Socially Optimal Charging Strategies for Electric Vehicles

Elena Yudovina and George Michailidis

Abstract—Electric vehicles represent a promising technology for reducing emissions and dependence on fossil fuels, and have started entering different automotive markets. In order to bolster their adoption by consumers and enhance their penetration rate, a charging station infrastructure needs to be deployed. This paper studies decentralized policies that assign electric vehicles to a network of charging stations with the goal to achieve little to no queueing. This objective is especially important for electric vehicles, whose charging times are fairly long. The social optimality of the proposed policies is established when each station is equipped with multiple charging slots. Further, convergence issues of the algorithm that achieves the optimal policy are examined. The results provide insight into the optimal location deployment of the charging infrastructure.

Index Terms—Electric vehicles, queueing, scheduling algorithm

I. INTRODUCTION AND MODEL

There has been an increasing penetration of Plug-in Hybrid and pure Electric Vehicles (PHEV/EV) over the least few years [1], due to developments in battery technology that have dramatically increased their range [2], advances in charging technologies that have reduced charging times [3], incentives that have lowered the acquisition and operation costs, and an overall desire to lower emissions [1]. On the other hand, EVs represent a potential source of disruption to normal grid operations if not integrated carefully because they need to connect to the distribution network to charge [4], [5]. The impact of electric vehicles on power grid operations will heavily depend on their market penetration [3]. Estimates vary widely, ranging from 3 to 18 million vehicles by 2025 and from 5 all the way to 40 million vehicles (approximately 20% of the total US market) by 2030 [1]. The disruptive impact is mainly due to the energy load they represent. The peak load of an average household without an EV is about 5 kW, so an EV using Level-1 (1.4 kW) charging technology is 30% of that load, and an EV using Level-2 (7.2 kW) technology is the equivalent of about 1.5 peak household loads [1], [3]. The injection of such large loads coupled with the possible uneven geographic distribution of EVs would strain the local distribution network [6], and possibly the entire grid in case the latter heavily relies on renewable generation resources [5].

Given the current predominance of PHEV vehicles, the literature has largely focused on scheduling at home overnight charging (see [7], [8], [9] and references therein). The proposed approaches treat the induced load as an aggregate and discuss different mechanisms for shifting it during night hours to take advantage of the underutilized electricity generation assets. However, with increasing penetration rates, it will become necessary to recharge EVs during day time [6], [10]. Charging EVs is a rather slow process, as even fast charging takes at least half an hour [1], thus requiring careful scheduling policies to provide the necessary quality of service to customers. Faster charging technologies (e.g. DC charging) could mitigate some of these effects, but as mentioned above, electric utilities have concerns about possible negative impacts of such technologies on the power grid, if deployed at large scale. The work to date on charging stations has mostly focused on modeling and optimizing the architecture of a single charging station (see [10] and references therein).

In this paper, we consider decentralized routing policies that assign EVs to a network of charging stations. Due to lengthy charging times, our primary focus is to ensure little to no customer queueing. Our results apply to the “many-server” regime, when each charging station is equipped with multiple charging slots (e.g. a commercial parking lot, a strip of street parking in a densly urban area), possibly in a mix of charging technologies (e.g. Level-2 and -3 charging, coupled with very fast DC charging infrastructure). We achieve this goal by introducing additional costs per unit time at the charging stations, which the drivers incorporate into their decision-making.

A. Sketch of operations

We provide next a sketch of the way the charging system will operate. Consider the EVs and charging stations within a specific geographical region, for example an urban area or a section of the highway system. Figure 1 shows a (not-to-scale) example. Note that different stations may be spatially far apart, so that not all of the charging slots are accessible to all cars.

