Dynamic Pricing for Electric Vehicle Extreme Fast Charging

Cheng Fang, Haibing Lu, Yuan Hong, Shan Liu, and Jasmine Chang

***Abstract*— Significant developments and advancement pertaining to electric vehicle (EV) technologies, such as extreme fast charging (XFC), have been witnessed in the last decade. However, there are still many challenges to the wider deployment of EVs. One of the major barriers is its availability of fast charging stations. A possible solution is to build a fast charging sharing system, by encouraging small business owners or even householders to install and share their fast charging devices, by reselling electricity energy sourced from traditional utility companies or their own solar grid. To incentivize such a system, a smart dynamic pricing scheme is needed to facilitate those growing markets with fast charging stations. The pricing scheme is expected to take into account the dynamics intertwined with pricing, demand, and environment factors, in an effort to maximize the long-term profit with the optimal price. To this end, this paper formulates the problem of dynamic pricing for fast charging as a Markov decision process and accordingly proposes several algorithmic schemes for different applications. Experimental study is conducted with useful and interesting insights.**

***Index Terms*— Fast charging, dynamic pricing, reinforcement learning, XFC, renewable energy.**

I. Introduction

THERE has been a growing interest for electric vehicles (EV), specifically over the past decade. It had taken five years to sell the first million electric cars earlier, while it only took 6 months for so in 2018. As a remarkable milestone, the Tesla Model 3 was sold over 100,000 in a single year of 2018. In the same year, EV sales account for 49% of total vehicle sales in Norway. EVs are perceived as being connected, fun, and practical. There are several advantages of EVs over conventional internal combustion engine vehicles. For example, EVs can respond quickly and have very good torque, thanks to electric motors. Electrical vehicles are often more digitally connected than conventional vehicles. Many electric vehicle charging stations allow customers to control charging from a smart phone app. Electric vehicles reduce the emissions that notoriously cause climate change and smog, in which sense we can improve public environment and health.

Several factors have been driving the significant developments and advancement of EV technologies. One part is attributed to government policies and financial incentives, including tax credits, for lowering the up-front costs of plug-in electric vehicles. For example, the U.S. government offers $2,500 to $7,500 tax credit per new electric vehicle purchased for use [1]. Recently, many governments have launched a variety of subsidy programs for supporting the installation of a charging infrastructure, and are starting to develop regulatory initiatives to support and manage an EV fleet. Another important part is due to the decreasing battery costs. The average cost of batteries has dropped from $1,000 per kWh to roughly $227 per kWh since 2010. Since batteries take a major portion of the EV's cost, such cost saving has helped make the vehicles themselves more affordable for a myriad of potential customers.

However, there are still many challenges encountered by the wider deployment of EVs. One is that EV charging takes much longer time than gasoline cars. Currently, EV charging equipment has two basic varieties. The first category operates based on alternating current (AC) and includes Levels 1 and 2, which takes 17 - 25 hours and 4 - 5 hours respectively, to fully charge an EV with a 100-mile battery. Levels 3, 4, and 5 take 44 minutes, 15 minutes, and 6 minutes respectively, to fully charge for 100 miles. Most commercial direct-current fast chargers (DCFC) are typically of Level 3, as of 2019. Tesla's proprietary network of Superchargers, is designed to serve Tesla vehicles exclusively and is close to Level 4. Level 5 ultra-fast DCFC, also called extreme fast charging, has not yet been deployed on a commercial basis, and is expected to be standard in the future.

Another major barrier to the wide adoption of EV is the availability of fast charging stations. The term, *range anxiety*, refers to the fear that a vehicle has insufficient range to reach its destination and would thus strand the vehicle's occupants [2]. It causes many people to be reluctant to purchase EVs, even if they are cheaper and their performance is better. The commercial success of the EV will require the

Manuscript received October 16, 2019; revised January 1, 2020, January 26, 2020, and March 1, 2020; accepted March 23, 2020. Date of publication April 6, 2020; date of current version December 24, 2020. This work was supported in part by the State Grid Corporation of China under Grant 5418-201958524A-0-0-00, in part by the First Class Discipline of Zhejiang-A (Zhejiang University of Finance and Economics-Statistics), in part by the National Natural Science Foundation (NSFC) Programs of China under Grant 71722014, and in part by the Youth Innovation Team of Shaanxi Universities Big data and Business Intelligent Innovation Team. The Associate Editor for this article was X. Chen. *(Corresponding author: Haibing Lu.)*

Cheng Fang is with the School of Data Science, Zhejiang University of Finance and Economics, Hangzhou 310018, China, and also with Santa Clara University, Santa Clara, CA 95053 USA (e-mail: fangcheng0628@163.com).

Haibing Lu is with the Department of Information Systems and Analytics, Santa Clara University, Santa Clara, CA 95053 USA (e-mail: hlu@scu.edu).

Yuan Hong is with the Department of Computer Science, Illinois Institute of Technology, Chicago, IL 60616 USA (e-mail: yhong26@iit.edu).

Shan Liu is with the School of Management, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: shanliu@xjtu.edu.cn).

Jasmine Chang is with the School of Management, New Jersey Institute of Technology, Newark, NJ 07102 USA (e-mail: jschang@njit.edu).

Digital Object Identifier 10.1109/TITS.2020.2983385

development of a charging infrastructure that is accessible, easy to use, and relatively inexpensive [3]. According to Open Charge Map,[1] the U.S. has about 21,000 charging stations, while Nanalyze[2] suggests that the U.S. needs about 2 million EV charging stations for the 40 or more EVs that are likely to be on the roads in their multitudes. Currently, many EV charging station companies, like ChargePoint and EVGo, are expanding their fast charging stations.

One possible solution to alleviate the shortage of charging stations is to encourage small businesses and even home owners to join the effort. For instance, many homes in California are installed with solar panels. They can sell their excess energies for EV charging by utilizing smart solar grid. To build such an energy sharing system, the platform needs to be economically sustainable. EV charging cost should be competitive to gas-powered cars to attract more customers. While, EV fast charging operators also need to make profit to make the industry sustainable. Currently, the main pricing models for commercial charging stations are stationary *pay-for-time* or *pay-per-kWh*. Some companies adopt monthly subscription pricing strategies, as an alternative means. However, from the station operators' perspective, dynamic pricing could generate more profits than the stationary pricing scheme, considering the dynamics intertwined with user demand, pricing, and environment factors. There are a lot of research studies on dynamic pricing. Its advantages include sale boost, profit maximization, reflecting demand, and providing insights to customer behavior. Dynamic pricing is also used in other sharing ecosystems, e.g., Uber and Lyft. In fact, dynamic pricing has been adopted by the traditional electricity marketing system, as utility companies offer time-to-use policy and implement demand charge to dis-encourage peak-time electricity usage. By having slightly higher price at peak time and lower price at off-peak time, we can smooth out customer traffic, which also reduces demand charge by electricity companies. A proper dynamic pricing model could improve the profitability of fast charging stations and thus incentivize more installations of charging stations.

