2023 62nd IEEE Conference on Decision and Control (CDC)
December 13-15, 2023. Marina Bay Sands, Singapore

Learning How to Price Charging in Electric Ride-Hailing Markets

Marko Maljkovic, Gustav Nilsson, and Nikolas Geroliminis

***Abstract*— With the electrification of ride-hailing fleets, there will be a need to incentivize where and when the ride-hailing vehicles should charge. In this work, we assume that a central authority wants to control the distribution of the vehicles and can do so by selecting charging prices. Since there will likely be more than one ride-hailing company in the market, we model the problem as a single-leader multiple-follower Stackelberg game. The followers, i.e., the companies, compete about the charging resources under given prices provided by the leader. We present a learning algorithm based on the concept of contextual bandits that allows the central authority to find an efficient pricing strategy. We also show how the exploratory phase of the learning can be improved if the leader has some partial knowledge about the companies' objective functions. The efficiency of the proposed algorithm is demonstrated in a simulated case study for the city of Shenzhen, China.**

I. Introduction

With the widespread adoption of electric vehicles (EVs) as an eco-friendly mode of transportation, the need for reliable and efficient charging infrastructure has emerged as a crucial factor dictating their overall usability [1]. To profitably manage large electric fleets in the near future, big ride-hailing companies such as Uber, Lyft, etc., would likely have to devise intelligent charging strategies dictated by the spatio-temporal distribution of the power supply. On the other hand, due to the ever-increasing electricity demand, coordinated charging of large fleets could have a strong positive impact on preventing imbalances and overloads in the power network. From the perspective of the local authorities, the interplay between the energy stakeholders, the ride-hailing service providers, and the heterogeneous demand opens the door for trading different services to achieve a societal optimum in terms of energy management and congestion levels in the region. As the company operators strive to reduce their operational expenses, which among others include charging costs, there is an inherent need to minimize queuing at charging stations due to the limited capacity of the shared infrastructure. In return, this creates a competitive environment between companies as stations in high-demand areas could be more prone to overcrowding.

We envision that the central authority acts as a regional entity of sufficient regulative power. As illustrated in Figure 1, we assume the central authority is in charge of determining the charging price at each charging station. Regardless of whether the central authority is the government, the power grid operator, etc., we assume it aims to steer the outcome of the competition between the ride-hailing companies towards the system optimum by offering discounted charging at certain stations. Consequently, through smartly designed prices, the central authority would hope to motivate the management of the ride-hailing companies to also charge their fleets in the more distant areas. As a result, the central authority would reduce the burden both on the power grid and the traffic network in the demand-attractive areas. In any case, with such a pricing-oriented structure, the interactions between the central authority and the ride-hailing market align well with the concept of Stackelberg games.

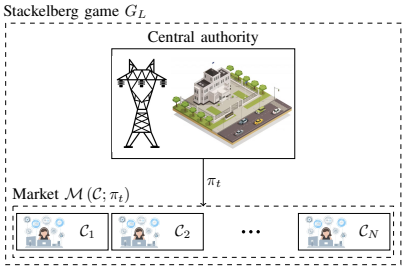

Fig. 1. Schematic sketch of the bi-level problem setting. The $|\mathcal{C}| = N$ ride-hailing companies form the market $\mathcal{M}(\mathcal{C})$, parametrized by the vector of charging prices determined by the central authority.

We aim to extend the analysis of the bi-level pricing game previously introduced in [2]–[4], where the system optimum is introduced by a predefined vehicle distribution profile. By adopting the operational cost model based on [5]–[7], the ride-hailing market becomes governed by a quadratic aggregative game. Upon the announcement of the charging prices, we assume that the rational company operators choose a no-regret strategy given by the Nash equilibrium (NE). Previous studies [2]–[4] show what can be achieved by means of collaborative information exchange between the agents on different hierarchy levels. Here, we analyze the pricing problem from the perspective of the leader and start by treating the ride-hailing market as a 'grey box' environment that provides little knowledge about the lower-level game. To begin with, we assume the central authority has no information about the functional form of the costs and constraints governing the operation of the companies. For such a scenario, we propose a contextual bandit (CB) algorithm [8] that learns how to price charging based on some intermediate aggregate information about the charging demand of the companies. In essence, the

M. Maljkovic, G. Nilsson, and N. Geroliminis are with the School of Architecture, Civil and Environmental Engineering, École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland. {marko.maljkovic, gustav.nilsson, nikolas.geroliminis}@epfl.ch.

This work was supported by the Swiss National Science Foundation under NCCR Automation, grant agreement 51NF40_180545.

An extended version containing all the proofs is available at https://arxiv.org/abs/2308.13460

environment encapsulates the pricing-induced interactions within the ride-hailing market and outputs just the attained NE. The algorithm then trains the parameters of the pricing module by observing the pairs of applied prices and observed NE. The problem of learning Stackelberg equilibria, under certain assumptions on the game structure, has already been studied in the literature [9]. In the context of pricing, there is a large body of research applying the multi-step counterpart of the contextual bandits, i.e., the reinforcement learning (RL) algorithm [10]–[12]. However, the setup of our problem does not facilitate a Markov decision process (MDP) based state transition as the prices applied at a particular instant of time have no influence on the future state of the environment described mainly by the charging demand of the vehicles. As such, our setup is not compatible with the RL-based methods.

