Reinforcement Learning Portfolio Optimization of Electric Vehicle Virtual Power Plants

Master Thesis

Author: Tobias Richter (Student ID: 558305)
Supervisor: Univ.-Prof. Dr. Wolfgang Ketter
Co-Supervisor: Karsten Schroer

Department of Information Systems for Sustainable Society
Faculty of Management, Economics and Social Sciences
University of Cologne

March 26, 2019

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Die Strafbarkeit einer falschen eidesstattlichen Versicherung ist mir bekannt, namentlich die Strafandrohung gemäß § 156 StGB bis zu drei Jahren Freiheitsstrafe oder Geldstrafe bei vorsätzlicher Begehung der Tat bzw. gemäß § 161 Abs. 1 StGB bis zu einem Jahr Freiheitsstrafe oder Geldstrafe bei fahrlässiger Begehung.

Tobias Richter

Köln, den 01.05.2019

Contents

List of Figures

List of Tables

List of Abbreviations

ANN Artificial Neural Network

DP Dynamic Programming

DSO Distribution System Operator

EPEX European Power Exchange

EV Electric Vehicle

GCRM German Control Reserve Market

GP Genetic Programming

MAW Mean Asymmetric Weighted Objective Function

MC Monte Carlo

MDP Markov Decision Process

PDF Probability Density Function

RES Renewable Energy Sources

RL Reinforcement Learning

TD Temporal-Difference

TSO Transmission System Operator

V2G Vehicle-to-Grid

VPP Virtual Power Plant

Summary of Notation

Capital letters are used for random variables, whereas lower case letters are used for the values of random variables and for scalar functions. Quantities that are required to be real-valued vectors are written in bold and in lower case (even if random variables).

$\doteq$	equality relationship that is true by definition
$\approx$	approximately equal
$\mathbb{E}[X]$	expectation of a random variable X, i.e., $\mathbb{E}[X] \doteq \sum_x p(x)x$
$\mathbb{R}$	set of real numbers
$\leftarrow$	assignment
ε	probability of taking a random action in an ε-greedy policy
α	step-size parameter
γ	discount-rate parameter
λ	decay-rate parameter for eligibility traces
s, s'	states
a	an action
r	a reward
$\mathcal{S}$	set of all nonterminal states
$\mathcal{A}$	set of all available actions
$\mathcal{R}$	set of all possible rewards, a finite subset of $\mathbb{R}$
$\subset$	subset of; e.g., $\mathcal{R} \subset \mathbb{R}$
$\in$	is an element of; e.g., $s \in \mathcal{S}$, $r \in \mathcal{R}$
t	discrete time step
$T, T(t)$	final time step of an episode, or of the episode including time step t
A_t	action at time t
S_t	state at time t, typically due, stochastically, to S_{t-1} and A_{t-1}
R_t	reward at time t, typically due, stochastically, to S_{t-1} and A_{t-1}
π	policy (decision-making rule)
$\pi(s)$	action taken in state s under *deterministic* policy π
$\pi(a\|s)$	probability of taking action a in state s under *stochastic* policy π
G_t	return following time t
$p(s', r \mid s, a)$	probability of transition to state s' with reward r, from state s and action a
$p(s' \mid s, a)$	probability of transition to state s', from state s taking action a
$v_\pi(s)$	value of state s under policy π (expected return)

$v_*(s)$	value of state s under the optimal policy
$q_\pi(s, a)$	value of taking action a in state s under policy π
$q_*(s, a)$	value of taking action a in state s under the optimal policy
V, V_t	array estimates of state-value function v_π or v_*
Q, Q_t	array estimates of action-value function q_π or q_*
d	dimensionality—the number of components of $\mathbf{w}$
$\mathbf{w}$	d-vector of weights underlying an approximate value function
$\hat{v}(s, \mathbf{w})$	approximate value of state s given weight vector $\mathbf{w}$
$\mu(s)$	on-policy distribution over states
$\overline{\mathrm{VE}}$	mean square value error

1 Introduction

1.1 Research Motivation

- (Lopes et al., 2011)

1.2 Research Question

1.3 Relevance

2 Background

2.1 Smart Electricity Markets

On electricity markets, actors participate in auctions to match the supply of electricity generation and the demand for electricity consumption. Participants place asks (sale offers) and bids (purchase orders). The electricity price is determined by an auction mechanism, which can take different forms depending on the type of market. Germany, like many other western countries, has a liberalized energy system in which the generation and distribution of electricity are decoupled. Multiple electricity markets exist in a liberalized energy system. They differ in the auction design and in their reaction time between the order contract and the delivery of electricity. Day-ahead markets and spot markets have a reaction time between a day and several hours, whereas in operating reserve markets the reaction time ranges from minutes to seconds. The auction mechanism design is essential for electricity markets (Kambil & van Heck, 1998). Electricity markets work according to the merit order principle in which resources are considered in ascending order of the energy price until the capacity demand is met. The clearing price is determined by the energy price, at the point where supply meets demand. Payment models differ in the markets: In contrast to day-ahead markets, where a uniform pricing schema is applied, in secondary reserve markets and intraday markets bidders, get compensated by the price they bid (pay-as-bid principle).

EV fleet operators can offer the capacity of their EV batteries on multiple markets at the same time to make use of the different market properties. On operating reserve markets,prices are usually more volatile and consequently more attractive for VPPs (Tomić & Kempton, 2007). Operating reserve markets also bear a higher risk for the fleet: Commitments have to be made one week in advance when customer demands are still uncertain. In order to not face penalties for unfulfilled commitments only a conservative amount of capacity can be offered to the market. On the other hand, spot markets allow participants to continuously trade electricity products up to five minutes prior to delivery. At this point in time, it is possible to predict available battery capacity of the fleet with high accuracy. This certainty creates the possibility to trade the remaining available capacity with low risk at the spot market. In the following, we will explain the market design of balancing markets and spot markets in more detail, since they are the markets we included in our research.

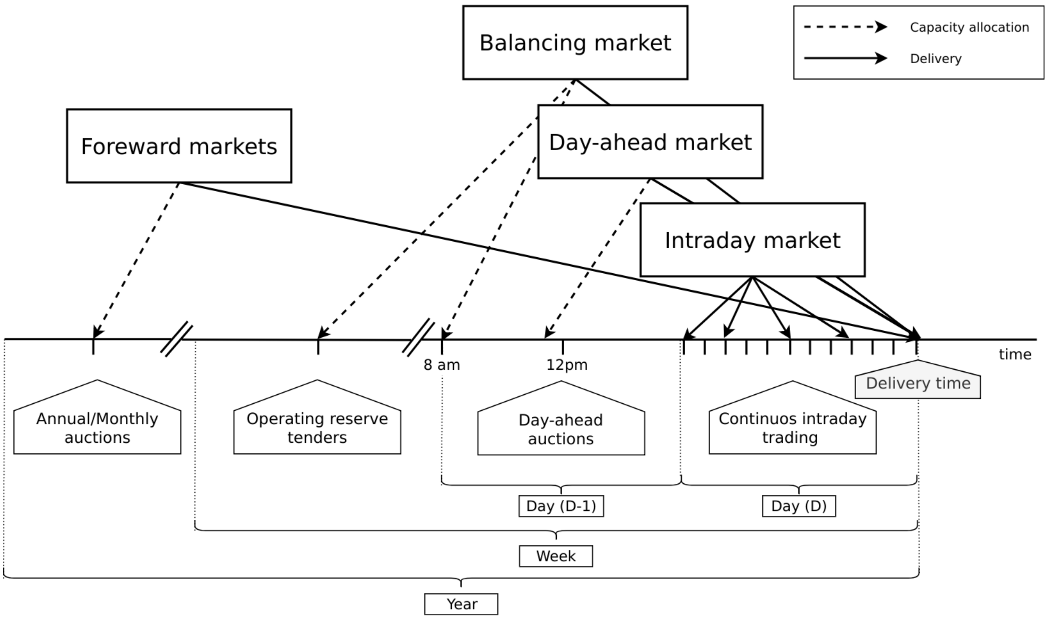

Figure 1: Interaction between electricity markets in relation to capacity allocation

2.2 Electricity Market Theory:

2.2.1 Balancing Market

The balancing market is a tool to balance frequency deviations in the power grid. It offers auctions for primary control reserve, secondary control reserve as well as tertiary control reserve (minute reserve), which primarily differ in the required ramp-up times of the participants. As depicted in Figure 1, the balancing market can be seen as the last link in a chain of electricity markets (van der Veen & Hakvoort, 2016). In this study, we will look at the German Control Reserve Market (GCRM), one of the largest frequency regulation markets in the world. However, the presented concepts can be easily transferred to other balancing markets in unbundled energy systems, since the market design is similar (Brandt et al., 2017). Transmission Systems Operators (TSO) procure their required control reserve via tender auctions at the GCRM. The market conducts daily auctions for the three types of control reserve. This thesis focuses on the secondary operating reserve auction, in which participants must be able to supply or absorb a minimum of 1MW of power over a 4-hour interval with a reaction time of 30 seconds.[1] Since EV batteries can absorb energy almost instantly, when they are connected to a charging station, they are suitable to provide such balancing services. Operating reserve providers have to be qualified by the TSO to participate in the market and are able to reliably provide the committed capacity. Although EV fleets are currently not qualified by the GCRM to be used as operating reserve, they could theoretically handle the minimum capacity requirements. Around 220 EVs would need to simultaneously charge at standard 4.6kW charging stations to provide 1MW of downward regulating capacity.

Up until 28th July 2018, auctions were held weekly, with two different segments each week (peak hours/non-peak hours). Afterwards, the auction mechanism changed to *daily* auctions of six four-hour segments of positive and negative control reserve.[2] Shorter auction cycles facilitate the integration of renewable energy generators into the secondary control reserve market, as they are dependent on accurate (short-term) capacity forecasts.

Positive control reserve is energy that is supplied to the grid, when the grid frequency falls below 50Hz. It can be provided by increasing the electricity generation or by reducing the grid load (i.e., electricity consumption). On the contrary, negative control reserve is required when the grid frequency rises above 50Hz and can be provided by adding grid load or reducing electricity generation. Since we do not consider V2G in this thesis, the EV fleets in our model are only able provide *negative control reserve*, which we will refer to as *control reserve* until the end of the thesis. Market participants submit bids in the following form to the market: (P^{bal}, p^c, p^e), where P^{bal} is the amount of electrical power that can be supplied on demand in kW, p^c is the capacity price for keeping the power available in $\frac{€}{\text{MW}}$ and p^e is the energy price for delivered energy in $\frac{€}{\text{MWh}}$. The TSO determines the target quantity of energy to acquire per timeslot, it usually acquires much higher regulation capacity to minimize risks and activates the capacity on demand. The TSO accepts the bids based on the capacity price in a merit order. Providers, whose bids were accepted, instantly get compensated for the provided capacity: $R^c = p^c \times P^{bal}$. At the time regulation capacity is needed, usually a day to a week later, the TSO activates the capacity according to a merit order of the ascending *energy prices* p^e. Hence, providers are also compensated according to the actual energy E^{bal} they supplied or consumed: $R^e = p^e \times E^{bal}$. Since provider get paid according to their submitted price p^e, instead of a market clearing price,

[1]See `https://regelleistung.net`, accessed on 15th February 2019, for further information on the market design and historical data.

[2]`https://www.bundesnetzagentur.de/SharedDocs/Pressemitteilungen/DE/2017/28062017_Regelenergie.html`, accessed 18th February, 2019

this type of auction is called *pay-as-bid* auction.

2.2.2 Spot Market

As mentioned in the previous chapter, the equilibrium of electricity supply and demand is ensured through a sequence of interdependent wholesale markets (cited) (Pape et al., 2016). Next to the balancing market at the end of the sequence, mainly two different types of spot markets exist, the day-ahead market and the intraday market. In this research, we consider the European Power Exchange (EPEX Spot) as it is the largest electricity market in Europe, with a total trading volume of approximately 567TWh in 2018[3], but most electronic spot markets in western economies work with similar market mechanisms.

In Germany, the most important spot market is the day-ahead market with a trading volume of over 234TWh in 2018[3]. Participants place asks and bids for hourly contracts of the following day on the *EPEX Spot Day-ahead Auction* market until the market closes at 12pm on the day before delivery (see Figure 1). The day-ahead market plays an essential role in integrating volatile renewable energy sources (RES) into the power system (Pape et al., 2016). Generators forecast the expected generation capacity for the next day and sell those quantities on the market (Karanfil & Li, 2017). After the market closes, the participants have the opportunity to trade the difference between the day-ahead forecast and the more accurate intraday forecast on the intraday market (Kiesel & Paraschiv, 2017). In this way, RES generators can cost effectively self-balance their portfolios, instead of relying on balancing services provided by the TSO, which imposes high imbalance costs on participants (Pape et al., 2016).

On the *EPEX Spot Intraday Continuous* market, electricity products are traded up until 5 minutes before physical delivery. Hourly contracts, as well as 15-minute and block contracts, can be traded. In contrast to the day-ahead auction, the intraday market is a continuous order-driven market. Participants can submit limit orders at any time during the trading window and equally change or withdraw the order at any time before the order is accepted. Limit orders are specified as price-quantity pairs: (P^{intr}, p^u), where P^{intr} is the traded amount of electrical power in kW and p^u is the price for the delivered energy unit (hour/quarter/block) in $\frac{€}{\text{MWh}}$. When an order to buy (bid) matches an order to sell (ask), the trade immediately gets executed. The order book is visible to all participants, hence it is known which unmatched orders exist at the time of interest. The intraday market has a trading volume of 82TWh, which is considerably smaller than day-ahead market's volume. Despite that, the intraday market plays a vital role to

[3] `https://www.epexspot.com/en/press-media/press/details/press/Traded_volumes_soar_to_an_all-time_high_in_2018`, accessed 19$^{\text{th}}$ February, 2019

the stability of the grid. All executed trades on the intraday market potentially reduce the activation of control reserve through the TSO.

Purchasing electricity on the continuous intraday market is attractive for EV fleets with uncertain mobility demand. Due to the intradays market's short time before delivery, EV fleet operators can rely on highly accurate forecasts of available battery capacity to charge, before submitting an order to buy. In this way, they can reliably charge at a potentially lower price at the intraday market than the regular industry tariff. In an integrated bidding strategy, EV fleet operators can, similarly to RES generators, balance out forecast errors of available battery capacity on the intraday market. Trades on the intraday market can complement bids that have been committed to other markets earlier (e.g., to the secondary operating reserve market).

2.3 EV Fleet Control in the Smart Grid

The increasing penetration of EVs has a substantial effect on electricity consumption patterns. During charging periods, power flows and grid losses increase considerably and challenge the grid. Operators have to reinforce the grid to ensure that transformers and substations do not overload (Sioshansi, 2012; Lopes et al., 2011). Loading multiple EVs in the same neighborhood, or worse, whole EV fleets at once, stress the grid. In these cases, even brown- or blackouts can occur. (Kim et al., 2012). Despite these challenges, it is possible to support the physical reinforcement by adopting smart charging strategies. In smart charging, EVs get charged when the grid is less congested to ensure grid stability. Smart charging reduces peaks in electricity demand, called *Peak Cutting*, and complement the grid in times of low demand, called *Valley Filling*. Smart charging has been researched thoroughly in the IS literature, in the following we will outline some of the most important contributions.