Dynamic pricing has been studied in many other business contexts. However, there are little studies on dynamic pricing for fast charging. The focus of this paper is to examine the dynamic pricing problem for fast charging stations, which is formulated as a Markov decision model. The main rational behind the model is the elasticity of demand to price change. By adjusting pricing dynamically, one can boost sales at different time periods, so to maximize the long-term profit. We present different algorithmic solutions to the formulated problem, including dynamic programming, Q-learning, and actor-critic. The presented solutions can account for different application scenarios to provide the best pricing policy. We also conduct numerical experiment to examine the proposed solution performance. The main contributions of this paper are summarized as follows:

- First of all, our research lays out the basic framework for the dynamic pricing problem pertaining to the fast-charging ecosystem, from the station owners' perspective.
- Second, we provide model-free reinforcement learning solutions that can learn optimal pricing policy from online experiences. It does not require the demand function to be defined in advance or necessarily time-invariant.
- Third, the proposed reinforcement learning solution can yield continuous pricing policy.

[1]https://openchargemap.org/site/

[2]https://www.nanalyze.com

II. Literature Works

Our research builds on three streams of research: (1) smart grid, (2) electricity pricing, and (3) reinforcement learning. In what follows, we review the related literature in detail and point out the contribution of our study to the literature.

Smart Grid is built on top of the traditional electricity grid, a network of transmission lines, substations, transformers and others that deliver electricity from power plant to individual homes or business facilities. The digital technology that allows for two-way communication between the utility and its customers, and the sensing along the transmission lines is what makes the grid smart [4]. The aspects that make the smart grid powerful in serving customers' electricity needs include: real-time monitoring at the transmission level, automation of various aspects of distribution systems, and smart-metering for electricity customers [5]. Renewable energies (e.g., solar and wind), besides fossil fuels, contributed 19.3% to humans' global energy consumption and 24.5% to their generation of electricity in 2015 [6]. Fluctuating renewable energy sources create the demand for Smart Energy System to integrate large penetrations of fluctuating renewable energy. Equipped with smart energy system, individual householders may benefit from the use of renewable energy, which can facilitate the growth of EV charging. Take the solar grid as an example. Solar grid is an electricity generating solar photovoltaics power system that is connected to the utility grid. The past decade has observed average annual growth rate of 50% of solar energy generation. Solar's increasing competitiveness against other technologies has quickly gained more than 2% of the total U.S. electrical generation [7]. Renewable energy systems are capable of powering houses and small businesses without any connection to the electricity grid [8]. Any excess electricity can be fed back into the grid, by which means the excess energy can be sold to utility companies or regular electric vehicle customers.

Price elasticity of demand is an economic measure that is used to quantify the responsiveness, or elasticity, of the demanded of a good or service to a change in its price. It is a key concept in economic theory. It has been well used in the electricity industry as a means of shaping electricity demand, which is also known as the *price-based demand response* [9]. Demand response provides an opportunity for consumers to play a significant role in the operation of the electric grid by reducing or shifting their electricity usage during peak periods in response to time-based rates or other forms of financial incentives. Price-based demand response has been studied for residential customers [10]. There are also studies on demand response for commercial and industrial customers [11]. However, most of existing studies are built

upon either strong assumptions or simulations, as there is no clear function linking price to demand [12]. Price elasticity of residential consumers at a macroscopic level has been studied across different countries, including the U.S., [13], U.K., [14], and many others.

As EV charging is a new and innovative practice, there are a few of studies on EV charging elasticity. But most researches support the argument that elasticity to EV charging is not the same as that to electricity consumption and a different set of pricing strategies should be developed [15]. In [16], smart grid consumers' price elasticity was estimated using a linear regression model with price changes as regressors and the corresponding shift in total demand as the response. In [17], case studies in California and Portland on hourly charging profiles for two utilities showed that the total consumption and timing of consumption vary by rate structure, and the elasticities have different values for different times and seasons. Many researches have pointed out that it is critical to consider the dynamics of consumer preferences over time. To this end, dynamic pricing has been advocated and studied for smart grid [18]. Recently, many researches have started to study the use of pricing and scheduling to improve utilization of charging stations with different goals, constraints, and procedures [19]–[26]. In contrast to the extant researches, our model utilizes reinforcement learning algorithms, which can account for the varying elasticity of demand for price in practice. Reinforcement learning has also been applied to energy management [27]–[29].

The principle of Reinforcement Learning (RL) is to earn by interacting with an environment. It has attracted a lot of public attention due to its success in artificial intelligence, e.g. Alpha-Go [30]. Nowadays, RL has been studied in many other disciplines, such as game theory [31] and finance [32]. RL has also been introduced to address dynamic pricing problem. In [33], Q-learning, a reinforcement learning algorithm is used to express anticipated future-discounted profits for possible prices and form a price-bot to adjust prices in response to the changing of market conditions. In [34], temporal difference is used to design information products' dynamic pricing from a yield management view. As opposed to high-profile reinforcement learning applications, such as Alpha-Go and self-driving car, this study does not contribute to the study of reinforcement learning. Rather, it is a novel application of reinforcement learning in the energy pricing setting. There are some differences in terms of problem characteristics. For example, Alpha-Go has a discrete action space, while self-driving car and our problem can be modeled with a continuous action space. Both Alpha-Go and self-driving car represent image as state and utilize deep neural network as approximate function. It is easy to define the state/action space for our problem and does not need complex approximation function.

III. Fast Charging Dynamic Pricing Problem

In this section, we will formally define the fast charging dynamic pricing problem and formulate it as a Markov Decision Process (MDP) problem.

A. Problem Description

We consider the EV fast charging pricing problem from the perspective of fast charging station operator who first purchases electricity from a utility company, and then acts as a retailer to sell it to customers. A fast charging station operator may own a network of stations, e.g., EVgo. Although the solution that we will present later can be easily adapted to a large EV electricity retailer, we start with examining a simple user case, where the operator has only one fast charging station with n chargers being installed. The station is open 24 hours a day, 7 days a week. The operator employs the pay-as-you-go pricing model. A customer is billed based on the length of time that the car is connected to the station. The pricing model is in fact adopted and implemented by Circuit Electrique (Canada).[3] However, the major commercial EV fast charging companies in the market do not adjust price frequently. In our setting, we allow the charging station operator to change charging rate, to serve the station owner's best financial interest. The dynamic pricing strategy, in fact, is utilized by the traditional oil industry, as gas stations change their prices on a daily basis (in the U.S., it often happens in the hours immediately following 6:00 pm Eastern Time for the gas stations to update price). In our model, we assume that the price is updated at a time interval. If the interval length is too short, it may cause resentment from customers due to high variance. On the other hand side, if the interval length is too long, the system may miss the opportunities to fully adapt to the market changes. Therefore, it is reasonable to assume to adjust the price on a hourly basis. We assume that customers are fully aware of the current price. For example, the price is broadcasted through a mobile app, e.g., the ChargePoint app. However, users do not have the knowledge of the future price at the next time period and we do not assume their forecasting ability.

The goal of the fast charging station operator is to maximize his/her long-term profit. The revenue primarily stemms from electricity sales. For the ease of mathematical modeling, we introduce some notations. First, let n_t denote the number of direct current chargers that are open in time period t. Although fast charging pricing is usually defined as price-per-minute, for the modeling convenience (without loss of generality), we let c_t^s denote the sales rate for direct current (fast) charging. u_t denotes the average utilization rate of those n_t direct current chargers. Thus, the sales on time period t is computed as:

$$c_t^s \times u_t \times n_t. \tag{1}$$

The cost of the station operator includes the fixed cost and the variable energy costs. Fixed cost for electric vehicle charger would include costs for installation, site preparation, utility service, transformer, equipment and others. There might be financial cost involved, if the operator leverages financial services to open the business. For simplicity, we assume that there is a fixed daily cost c^f.

