Dynamic Load Balancing of Selfish Drivers between Spatially Distributed Electrical Vehicle Charging Stations

Fernando Paganini and Andres Ferragut
Universidad ORT Uruguay

***Abstract*— This paper considers an electrical vehicle recharging infrastructure made up of physically separate stations serving spatially distributed requests for charge. Arriving EVs receive feedback on transport times to each station, and waiting times at congested stations, based on which they make a selfish selection. We present a fluid model of the resulting dynamics, in particular modeling queueing delays as a function of fluid queues, and two different models of client departures: given sojourn times, or given service times. In each case, the joint load balancing dynamics is related to a convex program, suitable variant of a centralized optimal transport problem. In particular, we show the correspondence between equilibrium points and the corresponding optima, and use Lagrange duality to interpret the convergence properties of the dynamics. The results have similarities and differences with classical work on selfish routing for transportation networks. We present illustrative simulations, which also explore the validity of the model beyond the fluid assumption.**

I. Introduction

In the roadmap towards the generalized use of Electrical Vehicles (EVs) (see e.g., [17]), a key component is the deployment of a suitable charging infrastructure. We are interested in particular on a network of charging stations, which may be distributed at parking lots over a certain geographical area, e.g. a city center. The deployment and operation of these facilities is a topic of active research [8]–[10], [18].

The coverage of a *fixed* demand for charging from the point of view of a central planner was analyzed in [11] with optimization tools, considering the spatially distributed nature of the problem. When station locations are part of the design we are in the realm of facility location problems (e.g., [3], [6]), which have also been applied to other forms of energy refueling (e.g. [7]). Given the stations, the assignment of a distributed demand yields versions of the *optimal transport* problem (cf. [14]).

Charging demand does not, however, materialize in a single batch; rather, we have a *dynamic* situation in which requests for service arise asynchronously in time and in different spatial locations. This traffic must be directed to an adequate charging station. If these routing decisions were in the hands of a central planner, we would have a *load balancing* problem similar in nature to those considered in computer networks (see e.g. [15] and references therein). However, compulsory routing may not be assumed here; rather, drivers will select a station consistent with their own incentives, typically to obtain the fastest possible service. In this regard, the problem has common features with the classical analysis of selfish routing in transportation networks (see, e.g. [2], [13]).

This paper draws from the above extensive background, but addresses some distinguishing features. In contrast with the transportation literature where *all* traffic is subject to selfish routing, and transport latency is modeled as a static function of resulting road flows, here we will assume EVs routing to charging stations to be a small portion of the traffic. Hence, transport delays of EVs to each charging station are given exogenously, but we must worry about congestion delays at the stations themselves, as it is habitual in networking. This mix of transport + queueing has interesting new features from a mathematical perspective. Furthermore, in contrast to other kinds of services, charging could be *partial*, limited by customer time availability, and still retain value; this aspect presents differences with respect to mainstream queueing analysis of load balancing.

In order to capture the aforementioned features, we model demand for EV charging service as arrival rates at different points, distributed across a region. EVs route selfishly to charging stations seeking the shortest time to service, using information on the transport time to each station, as well as congestion signals indicating queueing delay at the stations. The queues resulting from these inflows are modeled as fluid quantities; different departure models are considered: either a *sojourn time* is specified for EVs, independent of the service obtained, or a *service time* dictates departures. For each case, we present a convex optimization problem that characterizes the equilibrium allocation, a suitable variant of the centrally planned optimal transport problem. Dynamic convergence to equilibrium is established by analyzing the time evolution of the respective Lagrangian dual function.

The rest of the paper is organized as follows. In Section II we briefly review the relevant background, and present our general model. In Section III we analyze the dynamics for the sojourn time model of departures; we introduce an appropriate optimization problem and establish its connection with the equilibrium, as well as results on convergence. A similar analysis is carried out in Section IV for the service time model of departures. Simulations that support this theory are provided in Section V. In particular, we explore experimentally the behavior of our control system under discrete stochastic arrivals, showing good fit, and we illustrate the equilibrium behavior. Conclusions are given in Section VI.

This work was partially supported by ANII–Uruguay under grant FCE_1_2021_1_167301, and AFOSR-US, grant FA9550-23-1-0350.

E-mail: paganini@ort.edu.uy.

II. Background and Problem Formulation

We first define some notation. Let $e_j, j = 1, \dots, n$ be the canonical vectors in $\mathbb{R}^n$, and $\Delta_n = \mathrm{co}(\{e_j\})$ be their convex hull, i.e. the unit simplex. Define $\varphi : \mathbb{R}^n \to \mathbb{R}$ by

$$\varphi(x) = \min_j(x_j) = \min_{z \in \Delta_n} \sum_j x_j z_j. \tag{1}$$

This is a concave function. Its superdifferential (set of supergradient vectors) at a given point x is characterized by:

$$\partial\varphi(x) = \arg\min_{z \in \Delta_n} \sum_j x_j z_j = \mathrm{co}\big(\{e_k : k \in \arg\min(x_j)\}\big);$$

these are vectors in the unit simplex, equal to zero in non-minimizing coordinates. At points x where there is a single minimizing index k, $\partial\varphi(x) = \{\nabla\varphi(x)\} = \{e_k\}$.

Next, we will briefly review three independent areas of research relevant to this paper, as mentioned in the introduction. At the end of the section we will formulate our model for spatially distributed load balancing.

A. Optimal transport

The optimal transport problem has ancient roots in the 18th century work of Monge, and a fundamental relaxation by Kantorovich in the 20th; [14] is a recent reference. It concerns the mapping between two mass distributions, origin and destination, in a way that minimizes total transport cost. For our context, we will consider a *discrete* version:

- Demand originates at *locations* $i = 1, \dots m$, with respective quantities $\bar{q}_i$. For our application, these are discrete points in the plane (e.g. city corners) from where EVs request charging service.
- Supply is situated at *stations* $j = 1, \dots, n$, each offering $\bar{s}_j$ units of service, e.g. EV charging spots.
- The matrix c_{ij} specifies the transport cost between i and j; this often represents travel distance. In this paper we will associate it with travel *time*.