We will partition the EVs into a finite set of types \( i = 1, \ldots, I \) that encode their battery technology and charging station preferences. A vehicle that requires charging broadcasts a signal indicating its type, and receives responses from the charging stations within a prespecified distance. The charging stations are numbered \( j = 1, \ldots, J \). Each charging station may have multiple charging slots of the same technology; we are interested in the regime where there are many charging slots \( N_j \gg 1 \) of each type. The actual number of slots \( N_j \) would depend on the type of charging station under consideration,
ranging from a dozen or so for a neighborhood one in a dense urban area, to hundreds for a parking lot type. Thus, the network has many identical EVs, and many identical charging slots. We model the EVs’ preferences among the different charging stations through costs $c_i(j)$. We expect the cost to be lower if the station is along the direction the EV is headed to, higher if the station has slower technology, and infinite if the station cannot be reached with the remaining battery power. The stations respond with a price signal. The EV chooses a charging station based on its own preferences and the responses, and immediately proceeds there. We are interested in designing routing and pricing strategies that are socially optimal while not incurring queueing. The resulting policies will redirect drivers only when it is necessary for congestion control. Moreover, the formulated algorithm will be local in nature, involving only communication between nearby EVs and stations (cf. Figure 1, where A did not exchange any information with stations 4 and 5).

B. Discussion of approach

The social optimization approach to studying routing and congestion control has a rich history [11]. While it represents a simplification of the complex market mechanisms that govern driver–station interactions, it nonetheless provides fundamental insights into the problem. The solution of our proposed model exhibits desirable features for both drivers and charging station operators, as will be discussed in Section II. Although the price signals we find to be optimal are quite natural, it would be interesting to analyze the problem with charging stations as economic agents. Such analysis is beyond the scope of this paper, and has not been addressed in the literature. Rather, the problem addressed is a game between EVs and stations (cf. Figure 1, where A did not exchange any information with stations 4 and 5).

The optimization problem has similarities to the inventory and facility location problem: we are interested in designing routing and pricing strategies that are socially optimal while not incurring queueing. The resulting policies will redirect drivers only when it is necessary for congestion control. Moreover, the formulated algorithm will be local in nature, involving only communication between nearby EVs and stations (cf. Figure 1, where A did not exchange any information with stations 4 and 5).

The objective function corresponds to the rate at which costs are incurred. The constraints are for all arriving requests to be satisfied. We start by formulating a linear program, subsequently converted to an algorithm for assigning EVs to charging stations. Let $\lambda_i$ be the arrival rate of vehicles of type $i$, $\mu_{ij}$ the service rate of a vehicle of type $i$ at a station of type $j$, and suppose EVs of type $i$ are routed to station $j$ at rate $\lambda_{ij}$. (We will discuss stochastic process assumptions in Section III-A.) We consider the following basic linear program.

\[
\begin{align}
\text{minimize} & \quad \sum_{i,j} \lambda_{ij} c_i(j) \\
\text{s.t.} & \quad \lambda_i = \sum_j \lambda_{ij}, \forall i; \quad N_j = \sum_j \lambda_{ij} \mu_{ij}^{-1}, \forall j \\
& \quad \lambda_{ij} \geq 0, \forall i, j.
\end{align}
\]

The objective function corresponds to the rate at which costs are incurred. The constraints are for all arriving requests to be satisfied. We start by formulating a linear program, subsequently converted to an algorithm for assigning EVs to charging stations. Let $\lambda_i$ be the arrival rate of vehicles of type $i$, $\mu_{ij}$ the service rate of a vehicle of type $i$ at a station of type $j$, and suppose EVs of type $i$ are routed to station $j$ at rate $\lambda_{ij}$. (We will discuss stochastic process assumptions in Section III-A.) We consider the following basic linear program.

\[
\begin{align}
\text{minimize} & \quad \sum_{i,j} \lambda_{ij} c_i(j) \\
\text{s.t.} & \quad \lambda_i = \sum_j \lambda_{ij}, \forall i; \quad N_j = \sum_j \lambda_{ij} \mu_{ij}^{-1}, \forall j \\
& \quad \lambda_{ij} \geq 0, \forall i, j.
\end{align}
\]

The objective function corresponds to the rate at which costs are incurred. The constraints are for all arriving requests to be satisfied. We start by formulating a linear program, subsequently converted to an algorithm for assigning EVs to charging stations. Let $\lambda_i$ be the arrival rate of vehicles of type $i$, $\mu_{ij}$ the service rate of a vehicle of type $i$ at a station of type $j$, and suppose EVs of type $i$ are routed to station $j$ at rate $\lambda_{ij}$. (We will discuss stochastic process assumptions in Section III-A.) We consider the following basic linear program.