To the best of our knowledge, there are no works directly applicable to our problem of learning the central authority's pricing strategy in a Stackelberg game with limited information about the lower-level ride-hailing market. Naturally, we complement the proposed CB-based algorithm with an analysis of the pricing procedure with respect to different levels of information available to the leader. Namely, if the central authority has access to the cost functions but not to the constraints governing the ride-hailing market, we propose how to construct an initial exploration space for gathering high-quality training data when there is no previous domain-based knowledge. Finally, if the central authority also has information about the constraints in the ride-hailing market, we show that there is no need to formulate a learning problem since the bi-level Stackelberg pricing game reduces to a mathematical program with complementarity constraints (MPCC) [13] that can be recast into an instance of a mixed integer linear or quadratic problem (MILP/MIQP) [14] using the big-M reformulation [15].

The paper is outlined as follows: the rest of this section is devoted to introducing the basic notation. In Section II, we introduce the general problem setup before defining the structure of the electric ride-hailing market in Section III. In Section IV, we then present our main methodological and theoretical results. Finally, we conclude the paper with Sections V and VI, where we test our method in a numerical case study and propose ideas for future research.

Notation: Let $\mathbb{R}$ denote the set of real numbers, $\mathbb{R}_+$ the set of non-negative reals, and $\mathbb{Z}_{>0}$ the set of positive integers. Let $\mathbf{0}_m$ and $\mathbf{1}_m$ denote the all zero and all one vectors of length m respectively, and $\mathbb{I}_m$ the identity matrix of size $m \times m$. For a finite set $\mathcal{A}$, we let $\mathbb{R}^{\mathcal{A}}_{(+)}$ denote the set of (non-negative) real vectors indexed by the elements of $\mathcal{A}$ and $|\mathcal{A}|$ the cardinality of $\mathcal{A}$. Furthermore, for finite sets $\mathcal{A}$, $\mathcal{B}$ and a set of $|\mathcal{B}|$ vectors $x^i \in \mathbb{R}^{\mathcal{A}}_{(+)}$, we define $x \coloneqq \operatorname{col}\left((x^i)_{i\in\mathcal{B}}\right) \in \mathbb{R}^{|\mathcal{A}||\mathcal{B}|}$ to be their concatenation. For $A \in \mathbb{R}^{n\times n}$, $A \succ 0 (\succeq 0)$ is equivalent to $x^T A x > 0 (\geq 0)$ for all $x \in \mathbb{R}^{n\times n}$. We let $A \otimes B$ denote the Kronecker product between two matrices and for a vector $x \in \mathbb{R}^n$, we let $\operatorname{diag}(x) \in \mathbb{R}^{n\times n}$ denote a diagonal matrix whose elements on the diagonal correspond to vector x. For matrices M_i, such that $i \in \mathcal{A}$, we let $\operatorname{Diag}(M_i)_{i\in\mathcal{A}}$ denote their concatenation into a block-diagonal matrix. For a matrix P, let $\mathbf{tr}(P)$ denote its trace.

II. Problem statement

We consider a Stackelberg pricing game where the central authority, denoted as the leading agent L, wants to steer the decisions made by the agents in the ride-hailing market $\mathcal{M}(\mathcal{C})$. The market is defined by the operational management of a set of ride-hailing companies $i \in \mathcal{C}$ operating in a region where access to a set of shared charging stations $\mathcal{H}$ for electric vehicles is offered, with $|\mathcal{C}| = N$ and $|\mathcal{H}| = M$. We analyze the problem from the short-term perspective, i.e., for one snapshot of the day in which multiple drivers from the ride-hailing companies would like to recharge. We assume that the operator of each company is responsible for coordinating the charging of the respective electric fleet, and aims to do so in an attempt to minimize the one-step-ahead operational cost. Namely, each operator aims to match the vehicles that want to recharge with the stations so as to optimize a cost that encompasses the cost of charging, the monetary equivalent of the time spent queuing at the charging stations, and the expected revenue induced by a particular coordinated charging strategy. The central authority, on the other hand, parametrizes the optimization problem of each company by choosing the charging prices $\pi \in \mathcal{P} \subseteq \mathbb{R}^M$ for the stations in the region. For each $i \in \mathcal{C}$, let N_i be the number of vehicles that want to recharge and $n = \operatorname{col}((N_i)_{i\in\mathcal{C}})$. Then, the decision variable of company $i \in \mathcal{C}$ is given by $x^i \in \mathcal{X}_i \subseteq \mathbb{R}^M$, with $\|x^i\|_1 = N_i$ and $x^i_j \geq 0$ representing the number of vehicles assigned to station j. We let the sets $\mathcal{X}_i$ encode the constraints of each company. Since x^i is chosen as a real vector, and the number of vehicles that can be sent to each station is an integer, for a perfect match to exist between the vehicles and the stations, it suffices to choose polytopic constraints as previously discussed in [2]. Therefore, we assume the sets $\mathcal{X}_i$ are given by:

$$\mathcal{X}_i \coloneqq \left\{x^i \in \mathbb{R}^M \mid A_i x^i = b_i \wedge G_i x^i \leq h_i\right\}, \qquad (1)$$

with $A_i = \mathbf{1}_M^T$, $b_i = N_i$, $G_i \in \mathbb{R}^{m_i^{\text{ineq}}\times M}$ and $h_i \in \mathbb{R}^{m_i^{\text{ineq}}}$, for properly chosen $m_i^{\text{ineq}} \in \mathbb{Z}_{>0}$. If we define sets $\mathcal{X} \coloneqq \prod_{i\in\mathcal{C}} \mathcal{X}_i$ and $\mathcal{X}_{-i} \coloneqq \prod_{j\in\mathcal{C}\setminus i} \mathcal{X}_j$, then the joint strategy of all followers can be denoted as $x \coloneqq \operatorname{col}\left((x^i)_{i\in\mathcal{C}}\right) \in \mathcal{X}$ and for every $i \in \mathcal{C}$, we can define $x^{-i} \coloneqq \operatorname{col}\left((x^j)_{j\in\mathcal{C}\setminus i}\right) \in \mathcal{X}_{-i}$. The objective of each company is to minimize the cost