With the above definitions, the Kantorovich version of the optimal transport problem is to find the matrix Π of quantities π_{ij} transported between location i and station j, solving:

$$\min \sum_{ij} c_{ij}\pi_{ij}, \tag{2}$$

subject to:

$$\pi_{ij} \geq 0 \quad \forall i, j; \quad \sum_j \pi_{ij} = \bar{q}_i \quad \forall i; \quad \sum_i \pi_{ij} = \bar{s}_j \quad \forall j.$$

For feasibility, it is required that total supply matches total demand, i.e. $\sum_j \bar{s}_j = \sum_i \bar{q}_i$. Less restrictive variants of this optimization were discussed in [11], some of which will be used later in the paper.

It can be shown that if $\bar{s}_j$, $\bar{q}_i$ are integer, the above problem admits an integer solution Π, compatible with routing individual EVs to a single station (cf. [11]). We will not address such issues in this paper, and rather treat EV quantities as fluid (real valued); the effect of this approximation will be validated in simulation.

B. Load balancing between fluid queues

In communication and computing networks, *load balancing* refers to active control carried out by a dispatcher that distributes tasks between servers. The analysis is usually framed in terms of stochastic queues (see e.g. [15]); here we introduce a fluid model. Suppose a single dispatcher receives jobs at a rate r jobs/sec, and assigns each to a server from the set $j = 1, \dots, n$. If s_j is the current queue at location j, then a natural (and under some conditions, optimal [4]) strategy is to route to the shortest queue, with ties broken arbitrarily. To capture this control in a fluid model, the vector $a = (a_j)_{j=1}^n \in \mathbb{R}^n$ of arrival rates to each server must satisfy $a \in r\Delta_n$, with a_j nonzero only at minimizing coordinates of $\varphi(s) = \min(s_j)$. Namely:

$$a(s) \in r \ \arg\min_{z \in \Delta_n} \sum_j s_j z_j = r\partial\varphi(s). \tag{3}$$

Assuming we have specified as well a vector field for departure rates $d(s) = (d_j(s))$, we can write the following dynamics for the fluid queues:

$$\dot{s} = a(s) - d(s) \in r\,\partial\varphi(s) - d(s). \tag{4}$$

Remark 1: Eq. (4) is a *differential inclusion*, with a switching, discontinuous right-hand side. This complicates the rigorous treatment of solutions, requiring e.g., the notion of Filippov solutions (see e.g. [1]).

A natural workaround is to apply *regularization* techniques to approximate the switching dynamics by a smooth gradient flow (cf. [5], [12]). We briefly outline the idea here, let:

$$\varphi_\epsilon(s) := \max_y \left(\varphi(y) - \frac{1}{2\epsilon}\|s - y\|_2^2\right)$$

be the *Moreau envelope* of $\varphi(s)$; this is a smooth, concave function with an ordinary gradient $\nabla\varphi_\epsilon(s)$. It can be shown that this gradient coincides with $\nabla\varphi(s)$ at points where the minimizing coordinate s_k is at least ϵ away from the rest; in regions of approximate parity, $\nabla\varphi_\epsilon(s)$ gives a continuous interpolation to avoid switching. Replacing $\partial\varphi(s)$ by $\nabla\varphi_\epsilon(s)$ in eq. (4) turns it into an ordinary differential equation, which may be studied by simpler tools, and approximates the target behavior to any desired accuracy.

C. Selfish routing

The above load balancing rule assumes a central dispatcher with full routing authority. For our problem we are interested in *selfish* decisions. In his classical paper on road traffic [16], Wardrop studied the effect of these selfish driver choices in the resulting equilibrium flow. We recall briefly this approach, following the notation from [13].

In selfish routing models, each road is an edge e in a graph, carrying a flow f_e; the delay or latency associated with traveling on this road is modeled by an increasing function $l_e(f_e)$. A path P through the network has total latency $l_P(f) = \sum_{e \in P} l_e(f_e)$.

Each source-destination pair or commodity, indexed by i, has an input traffic rate r_i, which may be split among the set of compatible paths $\mathcal{P}_i$. Under these conditions:

- The (socially) optimal flow is the one that minimizes the number of vehicles in travel $C(f) = \sum_e f_e l_e(f_e)$, subject to supporting the input rates.
- The Nash-Wardrop equilibrium (WE) is a flow configuration where each commodity uses only minimal latency paths. The WE exists and is equivalent to minimizing the integrated latency

$$\sum_e \int_0^{f_e} l_e(\sigma) d\sigma. \tag{5}$$

These two notions do not coincide and a *price of anarchy* appears due to selfish decisions on the part of drivers. [13] contains an extensive analysis on bounding this price of anarchy. For *dynamic* studies of selfish routing beyond equilibrium, we refer to [2] and references therein.

D. The spatially distributed EV charging problem

We are now ready to present our spatial EV load balancing model, integrating all the different aspects considered above.

- Demand arises at locations $i = 1, \dots, m$ in the plane, each of which receives a rate r_i of charging requests.
- Charging stations are indexed by $j = 1, \dots, n$. The *transport delay* c_{ij} to reach station j from location i is considered exogenous, determined by distance and traffic conditions; it is available to drivers. For our analysis we will assume it is constant in time.
- Drivers are also informed of the current *queueing delay* $\mu_j(t)$ at station j. They self-route selfishly to the station j with minimum *total time to service* $c_{ij} + \mu_j$. Mathematically, if we denote by $a_{ij}(t)$ the rate of requests sent from location i to station j; the vector $a^i = (a_{ij})_{j=1}^n$ of rates originating from location i satisfies

$$a^i \in r_i \partial \varphi^i(\mu), \tag{6}$$

  which is analogous to (3), introducing the function

$$\varphi^i(\mu) := \min_j (c_{ij} + \mu_j). \tag{7}$$

- The total flow rate arriving into station j is

$$a_j := \sum_i a_{ij}. \tag{8}$$

- The state variable $s_j(t)$ represents the current assignment of station j; it evolves according to the fluid queue

$$\dot{s}_j = a_j - d_j, \quad j = 1, \dots, n, \tag{9}$$

  where d_j is the departure rate from station j. Two specific departure functions $d_j(s)$ will be analyzed, both with $d_j(0) = 0$ so queues in (9) remain non-negative.