\[
\begin{align}
\text{minimize} & \quad \sum_{i,j} \lambda_{ij} c_i(j) \\
\text{s.t.} & \quad \lambda_i = \sum_j \lambda_{ij}, \forall i; \quad N_j = \sum_j \lambda_{ij} \mu_{ij}^{-1}, \forall j \\
& \quad \lambda_{ij} \geq 0, \forall i, j.
\end{align}
\]

The objective function corresponds to the rate at which costs are incurred. The constraints are for all arriving requests to be satisfied. We start by formulating a linear program, subsequently converted to an algorithm for assigning EVs to charging stations. Let $\lambda_i$ be the arrival rate of vehicles of type $i$, $\mu_{ij}$ the service rate of a vehicle of type $i$ at a station of type $j$, and suppose EVs of type $i$ are routed to station $j$ at rate $\lambda_{ij}$. (We will discuss stochastic process assumptions in Section III-A.) We consider the following basic linear program.

\[
\begin{align}
\text{minimize} & \quad \sum_{i,j} \lambda_{ij} c_i(j) \\
\text{s.t.} & \quad \lambda_i = \sum_j \lambda_{ij}, \forall i; \quad N_j = \sum_j \lambda_{ij} \mu_{ij}^{-1}, \forall j \\
& \quad \lambda_{ij} \geq 0, \forall i, j.
\end{align}
\]

The objective function corresponds to the rate at which costs are incurred. The constraints are for all arriving requests to be satisfied. We start by formulating a linear program, subsequently converted to an algorithm for assigning EVs to charging stations. Let $\lambda_i$ be the arrival rate of vehicles of type $i$, $\mu_{ij}$ the service rate of a vehicle of type $i$ at a station of type $j$, and suppose EVs of type $i$ are routed to station $j$ at rate $\lambda_{ij}$. (We will discuss stochastic process assumptions in Section III-A.) We consider the following basic linear program.

\[
\begin{align}
\text{minimize} & \quad \sum_{i,j} \lambda_{ij} c_i(j) \\
\text{s.t.} & \quad \lambda_i = \sum_j \lambda_{ij}, \forall i; \quad N_j = \sum_j \lambda_{ij} \mu_{ij}^{-1}, \forall j \\
& \quad \lambda_{ij} \geq 0, \forall i, j.
\end{align}
\]
of charging; the corresponding virtual queue is incremented: $Q_{j*} \mapsto Q_{j*} + (\mu_{ij*})^{-1}$.

3. All virtual queues that are positive decrease at rate $N_j$ per time unit. Once a virtual queue reaches 0, it stays at level 0 until some EV is routed to it.

The algorithm runs in continuous time and no synchronization is necessary between different stations. Further, note the natural form of the routing decision: the EV driver is asked to add to her intrinsic costs $c_i(j)$ a certain cost per unit time $\beta Q_j(t)$. This cost will be greater for the stations that are in high demand, and lower for the stations that are less congested, which is natural from the point of view of the stations as well.

The key feature of the algorithm is the use of virtual queues to regulate the routing decisions (through structured price changes). The virtual queues play the role of congestion indicators, but without any actual queueing going on. Moreover, because virtual queue lengths will change with changing arrival patterns, the algorithm can automatically respond to changing arrival patterns. We illustrate the effects of a change in the arrival pattern in the simulations in Section IV.

However, the GPD algorithm has several disadvantages. While it adjusts to changes in the arrival pattern, it does so quite slowly. Moreover, because the signals sent to the EVs do not take into account the real-time occupancy level of the charging station, a driver may find herself being routed to a charging station with no free slots. While in general it is difficult to give a good trade-off between user preferences and potential queueing times, for the special case $c_i(j) \in (0, \infty)$ we present the Freest Charger Shortest Queue (FCSQ) algorithm. FCSQ is faster at adjusting to small changes in the arrival rate pattern and better at balancing the loads on different charger stations; further, the cost structure $c_i(j) \in (0, \infty)$ can be used for capacity planning.

