use $[n]$ to refer to the set of positive integers from 1 to n inclusive. For a vector or matrix A, its transpose is denoted $A^\top$. When A is square, its minimum and maximum eigenvalues are denoted $\lambda_{min}(A)$ and $\lambda_{max}(A)$ respectively. For a vector v and positive definite matrix P, we use $\|v\|$ to refer to the euclidean norm of v and $\|v\|_P$ to refer to $\sqrt{v^\top P v}$. For a d-dimensional vector or d-tuple v and positive integer i, we denote the ith element of v as v_i. In d dimensions, the non-negative orthant and positive orthant are referred to as $\mathbb{R}^d_+$ and $\mathbb{R}^d_{++}$ respectively. We use $\tilde{\mathcal{O}}$ to refer to Big-O notation that ignores logarithmic factors. A vector of zeros and a vector of ones are indicated by $\mathbf{0}$ and $\mathbf{1}$ respectively, where the size is inferred by context. For vectors u and v, the notation $u \succ v$ indicates that each element of u is greater than the corresponding element of v and $u \succeq v$ indicates that each element of u is greater than or equal to the corresponding element of v. For a set A, $\mathbf{int}A$ refers to the interior of A and $\mathbf{bd}A$ refers to the boundary of A. The domain of a function f is denoted by $\mathbf{dom}f$.

II. The Safe Price Response (SPR) Problem

In this section, we first describe the SPR problem, and then present an algorithm and theoretical performance guarantees to address this problem. The problem setup, algorithm, and regret analysis are presented in Sections II-A, II-B, and II-C.

A. Problem Setup

We pose a resource allocation problem involving repeated interactions between a central coordinator and n users. At each time step t in horizon $[T]$, there is an interaction between each user $i \in [n]$ and the central coordinator in which the central coordinator chooses a price γ_i^t and user i responds with a

resource consumption x_i^t. The physical limits of the system are specified by p linear constraints on the consumption vector $x^t = [x_1^t \ x_2^t \ ... \ x_n^t]^\top$. The objective of the central coordinator is to maximize the total user utility (defined later), while ensuring that the constraints are satisfied at every $t \in [T]$.

We adopt a parametric form for the price response function of the users. Specifically, we assume that the average resource consumption of user i in response to the price γ_i^t is given by its average price response function,

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

where $\theta_i^* \in \mathbb{R}_+^m$ is a nonzero parameter that is unknown to the central coordinator and $h_i : \mathbb{R} \to \mathbb{R}_+^m$ is a known continuous and non-increasing function where $\mathbf{dom} h_i = \mathbb{R}$. Equation (1) models each user's price response as an unknown mix of given *price response signatures*, where h_i specifies the set of possible price response signatures that may be present in the user population and θ_i^* specifies what (unknown) mix of these price response signatures make up user i's price response.

For example, in the electricity demand response set up, the total price response of each user to posted prices is composed of the sum of usage of individual flexible appliances (e.g., electric vehicle or dishwasher) and flexible appliances have a limited number of ways to respond to prices (which is justified given the automated nature of price response from home energy management systems, the limited types of flexible appliances, and the common electricity load patterns of electricity customers). For example, time shiftable loads with similar energy demand and similar deadlines would show similar price response signatures. The response of an electric vehicle to posted prices can be determined by the charging rate, the amount of required charge, and the charging deadline. If these parameters were known, the response can be fully determined. However, since this is not the case, we assume that each appliance can have one of a number of known price response signatures captured by $h_i(\gamma_i^t)$. The central coordinator does not know the exact combination of active price response signatures in each user's home (captured by θ_i^*) and as such, needs to learn this information by choosing prices and observing the electricity usage of the homes. We provide a more in-depth discussion of the electricity demand response example in Section IV.

The average price response function is non-increasing by definition, which is natural as consumption of a resource will generally not increase as price increases. Also, due to inherent stochasticities present in the users' behaviors, the central coordinator observes the average resource consumption with some additive noise μ_i^t. Specifically, on day t, the central coordinator observes the following response:

$$\bar{x}_i^t = x_i(\gamma_i^t; \theta_i^*) + \mu_i^t. \tag{2}$$

We assume the following noise model on μ_i^t, which is often used in similar problems (e.g. [23], [28], [30]).

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

In choosing the price vector $\gamma^t = [\gamma_1^t \ \gamma_2^t \ ... \ \gamma_n^t]^\top$ for the users at each time step $t \in [T]$, there are various objectives that the central coordinator might have depending on the specific application. For example, in infrastructure systems that supply critical resources, it is important that the allocation of resources is fair such that under-served communities are not charged high prices, or large consumers do not block access to resources. In any case, the central coordinator can design utility functions for each user to achieve the objective at hand. Utility functions map the resource consumption of a user to utility and have been extensively studied (e.g. [31]–[33]). Here, the utility function for user i is the strictly increasing function $f_i : \mathbb{R} \to \mathbb{R}$, which means that the total utility for the system at time step t is $\sum_{i=1}^n f_i(x_i^t)$.

Despite the high utility that comes with unrestricted consumption, there are physical limits on the system that restrict which consumption vectors are allowable. These limits are specified by p linear constraints on the users' consumptions, such that the set of feasible consumption vectors is compact and defined as

$$\bar{E} = \left\{ x \in \mathbb{R}^n : \sum_{i=1}^n a_{ji} x_i \leq c_j, \forall j \in [p] \right\}, \tag{3}$$

where $\{a_{ji}\}_{i \in [n], j \in [p]}$ and $\{c_j\}_{j \in [p]}$ are known to the central coordinator. Since the central coordinator only has access to noisy observations of the price response, it is in general impossible to design any method that enforces constraint (3) deterministically over the course of T days without additional (but unrealistic) assumptions. As such, we take the next alternative, which is to slightly relax this requirement of safety and ensure it with a high probability jointly throughout the T day operating time of our system. That is, the central coordinator needs to ensure that every consumption vector x^t is in $\bar{E}$ for all t in $[T]$ with high probability. Note that this is different from a regular chance constraint, which ensures constraint satisfaction with a certain probability per time step, meaning that the violation probability would compound as T grows. Since any feasible algorithm will ensure that all consumption vectors are in $\bar{E}$ with high probability, the following Lipschitz assumption on f_i only needs to hold for feasible consumption vectors.

Assumption 2. *For all $i \in [n]$, the utility function f_i is M-Lipschitz such that $|f_i(x_i^1) - f_i(x_i^2)| \leq M |x_i^1 - x_i^2|$ for all x^1, x^2 in $\bar{E}$.*

Given the model that has been specified so far, we can see that if the central coordinator had full information (i.e. knew $\{\theta_i^*\}_{\forall i \in [n]}$) they would choose the price for every time step as

$$\gamma^* \in \arg\max_{\gamma \in \bar{D}} \sum_{i=1}^n f_i(x_i(\gamma_i; \theta_i^*)), \tag{4}$$

where $\gamma = [\gamma_1 \ \gamma_2 \ ... \ \gamma_n]^\top$ and

$$\bar{D} = \left\{ \gamma \in \mathbb{R}^n : \sum_{i=1}^n a_{ji} x_i(\gamma_i; \theta_i^*) \leq c_j, \forall j \in [p] \right\}, \tag{5}$$

which we call the feasible price set. The central coordinator cannot simply solve (4) and choose γ^* immediately

because the θ_i^* are unknown to them. Instead, the central coordinator uses the information from previous time steps (i.e. $\{(\gamma^\tau, x^\tau)\}_{\tau=1}^{t-1}$) to choose the current price γ_i^t. The central coordinator's performance in this task is measured by how close the total realized utility is to the optimal utility over T time steps, which is referred to as regret:

$$R_T = \sum_{t=1}^{T} \sum_{i=1}^{n} \Big[f_i\big(x_i(\gamma_i^*; \theta_i^*)\big) - f_i\big(x_i(\gamma_i^t; \theta_i^*)\big) \Big] \tag{6}$$

The central coordinator's objective is to ensure that there is low regret and that, with high probability, every γ^t is in $\bar{D}$ for all $t \in [T]$.

Given the stated problem, we make a technical assumption on the price response function in the following.

Assumption 3. *For all $i \in [n]$, there exists a constant S such that $\|\theta_i^*\| \leq S$. Also, h_i satisfies $\lim_{\gamma_i \to -\infty} h_i(\gamma_i) = \boldsymbol{\infty}$ and $\lim_{\gamma_i \to \infty} h_i(\gamma_i) = \mathbf{0}$ for all $i \in [n]$.*

The first part of this assumption ensures that the norm of θ_i^* is bounded by S which is standard in the bandit learning literature, e.g. [30], [34]. In a real-world setting, an appropriate S can be found with domain knowledge. For example, in the demand response setup, an appropriate S can be chosen by finding the worst-case estimates of the size of each appliance in each home. The second part of Assumption 3 ensures that there is a price (which may be negative) that will persuade the user to consume any non-negative quantity of the resource. This will generally be satisfied in real-world settings because the price could be negative, i.e. the central coordinator would pay the user to consume the resource. Such negative prices are occasionally used in power systems, for example.