Remark 2: Eq. (8) does not distinguish between the time a vehicle arrives into the system and selects a station, and the time it effectively reaches it. Recalling that c_{ij} is the transportation delay, one might write instead $a_j(t) = \sum_i a_{ij}(t - c_{ij})$, which would give us a delay-differential equation, with heterogeneous delays and the consequent increase in complexity. Our simpler model is applicable provided transport delays are much smaller than EV service times.

To close the loop, it remains to model the dependence of queueing delay μ_j with station occupation s_j. If $s_j \leq \bar{s}_j$, there is no queueing. Otherwise, excess EVs $s_j - \bar{s}_j$ will wait for service slots to become available; since these are liberated at the departure rate d_j, a natural fluid model for the waiting time is:

$$\mu_j := \frac{[s_j - \bar{s}_j]^+}{d_j(s_j)}. \tag{10}$$

Our full model for the dynamics is thus given by (9), with arrival rates specified by (6–8), $\mu(s)$ given in (10), and the departure model $d_j(s_j)$.

Regarding these departures, we note that there are two possible reasons for leaving the system: completed service, or the customer running out of available time. The former is the usual assumption in queueing theory. Note, however, that for EV charging there is value for *partial service* (e.g. a partial recharge); it is thus arguable that in shared public facilities the customer may leave the system earlier for independent reasons. In the following sections both departure models will be analyzed separately.

As in Remark 1, our dynamics involves switching and a complete analysis would require either differential inclusions, or a regularization to obtain a smooth approximation. This paper does not address such issues: we will note below where our analysis may be subject to this limitation.

III. Sojourn time model

Our first model assumes sojourn time is an independent quantity, unrelated to received service. Customers have a *time budget* that must be split between traveling to the station, waiting for and receiving service. Travel and waiting constitute the cost, so routing according to (6) will maximize service time. These sojourn times may be random; in our fluid framework we simply let T represent the average sojourn time across customers.

Focusing on the population s_j assigned to each station, if there are no new arrivals, all the customers will leave after T time units, hence the appropriate departure rate is

$$d_j(s_j) = \frac{s_j}{T}. \tag{11}$$

Substituting in (10), our congestion signal becomes:

$$\mu_j(s_j) = \frac{[s_j - \bar{s}_j]^+}{d_j(s_j)} = T \left[1 - \frac{\bar{s}_j}{s_j}\right]^+. \tag{12}$$

Note again that μ_j has units of time; it varies in the interval $[0, T)$. It will be convenient to introduce as well the function

$$\begin{aligned} \phi_j(s_j) &:= \int_0^{s_j} \mu_j(\sigma) d\sigma \\ &= \begin{cases} 0 & s_j \leq \bar{s}_j, \\ T\left(s_j - \bar{s}_j - \bar{s}_j \log\left(\frac{s_j}{\bar{s}_j}\right)\right) & s_j > \bar{s}_j. \end{cases} \end{aligned} \tag{13}$$

This is a convex, monotonically increasing function, whose derivative $\phi_j'(s_j) = \mu_j(s_j)$ in (12). We next introduce a convex optimization problem involving this function.

A. Barrier optimization and its dual

In Section II-A we wrote the static transport optimization problem (2) assuming perfect balance. Consider now the following variation, in which the fixed supply constraint is replaced with a barrier cost:

$$\min \sum_{ij} c_{ij}\pi_{ij} + \sum_j \phi_j(s_j) \tag{14a}$$

$$\text{subject to:} \quad \pi_{ij} \geq 0 \ \forall i,j; \quad \sum_j \pi_{ij} = \bar{q}_i \ \ \forall i; \tag{14b}$$

$$\sum_i \pi_{ij} = s_j \ \ \forall j. \tag{14c}$$

The optimization variables are $\Pi = (\pi_{ij})$ and $s = (s_j)$. The barrier function $\phi_j(s_j)$ from (13) does not operate when s_j is below capacity; above that, a penalty term is applied. The problem is always feasible for $\bar{q}_i \geq 0$.

Our analysis of the preceding optimization and its connection to the dynamics will be based on duality. We write the Lagrangian with respect to the constraint (14c):

$$\begin{aligned} L(\Pi, s, \mu) &= \sum_{i,j} c_{ij}\pi_{ij} + \sum_j \phi_j(s_j) + \sum_j \mu_j \left[\sum_i \pi_{ij} - s_j\right] \\ &= \underbrace{\sum_{i,j} (c_{ij} + \mu_j)\pi_{ij}}_{L_1(\Pi,\mu)} + \underbrace{\sum_j \left[\phi_j(s_j) - \mu_j s_j\right]}_{L_2(s,\mu)}. \end{aligned} \tag{15}$$

Suggestively, we have denoted the multipliers by μ_j; to obtain the dual function we minimize the Lagrangian over the primal variables Π and s, each of which appears in a separate term, as indicated.

Note that constraints (14b) on Π are decoupled across i: the vector $\pi^i := (\pi_{ij})_{j=1}^n$ varies over $\bar{q}_i \Delta_n$, the unit simplex scaled by a constant factor. Reasoning as in (1), the corresponding minimum will be $\bar{q}_i \min_j (c_{ij} + \mu_j) = \bar{q}_i \varphi^i(\mu)$, using the notation in (7). Therefore:

$$D_1(\mu) = \min_{\Pi \in (14b)} L_1(\Pi, \mu) = \sum_i \bar{q}_i \varphi^i(\mu);$$

we further note that its superdifferential satisfies:

$$\partial D_1(\mu) = \left\{ \sum_i \hat{\pi}^i : \hat{\Pi} \in \arg\min_{\Pi \in (14b)} L_1(\Pi, \mu) \right\}. \tag{16}$$

The second sum $L_2(s,\mu)$ is unconstrained, and decoupled over s_j; we minimize each term $[\phi_j(s_j) - \mu_j s_j]$ separately. If $\mu_j \in (0,T)$ (interior to the range of $\phi_j'(s_j)$), we impose

$$\phi_j'(\hat{s}_j) = \mu_j \quad \Longrightarrow \quad \hat{s}_j = [\phi_j']^{-1}(\mu_j) = \frac{T\bar{s}_j}{T - \mu_j}. \tag{17}$$

Substitution of $\hat{s}_j$ into $\phi_j(s_j) - \mu_j s_j$ yields the minimum $T\bar{s}_j \log\left(1 - \frac{\mu_j}{T}\right)$, which is also valid for $\mu_j = 0$; the minimum is $-\infty$ for $\mu_j \notin [0,T)$. We conclude that

$$D_2(\mu) = \min_s L_2(s,\mu) = \sum_j T\bar{s}_j \log\left(1 - \frac{\mu_j}{T}\right),$$

for $\mu \in [0,T)^n$, and $-\infty$ outside this set.