We begin by introducing some additional notation. Assume that the solution of (1) is unique (it almost always is, see Section III). The pairs $(ij)$ for which $\lambda_{ij} > 0$ are called the basic activities. The FCSQ algorithm assumes that we know the set of basic activities; this set can be determined either by solving (1) statically, or by running the GPD algorithm in the background with a sufficiently small value of $\beta$. Because it is a discrete quantity, the set of basic activities will be unchanged if the arrival rates change slightly, which means that we can determine the correct set of basic activities even from imprecise measurements of the arrival rates.

FCSQ Algorithm:
1. Identify the set of basic activities (e.g. by running GPD in the background). Set $c_i(j) = \infty$ if $(ij)$ is not basic.
2. When an EV of type $i$ requests service, if some charger with $c_i(j) < \infty$ is available, route the vehicle to the charging station with the largest fraction of free chargers.
3. If no charger is available at the charging station, the charging station sends back an estimate of the time when the EV will begin charging. The EV joins the charging station where this time is earliest.

The second step relies on the good will of the drivers for implementation, but we could encourage this behavior using a price signal that depends on the percentage of occupied chargers at each station. In the third step, any reasonable estimate should work; for example, in Section IV, we add up the (future) charging times of all the EVs in the queue and divide by the number of chargers at the station.

The two main advantages of FCSQ over GPD are: (1) FCSQ is faster at reaching the new steady-state after a change in the arrival pattern; (2) FCSQ achieves load-balancing. This means that when there is queueing, the waiting times are approximately equal at all stations, which encourages compliance from the drivers and leads to faster dissipation of queues; it also means that extra capacity in the network can be “shared” by all users. A fairly minor disadvantage is that the basic activity tree needs to be computed; however, this can be accomplished as previously mentioned by running GPD in the background. The main disadvantage, however, is that FCSQ does not take into account user preferences beyond asking whether $c_i(j)$ is finite. However, an intermediate load-balancing algorithm can be formulated; see the extended version of this paper [14].

III. RESULTS

Next, we state our main results. We begin by introducing the main assumptions on the structure of the system. The stochastic process assumptions are standard in the queueing literature. They are satisfied by many common models, including renewal processes with finite second moment; Assumption 3 holds for almost all systems.

A. Assumptions

We begin by defining the underlying stochastic processes, indexed by $t \in \mathbb{R}$, $t \geq 0$.

$A_i(t) = \#\{\text{EVs of type } i \text{ arriving up to time } t\}$;

$A_{ij}(t) = \#\{\text{EVs of type } i \text{ that have been routed to station } j \text{ up to time } t\}$;

$D_{ij}(t) = \#\{\text{EVs of type } i \text{ that have completed service at station } j \text{ by time } t\}$.

While $A_i(t)$ is intrinsic to the system, the departures process $D_i(t)$ depends on the routing choices. The process $S_{ij}$ below, on the other hand, depends only on the sequence of service times of the EVs of type $i$ at station $j$: $S_{ij}(t) = \#\{\text{EVs of type } i \text{ whose total service requirement at a slot of type } j \text{ is } t\}$. To see more clearly the meaning of $S_{ij}()$, let $\Psi_{ij}(t) = \#\{\text{EVs of type } i \text{ in service at station } j \text{ at time } t\}$. Then

$D_{ij}(t) = S_{ij}(\int_0^t \Psi_{ij}(s)ds)$.

The employed notation $\Rightarrow$ indicates weak convergence with respect to the topology of uniform convergence on compact sets (u.o.c). That is, if the indexing variable is $t$, then the stated convergence in distribution holds uniformly for all values of $t \in [0, T]$, but the constants may depend on $T$.

We now formulate the stochastic process assumptions that should apply to $A_i()$ and $S_{ij}()$.

Assumption 1: $A_i()$ and $S_{ij}()$ satisfy functional law of large numbers approximations as follows:

$r^{-1} A_i(rt) \Rightarrow \lambda_i t$, $r^{-1} S_{ij}(rt) \Rightarrow \mu_{ij} t$, as $r \to \infty$.

Moreover, $\mathbb{E}[(A_i(t+1) - A_i(t))^2] < \infty$ and $\mathbb{E}[(S_{ij}(t+1) - S_{ij}(t))^2] < \infty$ uniformly in $t$. 