Valogianni et al. (2014) found that using intelligent agents to schedule EV charging substantially reshapes the energy demand and reduces peak demand without violating individual household preferences. Moreover, they showed that the proposed smart charging behavior reduces average energy prices and thus benefit households economically. In another study, Kara et al. (2015) investigated the effect of smart charging on public charging stations in California. Controlling for arrival and departure times, the authors presented beneficial results for the distribution system operator (DSO) and the owners of EVs. Their approach resulted in a price reduction in energy bills and a peak load reduction. An extension of the smart charging concept is Vehicle-to-Grid (V2G). When equipped with V2G devices, EVs can discharge their batteries back into the grid. Existing research has focused on this technology in respect to grid stabilization effects and arbitrage possibilities. For instance, Schill (2011) showed that the usage of EVs can decrease average consumer electricity prices. Excess EV battery capacity can be used to charge in off-peak hours and discharge in peak hours, when the prices are higher. These arbitrage possibilities reverse welfare effects of generators and increase the overall welfare and consumer surplus. Tomić and Kempton (2007) found that the arbitrage opportunities are especially prominent when a high variability in electricity prices on the target electricity market exists. The authors stated that short intervals between the contract of sale and the physical delivery of electricity increase arbitrage benefits. Consequently, ancillary service markets, like frequency control and operating reserve markets, are attractive for smart charging.

Peterson et al. (2010) investigated energy arbitrage profitability with V2G in the light of battery depreciation effects in the US. The results of their study indicate that large-scale use of EV batteries for grid storage does not yield enough monetary benefits to incentivize EV owners to participate in V2G activities. Considering battery depreciation cost, the authors arrived at an annual profit of only 6$ - 72$ per EV. Brandt et al. (2017) evaluated a business model for parking garage operators operating on the German frequency regulation market. When taking infrastructure costs and battery depreciation costs into account, they conclude that the proposed vehicle-grid integration is not profitable. Even with idealized assumptions about EV adoption rates in Germany and altered auction mechanisms, the authors arrived at negative profits. Kahlen et al. (2017) used EV fleets to offer balancing services to the grid. Evaluating the impact of V2G in their model, the authors conclude that V2G would only be profitable if reserve power prices were twice as high. Considering the results from the studies mentioned above, our research does not include V2G, since only marginal profits are expected.

In order to maximize profits, it is essential for market participants to develop successful bidding strategies. Several authors have investigated bidding strategies to jointly participate in multiple markets (Mashhour & Moghaddas-Tafreshi, 2011a; He et al., 2016). Mashhour and Moghaddas-Tafreshi (2011a) used stationary battery storage to participate in the spinning reserve market and the day-ahead market at the same time. The authors developed a non-equilibrium model, which solves the presented mixed-integer program with Genetic Programming (GP). Contrarily, we use a model-free RL agent that learns an optimal policy (i.e., a trading strategy) from actions it takes in the environment (i.e., bidding on electricity markets). Using a model-free approach is especially beneficial for us, since additional unknown variables and constraints (i.e., customer mobility demand) complicate the formulation of a mathematical model.

He et al. (2016) conducted similar research to Mashhour and Moghaddas-Tafreshi (2011a). The authors additionally incorporated battery life cycle in their profit maximization model, which proved to be a decisive factor. In contrast to the authors, we jointly participated in the secondary operating reserve and spot market with the *non-stationary* storage of EV batteries. Because shared EVs have to satisfy mobility demand, they have to be charged in any case, which allows us to safely exclude battery depreciation from our model. Further, we chose the intraday market over the day-ahead market, as it has the lowest reaction time among the spot markets, and thus potentially offers higher profits (Tomić & Kempton, 2007).

Previous studies often assume that car owners or households can directly trade on electricity markets. In reality, this is not possible due to the minimum capacity requirements of the markets, requirements that single EVs do not meet. For example, the German Control Reserve Market (GCRM) has a minimum trading capacity of 1MW to 5MW, depending on the specific market. In order to reach the minimum capacity, over 200 EVs would need to be connected to the grid via a standard 4.6kW charging station at the same time. Ketter et al. (2013) introduced the notion of electricity brokers, aggregators that act on behalf of a group of individuals or households to participate in electricity markets. Brandt et al. (2017) and Kahlen et al. (2014) successfully showed that electricity brokers can overcome the capacity issues by aggregating EV batteries. In addition to electricity brokers, we apply the concept of Virtual Power Plants (VPPs). VPPs are flexible portfolios of distributed energy resources, which are presented with a single load profile to the system operator, making them eligible for market participation and ancillary service provisioning (Pudjianto et al., 2007). Hence, VPPs allow providing regulation capacity to the market without knowing which exact sources provide the promised capacity until the delivery time (Kahlen et al., 2017). This concept is specially useful when dealing with EV fleets: VPPs enable carsharing providers to issue bids and asks based on an estimate of available fleet capacity, without knowing beforehand which exact EVs will provide the capacity at the time of delivery. Based on the battery charge and the availability of EVs, an intelligent agent decides in real-time which vehicles provide the capacity.

Centrally managed EV fleets make it possible for carsharing providers to use the presented concepts as a viable business extension. Free float carsharing is a popular concept that allows cars to be picked up and parked everywhere, and the customers are billed is by the minute. Free float carsharing offers flexibility to its users, saves resources, and reduces carbon emissions (Firnkorn & Müller, 2015). Most previous studies concerned with the usage of EVs for electricity trading, assumed that trips are fixed and known in advance, e.g., in Tomić and Kempton (2007). The free float concept adds uncertainty and nondeterministic behavior, which make predictions about future rentals a complex issue.

Kahlen et al. (2017) showed that it is possible to use free float carsharing fleets as VPPs to profitably offer balancing services to the grid. In their study, the authors compared cases from three different cities across Europe and the US. They used an event-based simulation, bootstrapped with real-world carsharing and secondary operating reserve market data from the respective cities. A central dilemma within their research was to decide whether an EV should be committed to a VPP or free for rent. Since rental profits are considerably higher than profits from electricity trading, it is crucial not to allocate an EV to a VPP when it could have been rented out otherwise. To deal with the asymmetric payoff, Kahlen et al. used stratified sampling in their classifier. This method gives rental misclassifications higher weights, reducing the likelihood of EVs to participate in VPP activities. The authors used a Random Forest regression model to predict the available balancing capacity on an aggregated fleet level. Only at the delivery time, the agent decides which individual EVs provide the regulation capacity. This heuristic is based on the likelihood that the vehicle is rented out and on its expected rental benefits.

In a similar study, the authors showed that carsharing companies can participate in day-ahead markets for arbitrage purposes (Kahlen et al., 2018). In the paper, the authors used a sinusoidal time-series model to predict the available trading capacity. Another central problem for carsharing providers is that committed trades, which can not be fulfilled, result in substantial penalties from the system operator or electricity exchange. In other words, fleet operators have to avoid buying any amount of electricity, which they can't be sure to charge with available EVs at the delivery time. To address this issue, the authors developed a mean asymmetric weighted (MAW) objective function. They used it for their time-series based prediction model, to penalize committing an EV to VPP when it would have been rented out otherwise. Because of the two issues mentioned above, Kahlen et al. (2018) could only make very conservative estimations and commitments of overall available trading capacity, resulting in a high amount of missed profits. This effect is especially prominent when participating in the secondary operating reserve market, since commitments have to be made one week in advance when mobility demands are still uncertain. Kahlen et al. (2017) stated that in 42% to 80% of the cases, EVs are *not* committed to a VPP when it would have been profitable to do so.

This thesis proposes a solution in which the EV fleet participates in the balancing market and intraday market simultaneously. With this approach, we align the potentially higher profits on the balancing markets, with more accurate capacity predictions for intraday markets (Tomić & Kempton, 2007). This research followed Kahlen et al. (2017), who proposed to work on a combination of multiple markets in the future.

2.4 Reinforcement Learning Controlled EV Charging

Previous research shows that intelligent agents equipped with Reinforcement Learning (RL) methods can successfully take action in the smart grid. The following chapter outlines different research approaches of RL in the domain of smart grids. For a more thorough description, mathematical formulations and common issues, of RL refer to Chapter 2.5.

Reddy and Veloso (2011a, 2011b) used autonomous broker agents to buy and sell electricity from DER on a proposed *Tariff Market*. The agents use Markov Decision Processes (MDPs) and RL to learn pricing strategies to profitably participate in the Tariff Market. To control for a large number of possible states in the domain, the authors used *Q-Learning* with derived state space features. Based on descriptive statistics, they defined derived price and market participant features. By engaging with its environment, the agent learns an optional sequence of actions (policy) based on the state of the agent. Peters et al. (2013) built on that work and further enhanced the method by using function approximation. Function approximation allows to efficiently learn strategies over large state spaces, by deriving a function that describes the states instead of defining discrete states. By using this technique, the agent can adapt to arbitrary economic signals from its environment, resulting in better performance than previous approaches. Moreover, the authors applied feature selection and regularization methods to explore the agent's adaption to the environment. These methods are particularly beneficial in smart markets because market design, structures, and conditions might change in the future. Hence, intelligent agents should be able to adapt to it (Peters et al., 2013).

Vandael et al. (2015) facilitated learned EV fleet charging behavior to optimally purchase electricity on the day-ahead market. Similarly to Kahlen et al. (2018), the problem is framed from the viewpoint of an aggregator that tries to define a cost-effective day-ahead charging plan in the absence of knowing EV charging parameters, such as departure time. A crucial point of the study is weighting low charging prices against costs that have to be paid when an excessive or insufficient amount of electricity is bought from the market (imbalance costs). Contrarily, Kahlen et al. (2018) did not consider imbalance cost in their model and avoid them by sacrificing customer mobility in order to balance the market (i.e., not showing the EV available for rent, when it is providing balancing capacity). Vandael et al. (2015) used a *fitted Q Iteration* to control for continuous variables in their state and action space. In order to achieve fast convergence, they additionally optimized the *temperature step* parameter of the Boltzmann exploration probability.

Dusparic et al. (2013) proposed a multi-agent approach for residential demand response. The authors investigated a setting in which 9 EVs were connected to the same transformer. The RL agents learned to charge at minimal costs, without overloading the transformer. Dusparic et al. (2013) utilized *W-Learning* to simultaneously learn multiple policies (i.e., objectives such as ensuring minimum battery charged or ensuring charging at low costs). Taylor et al. (2014) extended this research by employing Transfer Learning and *Distributed W-Learning* to achieve communication between the learning processes of the agents in a multi-objective, multi-agent setting. Dauer et al. (2013) proposed a market-based EV fleet charging solution. The authors introduced a double-auction call market where agents trade the available transformer capacity, complying with the minimum required State of Charge (SoC). The participating EV agents autonomously learn their bidding strategy with standard *Q-Learning* and discrete state and action spaces.

Di Giorgio et al. (2013) presented a multi-agent solution to minimize charging costs of EVs, a solution that requires neither prior knowledge of electricity prices nor future price predictions. Similar to Dauer et al. (2013), the authors employed standard *Q-Learning* and the ϵ-greedy approach for action selection. Vaya et al. (2014) also proposed a multi-agent approach, in which the individual EVs are agents that actively place bids in the spot market. Again, the agents use *Q-Learning*, with an ϵ-greedy policy to learn their optimal bidding strategy. The latter relies on the agents willingness-to-pay which depends on the urgency to charge. State variables, such as SoC, time of departure and price development on the market, determine the urgency to charge. The authors compared this approach with a centralized aggregator-based approach that they developed in another paper (Vaya & Andersson, 2015). Compared to the centralized approach, in which the aggregator manages charging and places bids for the whole fleet, the multi-agent approach causes slightly higher costs but solves scalability and privacy problems.

Shi and Wong (2011) consider a V2G control problem, while assuming real-time pricing. The authors proposed an online learning algorithm which they modeled as a discrete-time MDP and solved through *Q-Learning*. The algorithm controls the V2G actions of the EV and can react to real-time price signals of the market. In this single-agent approach, the action space compromises only charging, discharging and regulation actions. The limited action spaces makes it relatively easy to learn an optimal policy. Chis et al. (2016) looked at reducing the costs of charging for a single EV using known day-ahead prices and predicted next-day prices. A Bayesian ANN was employed for prediction and *fitted Q-Learning* was used to learn daily charging levels. In their research, the authors used function approximation and batch reinforcement learning, an offline, model-free learning method. Ko et al. (2018) proposed a centralized controller for managing V2G activities in multiple microgrids. The proposed method considers mobility and electricity demands of microgrids, as well as SoC of the EVs. The authors formulated a MDP with discrete state and action spaces and use standard *Q-Learning* with ϵ-greedy policy to derive an optimal charging policy. The approach takes microgrid autonomy and electricity prices into special consideration.

It should be noted that advanced RL methods and techniques are not the only solutions for problems in the smart grid, often basic algorithms and heuristics provide satisfactory results (Vázquez-Canteli & Nagy, 2019). Despite that, our paper considers RL as an optimal fit for the design of our proposed intelligent agent. Given the ability to learn user behavior (e.g., mobility demand) and the flexibility to adapt to the environment (e.g., electricity prices), RL methods are a promising way of solving complex challenges in smart grids.

2.5 Reinforcement Learning Theory

The following chapter will give an overview of the most important Reinforcement Learning (RL) concepts and will introduce the corresponding mathematical formulations. If not noted otherwise, the notation, equations, and insights are adopted from (Sutton & Barto, 2018), the de-facto reference book of RL research.

RL is an agent-based machine learning algorithm in which the agent learns to perform an optimal set of actions through interaction with its environment. The agents objective is to maximize the rewards it receives based on the actions it takes. Immediate rewards have to be weighted against long-term cumulative returns that also on its future actions. The RL problem is formalized as Markov Decision Processes (MDPs) which will be introduced in Chapter 2.5.1. A critical task of RL agents is to continuously estimate the value of the environments state. Values indicate the long-term desirability of a state, that is the total amount of reward the agent can expect to accumulate over the future, following a learned set of actions, called the policy. Policies and values are covered in Chapter 2.5.2, whereas the core mathematical foundations for evaluating policies and updating value functions are introduced in Chapter 2.5.3. When the model of the environment is fully known, the learning problem is reduced to a planning problem (Chapter 2.5.4) in which optimal policies can be computed with iterative approaches. Model-free RL approaches can be applied when rewards and state transitions are unknown, and the agent's behavior has to be learned from experience (Chapter 2.5.5). The last two chapter cover methods that solve the RL problem more efficiently, tackle new challenges and are widely used in practice and research.