$$J^i\left(x^i, x^{-i}; \pi\right) \coloneqq \hat{J}^i\left(x^i, x^{-i}\right) + \left(x^i\right)^T S_i \pi\,, \qquad (2)$$

where the first term $\hat{J}^i\left(x^i, x^{-i}\right)$ encapsulates the influence of other companies on the perceived cost, such as queuing at the station and lost income from passengers, and the second term describes the total charging price to be paid. Here, the matrix $S_i \in \mathbb{R}^{M\times M}$ is diagonal, i.e., $S_i = \operatorname{diag}\left(d^i\right) \succeq 0$, and every element $d^i_j \in \mathbb{R}_+$ of the vector $d^i \in \mathbb{R}^M_+$ can be interpreted as the expected average charging demand per one vehicle of the i-th company when choosing the station $j \in \mathcal{H}$. The π-parametetrized ride-hailing market can now

be described as a set of N optimization problems given by:

$$\mathcal{M}\left(\mathcal{C};\pi\right) \coloneqq \left\{ \min_{x^i \in \mathcal{X}_i} J^i\left(x^i, x^{-i};\pi\right), \forall i \in \mathcal{C} \right\}. \quad (3)$$

We assume the companies in the market $\mathcal{M}\left(\mathcal{C};\pi\right)$ collaborate to play a no-regret strategy according to a Nash equilibrium (NE) given in the following definition.

Definition 1: For any $\pi \in \mathcal{P}$, a joint strategy $x^* \in \mathcal{X}$ is a NE of the game played in $\mathcal{M}\left(\mathcal{C};\pi\right)$, if for all $i \in \mathcal{C}$ and all $x^i \in \mathcal{X}_i$ it holds that $J^i\left(x^{i*}, x^{-i*};\pi\right) \leq J^i\left(x^i, x^{-i*};\pi\right)$

Specifically, we focus on a subset of NE given by Definition 1, known as the variational Nash equilibria (v-NE).

Assumption 1: For any $\pi \in \mathcal{P}$, the companies in $\mathcal{M}\left(\mathcal{C};\pi\right)$ play a joint v-NE $x \in \mathcal{V}_\pi(\mathcal{M})$ described by the set $\mathcal{V}_\pi(\mathcal{M}) \coloneqq \{x \in \mathcal{X} | (y-x)^T F(x,\pi) \geq 0,\ \forall y \in \mathcal{X}\}$, where $F(x,\pi) \coloneqq \mathrm{col}((\nabla_{x^i} J^i(x^i, x^{-i};\pi))_{i\in\mathcal{C}})$.

In this paper, we assume the central authority is interested in controlling vehicle distribution among charging stations. The central authority chooses the prices $\pi \in \mathcal{P}$ in an attempt to force the company operators to coordinate charging such that the resulting total vehicle distribution matches a predefined one given by vector $\mathcal{Z} \in [0,1]^M$ with $\mathbf{1}^T\mathcal{Z} = 1$.

For every $i \in \mathcal{C}$, let $\Lambda_i \in \mathbb{R}^{M\times NM}$ be a selection matrix $\Lambda_i \coloneqq \left[\mathbf{0}_{M\times(i-1)M} \mid \mathbb{I}_{M\times M} \mid \mathbf{0}_{M\times(N-i)M}\right]$, $\Lambda \coloneqq \sum_{i\in\mathcal{C}} \Lambda_i$, $\Lambda_{-i} \coloneqq \Lambda - \Lambda_i$ and $\overline{\Lambda} \coloneqq \mathbf{1}_{N-1}^T \otimes \mathbb{I}_M$. Then, the central authority's optimization problem can be cast as

$$G_L \coloneqq \left\{ \begin{array}{c} \min\limits_{\pi\in\mathcal{P}} J^L\left(x^*,\pi\right) = \dfrac{1}{2}\left\|\Lambda x^* - \mathbf{1}^T n\mathcal{Z}\right\|_2^2 \\ \text{s.t. } x^* \in \mathcal{V}_\pi\left(\mathcal{M}\right) \end{array} \right\}. \quad (4)$$

If $\mathcal{K}$ denotes the space of available partial observations of the market state, then the central authority would like to obtain a functional $\pi : \mathcal{K} \to \mathcal{P}$ that would map each observation into properly chosen charging prices at stations $\mathcal{H}$. It is evident that the amount of shared information about the structure of the market, i.e., the cost functions $\hat{J}^i\left(x^i, x^{-i}\right)$ and constraint sets $\mathcal{X}_i$, directly dictates to what extent the problem (4) can be analytically solved. In cases when knowledge about the model is scarce, the central authority would try to learn the market behavior and how to price better through interactions with the ride-hailing market.

In the following section, we will define the elements of the underlying operational cost model of ride-hailing companies. However, in Section IV, we will start by treating this model as a black box and show that it is possible to design a general learning-based method capable of tackling such a market structure. We will then show how the initial exploration space $\mathcal{P}$ for the learning-based method can be designed should the companies be willing to disclose the information about the costs $\hat{J}^i\left(x^i, x^{-i}\right)$ and how the complete learning procedure collapses to an instance of a mixed integer linear program (MILP) if $\mathcal{X}_i$ is available as well.

