Reinforcement Learning Portfolio Optimization of Electric Vehicle Virtual Power Plants

Master Thesis

Author: Tobias Richter (Student ID: 5583705)
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

May 3, 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

Abstract

Integrating weather-dependent renewable energy sources into the electricity system imposes challenges on the power grid. Balancing services are needed which can be provided by virtual power plants (VPP) that aggregate distributed energy resources (DER) to consume or produce electricity on demand. Electric vehicle (EV) fleets can use the batteries of idle cars as combined storage to offer balancing services on smart electricity markets. However, there are risks associated with this business model extension. The fleet faces severe imbalance penalties if it can not charge the offered amount of balancing energy due to unpredicted mobility demand of the vehicles. Ensuring the fleet can fulfill all market commitments risks denying profitable customer rentals. We study the design of a decision support system that estimates these risks, dynamically adjusts the composition of a VPP portfolio, and profitably places bids on multiple electricity markets simultaneously. Here we show that a *Reinforcement Learning* (RL) agent can optimize the VPP portfolio by learning from favorable market conditions and fleet demand uncertainties. In comparison to previous research, in which the bidding risks were unknown and fleets could only offer conservative amounts of balancing power to a single market, our proposed approach increases the amount of offered balancing power by $48\% - 82\%$ and achieves a charging cost reduction of the fleet by 25%. In experiments with real-world carsharing data of 500 EVs, we found that the accuracy of mobility demand forecasting algorithms is crucial for a successful bidding strategy. Moreover, we show that recent advancements in RL decrease the convergence time and improve the robustness of the results. Our results demonstrate how modern RL algorithms can be successfully used for fleet management, VPP optimization, and demand response in the smart grid. We anticipate that DER, such as EVs, will play an essential role in providing reliable back up power for the grid and formulate market design recommendations to allow easier access to these resources.

Contents

List of Figures

List of Tables

List of Abbreviations

ANN Artificial Neural Network

DDQN Double Deep Q-Networks

DER Distributed Energy Resources

DP Dynamic Programming

EPEX European Power Exchange

EV Electric Vehicle

GCRM German Control Reserve Market

MDP Markov Decision Process

RES Renewable Energy Sources

RL Reinforcement Learning

SoC State of Charge

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.

$\doteq$	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
$p(s', r \,|\, s, a)$	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
$\overline{\text{VE}}$	mean square value error

1   Introduction

Global climate change is one of the most substantial challenges of our time. It necessitates reducing carbon emissions and shifting to sustainable energy production. However, the integration of renewable energy sources (RES) is a complex matter. Sustainable electricity production is dependent on the weather conditions, and difficult to integrate into the electrical power system. Changing weather conditions cause under- and oversupplies of intermittent solar and wind energy which are destabilizing the power grid. This thesis will explore a novel approach of integrating renewable energy into the power system by providing balancing power with virtual power plants (VPPs) of electric vehicles (EVs). The following chapter is structured as follows: First, the research motivation is described, the derived research questions are presented next, and finally, the relevance of this research explained.

1.1   Research Motivation

VPPs play an important role in stabilizing the grid by aggregating distributed energy resources (DER) to consume and produce electricity when it is needed (Pudjianto et al., 2007). At the same time, carsharing companies operate large, centrally managed EV fleets in major cities around the world. These EV fleets can be turned into VPPs by using their batteries as combined electricity storage. In this way, fleet operators can offer balancing services to the power grid or trade electricity on the open markets for arbitrage purposes. Fleet operators can charge the fleet by buying electricity and discharge the fleet by selling electricity when market conditions are favorable.

However, renting out EVs to customers is considerably more lucrative than using EV batteries for trading electricity (Kahlen et al., 2018). By making EVs available to be used as a VPP, fleet operators compromise customer mobility and potentially the profitability of the fleet. Knowing how many EVs will be available for VPP usage in the future is critical for a successful trading strategy. Accurate rental demand forecasts help determining the amount of electricity that can be traded on the market. EV fleet operators can also participate on multiple electricity markets simultaneously. They can take advantage of distinctive market properties, such as the auction mechanism or the lead time to delivery, to optimize their bidding strategy and reduce the bidding risks. Still, the ultimate risk remains that the fleet could made commitments to the markets it cannot fulfill due to unforeseen rental demand at the time of electricity delivery.

We state that participating in the balancing market and intraday market simultaneously can mitigate risks and increase profits of the fleet. In this research, we propose a portfolio optimizing strategy, in which the best composition of the VPP portfolio is dynamically determined using a *Reinforcement Learning* (RL) approach. RL methods have shown to be on par or better than human level decision making in a variety of domains (Mnih et al., 2015). An RL agent is able to adapt the bidding strategy to changing rental demands and market conditions. It learns from historical data, the observed environment, and realized profits to adjust the bidding quantities and dynamically determine bidding risks. The following fleet control tasks need to be performed by the agent in real-time: 1) Allocate available EVs to the VPP portfolio, 2) Determine the optimal VPP portfolio composition, and 3) Place bids and asks on multiple electricity markets with an integrated trading strategy.

1.2   Research Questions

Drawing upon the research motivation, this research aims to answer the following research questions:

1. *Can EV fleet operators create VPP portfolios to profitably trade electricity on the balancing market and intraday market simultaneously? What would an integrated bidding strategy look like for EV fleet operators that manage VPP portfolios?*

2. *Can an RL agent optimize VPP portfolios by learning the risks that are associated with bidding on the individual electricity markets?*

1.3   Relevance

From a scientific perspective, this thesis is relevant to the stream of agent-based decision making in smart markets (Bichler et al., 2010; Peters et al., 2013). It contributes to the body of Design Science in Information Systems (Hevner et al., 2004) and draws upon work grounded within a multitude of research areas: VPPs in smart electricity markets (Pudjianto et al., 2007), fleet management of (electric) carsharing as a new way of sustainable mobility (Brandt et al., 2017; Wagner et al., 2016), and advanced RL techniques for the smart grid (Vázquez-Canteli & Nagy, 2019). We specifically build on research that has been carried out by Kahlen et al. (2017, 2018). In their papers, the authors concentrate on trading EV battery capacity on one market at a time. As proposed by the authors, we take this research further and use a VPP of EVs to participate on multiple types of electricity markets simultaneously. In this way we create a VPP portfolio that

offers EV batteries as storage options to the markets with an integrated bidding strategy.

From a business perspective, this thesis is relevant to carsharing companies operating EV fleets, such as Car2Go or DriveNow. We show how these companies can increase their profits, using idle EVs as a VPP to trade electricity on multiple markets simultaneously. We propose the use of a decision support system (DSS), which allocates idle EVs to be used as VPP or to be available for rent. Furthermore, the DSS determines optimal capacity-price pairs to place bids on the individual electricity markets. Using an event-based simulation platform, we estimate the profitability of the proposed methods, using real-world data from German electricity markets and trip data from a German carsharing provider.

This thesis also contributes to the overall welfare of society. First, VPPs of EVs provide extra balancing services to the power grid. VPPs consume excess electricity almost instantly and stabilize the power grid. By integrating more intermittent RES into the grid, such balancing services will become indispensable in the future. Second, a reduction of electricity prices for the end-consumer is expected. Integrating VPPs into the power grid increases the efficiency of the whole system and, hence, will lower the prices. Kahlen et al. (2018) show results, where electricity prices decrease up to 3.4% on the wholesale market. We anticipate similar results in our research. Third, VPPs can lead to a decrease in CO_2 emissions. With an increasing share of renewable energy production, the supply of sustainable electricity can excess the total electricity demand at times of good weather conditions. A VPP can consume this electricity by charging the EV fleet and the sustainable energy production does not need to be curtailed. EVs that are equipped with special vehicle-to-grid (V2G) devices can feed the electricity back into the grid when there is more demand than supply. This mechanism increases the utilization of renewable electricity generation and reduces the total CO_2 emissions.