2.5.1 Markov Decision Processes

Markov Decision Processes (MDPs) are a classical formulation of sequential decision making and an idealized mathematical formulation of the RL problem. MDPs allow to derive exact theoretical statements about the learning problem and possible solutions. Figure 2 depicts the *agent-environment interaction*.

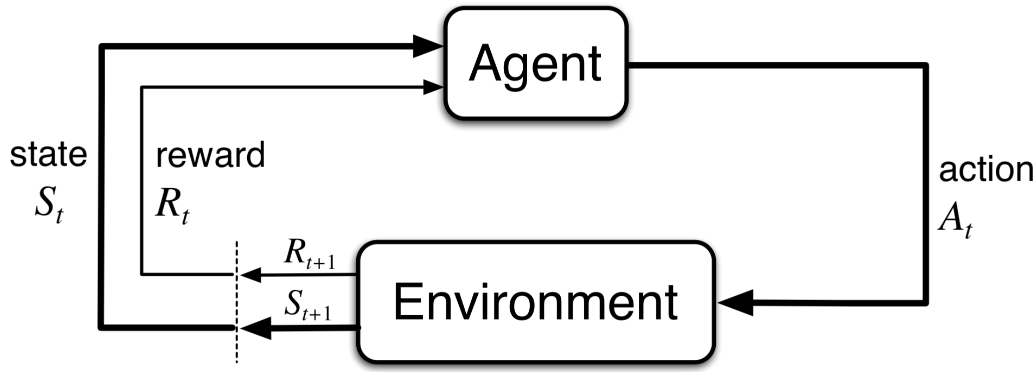

Figure 2: The agent-environment interaction in a Markov decision process (Sutton & Barto, 2018) [4]

In RL the agent and the environment continuously interact with each other. The agent takes actions that influence the environment, which in return presents rewards to the agent. The agent's goal is to maximize rewards over time, trough an optimal choice of actions. In each discrete timestep $t = 0, 1, 2, ..., T$ the RL agent interacts with the environment, which is perceived by the agent as a representation, called *state*, $S_t \in \mathcal{S}$. Based on the state, the agents selects an *action*, $A_t \in \mathcal{A}$, and receives a numerical *reward* signal, $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$, in the next timestep. Actions influence immediate rewards and successive states, and consequently also influence future rewards. The agent has to continuously make a trade-off between immediate rewards and delayed rewards to achieve its long-term goal.

The *dynamics* of a MDP are defined by the probability that a state $s' \in \mathcal{S}$ and a reward $r \in \mathcal{R}$ occurs, given the preceding state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$.

[4]**Figure 3.1** from "Reinforcement Learning: An Introduction" by Richard S. Sutton and Andew G. Barto is licencsed under CC BY-NC-ND 2.0 (https://creativecommons.org/licenses/by-nc-nd/2.0/)

In *finite* MDPs, the random variables R_t and S_t have well-defined probability density functions (PDF), which are solely dependent on the previous state and action. Consequently, it is possible to define ($\doteq$) the *dynamics* of the MDP as follows:

$$p(s', r|s, a) \doteq \Pr\{S_t = s', R_t = r|S_{t-1} = s, A_t = a\}, \tag{1}$$

for all $s', s \in \mathcal{S}$, $r \in \mathcal{R}$ and $a \in \mathcal{A}$. Note that each possible value of the state $\mathcal{S}_t$ depends only on the immediately preceding state $\mathcal{S}_{t-1}$. When a state includes all information of *all* previous states, the state possesses the so-called *Markov property*. If not noted otherwise, the Markov property is assumed throughout the whole chapter. The dynamics function allows computing the *state-transition probabilities*, another important characteristic of an MDP, as follows:

$$p(s'|s, a) \doteq \Pr\{S_t{=}s'|S_{t-1}{=}s, A_t{=}a\} = \sum_{r\in\mathcal{R}} p(s', r|s, a), \tag{2}$$

for $s', s{\in}\mathcal{S}$, $r{\in}\mathcal{R}$ and $a{\in}\mathcal{A}$.

The use of a *reward signal* R_t to formalize the agent's goal is a unique characteristic of RL. Each timestep the agent receives the rewards as a scalar value $\mathcal{R}_t \in \mathbb{R}$. The sole purpose of the RL agent is to maximize the long-term cumulative reward (as opposed to the immediate reward). The long-term cumulative reward can also be expressed as the *expected return* G_t:

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma R_{t+3} + \cdots \\ &= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \\ &= R_{t+1} + \gamma G_{t+1}, \end{aligned} \tag{3}$$

where γ, $0 \leq \gamma \leq 1$, is the *discount rate* parameter. The discount rate determines how "myopic" the agent is. If γ approaches 0, the agent is more concerned with maximizing immediate rewards. On the contrary, when $\gamma = 1$, the agent takes future rewards strongly into account, the agent is "farsighted".

2.5.2 Policies and Value Functions

An essential task of almost every RL agent is estimating *value functions*. These functions describe how "good" it is to be in a given state, or how "good" it is to perform an action in a given state. More formally, they take a state s or a state-action pair s, a as input and give the expected return G_t as output. The expected return is dependent on the actions the agent will take in the future. Consequently, value functions are formulated with respect to a *policy* π. A policy is a mapping of states to actions; it describes the probability that an agent performs a certain action, based on the current state. More formally, the policy is defined as $\pi(a|s) \doteq \Pr\{A_t = a | S_t = s\}$, a PDF of all $a \in \mathcal{A}$ for each $s \in \mathcal{S}$. RL approaches mainly differ in how the policy is updated, based on the agent's interaction with the environment.

In RL, value functions of states and value functions of state-action pairs are used. The *state-value function of policy* π is denoted as $v_\pi(s)$ and is defined as the expected return when starting in s and following policy π:

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t | S_t = s] \,, \text{ for all } s \in \mathcal{S} \tag{4}$$

The *action-value function of policy* π is denoted as $q_\pi(s,a)$ and is defined as the expected return when starting in s, taking action a and following policy π afterwards:

$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t | S_t = s, A_t = a] \,, \text{ for all } a \in \mathcal{A}, s \in \mathcal{S} \tag{5}$$

The *optimal policy* π_* has a greater (or equal) expected return than all other policies. The *optimal* state-value function and *optimal* action-value function are defined as follows:

$$v_*(s) \doteq \max_\pi v_\pi(s), \text{ for all } s \in \mathcal{S} \tag{6}$$

$$q_*(s,a) \doteq \max_\pi q_\pi(s,a), \text{ for all } s \in \mathcal{S}, a \in \mathcal{A} \tag{7}$$

The *optimal* action-value function describes the expected return when taking action a in state s following the optimal policy π_* afterwards. Estimating q_* to obtain an optimal policy is a substantial part of RL and has been known as *Q-learning* (Watkins & Dayan, 1992), which is described in Chapter 2.5.5.

2.5.3 Bellman Equations

A central characteristic of value functions is the recursive relationship between the values. Similar to Equation (3), current values are related to expected values of successive states. This relationship is heavily used in RL and has been formulated as *Bellman equations* (Bellman, 1957). The Bellman equation for $v_\pi(s)$ is defined as follows:

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi[G_t | S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) \Big[r + \gamma v_\pi(s') \Big], \end{aligned} \tag{8}$$

where $a \in \mathcal{A},\ s, s' \in \mathcal{S},\ r \in \mathcal{R}$. In other words, the value of a state equals the immediate reward plus the expected value of all possible successor states, weighted by their probability of occurring. $v_\pi(s)$ is the only solution to its Bellman equation. The Bellman equation of the optimal value function v_* is called the *Bellman optimality equation*:

$$\begin{aligned} v_*(s) &\doteq \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a) \\ &= \max_a \mathbb{E}_{\pi_*}[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\ &= \max_a \mathbb{E}_{\pi_*}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a] \\ &= \max_a \sum_{s',r} p(s', r|s,a) \Big[r + \gamma v_*(s') \Big] \end{aligned} \tag{9}$$

where $a \in \mathcal{A},\ s, s' \in \mathcal{S},\ r \in \mathcal{R}$. In other words, the value of a state under an optimal policy equals the expected return for the best action from that state. Note that the Bellman optimality equation does not refer to a specific policy, it has a unique solution independent from one. It can be seen as an equation system, which can be solved when the dynamics of the environment p are known. Similar Bellman equations to Equations (8) and (9) can also be formed for $q_\pi(s,a)$ and $q_*(s,a)$. Bellman equations form the basis for computing and approximating value functions and were an important milestone in RL history. Most RL methods are *approximately* solving the Bellman optimality equation, by using experienced state transitions instead of expected transition probabilities. The most common methods will be explored in the following chapters.

2.5.4 Dynamic Programming

Dynamic Programming (DP) is a method to compute optimal policies, the primary goal of every RL method. DP makes use of value functions to facilitate the search for good policies. Once an optimal value function, (i.e., one that satisfies the Bellman optimality equation) is found, optimal policies can be easily obtained. Despite the limited utility of DP in real-world settings, it provides the theoretical foundation for all other RL methods. In fact, all of the RL methods try to achieve the same goal, but without the assumption of a perfect model of the environment and less computational effort. Because DP assumes full knowledge of the environment, it is known as *planning*, in which optimal solutions are *computed*. In *control* problems (Chapter 2.5.5), optimal solutions are *learned* from an unknown environment.

The two most popular DP algorithms that compute optimal policies are called *policy iteration* and *value iteration*. These methods perform "sweeps" through the whole state set and update the estimated value of each state via an *expected update* operation. In policy iteration, a value function for a given policy v_π needs to be computed first, a step called *policy evaluation.* A sequence of approximated value functions $\{v_k\}$ are updated using the Bellman equation for v_π (Eq. 8) until convergence to v_π is achieved. After computing the value function for a given policy, it is possible to modify the policy and see if the value $v_\pi(s)$ for a given state increases (*policy improvement*). A way of doing this, is evaluating the action-value function $q_\pi(s, a)$ by *greedily* taking the best short-term action $a \in A$ at a given timestep. Alternating between these two steps monotonically improves the policies and the value functions until they converge to the optimum. This algorithm is called *policy iteration*:

$$\pi_0 \xrightarrow{\text{E}} v_{\pi_0} \xrightarrow{\text{I}} \pi_1 \xrightarrow{\text{E}} v_{\pi_1} \xrightarrow{\text{I}} \pi_2 \xrightarrow{\text{E}} \dots \xrightarrow{\text{I}} \pi_* \xrightarrow{\text{E}} v_*, \tag{10}$$

where $\xrightarrow{\text{E}}$ denotes a policy evaluation step, $\xrightarrow{\text{I}}$ denotes a policy improvement step. π_* and v_* are the optimal policy and optimal value function, respectively. Note that in each iteration of the policy iteration algorithm, a policy evaluation has to be performed, which requires multiple sweeps through the state space. In *value iteration*, the policy evaluation step is stopped after one sweep. In this case, the two previous steps can be combined into one single update step:

$$\begin{aligned} v_{k+1}(s) &\doteq \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a] \\ &= \max_a \sum_{s',r} p(s', r|s, a) \Big[r + \gamma v_k(s') \Big], \end{aligned} \tag{11}$$

where $a \in \mathcal{A}$, $s, s' \in \mathcal{S}$, $r \in \mathcal{R}$. It can be shown, that for any given v_0, the sequence v_k converges to the optimal value function v_*. In value iteration, the Bellman optimality equation (9) is simply turned into an update rule. Both of the algorithms can be effectively used to compute optimal values and value function in finite MDPs with a perfect model of the environment.

2.5.5 Temporal-Difference Learning

The previous chapter dealt with solving a *planning* problem, that is computing an optimal solution (i.e., an optimal policy π_*) of an MDP when a perfect model of the environment is known. In the following chapters, we will look at *model-free* prediction and *model-free* control. As opposed to planning, model-free methods learn from experience and require no prior knowledge of the environment. Remarkably, these methods can still achieve optimal behavior.

The *TD prediction problem* is concerned with estimating state-values v_π using past experiences of following a given policy π. TD methods update an estimate V of v_π in every timestep. At time $t+1$ they immediately perform an update operation on $V(S_t)$. Because of the step-by-step nature of TD learning, it is categorized as *online learning*. Also note that TD methods perform update operations on value estimates based on other learned estimates, a procedure called *bootstrapping*. In simple TD prediction, the value estimates V are updated as follows:

$$V(S_t) \leftarrow V(S_t) + \alpha\big[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\big], \tag{12}$$

where α is a constant *step-size* parameter and γ is the *discount rate*. Here, the update of the state-value is performed using the observed reward R_{t+1} and the estimated value $V(S_{t+1})$.

When a model is not available, it is useful to estimate *action-values*, instead of *state-values*. If the environment is completely known, it is possible for the agent to look one step ahead and select the best action. Without that knowledge, the value of each action in a given state needs to be estimated. The latter constitutes a problem, since not every *state-action* pair will be visited when the agent follows a deterministic policy. A deterministic policy $\pi(a|s)$ returns exactly one action given the current state, hence the agent will only observe returns for one of the actions. In order to evaluate the value function for all *state-action* pairs q_π, continuous *exploration* needs to be ensured. In other words, the agent has to explore state-action pairs which are seemingly disadvantageous given the current policy. This dilemma is also known as the *exploration-exploitation* trade-off. One way to achieve exploration is using *stochastic* policies for the action selection. Stochastic policies have a non-zero probability of selecting each action in each state. A typical stochastic policy is the *ϵ-greedy policy*, which selects the action with the highest estimated value, except for a probability ϵ, it selects an action at random.

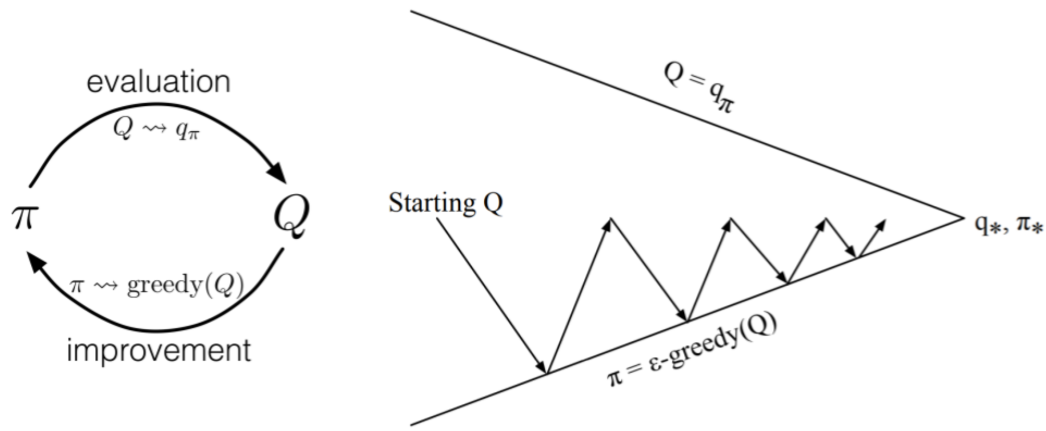

Figure 3: On-policy control with Sarsa (Sutton & Barto, 2018). [5]

There are two approaches to make use of stochastic policies to ensure all actions are chosen infinitely often. On-policy methods improve the (stochastic) decision policy, by continually estimating q_π in regard to π, while simultaneously driving π towards q_π, e.g., with a ϵ-greedy action selection. Figure 3 depicts this learning process. Off-policy methods improve the deterministic decision policy, by using a second stochastic policy to generate behavior. The first policy is becoming the optimal policy by evaluating the exploratory behavior of the second policy. Off-policy approaches are considered more powerful than on-policy approaches and have a variety of additional use cases. On the other side, they often have a higher variance and take more time to converge to an optimum.