Assumption 1 is sufficient for results concerning stability of systems. For more fine-grained description, we require more detailed assumptions on the deviations of the processes from their means. These assumptions also hold for renewal processes with finite second moment.

Assumption 2: $A_i(\cdot)$ and $S_{ij}(\cdot)$ satisfy a functional central limit theorem as follows:

$$r^{-1/2}(A_i(rt) - \lambda_i rt) \xrightarrow{d} W_i(t),$$
$$r^{-1/2}(S_{ij}(rt) - \mu_{ij} rt) \xrightarrow{d} W_{ij}(t), \text{ as } r \to \infty,$$

where $W_i$ and $W_{ij}$ are independent Brownian motions with some finite variances (not necessarily all the same).

In addition to the stochastic process assumptions above, we will require a technical assumption on the nature of the solution to the linear program (1).

Assumption 3: The optimal solution $(\lambda^*_i)$ to the linear program (1) is unique. The optimal set of dual variables $(q^*_j)$ corresponding to the capacity constraints is unique. The number of basic activities is $I + J - 1$.

This assumption is commonly made in the queueing literature, and the set of parameters $\lambda_i$, $\mu_{ij}$, and $c_i(j)$ for which Assumption 3 fails have measure 0 in $\mathbb{R}^{I+2IJ}$ (see [15] for a proof).

In other words, if a given set of parameters does not satisfy Assumption 3, some small perturbation of the parameters will cause the assumption to hold. (One setting in which the assumption typically does not hold is when $c_i(j) \in \{0, \infty\}$; however, as this is primarily a technical assumption, we do not expect the performance of FCSQ to be affected by its failure to hold.)

B. Statement of results

Our main result concerns the behavior of the GPD algorithm when $N_j \to \infty$ (the system has many charging slots) and $\beta \to 0$ simultaneously. Specifically, we consider a sequence of systems indexed by $r$, with charging station sizes $N_j = rN_j + O(1)$ (the $O(1)$ is to ensure $N_j$ is an integer), arrival rates $\lambda^*_j = r\lambda_j$, and constant charging rates $\mu^*_ij = \mu_{ij}$.1 While the results we derive are only valid in the asymptotic limit $r \to \infty$ (i.e. $N_j \to \infty$), the results give insight into the performance of finite systems, which would otherwise be very difficult to analyze precisely. In other contexts in many-server literature (e.g. [16]), it has been found that the asymptotics are in good agreement with reality for values of $r$ as small as $r = 10$.

Let $\lambda^*_j$ and $q^*_j$ be the unique optimal solution to (1) with parameters $\lambda_j$, $\mu_{ij}$, and $N_j$. In the $r$th system above, we would like the actual routing decisions to give $A^*_ij(t) = r\lambda^*_ij t$. We show that this is possible given the right choice of $\beta$ for the GPD algorithm.

The first (supplementary) result shows that in steady-state, the virtual queues will be large and proportional to the optimal dual variables $q^*_j$, namely $Q^*_j \approx (\beta^2)^{-1}q^*_j$.

Theorem 1: Consider the sequence of systems indexed by $r$ as above, where system $r$ runs the GPD algorithm

with $\beta^r = r^{-3/4}$. Suppose the arrival and service completion processes are Poisson, and the parameters $\mu_{ij}$ are rational, so that the virtual queueing system can be modeled by a countable state-space Markov process. For each $r$, consider the associated stationary version of the process, and let $Q^*_j(t)$ be the stationary version of $Q^*_j(t)$. Then $r^{-1/2}(Q^*_j(t) - (\beta^r)^{-1}q^*_j) \to 0\forall j$.

Remark: While this theorem is stated only for Markovian processes, we expect the result that $Q^*_j(t) \approx (\beta^r)^{-1}q^*_j$ for large $r$ and $t$ to hold more broadly. The countable state-space Markov process formulation is used to avoid technicalities in dealing with stationary versions of processes.