2   Background

The following chapter describes the relevant research and the theoretical background of this thesis. First, we introduce smart electricity markets, give detailed explanations of two particular types of electricity markets, and explain applications for EV fleets in the smart grid. Next, we discuss the related RL literature in the context of controlled EV charging. Finally, the methodological background is presented in the form of a comprehensive review of the most important concepts and formulations in RL theory.

2.1   Smart Electricity Markets

On electricity markets, participants place asks (sale offers) and bids (purchase orders) in auctions to match the supply of electricity generation and the demand for electricity consumption. 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. Most electricity markets work according to the merit order principle in which resources are considered in an 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 an uniform pricing schema is applied, in secondary reserve markets and intraday markets, participants get compensated by the price they bid (pay-as-bid principle).

In smart electricity markets, EV fleet operators can offer the capacity of their EV batteries to different types of markets. On operating reserve markets, prices are usually more volatile and consequently more attractive for VPPs (Tomić & Kempton, 2007). But operating reserve markets also bear a higher risk for the fleet: Commitments have to be made up to one week in advance when customer demands are still uncertain. In order to avoid 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. Just minutes before delivery it is

possible to predict the available battery capacity of the fleet with a high accuracy. This certainty creates the opportunity to aggressively trade the available battery capacity with a low risk at the spot market.

2.2   Electricity Market Theory

In the following chapter, we will explain the market design, market properties and bidding procedures of the balancing market and spot market in more detail, as they are the markets we include in our research.
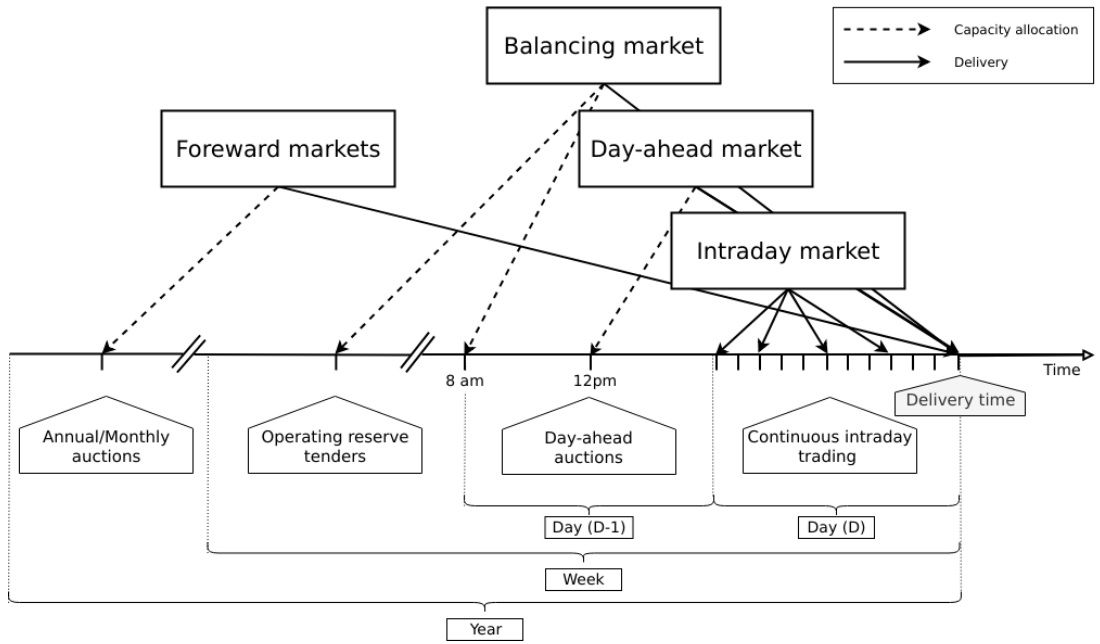
Figure 1: Interaction between electricity markets in relation to capacity allocation.

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 is conceptually placed at the intersection between the open electricity markets and the physical power system (van der Veen & Hakvoort, 2016). It can be seen as the last link in a chain of electricity markets, although the lead time of capacity allocation is considerably longer than the at the day-ahead market or the intraday market.

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 liberalized 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 1 MW of power over a 4-hour interval within 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 that require fast ramp-up times. Operating reserve providers have to be qualified by the TSO to participate in the market and prove that they are able to reliably provide balancing power. Although EV fleets are currently not qualified by the GCRM to be used as operating reserve, they could theoretically handle the minimum capacity requirements if they were large enough. Around 220 EVs would need to simultaneously charge at standard 4.6 kW charging stations to provide 1 MW of downward regulating capacity.

Up until July 28, 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 50 Hz. 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 50 Hz 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 plainly refer to as *balancing power* until the end of the thesis.

Market participants submit bids in the form (P^{bal}, p^c, p^e) to the market, 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 $\frac{\text{€}}{\text{MW}}$ and p^e is the energy price for delivered energy in $\frac{\text{€}}{\text{MWh}}$. The TSOs determine the target quantity of balancing power to acquire per market period based on historical balancing data. They usually acquire much higher regulation capacity to minimize risks and activate the needed amount of capacity on demand. The TSOs accept the

---

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

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

bids based on the capacity price in a merit order. Providers whose bids were accepted instantly get compensated for the provided capacity 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. Since providers get paid according to the submitted prices p^c and p^e (instead of a market clearing price) 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 (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 567 TWh in 2018[3], but most spot markets in western economies work with similar market mechanisms.

Germany's most important spot market is the day-ahead market with a trading volume of over 234 TWh 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 important role in integrating volatile 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 precise intraday forecast (forecasting error) 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

---

[3]`https://www.epexspot.com/en/press-media/press/details/press/Traded_volumes`
`_soar_to_an_all-time_high_in_2018`, accessed February 19, 2019

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 $\frac{\text{€}}{\text{MWh}}$. When an order to buy (bid) matches an order to sell (ask), the trade gets executed immediately. 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 82 TWh, which is considerably smaller than the day-ahead market's volume. Despite that, the intraday market plays a vital role to the stability of the grid. All executed trades on the intraday market potentially reduce the needed balancing power of the TSOs and thus can be seen as a type of balancing power (Pape et al., 2016).

Purchasing electricity on the continuous intraday market is attractive for EV fleets with uncertain mobility demand. The short time before delivery allows EV fleet operators to rely on highly accurate forecasts of available battery capacity, 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. For example can trades on the intraday market complement bids that have been committed earlier to the balancing 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). Charging multiple EVs in the same neighborhood, or worse, whole EV fleets, can even cause brown- or blackouts (Kim et al., 2012). Despite these challenges, it is possible to support the physical reinforcement of the grid 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 (*peak cutting*) and complement the grid in times of low demand (*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 and the owners of the EVs. Their approach resulted in a price reduction in energy bills and a peak load reduction of 37%. An extension of the smart charging concept is 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 V2G 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 the 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 the 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 six to seventy-two dollar 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 negative results from the studies mentioned above, we did not include the V2G concept in our research.

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 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 devel-

oped a non-equilibrium model, which solves the presented mixed-integer program with genetic programming. Contrarily to their study, we investigate the use of 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). He et al. (2016) conducted similar research, in which they additionally incorporated battery depreciation costs in a profit maximization model, which proved to be a decisive factor. In contrast their study, we jointly participate in the secondary operating reserve and intraday market with the *non-stationary* storage of EV batteries. Also, we can exclude battery depreciation from our model, since shared EVs have to satisfy mobility demand and be charged in any case.

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 1 MW to 5 MW, while a typical EV has a charging power of 3.6 kW to 22 kW. 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) showed that electricity brokers can overcome the capacity issues by aggregating EV batteries. In addition to electricity brokers, we apply the concept of VPPs. VPPs are flexible portfolios of DER, 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 balancing power to the markets without knowing which exact sources provide the promised capacity until the delivery time. 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 level 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 mobility concept that allows cars to be picked up and parked everywhere This concept 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, see for example 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 to not 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 not be sure to charge with their available EVs at the delivery time. To address this issue, the authors developed a mean asymmetric weighted objective function. They used it for their time-series based prediction model to penalize committing an EV to a 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 showed that intelligent agents equipped with 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 *Tariff Market*. The agents are modelled with a Markov decision process (MDP) and use RL to determine 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 price features and market participant features. By engaging with its environment, the agent learns an optional sequence of actions, called policy.