For coordinates where $\mu_j > 0$, $D_2(\mu)$ is differentiable: in fact $\frac{\partial D_2}{\partial \mu_j} = -\hat{s}_j$, with $\hat{s}_j$ in (17). At $\mu_j = 0$, the right partial derivative is $-\bar{s}_j$; hence, any $-\hat{s}_j \geq -\bar{s}_j$ is a valid supergradient. We thus have the superdifferential:

$$\begin{aligned} \partial D_2(\mu) = \big\{ -\hat{s} \in \mathbb{R}^n : \hat{s}_j = \frac{T\bar{s}_j}{T - \mu_j} \text{ if } \mu_j > 0; \\ \hat{s}_j \leq \bar{s}_j \text{ if } \mu_j = 0 \big\}. \end{aligned} \tag{18}$$

The overall dual function $D(\mu) = D_1(\mu) + D_2(\mu)$ is bounded above (note $D(\mu) \to -\infty$ as $\mu_j \uparrow T$). A maximizing point μ^*, together with the minimizing variables $\hat{\Pi}, \hat{s}$ described above provide a saddle point of the Lagrangian in (15).

B. Equilibrium characterization

We proceed to relate the modified transport problem (14) to our model of dynamic load balancing under selfish routing and departures determined by sojourn times. For convenience, it is summarized below:

$$\dot{s} = a - d = \sum_{i=1}^m a^i - \frac{s}{T}; \tag{19a}$$

$$\mu_j(s_j) = \phi_j'(s_j) = T\left[1 - \frac{\bar{s}_j}{s_j}\right]^+ \quad \forall j. \tag{19b}$$

$$a^i \in r_i \partial\varphi^i(\mu) \quad \forall i. \tag{19c}$$

We will denote by $A = (a_{ij})$ the matrix of route flows.

An equilibrium of the above dynamics requires a choice of occupation states s_j^*, congestion delays $\mu_j^* = \phi_j'(s_j^*)$, and route flows A^* consistent with (19c) for $\mu = \mu^*$, such that the right-hand side of (19a) is zero.

Theorem 1: The following are equivalent:

(i) (s^*, A^*, μ^*) is an equilibrium point of (19), under constant r_i.

(ii) (s^*, Π^*, μ^*) is a saddle point of the Lagrangian L in (15), with $\Pi^* = A^*T$, $\bar{q}_i = r_i T$.

In particular, an equilibrium of the dynamics always exists.

Proof: Starting from (i), conditions (19c) imply that A^* minimizes $\sum_{ij}(c_{ij} + \mu_j^*)a_{ij}$, subject to $a_{ij} \geq 0$ and $\sum_j a_{ij} = r_i$ for each i. Multiplication by T yields $\pi_{ij}^* = Ta_{ij}^*$ that is a minimizer for $L_1(\Pi, \mu^*)$, under constraints (14b). Also, (19b) implies that s^* is a minimizer of $L_2(s, \mu^*)$, as discussed around equations (17)-(18). So $L(\Pi, s, \mu^*)$ is at a minimum, as required for a saddle point.

From the equilibrium in (19a) we have $\sum_i a_{ij}^* = \frac{s_j^*}{T}$; multiplying by T we conclude that $\sum_i \pi_{ij}^* = s_j^*$ for each j, i.e. (π^*, s^*) is primal feasible for Problem (14). $L(\pi^*, s^*, \mu)$ in (15) is independent of μ and therefore at a maximum in this variable. The saddle condition is established.

The steps are reversible: starting with a saddle point (Π^*, s^*, μ^*) of L, defining $A^* = \frac{1}{T}\Pi^*$ we obtain a selection of rates satisfying (19c) for multipliers μ^*; also, the saddle condition on s^* is consistent with (19b). Due to primal feasibility we have $a_j^* = s_j^*/T$, as required for equilibrium.

For the final statement, we have already justified the existence of a saddle point. ∎

We have found that equilibrium flows for dynamic load balancing map to solutions of an optimal transport problem. Further interpretations of this relationship are as follows:

- The demand quantity $\bar{q}_i = r_i T$, product of arrival rate and sojourn time, is the average number of customers present in the system originating from location i.
- The transported quantities $\pi_{ij} = T a_{ij}$ represent the average number of customers from location i assigned to station j.
- The total assignment s_j to station j is associated with the corresponding steady-state queue.
- The dual variable or shadow price μ_j of station j is associated with the queueing delay at the station.

Remark 3: The preceding analysis has some similarity with the results of selfish routing reviewed in Section II-C. In both cases, the equilibrium resulting from selfish balancing actions is characterized by an optimal cost, involving the *integral* of a latency cost function (compare (13) with (5)).

The main difference is that latencies in Section II-C were functions of network *flows*. For the *transport* latencies c_{ij}, the resulting cost term would be $\sum_{i,j} c_{ij} a_{ij}$; this is equal to our transport cost up to a constant factor. However, for the *waiting* cost, our latency is a function of station *queues*, not directly mapped to flows as in the other model.

Nevertheless, a *price of anarchy* arises here as well: the total waiting cost experienced by consumers is $\sum_j \mu_j s_j$, different from $\sum_j \phi_j(s_j)$; the equilibrium will not deliver the social optimum (minimum overall time to service).