As defined thus far, the problem does not provide the central coordinator with enough information to choose initial prices that satisfy the constraints. To remedy this, we ensure by assumption that the prior knowledge on θ^*, i.e. the fact that $\|\theta_i^*\| \leq S$ for all i in $[n]$, is enough information for the central coordinator to construct a set of prices that is strictly within $\bar{D}$. To state such an assumption, we first define the initial confidence set for $\theta^* = (\theta_1^*, \theta_2^*, ..., \theta_n^*)$ as $C^0 = C_1^0 \times C_2^0 \times ... \times C_n^0$, where

$$C_i^0 = \{\theta_i \in \mathbb{R}_+^m : \|\theta_i\| \leq S\} \tag{7}$$

for all i in $[n]$. Since θ^* is known to be in C^0, it follows that

$$D^0 := \Bigg\{ \gamma \in \mathbb{R}^n : \sum_{i=1}^{n} a_{ji} \theta_i^T h_i(\gamma_i) \leq c_j - \zeta, \quad \forall j \in [p], \forall \theta \in C^0 \Bigg\} \tag{8}$$

is a subset of $\bar{D}$ for any $\zeta \geq 0$. In Assumption 4, we assume that D^0 is nonempty for some ζ, providing the algorithm with a set of prices that are initially known to strictly satisfy the constraints. We will also consider a set that is known to be larger than $\bar{D}$ given that $\theta^* \in C^0$,

$$\hat{D}^0 := \Bigg\{ \gamma \in \mathbb{R}^n : \exists \theta \in C^0 \text{ s.t. } \sum_{i=1}^{n} a_{ji} \theta_i^T h_i(\gamma_i) \leq c_j, \quad \forall j \in [p] \Bigg\}. \tag{9}$$

Algorithm 1 Safe Price Response Algorithm

Input: $\{h_i\}_{i\in[n]}$, $\{a_{ji}\}_{i\in[n],j\in[p]}$, $\{c_j\}_{j\in[p]}$, $\{f_i\}_{i\in[n]}$, S, L

for $t = 1$ to T' **do**
  Broadcast $\gamma^t \sim \text{Unif}(D^0)$ to the users.
  Observe noisy consumption $\bar{x}^t$.
end for
Construct confidence set $C^{T'}$ with (12).
Construct safe price set $D^{T'}$ with (14).
for $t = T' + 1$ to T **do**
  Find optimistic price γ^t with (15).
  Broadcast γ^t to the users.
  Observe noisy consumption $\bar{x}^t$.
  Update confidence set C^t with (12).
  Update safe price set D^t with (14).
end for

Since each price in $\hat{D}^0$ only needs to satisfy the constraints for *some* $\theta \in C^0$, any algorithm that incorporates the information that $\theta^* \in C^0$ will only choose prices that are in $\hat{D}^0$. Therefore, by assuming that the norm of h_i is bounded for any price in $\hat{D}^0$, Assumption 4 ensures that the norm of h_i is bounded by a constant L for any prices that are chosen by the algorithm. In the real-world, such a constant L can be simply calculated given that h_i and $\hat{D}^0$ are known to the central coordinator.

Assumption 4. *There exists positive constants ζ and κ such that the initial safe set D^0 is nonempty and $|a_{ji}| \leq \kappa$ for all i in $[n]$ and j in $[p]$. Additionally, there exists a positive constant L such that $\max_{i\in[n]} \|h_i(\gamma_i)\| \leq L$ for all $\gamma \in \hat{D}^0$. Also, there does not exist a nonzero $\alpha_i \in \mathbb{R}^m$ for each $i \in [n]$ such that $\alpha_i^\top h_i(\gamma_i) = 0$ for all $[\gamma_1 \ \gamma_2 \ ... \ \gamma_n]^\top$ in D^0.*

In addition to what has already been discussed, Assumption 4 also specifies that 1) $|a_{ji}|$ is bounded by some constant κ for all i, j and that 2) the elements of h_i are linearly independent on D^0. Note that point 1 is mild given that the value of $|a_{ji}|$ is known and will be finite in any real-world application and therefore such a κ can simply be calculated. The linear independence assumption of point 2 specifies that the elements of h_i are not scalar multiples of each other for all prices in D^0. This ensures that sampling D^0 will provide sufficient information about every dimension of θ_i^*. In practice, this requires that the selected price response signatures are sufficiently different, which is a design choice.

With the problem established, we develop an appropriate algorithm in the next section.

B. Proposed Algorithm

The proposed algorithm (Algorithm 1) first performs pure exploration by choosing prices in the initial safe set D^0 for an appropriately chosen duration T', and then for the remaining time steps, chooses the prices via the optimism in the face of uncertainty (OFU) paradigm restricted to prices that are known to satisfy the constraints. As proven in the analysis, this algorithm achieves sublinear regret while ensuring that the prices are in the feasible price set $\bar{D}$ for all time steps with high probability.

In order to both implement OFU and determine which prices are safe, the proposed algorithm uses previous price response information $\{(\gamma^\tau, x^\tau)\}_{\tau=1}^{t-1}$ to construct confidence sets in which θ_i^* lie with high probability. Given the regularized least-squares estimator for θ_i^* at time step t with regularization paramater $\nu > 0$,

$$\hat{\theta}_i^t = [V_i^t]^{-1} \sum_{s=1}^{t} h_i(\gamma_i^s)\bar{x}_i^s, \tag{10}$$

where the gram matrix is

$$V_i^t = \nu I + \sum_{s=1}^{t} h_i(\gamma_i^s) h_i(\gamma_i^s)^\top, \tag{11}$$

we use a modified version of the confidence set developed in [30].

Theorem 1. *(Theorem 1 in [30] modified for multiple users) Let Assumptions 1 and 3 hold. Recall the definition of V_i^t in (11). Then for all i in $[n]$ and $t \geq 0$, we have with probability at least $1-\delta$ that θ_i^* lies in the set*

$$C_i^t = \left\{ \theta_i \in \mathbb{R}_+^m : \left\| \theta_i - \hat{\theta}_i^t \right\|_{V_i^t} \leq \sqrt{\beta^t}, \|\theta_i\| \leq S \right\} \tag{12}$$

where

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

The pure exploration phase of the algorithm is used to control the minimum eigenvalue of $V_i^{T'}$ and hence control the size of the confidence set $C^{T'}$. In order to shrink the confidence set in a controlled manner, the algorithm samples prices IID from the initial safe set D^0. Formally, we can state that as $\gamma^t \overset{\text{iid}}{\sim} \text{Unif}(D^0)$ for all t in $[1, T']$. As proven in Lemma 1, this exploration strategy ensures that the parameter λ_- (defined in (13)) is strictly greater than zero. This means that the confidence set $C^{T'}$ will shrink with T' and therefore guarantees that the algorithm will have sublinear regret.

Lemma 1. *Let Assumption 4 hold. Then, with Algorithm 1 we have that*

$$\lambda_- := \min_{i \in [n]} \left[\lambda_{min} \Big(\mathbb{E}\left[h_i(\gamma_i^t) h_i(\gamma_i^t)^\top \right] \Big) \right] > 0, \tag{13}$$

for all t in $[1, T']$.

Proof. The proof is given in Lemma 6 in Appendix A. □

For time steps after the pure exploration phase, the algorithm chooses actions optimistically within a conservative inner approximation of the feasible price set, which we call the safe price set. The safe price set is defined as

$$D^t = \left\{ \gamma \in \mathbb{R}^n : \sum_{i=1}^{n} a_{ji} x_i(\gamma_i; \theta_i) \leq c_j, \forall j \in [p], \forall \theta \in C^t \right\} \tag{14}$$

where $C^t = C_1^t \times C_2^t \times ... \times C_n^t$. Equation (14) implies that for any $\gamma \in D^t$ and any $\theta \in C^t$, it holds that $[x_1(\gamma_1, \theta_1), ..., x_n(\gamma_n, \theta_n)]^\top$ is in the feasible consumption set $\bar{E}$. Since θ^* is in C^t for all $t \in [T]$ with high probability (due to Theorem 1), any $\gamma \in D^t$ will yield a feasible consumption vector with the same probability. Therefore, the algorithm ensures that the price vectors at all time steps are feasible with high probability by choosing each price vector γ^t from the safe price set D^t. Among the price vectors in D^t, the algorithm chooses one that is optimistic, i.e. the algorithm finds a γ^t such that

$$(\gamma^t, \tilde{\theta}^t) \in \underset{(\gamma,\theta) \in D^{t-1} \times C^{t-1}}{\arg\max} \sum_{i=1}^{n} f_i\left(h_i(\gamma_i)^\top \theta_i \right). \tag{15}$$

For each time step after the pure exploration phase (when $t > T'$) the algorithm broadcasts the optimistic price found with (15), observes the noisy consumption $\bar{x}^t$ and then updates the confidence set C^t. In the next section we provide theoretical regret guarantees for the proposed algorithm.