The variable energy cost is what the station operator needs to pay to the contracted utility company (e.g. the Pacific Gas and Electric Company) for the consumed electricity. The electricity bill includes (1) the *energy charges* for the

[3]https://lecircuitelectrique.com/

total amount of consumed electricity and (2) the demand *surcharges* (if any) for the "peak usage" of electricity. For energy charges, an utility company typically sets different rates for residential charging and commercial charging. In our case, the commercial charging rate is applied. Commercial rate typically offer lower volumetric charges than residential rate. According to the U.S. Energy Information Administration, the average of commercial tariff is approximately \$0.145 per kWh in 2016, compared to \$0.176 for residential. Energy charges are published about utility companies and may vary by seasons. We assume that the energy price is fixed within a day. We are aware of that some utility companies employ the time-to-use pricing policy, which sets different prices for different time periods in a day. We can easily adapt to the scenario by dividing a day into multiple time periods, such that the energy rate is stable within each period. For modeling convenience later, we denote c_t^c as the unit energy cost at the time period t.

In addition to cost on consumed electricity volume, the commercial rate typically incurs demand charges, which is approximately \$10 - \$17 per peak kW in the U.S. in 2016. Demand charges are usually based on the highest 15-minute average usage recorded on the demand meter within a given month. Typically, if a facility tends to use a lot of power intensively over short periods, the demand charges will comprise a larger portion of the bill. If the facility uses power at a stable rate consistently throughout the month, the demand charges will generally be a smaller part of the bill. In reality, demand charges make up a significant portion of commercial and industrial consumer's total electricity costs, typically between 30 and 70 percent. In our model, we assume that the demand charge is fixed as the number of chargers are given. Hence, the demand charge is modeled as a part of the fixed cost.

We assume that the number of customer arrivals at each time period follows some probability distortion. There is no limit on the queue length. However, there is a backlog rate, which refers to the ratio of people in the queue who are willing to wait till the next time period, denoted by β. Suppose there are n customers in the queue at time period t, while only m customers can be served at one time period. Then the backlogged $(n-m)*\beta$ customers, along with newly arrivals, will stay in the queue at the next time period $t+1$.

We also assume that customers respond to price change, which is referred to as price elasticity of demand in economic theory. It is computed as the percentage change in quantity demanded divided by the percentage change in price:

$$e = \frac{\%\ change\ in\ sales}{\%\ change\ in\ price}. \tag{2}$$

In the context of electric vehicle, customers are sensitively responsive to the price change on the retail price of energy. Strategically, the fast charging operator may choose to lower the retail price to induce more customers in an attempt to boost revenue during off-peak season, while raise the retail price at peak time. In such a fashion, utilization rate throughout a long time period could be smoothened out, which could maximize the the operator's long-term profit. Suppose d_t is the demand at time period t which is drawn upon some distribution with original price. Now given the price change of Δc_{t+1}^s and the average customer elasticity e, the demand at time step $t+1$ is updated as:

$$d_{t+1} \leftarrow e\frac{\Delta c_{t+1}^s}{c_t^s} + d_t. \tag{3}$$

By adjusting the price, the station operator can control the number of consumers in the queue. For example, if the operator expects less arrivals in the subsequent time periods, s/he can raise the price; otherwise, if more arrivals are expected, the station operator can lower the price. By strategically adjusting the price, the station operator aims to maximize the expected overall revenue. It is our objective to device the best pricing strategy.

B. Markov Decision Process Formulation

The goal of the EV fast charging dynamic pricing problem is to maximize the profit of the fast charging station operator. The problem can be formulated as a *Markov decision process* (MDP). Before we present our Markov decision process formulation, we first introduce some basic concepts. A Markov decision process is a discrete-time stochastic control process and provides a mathematical framework for modeling decision making in situations where outcomes are partly random and partly under the control of a decision maker. At each time step, the system is in some state s, and the decision maker may choose any action a that is available in state s. The process responds at the next time step by randomly moving into a new state s', and accordingly it yields the decision maker a corresponding reward r. The probability that the process moves into its new state s' is influenced by the chosen action a. Specifically, it is characterized by the state transition function $p(s'|s,a)$. Thus, the next state s' depends on the current state s and the decision maker's action a as well. Due to the memoryless feature of MDP, the transition is merely dependent on s and a, however it is independent of any previous states and actions.

As we can see, there are some basic elements for the Markov decision process model, i.e. state, action, reward, and transition probability. Now we proceed to instantiate the basic model elements pertaining to the fast charging dynamic pricing problem.

1) State: State is the phenomenon representation of the environment and is formally a set of history and current occurrence. State will be leveraged for agent (e.g., the fast charging station owner) to make action decision. However, agent might not be able fully observe the environment. So in our problem setting, state is the information about the environment that the fast charging station owner can collect to help him/her to make strategic decision to maximize long-term profit. In our problem setting, there can be several groups of features at time period t to describe state s_t, such as pricing features, sales features, customer traffic features, and competitiveness features. Pricing features contain payment information. Sales features include consumed electricity volume v_t, number of chargers that are available at time t, and utilization rate of chargers u_t. It may also include how much time for each car spend at the charging

station. Customer traffic features are used to represent user demand information, which is not easy to accurately capture in reality. Traffic features may include the number of cars passing by at time t and even remaining electricity quantity for each EV, if EVs are equipped with cloud-based connectivity. Competitiveness features describe what alternative options that EV drivers have. They may include availability, type, and pricing of nearby charging stations, charging options at home, etc. However, the above is the conceptal model. In reality, it might not be feasible to include all features, due to difficulties in data collection and computational challenge.

2) Action: Action is what an agent can do after observing the system state at each time step and it will be fed back to the environment to impact the next environment state. Once the physical site of a fast charging station is built, the operator can take a variety of operational actions to impact sales. For example, an agent can leverage some pricing strategies, such as real-time price announced at each time step t, sales promotion, membership or subscription based pricing model, etc. Another thing that an agent can do is changing the number of open chargers. Recall that demand charge accounts for a large portion of electricity bill. So during off-peak seasons, it might be optimal to keep less chargers open to the public to save cost. Although there are many other actions that an operator can take, our study just focuses on the action of changing price.

3) Reward: As a distinctive feature of MDPs, reward, paid by the environment to the agent, defines the purpose or goal of the agent. At each time period t, the reward is a simple number, $r_t \in \mathcal{R}$. The agent's goal is to maximize the total amount of rewards it receives over the long run. Let G_t denote the total discounted future reward starting from time step t, which is

$$G_t = \sum_{k=t+1}^{\infty} \gamma^{k-t-1} r_k, \tag{4}$$

where the constant $\gamma \in [0\,1]$ is the discount rate. The discount rate is a mathematical trick to make an infinite sum finite. It has practical meanings as well. If the reward is money, the money that were received k times steps in the future would have less value than now, due to the discount factor. It also reflects how the agent values the future reward. If $\gamma = 0$, the agent is "myopic" in being concerned only with maximizing immediate rewards, as its objective is to maximize r_{t+1}. If γ approaches 1, the agent becomes more farsighted and takes future rewards into account more strongly. In our setting, the reward r_t is the profit that the fast charging operator makes at time t. Definitely, there are many other ways to model the reward function for different objectives. For instance, if the goal is to smooth out electricity usage, the utilization at each time step can be used to model the reward. Road traffic can obviously influence revenue and profit. To signal the system to target consistent profit, we can use the ratio of profit over traffic as the reward. In our study, the objective is to max G_t while in period t.