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 learning a function representation of state values instead of single values of discrete states. By using this technique, the agent can adapt to arbitrary economic signals from its environment, resulting in a 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, and 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 nine 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 autonomous microgrids, as well as SoC of the EVs. The authors formulated an MDP with discrete state and action spaces and use standard *Q-Learning* with an ϵ-greedy policy to derive an optimal charging policy.

It should be noted that RL methods are not the only solution for problems in the smart grid, often basic algorithms and heuristics provide satisfactory results (Vázquez-Canteli & Nagy, 2019). Another possibility are stochastic programming approaches, which rigorously model the problem as an optimization problem under uncertainty that can be analytically solved, see for example Pandžić et al. (2013) and Nguyen and Le (2016). Uncertainties (e.g., renewable energy generation or electricity prices) are modeled by a set of scenarios that are based on historical data. Despite that, we consider 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 the smart grid.

2.5   Reinforcement Learning Theory

The following chapter will give an overview of the most important RL concepts and will introduce the corresponding mathematical formulations. If not noted otherwise, the notation, equations, and explanations are adapted from Sutton and 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 depend on future actions of the agent. 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 environment's state. State values indicate the long-term desirability of a state, that is the total amount of reward the agent can expect to accumulate in 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 (see 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 (see Chapter 2.5.5). The last two chapters cover novel 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

MDPs are a classical formulation of sequential decision making and an idealized mathematical formulation of the RL problem. They allow to derive exact theoretical statements about the learning problem and possible solutions. Figure 2 depicts the *agent-environment interaction*.
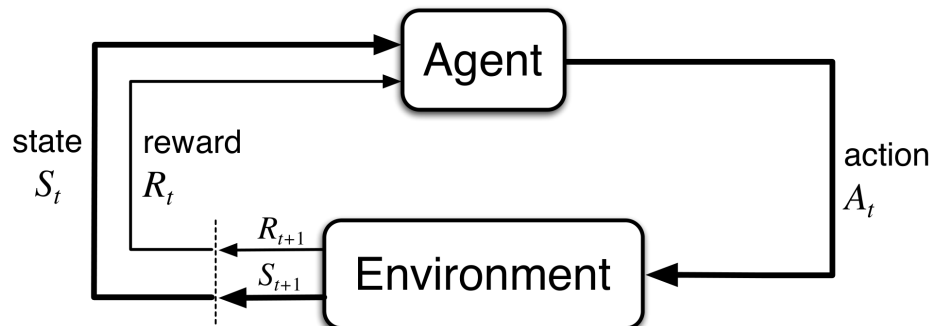
Figure 2: The agent-environment interaction in a Markov decision process (Sutton & Barto, 2018). [4]

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 an 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 $a \in \mathcal{A}$. In *finite* MDPs, the random variables R_t and S_t have well-defined probability density functions, which are solely dependent on the previous state and the agent's action.

---

[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/)

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 S_t depends only on the immediately preceding state 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 (1) allows computing the *state-transition probabilities*, another important characteristic of an MDP:

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

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 $R_t \in \mathbb{R}$. 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 dependent on the current state. More formally, the policy is defined as $\pi(a|s) \doteq \Pr\{A_t = a|S_t = s\}$, a probability density function of all $a \in \mathcal{A}$ for each

$s \in \mathcal{S}$. The various RL approaches mainly differ in how the policy is updated, based on the agent's interaction with the environment.

In RL, not only value functions of states but also 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 their recursive relationship between the state (or action) values. Similar to rewards in (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:

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

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. The value function $v_\pi(s)$ is the unique solution

to its Bellman equation. The Bellman equation of the optimal value function v_* is called the *Bellman optimality equation*:

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

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 (1) are fully known. Similar Bellman equations as (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 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 (see 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 (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 $a \in A$ without forward looking behaviour. 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*:

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

where $\xrightarrow{\text{E}}$ denotes a policy evaluation step, $\xrightarrow{\text{I}}$ denotes a policy improvement step. π_* and v_* denote 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 and is computationally expensive. In contrast, the *value iteration* algorithm stops the policy evaluation step after one sweep. In this case, the two previous steps can be combined into one single update step:

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

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 fully known 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 model of the environment is fully 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 t. 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 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 given probability ϵ, it selects an action at random.

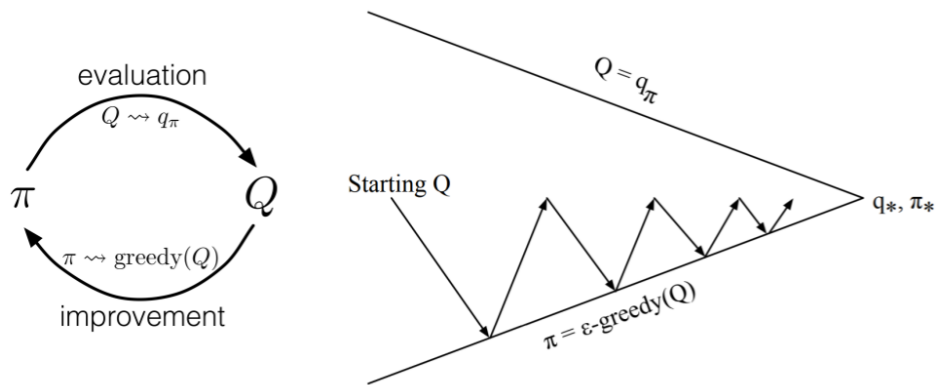Figure 3: On-policy methods improve stochastic policies by iteratively updating and evaluating the same policy and its Q-values to converge to the optimal policy π_* (Sutton & Barto, 2018). [5]

---

[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/)

There are two approaches to make use of stochastic policies to ensure all actions are chosen infinitely often. Figure 3 depicts the learning process of *on-policy* methods. They improve the stochastic decision policy, by continually estimating q_π in regard to π (left panel), while simultaneously driving π towards q_π, for example with an ϵ-greedy action selection (right panel). In contrast, *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, such as being able to learn from data generated by a human expert. On the other side, they have shown to 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 (12). Instead of considering state transitions, state-action transitions are considered in this case. The update rule is constructed as follows:

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

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 an 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 rather restrictive assumption considering the enormous state and actions spaces of real-world applications.

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

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

Here, the action-value estimates 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 (Watkins, 1989).

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

learn from experience, and can be easily applied to real-world problems with
low computational effort. Moreover, the mathematical complexity of the pre-
sented approaches is limited, and they can be easily implemented into computer
programs. TD 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 computa-
tional 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. 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 agent's experience with a small subset of visited
states is generalized to approximate values of the whole state set. Function ap-
proximation has been widely studied in supervised machine learning. Gradient
methods, as well as linear and non-linear model 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 also decreases 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 the 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* $\overline{\text{VE}}$, which

weights the prediction error with the state distribution $\mu(s)$:

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

Minimizing $\overline{\text{VE}}$ 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 (ANNs) have shown great success as function approximators, which coined the term *deep reinforcement learning* (Mnih et al., 2015; Silver et al., 2016). A simplified ANN that approximates the action-value function $q_\pi(s, a)$ can be found in Figure 4. In this example, the network estimates Q-values of the combination of four states and two actions. 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. For the model of our research, we use *double deep Q-Networks* (van Hasselt et al., 2016), which we will introduce in Chapter 4.3.2. Despite that, RL theory is mostly limited to tabular and linear approximation methods. Refer to Bengio (2009) for a comprehensive review of deep learning methods.