C. Regret Analysis

In this section, we prove that, with high probability, the regret of the proposed algorithm is $\tilde{\mathcal{O}}(T^{2/3})$ as given by Theorem 2. This regret bound is comparable with similar safe learning algorithms as [23] and [28] give the same order bound.

Theorem 2. *Let Assumptions 1–4 hold. Then with probability at least $1-2\delta$, we have that the regret of Algorithm 1 satisfies*

$$\begin{aligned} R_T \leq & nM \max(LS, 1) \sqrt{8(T-T')\beta^T m \log\left(1 + \frac{TL^2}{m\nu}\right)} \\ & + 2MnLST' + \frac{4\sqrt{2}(T-T')\kappa M n^2 L^2 S \sqrt{\beta^T}}{\zeta \sqrt{2\nu + \lambda_- T'}} \end{aligned}$$

when $T' \geq t_\delta = \frac{8L^2}{\lambda_-} \log(\frac{nm}{\delta})$. In particular, choosing $T' = \max(n^{2/3}T^{2/3}, t_\delta)$ ensures that $R_T \in \tilde{\mathcal{O}}(n^{5/3}T^{2/3})$.

The complete proof of Theorem 2 is given in Appendix A of the full online version of this paper in [35]. This proof relies on a decomposition of the instantaneous regret that separates (I) the instantaneous regret due to the difference between the safe price set D^t and the true price set $\bar{D}$, and (II) the instantaneous regret due to the size of the confidence set for θ_i^*. Given the definition of instantaneous regret,

$$r_t = \sum_{i=1}^{n} \left[f_i\left(h_i(\gamma_i^*)^\top \theta_i^* \right) - f_i\left(h_i(\gamma_i^t)^\top \theta_i^* \right) \right], \tag{16}$$

we have the decomposition $r_t = r_t^I + r_t^{II}$, where

$$\begin{aligned} r_t^I &= \sum_{i=1}^{n} \left[f_i\Big(h_i(\gamma_i^*)^\top \theta_i^* \Big) - f_i\Big(h_i(\gamma_i^t)^\top \tilde{\theta}_i^t \Big) \right], \\ r_t^{II} &= \sum_{i=1}^{n} \left[f_i\Big(h_i(\gamma_i^t)^\top \tilde{\theta}_i^t \Big) - f_i\Big(h_i(\gamma_i^t)^\top \theta_i^* \Big) \right]. \end{aligned} \tag{17}$$

Establishing the bound on r_t^{II} uses similar techniques to the stochastic linear bandit analysis, such as [30]. Bounding r_t^I is somewhat more challenging and existing theory proves to be largely insufficient.

We bound r_t^I for all time steps greater than T' in Lemma 2.

Lemma 2. *Let Assumptions 1–4 hold. Then, the set of prices chosen by Algorithm 1 for time steps t greater than $T' \geq t_\delta$, $\{\gamma_i^t\}_{\forall i \in [n], t > T'}$, satisfies*

$$r_t^I = \sum_{i=1}^{n} \left[f_i\Big(h_i(\gamma_i^*)^\top \theta_i^*\Big) - f_i\Big(h_i(\gamma_i^t)^\top \tilde{\theta}_i^t\Big)\right] \leq \frac{4\sqrt{2}\kappa M n^2 L^2 S \sqrt{\beta^T}}{\zeta \sqrt{2\lambda + \lambda_- T'}},$$

with probability at least $1 - 2\delta$.

Proof sketch: The complete proof of Lemma 2 is given in Appendix A. This proof draws inspiration from [28] in that it considers a line segment between a point in the initial safe set and the optimal solution, and then tracks the growth of the safe set along this line segment by relating it to the shrinkage of the paramater confidence set (i.e. C^t in this case). Despite the influence from [28], our problem requires more complex work and new techniques to handle two primary challenges: (a) the fact that there are nonlinear basis functions (i.e. h_i), and (b) the fact that there multiple constraints that jointly apply to multiple users.

Due to challenge (a), we cannot take the natural approach of using a line segment in the price domain (i.e. a line segment from a point in D^0 to γ^*), because the constraint is nonlinear with respect to the price. To work around this issue, we consider a line segment in the domain of $h = [h_1(\gamma_1)^\top \ h_2(\gamma_2)^\top \ \ldots \ h_n(\gamma_n)^\top]^\top$. As a result, each of the constraints are linear with respect to any point on the line segment, making it feasible to bound. However, this introduces additional challenges because there needs to be careful consideration of which values the h vector can take given the range of each h_i function.

Challenge (b) makes it difficult both to determine which constraint this line segment crosses and to bound the growth of the safe set across multiple users. To address this challenge, we use the increasing property of f_i to show that at least one constraint is tight on the optimal solution and use this fact to relate the growth of the safe set to the minimum eigenvalue of the gram matrix.

In the next section, we extend the work from the price response problem to a setting where the utility function is a property of the user rather than being chosen by the central coordinator.

III. Safe Utility Maximization (SUM) Problem

In this section, we consider the setting where the user utility functions $f_i(\cdot)$ are not designed by the central coordinator, but are instead defined by the price response of the users under the assumption that the users are self-interested agents. In particular, the price response function $x_i(\gamma_i, \theta_i)$ now corresponds to the profit-maximizing consumption, with the *profit* due to consumption x_i taken to be the utility $f_i(x_i, \theta_i)$ minus the cost $\gamma_i x_i$. That is, we want f_i to be defined such that

$$x_i(\gamma_i, \theta_i) = \arg\max_{x_i} (f_i(x_i, \theta_i) - \gamma_i x_i). \tag{18}$$

This setting is especially useful because it captures the behavior of rational self-interested agents, which are prevalent in safety-critical infrastructure systems. For example, electricity customers will choose an electricity consumption that maximizes the benefit (or utility) that they get from the electricity minus the costs of the electricity (e.g. [1], [11] use such a model). We will specify the specific structure of utility functions that can concurrently satisfy (18) and (1).

Given that our problem is utility-maximizing and satisfies this profit-maximizing property, it can also be viewed as a safe version of the dual NUM problem (see [9]) with a specific structure for the utility function and noisy observations of the consumption. Our problem is considered to be *safe* because, unlike conventional dual NUM, it ensures that the resource constraints are satisfied at each time step. Therefore, our work may find further application in areas in which dual NUM algorithms have traditionally been used, as well as in safety-critical areas that may benefit from NUM-type algorithms.

The problem setup, proposed algorithm and regret analysis are given in Sections III-A, III-B and III-C, respectively.

A. Problem Setup

In this setting, the utility functions are not known to the central coordinator. Instead, the utility functions, denoted by $\underline{f}_i$, are defined in terms of the price response function $x_i(\cdot)$ given that the price response is the profit-maximizing consumption. That is, the utility function for user i, denoted $\underline{f}_i : \mathbb{R}_{++} \times \mathbb{R}^m \to \mathbb{R}$, is differentiable with respect to the first argument and is implicitly defined as

$$x_i(\gamma_i, \theta_i) = \arg\max_{x_i \in \mathbb{R}_{++}} \left(\underline{f}_i(x_i, \theta_i) - \gamma_i x_i\right), \tag{19}$$

where $x_i(\cdot)$ is the price response function in (1). It follows from its definition that $\underline{f}_i(x_i, \theta_i)$ represents the utility that a self-interested user gets from a consumption of x_i, given that her price response function is $x_i(\cdot, \theta_i)$.

To ensure that $\underline{f}_i$ is well defined and satisfies the same properties of the utility functions as in the SPR setting (e.g. increasing, Lipschitz), we make the following modifications to the price response functions. We first specify that h_i is differentiable and strictly decreasing. This is more restrictive than the SPR setting as h_i is specified as continuous and non-increasing in that case. These restrictions on h_i ensure that $\underline{f}_i(\cdot, \theta_i)$ is unique up to an additive constant for a given price response function $x_i(\cdot, \theta_i)$ as proven in Appendix B of the full online version [35]. We also make the following assumption on the price response, which is a stronger version of Assumption 3 from the SPR setting.