III. Electric Ride-Hailing Market

Having defined the optimization problem of every agent in the system, we now proceed to define the elements of the

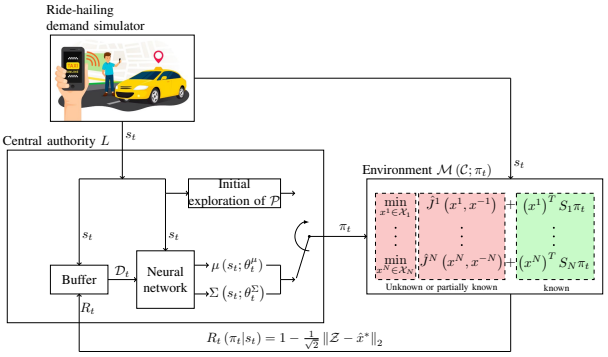

Fig. 2. Schematic sketch of the contextual bandit learning setup.

ride-hailing company's cost in more detail. Every company operator is interested in minimizing the one-step-ahead cost under the feasibility constraints imposed by the battery status of its vehicles. Inspired by the objective functions analyzed in [2], [3], [5]–[7], in this paper we analyze a cost $J^i\left(x^i, x^{-i};\pi\right) = J_1^i\left(x^i, x^{-i}\right) + J_2^i\left(x^i\right) + J_3^i\left(x^i,\pi\right)$, where $J_1^i\left(x^i, x^{-i}\right)$ denotes the expected queuing cost, $J_2^i\left(x^i\right)$ is the negative expected revenue and $J_3^i\left(x^i,\pi\right) = \left(x^i\right)^T S_i\pi$ is the charging cost introduced in (2).

The expected queuing cost model of the company $i \in \mathcal{C}$ effectively depends on the total vehicle distribution as $J_1^i\left(x^i, x^{-i}\right) = \left(x^i\right)^T C\left(x^i + \overline{\Lambda}x^{-i} - \tau\right)$, where $C \in \mathbb{R}^{M\times M}$ is a scaling matrix, i.e., $C = \mathrm{diag}\left(c^i\right) \succeq 0$, such that the element $c_j^i \in \mathbb{R}_+$ depicts how expensive it is for a vehicle to queue in the region around the station $j \in \mathcal{H}$ and $\tau \in \mathbb{R}^M$ is the vector of charging station capacities. The more the capacity of the station is exceeded, the higher the cost per vehicle should be. Hence, to calculate the total queuing cost for the whole fleet, we take the inner product between the vector describing the fleet's distribution x^i, and the incurred cost per vehicle for choosing a particular station.

The negative expected revenue is modeled as $J_2^i\left(x^i\right) = \left(e_i^{\text{arr}}\right)^T x^i - \left(e_i^{\text{pro}}\right)^T x^i$, where $e_i^{\text{arr}} \in \mathbb{R}^M$ is the average cost of a vehicle being unoccupied while traveling to a station and the vector $e_i^{\text{pro}} \in \mathbb{R}^M$ is the expected profit in regions around charging stations estimated from historical data. Therefore, the part $\hat{J}^i\left(x^i, x^{-i}\right)$ of the total cost in (2) can be simplified to a quadratic form given by

$$\hat{J}^i\left(x^i, x^{-i}\right) = \frac{1}{2}\left(x^i\right)^T P_i x^i + \left(x^i\right)^T Q_i x^{-i} + r_i^T x^i, \quad (5)$$

where $P_i \coloneqq 2C$, $Q_i \coloneqq C\overline{\Lambda}$ and $r_i \coloneqq e_i^{\text{arr}} - e_i^{\text{pro}}$. For such a game structure, the following proposition guarantees the existence and uniqueness of a Nash equilibrium.

Proposition 1: For any $\pi \in \mathcal{P}$, let the $\pi-$parametrized ride-hailing market $\mathcal{M}\left(\mathcal{C};\pi\right)$ be defined as in (3). Moreover, for every $i \in \mathcal{C}$ let the constraint sets $\mathcal{X}_i$ be defined as in (1) and the company operator's objective be defined by (5). Under Assumption 1, there is a unique v-NE joint strategy $x^* \in \mathcal{X}$ describing the interactions between the ride-hailing companies in the market $\mathcal{M}\left(\mathcal{C};\pi\right)$.

The proof is given in the extended version of our paper.

Given the aggregative structure of the game, several pricing mechanisms and computational methods to find the Nash and local Stackelberg equilibria have already been analyzed in the literature [2]–[4]. However, the underlying assumption in these works entails that all the agents are motivated to work toward the societal optimal described by the central authority's objective. Both the central authority and the ride-hailing companies work together to iteratively compute the local Stackelberg equilibrium. In this paper, we focus on the pricing problem just from the perspective of the central authority. Namely, we aim to investigate what happens if the central authority has no or partial access to $\hat{J}^i\left(x^i, x^{-i}\right)$ and $\mathcal{X}_i$ and hence has to treat the market dynamics as a black-box that outputs the attained v-NE for a particular pricing vector. Therefore, we turn to learning-based methods that involve interacting with the unknown environment and proceed to introduce the framework in the following section.

IV. Learning the charging prices

At a particular time step t, let us assume that the central authority has information about the average charging demand per vehicle and the negative expected revenue, encoded by the vectors $s_t^d \coloneqq \operatorname{col}\left((d^i)_{i\in\mathcal{C}}\right)$ and $s_t^r \coloneqq \operatorname{col}\left((r^i)_{i\in\mathcal{C}}\right)$. Then, the observed state vector of the central authority can be described by $s_t = \operatorname{col}\left(\{s_t^d, s_t^r\}\right) \in \mathcal{K} \subseteq \mathbb{R}^{2NM}$. Typical for learning-based optimization problems, in order to encourage exploration of the space of prices, the central authority chooses a pricing vector $\pi_t \in \mathcal{P} \subseteq \mathbb{R}^M$ based on the probabilistic pricing policy $\rho\left(\pi|s\right)$. Because the central authority is only focused on optimizing the one-step-ahead cost $J^L\left(x^*(\pi_t), \pi_t\right)$, and there is no clear Markov Decision Process governing the relation between the states s_t and s_{t+1}, the optimization problem of the central authority falls under the category of contextual bandits. Hence, we will now present a framework that assumes no prior knowledge about the structure of the electric ride-hailing market.