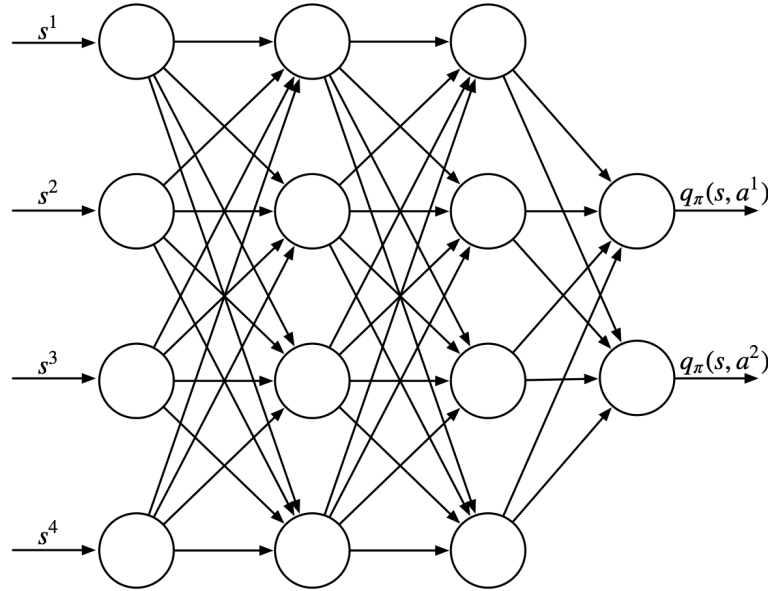Figure 4: A sample ANN consisting of four input nodes, two fully connected hidden layers and two output nodes. When approximating the action-value function $q_\pi(s, a)$ the number of input nodes equals the size of the state space and the number of output nodes the size of the action space. The learned connection weights $\mathbf{w}$ on the arrows between the layers are ommitted in this figure. [6]

2.5.7   Further Topics

The previous chapters provided a detailed overview of the most important concepts and mathematical foundations in RL. The literature has produced many more methods 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 $Q(\lambda)$ (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]Adapted from **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. We find that Germany is a suitable test bed for our experiments, 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 comparatively high renewable energy content in the energy mix causes electricity prices to be more volatile, which makes Germany an attractive location for VPP activities.

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 EV fleets. This thesis 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. The free-float concept adds additional complexity to our research, since the mobility patterns are inherently uncertain and difficult to predict. Customers of Car2Go or DriveNow 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 the carsharing service Car2Go, and collected freely available balancing market data from the GCRM platform website.[9] The data of the EPEX Spot market have kindly been provided by ProCom GmbH[10] for research purposes. In the next chapters the different data sets are described, as well the most important processing steps outlined.

3.1   Electronic Vehicle Fleet Data

The Car2Go data set consists of GPS data of around 500 Smart ED3 Fortwo vehicles in Stuttgart. These subcompact cars are equipped with a 17.6 kWh battery and a standard 3.6 kW on-board charger. They fully charge in about six

---

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

Table 1: Sample Raw Car2Go Data

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

Number Plate	Trip	Start Time	Start Latitude	Start Longitude	Start SoC (%)
S-GO2471	1	24.12.2017 20:10	9.19121	48.6890	94
S-GO2471	2	24.12.2017 23:10	9.15922	48.7885	71
S-GO2471	3	25.12.2017 06:50	9.17496	48.7493	66

Number Plate	Trip	End Time	End Latitude	End Longitude	End SoC (%)
S-GO2471	1	24.12.2017 23:05	9.15922	48.7885	71
S-GO2471	2	25.12.2017 00:40	9.17496	48.7493	62
S-GO2471	3	25.12.2017 08:25	9.2167	48.7875	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

to seven hours and can reach a maximum driving distance of 145 km according to the manufacturer. When equipped with an additional 22 kW 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 data set 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, which are not displayed in the table. Especially relevant for our research is the SoC (%) and information whether the EV is plugged into a charging station or parked at a regular parking spot. Note that the data only contain EVs that are *available for rent*, meaning that they are currently not rented out by a customer. EVs which are plugged into a charging station are also not available until they have charged up to approximately 70% SoC. For further analysis, individual trips had to be reconstructed using the GPS data of the cars, moreover we inferred travelled trip distances and rental prices based on Car2Go specifications. See Appendix B for a detailed listing of the executed preprocessing steps.

3.2   Balancing Market Data

We use market balancing data from the German secondary reserve market, which contain weekly lists of anonymized bids between June 1, 2016 and January 1, 2018, and a data set of activated control reserve in Germany during the same period. The bidding data consist of the traded electricity product, the offered balancing power P^{bal} (MW), the capacity price p^c ($\frac{€}{MW}$), and the energy price p^e ($\frac{€}{MWh}$) of each bid. Four different products are traded, which are a combination of positive control reserve (POS) or negative control reserve (NEG) and the provided time segment, peak hours (HT) or non-peak hours (NT). 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, either partially or fully, is listed. Bids, which were not accepted by the TSOs are not listed. An exemplary excerpt of the data set is displayed in Table 3.

Table 3: List of bids of the German secondary reserve market for the tender period December 4, 2017 to December 11, 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

Continued from previous page

Product	Capacity Price	Energy Price	Payment	Offered	Accepted
. . .	. . .	. . .	. . .	. . .	. . .
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 for 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 $\overline{p}^c$ ($\frac{\text{\euro}}{\text{MW}}$) as the capacity price of the bid that was just barely accepted, whereas $\overline{p}^e$ ($\frac{\text{\euro}}{\text{MWh}}$) 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 (offered capacity of bids) meets demand (activated balancing power), the critical price $\overline{p}^e$ was determined.

Example: The critical prices for the secondary operating reserve tender interval of December 6, 2017 between 08:00 and 08:15 were obtained as follows: Three suppliers submitted a negative reserve capacity of 5, 15 and 22 MW respectively (see Table 3, first three rows). The critical capacity price $\overline{p}^c = 200.3 \frac{\text{\euro}}{\text{MW}}$ is determined by the capacity price of the last (third) accepted bid in that time segment. The TSO reported that 18 MW of control reserve were activated between 08:00 and 08:15. Hence, the second bid determines the critical energy price $\overline{p}^e = 251 \frac{\text{\euro}}{\text{MWh}}$, as control reserve capacity gets activated according to an ascending order of the submitted energy prices. The second bidder would get compensated for the offered balancing capacity plus the actually provided balancing energy:

$(10.73\frac{\text{€}}{\text{MW}} \times 15 \text{ MWh}) + (251\frac{\text{€}}{\text{MWh}} \times 13 \text{ MWh} \times 0.25 \text{ h}) = 976.7$ €. Note that the second bidder would get compensated for providing 13 MW for 15 minutes (0.25 h), instead of the submitted 15 MW, since 5 MW of the total 18 MW of activated control capacity were already provided by the first bidder.
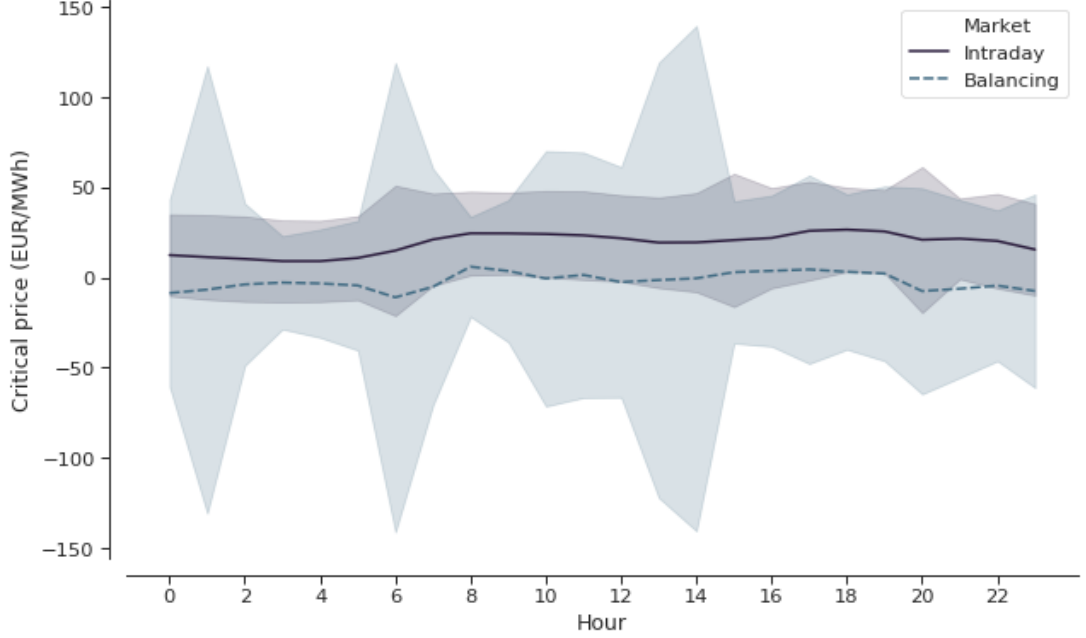

Figure 5: Daily critical electricity market prices (average, standard deviation) from June 2016 to January 2018. Highly volatile prices, e.g., between 12:00 and 14:00, demonstrate the benefits of an integrated trading strategy, which considers trading on both markets depending on the market conditions.