Assumption 5. *(Replaces Assumption 3) There exists positive constants S and ρ such that $\|\theta_i^*\| \leq S$ and $\mathbf{1}^\top \theta_i^* \geq \rho$. Also, the domain of h_i is the positive reals where $\lim_{\gamma_i \to 0^+} h_i(\gamma_i) = \boldsymbol{\infty}$ and $\lim_{\gamma_i \to \infty} h_i(\gamma_i) = \mathbf{0}$.*

The first part of Assumption 5 is rather mild as it only adds the condition that $\mathbf{1}^\top \theta_i^* \geq \rho$ to Assumption 3. However, the second part of Assumption 5 is somewhat stronger than the equivalent part of Assumption 3 because it ensures that there exists a *positive* price (versus a real-valued price in Assumption 3) that will compel the user to consume any non-negative quantity of the resource. Examples of basis functions

that satisfy these assumptions (where $m = 1$, i.e. θ_i is a scalar) are $h_i(\gamma_i) = 1/\gamma_i$ which corresponds to $x_i(\gamma_i, \theta_i) = \theta_i/\gamma_i$ and $\underline{f}_i(x_i, \theta_i) = \theta_i \log(x_i)$, as well as $h_i(\gamma_i) = 1/\sqrt{\gamma_i}$ which corresponds to $x_i(\gamma_i, \theta_i) = \theta_i/\sqrt{\gamma_i}$ and $\underline{f}_i(x_i, \theta_i) = -\theta_i^2/x_i$. [1]

Now that the utility function for this setting has been specified, we define the optimal price vector for this setting

$$\underline{\gamma}^* \in \arg\max_{\gamma \in \bar{D}} \sum_{i=1}^{n} \underline{f}_i(x_i(\gamma_i, \theta_i^*), \theta_i^*), \tag{20}$$

where $\bar{D}$ is the feasible price set defined in (5). Using the definition of optimal prices, we can then define the regret due to the prices $\{\underline{\gamma}_i^t\}_{i\in[n],t\in[T]}$ as

$$\underline{R}_T = \sum_{t=1}^{T} \sum_{i=1}^{n} \left[\underline{f}_i(x_i(\underline{\gamma}_i^*, \theta_i^*), \theta_i^*) - \underline{f}_i(x_i(\underline{\gamma}_i^t, \theta_i^*), \theta_i^*) \right]. \tag{21}$$

Note that the only difference between the definition of both the optimal price and regret in this setting versus the SPR setting is that $\underline{f}_i$ is used in place of the SPR utility function f_i.

Given that Assumption 5 provides an additional restriction on the parameter θ_i^* (i.e. the condition that $\mathbf{1}^\top \theta_i^* \geq \rho$), we need to define the initial confidence set for this setting (equivalent to (7)),

$$\underline{C}_i^0 = \{\theta_i \in \mathbb{R}_+^m : \|\theta_i\| \leq S, \mathbf{1}^\top \theta_i \geq \rho\} \tag{22}$$

with $\underline{C}^0 = \underline{C}_1^0 \times \underline{C}_2^0 \times ... \times \underline{C}_n^0$. As in the SPR setting, we use $\underline{C}^0$ to define a set of prices that is contained in $\bar{D}$ and a set of prices that contains $\bar{D}$ which are $\underline{D}^0 = \{\gamma \in \mathbb{R}^n : \sum_{i=1}^n a_{ji}\theta_i^T h_i(\gamma_i) \leq c_j - \zeta, \forall j \in [p], \forall \theta \in \underline{C}^0\}$ and $\hat{\underline{D}}^0 := \left\{\gamma \in \mathbb{R}^n : \exists \theta \in \underline{C}^0 \text{ s.t. } \sum_{i=1}^n a_{ji}\theta_i^T h_i(\gamma_i) \leq c_j, \forall j \in [p]\right\}$, respectively. In the following assumption, we assume that $\underline{D}^0$ is nonempty, $h_i(\gamma_i)$ has bounded norm for all $\gamma \in \hat{\underline{D}}^0$ and that the elements of h_i are linearly independent (equivalent to Assumption 4).

Assumption 6. *(Replaces Assumption 4) There exists positive constants ζ and κ such that $\underline{D}^0$ is nonempty and $|a_{ji}| \leq \kappa$ for all i in $[n]$ and j in $[p]$. Additionally, there exists a positive constant L such that $\max_{i\in[n]} \|h_i(\gamma_i)\| \leq L$ for all $\gamma \in \hat{\underline{D}}^0$. Also, there does not exist a nonzero $\alpha_i \in \mathbb{R}^m$ for each $i \in [n]$ such that $\alpha_i^\top h_i(\gamma_i) = 0$ for all $[\gamma_1\ \gamma_2\ ...\ \gamma_n]^\top$ in $\underline{D}^0$.*

Note that Assumption 6 is the same as Assumption 4 from the SPR setting, except it incorporates the additional prior information that $\mathbf{1}^\top \theta_i \geq \rho$ for all i in $[n]$.

We then use $\underline{C}_i^0$ to state an assumption which ensures that $\underline{f}_i$ is Lipschitz. This assumption uses the inverse of the price response function, which we denote $g_i(x_i, \theta_i)$ such that $x_i(g_i(x_i, \theta_i), \theta_i) = x_i$ for any x_i in $\mathbb{R}_{++}$ and $g_i(x_i(\gamma_i, \theta_i), \theta_i) = \gamma_i$ for any γ_i in $\mathbb{R}_{++}$. In Lemma 11 in Appendix B1 of the full online version of this paper [35], $g_i(x_i, \theta_i)$ is proven to exist and to be equal to $\frac{\partial}{\partial x_i}\underline{f}_i(x_i, \theta_i)$ for all x_i in $\mathbb{R}_{++}$.

Assumption 7. *For all x in $\bar{E}$ and θ in $\underline{C}^0$, there exists positive constants Γ, L and K such that $g_i(x_i, \theta_i) \leq \Gamma$, $\|h_i(g_i(x_i, \theta_i))\| \leq L$, and $h_i'(g_i(x_i, \theta_i)) \preceq -\mathbf{1}K$ for all $i \in [n]$. Additionally, there exists a point x^0 in E such that $\underline{f}_i(x_i^0, \cdot)$ is η-Lipschitz on $\underline{C}_i^0$ for all i in $[n]$.*

[1] We give examples with $m = 1$ because in more complicated settings, there may not be a closed-form expression for the utility function. See (51) in the full online version [35] for an integral expression for the utility function.

Algorithm 2 Safe Utility Maximization Algorithm

Input: $\{h_i\}_{i\in[n]}$, $\{a_{ji}\}_{i\in[n],j\in[p]}$, $\{c_j\}_{j\in[p]}$, $\{\underline{f}_i\}_{i\in[n]}$, S, L, ρ, x^0, f^0

for $t = 1$ to T' **do**
  Broadcast $\underline{\gamma}^t \sim \text{Unif}(\underline{D}^0)$ to the users.
  Observe noisy consumption $\bar{x}^t$.
end for
Construct confidence set $\underline{C}^{T'}$ with (23).
Construct safe price set $\underline{D}^{T'}$ with (24).
for $t = T' + 1$ to T **do**
  Choose some $\check{\underline{\theta}}^t$ in $\underline{C}^{t-1}$.
  Find optimistic price $\underline{\gamma}^t$ with (25).
  Broadcast $\underline{\gamma}^t$ to the users.
  Observe noisy consumption $\bar{x}^t$.
  Update confidence set $\underline{C}^t$ with (23).
  Update safe price set $\underline{D}^t$ with (24).
end for

The first part of Assumption 7 provides bounds on $g_i(x_i, \theta_i)$, $h_i(g_i(x_i, \theta_i))$ and $h_i'(g_i(x_i, \theta_i))$ for values of x_i and θ_i that the central coordinator might use as arguments for the utility function when estimating the optimal utility (i.e. the central coordinator initially knows that θ^* is in $\underline{C}^0$ and that the optimal x is in $\bar{E}$). The second part of Assumption 7 is mild as it only states that $\underline{f}_i(x_i, \cdot)$ is Lipschitz for a single value of $x_i = x_i^0$. Along with the other assumptions, this is sufficient to ensure that $\underline{f}_i(x_i, \cdot)$ is Lipschitz for all $x \in \bar{E}$. In order to bound x^0 in the analysis, we use the fact that $\bar{E}$ is bounded by definition (in Section II-A) to define the positive constant ξ as satisfying $|x^1 - x^2| \preceq \mathbf{1}\xi$ for all x^1 and x^2 in $\bar{E}$. In the next section, we propose an algorithm to address this problem setup.