A. Contextual bandit framework

To cast the central authority's learning problem in the standard form of the contextual bandits, we propose a parametrized form of a Gaussian pricing policy given by

$$\rho_\theta\left(\pi|s\right) \coloneqq \mathcal{N}\left(\mu\left(s;\theta^\mu\right), \Sigma\left(s;\theta^\Sigma\right)\right) \ . \tag{6}$$

With this in mind, the central authority now aims to iteratively update $N_\theta \in \mathbb{Z}_{>0}$ trainable parameters $\theta = \left[\theta^\mu, \theta^\Sigma\right] \in \Omega \subseteq \mathbb{R}^{N_\theta}$ through interactions with the ride-hailing market in an attempt to maximize the instantaneous reward $R_t\left(\pi|s\right) \in [0,1]$ that describes the quality of the chosen prices. For a particular choice of θ, the objective of the central authority is to maximize the objective $J(\theta) \coloneqq \mathbb{E}_{\rho_\theta(\pi|s)}\left[R\left(\pi|s\right)\right]$ via policy gradient method first introduced in [16]. Namely, to update the parameters of the stochastic policy $\rho_\theta\left(\pi|s\right)$ via gradient descent, we aim to utilize a well-known identity concerning $\nabla_\theta J(\theta)$ and given by

$$\nabla_\theta \mathbb{E}_{\rho_\theta(\pi|s)}\left[R\left(\pi|s\right)\right] = \mathbb{E}_{\rho_\theta(\pi|s)}\left[R\left(\pi|s\right)\nabla_\theta \log \rho_\theta\left(\pi|s\right)\right] . \tag{7}$$

The right-hand side of (7) is particularly useful if we assume that at time $t = T$, the central authority has access to the history buffer $\mathcal{D}_t$ of observed interactions described by triplets $z_t \coloneqq (s_t, \pi_t, R_t(\pi_t|s_t))$, i.e., $\mathcal{D}_t \coloneqq \{z_t\}_{t\le T}$. In that case, the gradient of the central authority's objective, $\nabla_\theta J(\theta) = \nabla_\theta \mathbb{E}_{\rho_\theta(\pi|s)}\left[R\left(\pi|s\right)\right]$, can be estimated by approximating the expectation on the right-hand side of (7) via sampling from $\mathcal{D}_t$. To match the objective of the leader introduced in (4), we set the reward function $R\left(\pi|s\right) \coloneqq 1 - \frac{1}{\sqrt{2}}\left\|\mathcal{Z} - \hat{x}^*\right\|_2 \in [0,1]$ such that $x^* \in \mathcal{V}_\pi\left(\mathcal{M}\right)$ and $\hat{x}^* = \Lambda x^*/\mathbf{1}^T n$. By combining (6) and (7) at time step t, the central authority updates the parameters θ according to

$$\arg\max_{\theta\in\Omega} \sum_{z_k\in\mathcal{D}_t} R_k(\cdot) \sum_{j\in\mathcal{H}} \left(\log\frac{1}{\sigma_j(\cdot;\theta)} - \frac{\left(\pi_{k,j} - \mu_j(\cdot;\theta)\right)^2}{2\sigma_j^2(\cdot;\theta)}\right) ,$$

where $\sigma_j(\cdot;\theta) \in \mathbb{R}_+$ represents the jth diagonal element of $\Sigma(s;\theta^\Sigma)$, $\pi_{k,j}$ represents the jth element of the central authority's pricing vector and $\mu_j(\cdot;\theta) \in \mathbb{R}_+$ is the jth element of the mean vector $\mu(s;\theta^\mu)$.

The complete schematic overview of the learning setup is presented in Figure 2. The ride-hailing EVs serve demand based on the real taxi data from the city of Shenzhen [17] and the ride-hailing simulator then provides the state of the fleets s_t as an exogenous input to the central authority and the environment. The green part of the environment encapsulates the pricing part of the market that is inherently known to the central authority. On the other hand, the knowledge about the red part depends directly on the willingness of the companies to share information about their operational management. To start the training procedure, the buffer of the observed triplets needs to be filled with some historical data. Needless to say, the quality of the learned pricing policy depends directly on the quality of the observed data. In this paper, we obtain the historical data through initial random exploration of the pricing space $\mathcal{P}$. Generally speaking, if there is no prior knowledge about the structure of the model, one would have to explore the space $\mathcal{P} = \mathbb{R}^M$ as much as possible. However, for the structure of the market described in Section III, we will propose a method to construct a bounded search space based on the structure of $\hat{J}^i\left(x^i, x^{-1}\right)$ that guarantees attainability of all the interior v-NE x^*, i.e., v-NE satisfying $G_i x^{i*} < h_i$, that would be induced by exploring $\mathbb{R}^M$.