4) Transition Probability: The key assumption of an MDP model is the transition probability $p(s'|s, a)$, while the environment state changing from s to s' under agent's action a. $p(s'|s, a)$ is also referred to as the model of the environment. One might be able to obtain or estimate the model, depending on specific problems. The goal of an agent is to collect as much reward as possible. So to act optimally to reach the best performance, the agent must reason about the long term consequences of its actions (i.e., maximize future income), although the immediate reward associated with this might not be significant. Thus, the MDP model is particularly well-suited to problems that include a long-term versus short-term reward trade-off. Our problem setting posses its unique features which differentiate it from many successful applications of MDP. For instance, its transition probability cannot be easily generated in a simulated environment, although it is feasible under some strong assumptions. For applications like chess, it is clear on how the game state changes under each move. For our pricing problem, if the state space include the aforementioned features (e.g., real-time traffic, competitor strategies, etc.), many of them are not fully observed, and it becomes infeasible to estimate the transition probability. For such a case, RL can be a natural choice. RL is also preferred when the elasticity factor is delayed, as it does not require station transition probabilities to be known in advance. RL approaches, such as q-learning and actor-critic, learn from observations and update/improve state-action values when more samples are collected. Even if the elasticity is delayed, given enough learning time, the real values will be captured.

IV. Algorithms

In this section, we present several solutions to the formulated dynamic pricing MDP problem, including dynamic programming, Q-learning, and actor-critic, addressing different scenarios. Dynamic programming is applicable to Markov decision problem only when transition probability is available and the problem scale is relatively small. It is hence called model-based solution. In contrast, Q-learning does not requires transition probability, therefore it is called model-free solution. However, it does requires the action space to be finite. In our setting, pricing choices is limited to be discrete. Without much limitation, actor-critic can handle infinite action space and supports continue pricing.

A. Dynamic Programming

Dynamic programming refers to a collection of algorithms that can be used to compute optimal policies, for a perfect model environment. A perfect model means state, action and reward spaces are well defined and state transition probabilities are known. Dynamic programming might not be tractable for large-scale problems. In other words, exact solution for dynamic programming is only possible for a finite MDP. In the finite MDPs, the sets of states, actions, and rewards ($\mathcal{S}$, $\mathcal{A}$, $\mathcal{R}$) all have a finite number of elements. In this case, the random variables r_t and s_t have well defined discrete probability distributions dependent only on the preceding state and action. Therefore, for particular values of these random variables $s' \in \mathcal{S}$ and $r \in \mathcal{R}$, there is a probability of those values occurring at time t while observing particular values of the preceding state and action.

As introduced above, in our problem setting, the state space can characterize a variety of features, e.g., pricing, revenue, traffic, and competitors. To solve our formulated model via dynamic programming approach, we need to reduce the state space size. As a matter of fact, it is cumbersome or even intractable to collect information on real-time traffic and competitor information. So finite MDPs with dynamic programming might be a practical solution. We might simply consider a state as (u, t), where $u \in [0, 1]$ is the utilization rate and t is time step. To obtain finite states, we can discretize the utilization value space $[0, 1]$ into the set of intervals such as $\{[0, 0.1), [0.1, 0.2), \ldots, [0.9, 1]\}$. Time steps can be of the hourly basis $\{0, 1, \ldots, 23\}$. The action space $\mathcal{A}$ can be discretized as well. The pricing problem has a cold-start issue that is how to determine the price and number of open chargers for time $t = 0$. If we consider the action at time step 0 as an variable, it would make the solution space too large and cause difficulty in finding a good solution. Therefore, a compromising and also practical-solvable solution is to set the price c_0^s and number of open chargers n_0 at time step $t = 0$, according to historical pricing and sales data. Then the action at subsequent time step t is Δc_t^s, which denotes the price change ratio.

The challenge remains to obtain the state transition probability $p(s_{t+1}|s_t, a_t)$. If we allow to conduct an online experiment, $p(s_{t+1}|s_t, a_t)$ can be estimated by trying various action values for a large number of times. However, the experiment would be expensive and put the business at the risk of great financial loss. Dynamic programming is typically used in the setting where transition probability can be derived. One turnaround is to estimate state transition probability with reasonable assumptions, which is a common practice for statistical modeling approaches. To estimate utilization rate of an EV fast charger station at time step t, first we need to assess demand d_t, which can be done by utilizing many resources, e.g. utilization rate at a nearby gas station, real-time public traffic information, and historical information of some existing fast-charging networks. Furthermore, d_t should follow a certain distribution that can be estimated from historical data. Suppose we fix electricity price and number of open chargers. Then real sales is the minimum of d_t and capacity at that fast charging station.

So far, we have collected all the necessary bricks to build the MDP model with finite state and action space and known state transition probability. Recall that dynamic programming refers to a collection of algorithms that can be used to compute optimal policies. We employ the value iteration algorithm to find the optimal policy, which is summarized in Algorithm 1. Value iteration is derived from the classic Bellman optimality equation:

$$\begin{aligned} v_*(s) &= \max_a E\Big[R_{t+1} + \gamma\, v_*(S_{t+1}|S_t = s, A_t = a)\Big] \\ &= \max_a \sum_{s',r} p(s', r|s, a)[r + \gamma\, v_*(s')]. \end{aligned} \tag{5}$$

Intuitively, the Bellman optimality equation express the fact that the value of a state under an optimal policy must equal the expected return for the best action from that state. Value iteration starts with an initial assignment to all state values and then iteratively update all state values according to the Bellman optimality equation until state values converge. As described in Algorithm 1, a small threshold θ is used to terminate the algorithm when an accurate enough solution is found. The smaller θ (e.g., $\theta = 0.001$), the more accurate of the solution.

Algorithm 1 Value Iteration

Data: a small threshold $\theta > 0$ determining accuracy of estimation

Result: policy $\pi(s)$, where state s is a pair of values (n, h), where n is the number of cars being serviced and h is hour

Initialize $V(s)$ arbitrarily except that **repeat**
  $\Delta \leftarrow 0$;
  for $s \in \mathcal{S}$ **do**
    $v \leftarrow V(s)$;
    $V(s) \leftarrow \max_a \sum_{s',r} p(s', r|s, a)[r + \gamma\, V(s')]$;
  end
until $\Delta < \theta$;
$\pi(s) \leftarrow \arg\max_a \sum_{s',r} p(s', r|s, a)[r + \gamma\, V(s')]$

B. Q-Learning

The value iteration algorithm introduced above is elegant, but has several limitations. One is the requirement of state transition probability, as needed by the step $\max_a \sum_{s',r} p(s', r|s, a)[r + \gamma\, V(s')]$. The EV fast charging dynamic pricing problem does not have obvious state transition probabilities. To use the value iteration algorithm, we need to reduce and discretize the state and action spaces, and add strong assumptions on statistical distribution of customer demand at different time steps, and price elasticity of demand, so to derive the state transition probability. However, those assumptions might not be accurate or hold in time, as state transition probability can be non-stationary in reality. If we are able to learn optimal policies with online experiences and gradually improve policy as more instances are observed, then we can overcome the limitations.