B. Proposed Algorithm

The proposed algorithm for this setting (Algorithm 2) operates nearly the same as the SPR algorithm with the exception being that the confidence set and the optimistic price are defined differently. Incorporating Assumption 5 into (12), the confidence set for θ_i^* in this setting is

$$\underline{C}_i^t = \Bigg\{\theta_i \in \mathbb{R}_+^m : \left\|\theta_i - \hat{\theta}_i^t\right\|_{V_i^t} \leq \sqrt{\beta^t}, \|\theta_i\| \leq S, \mathbf{1}^\top \theta_i \geq \rho\Bigg\}. \tag{23}$$

The safe price set is then

$$\underline{D}^t = \left\{\gamma \in \mathbb{R}^n : \sum_{i=1}^{n} a_{ji} x_i(\gamma_i; \theta_i) \leq c_j, \forall j \in [p], \forall \theta \in \underline{C}^t\right\} \tag{24}$$

where $\underline{C}^t = \underline{C}_1^t \times \underline{C}_2^t \times ... \times \underline{C}_n^t$. Using these definitions, the optimistic price is found by first choosing some $\check{\underline{\theta}}^t$ in $\underline{C}^{t-1}$ and then solving

$$(\underline{\gamma}^t, \tilde{\underline{\theta}}^t) \in \underset{(\gamma,\theta)\in \underline{D}^{t-1}\times \underline{C}^{t-1}}{\arg\max} \sum_{i=1}^{n} \underline{f}_i\left(h_i(\gamma_i)^\top \theta_i, \check{\underline{\theta}}_i^t\right) \tag{25}$$

Note that this price update is nearly the same as the price update in the SPR setting, with the exception being that $\underline{f}_i(\cdot, \check{\underline{\theta}}_i^t)$ is used instead of $f_i(\cdot)$ where $\check{\underline{\theta}}^t$ is chosen arbitrarily in $\underline{C}^{t-1}$. Although the optimistic price update (25) maximizes the approximate utility $\sum_{i=1}^n \underline{f}_i(\cdot, \check{\underline{\theta}}_i^t)$ rather than the true utility $\sum_{i=1}^n \underline{f}_i(\cdot, \theta_i^*)$ (since θ^* is unknown), the algorithm still enjoys sublinear regret because the difference between $\underline{f}_i(x_i, \check{\underline{\theta}}_i^t)$ and $\underline{f}_i(x_i, \theta_i^*)$ shrinks with time horizon. This is due to the pure exploration phase, which ensures that C^{t-1} shrinks with the time horizon and therefore that the distance between $\check{\underline{\theta}}^t$ and θ^* shrinks as well. The complete regret bound for this algorithm is given in the next section.

C. Regret Analysis

In this section, we extend the SPR regret analysis in Section II-C to this setting. The main result is Theorem 3, which gives the regret bound for this setting.

Theorem 3. *Let Assumptions 1, 5–7 hold. Then, with probability at least* $1 - 2\delta$*, we have that*

$$\begin{aligned}\underline{R}_T \leq & n\Gamma \max(LS, 1)\sqrt{8(T-T')\beta^T m \log\left(1 + \frac{TL^2}{m\nu}\right)} \\ & + 2\Gamma nLST' + \frac{4\sqrt{2}(T-T')\kappa\Gamma n^2 L^2 S\sqrt{\beta^T}}{\zeta\sqrt{2\nu + \lambda_- T'}} \\ & + \frac{4\sqrt{2}n(T-T')(\eta + \frac{\xi L}{\rho K})\sqrt{\beta^T}}{\sqrt{2\nu + \lambda_- T'}}.\end{aligned}$$

when $T' \geq t_\delta = \frac{8L^2}{\lambda_-}\log(\frac{nm}{\delta})$*. In particular, choosing* $T' = \max(n^{2/3}T^{2/3}, t_\delta)$ *guarantees that* $\underline{R}_T \in \tilde{\mathcal{O}}(T^{2/3}n^{5/3})$.

The proof of Theorem 3 is in Appendix B of the full online version of this paper in [35]. We can see that the first three terms of the regret bound in Theorem 3 match the bound for the SPR setting (Theorem 2) except that Γ appears in this bound where M appears in the SPR bound. The fourth term in the regret bound comes from the error in the second argument of the utility function, i.e. the difference between $\underline{f}_i(x_i, \check{\underline{\theta}}_i^t)$ and $\underline{f}_i(x_i, \theta_i^*)$ for some x_i. However, the bound in Theorem 3 is still the same order as Theorem 2 from the SPR setting.

IV. Application to Demand Response in Smart Grid

In this section, we apply the SPR and SUM algorithms to demand response (DR) in smart grid. DR is a mechanism by which an aggregator (or other organization that supplies power) can modify the electricity usage of its customers, sometimes through variable pricing. This is advantageous because it can reduce the costs for the aggregator and its customers, and improve reliability [36]. One popular type of DR program is day-ahead real-time pricing (RTP), where each day the aggregator posts prices for each time interval in the next day. In choosing these prices, the aggregator aims to ensure that the utility provided by the electricity consumption is high for the users (i.e., they are satisfied), while maintaining low costs for providing that electricity. It is also paramount that the prices be chosen such that the customers' consumption does not violate the physical limits of the grid to avoid service outages and repair costs. Our approach to day-ahead RTP, based on the algorithms developed in this paper, achieves high utility and ensures that cost constraints and grid constraints are satisfied without knowing the specific flexibility or responsiveness of the customer's load beforehand. This is illustrated in Fig. 1. For the remainder of this section, we formulate a day-ahead RTP problem that is utility maximizing and safe with regard to system constraints, and then show, through simulation, that the SPR and SUM algorithms are effective for this problem.

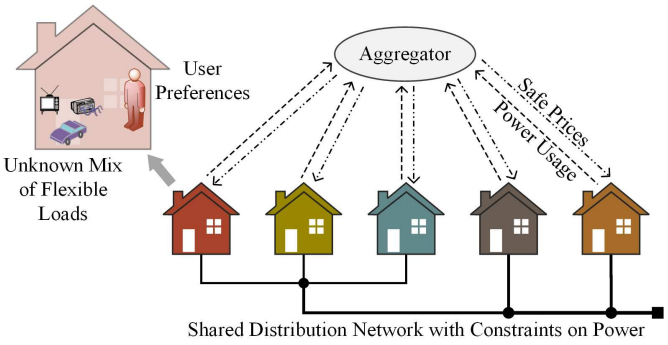

Fig. 1. The aggregator does not specifically know how each electricity customer will respond to prices beforehand as each customer has an unknown mix of flexible loads and unique user preferences. Despite having limited knowledge as such, the aggregator needs to choose prices such that the utility is high and the distribution network constraints are satisfied.

A. Demand Response Formulation

Each day (time step) t, the aggregator (central coordinator) posts prices $\gamma_{i,v}^t$ for each time period v in $[V]$ and customer (user) i in $[n]$. Customer i then responds with a noisy power consumption for each period in the day[2], denoted $\bar{x}_i(\gamma_i^t; \theta_i) = [\bar{x}_{i,1}(\gamma_i^t; \theta_i) \ ... \ \bar{x}_{i,V}(\gamma_i^t; \theta_i)]^\top$ where $\bar{x}_{i,v}(\gamma_i^t; \theta_i) = x_{i,v}(\gamma_i^t; \theta_i) + \mu_{i,v}^t$ and $\gamma_i^t = [\gamma_{i,1}^t \ ... \ \gamma_{i,V}^t]$. As before, we use a parametrically linear model for the consumption $x_{i,v}(\gamma_i^t; \theta_i) = h_{i,v}(\gamma_i^t)^\top \theta_i$ and take the noise $\mu_{i,v}^t$ to be conditionally subgaussian. Note that the consumption at each period v is allowed to depend on the price at all the periods in the day to account for inter-temporal flexibility. The central coordinator observes the noisy consumption $\bar{x}_i(\gamma_i^t, \theta_i)$ for all customers on each day, and uses this to inform the choice of prices on future days.

When choosing prices, the aggregator aims to maximize utility while satisfying grid constraints and cost constraints. We denote the utility function for customer i as the increasing function $U_i : \mathbb{R}^V \to \mathbb{R}$. The grid constraints are on the

[2]Although the consumption $\bar{x}_i$ and price γ_i^t are vectors in the demand response formulation, the results from the SPR and SUM formulation, where they are scalars, can be easily extended to this case. In stating our theoretical results, this vector case is not adopted for brevity of notation.

nodal voltages u_v^t and the distribution line power flows p_v^t: $u_{min} \leq u_v^t \leq u_{max}$ and $p_v^t \leq S_{max}$ for all $v \in [V]$ and $t \in [T]$. The nodal voltages and power flows are related to the consumption via the power flow model. In particular, we use the *LinDistFlow* model [37] for a feeder network to express the reliability constraints linearly with respect to the consumption, of the form $\sum_{i=1}^n a_{ji} x_i(\gamma_i^t) \preceq \mathbf{1} c_j$ for all j in $[p]$, t in $[T]$. Cost constraints can be implemented by specifying a limit on the total power supplied to the users at each period in the day according to the supply price and cost limit at that period. With the objective and constraints defined, we have that the optimal prices satisfy

$$\begin{aligned} \gamma^* \in \underset{\gamma \in \mathbb{R}^{V \times n}}{\arg\max} & \sum_{i=1}^n U_i(x_i(\gamma_i)) \\ \text{s.t.} & \sum_{i=1}^n a_{ji} x_i(\gamma_i) \preceq \mathbf{1} c_j, \ \forall j \in [p] \end{aligned} \tag{26}$$

In the next section, we discuss the price response model that is used to define the price response function.