3.3   Intraday Market Data

The data from the EPEX Spot Intraday Continuous encompass order books and executed trades from June 1, 2016 until December 1, 2018. An extract of the data can be found in Appendix A. The list of trades contains information on the unit price p^u ($\frac{\text{€}}{\text{MWh}}$), 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 the decreased flexibility at that point in time.

On the spot market, electricity trades can have a very short lead time of up to 5 minutes before delivery. 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 minutes ahead. Similarly to the balancing market, the critical price $\overline{p}^u$ has to be determined for all intraday trading intervals. The critical unit price $\overline{p}^u$ is defined as the lowest price of all executed trades.

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

where $\mathcal{T}$ is the set of all trades in a bidding interval. *Example:* The critical unit price of the trades listed in Appendix A is $\overline{p}^u = 51.00 \frac{\text{€}}{\text{MWh}}$ (trades 8031392, 8031387 and 8031375). All buyers that submitted bids with a price higher than the critical unit price successfully procured electricity. Hence, accurate forecasts of the critical price allow to optimize the bidding behavior.

In Figure 5, a market comparison of the calculated critical prices of both presented data sets is depicted. The high standard deviation (error bands) of the prices illustrates the price volatility and highlights the possibility to profitably combine trading on both markets simultaneously.

4   Model

We start by giving a high-level overview of the model, which proposes a solution for utilizing a VPP portfolio of EVs to profitably provide balancing services to the grid. A control mechanism procures energy from multiple electricity markets, allocates available EVs to the VPP, and intelligently dispatches EVs to charge the acquired amount of energy. The model employs an RL agent that learns an optimal bidding strategy by interacting with the electricity markets and reacts to changing rental demands of the EV fleet.

We formulate the research problem as a *controlled EV charging* problem. The EV fleet operator represents the *controller*, which aims to charge the fleet at minimal costs. Therefore, it first 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. 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.

The rest of the chapter is structured as follows: The information and market assumptions are listed first. Next, we lay out the developed control mechanism, which we also illustrate with a comprehensive example. Finally, the RL approach is described, where we give details about the defined MDP and the employed learning algorithm. The required notation for the mathematical formulations are displayed in Table 4.

Table 4: 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	-

<div align="right">Continued on next page</div>

Continued from previous page

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

4.1   Assumptions

In order to evaluate and operationalize our model, the following assumptions about the available information and the electricity market mechanisms were taken:

I-1. Mobility demand information

> The controller is able to forecast the mobility demand of the EV fleet with different time-horizons based on historical data. More specifically, the controller can predict the amount of plugged-in EVs and consequently the available charging power P_t^{fleet} of the fleet in control period t. The prediction accuracy is increasing with shorter time horizons, from uncertain predictions one week ahead to very accurate predictions 30 minutes ahead. Past research presented such mobility demand forecast algorithms in the context of free-float carsharing (Kahlen et al., 2018, 2017; Wagner et al., 2016).

I-2. Critical electricity price information

> The controller is able to forecast electricity prices of spot and balancing markets based on historical data. More specifically, the controller 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 (see Chapter 3.2 and Chapter 3.3 for the critical price definitions). Electricity price forecasting is an extensively studied research area with well-advanced prediction algorithms (Weron, 2014; Avci et al., 2018).

We are confident that making the above information assumptions is feasible. Assuming available forecasting information is common practice in the VPP and EV fleet charging literature, see for example Vandael et al. (2015), Mashhour and Moghaddas-Tafreshi (2011b), Tomić and Kempton (2007), and Pandžić et al. (2013).

M-1. Balancing market mechanism

> We assume that the controller is able to submit bids of any quantity for single 15-minute market periods 7 days ahead to the balancing market. Since the critical capacity and energy prices are available by I-2, the controller submits bids in the form $(P^{bal}, \overline{p}^c, \overline{p}^e)$. By construction of the critical prices, the submitted bid will always get accepted by the TSOs and the offered balancing capacity fully activated.

M-2. Intraday market mechanism

> The controller submits bids in the form $(P^{intr}, \overline{p}^u)$ to the intraday market 30 minutes ahead. We assume that the order to buy will always get matched

until the minimal lead time of the trade (e.g., 5 minutes on the EPEX Spot Intraday Continuous). In reality, this is not always the case since trades are executed immediately and it is not guaranteed that a matching order to sell is submitted between the bidding time and the minimal lead time.

In summary, we are assuming that the controller always submits the optimal bids at the right time. In other words, every bid leads to the successful procurement of the desired amount of electricity. This assumption provides an upper bound for the fleet profits from trading EV battery storage on the electricity markets. However, this upper bound is only influenced by the accuracy of the electricity price forecasting algorithm, which we take as given by I-2. Incorporating the electricity forecasting aspect into our work, would well exceed the scope of this thesis. Furthermore, we assume that the controller is a price-taker. Due to the limited size of its bids, it is lacking the market share to influence prices on the markets. Similar assumptions have been made by Brandt et al. (2017) and Vandael et al. (2015).

4.2   Control Mechanism

The control mechanism constitutes the core of this research. It can be seen as a decision support system that can be deployed at an EV fleet operator to centrally control the charging of the fleet. Figure 6 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 places 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 on the intraday market for the market period 30 minutes ahead. The third phase, *Dispatch Phase*, takes places in every control period of $\Delta t = 5$ minutes. The controller has to dispatch available EVs to charge the procured electricity from the markets. The phase involves allocating individual EVs to the VPP and potentially refusing customer rentals to assure that all market commitments can be fulfilled.

The following sections highlight important parts of the three phases, formalize the model mathematically and provide detailed explanations and illustrative examples.