C. Convergence

Beyond the characterization of the equilibrium, we aim at using optimization arguments to establish convergence of the dynamics (19). The main observation is that when (A, s, μ) follow these dynamics, we have

$$T\dot{s} = Ta - s \in \partial D(\mu), \tag{20}$$

subdifferential of the dual function of Problem 14, with $\bar{q}_i = r_i T$. This follows from our characterization of superdifferentials: for a^i satisfying (19c), $Ta = \sum_i T a^i = \sum_i \pi^i \in \partial D_1(\mu)$, as follows from (16); similarly, if s, μ are related by (19b), $-s$ is a supergradient of $D_2(\mu)$, invoking (18).

Note also that from (19b) we have

$$\dot{\mu} = \mathrm{diag}\left(\phi_j''(s_j)\right) \dot{s}.$$

If the dual function $D(\mu)$ were differentiable, we could apply the chain rule and write:

$$\frac{d}{dt} D(\mu(s(t))) = (\nabla D)^T \dot{\mu} = (\nabla D)^T \mathrm{diag}\left(\phi_j''(s_j)\right) \dot{s}. \tag{21}$$

However, we only have piecewise differentiability of $D(\mu)$; we will proceed under the simplification that ∇D can be replaced by the supergradient in (20).[1]

Also note that[2] $\phi_j''(s_j) = \frac{\bar{s}_j}{s_j^2} > 0$ for $s_j > \bar{s}_j$, and $\phi_J''(s_j) = 0$ for $s_j < \bar{s}_j$. We arrive at:

$$\frac{d}{dt} D(\mu(t)) = \sum_{j: s_j > \bar{s}_j} T \bar{s}_j \left[\frac{\dot{s}_j}{s_j}\right]^2 \geq 0. \tag{22}$$

We state the following conclusion:

Proposition 2: The dual function $D(\mu)$ for Problem 14 (with $\bar{q}_i = r_i T$) is non-decreasing along trajectories $\mu(t)$ arising from (19).

The preceding derivation does not qualify as a full proof of Proposition 2, given the simplification performed. Accepting the validity of (22), we have the following consequence:

Proposition 3: Consider a trajectory of the dynamics (19), such that $\frac{d}{dt} D(\mu(t)) \equiv 0$ for all time. Then $\mu(t) \equiv \mu^*$. Also:

- for any $j : \mu_j^* > 0$, $s_j(t) \equiv s_j^*$;
- for any $j : \mu_j^* = 0$, $s_j(t) \to s_j^*$;

the resulting s^* is an equilibrium of the dynamics.

Proof: The equality condition in (22) implies that $\dot{s}_j \equiv 0$ and so $s_j \equiv s_j^*$, constant in time for all $j : s_j > \bar{s}_j$; therefore $\mu_j = \phi_j'(s_j) \equiv \mu_j^* > 0$ for such stations. Non-saturated stations will have $\mu_j = 0$, and cannot exit from this condition without violating (22). Therefore, $\mu(t) \equiv \mu^*$.

As a consequence, the prices $c_{ij} + \mu_j^*$ seen by arriving customers remain constant, which implies[3] constant assigned rates a_{ij}^* and station arrival rates: $a_j(t) \equiv a_j^*$ for all j.

Stations with $\mu_j(t) \equiv 0$ ($s_j(t) \leq \bar{s}_j$) need not be in equilibrium; however they receive a constant rate a_j^* and thus evolve according to the first-order linear dynamics

$$\dot{s}_j = a_j^* - s_j/T,$$

converging exponentially to $s_j^* = T a_j^* \in [0, \bar{s}_j]$. Combined with $s_j \equiv s_j^*$ whenever $\mu_j^* > 0$, the limiting behavior is thus an equilibrium point of the dynamics. ∎

With the stated limitations in the analysis of switching, we see that $D(\mu(t))$ is increasing and bounded above, so it must approach a finite limit as $t \to \infty$. Furthermore, Proposition 3 is suggestive of a LaSalle invariance argument to establish global convergence to the equilibrium set. This method is, however, also difficult to formalize in the presence of a discontinuous field. Hence, convergence is at this point a (plausible) conjecture. Below, we observe it in simulations.

A variant of the dynamics, more amenable to standard Lyapunov analysis, could be to replace switching by a regularization, as outlined in Remark 1.

IV. Service time model

In this section we turn to a more traditional departure model in queueing systems: tasks depart when completing a *service* requirement, specified by T_0 in units of service *time*. Here again, in a fluid model we apply this quantity to all participants. The number of tasks *in service* at station j is $\min(s_j, \bar{s}_j)$; hence the departure rate is

$$d_j(s_j) = \frac{\min(s_j, \bar{s}_j)}{T_0}. \tag{23}$$

[1] It would suffice to guarantee that differentiability holds for almost all t along trajectories; however this is still potentially restrictive.

[2] We ignore the isolated point of non-differentiability of $\phi_j'(s_j)$.

[3] Assuming the tie-breaking rule is time-invariant.

With this choice, our queueing delay from (10) becomes

$$\mu_j(s_j) = \frac{[s_j - \bar{s}_j]^+}{d_j(s_j)} = T_0 \left[\frac{s_j}{\bar{s}_j} - 1\right]^+. \tag{24}$$

A. Constrained optimization and its dual

We now formulate an alternate transport optimization problem, also a variant of (2) in Section II-A:

$$\min \sum_{ij} c_{ij}\pi_{ij} \tag{25a}$$

$$\text{subject to:} \quad \pi_{ij} \geq 0 \ \forall i,j; \quad \sum_j \pi_{ij} = \bar{q}_i \quad \forall i; \tag{25b}$$

$$\sum_i \pi_{ij} \leq \bar{s}_j \quad \forall j. \tag{25c}$$

Instead of a barrier penalty for exceeding capacity as in (14), here we impose hard constraints; for feasibility, our problem will require the condition $\sum_i \bar{q}_i \leq \sum_j \bar{s}_j$.