B. Price Response Model

In order to capture the consumption behavior of a customer in response to electricity prices (i.e. specify h_i in (1)), we use the price response model developed in [38], which itself uses the appliance model in [39]. This appliance model considers clusters of appliances which each have a set of feasible consumption profiles (a consumption profile specifies the consumption from those appliances for each period in the day). For example, one cluster might represent electric vehicles (EV) which need to be fully charged within a specific time frame subject to power limits. Depending on how tight the time frame is, there might be several different consumption profiles for the electric vehicles that would satisfy these charging requirements. Refer to [39] for further discussion on modeling other appliance types.

Given that there are several possible consumption profiles for each appliance cluster, the price response model in [38] considers two mechanisms by which price impacts a customer's power consumption: (1) the cost-minimizing appliance scheduling by the home energy management system (HEMS) and (2) the adjustment of the customer's preferences in response to electricity prices. Mechanism (1) specifies that the HEMS will choose the consumption profile for each appliance cluster that minimizes the cost of operating that appliance while mechanism (2) specifies that the customer's usage of each appliance cluster varies according to price equally for all periods in the day. We assume that the way in which the HEMS schedules appliances and the customer's preference adjustment function are known, while the number of appliances that each customer has in each appliance cluster (specified by each element of θ_i) is unknown. Note that our approach and algorithms could accommodate a more general model, but we use this one to provide an example of how the approach and algorithm can be used.

C. Test Setup

To evaluate the performance of our algorithms in the demand response problem through simulation, we use a real radial distribution network with $n = 37$ customers as specified in [38] (originally from [40]). For this distribution system, we use the power limits specified for each line given in [38] and the nodal voltage limits of 0.95 and 1.05 p.u. (with 12.5kV base) as given in [40]. We use $T = 365$ days and $V = 3$ periods, with $m = 2$ different appliance clusters. One appliance cluster is for appliances that operate at the same time regardless of price and includes lighting (200 W, on for intervals $\{2, 3\}$) and cooking (500 W, on for interval 3). The other appliance cluster is for flexible appliances that can be scheduled at several different times and includes EV charging (500 W, on for 1 interval in $\{1, 3\}$), washer/drier (300 W, on for 1 interval in $\{2, 3\}$), HVAC (600 W, on for 1 interval in $\{1, 2, 3\}$) and entertainment (200 W, on for 1 interval in $\{2, 3\}$). We use shifted sigmoids ($1/(1 + e^{\gamma_i^\top \mathbf{1} - 5})$) for the preference adjustment functions of the clusters. Also, we use $U_i(x_i) = b_i \log(x_i + 1)$ where $b_i \sim \mathcal{U}[0, 1]$ (for the SPR Algorithm), choose the unknown parameter $[\theta_i]_k \sim \mathcal{U}[0.5, 1]$ for each k in $[m]$, take the variance proxy as $\sigma = 1.5$ for the SPR experiments and $\sigma = 3$ for the SUM experiments, where $\mu_{i,v}^t \sim \mathcal{N}(\sigma)$. For the algorithm parameters, we use $\nu = 10$ for the regularization parameter, $T' = 52 \approx T^{2/3}$ for the duration of the safe exploration phase and $\delta = 0.01$ for the confidence parameter. To make the algorithms tractable, we consider a finite set of prices in that the set of possible prices for each user at each time step is $\{2, 5\}$. We also divide the users in to three groups and use the same price for each of these groups to reduce the number of prices that need to be considered. These groups are (1) users 11-25, (2) users 29-35, and (3) users 1-10, 26-28. To reduce the computational load of the SUM experiments, the users in each group in the SUM experiments are assumed to have the same price response function. We also calculate the SUM utility with a smoothed version of the price response as detailed in Appendix D in the full online version in [35]. Also, note that we do not incorporate the assumed bounds on θ_i (i.e. $\|\theta_i\| \leq S$, $\mathbf{1}^\top \theta_i \geq \rho$) when calculating the confidence sets for θ_i.

D. Simulation Results

The SPR Algorithm (Algorithm 1) was implemented for the specified demand response problem and simulated for five trials, each with different realizations of the noise. For all five trials, there were zero constraint violations. The rolling sum of the instantaneous regret for each trial is shown in the left-hand side of Fig. 2.

Similarly, the SUM Algorithm (Algorithm 2) was implemented for the demand response problem. In five trials, there were zero constraint violations. The rolling sum of the instantaneous regret for each trial is shown in the right-hand side of Fig. 2.

Fig. 2 provides experimental evidence that the regret of both algorithms grows sub-linearly with respect to time. This provides validation of the stated theoretical results, and

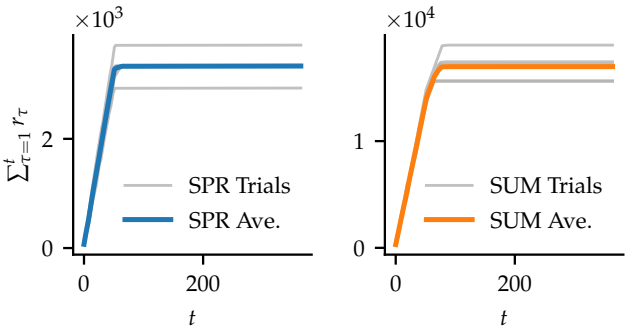

Fig. 2. The rolling sum of the instantaneous regret of the Safe Price Response (SPR) and Safe Utility Maximization (SUM) algorithms for five trials of the demand response problem.

demonstrates the algorithms' performance in a realistic demand response setting.

V. Conclusion

In this paper, we present two novel safe optimization problems with applications to pricing design for safety-critical infrastructure systems. We propose algorithms for each of these problems and prove in our analysis that they both enjoy sublinear regret. We then demonstrate the real-world applicability of these algorithms by simulating them being used for DR pricing in a distribution network. These simulations also provide numerical validation for the theoretical results.

Despite the efficacy of our approach, there are some limitations. Firstly, our problem formulation requires that a set of safe prices is initially known by the algorithm. This is a fundamental limitation of any problem formulation with uncertain constraints that need to be satisfied at every time step because, in such a setting, the algorithm cannot ensure safety in the initial rounds without prior information. That said, it may be challenging in some real-world settings to accurately determine a set of safe prices without some (potentially unsafe) experimentation.

Another limitation of our approach is that it only considers linear constraints. As a result, this work cannot be directly applied to safe pricing settings with nonlinear constraints. An important such example is power flow constraints in the smart grid, which can be more accurately specified by nonlinear constraints. We leave the problem of safe pricing under nonlinear constraints as future work.

References

[1] 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.

[2] 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.

[3] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[4] 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.

[5] 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.

[6] E. Bitar and S. Low, "Deadline differentiated pricing of deferrable electric power service," in *2012 IEEE 51st IEEE conference on decision and control (CDC)*. IEEE, 2012, pp. 4991–4997.

[7] S. Han, U. Topcu, and G. J. Pappas, "An approximately truthful mechanism for electric vehicle charging via joint differential privacy," in *2015 American Control Conference (ACC)*. IEEE, 2015, pp. 2469–2475.

[8] B. Sun, X. Tan, and D. H. Tsang, "Eliciting multi-dimensional flexibilities from electric vehicles: A mechanism design approach," *IEEE Transactions on Power Systems*, vol. 34, no. 5, pp. 4038–4047, 2018.

[9] 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.

[10] 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.

[11] 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.

[12] B. Turan and M. Alizadeh, "Safe dual gradient method for network utility maximization problems," *arXiv preprint arXiv:2208.04446*, 2022.

[13] A. Nedić and A. Olshevsky, "Distributed optimization over time-varying directed graphs," *IEEE Transactions on Automatic Control*, vol. 60, no. 3, pp. 601–615, 2014.

[14] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, "Safe learning in robotics: From learning-based control to safe reinforcement learning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 411–444, 2022.

[15] A. Gahlawat, P. Zhao, A. Patterson, N. Hovakimyan, and E. Theodorou, "L1-gp: L1 adaptive control with bayesian learning," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 826–837.

[16] G. Chowdhary, H. A. Kingravi, J. P. How, and P. A. Vela, "Bayesian nonparametric adaptive control using gaussian processes," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 3, pp. 537–550, 2014.

[17] F. Berkenkamp and A. P. Schoellig, "Safe and robust learning control with gaussian processes," in *2015 European Control Conference (ECC)*. IEEE, 2015, pp. 2496–2501.

[18] T. Koller, F. Berkenkamp, M. Turchetta, and A. Krause, "Learning-based model predictive control for safe exploration," in *2018 IEEE conference on decision and control (CDC)*. IEEE, 2018, pp. 6059–6066.

[19] K. P. Wabersich and M. N. Zeilinger, "Linear model predictive safety certification for learning-based control," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 7130–7135.