Q-learning is a model-free RL algorithm [35]. Q-learning is an off policy algorithm that seeks to find the best action to take given the current state. It is considered as an off policy because the Q-learning function learns from actions that are likely outside the current policy, in a sense of taking random actions. Therefore, a policy is not designated or needed. More specifically, Q-learning seeks to learn a policy that maximizes the total reward from online experiences.

Recall that dynamic programming aims to find state value $v(s)$, i.e., the total expected discounted future reward at state s. Given $v(s)$, the optimal policy $\pi(a|s)$ is $\text{argmax}_a \sum_s' p(s', r|s, a)r$. While, Q-learning computes the state-action value $Q(s, a)$, which refers to the expected overall future rewards at state s given action a. So if all state-action values are given, then the optimal policy is $\pi(a|s)$ is $\text{argmax}_a Q(s, a)$. It immediately tells which price changing action to take at any state s. As Q-learning is a model-free method, $Q(s, a)$ is learned and updated from online experiences. Initially, $Q(s, a)$ can be set with any arbitrary value.

At each time period in a day, we choose the action a from the action space $\mathcal{A}$ that has the maximum value of $Q(s, a)$. As $Q(s, a)$ is randomly generated initially, such a best action might not be the optimal. To explore the action space, ϵ-greedy can be employed, where ϵ is a constant of a small value. It is to choose the best action based on the current state-action values $Q(s, a)$ with certain probability (e.g. 90%) or a random action (e.g. with 10%). By gradually reducing the probability of choosing a random action, the state-action value $Q(s, a)$ converges. The whole process can be conducted in a real environment without knowing state transition probability. The pseudo-code is provided in algorithm 2. The algorithm has several parameters. ϵ is the parameter, controlling the trade-off between exploration and exploitation. It it typically set as 0.9. γ is the discount rate, which is also used in iterative method. α is called step size, which can account for non-stationary problems. If α decreases proportionally to the number of episodes, then the algorithm takes account of all history. If α is set as a constant value (e.g. 0.1), then $Q(s, a)$ is updated mainly based on recent experiences, which addresses issues with non-stationary transition probability. The Q-learning algorithm is described in Algorithm 2.

Algorithm 2 Q-Learning

Data: step size $\alpha \in (0, 1]$, small $\epsilon > 0$
Result: Initialize $Q(s, a)$, for all $s \in \mathcal{S}$, arbitrarily except that $Q(terminal, .) = 0$
for *each episode* **do**
  Initialize $Q(s, a)$ with random small values;
  for *each step of episode* **do**
    Choose a from S using policy derived from Q, using ϵ-greedy;
    Take action a, observe r and s';
    $Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_b Q(s', b) - Q(s, b)]$;
    $s \leftarrow s'$

C. Actor-Critic Method for Continuous Pricing

The previous algorithms assume that the action space of pricing change is finite and discrete. In reality, we may want to expand the action space to allow continuous pricing. Discrete pricing is like coarse-grained pricing, while continuous pricing is more of fine-grained pricing. The fine-grained pricing scheme gives more choices to users and hence attract more customers, as opposed to coarse-grained pricing would turn away potential customers, as they do not see the price choice that is appealing to them. With fine-grained (continuous) pricing, the overall sales/profits can be improved. To achieve continuous pricing, one way is to have a very large pricing space, so to be close to continuous pricing. If so, the algorithm needs to learn a large number of state-action values, which is not feasible. Actor-critic method can handle continuous action space [36], which uses function approximation. Instead of storing $Q(s, a)$ in a table for all possible state s and action a, we represent state-action value as a function $Q(s, a|\theta^Q)$, where θ^Q is function parameter to be learned. The policy is also represented as a function $a = \mu(s|\theta^\mu)$, where θ^μ is function parameter. It is the actor part of the model. With given θ^μ, the function $\mu(s|\theta^\mu)$ can return action for any state. Now the goal is to learn θ^Q and θ^μ. The actor-critic algorithm is described in Algorithm 3. We use polynomial functions for $Q(s, a|\theta^Q)$ and $\mu(s|\theta^\mu)$.

Algorithm 3 Actor-Critic

Result: critic network $Q(s, a|\theta^Q)$ and actor $\mu(s|\theta^\mu)$
Randomly initialize critic network $Q(s, a|\theta^Q)$ and actor $\mu(s|\theta^\mu)$ with weight θ^Q and θ^μ;
Initialize target network Q' and μ' to be Q and μ respectively;
for *each episode* **do**
  Initialize a random process $\mathcal{N}$ for action exploration;
  Receive initial observation state s;
  for *each step of episode* **do**
    Select action $a_t = \mu(s_t|\theta^\mu) + \mathcal{N}_t$ according to the current policy and exploration noise;
    Execute action a_t and observe reward r_t and observe new state s_{t+1};
    Store transition (s_t, a_t, r_t, s_{t+1});
    Sample a random mini-batch of N transitions (s_i, a_i, r_i, s_{i+1});
    $y_i \leftarrow r_i + \gamma\, Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'})$;
    Update critic by minimizing the loss $\frac{1}{N}\sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$;
    Update the actor policy using the sampled policy gradient: $\Delta_{\theta^\mu} J = 1/N \sum_i \Delta_a Q(s, a|\theta^Q) \Delta_{\theta^\mu} \mu(s|\theta^\mu)$;
    Update critic network and actor: $\theta^{Q'} \leftarrow \tau\theta^Q + (1-\tau)\theta^{Q'}$, $\theta^{\mu'} \leftarrow \tau\theta^\mu + (1-\tau)\theta^{\mu'}$;

In the above sections, we assume that the elasticity is stable. But in reality, a customer's sensitivity to price might change due to various reasons. As such, the state transition probability becomes non-stationary. For non-stationary problems, to make the dynamic programming approach work, we have to keep track of elasticity and updating state transition probabilities. It is costly and not practically feasible. Q-learning and actor-critic does not require state transition probabilities as input and thus can handle non-stationary problems. Consider Q-learning. It updates the state-action value $Q(s, a)$ by $Q(s, a) + \alpha[r + \gamma max_b Q(s', b) - Q(s, b)]$, where α is the step-size parameter. If α is $1/n$, where n is the number of steps, it updates $Q(s, a)$ as the average of all past observations. When α is a constant, e.g. 0.01, it always updates $Q(s, a)$ as the mean of the past 100 observations. As such, the non-stationary factor can be effectively captured and fed into the model.

V. Numerical Study

In this section, we conduct an experimental study to evaluate the three algorithmic solutions of dynamic programming, Q-learning, and actor-critic methods.

A. Dynamic Programming

Dynamic programming is model-based solution and only applicable to the finite MDPs with known state transition probabilities. To realize the finite MDPs model, we need to estimate $p(s', r|s, a)$, which is the transition probability from an utilization rate internal to another interval at a time step t under price change ratio $a \in \mathcal{A}$. Given an online environment, it is feasible to obtain such transition probabilities by collecting enough observations and trying with different price change polices. However, we do no have such an environment. Even if the environment is available, it may take a long time to obtain accurate estimates. One circumvention is to add some realistic assumptions and then generate transition probabilities with simulations.