4.2.1   Fleet Charging Power Prediction

In a first step, the controller has to predict the available fleet charging power for the market period of interest (see (A) in Figure 6). The actual available fleet
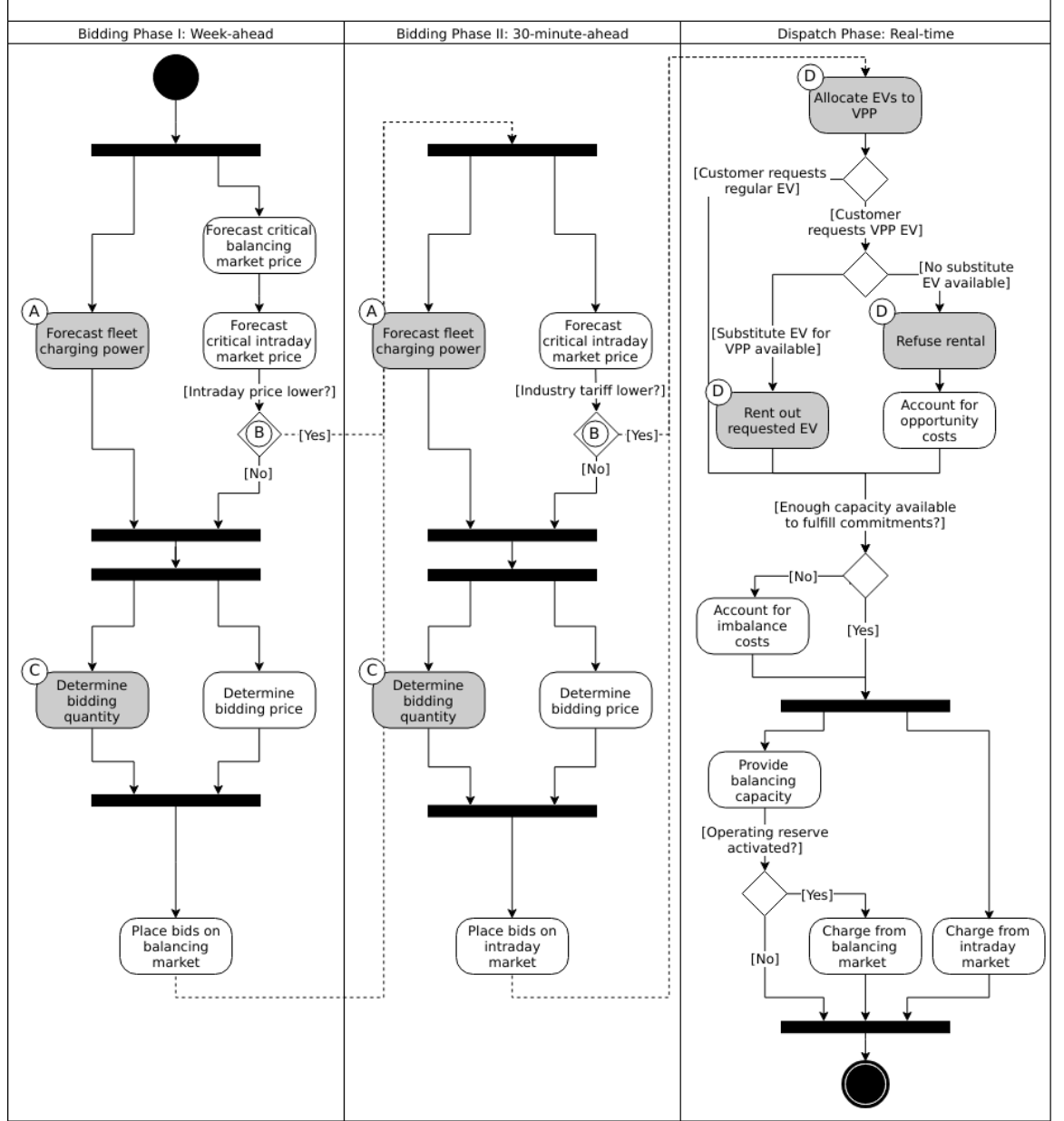
Figure 6: The central control mechanism of the fleet. In order to profitably charge the fleet, the controller has to take decisions in three distinct phases before electricity delivery. The controller needs to predict the available charging power, decide on the bidding quantities to submit to the markets, and intelligently dispatch EVs to charge the procured amount of electricity.

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 $t+1$.

By assumption I-1, the predicted charging power of the fleet is available to the controller. However, when the controller procures electricity from the markets, the fleet has to charge with the committed charging power throughout the whole market period h, otherwise imbalance costs occur. To address this risk, we define the predicted fleet charging power in a market period as the minimal predicted fleet charging power of all control periods in that 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, t its first control period and T the number of control periods in a market 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 (see ⒝ in Figure 6). Therefore, it compares the costs for charging electricity from the balancing market with the costs for charging from the intraday market. The cost function for procuring electricity from the balancing market is defined as follows:

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

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 term from kW to MW (see (17), second part). 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 $\overline{p}^c \in \mathbb{R}_0^+$ is never negative and therefore never results in costs for the fleet. The cost function for charging from the intraday market is defined similarly to (17):

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

Again, depending on the market situation, $\overline{p}^u \in \mathbb{R}$ can either be 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 amount of charged energy affects the costs.

The costs of charging influence the controllers market decision and the composition of the VPP portfolio. Depending on the charging costs and the associated risks with bidding on the markets the controller decides on the bidding quantities it should submit to each market. The next section eludes this core challenge of the controlled charging problem.

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 (see $\textcircled{C}$ in Figure 6). The bidding quantity determines the profits that can be made by charging at a cheaper market price than the flat industry tariff. On one hand, the controller aims to maximize its profits by procuring as much electricity as possible from the markets. On the other hand, 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 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 their batteries as a VPP. 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 of energy at all, even with refusing rentals. In case (b), the fleet also faces opportunity costs of lost rentals when individual EVs do not 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 bidding only 30 minutes ahead on the intraday market. This is because predictions are more uncertain with a larger time horizon. To account for all the mentioned risks, we introduce a *risk factor* $\lambda \in \mathbb{R}_{0 \leq \lambda \leq 1}$, 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:

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

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 (see (19), first condition). If the controller can charge cheaper at the regular industry tariff p^{ind} (see (19), second condition), 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 $\widehat{P}_h^{fleet}$:

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

where h is the market period of interest 30 minutes ahead. Here the undiscounted bidding quantity equals $\widehat{P}_h^{fleet} - P_h^{bal}$, since the amount of electricity that the controller procured from the balancing market does not need to be bought from intraday market again. When the controller submits bids to the intraday market in the second decision phase, it is able to correct bidding errors it made in the first decision phase, and optimize the bidding strategy of the 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 (see Ⓓ in Figure 6). Therefore the controller needs to detect how many EVs are eligible to be used as VPP in the control period t. An EV i is eligible if (a) it is connected to a charging station ($\mathbf{c}_i = 1$), and (b) it has enough free battery storage available ($\omega_i \leq \Omega - \gamma \Delta t$) to charge the next control period. Hence, the VPP is defined as:

$$VPP \doteq \left\{ i \in \mathcal{F} \mid \mathbf{c}_i = 1 \vee \omega_i \leq \Omega - \gamma \Delta t \right\},$$
(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. *Example:* Assuming a charging power of $\gamma = 3.6$ kW, an EV battery capacity of $\Omega = 17.6$ kWh, and control periods of 5 minutes, the amount of energy charged in one control period is 3.6 kW $\times \frac{5}{60}$h $= 0.3$ kWh. Hence, the maximally charged electricity ω_i of an EV to be eligible for VPP use is $17.6 - 0.3 = 17.3$ kWh.

Remember that the fleet has to provide the total 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 wants 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 the lost rental profits $\rho_{t,i}$. If the VPP's 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 the full amount of committed energy can be charged by the EVs of the VPP.

4.2.5   Evaluating the Bidding Risk

The controllers main 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:

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

where θ_λ is the set of the risk factors λ_h^{bal}, $\lambda_h^{intr} \in \mathbb{R}_{0 \leq \lambda \leq 1}$ 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 dependent on the chosen risk factors λ_h^{bal}, λ_h^{intr} (see (19) and (20)), which are omitted here for simplicity. 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:

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

A goal of this thesis is to develop a model that can be applied to previously unknown settings and learn from uncertain environments in the smart grid, such as new mobility contexts and smart electricity markets. As discussed in Chapter 2.4, RL is a suitable approach to achieve this goal and solve the proposed controlled charging problem (23). Before we introduce the developed RL approach in Chapter 4.3, we illustrate the three decision phases of the previously described control mechanism with a comprehensive example in the next section.

4.2.6   Example: Decision Phases

On August 9, 2017 at 3pm, the controller enters the first bidding phase for the market period $h = $ *16.08.2017 15:00-15:15*. It predicts that in that interval

250 EVs are connected to a charging station, resulting in 900 kW available fleet charging power ($\widehat{P}_h^{fleet} = 900$ kW), given the charging power of 3.6 kW per EV. Assuming the critical prices are $\overline{p}_h^c = 5\frac{\text{€}}{\text{MW}}$, $\overline{p}_h^e = -10\frac{\text{€}}{\text{MWh}}$, and $\overline{p}_h^u = 10\frac{\text{€}}{\text{MWh}}$ in that market period, the controller now evaluates the cheapest charging option. The flat industry electricity tariff is assumed to be $p^{ind} = 0.15\frac{\text{€}}{\text{kWh}}$. The costs for charging with the predicted amount of available power 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 ($900 \text{ kW} \times 0.25 \text{ h} \times 0.15\frac{\text{€}}{\text{kWh}} = 33.75$ €). In this example, the fleet operator will even get paid 6.25 € for charging the fleet by choosing the balancing market.