B. Characterizing the exploration space

For a set of pricing policies $\mathcal{P} \subseteq \mathbb{R}^M$, let the set of $\mathcal{P}$-induced interior v-NE of the market game $\mathcal{M}\left(\mathcal{C};\pi\right)$ be $\mathcal{V}_\mathcal{P} \coloneqq \left\{x^* \in \bigcup_{\pi\in\mathcal{P}} \mathcal{V}_\pi\left(\mathcal{M}\right) \,|\, G_i x^{i*} < h_i,\ \forall i \in \mathcal{C}\right\}$. Let $\overline{c} \in \mathbb{R}_{>0}$ denote the largest diagonal element of C, $\overline{N} = \max_{i\in\mathcal{C}} N_i$, $\underline{N} = \min_{i\in\mathcal{C}} N_i$ and $N_{\text{tot}} = \sum_{i\in\mathcal{C}} N_i$. According to the definition of P_i in (5), we can drop the subscript i and write $P_i \coloneqq P$ for every $i \in \mathcal{C}$. Furthermore, let $\alpha \in \mathbb{R}_{>0}$ be defined via $\alpha^{-1} = \mathbf{tr}(P^{-1})$ and let $\Psi \in \mathbb{R}^{M\times M}$ be $\Psi \coloneqq \mathbb{I}_M - \alpha \mathbf{1}_M \mathbf{1}_M^T P^{-1}$. For every $i \in \mathcal{C}$, let $\overline{r}_i \coloneqq \Psi r_i$ and $\overline{r}_{i,j} \in \mathbb{R}$ be its jth element. Let $\overline{r}_{\max} = \max_{i\in\mathcal{C}, j\in\mathcal{H}} \overline{r}_{i,j}$, $\overline{r}_{\min} = \min_{i\in\mathcal{C}, j\in\mathcal{H}} \overline{r}_{i,j}$, $\overline{z} \coloneqq \left(\overline{c} - \frac{\alpha}{2}\right) N_{\text{tot}} + \left(\overline{c} + \frac{\alpha}{2}\right)\overline{N}$ and $\underline{z} \coloneqq \frac{\alpha}{2}\left(\underline{N} - N_{\text{tot}}\right)$. Then the following theorem provides a way to construct a polytopic exploration space $\mathcal{P}$.

Theorem 1 (Relaxed exploration space): Let the market $\mathcal{M}(\mathcal{C};\pi)$ be defined as in (3). Moreover, for every $i \in \mathcal{C}$, let the sets $\mathcal{X}_i$ be defined as in (1) and the company operator's objective be defined by (5). If $\overline{\mathcal{P}}_1 = \mathbb{R}^M$, then choosing a polytopic set $\overline{\mathcal{P}}_2 := \{\pi \in \mathbb{R}^M \mid \gamma \mathbf{1}_{MN} \leq G_\pi \pi \leq \Gamma \mathbf{1}_{MN}\}$ with $\gamma := \alpha \underline{N} - \overline{r}_{\max} - \overline{z} \in \mathbb{R}$, $\Gamma := \alpha \overline{N} - \overline{r}_{\min} - \underline{z} \in \mathbb{R}$ and $G_\pi^T := [S_1^T \Psi^T \mid \ldots \mid S_N^T \Psi^T]$, yields $\mathcal{V}_{\overline{\mathcal{P}}_1} = \mathcal{V}_{\overline{\mathcal{P}}_2}$.

The proof is given in the extended version of our paper.

Uniform sampling from a polytopic constraint is, in general, a hard problem. However, we show in the extended version how $\mathcal{P}_2$ can be extended to a box superset. Finally, if the central authority also has knowledge about the inequality constraints of the ride-hailing companies, we can show that the learning process can be completely reduced to an instance of a mixed integer program.

C. Complete knowledge about the market

If the central authority has full knowledge about the cost functions and constraint sets in $\mathcal{M}(\mathcal{C})$, then the complete learning procedure can be avoided by solving an MPCC.

Theorem 2 (Global optimum feasibility check): Let the $\pi-$parametrized ride-hailing market $\mathcal{M}(\mathcal{C};\pi)$ be defined as in (3). Moreover, for every $i \in \mathcal{C}$, let the constraint sets $\mathcal{X}_i$ be defined as in (1) and the company operator's objective be defined by (5). There exists a vector $\pi \in \mathbb{R}^M$ such that $J^{L*}(\cdot,\pi) = 0$ if and only if there exists $\beta > 0$ such that following feasibility MILP has a solution

$$
\begin{aligned}
\underset{x^*,\lambda^*,\nu^*,\pi,\mathbf{m}}{\text{minimize}} \quad & 1 \\
\text{subject to} \quad & \overline{\mathbf{P}}_1 x^* + \overline{\mathbf{P}}_2 \nu^* + \overline{\mathbf{P}}_3 \lambda^* = \overline{\mathbf{r}} + \overline{\mathbf{S}}\pi \quad , \\
& \overline{\mathbf{A}} x^* = \overline{\mathbf{b}} \quad , \\
& \Lambda x^* = \mathbf{1}_M^T n \mathcal{Z} \quad , \quad \text{(8a)} \\
& \mathbf{0}_L \leq \lambda^* \leq \beta \mathbf{m} \quad , \\
& \mathbf{0}_L \leq -\left(\overline{\mathbf{G}} x^* - \overline{\mathbf{h}}\right) \leq \beta(\mathbf{1}_L - \mathbf{m}), \\
& \mathbf{m} \in \{0,1\}^L
\end{aligned}
$$

where $L = \sum_{i \in \mathcal{C}} m_i^{\text{ineq}}$, $\overline{\mathbf{P}}_1 = \mathbb{I}_N \otimes C + \mathbf{1}_N \mathbf{1}_N^T \otimes C$, $\overline{\mathbf{P}}_2 = \text{Diag}(\mathbf{1}_M)_{i \in \mathcal{C}}$, $\overline{\mathbf{P}}_3 = \text{Diag}(G_i^T)_{i \in \mathcal{C}}$, $\overline{\mathbf{r}} = \text{col}((r_i)_{i \in \mathcal{C}})$, $\overline{\mathbf{A}} = \text{Diag}(\mathbf{1}_M^T)_{i \in \mathcal{C}}$, $\overline{\mathbf{b}} = \text{col}((N_i)_{i \in \mathcal{C}})$, $\overline{\mathbf{G}} = \text{Diag}(G_i)_{i \in \mathcal{C}}$, $\overline{\mathbf{h}} = \text{col}((h_i)_{i \in \mathcal{C}})$, $\overline{\mathbf{S}}^T = [S_1 \mid \ldots \mid S_M]$.