A popular on-policy TD control method is Sarsa, developed by Rummery and Niranjan (1994). In the prediction step, the action-value function $q_\pi(s,a)$ of all actions and states has to be estimated for the current policy π. The estimation can be done similar to TD prediction of state values (Eq. 12). Instead of considering state transitions, state-action transitions are considered in this case. The update rule is constructed as follows:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\big[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\big] \tag{13}$$

After every transition from a state S_t, an update operation using the events $(S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1})$ is performed. This quintuple also constituted the name Sarsa. The on-policy control step of the algorithm is straightforward, and uses a ϵ-greedy policy improvement, as described in the previous paragraph. It has been shown that Sarsa converges to the optimal policy π_* under the assumption of infinite visits to all state-action pairs.

A breakthrough in RL has been achieved when Watkins and Dayan (1992) developed the *off-policy* TD control algorithm, called Q-learning. The update rule is defined as follows:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\big[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\big] \tag{14}$$

Here, the estimated action-values Q are updated towards the highest estimated action-value of the next time step. In this way, Q directly approximates the optimal action-value function q_*, independently of the policy the agent follows. Due to this simplification, Q-learning is a widely used model-free method, and its convergence can be proved easily.

This chapter covered the most important RL methods. They work online,

[5] The in-text figure of **Chapter 5.3** from "Reinforcement Learning: An Introduction" by Richard S. Sutton and Andew G. Barto is licencsed under CC BY-NC-ND 2.0 (https://creativecommons.org/licenses/by-nc-nd/2.0/)

learn from experience, and can be easily applied to real-world problems with low computational effort. Moreover, the mathematical complexity of the presented approaches is limited, and they can be easily implemented into computer programs. Temporal-Difference learning is a *tabular* method, in which Q-values are stored and updated in a lookup table. If the state and action spaces are continuous or the number of states and actions is very large, a table representation is computational infeasible and the speed of convergence is drastically reduced. In this case, a *function approximator* can replace the lookup table. The next chapter will briefly cover function approximation, as well as other advancements in RL.

2.5.6 Approximation Methods

Up to this point, only tabular RL methods have been covered, which form the theoretical foundation of RL in general. But in many real-world use cases, the state space is enormous and it is improbable to find an optimal value function with tabular methods. Not only is it a problem to store such a large table in the memory, but also would it take an almost infinite amount of time to fill every entry with meaningful results. Contrarily, *function approximation* tries to find a function that approximates the optimal value function as closely as possible, with limited computational resources. The experience with a small subset of visited states is generalized to approximate values of the whole state set. Function approximation has been widely studied in supervised machine learning: Gradient methods, as well as linear and non-linear models have shown good results for RL.

The approximated value of a state s is denoted as the parameterized functional form $\hat{v}(s, \mathbf{w}) \approx v_\pi(s)$, given a weight vector $\mathbf{w} \in \mathbb{R}^d$. Function approximation methods are approximating v_π by learning (i.e., adjusting) the weight vector $\mathbf{w}$ from the experience of following the policy π. By assumption, the dimensionality d of $\mathbf{w}$ is much lower than the number of states, which is the reason for the desired generalization effect: Adjusting one weight affects the values of many states. However, optimizing an estimate for one state negatively affects the accuracy of the estimates for other states. This effect motivates the specification of a state distribution $\mu(s)$, which represents the importance of the prediction error for each state. In on-policy prediction, $\mu(s)$ is often selected to be proportion of time spend in each state s. The prediction error of a state is defined as the squared difference between the predicted (i.e., approximated) value $\hat{v}(s, \mathbf{w})$ and the true value $v_\pi(s)$. Consequently, the objective function of the supervised learning problem can be defined as the *Mean Squared Value Error* $\overline{\mathrm{VE}}$, which weights the prediction error with the state distribution $\mu(s)$:

$$\overline{\mathrm{VE}}(\mathbf{w}) \doteq \sum_{s\in\mathcal{S}} \mu(s)\Big[v_\pi(s) - \hat{v}(s,\mathbf{w})\Big]^2, \text{ where } \mathbf{w} \in \mathbb{R}^d \tag{15}$$

Minimizing $\overline{\mathrm{VE}}$ in respect to $\hat{v}$ will yield a value function, which facilitates finding a better policy — the primary goal of RL. Remember that $\hat{v}$ can take any form of a linear or non-linear function of the state s.

In practice, deep artificial neural networks networks (ANNs) have shown great success as function approximators, which coined the term *Deep Reinforcement Learning* (Mnih et al., 2015; Silver et al., 2016). A simple feedforward ANN can be found in Figure 4. ANNs have the advantage that they can theoretically approximate any continuous function by adjusting the connection weights of the network $\mathbf{w} \in \mathbb{R}^{d\times d}$ (Cybenko, 1989). Advancements in the field of *Deep Learning* facilitated remarkable performance improvements in RL applications. Despite that, the RL theory is mostly limited to tabular and linear approximation methods. Refer to Bengio (2009) for a comprehensive review of deep learning methods.

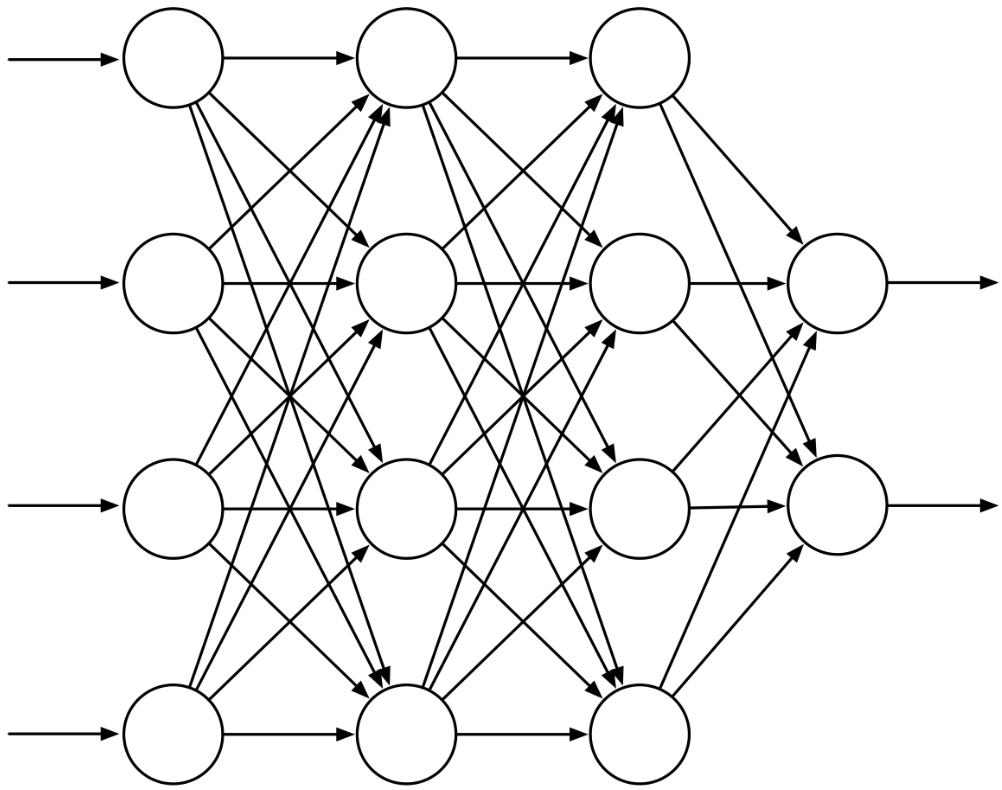

Figure 4: A sample ANN consisting of four input nodes, two fully connected hidden layers and two output nodes (Sutton & Barto, 2018). [6]

2.5.7 Further Topics

The previous chapters provided a detailed overview of the most important concepts and mathematical foundations in RL. In the research there are many more topics that were not covered here. *Eligibility traces* offer a way to more general learning and faster convergence rates. Almost any TD method can be extended to use eligibility traces, a popular methods is called Watkins's Q(λ) (Watkins, 1989). *Fitted-Q Iteration* (Ernst et al., 2003) combined Q-learning and fitted value iteration with batch-mode RL. In batch-mode the whole dataset is available offline, contrary to online RL where the data is acquired by the agent's action in its the environment. *Actor-critic* methods (Sutton, 1984) directly learn a parameterized policy instead of action-values, which inherently allow continuous state spaces and learning appropriate levels of exploration. Simultaneously to learning the policy, they approximate a state-value function, which serves as a "critic" to the learned policy, the "actor". In the current theory most RL models are single-agent models. For certain real-world applications multi-agent RL algorithms are necessary to coordinate interaction between the agents. When multiple learning agents interact with a non-stationary environment, convergence and stability are a serious problem. *W-learning* (Humphrys, 1996) is an multi-agent approach that aims to solve these difficulties.

[6] **Figure 9.14** from "Reinforcement Learning: An Introduction" by Richard S. Sutton and Andew G. Barto is licencsed under CC BY-NC-ND 2.0 (https://creativecommons.org/licenses/by-nc-nd/2.0/)

3 Empirical Setting

This research is embedded in the German carsharing and electricity markets. Germany is a suitable testbed, since it has a comparably high share of renewables in its energy mix and is pushing for an energy turnaround (German: *Energiewende*) since 2010 (BMU, 2010) The high renewable energy content in the energy mix causes electricity prices to be volatile, which makes Germany an attractive location for the use of VPPs.

Germany is home to the carsharing providers Car2Go[7] and DriveNow[8], which operate large EV fleets across the globe. It has been argued that electric carsharing can simultaneously solve several traditional mobility and environmental problems and are an important element of future smart cities (Firnkorn & Müller, 2015). Further, it is widely regarded that the future of mobility will be electric, shared, smart and eventually autonomous (Burns, 2013; Sterling, 2018). Carsharing providers are already contributing to the first two points by operating large fleets of electric vehicles. This research addresses the third point: Using electric carsharing fleets to smartly participate in electricity markets. Carsharing providers, like Car2Go and DriveNow, operate their carsharing fleets in a free-float model, which allows customers to pick up and drop vehicles at any place within the operating zone of the provider. Customers pay by the minute and are offered incentives to park the EVs at charging stations at the end of their trip.

We obtained real-world trip data from Daimler's carsharing service Car2Go. Additionally, we collected freely available balancing market data from the GCRM platform website `https://regelleistung.net`. The data of the EPEX Spot market have kindly been provided by ProCom GmbH[9] for research purposes. In the next chapters the different datasets are described, as well the most important processing steps outlined.

3.1 Electronic Vehicle Fleet Data

The Car2Go dataset consists of GPS data of around 500 Smart ED3 Fortwo vehicles in Stuttgart. These subcompact cars are equipped with a 17.6kWh battery and a standard 3.3kW on-board charger. They fully charge in about six to seven hours and can reach a maximum driving distance of 145km according to the manufacturer. When equipped with an additional 22kW fast charger the charging time reduces to about an hour.

In Table 1 the raw data is displayed, as we have obtained it by Car2Go. The

[7] `https://www.car2go.com`
[8] `https://www.drive-now.com`
[9] `https://procom-energy.de`

dataset contains spatio-temporal attributes, such as timestamp, coordinates, and the address of the EVs in 5 minute intervals. Additionally, status attributes of the interior and exterior are given (not displayed). Especially relevant for our research is the state of charge (SoC, in %) and information whether the EV is plugged into one of the 380 charging stations in Stuttgart. Note that the data only contain EVs that are *available for rent*, i.e., they are not currently rented out by a customer. EVs which are parked at a charging station are also not available until they have charged up approximately 70 SoC. Individual trips have to be reconstructed using the GPS data of the cars. The following preprocessing steps have been taken to prepare the data for further analysis. Table 2 depicts the dataset after all processing steps.

1. Drop unused data columns

    - *ID*: Number plate is already a unique identifier for every EV.
    - *Address*: Different addresses were given from same coordinates. *Latitude, Longitude* was used for locational data instead.
    - *Interior, Exterior*: Status attributes were not used in the analysis of this research. Although they could form interesting features for rental predictions.
    - *Engine Type*: All EVs in Stuttgart are electric vehicles.

2. Decrease GPS resolution to 10 meters

    The GPS accuracy of private industry sensors is approximately 5 meters under open sky, and worse near buildings, bridges and trees[10]. Rental trips are identified by changing GPS locations of the EV (See next point). To reduce the number of false identified trips, due to GPS measurement errors, the resolution is decreased.

3. Determine rental trips

    We infer that a customer rented an EV, if the position coordinates change between two data points of the same EV (see Table 1, 4th to 5th row). Note that we assume that customer do not undertake trips, which begin and end at the exact same location.

4. Infer charging stations

    The GPS location of the EVs is matched with the GPS locations, where an EV has been charged at least once in the dataset. We observed that

[10] See `https://www.gps.gov/systems/gps/performance/accuracy`, accessed 23th February 2019.

the raw data do not show EVs that are parked at charging stations, but are not plugged in. This research assumes that all EVs, which are parked at charging stations are also plugged in. That is a valid assumption, since in Germany cars are only allowed to park at charging station if they are connected to it.

5. Clean data

   - *Service trips*: 999 rental trips were removed that had a trip duration longer than the maximum allowed rental time of two days. We assume that these trips were *service trips* undertaken by Car2Go. When the EVs returned with a higher SoC (e.g., they have been charged at the car repair shop), the previous trip had to be altered to end at a charging station to ensure charging consistency.
   - *Incorrectly charged EVs*: 999 EVs were removed that show incorrect charging behavior. The data of these EVs showed an increase of more than 20% SoC between trips or on trips, while not being located at a charging station.

3.2 Balancing Market Data

In this research, we use market balancing data from the German secondary reserve market. The following chapter will give an overview of the dataset and preprocessing steps that were taken. The data encompasses weekly lists of anonymized bids between 01.06.2016 and 01.01.2018 and a dataset of activated control reserve in Germany during the same period. For a detailed description about the market design of balancing markets refer to Chapter 2.2.1.

The bidding data consists of the traded electricity product, the offered capacity P^{bal} (MW), the capacity price p^c ($\frac{€}{\mathrm{MW}}$), and the energy price p^e ($\frac{€}{\mathrm{MWh}}$) of each bid. Four different products are traded, which are a combination of positive control reserve (feed electricity into the grid) or negative control reserve (take electricity from the grid) and the provided time segment (peak or non-peak hours). Since negative prices are allowed on the secondary operating reserve market, the payment direction is included as well. Moreover, information about the amount of capacity that was accepted, i.e., either partially or fully, is listed. Bids, which were not accepted by the TSOs are not listed. An exemplary excerpt of the dataset is displayed in Table 3.