The Lagrangian with respect to the supply constraints (25c), with multipliers $\mu_j \geq 0$ takes the form

$$\begin{aligned}\tilde{L}(\Pi, \mu) &= \sum_{i,j} c_{ij}\pi_{ij} + \sum_j \mu_j \left[\sum_i \pi_{ij} - \bar{s}_j\right] \\ &= \sum_{i,j} (c_{ij} + \mu_j)\pi_{ij} - \sum_j \mu_j \bar{s}_j. \end{aligned} \tag{26}$$

The minimum over Π (the only primal variables in this problem) under constraints (25b) can be solved analogously to the previous section, giving the dual function

$$\tilde{D}(\mu) = \sum_i \bar{q}_i \min_j (c_{ij} + \mu_j) - \sum_j \mu_j \bar{s}_j. \tag{27}$$

The corresponding subdifferential can be expressed in terms of any $\hat{\Pi} \in \arg\min \tilde{L}(\Pi, \mu)$:

$$\partial \tilde{D} = \sum_i \bar{q}_i \partial \varphi^i - \bar{s} = \sum_i \hat{\pi}^i - \bar{s}. \tag{28}$$

Under the feasibility condition for the primal problem, $\tilde{D}(\mu)$ is bounded in $\mu \geq 0$, and has a global maximum a certain μ^*. Together with any

$$\Pi^* \in \arg\min \tilde{L}(\Pi, \mu^*), \ \text{subject to (25b)}, \tag{29}$$

it defines a saddle point of the Lagrangian. A saddle point (Π^*, μ^*) is characterized by the following requirements: Π^* must be primal feasible for (25) and satisfy (29); $\mu^* \geq 0$, and complementary slackness must hold:

$$\mu_j^* \Big(\sum_i \pi_{ij}^* - \bar{s}_j \Big) = 0 \quad \forall j. \tag{30}$$

B. Equilibrium characterization

We now relate Problem (25) to the load-balancing dynamics under the service time departure model, summarized below:

$$\dot{s} = a - d = \sum_{i=1}^m a^i - \frac{\min(s_j, \bar{s}_j)}{T_0}; \tag{31a}$$

$$\mu_j(s_j) = T_0 \left[\frac{s_j}{\bar{s}_j} - 1\right]^+. \quad \forall j. \tag{31b}$$

$$(a_{ij})_{j=1}^n =: a^i \in r_i \partial \varphi^i(\mu) \quad \forall i. \tag{31c}$$

Theorem 4: The following are equivalent:

(i) (s^*, A^*, μ^*) is an equilibrium point of (31), under constant r_i.

(ii) (Π^*, μ^*) is a saddle point of the Lagrangian $\tilde{L}$ in (26), with $\Pi^* = A^* T_0$, $\bar{q}_i = r_i T_0$, and s^* is given by:

$$s_j^* = \begin{cases} \bar{s}_j \left(1 + \frac{\mu_j^*}{T_0}\right) & \text{if } \mu_j^* > 0; \\ \sum_i \pi_{ij}^* & \text{if } \mu_j^* = 0. \end{cases} \tag{32}$$

An equilibrium exists provided that $\sum_i r_i \leq \sum_j \bar{s}_j / T_0$.

Proof: Starting with (i), note that (31c) implies that A^* minimizes $\sum_{ij}(c_{ij} + \mu_j^*)a_{ij}$, subject to $a_{ij} \geq 0$ and $\sum_j a_{ij} = r_i$ for each i. Multiplication by T_0 yields that $\Pi^* = A^* T_0$ satisfies (25b) with $\bar{q}_i = r_i T_0$, and (29) holds.

Also, by the equilibrium condition we have:

$$\sum_i \pi_{ij}^* = T_0 a_j^* = T_0 d_j(s_j^*) = \min(s_j^*, \bar{s}_j) \leq \bar{s}_j; \tag{33}$$

hence π^* satisfies (25c) and is primal feasible.

$\mu^* \geq 0$ holds by (31b). For complementary slackness, suppose that $\mu_j^* > 0$. Then (31b) gives $s_j^* = \bar{s}_j \left(1 + \frac{\mu_j^*}{T_0}\right)$. In particular $s_j^* > \bar{s}_j$, so there is no gap in the rightmost inequality of (33); we have $\sum_i \pi_{ij}^* = \bar{s}_j$ and the second factor in (30) vanishes.

The above also establishes the first case of (32). In the alternative case $\mu_j^* = 0$, $\min(s_j^*, \bar{s}_j) = s_j^*$ from (31b), and therefore from (33) we obtain $\sum_i \pi_{ij}^* = s_j^*$ as claimed.

Now start from (ii), with a saddle point (Π^*, μ^*) and s^* as in (32). Choosing $A^* = \Pi^*/T_0$, it follows directly from (29) that (31c) holds for $\mu = \mu^*$. For convenience denote

$$\hat{s}_j := \sum_i \pi_{ij}^* \leq \bar{s}_j; \tag{34}$$

we claim that the following conditions both hold at each j:

$$\mu_j^* = T_0 \left[\frac{s_j^*}{\bar{s}_j} - 1\right]^+; \tag{35a}$$

$$\hat{s}_j = \min(s_j^*, \bar{s}_j). \tag{35b}$$

Indeed, if $\mu_j^* = 0$, (32) and (34) give $s_j^* = \hat{s}_j \leq \bar{s}_j$; both conditions above can be checked in this case.

If $\mu_j^* > 0$, now by (32) we have $s_j^* > \bar{s}_j$; this immediately yields (35a). Also, in this case by complementary slackness (30) we must have $\hat{s}_j = \sum_i \pi_{ij}^* = \bar{s}_j$ and then (35b) holds.

Note finally that (35a) is just a re-statement of (31b), and that (35b) and (23) imply that the departure rate at s_j^* is

$$d_j(s_j^*) = \frac{\hat{s}_j}{T_0} = \frac{1}{T_0} \sum_i \pi_{ij}^* = a_j^*,$$

consistent with equilibrium in (31a).