The second result shows that after the GPD algorithm has converged, it results in a good routing pattern. Specifically, suppose the system is critically loaded, i.e. $\sum_{j} \lambda^*_j \mu^{-1}_{ij} = N_j$ for all $j$ in (1). We show that $A^*_ij(t) \approx r\lambda^*_ij(t) + \sqrt{7}(\text{Brownian motion})$.

Theorem 2: Consider the sequence of systems indexed by $r$ as above, and suppose that $r^{-1/2}(Q^*_j(t) - (\beta^r)^{-1}q^*_j) \to 0\forall j$. Suppose further that the system is critically loaded. Then

$$r^{-1/2}(A^*_ij(\cdot) - r\lambda^*_ij(\cdot)) \xrightarrow{d} H(W), \text{ as } r \to \infty,$$

where $W$ is the Brownian motion identified in Assumption 2, and $H : \mathbb{R}^t \to \mathbb{R}^{I+2IJ}$ is a linear mapping defined as follows. For a vector $v = (v_1, \ldots, v_I) \in \mathbb{R}^I$, the image $w = H(v)$ with coordinates indexed by basic activities $(ij)$ satisfies

$$\sum_j w_{ij} = v_i, \forall i;$$
$$\mu^{-1}_{ij} \mu^{-1}_{ij} w_{ij} = \mu^{-1}_{ij} \sum_{ij'} \mu^{-1}_{ij} w_{ij}, \forall (ij), (ij').$$

Remark on $\beta$: In the above two theorems, we could choose $\beta = f(r)$ for any function $f(r)$ satisfying $rf(r) \to \infty$ and $r^{1/2}f(r) \to 0$. The value of $\beta$ controls the rate of change of $Q^*_j(t)$; because the size of the fluctuations is $O(\sqrt{r})$, larger values of $\beta$ correspond to systems that react faster to changes. However, large values of $\beta$ mean greater relative size of fluctuations of $Q^*_j$ around its equilibrium level, so a larger value of $r$ will be required before asymptotics become valid.

For finite values of $N_j$ and $\beta$, if the system has sufficient capacity to process all the EVs, then the GPD algorithm will keep the system stochastically stable. Specifically, if $N_j > \sum_j (\mu_{ij})^{-1}q^*_j$ for some feasible solution of (1), and $r^{-1/2}A^*_ij(rt) \to 0$. The moment we have infinitely many charging slots available at station $j$. Then the number of charging slots occupied by vehicles of type $i$ can be approximated by an Ornstein-Uhlenbeck process centered on $r^{1/2} \lambda^*_ij$. The steady-state deviation from $r^{1/2} \lambda^*_ij$ will be normal, with variance that can be evaluated explicitly in terms of the Brownian motions appearing in Assumption 2. Thus, it is possible to plan $O(r^{1/2})$ excess charging slots for any desired probability of queueing.
Note that if we use the GPD algorithm, we need to plan for excess capacity at every station. The FCSQ algorithm achieves load balancing, which allows excess capacity to be placed anywhere within a connected component of the basic activity graph. While we do not have precise analytic guarantees for FCSQ, these are available for similar algorithms (see [17], [18]). In particular, if $\mu_{ij} = \mu_j$ (charging time depends only on charger technology), [18] shows for a very similar algorithm that the steady-state deviations in the routing pattern for large $r$ are of size $O(r^{1/2})$, and thus putting $O(r^{1/2})$ extra charging slots anywhere in the network suffices to provide guarantees on the probability of queueing. (Typically different charging stations will come into the “pool” of excess capacity with different multipliers.) By the results in [15], steady-state queues cannot be smaller than $O(r^{1/2})$; the small queue size resulting form FCSQ is born out in simulations.

IV. SIMULATIONS AND DISCUSSION

We use a toy network to compare the performance of GPD and FCSQ algorithms. There are two EV classes, three charger classes, and each EV can use two of the charger classes. The service rates are $\mu = \begin{bmatrix} 3 & 3 & 0 \\ 0 & 1 & 2 \end{bmatrix}$. We set $N_j = 20$ for all $j$ and simulate 10,000 EV arrivals. The first 5,000 arrivals are generated using arrival rates $\lambda = (2.5, 2.2) \times 20$; for the second 5,000 arrivals, we use $\lambda = (2.2, 2.5) \times 20$ to illustrate the effect of a change in the arrival pattern. (The basic activities are the same throughout the simulation.) The system is quite heavily loaded: for the first arrival pattern, we need 91% of all the charging slots, while for the second we need 97%.