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 $P_h^{bal} = \widehat{P}_h^{fleet} \times (1 - \lambda_h^{bal}) = 900 \text{ kW} \times 0.7 = 630$ kW, submits a bid to the market, and accounts for charging costs of $C_h^{bal}(630 \text{ kW}) = -4.725$ €.

One week later, 30 minutes before electricity delivery time, the controller enters the second bidding phase. Due to the short time horizon, it predicts with high accuracy that only $\widehat{P}_h^{fleet} = 810$ kW (instead of 900 kW) is available for the market period *16.08.2017 15:00-15:15*. 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, the RL agent determines a intraday bidding risk of $\lambda_h^{intr} = 0.05$, and sets the bidding quantity to $P_h^{intr} = (810 \text{ kW} - 630 \text{ kW}) \times (1 - 0.05) = 171$ kW. The controller procures 171 kW from the intraday market and accounts for the charging costs of $C_h^{intr}(171 \text{ kW}) = 0.4275$ €.

At electricity delivery time, August 16, 2017 at 3pm, the controller detects 255 EVs that are eligible for VPP use; EVs that are connected to a charging station and have enough battery capacity left to charge during the next control period. It assigns 223 EVs to provide the total committed 801 kW 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 of the VPP are available to charge instead. The third rental has be to refused, since no EV can substitute the charging power. Hence, the controller has to account for the opportunity costs of that 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 an MDP, and second, the 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 costs, by setting the risk factors λ^{bal}, λ^{intr} each market period h. The risk factors determine the bidding quantities P_h^{bal}, P_h^{intr} that the controller submits to the balancing and intraday market, which in the end determine the fleet costs. The RL agent decides on the risk factors (i.e., takes an action) based on the observed state S_h every market period h (usually denoted as time step t in the RL literature). The optimal set of risk factors is learned by the RL agent through 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 influence the risk of bidding on the markets:

1. The bidding period's time of the day

   In times of volatile customer rental demand, for example during rush hour, the uncertainty on the guaranteed amount of available EVs increases. Bidding for these periods involves a higher 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 that means, larger VPPs are exposed to smaller risks: They have an increased probability that "lost" charging power, due to unforeseen EV rentals, can be substituted by other EVs of the VPP.

Since forecasts of available charging power are available by I-1, we define the predicted VPP size $|\widehat{VPP}|_h$ 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 brackets $\lceil x \rceil$, some readers might not be familiar with, mean the smallest integer equal or greater than x, which can also be written as ceil(x). *Example:* When the controller predicted 910 kW

available charging power, the required future size of the VPP to charge with the predicted power is ceil(910 kW/3.6 kW) = 253.

Based on the listed factors, we define the state space as the set of all valid values of the following tuple:

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

where:

- $t(h)$ is the bidding period's daytime in hours, with discrete values in the range $\left[0,\ 23\right] \in \mathbb{N}$.
- $|VPP|_t$ is the current VPP size, with discrete values in the range $\left[0,\ |\mathcal{F}|\right] \in \mathbb{N}$.
- $|\widehat{VPP}|_{h+2}$ is the predicted VPP size 30 minutes ahead, with discrete values in the range $\left[0,\ |\mathcal{F}|\right] \in \mathbb{N}$.
- $|\widehat{VPP}|_{h+(7\times H)}$ is the predicted VPP size 7 days ahead, with discrete values in the range $\left[0,\ |\mathcal{F}|\right] \in \mathbb{N}$.

Considering all possible combinations of the values, the state space encompasses $24\times|\mathcal{F}|^3$ states. When assuming a fleet size $|\mathcal{F}|$ of 500 EVs, that are 3×10^9 different states.

The agent takes actions by determining the risk that is associated with bidding on the electricity markets at each market period h. Hence, the action space is constituted by all combinations of valid values of the risk factors $\lambda^{bal}, \lambda^{intr}$:

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

where:

- λ^{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^{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.

Considering 20 discrete increments of the action values, the action space encompasses $20^2 = 400$ actions. The state space and action space were consciously discretized to achieve faster learning rates. Convergence in continuous spaces is theoretically achievable, but computationally more complex (Sutton & Barto, 2018). In order to facilitate faster learning in real-world settings, where long training periods are not desirable, we chose to not pursue this direction further.

The reward signal is naturally defined as the fleet costs that occurred in the last time step:

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

where C_h^{fleet} are the total accumulated fleet costs until the market period h. When

accumulating the occurred rewards for all time steps, we arrive at the total fleet costs, which we aim to minimize. See (22) for a complete formulation of the cost function. The agent's actions directly determine the occurred costs or profits, and are presented to the agent in form of a positive or negative reward signal. The particular challenge in the proposed RL problem is the significantly *delayed reward*. Choosing a risk factor in time step h determines the reward up to 672 time steps later (7 days, with 15-minute time steps), when the electricity from the balancing market has to be charged.

4.3.2   Learning Algorithm

This research proposes to solve the presented RL problem, with the double deep Q-Network algorithm (DDQN), developed by van Hasselt et al. (2016). DDQN is a state-of-the-art, model-free RL approach that uses a deep neural network as a function approximator to estimate optimal Q-values (see Chapter 2.5.6 for a explanation of function approximation methods). It combines the revolutionary deep Q-Network (DQN), originally proposed by Mnih et al. (2015) with double Q-Learning (van Hasselt, 2010). In double Q-Learning, experiences are randomly selected to update two different value functions to select and evaluate actions (in contrast to just one function for both tasks). DDQN has shown to reduce overoptimistic action-value estimates of the DQN algorithm, resulting in more stable and reliable learning results (van Hasselt et al., 2016). Combined with the *dueling network* architecture, proposed by Wang et al. (2015), this approach outperforms existing deep RL methods. Dueling networks lead to faster convergence rates in control problems with large action spaces than traditional single stream approaches. This property is especially beneficial for our proposed RL problem, as the defined action space (400 possible actions) is still comparably large in comparison to classical control problems (Sutton, 1996; Barto et al., 1983). In Figure 7, the conventional single stream approach (top) versus the dueling architecture (bottom) is depicted. The dueling architecture consists of a neural network of any shape with two streams that separately estimate the state-value and the action advantages. These estimates are later combined into Q-values (see Figure 7, green layer):

$$Q(s,a) = V(s) + \left(A(s,a) - \frac{1}{|\mathcal{A}|} \sum_{a'} A(s,a') \right) , \qquad (28)$$

where V and A are estimates of the value function and action advantages respectively, represented by the two different streams in the network. By subtracting the mean action advantages (see (28), last term), identifiability (V and A can be recovered, given Q) and stability of the optimization is ensured. The separated