In what follows, we describe our data simulation process. According to the 2017 EV analysis report by RMI.org, users of EVGo's Level 3 network in California, for instance, average just 5-12 kWh per session (or enough to drive an additional 15-36 miles), which is about 10 minutes. For XFC charging, it can charge about 100 miles with the same 10 minutes. We assume that users of XFC charging exhibit the similar behavior. If so, we simply assume that a XFC charger can accommodate 6 customers per hour and there are 4 chargers. Hence, if there are 24 customers being served at a time step, the utilization rate at that time step is assumed to be 100%. We also assume that demand, i.e. the number of arrivals, at each time step follows a Poisson distribution, which is a common practice for service-type problems. Demand may fluctuate throughout a day. It is not easy to capture the real demand for a specific fast charging site, as the real demand is impacted by many unobservable factors. However, it is reasonable to assume that demand at each time step is linearly correlated to real-time traffic, which can be captured. To make the simulation more realistic, we use the highway traffic data collected by the Wisconsin Department of Transportation,[4] which collects continuous count data from 318 permanent data collection locations primarily located on the State Trunk Highway System. Data at continuous county sites are scheduled to be collected in hourly intervals each day of the year. For example, the normalized hourly traffic from 0:00 am to 23:00 pm at the site name of "030010, 15008, NW" in the Barron county on Monday are: $\{0.5, 0.4, 0.4, \ldots, 1.9, 1.2, 0.8\}$. For the simulation purpose, we multiply the normalized values by 4. The resultant mean traffic at each time period is depicted in Figure 1.

We break utilization rate into 5 buckets and consider five price changing actions $\{-0.2, -0.1, 0, 0.1, 0.2\}$. To avoid the cold-start issue, we assume that the price at the first time period is given, which in reality, can be obtained from experiences. The policy to be determined is the choice of action at each subsequent time period. To apply dynamic programming, we need to estimate transition probability among states, i.e. (utilization, time period), under each action. As it is difficult to calculate such state transition probability analytically, we resort to Monte Carlo simulation. In particular, we first simulate many episodes (days) and then estimate transition probability based on generated samples. At each time period, we sample the number of arrivals according to the Poisson distribution $P(d_t) = e^{-\lambda_t} \frac{\lambda_t{}^{d_t}}{d_t!}$, where λ_t is the average number of arrivals at time t, and their values can be retrieved from Figure 1. Then, we randomly pick an action and apply backlog rate β, and repeat the steps until the episode completes. Note that the Poisson distribution is a common choice for modeling the probability of a given number of events occurring in a fixed interval of time, e.g. the number of jumps in a stock price in a given time interval and the number of deaths per year in a given age group. We can use other alternatives, such as the negative binomial distribution or Conway-Maxwell-Poisson distribution for the data simulation purpose. The choice of distribution function does not impact our pricing learning algorithm.

[4]https://wisconsindot.gov

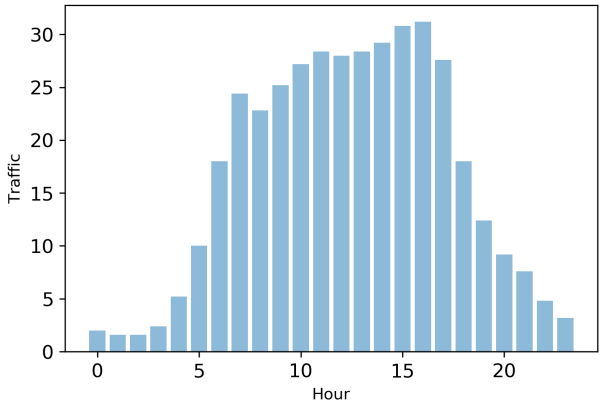

Fig. 1. Mean hourly traffic.

Algorithm 4 Estimating State Transition Probability via Monte Carlo

Data: average demand at the h-th time period λ_h, backlog rate β, elasticity e, service capacity m, price at the first time period

Result: $p(s_{t+1}|s_t, a)$, where state s_t is a pair of values (u_h, h) with u being the utilization rate of the charging station and h being hour, and action a is in the set of price changes $\mathcal{A}$

Initialize $p(s_{t+1}|s_t, a)$ to be 0;
Initialize b_0 (backlog at $t = 0$) to be 0;
Initialize a_0 (price change action at $t = 0$) to be 0;
for *each hour h in {0,...,23}* **do**
  Sample a_h from $\mathcal{A}$ (a is 0 for the first time period);
  Sample d_h from Poisson distribution $P(d) = e^{-\lambda_h} \frac{\lambda_h{}^{d}}{d!}$, where λ_h is the expected number of arrivals at time period h;
  Utilization rate at the hth time period is $\min((1 + ea_h)d_h + b_h, m)/m$, where $(1 + ea_h)d_h + b_h$ is the total number of customers in the queue;
  Backlog b_{h+1} is $\beta \max(d_h - m, 0)$;
Repeat the above steps by a large number of iterations to obtain a good estimate of $p((u', h + 1)|(u, h), a)$;

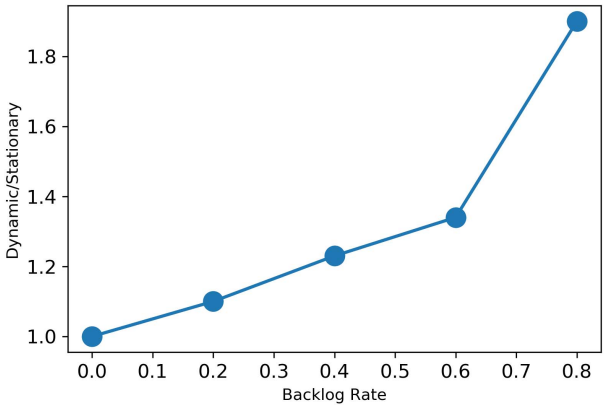

Fig. 2. Dynamic pricing versus stationary pricing.

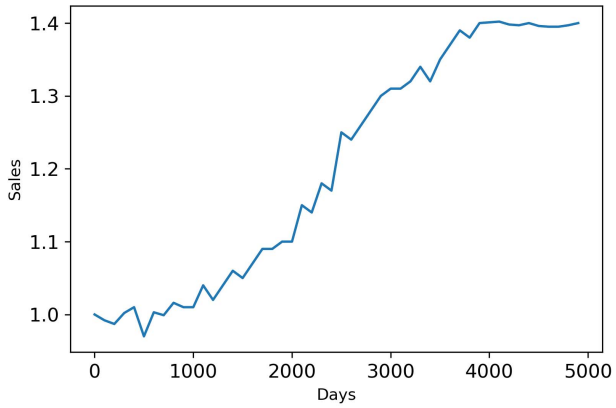

Fig. 3. Convergence of Q-learning.