Fig. 2 depicts the delays observed when running the GPD algorithm and the FCSQ algorithm on the same pattern of arrival events. Note the effect of the change in the arrival pattern, which occurs at the vertical line. Using FCSQ reduces the largest delay significantly; there are long stretches of time with no queueing anywhere in the system; and whenever there is queueing, it is nearly the same at all stations. The FCSQ algorithm also reaches the new steady-state much faster.

For EV traffic, we expect significant diurnal fluctuation in the arrival pattern. While GPD can mitigate this to some extent, performance may be improved if the parameters $Q^*_j$ are reset “by hand” several times during the day to values that are close to optimal. GPD algorithm will then “take up the slack” in approaching the optimum.

We recall again that under FCSQ, the probability of delay depends on the aggregate excess capacity across the entire connected component of basic activities tree. This consideration allows excess capacity to be placed strategically, rather than having to allocate excess capacity at each charging station. The effect is greatest if the connected components of the basic activity tree are large, which is a function of the arrangement of the charging stations. It would therefore be interesting to determine what arrangement of the charging stations results in the most connected basic activity graph and is therefore most flexible.

An important extension of this work would be to incorporate the effect of placing large batteries [10] at charging stations to smooth the peak demand produced by charging many vehicles at once. This introduces an additional algorithmic challenge by introducing demand mobility in time in addition to in space.

V. ACKNOWLEDGMENTS

The authors would like to thank the referees for helpful suggestions in improving the paper.

APPENDIX

While the GPD algorithm employed is a special case of the algorithm described in [13], there are certain technical differences between the settings. Consequently, the analytic guarantees presented in Theorems 1 and 2 are not automatic. The derivation of the stationary result of Theorem 1 is very similar to the proof of [15, Theorem 6.3, 6.4]. The overall technique of proof of Theorem 2 follows the proof of [15, Theorem 6.3], but there are substantial deviations, and for completeness we give a sketch of the full argument.

Sketch of proof of Theorem 2: Fix a time interval $[0, T]$. First, we show that only the basic activities are used
during the entire time interval. This would be the case if we had \( \beta r Q_j'(t) = q_j^* \). Indeed, complementary slackness for the solutions of (1) implies that \( \lambda_{ij}^* > 0 \) for precisely those pairs \((i,j)\) where \( c_j(i) + \mu_{ij}^{-1} q_j^* \) is minimized. Consequently, our first result will be to show that \( \beta r Q_j'(t) \approx q_j^* \) on all of \([0, T]\).

**Lemma 3:** If \( \beta r Q_j'(0) - q_j^* \Rightarrow 0 \), then \( \beta r Q_j'(t) - q_j^* \Rightarrow 0 \), \( j \in J \), as \( r \to \infty \), uniformly in \( t \in [0, T] \).

This result corresponds to [15, Theorem 6.1], but its statement (and proof) in our case is different, mainly because we have no constraints on the \( \sum_j Q_j \). We postpone the proof in order to proceed with the rest of our argument.

Consider two choices \( ij \) and \( ij' \) for a vehicle of type \( i \), where \( ij \) is basic and \( ij' \) is not. This means that \( c_j(i) + \mu_{ij}^{-1} q_j^* < c_j(i') + \mu_{ij'}^{-1} q_j^* \). Because of Lemma 3, for all large enough \( r \), with high probability \( c_j(i) + \mu_{ij}^{-1} \beta r Q_j'(t) < c_j(i') + \mu_{ij'}^{-1} \beta r Q_j'(t) \) at all times \( t \in [0, T] \), and thus no vehicles of type \( i \) will be routed along the non-basic activity to station \( j' \).

It remains to determine the numbers of EVs routed along each basic activity. By flow conservation, \( A_j'(t) - r \lambda_{ij}^* t \approx r^{1/2} W_j(t) + o(r^{1/2}) \). To show the second part of (2), we present the following lemma.