Table 1: Sample Raw Car2Go Data in Stuttgart

Number Plate	Timestamp	Latitude	Longitude	Street	Zip Code	Charging	SoC (%)
S-GO2471	24.12.2017 20:00	9.19121	48.68895	Parkplatz Flughafen	70692	no	94
S-GO2471	...	...	...	...	...	....	...
S-GO2471	24.12.2017 20:05	9.19121	48.68895	Parkplatz Flughafen	70692	no	94
S-GO2471	24.12.2017 20:10	9.19121	48.68895	Parkplatz Flughafen	70692	no	94
S-GO2471	24.12.2017 23:05	9.15922	48.78848	Salzmannweg 3	70192	no	71
S-GO2471	24.12.2017 23:10	9.15922	48.78848	Salzmannweg 3	70192	no	71
S-GO2471	25.12.2017 00:40	9.17496	48.74928	Felix-Dahn-Str. 45	70597	yes	62
S-GO2471	25.12.2017 00:45	9.17496	48.74928	Felix-Dahn-Str. 45	70597	yes	64
S-GO2471	...	...	...	...	...	....	...
S-GO2471	25.12.2017 06:50	9.17496	48.74928	Felix-Dahn-Str. 45	70597	no	100
S-GO2471	25.12.2017 08:25	9.2167	48.78742	Friedenaustraße 25	70188	no	42

Table 2: Sample Processed Car2Go Trip Data in Stuttgart

Number Plate	Trip	Start Time	Start Latitude	Start Longitude	Start SoC (%)
S-GO2471	1	24.12.2017 20:10	9.19121	48.68895	94
S-GO2471	2	24.12.2017 23:10	9.15922	48.78848	71
S-GO2471	3	25.12.2017 00:50	9.17496	48.74928	66
Number Plate	**Trip**	**End Time**	**End Latitude**	**End Longitude**	**End SoC (%)**
S-GO2471	1	24.12.2017 23:05	9.15922	48.78848	71
S-GO2471	2	25.12.2017 00:40	9.17496	48.74928	62
S-GO2471	3	25.12.2017 03:25	9.2167	48.78742	42
Number Plate	**Trip**	**Trip Duration (min)**	**Trip Distance (km)**	**Trip Charge (%)**	**End Charging**
S-GO2471	1	175	33.35	23	no
S-GO2471	2	90	13.05	9	yes
S-GO2471	3	155	29	20	no

Table 3: List of Bids of the German Secondary Reserve Market for the tender period 04.12.2017 - 11.12.2017.

Product	Capacity Price	Energy Price	Payment	Offered	Accepted
NEG-HT	0	1.1	TSO to bidder	5	5
NEG-HT	10.73	251	TSO to bidder	15	15
NEG-HT	200.3	564	TSO to bidder	22	22
...	...	...	...	...	...
NEG-NT	0	21.9	Bidder to TSO	5	5
NEG-NT	0	22.4	Bidder to TSO	5	5
...	...	...	...	...	...
POS-NT	696.6	1200	TSO to bidder	5	5
POS-NT	717.12	1210	TSO to bidder	10	7

In this study, we assume that bidding on 15-minute intervals in secondary operating reserve auctions will be possible in future energy markets. As mentioned in Chapter 2.2.1, the market design of the GCRM secondary operating reserve tender was adjusted in 2017. Daily tenders with 4-hour bidding intervals were introduced in favor of weekly tenders with only two time segments. This change represents the trend by the TSOs to change the market design in order to better include RES into the operating reserve markets (Agricola et al., 2014). Due to the volatility of renewable electricity generation, providers are naturally dependent on accurate short-term forecasts, which are only possible with short tender periods and fine-grained bidding intervals.

In order to estimate the upper bound of profits that the EV fleet can earn by participating in the secondary operating reserve market, the *critical prices* $\overline{p}^c$ and $\overline{p}^e$ were determined for each auctioned interval. Following Brandt et al. (2017), we define $\overline{p}^c$ ($\frac{€}{\text{MW}}$) as the capacity price of the bid that was just barely accepted, whereas $\overline{p}^e$ ($\frac{€}{\text{MWh}}$) is the highest energy price that was payed for activated control reserve during that interval. For every 15-minute interval within the given tender period of one week, the activated control reserve in that interval was matched with the accepted bids in that tender period. At the point where supply, i.e., offered capacity of bids, met demand, i.e., activated control reserve, the critical price $\overline{p}^e$ was determined.

Example: The assumed critical prices for the secondary operating reserve tender interval of the 6$^{\text{th}}$ December 2017 between 08:00 and 08:15 are obtained as follows: Three suppliers submitted a reserve capacity of 5MW, 15MW and 22MW respectively (see Table 3). The critical capacity price $\overline{p}^c = 200.3\frac{€}{\text{MW}}$ is determined by the capacity price of the last (third) accepted bid in that time segment. The TSO reported that 18MW of control reserve were activated between 08:00 and 08:15. Hence, the second bid determines the critical energy price $\overline{p}^e = 251\frac{€}{\text{MWh}}$, as control reserve capacity gets activated according to ascending order of the submitted energy prices. In this example the second bidder would get compensated with: $R = R^c + R^e = (10.73\frac{€}{\text{MW}} \times 15\text{MWh}) + (251\frac{€}{\text{MWh}} \times 13\text{MWh} \times 0.25\text{h}) = 976.7€$. Note that the second bidder get compensated for providing 13MW for 15 minutes (0.25h), instead of the submitted 15MW, since in total only 18MW of control capacity was activated, which was partly fulfilled by the 5MW of the first bidder.

3.3 Spot Market Data

The data from the EPEX Spot Intraday Continuous encompass order books and executed trades from 01.06.2016 until 01.01.2018. The list of trades contain information on the unit price p^u ($\frac{€}{\text{MWh}}$), the quantity (kW) and the traded product (hourly, quarterly or block). In this research, we focus on quarterly product times (15-minute intervals), as they provide the highest flexibility. Fleet controllers can promptly react to fluctuant electricity demand of the EV fleet by accurately adjusting the bid quantities. Future research could also consider other electricity products if lower prices justify decreased flexibility at that point in time. Additionally, the TSOs of the buyer and seller are listed in the dataset. They are only relevant if special conditions between TSO apply, e.g., when delivering electricity to other countries.

On the spot market, electricity trades can have a very short lead time of up to 5 minutes before delivery (see Table 4). This market characteristic is beneficial for our proposed trading strategy, since it allows the EV to procure electricity in almost real time. The controller can submit bids to the market, with accurate estimations of available charging capacity up to five minute-ahead. Similarly to the balancing market, the critical price $\overline{p}^u$ has to be determined for all intraday trading intervals. $\overline{p}^u$ is defined as the lowest unit price of all executed trades $\mathcal{T}$ in a bidding interval:

$$\overline{p}^u \doteq \min_{t \in \mathcal{T}} p^u_t$$

Note that $\overline{p}^u$ is essentially the optimal bidding price, it provides on upper bound for the fleets profits when bidding on the intraday market. We assume that bids with $p_u \geq \overline{p}^u$ will always get matched. In reality this is not always the case, since trades are executed immediately and it is not guaranteed that unmatched asks at $\overline{p}^u$ exist at the time of commitment. For a detailed description of the intraday continuous market see Chapter 2.2.2.

Table 4: List of Trades of the EPEX Spot Intraday Continuous Market

Execution time	ID	Unit Price	Quantity	Buyer Area	Seller Area	Product	Product Time	Delivery Date
2017-12-04 06:54:55	8031392	51.00	5500	Amprion	Amprion	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:53:26	8031391	59.00	10000	TenneT	TenneT	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:53:26	8031390	58.90	10000	TenneT	TenneT	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:53:15	8031389	52.30	7000	50Hertz	50Hertz	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:53:13	8031386	59.00	500	TenneT	TenneT	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:53:13	8031387	51.00	3600	Amprion	Amprion	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:53:13	8031388	52.00	1400	Amprion	Amprion	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:53:02	8031385	58.90	11000	TenneT	TenneT	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:52:38	8031380	60.00	10000	Amprion	Amprion	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:52:38	8031381	57.50	8000	Amprion	Amprion	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:52:38	8031382	58.00	2000	Amprion	Amprion	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:52:38	8031383	58.90	4000	TenneT	TenneT	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:52:38	8031384	60.00	4000	Amprion	Amprion	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:52:27	8031379	52.30	8000	50Hertz	50Hertz	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:51:33	8031378	66.00	5000	TransnetBW	TransnetBW	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:51:28	8031377	54.00	8000	Amprion	Amprion	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:51:24	8031376	54.00	7000	TenneT	TenneT	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:49:34	8031375	51.00	4000	TenneT	TenneT	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:49:26	8031374	54.00	5000	50Hertz	50Hertz	Q	07:15 - 07:30	2017-12-04
2017-12-04 06:49:23	8031373	55.10	8000	50Hertz	50Hertz	Q	07:15 - 07:30	2017-12-04

4 Model: FleetRL

The following chapter will introduce the model of this research. In its essence, we propose a solution for EV fleet providers to utilize a VPP portfolio to profitably provide balancing services to the grid on multiple markets. A control mechanism procures energy from electricity markets, allocates available EVs to VPPs, and intelligently dispatches EVs to charge the acquired amount of energy. The model uses a RL agent that learns an optimal bidding strategy by interacting with the electricity markets and reacts to changing rental demand of the EV fleet. This chapter is structured as follows: The information assumptions are listed first, the control mechanism is explained next, and finally the RL approach is described in detail. For a table of notation refer to Table 5.

We formulate the problem as a *controlled EV charging* problem. The EV fleet operator represents the *controller*, which aims to charge the fleet at minimal costs. First, the controller predicts the amount of energy it can charge in a given *market period* h. The length of the market period Δh and the market closing time depend on the considered electricity market. Second, the controller places bids on one or multiple markets to procure the predicted amount of energy. Lastly, at electricity delivery time, the controller communicates with the EV fleet to control the charging in real-time. Online EV *control periods* t are typically shorter than market periods. In the empirical case that we consider, the market periods are 15 minutes long, while the EV control periods last 5 minutes. Nonetheless, the presented approach generalizes to other period lengths. During each control period, the controller has to take decisions which individual EVs it should dispatch to charge the procured amount of electricity. In times of unforeseen rental demand, this decision implies trading off commitments to the markets with compromising customer mobility by refusing customer rentals.

Table 5: Table of Notation

Symbol	Description	Unit
t	Control period.	-
h	Market period.	-
T	Number of control periods in a market period.	-
H	Number of market periods in day.	-
N_h	Total number of market periods.	-
Δt	Length of control period.	hours
Δh	Length of market period.	hours
P_h^{bal}	Amount of balancing power offered on the balancing market.	kW

Continued on next page

Continued from previous page

Symbol	Description	Unit
p_h^c	Capacity price offered on the balancing market.	$\frac{€}{\text{MW}}$
p_h^e	Energy price offered on the balancing market.	$\frac{€}{\text{MWh}}$
$\overline{p}_h^c$	Critical capacity price in market period h.	$\frac{€}{\text{MW}}$
$\overline{p}_h^e$	Critical energy price in market period h.	$\frac{€}{\text{MWh}}$
P_h^{intr}	Amount of power offered for the unit on the intraday market.	kW
p_h^u	Unit price offered on the intraday market.	$\frac{€}{\text{MWh}}$
$\overline{p}_h^u$	Critical unit price in market period h.	$\frac{€}{\text{MWh}}$
E_h^{bal}	Amount of energy charged from balancing market in market period h.	MWh
E_h^{intr}	Amount of energy charged from the intraday market in market period h.	MWh
p^i	Industry tariff	$\frac{€}{\text{kWh}}$
P_t^{fleet}	Amount of available fleet charging power in control period t.	kW
$\widehat{P}_t^{fleet}$	Predicted amount of available fleet charging power in control period t.	kW
$C_h^{bal}(P)$	Cost function for procuring electricity from the balancing market.	€
$C_h^{intr}(P)$	Cost function for procuring electricity from the intraday market.	€
$\rho_{t,i}$	Opportunity costs of lost rental of EV i in control period t.	€
β_h	Imbalance costs in market period h.	€
λ_h^{bal}	Balancing market risk factor.	-
λ_h^{intr}	Intraday market risk factor.	-
θ_λ	Set of risk factors for all market periods $h \in \{1, ..., N_h\}$.	-
$C^{fleet}(\theta_\lambda)$	Cost function for the fleets total costs over all market periods h.	€
C_h^{fleet}	Total accumulated fleet costs at market period h.	€
i	Electric Vehicle.	-
$\mathbf{c}_i$	Dummy variable if EV is connected to a charging station.	0/1
ω_i	Amount of electricity stored in EV.	kWh
Ω	Maximum battery capacity of EV.	kWh
δ	Charging power of EV at the charging station.	kW
$\mathcal{F}$	Set of all EVs in the fleet	-

Continued on next page

Continued from previous page

Symbol	Description	Unit
$\lvert\mathcal{F}\rvert$	Total number of EVs in the fleet.	-

4.1 Required Information Assumptions

The following information is assumed to be available:

1. The controller is able to forecast the mobility demand of the EV fleet at different time-horizons based on historical data. More specifically, it can predict the amount of plugged-in EVs and consequently the available charging power P_t^{fleet} of the fleet at control period t. The prediction's accuracy is increasing with shorter time horizons, from uncertain predictions one week ahead to very accurate predictions 30 minutes ahead. Past research presented successful mobility demand forecast algorithms in the context of free-float carsharing (Kahlen et al., 2018, 2017; Wagner et al., 2016).

2. The controller is able to forecast electricity prices of spot and balancing markets based on historical data. More specifically, it can estimate the critical prices $\overline{p}_h^c$, $\overline{p}_h^e$, and $\overline{p}_h^u$ for each market period with perfect accuracy. The critical prices form an essential piece of information for the proposed bidding strategy; bids equal or below the critical price will get accepted and result in successful electricity procurement. Electricity price forecasting is an extensively studied research area, with well-advanced prediction algorithms (Weron, 2014; Avci et al., 2018).

We are confident that taking the above assumptions is viable, assuming available forecasting information is common practice in the VPP and EV fleet charging literature, see e.g.: Brandt et al. (2017); Vandael et al. (2015); Mashhour and Moghaddas-Tafreshi (2011b); Tomić and Kempton (2007); Pandžić et al. (2013).