[20] 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.

[21] ——, "Safe non-smooth black-box optimization with application to policy search," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 980–989.

[22] T. Chen and G. B. Giannakis, "Bandit convex optimization for scalable and dynamic iot management," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 1276–1286, 2018.

[23] S. Chaudhary and D. Kalathil, "Safe online convex optimization with unknown linear safety constraints," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 6, 2022, pp. 6175–6182.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] S. Hutchinson, B. Turan, and M. Alizadeh, "A safe pricing mechanism for distributed resource allocation with bandit feedback," in *2022 61st IEEE Conference on Decision and Control (CDC)*. IEEE, 2022.

[30] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," *Advances in neural information processing systems*, vol. 24, 2011.

[31] 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.

[32] 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.

[33] T. Lan, D. Kao, M. Chiang, and A. Sabharwal, *An axiomatic theory of fairness in network resource allocation*. IEEE, 2010.

[34] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," 2008.

[35] S. Hutchinson, B. Turan, and M. Alizadeh, "Safe pricing mechanisms for distributed resource allocation with bandit feedback," 2023. [Online]. Available: https://arxiv.org/abs/2307.15785

[36] Q. Qdr, "Benefits of demand response in electricity markets and recommendations for achieving them," *US Dept. Energy, Washington, DC, USA, Tech. Rep*, vol. 2006, 2006.

[37] M. E. Baran and F. F. Wu, "Optimal capacitor placement on radial distribution systems," *IEEE Transactions on power Delivery*, vol. 4, no. 1, pp. 725–734, 1989.

[38] N. Tucker, A. Moradipari, and M. Alizadeh, "Constrained thompson sampling for real-time electricity pricing with grid reliability constraints," *IEEE Transactions on Smart Grid*, vol. 11, no. 6, pp. 4971–4983, 2020.

[39] M. Alizadeh, A. Scaglione, A. Applebaum, G. Kesidis, and K. Levitt, "Reduced-order load models for large populations of flexible appliances," *IEEE Transactions on Power Systems*, vol. 30, no. 4, pp. 1758–1774, 2014.

[40] P. Andrianesis, M. Caramanis, R. D. Masiello, R. D. Tabors, and S. Bahramirad, "Locational marginal value of distributed energy resources as non-wires alternatives," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 270–280, 2019.

[41] T. Lattimore and C. Szepesvári, *Bandit algorithms*. Cambridge University Press, 2020.

APPENDIX

A. Proof of Lemma 2.[3]

Note that, in this section, many of the statements only hold with high probability (i.e. they rely on Theorem 1 and Lemma 7). For brevity, this is not referenced at each step but the probability of the complete bound is discussed at the end.

First, we define some sets that allow us to more easily bound the growth of the safe decision set. We first have an expanded confidence set for θ_i^* that is centered at θ_i^* instead of $\hat{\theta}_i$:

$$\tilde{C}_i^t = \{\theta_i \in \mathbb{R}_+^m : \|\theta_i - \theta_i^*\|_{V_i^t} \leq 2\sqrt{\beta^t}, \|\theta_i\| \leq S\}. \quad (27)$$

Note that $C_i^t \subseteq \tilde{C}_i^t$ as we can use the triangle inequality with any $\theta_i \in C_i^t$ to get $\|\theta_i - \theta_i^*\|_{V_i^t} = \|\theta_i - \hat{\theta}_i + \hat{\theta}_i - \theta_i^*\|_{V_i^t} \leq \|\theta_i - \hat{\theta}_i\|_{V_i^t} + \|\hat{\theta}_i - \theta_i^*\|_{V_i^t} \leq 2\sqrt{\beta^t}$. We then use this expanded confidence set to define a shrunk safe price set,

$$\tilde{D}^t = \{\gamma \in \mathbb{R}^n : \sum_{i=1}^n a_{ji} x_i(\gamma_i; \theta_i) \leq c_j, \forall \theta \in \tilde{C}^t, \forall j \in [p]\}, \quad (28)$$

where $\tilde{C}^t = \tilde{C}_1^t \times \tilde{C}_2^t \times ... \times \tilde{C}_n^t$. The remaining analysis deals with the h vector, which is defined as $h = [h_1(\gamma_1)^\top \ h_2(\gamma_2)^\top \ ... \ h_n(\gamma_n)^\top]^\top$. Accordingly, we define a safe set for the h vector:

$$\tilde{G}^t = \{[h_1(\gamma_1)^\top \ ... \ h_n(\gamma_n)^\top]^\top : [\gamma_1 \ ... \ \gamma_n]^\top \in \tilde{D}^t\}. \quad (29)$$

[3]The missing proofs in this section are given in the full online version [35].

We also have the initial safe set for the h vector:

$$G^0 = \{[h_1(\gamma_1)^\top \ ... \ h_n(\gamma_n)^\top]^\top : [\gamma_1 \ ... \ \gamma_n]^\top \in D^0\} \quad (30)$$

Note that, by definition, $\tilde{C}^t \subseteq C^0$ for all $t \geq 1$ which implies that $D^0 \subseteq \tilde{D}^t$ for all $t \geq 1$ and that $G^0 \subseteq \tilde{G}^t$ for all $t \geq 1$.

Next, we consider a line from a point in G^0 to the optimal h vector. Let $h^* = [h_1(\gamma_1^*)^\top \ ... \ h_n(\gamma_n^*)^\top]^\top$ and h^0 be any element in G^0. Then, we can use α^t to track the safe set along a line as $\alpha^t = \max\{\alpha \in [0,1] : \alpha h^* + (1-\alpha)h^0 \in \tilde{G}^t\}$. Let $z^t = \alpha^t h^* + (1-\alpha^t)h^0$ and $z_i^t = \alpha^t h_i^* + (1-\alpha^t)h_i^0$. We can then bound r_{t+1}^I with α^t by using the fact that $\sum_{i=1}^n f_i(h_i(\gamma_i^{t+1})^\top \tilde{\theta}_i^{t+1}) \geq \sum_{i=1}^n f_i([z_i^t]^\top \theta_i^*)$ (which follows from (15)).

$$\begin{aligned} r_{t+1}^I &= \sum_{i=1}^n [f_i(h_i(\gamma_i^*)^\top \theta_i^*) - f_i(h_i(\gamma_i^{t+1})^\top \tilde{\theta}_i^{t+1})] \\ &\leq \sum_{i=1}^n [f_i(h_i(\gamma_i^*)^\top \theta_i^*) - f_i([z_i^t]^\top \theta_i^*)] \\ &\leq M \sum_{i=1}^n |h_i(\gamma_i^*)^\top \theta_i^* - [z_i^t]^\top \theta_i^*| \\ &= M \sum_{i=1}^n |(h_i^* - h_i^0)^\top \theta_i^*|(1-\alpha^t) \\ &\leq M \sum_{i=1}^n \|h_i^* - h_i^0\| \|\theta_i^*\|(1-\alpha^t) \\ &\leq 2MnLS(1-\alpha^t) \end{aligned} \quad (31)$$

Then, we establish a lower bound on α_t by first showing that at least one constraint is tight on the optimal solution. We first show that the problem can be expressed as optimization over the consumption in Lemma 3.

Lemma 3. *Let Assumption 3 hold. Then the optimal consumption* $x^* = [h_1(\gamma_1^*)^\top \theta_1^* \ ... \ h_n(\gamma_n^*)^\top \theta_n^*]^\top$ *satisfies* $x^* \in \arg\max_{x \in E} \sum_{i=1}^n f_i(x_i)$, *where* $E = \left\{x \in \mathbb{R}_{++}^n : \sum_{i=1}^n a_{ji} x_i \leq c_j, \ \forall j \in [p]\right\}$.

Proof. Consider the set of feasible consumption vectors, $\tilde{E} = \left\{\left[h_1(\gamma_1)^\top \theta_1^* \ ... \ h_n(\gamma_n)^\top \theta_n^*\right]^\top : \gamma \in \bar{D}\right\}$. We can see that $\arg\max_{x \in \tilde{E}} \sum_{i=1}^n f_i(x_i)$ contains $[h_1(\gamma_1^*)^\top \theta_1^* \ ... \ h_n(\gamma_n^*)^\top \theta_n^*]^\top$, where γ^* is defined in (4). Therefore, it only remains to be shown that $\tilde{E}$ is equal to E. To do so, we consider the range of values that $h_i(\gamma_i)^\top \theta_i^*$ can take for $\gamma \in \bar{D}$ and hence the values that are in $\tilde{E}$. Since $h_i(y)$ is continuous, $h_i(y)^T \theta_i^*$ is continuous as a function of y. Also, Assumption 3 and the fact that θ_i^* is nonzero imply that $\lim_{y \to -\infty} h_i(y)^T \theta_i^* = \infty$ and that $\lim_{y \to \infty} h_i(y)^T \theta_i^* = 0$. Therefore, by the Intermediate Value Theorem, for every $x_i \in \mathbb{R}_+$, there exists a $y \in \mathbb{R}$ such that $x_i = h_i(y)^T \theta_i^*$. Thus, $\tilde{E}$ contains all $x \succ 0$ such that $\sum_{i=1}^n a_{ji} x_i \leq c_j$ for all j in $[p]$ and therefore equals E. □

Using this result, we then show that at least one constraint is tight on the optimal solution in Lemma 4.