**Lemma 4:** For \( \beta r Q_j'(0) - (\beta r)^{-1} q_j^* = o(r^{1/2}) \), then 
\[
r^{-1/2} \left( (c_j(i) + \mu_{ij}^{-1} \beta r Q_j'(t)) - (c_j(i') + \mu_{ij'}^{-1} \beta r Q_j'(t)) \right) \Rightarrow 0, \forall (ij), (ij') \text{ basic, as } r \to \infty, \text{ uniformly in } t \in [0, T].
\]

Lemma 4 implies that for all pairs of stations \( j, j' \) which have a common vehicle type \( i \) in the basic activity tree,
\[
\mu_{ij}^{-1} - \mu_{ij'}^{-1} = o(r^{1/2}).
\]

For renormalization, in heavy traffic, this is precisely the second half of (2).

We now sketch the proofs of Lemmas 3 and 4.

**Proof of Lemma 3:** We use local fluid limits. A general introduction to the technique may be found in [19, Section 8]. Select a time \( T_0 \) as described below, and define the local-fluid-scaled processes \( q_j^{(r,m)}(u) = \beta r Q_j'(t^m + \beta r u) - q_j^*, u \in [0, T_0], m = 1, \ldots, \beta r \). Standard techniques imply the following functional law of large numbers result. As \( r \to \infty \), any sequence of \( q_j^{(r,m)}(\cdot) \) has a convergent subsequence, whose limit is a Lipschitz process \( q_j^{(r)}(\cdot) \) satisfying certain differential equations almost everywhere. Some of these equations are
\[
d/dt a_j^{(r)}(t) = \sum_i \lambda_i(t) - N_j(t), \quad d/dt a_j^{(r)}(t) = 0 \quad \text{unless } c_j(i) + \mu_{ij}^{-1} q_j^{(r)} \text{ is smallest over } j'.
\]
Here, \( a_j^{(r)}(t) \) is a local-fluid-scaled quantity responsible for routing choices. We next show
\[
\frac{1}{\beta r} \sup_{t \in [0, T_0]} \max_j q_j^{(r)}(t) \to c_j \quad \text{as } r \to \infty
\]
and the derivative is defined. Consider the set of stations \( J \) achieving \( \max_j q_j^{(r)}(t) \). The arrivals to these stations at time \( t \) are only those vehicle types that have no other basic activity. If \( \max_j q_j^{(r)}(t) > 0 \), Assumption 3 implies that
\[
\sum_j a_j^{(r)}(t) < \sum_j \mu_{ij} \lambda_i(t) - \epsilon = \sum_j N_j(t) - \epsilon \quad \text{for some } \epsilon > 0.
\]
Consequently, if \( a_j^{(r)}(0) > 0 \) for all \( j \), there will be a time \( T_0 \) such that \( q_j^{(r)}(T_0) = 0 \) for all \( j \) and then \( q_j^{(r)} \) necessarily stays at 0. Thus, for a single value of \( m \) and sufficiently large \( r \), the vector \( q_j^{(r,m)}(u) \) is very likely to stay close to 0. To conclude this for all \( m \) simultaneously, we use a diagonalization argument: pick the subsequence of intervals \( m(r) \) along which \( q_j^{(r,m)}(u) \) first crosses level \( \delta < 1 \), then pick a subsequence along which \( q_j^{(r,m)}(0) \) converges, and observe that the corresponding local fluid limit contradicts our previous assertions.

**Sketch of proof of Lemma 4:** This result is also proved using local fluid limits, but using a somewhat different scaling. Pick a time \( T_0 \), and define \( \tilde{q}_j^{(r,m)}(u) = r^{-1/2} \beta r Q_j'(t^m + r^{-1/2} u) - Q_j'(t^m) \), \( u \in [0, T_0), m = 1, \ldots, \sqrt{r} \). Note that we subtract \( Q_j'(t^m) \) rather than \( (\beta r)^{-1} q_j^* \). Arguing similarly to the proof of Lemma 3, we first show that the largest positive difference between quantities of the form \( c_j(i) + \mu_{ij}^{-1} \beta r Q_j' \) must decrease for one \( m \), and then conclude that it will stay near 0 on all of \([0, T]\).