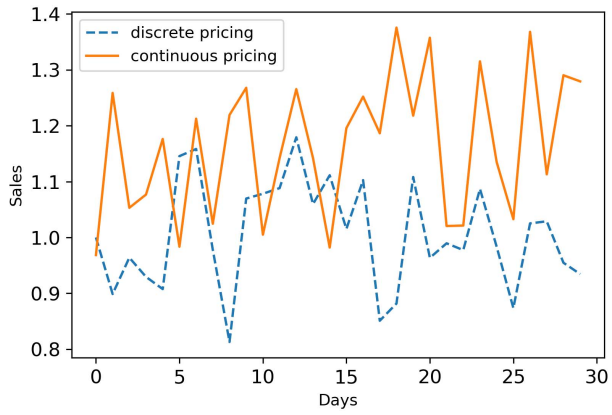

Fig. 4. Continuous pricing (Actor-Critic) v.s. discrete pricing (Q-learning).

After obtaining state transition probability, the iterative method for dynamic programming can converge such that the optimal dynamic pricing policy can be obtained. We consider the stationary pricing as a benchmark, for which we keep the price at the first time period throughout the day. To show the efficiency of the dynamic pricing strategy, we compare the expected total revenue per day, based on the optimal pricing policy obtained by dynamic programming, compared with that under the stationary pricing policy. The result is depicted in Figure 2, where the x-axis is the backlog rate and the y-axis is ratio of revenues between dynamic pricing over stationary pricing. The result shows dynamic pricing performs better than stationary pricing consistently. Importantly, the higher the backlog rate, the better performance of the dynamic pricing policy. The underlying rational is that with a high backlog rate, there is an advantage from a systematic pricing policy to benefit from balanced demands throughout the day.

B. Q-Learning

As stated before, dynamic programming can only apply to problems with known state transition probability. In the previous experiment study, we need to add many assumptions in order to use Monte Carlo method to simulate state transition probability. In reality, many assumptions may not hold. For example, the distribution on number of arrivals at each time period may not be stationary or even known. Price elasticity of demand is not known either. It may fluctuate and be subject to many other environment factors. In such case, we need a model-free method, e.g., Q-learning, which learns transition probability from online experiences. To evaluate the performance of Q-learning approach, we use the same simulation environment as the previous example. The difference is that we apply the Q-learning algorithm, as episodes are being sampled, without waiting until the end of simulation to estimate state transition probability. We set the backlog rate as 0.6. We set the learning rate α as 0.1 and the discount rate γ as 0.9, which conforms with the common practice. The result is reported in Figure 3. We divide the numbers by the first day of sale for normalization. The graph shows that the sales fluctuate initially and then gradually evolve to a stable state. The downside is that it takes over 4000 days to reach a relatively stable state, which is not acceptable in reality. To overcome the obstacle, a practical solution is to start with some simulation environment, which is built on past experiences, and obtain a decent policy. Then we use that policy as a starting policy for Q-learning and improve it in the real environment. It coincides with the popular concept, transfer learning, in machine learning.

C. Actor-Critic

In the previous Q-learning experiment, we let action of price change to be one of $\{-0.2, -0.1, 0, 0.1, 0.2\}$. It limits the power of reinforcement learning, as action space is obviously larger than five. By allowing a number of actions, it would cause the convergence problem to Q-learning. Gradient policy method addresses the issue and allows infinite action space. We initialize both critic and actor network functions as quadratic polynomial functions with parameters to be random small values that are close to zero. We set the discount factor γ as 0.95 and τ as $1/n$, where n is the number of episodes, and the mini-batch size as 50. To evaluate gradient policy method, we use the same simulation environment as before. The only change is to make action space continuous as $[-0.2, 0.2]$, as opposed to five discrete actions. We apply the actor-critic method to find a continuous pricing policy and compare it with the discrete solution obtained by Q-learning. For both methods, we generate a sufficiently large number

of episodes so to obtain stable models. Then we compare performance of the derived stable models. For each model, we apply the resultant stable policy to 30 newly sampled days (episodes) and then compare their performance in terms of daily sales. For a better visualization, we normalize the results by dividing them by the first day sale of Q-learning model. The result is depicted in Figure 4. The continuous pricing scheme (i.e., actor-critic) performs better than the discrete pricing scheme (i.e., Q-learning), although it does not hold for all days. But one limitation of actor-critic, which is not shown in Figure 4, is that in our experiment, the actor-critic method requires a lot more episodes to obtain a relatively stable policy.

VI. Discussion and Future Work

In this paper, we study the dynamic pricing problem encountered by the electric vehicle fast charging stations from the business owner perspective. Our research sheds light on the economic incentive and profit maximization for expansion of fast charging infrastructure. We model the problem as the Markov decision problem and present three different algorithms that account for different application scenarios. Our experimental results show the advantage of dynamic pricing over stationary pricing. They also provide some managerial insights. For example, the outperformance of the dynamic pricing over the stationary pricing scheme is contingent on the backlog rate. There are still many research problems that can be extended from this study. Besides discrete and continuous pricing, there are other pricing schemes, such as membership and promotional discount. It would be interesting to incorporate other pricing schemes into the Markov decision process model. In the experimental study, we did not include other features as explained in the paper while modeling the states. It would be more meaningful and practical to incorporate the environmental features (e.g., competitor, weather, and news), and test the trained model in the real environment.

References

[1] Y. Zhou, M. Wang, H. Hao, L. Johnson, H. Wang, and H. Hao, "Plug-in electric vehicle market penetration and incentives: A global review," *Mitigation Adaptation Strategies Global Change*, vol. 20, no. 5, pp. 777–795, Jun. 2015.

[2] J. Neubauer and E. Wood, "The impact of range anxiety and home, workplace, and public charging infrastructure on simulated battery electric vehicle lifetime utility," *J. Power Sources*, vol. 257, pp. 12–20, Jul. 2014.

[3] H. Lee and A. Clark, "Charging the future: Challenges and opportunities for electric vehicle adoption," Tech. Rep., 2018.

[4] H. Farhangi, "The path of the smart grid," *IEEE Power Energy Mag.*, vol. 8, no. 1, pp. 18–28, Jan./Feb. 2010.

[5] S. Blumsack and A. Fernandez, "Ready or not, here comes the smart grid!" *Energy*, vol. 37, no. 1, pp. 61–68, Jan. 2012.

[6] R. Adib *et al.*, "Renewables 2015 global status report," REN21 Secretariat, Paris, France, Tech. Rep., 2015, p. 84, vol. 83.

[7] E. Kabir, P. Kumar, S. Kumar, A. A. Adelodun, and K.-H. Kim, "Solar energy: Potential and future prospects," *Renew. Sustain. Energy Rev.*, vol. 82, pp. 894–900, Feb. 2018.

[8] C. Wan, J. Zhao, Y. Song, Z. Xu, J. Lin, and Z. Hu, "Photovoltaic and solar power forecasting for smart grid energy management," *CSEE J. Power Energy Syst.*, vol. 1, no. 4, pp. 38–46, Dec. 2015.

[9] P. Siano, "Demand response and smart grids—A survey," *Renew. Sustain. Energy Rev.*, vol. 30, pp. 461–478, Feb. 2014.

[10] K. Valogianni and W. Ketter, "Effective demand response for smart grids: Evidence from a real-world pilot," *Decis. Support Syst.*, vol. 91, pp. 48–66, Nov. 2016.

[11] F. Papier, "Managing electricity peak loads in make-to-stock manufacturing lines," *Prod. Oper. Manage.*, vol. 25, no. 10, pp. 1709–1726, Oct. 2016.