4.2 Control Mechanism

The central control mechanism constitutes the core of this research. It can be seen as a decision support system that can be deployed at a EV fleet operator to control the charging its fleet. Figure 5 depicts the control mechanism, which is divided into three distinct phases:

The first phase, *Bidding Phase I*, takes place just before the closing time of the balancing market, once every week (e.g., Wednesdays at 3pm at the GCRM). In this phase, the controller can place bids for every market period h of the following week on the balancing market. The second phase, *Bidding Phase II*, takes places in every market period of $\Delta h = 15$ minutes. At this point, the controller has the opportunity to place bids for the market period 30 minutes ahead. By submitting bids 30 minutes ahead of time, the controller assures that the bid will be matched until the lead time of the market (e.g, 5 minutes on the EPEX Spot Intraday Continuous). The third phase, *Dispatch Phase*, takes places in every control period of $\Delta t = 5$ minutes. In this phase the controller has to dispatch available EVs to charge the procured electricity from the markets. This phase involves allocating individual EVs to the VPP and eventually refusing customer rentals to assure that all commitments can be fulfilled.

The following chapters will highlight the important parts of the various phases and provide detailed explanation and mathematical formulations.

4.2.1 Fleet Charging Power Prediction

In a first step, the controller has to predict the available fleet charging power $\widehat{P}_h^{fleet}$ for all market periods h of the next week. The actual available fleet charging power P_t^{fleet} in a control period t is given by the number of EVs that are connected to a charging station, with enough free battery capacity to charge the next control period t+1. As mentioned in the Chapter 4.1, the controller is able to predict the available fleet charging power $\widehat{P}_t^{fleet}$ for all control periods t with different levels of accuracy dependent on the time horizon of t.

When the controller procures electricity from the markets, the fleet has to charge with the committed charging power during all T control periods of the market period h. To minimize the risk of not being able to charge the committed amount of energy during the whole market period, and consequently causing imbalance costs, the predicted fleet charging power in a market period is defined as the minimal predicted fleet charging power of all T control periods in a market period.

$$\widehat{P}_h^{fleet} \doteq \min_{n \in \{1,..,T\}} \widehat{P}_{t+n}^{fleet} \ , \tag{16}$$

where h is the market period of interest and t its first control period.

4.2.2 Market Decision

In a second step, the controller has to decide from which market it should procure the desired amount of energy. Therefore, it compares the costs for charging electricity from the balancing market and the intraday market. The cost function

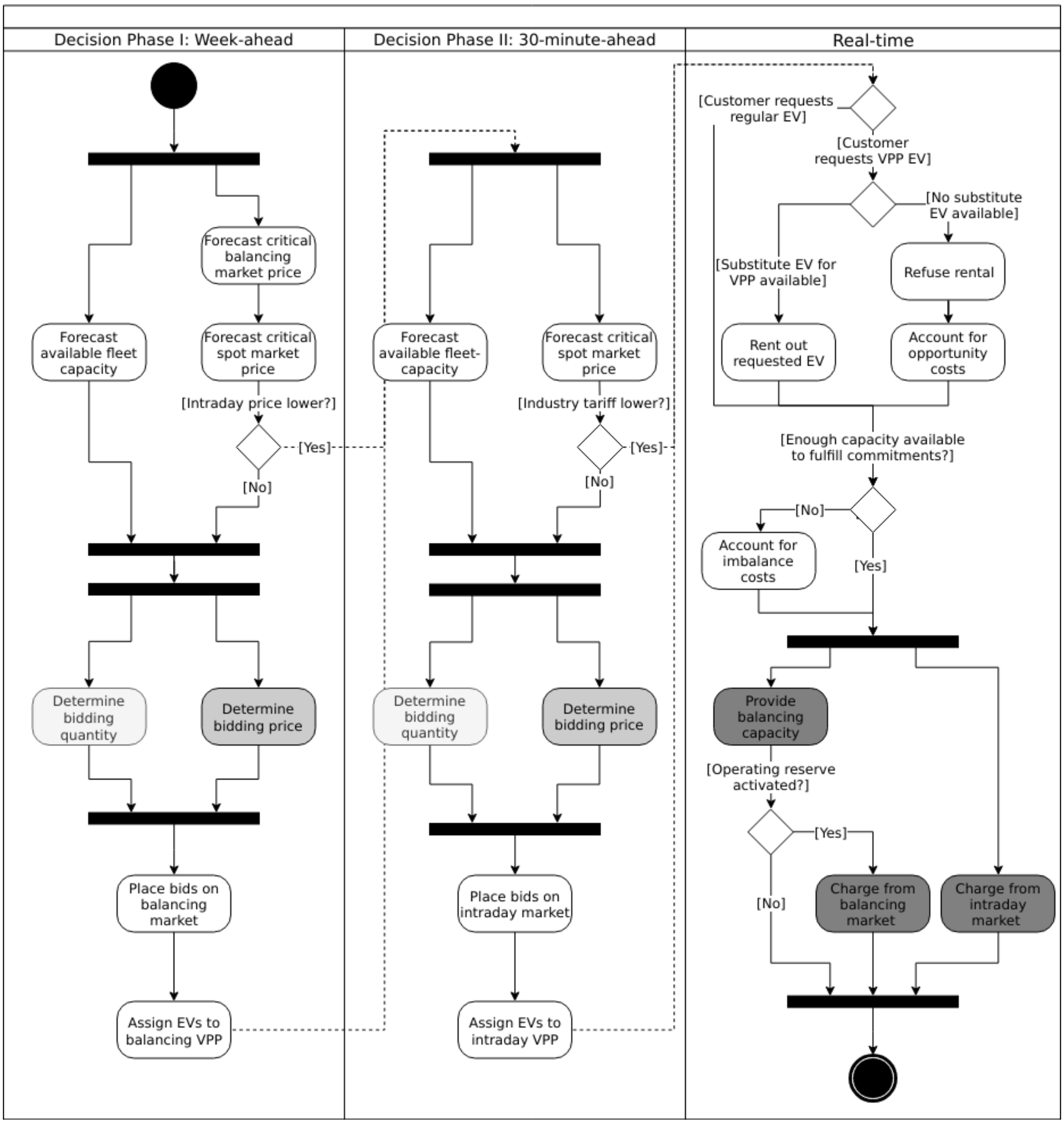

Figure 5: EV VPP Control Mechanism

for charging electricity from the balancing market is defined as follows:

$$\begin{aligned} C_h^{bal}(P) &\doteq -(P \times 10^{-3} \times \overline{p}_h^c) + (E_h^{bal} \times \overline{p}_h^e) \\ &= -(P \times 10^{-3} \times \overline{p}_h^c) + (P\frac{\Delta h}{10^3} \times \overline{p}_h^e) \ , \end{aligned} \tag{17}$$

where P (kW) is the amount of offered balancing power. The first term of the equation corresponds to the compensation the controller retrieves for keeping the balancing capacity available, while the second term corresponds to the costs for charging the activated balancing energy E_h^{bal} (MWh). Energy is power over time, hence E_h^{bal} can be substituted with P times the market periods length Δh, divided by the unit conversion from kW to MW. As mentioned in the Chapter 4.1, the critical prices $\overline{p}^c, \overline{p}^e, \overline{p}^u$ are assumed to be available for all market periods. Note that the critical energy price $\overline{p}^e \in \mathbb{R}$, can also take negative values, resulting in profits for the fleet, while the critical capacity price $\overline{p}^c \in \mathbb{R}_0^+$ can not take negative values and therefore always results in profits for the fleet.

The cost function for charging from the intraday market is defined similarly:

$$\begin{aligned} C_h^{intr}(P) &\doteq E_h^{intr} \times \overline{p}_h^u \\ &= P\frac{\Delta h}{10^3} \times \overline{p}_h^u \end{aligned} \tag{18}$$

Again, depending on the market situation, $\overline{p}^u \in \mathbb{R}$ can be either negative or positive, resulting in costs or profits for the fleet. Contrarily to the balancing market, on the intraday market the fleet does not get compensated for keeping the charging power available; only the charged energy affects the costs. If the costs for charging from the balancing market 7 days ahead $C_{h+(7\times H)}^{bal}(\widehat{P}_{h+(7\times H)}^{fleet})$ are higher than the costs of charging from the intraday market at the same market period $C_{h+(7\times H)}^{intr}(\widehat{P}_{h+(7\times H)}^{fleet})$, the controller does not place bids on the balancing market.

4.2.3 Determining the Bidding Quantity

In a third step, the controller has to take a decision on the amount of energy it should procure from the markets. Determining the bidding quantity is a central piece of the controlled charging problem. The bidding quantity determines the profits that can be made, by charging at a cheaper market price than the flat industry tariff. In order to maximize its profits, the controller aims to procure as much electricity as possible from the markets. In order to optimally place bids on the electricity markets, it needs to balance the risk of (a) procuring more energy that it can maximally charge and (b) not procuring enough energy from the market to sufficiently charge the fleet.

In the first case (a), the fleet is facing costs of compromising customer mobility, or worse, high imbalance penalties from the markets. Renting out EVs is considerably more profitable than using EVs as a VPP to participate on the electricity markets. Refusing customer rentals, in order to fulfill market commitments, induces opportunity costs of lost rentals ρ on the fleet. Imbalance costs β occur, when the fleet can not charge the committed amount energy at all, even with refusing rentals. In the second case (b), the fleet also faces opportunity costs of lost rentals when individual EVs do now have enough SoC for planned trips of arriving customers.

The controller faces additional risks by bidding one week ahead on the balancing market, in contrast to only 30 minutes ahead on the intraday market, as the predictions of available charging power are more uncertain with the larger time horizon. To account for all the mentioned risks, we introduce a *risk factor* $\lambda \in \mathbb{R}_{0\leq\lambda\leq1}$, where $\lambda = 0$ indicates no risk, and $\lambda = 1$ indicates a high risk. The controller determines the bidding quantity P_h^{bal} by discounting the predicted available fleet charging power $\widehat{P}_h^{fleet}$ with the possible risk λ_h of imbalance or opportunity costs:

$$P_h^{bal} \doteq \begin{cases} 0, & \text{if } C_h^{bal}(\widehat{P}_h^{fleet}) \geq E_h^{bal} 10^3 \times p^i \\ 0, & \text{if } C_h^{bal}(\widehat{P}_h^{fleet}) \geq C_h^{intr}(\widehat{P}_h^{fleet}) \\ \widehat{P}_h^{fleet} \times (1-\lambda_h^{bal}), & \text{otherwise} \end{cases} \tag{19}$$

where h is the market period of interest one week ahead. If the controller can buy electricity at the intraday market at a lower price, it does not place a bid at the balancing market. If the controller can charge cheaper at the regular industry tariff p^i, it does not place a bid either. In all other cases, the controller submits P_h^{bal} to the market.

The bidding quantity for the intraday market P_h^{intr} depends on the previously committed charging power P_h^{bal} and the newly predicted charging power $\widehat{P}_h^{fleet}$:

$$P_h^{intr} \doteq \begin{cases} 0, & \text{if } C_h^{intr}(\widehat{P}_h^{fleet} - P_h^{bal}) \geq E_h^{intr} 10^3 \times p^i \\ (\widehat{P}_h^{fleet} - P_h^{bal}) \times (1-\lambda_h^{intr}), & \text{otherwise} \end{cases} \tag{20}$$

where h is the market period of interest 30 minutes ahead. Note that any amount of electricity that the controller procured from the balancing market, does not need to be bought from intraday market for the same market period. Since the predicted charging power $\widehat{P}_h^{fleet}$ is expected to be more accurate 30 minutes ahead than one week ahead, the controller is able to correct bidding errors it made in the first decision phase, and optimally charge the whole EV fleet.

4.2.4 Dispatching Electronic Vehicle Charging

In the last step, at electricity delivery time, the EVs have to be assigned to the VPP and be *dispatched* to charge. Therefore the controller first needs to detect how many EVs are eligible to be used as VPP per control period t. EVs are eligible if they (a) are connected to a charging station ($\mathbf{c}_i = 1$), and (b) have enough free battery storage available ($\Omega - \omega_i$) to charge the next control period. Hence, the VPP is defined as:

$$VPP \doteq \{i \in \mathcal{F} \mid \mathbf{c}_i = 1 \vee \Omega - \omega_i \geq \gamma \Delta t\} \ , \tag{21}$$

where $\gamma \Delta t$ (kWh) denotes the amount of energy that can be charged with the charging speed of γ (kW) in control period t. γ is limited by either the EVs build-in charger, or the charging power of the connected charging station. In this model we assume γ is equal for all considered EVs and charging stations.

Remember that the fleet has to provide the committed charging power $P_h^{bal} + P_h^{intr}$ across all control periods t of the market period h, independent of which individual EVs are actually charging the electricity. This fact allows the controller to dynamically dispatch EVs every control period and react to unforeseen rental demand. If a customer want to rent out an EV that is assigned to the VPP, the controller only has to refuse the rental, if no other EV is available to charge instead. When no replacement EV is available, the controller has to account for lost rental profits $\rho_{t,i}$. If the VPPs total amount of available charging power $|VPP|_t \times \gamma$ is not sufficient to provide the total market commitments $P_h^{bal} + P_h^{intr}$, the fleet gets charged imbalance costs β_h. Otherwise all the committed energy can be charged by the VPP.

4.2.5 Evaluating the Bidding Risk

The controllers central goal is to choose the risk factors λ_h^{bal}, λ_h^{intr} for every market period h, that minimize the cost of charging, while avoiding the risks of lost rental profits $\rho_{t,i}$ or imbalance costs β_h. The total fleet costs are defined as follows:

$$C^{fleet}(\theta_\lambda) \doteq \sum_{h}^{N_h} \left[C_h^{bal}(P_h^{bal}) + C_h^{intr}(P_h^{intr}) + \beta_h + \sum_{t}^{T} \sum_{i}^{|\mathcal{F}|} \rho_{t,i} \right] \ , \tag{22}$$

where $\theta_\lambda \in \mathbb{R}_{0 \leq \lambda \leq 1}^{2 \times N_h}$ is the matrix of the risk factors λ_h^{bal}, λ_h^{intr} for all considered market periods N_h. $\mathcal{F}$ denotes the set of all EVs i in the fleet and $|\mathcal{F}|$ the fleet size. The costs for charging $C_h^{bal}(P_h^{bal})$, $C_h^{intr}(P_h^{intr})$ are clearly dependent on the chosen risk factors λ_h^{bal}, λ_h^{intr} (see Eq. 19, Eq. 20). In summary, the problem can be formulated as minimizing the total costs of the fleet, by choosing the optimal set of risk factors θ_λ:

$$\begin{aligned} \underset{\theta_\lambda}{\text{minimize}} \quad & C^{fleet}(\theta_\lambda) \\ \text{subject to} \quad & 0 \leq \lambda_h^{bal} \leq 1, \ \forall \lambda_h^{bal} \in \theta_\lambda \\ & 0 \leq \lambda_h^{intr} \leq 1, \ \forall \lambda_h^{intr} \in \theta_\lambda \end{aligned} \tag{23}$$

Solving this optimization problem with common methods like stochastic programming is a difficult task, assuming that complete information of available charging power and future electricity market prices is not always available. Since one goal of this research is to develop a model that can be applied to previously unknown settings and learn from uncertain environments, as mobility and electricity markets, we chose to solve the problem with a RL learning approach that is explained in detail in Chapter 4.3.