Lemma 4. *Let Assumption 3 hold. Then, there exists a constraint* j *in* $[p]$ *such that* $\sum_{i=1}^n a_{ji} x_i^* = c_j$, *where* x^* *is an optimal consumption as defined in Lemma 3.*

Lemma 5. *Let Assumption 3 hold. Then with probability at least $1-\delta$, there exists j in $[p]$ such that*

$$\max_{\theta\in\tilde{C}^t}\sum_{i=1}^{n} a_{ji}\theta_i^\top z_i^t = c_j. \tag{32}$$

To lower bound α^t, we use Lemma 5 which gives that there is a j in $[p]$ such that $\max_{\theta\in\tilde{C}^t}\sum_{i=1}^n a_{ji}\theta_i^T z_i^t = c_j$. Therefore,

$$\begin{aligned} c_j &= \max_{\theta\in\tilde{C}^t}\sum_{i=1}^{n} a_{ji}[z_i^t]^\top\theta_i \qquad (33)\\ &\le \sum_{i=1}^{n}\max_{\theta_i\in\tilde{C}_i^t} a_{ji}\alpha^t[h_i^*]^\top\theta_i + \sum_{i=1}^{n}\max_{\theta_i\in\tilde{C}_i^t} a_{ji}(1-\alpha^t)[h_i^0]^\top\theta_i\\ &= \alpha^t\underbrace{\sum_{i=1}^{n}\max_{\theta_i\in\tilde{C}_i^t} a_{ji}[h_i^*]^\top\theta_i}_{b^*} + (1-\alpha^t)\underbrace{\sum_{i=1}^{n}\max_{\theta_i\in\tilde{C}_i^t} a_{ji}[h_i^0]^\top\theta_i}_{b^0} \end{aligned}$$

Next, we bound b^*. In order to do so, we need the following lemmas, together which lower bound the minimum eigenvalue of the gram matrix after the pure exploration phase.

Lemma 6. *(Duplicate of Lemma 1) Let Assumption 4 hold. Then, with Algorithm 1 we have that*

$$\lambda_- := \min_{i\in[n]}\left[\lambda_{min}\Big(\mathbb{E}\left[h_i(\gamma_i^t)h_i(\gamma_i^t)^\top\right]\Big)\right] > 0, \tag{34}$$

for all t in $[1, T']$.

Lemma 7. *(Lemma 1 in [28] modified for multiple users) It holds with probability at least $1-\delta$ that $\lambda_{min}(V_i^{T'}) \ge \nu + \frac{\lambda_- T'}{2}$, for $T' \ge t_\delta := \frac{8L^2}{\lambda_-}\log(\frac{nm}{\delta})$ and all i in $[n]$.*

We can then establish a bound on b^* for $T \ge t+1 > T'$:

$$\begin{aligned} b^* &\overset{(a)}{\le} \sum_{i=1}^{n}\Big(a_{ji}[h_i^*]^\top\theta_i^* + 2|a_{ji}|\sqrt{\beta^t}\|h_i^*\|_{[V_i^t]^{-1}}\Big)\\ &\overset{(b)}{=} c_j + \sum_{i=1}^{n} 2|a_{ji}|\sqrt{\beta^t}\|h_i^*\|_{[V_i^t]^{-1}}\\ &\overset{(c)}{\le} c_j + \sum_{i=1}^{n}\frac{2|a_{ji}|\sqrt{\beta^t}\|h_i^*\|}{\sqrt{\lambda_{\min}(V_i^{T'+1})}}\\ &\overset{(d)}{\le} c_j + \sum_{i=1}^{n}\frac{2|a_{ji}|\sqrt{2\beta^t}\|h_i^*\|}{\sqrt{2\nu+\lambda_- T'}}\\ &\overset{(e)}{\le} c_j + \frac{2n\kappa L\sqrt{2\beta^T}}{\sqrt{2\nu+\lambda_- T'}} =: c_j + \ell_t \end{aligned} \tag{35}$$

The step (a) is due to the closed form solution for the support function of an ellipsoid (e.g. [41] Eq. 19.13), (b) is due to Lemma 5, (c) is due to the fact that $\lambda_{\min}(V_i^t) \ge \lambda_{\min}(V_i^{T'})$ for $t \ge T'$, (d) is due to Lemma 7 and (e) is due to Assumptions 3, 4, and the fact that β^t increases with t. Using the same process, we can see that $b^0 \le c_j^0 + \ell^t$, where $c_j^0 = \sum_{i=1}^n a_{ji}\theta_i^T h_i^0$.

We can now return to bounding α^t in (33). Because $G^0 \subseteq \tilde{G}^t$ for all $t \ge 1$, we know that h^0 is in $\tilde{G}^t$ for all $t \ge 1$. We use the following Lemma to show that $b^* \ge c_j$.

Lemma 8. *Assume the same as Lemma 5 and let j be a constraint satisfying Lemma 5. Then, we have that $\max_{\theta\in\tilde{C}^t}\sum_{i=1}^n a_{ji}\theta_i^\top h_i^* \ge c_j$.*

Therefore, we know that $b^* \ge c_j$ and $b^0 \le c_j$ (as $[h_1^{0,\top}\theta_1 \ \dots \ h_n^{0,\top}\theta_n]^\top$ is in E for all θ in C^0), which imply that $b^* - b^0 \ge 0$ and $\tilde{c}_j = c_j - b^0 \ge 0$. With these facts in mind and the bounds on b^* and b^0, we have that

$$\begin{aligned} &c_j \le \alpha^t b^* + (1-\alpha^t)b^0\\ &\Rightarrow \alpha^t \ge \frac{c_j - b^0}{b^* - b^0} = \frac{\tilde{c}_j}{b^* - b^0} \ge \frac{\tilde{c}_j}{c_j + \ell^t - b^0} = \frac{\tilde{c}_j}{\tilde{c}_j + \ell^t} \end{aligned} \tag{36}$$

From the bound on b^0, we also have that

$$\tilde{c}_j = c_j - b^0 \ge c_j - c_j^0 - \ell^t \ge \zeta - \ell^t, \tag{37}$$

where the last inequality follows from Assumption 4 which implies that for any h in G^0 (such as h^0), it holds that $c_j - \sum_{i=1}^n a_{ji}h_i^\top\theta_i^* \ge \zeta$ for all j in $[p]$. Thus, $1-\alpha^t \le \frac{\ell^t}{\tilde{c}_j+\ell^t} \le \frac{\ell^t}{\zeta}$. Therefore, we have the bound on r_{t+1}^I for all t such that $T \ge t+1 > T' \ge t_\delta$:

$$r_{t+1}^I \le \frac{4\sqrt{2}\kappa M n^2 L^2 S\sqrt{\beta^T}}{\zeta\sqrt{2\nu+\lambda_- T'}} \tag{38}$$

Note that this bound only holds if both the events in Theorem 1 and Lemma 7 hold. Since each of these events happen with probability at least $1-\delta$, they jointly hold with probability at least $1-2\delta$ by the union bound.

SPENCER HUTCHINSON is currently pursuing the Ph.D. degree in electrical and computer engineering from the University of California, Santa Barbara in Santa Barbara, CA, USA. He received the B.S. degree in electrical engineering from Colorado School of Mines in 2021. His research interests include the design and analysis of optimization and learning algorithms for the control of human-cyber-physical systems.

BERKAY TURAN is pursuing the Ph.D. degree in Electrical and Computer Engineering at the University of California, Santa Barbara. He received the B.Sc. degree in Electrical and Electronics Engineering as well as the B.Sc. degree in Physics degree from Boğaziçi University, Istanbul, Turkey, in 2018. The overarching goal of his research is to design network control, optimization, and learning frameworks to promote efficiency and resiliency in societal-scale cyber-physical systems.

MAHNOOSH ALIZADEH is an associate professor of Electrical and Computer Engineering at the University of California Santa Barbara. She received the B.Sc. degree ('09) in Electrical Engineering from Sharif University of Technology and the M.Sc. ('13) and Ph.D. ('14) degrees in Electrical and Computer Engineering from the University of California Davis. From 2014 to 2016, she was a postdoctoral scholar at Stanford University. Her research interests are focused on designing network control, optimization, and learning frameworks to promote efficiency and resiliency in societal-scale cyber-physical systems. Dr. Alizadeh is a recipient of the NSF CAREER award.