The proof is given in the extended version of our paper.

However, if the exact minimization of the leader's objective $J^L(\cdot,\pi)$ is not possible, i.e., the feasibility MILP in Theorem 2 does not have a solution, one can search for the optimal pricing vector $\pi \in \mathcal{P}$ by solving a MIQP obtained by removing the constraint (8a) and changing the objective in the feasibility MILP to $\|\Lambda x^* - \mathbf{1}^T n \mathcal{Z}\|_2^2$. Generally speaking, in the case of full knowledge about the structure of the market $\mathcal{M}(\mathcal{C})$, avoiding the learning process is done at the expense of having to find a proper parameter $\beta > 0$ for the big-M reformulation. This is an NP-hard problem [18], tackled in reality via different heuristics [14]. However, for a dynamic scenario in which the parameters describing the cost functions J^i of the ride-hailing companies change in each iteration based on the state of the respective ride-hailing fleet, applying such a method might be less favorable than having a trained pricing agent.

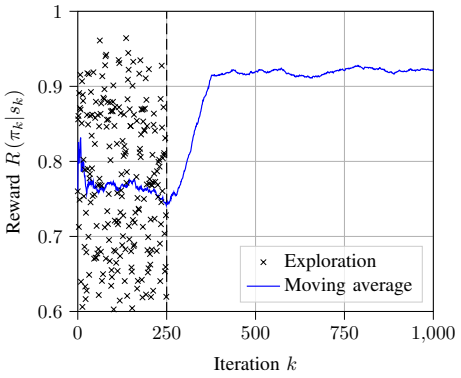

Fig. 3. Evolution of the reward signal. The solid blue line represents the moving average of the last 100 iterations.

In the next section, we will present the results obtained in a simulated case study based on real taxi data from Shenzhen.

V. Case Study

Let us assume there exist three ride-hailing companies $\mathcal{C} = \{\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3\}$ with fleet sizes $N_{\text{fleet}} = [450, 400, 350]^T$ that serve the ride-hailing demand in the Shenzhen region with four public charging stations $\mathcal{H} = \{\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3, \mathcal{H}_4\}$. The stations are located in parts of Shenzhen with different demands for ride-hailing services and are described by the vector of capacities $\tau = [15, 60, 35, 50]^T$. After a 3-hour long simulation, representing one of the two peak-hour periods during the day when the companies serve the real taxi demand obtained from [17], the vehicles whose battery level is lower than 55% are considered interested in charging. To prevent the ride-hailing vehicles from flocking in the most demand-attractive parts of Shenzhen, the desired distribution of the ride-hailing vehicles $\mathcal{Z}$ is formed so as to match the spatial distribution of the ride-hailing service requests. To approximate this distribution, the city is divided into four cells according to the Voronoi [19] partitioning of the map, with the stations chosen as the centroids of the Voronoi cells. The distribution $\mathcal{Z}$ is chosen to correspond to the total number of requests in each cell which results in $\mathcal{Z} = [0.37, 0.19, 0.27, 0.17]^T$. We run the contextual bandit for a total of $N_{\text{iter}} = 1000$ iterations, with the first $N_{\text{exp}} = 250$ used for random exploration. In the t-th iteration, the central authority samples a batch $\mathcal{B}_t$ of $|\mathcal{B}_t| = 32$ triplets from the observations in the current buffer $\mathcal{D}_t$ and uses them to perform $N_{\text{epoch}} = 20$ epochs of the parameter update procedure. After the update has been completed, for the exogenously given state s_t, the agent performs the forward propagation to obtain $\mu(s_t;\theta_t^\mu)$ and $\Sigma(s_t;\theta_t^\Sigma)$, and then samples the pricing policy $\pi_t \sim \mathcal{N}(\mu(s_t;\theta_t^\mu), \Sigma(s_t;\theta_t^\Sigma))$ for the current iteration. The pricing is then applied in the market $\mathcal{M}(\mathcal{C})$, the resulting distribution of vehicles $\hat{x}^*$ among the stations is observed, the reward $R_t(\pi_t|s_t)$ is calculated and the resulting triplet is

TABLE I

CHARGING PRICES AND ATTAINED VEHICLE DISTRIBUTIONS

Station $\mathcal{H}$	$\mathcal{Z}$	Contextual Bandit		Feasibility MILP	
		π_j	$\hat{x}_j^*$	π_j	$\hat{x}_j^*$
$\mathcal{H}_1$	0.37	3.69	0.35	3.39	0.37
$\mathcal{H}_2$	0.19	2.37	0.20	2.20	0.19
$\mathcal{H}_3$	0.27	2.87	0.26	2.83	0.27
$\mathcal{H}_4$	0.17	1.34	0.19	1.58	0.17