For the final statement, note the feasibility conditions for existence of saddle points. ∎

We provide the following comments for our result:

- The stability condition (for existence of equilibrium) has a natural interpretation: since vehicles do not depart without completing service, their total arrival rate $\sum_i r_i$ cannot exceed the maximum total service rate; note that service rates at station j are bounded by $\bar{s}_j/T_0$.
- Quantities in the optimization have a different interpretation from those in the previous section. $\bar{q}_i = r_i T_0$ now represents average number of *service slots* requested from location i. The resulting matrix Π^* distributes these requests between stations, resulting in a number $\hat{s}_j$ as in (34) of service slots occupied in each station.
- $\hat{s}_j$ may *differ* from s_j^*, equilibrium population of EVs at station j under load balancing. The difference appears in saturated stations with queueing delay $\mu_j^* > 0$.

Remark 4: If there were a centralized dispatcher splitting requests according to $A^* = \Pi^*/T_0$, queueing delay could be completely avoided. This is the *price of anarchy* in our dynamics: station delays must build up to coax selfish drivers to the optimal transport allocation.

C. Convergence

We sketch the analysis which parallels the one in the previous section. We differentiate the dual function along trajectories of the dynamics (31), replacing $\nabla \tilde{D}(\mu)$ by a valid supergradient. Invoking (28), we note that (A, S, μ) are constrained by (31c)- (31b), then

$$T_0 a - \bar{s} \in \partial \tilde{D}(\mu);$$

this leads to:

$$\frac{d}{dt}\tilde{D}(\mu(t)) = \sum_j [T_0 a_j - \bar{s}_j]\dot{\mu}_j(t) = \sum_{s_j > \bar{s}_j} T_0 \dot{s}_j \frac{T_0}{\bar{s}_j}\dot{s}_j \geq 0.$$

Above, we first eliminate terms where $s_j(t) < \bar{s}_j$, which have $\dot{\mu}_j(t) = 0$. For saturated stations, the positive part in (31b) is inactive, and $T_0 a_j - \bar{s}_j = T_0 \dot{s}_j$, invoking (31a). The following statements parallel Propositions 2 and 3. Again, we are not carefully addressing discontinuous switching.

Proposition 5: The dual function $\tilde{D}(\mu)$ for Problem (25) (with $\bar{q}_i = r_i T_0$) is non-decreasing along trajectories $\mu(t)$ arising from the dynamics (31).

Proposition 6: Consider a trajectory of the dynamics (31), such that $\frac{d}{dt}\tilde{D}(\mu(t)) \equiv 0$ for all time. Then $\mu(t) \equiv \mu^*$. Also:

- for any $j : \mu_j^* > 0$, $s_j(t) \equiv s_j^*$;
- for any $j : \mu_j^* = 0$, $s_j(t) \to s_j^*$;

the resulting s^* is an equilibrium of the dynamics.

V. Stochastic Simulations

In this section we provide simulations of our load balancing dynamics for illustration purposes, and also to contrast our fluid models with more realistic scenarios, involving discrete EVs with stochastic arrival locations and times, as well as random sojourn times.

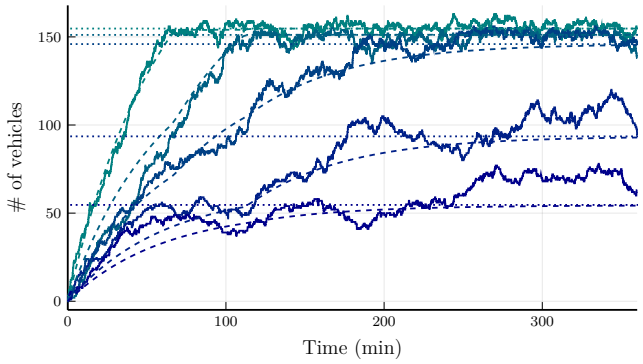

Fig. 1. Time evolution of station occupations for the sojourn time model in the simulated stochastic system (solid), the fluid solution (dashed) and the predicted equilibrium of Theorem 1 (dotted).

Our spatial domain is taken to be a square region, where recharge requests arrive as a Poisson process of overall rate $r = 10$ EVs/min, spawning in a random spatial location, uniformly chosen in the region. Travel times c_{ij} are modeled as the Euclidean distance divided by a speed v chosen such that the maximum travel time across the region is 10 minutes. We also fix 5 charging station locations at random in the region, each with capacity for charging $\bar{s}_j = 150$ EVs simultaneously. In Figure 2 we mark the positions of these charging stations.

In our first example, we follow the sojourn time model of Section III, and choose EV sojourn times as exponentially distributed with mean $T = 60$ minutes; Therefore, in steady state there should be an average of $rT = 600$ EVs in the system, split among the service points. Note that $rT < \sum_j \bar{s}_j$, so the system can cope well with demand, but due to the random location of chargers, some of them may experience congestion.

Upon arrival, vehicles are routed to stations in accordance to the minimum time rule based on the current congestion prices μ_j given by (12), and the charging station occupation s_j is updated accordingly. In order to simulate our fluid model, with finite arrival locations m, we discretize the space in a grid of 10000 points.

Figure 1 shows the evolution of the stochastic occupations and their comparison with the solutions of the fluid model under (19) through an ordinary differential equation solver.

Initially, all stations are uncongested and thus EVs are routed to the closest station. However, with our arrival and stations pattern, the rightmost station which is far from the others gets congested. This forces the $\mu_1(t)$ upwards, and thus the other two closer stations start to receive more traffic, until eventually the first two stations experience congestion, the third one operates near congestion and the other two remain operating below capacity. The fluid dynamics reaches the equilibrium predicted by the optimization problem (14), shown in Figure 1 in dotted lines.

In order to better visualize this congestion pattern, in Figure 2 we represent the attraction regions for each station in equilibrium. We note that the rightmost cell has shrunk

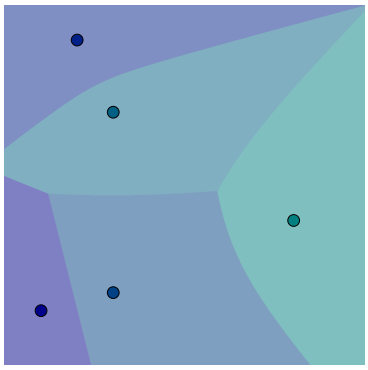

Fig. 2. Charging station positions and attraction regions in equilibrium for the sojourn time model.