4.2.6 Example

At 3pm on the 9th of August 2017, the controller enters the first bidding phase for procuring electricity one week ahead, the market period h = *16.08.2017 15:00-15:15*. It predicts that at that point in time 250 EVs are connected to a charging station, resulting in 900kW available fleet charging power ($\widehat{P}_h^{fleet} = 900\text{kW}$), given the charging power of 3.6kW per EV. Assuming the available critical prices $\overline{p}_h^c = 5\frac{€}{\text{MW}}$, $\overline{p}_h^e = -10\frac{€}{\text{MWh}}$, and $\overline{p}_h^u = 10\frac{€}{\text{MWh}}$ for that market period, the controller now evaluates the cheapest charging option. The flat industry electricity tariff is assumed to be $p_i = 0.15\frac{€}{\text{kWh}}$. The costs for charging with the maximal amount of power $\widehat{P}_h^{fleet}$ from the balancing market ($C_h^{bal}(900\text{kW}) = -6.25€$) are less than charging from the intraday market ($C_h^{intr}(900\text{kW}) = 2.25€$) or charging at the industry tariff ($900\text{kW} \times 0.25\text{h} \times 0.15\frac{€}{\text{kWh}} = 33.75€$). In this example, by choosing the cheapest option, the balancing market, the fleet operator will even get compensated for charging its fleet.

In the next step, the controller has to submit bids to the balancing market. The RL agent determined that the risk of bidding on the balancing market is $\lambda_h^{bal} = 0.3$. Consequently, the controller sets the bidding quantity to $P_h^{bal} = \widehat{P}_h^{fleet} \times (1 - \lambda_h^{bal}) = 900\text{kW} \times 0.7 = 630\text{kW}$ and submits a bid to the market. Since we are assuming that bids at the critical price, will always get accepted, the controller procures 630kW from the balancing market and updates its account with $C_h^{bal}(630\text{kW}) = -4.725€$.

One week later, at 2:30pm on the 16th of August 2017, the controller enters the second bidding phase. With a time horizon of 30 minutes, it predicts less available fleet charging power of $\widehat{P}_h^{fleet} = 810\text{kW}$ for the same market period *16.08.2017-15:00*. By trading at the intraday market, the controller can now charge the remaining available EVs with a low risk of procuring more energy than it can maximally charge. At this point in time, the RL agent determines a remaining risk of $\lambda_h^{intr} = 0.05$, and sets the bidding quantity to $P_h^{intr} = (810\text{kW}-630\text{kW}) \times (1-0.05) = 171\text{kW}$. Hence, the controller procures 171kW from the intraday market and updates its account with $C_h^{intr}(171\text{kW}) = 0.4275$€.

At electricity delivery time, the 16th of August 2017 at 3:00pm, the controller detects 255 available EVs; EVs which are connected to a charging station and have enough battery capacity left to be charged in the next control period. It assigns 223 EVs to provide the committed 801kW charging power for the market period time Δh of 15 minutes. During that time, three customers want to rent out EVs that are allocated to the VPP. The first two rentals are accepted, because two other EVs are available to charge instead. The third rental has be to refused, since no EV is remaining as substitution. The controller has to account for the opportunity costs of the lost rental $\rho_{t,i}$.

4.3 Reinforcement Learning Approach

In the following chapter the developed RL approach is outlined. First, we define the charging problem as a MDP, and second, the developed learning algorithm is explained.

Remember that the goal of the controlled charging problem is to choose a set of risk factors θ_λ that minimize the fleets total costs across all market periods. The controller is able to influence the charging costs, by setting the risk factors λ_h^{intr}, λ_h^{bal}, which determine the bidding quantities P_h^{bal}, P_h^{intr} that the controller submits to the balancing and intraday market. The RL agent decides on the risk factors (i.e., takes an action) based on the observed state $\mathcal{S}$ every timestep. Bids are submitted every market period h, thus the *timestep* (also called iteration) of the RL problem is set to h, with a length of Δh. The RL agent learns the optimal set of risk factors, by estimating a policy $\pi(a|s)$ that maps every state $s \in \mathcal{S}$ to an action $a \in \mathcal{A}$.

4.3.1 Markov Decision Process Definition

MDPs are defined by the state space $\mathcal{S}$, the action space $\mathcal{A}$, a set of reward signals $\mathcal{R}$ and the state-transition probabilities $p(s'|a,s)$. When $p(s'|a,s)$ is unknown, as it is in our case, it is possible to use a *model-free* approach (see Chapter 2.5.5).

The state space compromises the observed information the agent uses to decide on the action it is going to take. We observed the following factors that are associated with the bidding risk:

1. The bidding period's time of the day

Times of volatile customer activity (e.g., rush hour), increase uncertainty of the amount of available EVs. Bidding for these periods involves more risk of not being able to fulfill market commitments.

2. The current and estimated future size of the VPP

   Large VPPs benefit from the *risk-pooling* effect (Kahlen et al., 2017): Intuitively, larger VPPs have an increased probability that "lost" charging power, due to unforeseen rentals, can be substituted by the rest of the VPP. Hence, we expect that VPP's size is negatively correlated with the associated bidding risk.

Since forecasts of available charging power are already available, we define the predicted size of a VPP $|\widehat{VPP}|_h$ as the as the necessary amount of EVs to provide the predicted charging power $\widehat{P}^{fleet}$ in time period h:

$$|\widehat{VPP}|_h \doteq \left\lceil \frac{\widehat{P}_h^{fleet}}{\gamma} \right\rceil , \tag{24}$$

where γ is the charging power per EV. The state space is then defined as the set of all valid values of the elements of the following tuple:

$$\mathcal{S} \doteq \Big\langle t(h), |VPP|_h, |\widehat{VPP}|_{h+2}, |\widehat{VPP}|_{h+(7\times H)} \Big\rangle \tag{25}$$

where:

- $t(h)$ is the current daytime in hours, with discrete values in the range $\big[0,\ 23\big] \in \mathbb{N}$.
- $|VPP|_t$ is the current VPP size, with discrete values in the range $\big[0,\ |\mathcal{F}|\big] \in \mathbb{N}$.
- $|\widehat{VPP}|_{h+2}$ is the predicted VPP size 30 minutes ahead, with discrete values in the range $\big[0,\ |\mathcal{F}|\big] \in \mathbb{N}$.
- $|\widehat{VPP}|_{h+(7\times H)}$ is the predicted VPP size 7 days ahead, with discrete values in the range $\big[0,\ |\mathcal{F}|\big] \in \mathbb{N}$.

The state space encompasses $|\mathcal{F}|^3 \times 24$ states. Assuming a fleet size of 500 EVs that are 3×10^9 states.

The action space is constituted by all valid values of the risk factors $\lambda_h^{bal}, \lambda_h^{intr}$ constitute the action space:

$$\mathcal{A} \doteq \big\{\lambda_h^{bal}, \lambda_h^{intr} \in \mathbb{R}_{0\leq\lambda\leq 1}\big\} \tag{26}$$

where:

- λ_h^{bal} is the risk factor for bidding on the balancing market 7 days ahead, with discrete values in the range $\left[0, 1\right]$ in 0.05 increments.
- λ_h^{intr} is the risk factor for bidding on the intraday market 30 minutes ahead, with discrete values in the range $\left[0, 1\right]$ in 0.05 increments.

The action space encompasses $20^2 = 400$ actions. The state space and action space were consciously discretized to facilitate faster convergence rates of the RL agent. Continuous spaces are theoretically possible with current RL algorithms, but they are computationally more complex and require function approximation methods to converge (Sutton & Barto, 2018).

We define the reward signal as the total fleet costs that occurred in the last timestep (see Eq. 22).

$$R_h = C_h^{fleet} - C_{h-1}^{fleet} \ , \tag{27}$$

where C_h^{fleet} are the total accumulated fleet costs until market period t. Depending on the risk factors, which determine bidding quantities, imbalance costs and lost rental costs, the agents get positive or negative rewards based on the actions it took. Since the result of

4.3.2 Learning Algorithm

- (Dauer et al., 2013),
- (Di Giorgio et al., 2013)
- (Vaya et al., 2014)
- (Vandael et al., 2015)
- Cost function?
- Deep double Q learning (van Hasselt et al., 2016)
  - Image: (Wang et al., 2015)
  - double q learning combined with deep q learning:
  - off-policy approach

The RL agent was written in Python 3.6, and used the library Keras (`https://keras.io/`). Keras is "a high-level neural networks API, [...] capable of running on top of TensorFlow [...] with a focus on enabling fast experimentation".

- OpenAi Gym?

- Hardware: Google Colab
  - GPU: Nvidia Tesla K80 , 2880 x 2 CUDA cores, 12GB GDDR5 VRAM
  - CPU: Intel(R) Xeon(R) CPU @ 2.30GHz (1 core, 2 threads), 45MB Cache
  - RAM: ~12.6 GB Available
- Training:
  - Time
  - Hyperparameters? - In results or here?

5 Simulation Platform: FleetSim

- Simulation Platform to allow evaluating the performance of intelligent agents in smart charging/balancing the grid of EV fleets. Allows to test out bidding strategies and control mechanisms in a realistic EV fleet setting.
- Real life comparison graph: 10%.

5.1 Event-based Simulation

- Not only t+1
- Event e.g. denying rental, charge/little charge/no-charge has effects for the whole simulation
- Simpy
- Python

5.2 Architecture / Components

5.3 Modular Expandability

- Plug-in different Market designs
- Use different real-world data
- Change Fleet parameters
- Develop new strategy
-

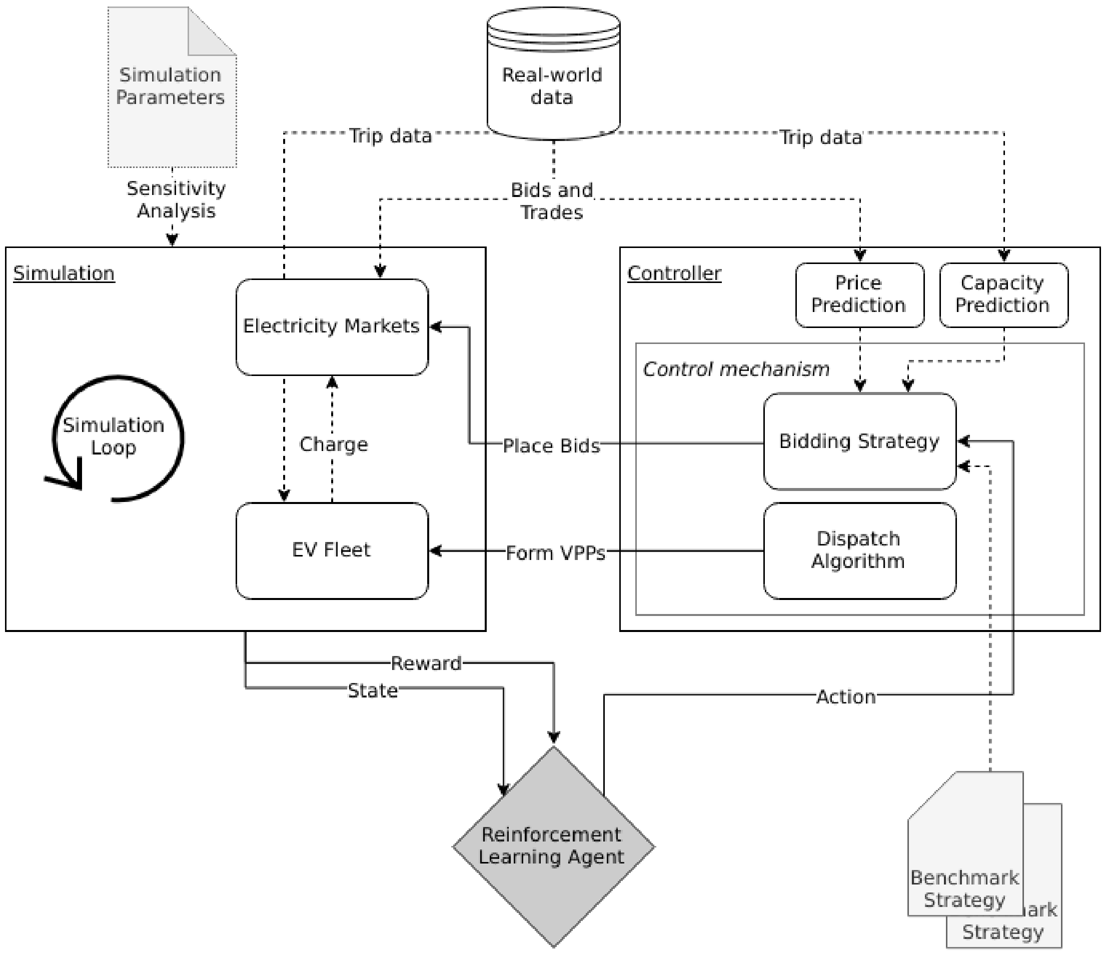

Figure 6: Architecture of FleetSim

6 Results

6.1 Simulation Settings

- Results heavily dependent on industry charging price, since on average the balancing prices are 50% cheaper, and intraday 30% cheaper.
- BMWi. (n.d.). Prices of electricity for the industry in Germany from 2008 to 2017 (in euro cents per kilowatt hour). In Statista - The Statistics Portal. Retrieved March 18, 2019, from `https://www.statista.com/statistics/595803/electricity-industry-price-germany/`.

6.2 FleetRL

- long-delayed rewards make RL hard (!?)
- Compare RL Algos eg. Q-learning vs DDQN –> deep learning makes difference in practice for complex system

6.3 Sensitivity Analysis

- Prediction Accuracy
- Charging infrastructure

7 Conclusion

7.1 Contribution

- Compare to most simular studies:

(Kahlen et al., 2018; Vandael et al., 2015) etc..

- Business model for EV fleet owners with better results than previous studies
- Environmental impact by providing balancing power
- Decision Support System for controlled EV charging from multiple markets
- RL Algorithm that is designed to work in previously unknown environments and thus suited to deploy in real life settings of all kinds of EV fleets in all kinds of cities. E.g. scooters also?
- Event-based Simulation Platform to evaluate bidding strategies and RL agents, facilitate research

7.2 Limitations

- Model:
  - Bidding Mechanism: one week ahead, always accepted
  - Policy & Regulation: EVs not allowed to provide balancing power, minimum bidding quantities 1MW.
  - Markets: Fleet is a price-taker, what about larger fleets? Simulate market influence
- RL: See (Vázquez-Canteli & Nagy, 2019) conclusion for limitations.

7.3 Future Research

- Model: Current market design, i.e. daily w/ 4h slots. German "Mischpreisverfahren"
- RL: Long-delayed rewards, different reward structure, memory based

References

Agricola, A., Seidl, H., & Mischinger, S. (2014). DENA Ancillary Services Study 2030. Security and Reliability of a Power Supply With a High Percentage of Renewable Energy [Technical Report].