[12] K. Spees, "Impacts of responsive load in PJM: Load shifting and real time pricing," in *Proc. Develop. Delivering Affordable Energy 21st Century, 27th USAEE/IAEE North Amer. Conf.* Cleveland, OH, USA: International Association for Energy Economics, Sep. 2007.

[13] P. C. Reiss and M. W. White, "Household electricity demand, revisited," *Rev. Econ. Stud.*, vol. 72, no. 3, pp. 853–883, Jul. 2005.

[14] K. King and P. Shatrawka, "Customer response to real-time pricing in great Britain," in *Proc. ACEEE 1994 Summer Study on Energy Efficiency in Buildings*. 1994, pp. 194–203.

[15] C. Luo, Y.-F. Huang, and V. Gupta, "Stochastic dynamic pricing for EV charging stations with renewable integration and energy storage," *IEEE Trans. Smart Grid*, vol. 9, no. 2, pp. 1494–1505, Mar. 2018.

[16] V. Gomez, M. Chertkov, S. Backhaus, and H. J. Kappen, "Learning price-elasticity of smart consumers in power distribution systems," in *Proc. IEEE 3rd Int. Conf. Smart Grid Commun. (SmartGridComm)*, Nov. 2012, pp. 647–652.

[17] M. Biviji, C. Uçkun, G. Bassett, J. Wang, and D. Ton, "Patterns of electric vehicle charging with time of use rates: Case studies in California and Portland," in *Proc. ISGT*, Feb. 2014, pp. 1–5.

[18] A. R. Khan, A. Mahmood, A. Safdar, Z. A. Khan, and N. A. Khan, "Load forecasting, dynamic pricing and DSM in smart grid: A review," *Renew. Sustain. Energy Rev.*, vol. 54, pp. 1311–1322, Feb. 2016.

[19] Z. Hu, K. Zhan, H. Zhang, and Y. Song, "Pricing mechanisms design for guiding electric vehicle charging to fill load valley," *Appl. Energy*, vol. 178, pp. 155–163, Sep. 2016.

[20] Y. Zhang, P. You, and L. Cai, "Optimal charging scheduling by pricing for EV charging station with dual charging modes," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 9, pp. 3386–3396, Sep. 2019.

[21] X. Wang, C. Yuen, N. U. Hassan, N. An, and W. Wu, "Electric vehicle charging station placement for urban public bus systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 1, pp. 128–139, Jan. 2017.

[22] D. Schurmann, J. Timpner, and L. Wolf, "Cooperative charging in residential areas," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 4, pp. 834–846, Apr. 2017.

[23] S. Huang, L. He, Y. Gu, K. Wood, and S. Benjaafar, "Design of a mobile charging service for electric vehicles in an urban environment," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 787–798, Apr. 2014.

[24] C. Yang, W. Lou, J. Yao, and S. Xie, "On charging scheduling optimization for a wirelessly charged electric bus system," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 6, pp. 1814–1826, Jun. 2018.

[25] T. Zhang, X. Chen, Z. Yu, X. Zhu, and D. Shi, "A Monte Carlo simulation approach to evaluate service capacities of EV charging and battery swapping stations," *IEEE Trans. Ind. Informat.*, vol. 14, no. 9, pp. 3914–3923, Sep. 2018.

[26] X. Chen, S.-C. Wong, and C. K. Tse, "Adding randomness to modeling Internet TCP-RED systems with interactive gateways," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 57, no. 4, pp. 300–304, Apr. 2010.

[27] B.-G. Kim, Y. Zhang, M. van der Schaar, and J.-W. Lee, "Dynamic pricing and energy consumption scheduling with reinforcement learning," *IEEE Trans. Smart Grid*, vol. 7, no. 5, pp. 2187–2198, Sep. 2016.

[28] T. Liu, X. Hu, S. E. Li, and D. Cao, "Reinforcement learning optimized look-ahead energy management of a parallel hybrid electric vehicle," *IEEE/ASME Trans. Mechatronics*, vol. 22, no. 4, pp. 1497–1507, Aug. 2017.

[29] R. Xiong, J. Cao, and Q. Yu, "Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle," *Appl. Energy*, vol. 211, pp. 538–548, Feb. 2018.

[30] D. Silver *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, Jan. 2016.

[31] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Machine Learning Proceedings 1994*. Amsterdam, The Netherlands: Elsevier, 1994, pp. 157–163.

[32] J. J. Choi, D. Laibson, B. C. Madrian, and A. Metrick, "Reinforcement learning and savings behavior," *J. Finance*, vol. 64, no. 6, pp. 2515–2534, Dec. 2009.

[33] J. O. Kephart, J. E. Hanson, and A. R. Greenwald, "Dynamic pricing by software agents," *Comput. Netw.*, vol. 32, no. 6, pp. 731–752, May 2000.

[34] M. Schwind and O. Wendt, "Dynamic pricing of information products based on reinforcement learning: A yield-management approach," in *Proc. Annu. Conf. Artif. Intell.* Berlin, Germany: Springer, 2002, pp. 51–66.
[35] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[36] T. P. Lillicrap *et al.*, "Continuous control with deep reinforcement learning," 2015, *arXiv:1509.02971*. [Online]. Available: http://arxiv.org/abs/1509.02971

Cheng Fang received the B.S. and M.S. degrees from Xi'an Jiaotong University and the Ph.D. degree from the Shanghai University of Finance and Economics. He is currently an Assistant Professor with the School of Data Sciences, Zhejiang University of Finance and Economics.

Haibing Lu received the B.S. and M.S. degrees in mathematics from Xi'an Jiaotong University, China, in 1998 and 2002, respectively, and the Ph.D. degree from Rutgers University in 2011. He is currently an Associate Professor with the Department of Information Systems and Analytics, Leavey School of Business, Santa Clara University. His research is at the confluence of privacy/security, data analytics, and information systems.

Yuan Hong received the Ph.D. degree in information technology from Rutgers, The State University of New Jersey. He is currently an Assistant Professor with the Department of Computer Science, Illinois Institute of Technology. His research interests primarily lie at the intersection of privacy, security, optimization, and data mining. He has coauthored more than 30 refereed publications in the above areas and his research was supported by the National Science Foundation.

Shan Liu received the Ph.D. degree in management science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2009. He is currently a Professor of information systems and e-commerce and the Associate Dean of the School of Management, Xi'an Jiaotong University, Xi'an, China. He has authored or coauthored more than 50 refereed articles, including articles that have appeared or accepted in the *Journal of Operations Management*, the IEEE TRANSACTIONS ON ENGINEERING MANAGEMENT, *Information Systems Journal*, the *European Journal of Information Systems*, the *European Journal of Operational Research*, and *Information and Management*. His research interests include IT project management, E-commerce, and data analytics.

Jasmine Chang received the M.A. degree in international economics, the M.B.A. degree in finance and supply chain management, and the Ph.D. degree in supply chain management from the Rutgers Business School. She is currently an Assistant Professor of business data science and finance with the Martin Tuchman School of Management, NJIT. Her research interests include technological applications for supply chain and finance, business sustainability, text analytics, and healthcare information technology. Additionally, she is certified with CPIM (Certified in *Production and Inventory Management*) and SAP, and has consulting experience for ERP Transportation and Logistics Management System development.