stored. The performance of the contextual bandit is depicted in Figure 3, where the blue line shows how the moving average of 100 samples of the reward signal evolves with respect to the iteration number. The black marks on the left-hand side of the dashed line denoting the 250-th iteration show the attained rewards during the training data collection phase. It is important to note here that the exogenous state of the system s_k is different at every iteration meaning that the system aims to map a whole distribution of ride-hailing market states into reasonably good pricing vectors. When learning begins, we can observe that the moving average curve seemingly converges after approximately 150 iterations to a value that yields good matching between the desired and attained vehicle distribution. We further support this by plotting the evolution of attained vehicle distributions in the extended version of the paper. Regarding the collection of the training data during the initial exploration phase, based on the results from [3], [4], we a priori chose $\mathcal{P} = [0.0, 5.0]^4$. Since the exogenously generated state s_k determines the parameters of the cost functions in the market for every iteration, one should verify that the sampled $\pi \in \mathcal{P}$ also belongs to the set $\overline{\mathcal{P}}_2$ given by Theorem 1. We observed that for the parameters of our case study, there was a significant gap between $\overline{\mathcal{P}}_2$ and $\mathcal{P} = [0.0, 5.0]^4$, which stresses the fact that the bounds derived in Theorem 1 are not tight. In the future, we aim to investigate how and if these bounds can be improved. Finally, for a particular state s for which the feasibility MILP in Theorem 2 was feasible, we compare the performance of the pricing policy obtained by the feasibility MILP and the fully trained contextual bandit. The comparison of the charging prices and attained vehicle distributions is displayed in Table I. The illustrated numerical values are obtained for $\pi = \mu(s; \theta^\mu)$. Clearly, the MILP algorithm achieves $R_{\text{MILP}} = 1.0$. The CB-agent, on the other hand, achieves $R_{\text{CB}} = 0.974$, which represents a minor discrepancy between the desired and attained distributions.

VI. CONCLUSIONS

In this paper, we present a learning algorithm based on the concept of contextual bandits. It is capable of learning how to map an exogenously given distribution of the ride-hailing fleet states into an adequate charging price vector for a ride-hailing market described by a quadratic aggregative game. For the case when no domain knowledge is available, but the central authority has access to the form of the cost functions in the market, we propose a polytopic search space for generating the training data in the initial phase of the process. If the central authority has full knowledge of the market structure, we show that it suffices to solve an instance of a MILP/MIQP in order to find the optimal pricing.

In the future, we aim to investigate if the bounds on the search space in Theorem 1 can be tightened and if more complex instances of the pricing games could be solved.

REFERENCES

[1] O. N. Nezamuddin, C. L. Nicholas, and E. C. d. Santos, "The problem of electric vehicle charging: State-of-the-art and an innovative solution," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–11, 2021.

[2] M. Maljkovic, G. Nilsson, and N. Geroliminis, "A pricing mechanism for balancing the charging of ride-hailing electric vehicle fleets," in *2022 European Control Conference (ECC)*, 2022, pp. 1976–1981.

[3] ——, "Hierarchical pricing game for balancing the charging of ride-hailing electric fleets," *IEEE Transactions on Control Systems Technology*, pp. 1–16, 2023.

[4] ——, "On finding the leader's strategy in quadratic aggregative stackelberg pricing games," in *2023 European Control Conference (ECC)*, 2023, pp. 1–6.

[5] Y. Yu, C. Su, X. Tang, B. Kim, T. Song, and Z. Han, "Hierarchical game for networked electric vehicle public charging under time-based billing model," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 1, pp. 518–530, 2021.

[6] W. Tushar, W. Saad, H. V. Poor, and D. B. Smith, "Economics of electric vehicle charging: A game theoretic approach," *IEEE Transactions on Smart Grid*, vol. 3, no. 4, pp. 1767–1778, 2012.

[7] E. Zavvos, E. H. Gerding, and M. Brede, "A comprehensive game-theoretic model for electric vehicle charging station competition," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 12 239–12 250, 2022.

[8] D. Bouneffouf, I. Rish, and C. Aggarwal, "Survey on applications of multi-armed and contextual bandits," in *2020 IEEE Congress on Evolutionary Computation (CEC)*, 2020, pp. 1–8.

[9] T. Fiez, B. Chasnov, and L. Ratliff, "Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 3133–3144.

[10] Y. Lu, Y. Liang, Z. Ding, Q. Wu, T. Ding, and W.-J. Lee, "Deep reinforcement learning-based charging pricing for autonomous mobility-on-demand system," *IEEE Transactions on Smart Grid*, vol. 13, no. 2, pp. 1412–1426, 2022.

[11] T. Qian, C. Shao, X. Li, X. Wang, Z. Chen, and M. Shahidehpour, "Multi-agent deep reinforcement learning method for ev charging station game," *IEEE Transactions on Power Systems*, vol. 37, no. 3, pp. 1682–1694, 2022.

[12] A. Abdalrahman and W. Zhuang, "Dynamic pricing for differentiated pev charging services using deep reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 2, pp. 1415–1427, 2022.

[13] H. Scheel and S. Scholtes, "Mathematical programs with complementarity constraints: Stationarity, optimality, and sensitivity," *Mathematics of Operations Research*, vol. 25, no. 1, pp. 1–22, 2000.

[14] T. Kleinert, M. Labbé, I. Ljubić, and M. Schmidt, "A survey on mixed-integer programming techniques in bilevel optimization," *EURO Journal on Computational Optimization*, vol. 9, p. 100007, 2021.

[15] J. Fortuny-Amat and B. McCarl, "A representation and economic interpretation of a two-level programming problem," *The Journal of the Operational Research Society*, vol. 32, no. 9, pp. 783–792, 1981.

[16] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, 1992.

[17] C. V. Beojone and N. Geroliminis, "On the inefficiency of ride-sourcing services towards urban congestion," *Transportation Research Part C: Emerging Technologies*, vol. 124, p. 102890, 2021.

[18] T. Kleinert, M. Labbé, F. Plein, and M. Schmidt, "Technical Note—There's No Free Lunch: On the Hardness of Choosing a Correct Big-M in Bilevel Optimization," *Operations Research*, vol. 68, no. 6, pp. 1716–1721, 2020.

[19] J. M. Kang, *Voronoi Diagram*. Boston, MA: Springer US, 2008, pp. 1232–1235.