Avci, E., Ketter, W., & van Heck, E. (2018). Managing Electricity Price Modeling Risk Via Ensemble Forecasting: the Case of Turkey. *Energy Policy*, 390-403. Retrieved from `https://doi.org/10.1016/j.enpol.2018.08.053` doi: 10.1016/j.enpol.2018.08.053

Bellman, R. E. (1957). *Dynamic Programming* [Ph.D. Thesis]. Courier Dover Publications.

Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends® in Machine Learning*. Retrieved from `https://doi.org/10.1561/2200000006` doi: 10.1561/2200000006

BMU. (2010). *Energy Concept for an Environmentally Sound, Reliable and Affordable energy* [Technical Report]. Federal Ministry for the Environment, Nature Conservation and Nuclear Safety.

Brandt, T., Wagner, S., & Neumann, D. (2017). Evaluating a Business Model for Vehicle-Grid Integration: Evidence From Germany. *Transportation Research Part D: Transport and Environment*, 488-504. Retrieved from `https://doi.org/10.1016/j.trd.2016.11.017` doi: 10.1016/j.trd.2016.11.017

Burns, L. D. (2013). Sustainable Mobility: a Vision of Our Transport Future. *Nature*. Retrieved from `https://doi.org/10.1038/497181a` doi: 10.1038/497181a

Chis, A., Lunden, J., & Koivunen, V. (2016). Reinforcement Learning-Based Plug-In Electric Vehicle Charging With Forecasted Price. *IEEE Transactions on Vehicular Technology*. Retrieved from `https://doi.org/10.1109/tvt.2016.2603536` doi: 10.1109/tvt.2016.2603536

Cybenko, G. (1989). Approximation By Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals, and Systems*. Retrieved from `https://doi.org/10.1007/bf02551274` doi: 10.1007/bf02551274

Dauer, D., Flath, C. M., Strohle, P., & Weinhardt, C. (2013). Market-Based EV Charging Coordination. In *Ieee/wic/acm international joint conferences on web intelligence (wi) and intelligent agent technologies (iat).* Retrieved from `https://doi.org/10.1109/wi-iat.2013.97` doi: 10.1109/wi-iat.2013.97

Di Giorgio, A., Liberati, F., & Pietrabissa, A. (2013). On-board stochastic control of Electric Vehicle recharging. In *52nd ieee conference on decision and control.* Retrieved from `https://doi.org/10.1109/cdc.2013.6760789` doi: 10.1109/cdc.2013.6760789

Dusparic, I., Harris, C., Marinescu, A., Cahill, V., & Clarke, S. (2013). Multi-agent residential demand response based on load forecasting. In *IEEE Conference on Technologies for Sustainability (SusTech).* Retrieved from `https://doi.org/10.1109/sustech.2013.6617303` doi: 10.1109/sustech.2013.6617303

Ernst, D., Geurts, P., & Wehenkel, L. (2003). Iteratively extending time horizon reinforcement learning. In *European conference on machine learning.*

Firnkorn, J., & Müller, M. (2015). Free-Floating Electric Carsharing-Fleets in Smart Cities: the Dawning of a Post-Private Car Era in Urban Environments? *Environmental Science & Policy*, 30-40. Retrieved from `https://doi.org/10.1016/j.envsci.2014.09.005` doi: 10.1016/j.envsci.2014.09.005

He, G., Chen, Q., Kang, C., Pinson, P., & Xia, Q. (2016). Optimal Bidding Strategy of Battery Storage in Power Markets Considering Performance-Based Regulation and Battery Cycle Life. *IEEE Transactions on Smart Grid*, 2359-2367. Retrieved from `https://doi.org/10.1109/tsg.2015.2424314` doi: 10.1109/tsg.2015.2424314

Humphrys, M. (1996). Action Selection Methods Using Reinforcement Learning. In *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior.*

Kahlen, M., Ketter, W., & Gupta, A. (2017). Fleetpower: Creating Virtual Power Plants in Sustainable Smart Electricity Markets. *SSRN Electronic Journal*. Retrieved from `https://doi.org/10.2139/ssrn.3062433` doi: 10.2139/ssrn.3062433

Kahlen, M., Ketter, W., & van Dalen, J. (2014). Balancing With Electric Vehicles: a Profitable Business Model.

Kahlen, M., Ketter, W., & van Dalen, J. (2018). Electric Vehicle Virtual Power Plant Dilemma: Grid Balancing Versus Customer Mobility. *Production and Operations Management*. Retrieved from `https://doi.org/10.1111/poms.12876` doi: 10.1111/poms.12876

Kambil, A., & van Heck, E. (1998). Reengineering the Dutch Flower Auctions: a Framework for Analyzing Exchange Organizations. *Information Systems Re-*

search, 1-19. Retrieved from https://doi.org/10.1287/isre.9.1.1 doi: 10.1287/isre.9.1.1

Kara, E. C., Macdonald, J. S., Black, D., Bérges, M., Hug, G., & Kiliccote, S. (2015). Estimating the Benefits of Electric Vehicle Smart Charging At Non-Residential Locations: a Data-Driven Approach. *Applied Energy*, 515-525. Retrieved from https://doi.org/10.1016/j.apenergy.2015.05.072 doi: 10.1016/j.apenergy.2015.05.072

Karanfil, F., & Li, Y. (2017). The Role of Continuous Intraday Electricity Markets: the Integration of Large-Share Wind Power Generation in Denmark. *The Energy Journal*. Retrieved from https://doi.org/10.5547/01956574.38.2.fkar doi: 10.5547/01956574.38.2.fkar

Ketter, W., Collins, J., & Reddy, P. (2013). Power Tac: a Competitive Economic Simulation of the Smart Grid. *Energy Economics*, 262-270. Retrieved from https://doi.org/10.1016/j.eneco.2013.04.015 doi: 10.1016/j.eneco.2013.04.015

Kiesel, R., & Paraschiv, F. (2017). Econometric Analysis of 15-minute Intraday Electricity Prices. *Energy Economics*. Retrieved from https://doi.org/10.1016/j.eneco.2017.03.002 doi: 10.1016/j.eneco.2017.03.002

Kim, E. L., Tabors, R. D., Stoddard, R. B., & Allmendinger, T. E. (2012). Carbitrage: Utility Integration of Electric Vehicles and the Smart Grid. *The Electricity Journal*, 16-23. Retrieved from https://doi.org/10.1016/j.tej.2012.02.002 doi: 10.1016/j.tej.2012.02.002

Ko, H., Pack, S., & Leung, V. C. M. (2018). Mobility-Aware Vehicle-To-Grid Control Algorithm in Microgrids. *IEEE Transactions on Intelligent Transportation Systems*. Retrieved from https://doi.org/10.1109/tits.2018.2816935 doi: 10.1109/tits.2018.2816935

Lopes, J. A. P., Soares, F. J., & Almeida, P. M. R. (2011). Integration of Electric Vehicles in the Electric Power System. *Proceedings of the IEEE*, 168-183. Retrieved from https://doi.org/10.1109/jproc.2010.2066250 doi: 10.1109/jproc.2010.2066250

Mashhour, E., & Moghaddas-Tafreshi, S. M. (2011a). Bidding Strategy of Virtual Power Plant for Participating in Energy and Spinning Reserve Markets-Part II: Numerical Analysis. *IEEE Transactions on Power Systems*, 957-964. Retrieved from https://doi.org/10.1109/tpwrs.2010.2070883 doi: 10.1109/tpwrs.2010.2070883

Mashhour, E., & Moghaddas-Tafreshi, S. M. (2011b). Bidding Strategy of Virtual Power Plant for Participating in Energy and Spinning Reserve Markets-Part I: Problem Formulation. *IEEE Transactions on Power Systems*, 949-956. Retrieved from `https://doi.org/10.1109/tpwrs.2010.2070884` doi: 10.1109/tpwrs.2010.2070884

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … Hassabis, D. (2015). Human-Level Control Through Deep Reinforcement Learning. *Nature*. Retrieved from `https://doi.org/10.1038/nature14236` doi: 10.1038/nature14236

Pandžić, H., Morales, J. M., Conejo, A. J., & Kuzle, I. (2013). Offering Model for a Virtual Power Plant Based on Stochastic Programming. *Applied Energy*. Retrieved from `https://doi.org/10.1016/j.apenergy.2012.12.077` doi: 10.1016/j.apenergy.2012.12.077

Pape, C., Hagemann, S., & Weber, C. (2016). Are Fundamentals Enough? Explaining Price Variations in the German Day-Ahead and Intraday Power Market. *Energy Economics*. Retrieved from `https://doi.org/10.1016/j.eneco.2015.12.013` doi: 10.1016/j.eneco.2015.12.013

Peters, M., Ketter, W., Saar-Tsechansky, M., & Collins, J. (2013). A reinforcement learning approach to autonomous decision-making in smart electricity markets. *Machine learning*, 5–39.

Peterson, S. B., Whitacre, J., & Apt, J. (2010). The Economics of Using Plug-In Hybrid Electric Vehicle Battery Packs for Grid Storage. *Journal of Power Sources*, 2377-2384. Retrieved from `https://doi.org/10.1016/j.jpowsour.2009.09.070` doi: 10.1016/j.jpowsour.2009.09.070

Pudjianto, D., Ramsay, C., & Strbac, G. (2007). Virtual Power Plant and System Integration of Distributed Energy Resources. *IET Renewable Power Generation*, 10. Retrieved from `https://doi.org/10.1049/iet-rpg:20060023` doi: 10.1049/iet-rpg:20060023

Reddy, P. P., & Veloso, M. M. (2011a). Learned Behaviors of Multiple Autonomous Agents in Smart Grid Markets. In *Aaai.*

Reddy, P. P., & Veloso, M. M. (2011b). Strategy learning for autonomous agents in smart grid markets. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence.*

Rummery, G. A., & Niranjan, M. (1994). *On-line Q-learning using connectionist systems.* University of Cambridge, Department of Engineering Cambridge, England.

Schill, W.-P. (2011). Electric Vehicles in Imperfect Electricity Markets: the Case of Germany. *Energy Policy*, 6178-6189. Retrieved from `https://doi.org/10.1016/j.enpol.2011.07.018` doi: 10.1016/j.enpol.2011.07.018

Shi, W., & Wong, V. W. (2011). Real-time vehicle-to-grid control algorithm under price uncertainty. In *Ieee international conference on smart grid communications (smartgridcomm).* Retrieved from `https://doi.org/10.1109/smartgridcomm.2011.6102330` doi: 10.1109/smartgridcomm.2011.6102330

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D. (2016). Mastering the Game of Go With Deep Neural Networks and Tree Search. *Nature*. Retrieved from `https://doi.org/10.1038/nature16961` doi: 10.1038/nature16961

Sioshansi, R. (2012). The Impacts of Electricity Tariffs on Plug-In Hybrid Electric Vehicle Charging, Costs, and Emissions. *Operations Research*, 506-516. Retrieved from `https://doi.org/10.1287/opre.1120.1038` doi: 10.1287/opre.1120.1038

Sterling, D. (2018). *Three Revolutions: Steering Automated, Shared, and Electric Vehicles to a Better Future.* Island Press/Center for Resource Economics. Retrieved from `https://doi.org/10.5822/978-1-61091-906-7` doi: 10.5822/978-1-61091-906-7

Sutton, R. S. (1984). Temporal Credit Assignment in Reinforcement Learning [Ph.D. Thesis]. *University of Massechusetts, Amherst*.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

Taylor, A., Dusparic, I., Galvan-Lopez, E., Clarke, S., & Cahill, V. (2014). Accelerating Learning in multi-objective systems through Transfer Learning. In *International joint conference on neural networks (ijcnn).* Retrieved from `https://doi.org/10.1109/ijcnn.2014.6889438` doi: 10.1109/ijcnn.2014.6889438

Tomić, J., & Kempton, W. (2007). Using Fleets of Electric-Drive Vehicles for Grid Support. *Journal of Power Sources*, 459-468. Retrieved from `https://doi.org/10.1016/j.jpowsour.2007.03.010` doi: 10.1016/j.jpowsour.2007.03.010

Valogianni, K., Ketter, W., Collins, J., & Zhdanov, D. (2014). Effective Management of Electric Vehicle Storage Using Smart Charging. In *Aaai* (pp. 472–478).

Vandael, S., Claessens, B., Ernst, D., Holvoet, T., & Deconinck, G. (2015). Reinforcement Learning of Heuristic EV Fleet Charging in a Day-Ahead Electricity Market. *IEEE Transactions on Smart Grid*, 1795-1805. Retrieved from `https://doi.org/10.1109/tsg.2015.2393059` doi: 10.1109/tsg.2015.2393059

van der Veen, R. A., & Hakvoort, R. A. (2016). The Electricity Balancing Market: Exploring the Design Challenge. *Utilities Policy*. Retrieved from `https://doi.org/https://doi.org/10.1016/j.jup.2016.10.008` doi: https://doi.org/10.1016/j.jup.2016.10.008

van Hasselt, H., Guez, A., & Silver, D. (2016). Deep Reinforcement Learning with Double Q-Learning. In *Aaai*.

Vaya, M. G., & Andersson, G. (2015). Optimal Bidding Strategy of a Plug-In Electric Vehicle Aggregator in Day-Ahead Electricity Markets Under Uncertainty. *IEEE Transactions on Power Systems*. Retrieved from `https://doi.org/10.1109/tpwrs.2014.2363159` doi: 10.1109/tpwrs.2014.2363159

Vaya, M. G., Rosello, L. B., & Andersson, G. (2014). Optimal bidding of plug-in electric vehicles in a market-based control setup. In *Power systems computation conference*. Retrieved from `https://doi.org/10.1109/pscc.2014.7038108` doi: 10.1109/pscc.2014.7038108

Vázquez-Canteli, J. R., & Nagy, Z. (2019). Reinforcement Learning for Demand Response: a Review of Algorithms and Modeling Techniques. *Applied Energy*, 1072-1089. Retrieved from `https://doi.org/10.1016/j.apenergy.2018.11.002` doi: 10.1016/j.apenergy.2018.11.002

Wagner, S., Brandt, T., & Neumann, D. (2016). In Free Float: Developing Business Analytics Support for Carsharing Providers. *Omega*, 4-14. Retrieved from `https://doi.org/10.1016/j.omega.2015.02.011` doi: 10.1016/j.omega.2015.02.011

Wang, Z., Schaul, T., Hessel, M., Van Hasselt, H., Lanctot, M., & De Freitas, N. (2015). Dueling Network Architectures for Deep Reinforcement Learning. *arXiv preprint arXiv:1511.06581*.

Watkins, C. J. C. H. (1989). Learning From Delayed Rewards [Ph.D. Thesis]. *King's College, Cambridge*.

Watkins, C. J. C. H., & Dayan, P. (1992). Q-Learning. *Machine Learning*. Retrieved from `https://doi.org/10.1007/BF00992698` doi: 10.1007/BF00992698

Weron, R. (2014). Electricity Price Forecasting: a Review of the State-Of-The-Art With a Look Into the Future. *International Journal of Forecasting*, 1030-1081. Retrieved from `https://doi.org/10.1016/j.ijforecast.2014.08.008` doi: 10.1016/j.ijforecast.2014.08.008