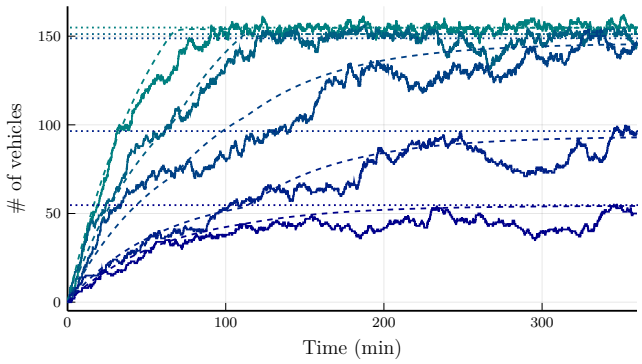

Fig. 3. Time evolution of station occupations for the service time model in the simulated stochastic system (solid), the fluid solution (dashed) and the predicted equilibrium of Theorem 4 (dotted).

with respect to the Voronoï tessellation which would appear in the uncongested case.

We now turn our attention to the service time model from Section IV. Arrivals are as before, but now EVs only leave the system after completing their charge, which is randomly chosen from an exponential distribution with mean $T_0 = 60$ min. Note that, with this choice of parameters, the total demand $\sum_i \bar{q}_i = rT_0 < \sum_i \bar{s}_i$ so the stability condition for existence of equilibrium in Theorem 4 is satisfied. As before, upon arrival, vehicles are routed according to the minimum time rule, but with the congestion price given by (24) to represent queueing delay.

In Figure 3 we plot again the evolution of station occupations, starting from an empty system. We can see that the fluid model again captures the right trend from the stochastic occupations. In steady state, the system converges to the equilibrium given in Theorem 4, also shown in the Figure, where two stations become congested. The steady state attraction regions in this case are similar to the ones in Figure 2, and we omit them.

VI. Conclusions and Future Work

For a distributed EV charging infrastructure, we have analyzed through a fluid model the load balancing dynamics of selfish users responding to information on delay to service, comprised of transport and queueing delays. Two departure models were considered, depending on whether sojourn times or service times are taken to be given. Under stationary arrivals, we established in each case a connection, in terms of equilibrium and dynamic evolution, with an appropriate convex optimization problem. A complete treatment of the switching dynamics in our model is left for future work.

Other future directions are: (i) the more extensive analysis of the price of anarchy and mitigation strategies for these problems; (ii) the consideration of *elastic* demand, where some arriving load is curtailed due to the unwillingness of customers to accept the current level of delay.

References

[1] F. M. Ceragioli, "Discontinuous ordinary differential equations and stabilization," *PhD Theis, Universita degli Studi di Firenze*, 1999.

[2] G. Como and R. Maggistro, "Distributed dynamic pricing of multiscale transportation networks," *IEEE Transactions on Automatic Control*, vol. 67, no. 4, pp. 1625–1638, 2021.

[3] M. Daskin, "Network and discrete location: Models, algorithms and applications," *Journal of the Operational Research Society*, vol. 48, no. 7, pp. 763–764, 1997.

[4] A. Ephremides, P. Varaiya, and J. Walrand, "A simple dynamic routing problem," *IEEE transactions on Automatic Control*, vol. 25, no. 4, pp. 690–693, 1980.

[5] G. França, D. P. Robinson, and R. Vidal, "Gradient flows and proximal splitting methods: A unified view on accelerated and stochastic optimization," *Physical Review E*, vol. 103, no. 5, p. 053304, 2021.

[6] J. Krarup and P. M. Pruzan, "The simple plant location problem: Survey and synthesis," *European Journal of Operational Research*, vol. 12, no. 1, pp. 36–81, 1983.

[7] M. Kuby and S. Lim, "The flow-refueling location problem for alternative-fuel vehicles," *Socio-Economic Planning Sciences*, vol. 39, no. 2, pp. 125–145, 2005.

[8] Z. J. Lee, G. Lee, T. Lee, C. Jin, R. Lee, Z. Low, D. Chang, C. Ortega, and S. H. Low, "Adaptive charging networks: A framework for smart electric vehicle charging," *IEEE Transactions on Smart Grid*, vol. 12, no. 5, pp. 4339–4350, 2021.

[9] J. Liu, "Electric vehicle charging infrastructure assignment and power grid impacts assessment in Beijing," *Energy policy*, vol. 51, pp. 544–557, 2012.

[10] J. C. Mukherjee and A. Gupta, "A review of charge scheduling of electric vehicles in smart grid," *IEEE Systems Journal*, vol. 9, no. 4, pp. 1541–1553, 2014.

[11] F. Paganini, E. Espíndola, D. Marvid, and A. Ferragut, "Optimization of spatial infrastructure for EV charging," in *61st IEEE Conference on Decision and Control*, 2022.

[12] N. Parikh, S. Boyd *et al.*, "Proximal algorithms," *Foundations and trends® in Optimization*, vol. 1, no. 3, pp. 127–239, 2014.

[13] T. A. Roughgarden, *Selfish routing*. PhD Dissertation, Cornell University, 2002.

[14] F. Santambrogio, "Optimal transport for applied mathematicians," *Birkäuser, NY*, vol. 55, no. 58-63, p. 94, 2015.

[15] M. Van der Boor, S. C. Borst, J. S. Van Leeuwaarden, and D. Mukherjee, "Scalable load balancing in networked systems: A survey of recent advances," *SIAM Review*, vol. 64, no. 3, pp. 554–622, 2022.

[16] J. G. Wardrop, "Road paper. some theoretical aspects of road traffic research." *Proceedings of the institution of civil engineers*, vol. 1, no. 3, pp. 325–362, 1952.

[17] White House, "FACT SHEET: The Biden-Harris Electric Vehicle Charging Action Plan," www.whitehouse.gov/briefing-room/statements-releases/2021/12/13/fact-sheet-the-biden-harris-electric-vehicle-charging-action-plan/, 2021.

[18] M. Zeballos, A. Ferragut, and F. Paganini, "Proportional fairness for EV charging in overload," *IEEE Transactions on Smart Grid*, vol. 10, no. 6, pp. 6792–6801, 2019.