streams allow to learn which states are valuable without having to learn each state-action interaction individually. Like this, a general state-value is learned that can be shared across many different actions, leading to a faster convergence of dueling architectures (Wang et al., 2015).
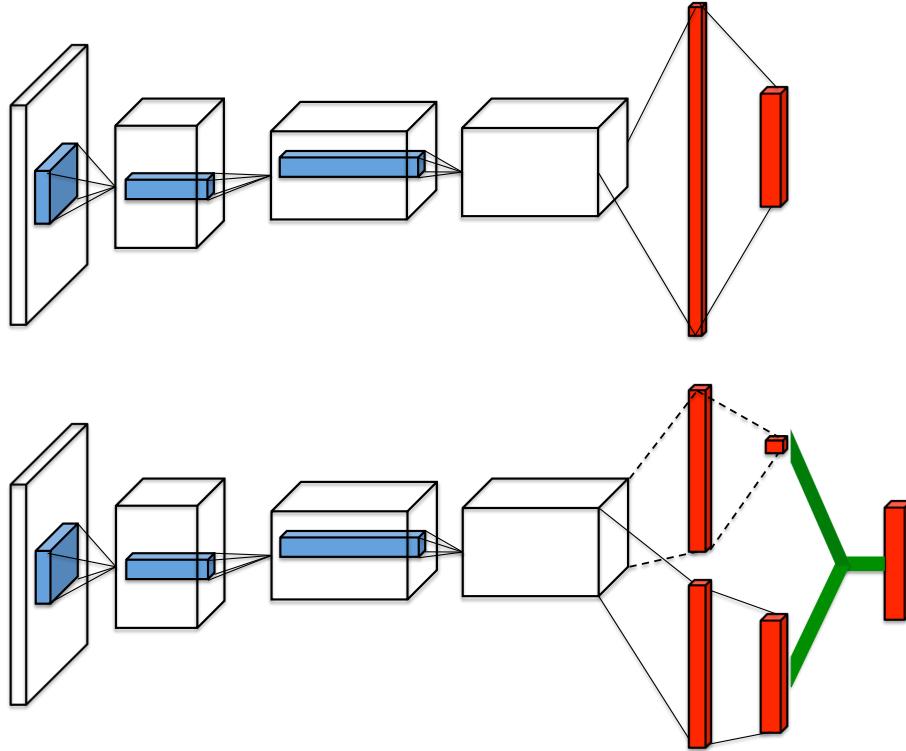

Figure 7: The dueling network architecture (Wang et al., 2015). The two separate streams achieve faster convergence rates by learning individual state values and action advantages first, and only later combining them into Q-values.

Our agent uses the dueling DDQN algorithm with a standard ANN architecture. The ANN consist of four input nodes, three fully-connected hidden layers with ReLU (Nair & Hinton, 2010) activation functions, and a linear output layer with two nodes. Further, an ϵ-greedy policy with a linear decreasing exploration rate was used. The exact network definition can be found in Appendix D. We implemented the RL agent with the neural networks API Keras[11, 12], which is a high-level abstraction layer of TensorFlow. TensorFlow is the de-facto standard for robust and scalable machine learning in industry and research (Abadi et al., 2016). Further, we used the shared research environment Google Colaboratory to train and evaluate the agent. It offers free access to computing resources that are optimized for training machine learning models.[13]

---

[11]`https://www.keras.io`

[12]`https://github.com/keras-rl/keras-rl`

[13]Google Colaboratory (`https://colab.research.google.com`) provides a NVIDIA Tesla K80 GPU, with 2880×2 CUDA cores and 12GB GDDR5 VRAM, and can be used up to 12 hours of consecutive training time.

5   Results

The following chapter will cover the main results of this research. First, our developed simulation environment is presented. Second, we examine the economic sustainability of an integrated bidding strategy that manages a portfolio of EVs. Here we compare the performance of the proposed control mechanism against predefined benchmark strategies. Third, the RL approach that aims to optimize the VPP portfolio is evaluated. In the last section, we show results of a sensitivity analysis on the prediction accuracy, which proved to be a decisive factor on the overall performance of the developed model.

5.1   Simulation Environment

As part of this research, we developed an event-based simulation platform called *FleetSim*. The platform allows researchers to develop and test out different smart charging methods and bidding strategies in realistic environment based on real-world data. In FleetSim, intelligent agents centrally control the charging operation of an EV fleet and are responsible to meet the customer mobility demand with sufficiently charged vehicles. At the same time, the agents can create a VPP of EVs, provide balancing services to the grid, and take part in electricity trading. Rental trips of the fleet are simulated on an individual level. Charging decisions for individual EVs at a particular point in time have consequences for the rest of the simulation. For example can an insufficiently charged EV battery cause a ripple effect of lost rentals when the EV does not get parked at a charging station again. The performance of the agents is evaluated based on the saved charging costs of buying cheaper electricity on the markets than charging with the regular industry tariff. Also, the opportunity costs of losing rentals and the imbalances costs are considered.

FleetSim facilitates easy sensitivity analyses, adaption to future market designs, and integration of novel data sets through its modular architecture and expandable design (see Figure 8). We consider FleetSim a research platform for sustainable and smart mobility, similar to PowerTac (Ketter et al., 2016). It builds on SimPy[14], a process-based discrete-event simulation framework. Fleet-Sim is available open source[15] and can be readily installed as a Python package.

---

[14]https://pypi.org/project/simpy/
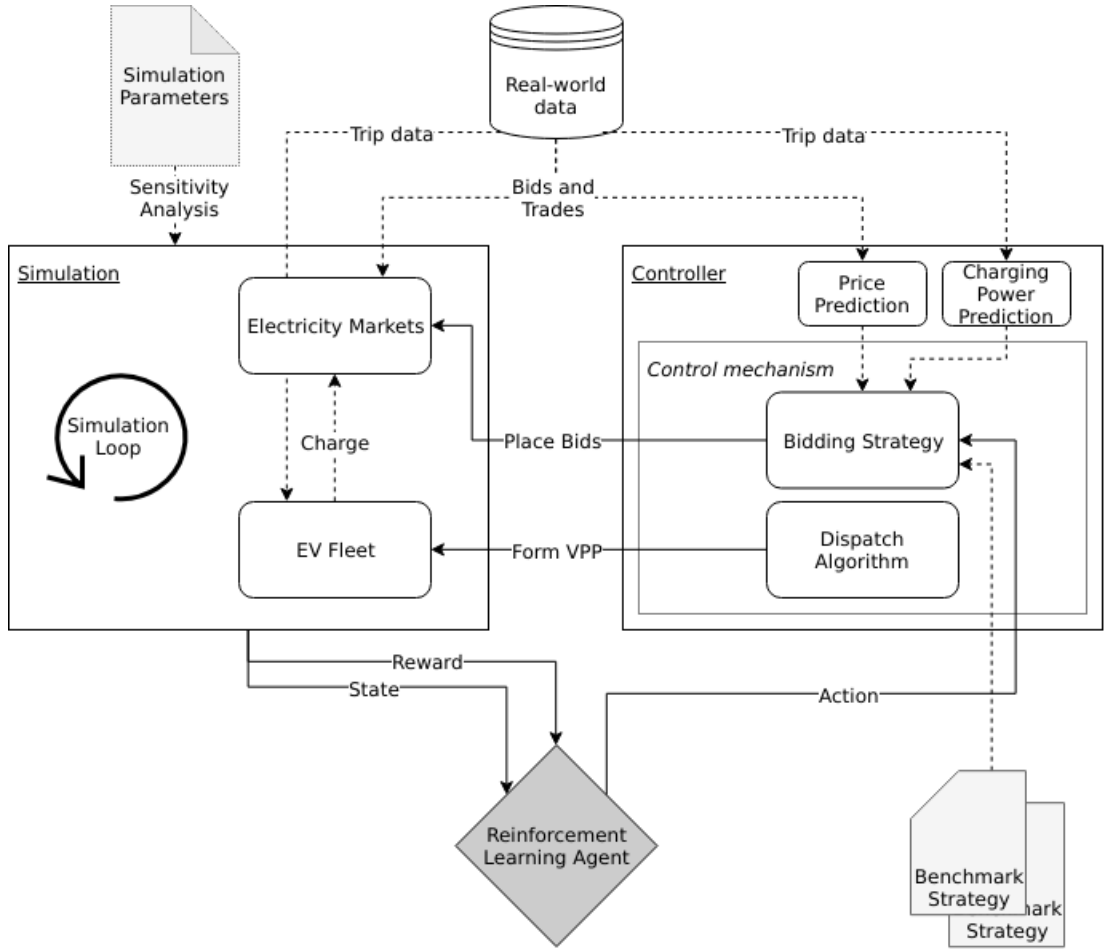[15]https://github.com/indyfree/fleetsim

Figure 8: The FleetSim platform. The modular design facilitates the straightforward integration of novel bidding strategies, future market designs, and new data sets. Sensitivity analyses can be easily performed by changing centrally stored simulation parameters.

In order to simplify comparability and focus real-world applicability of the analysis, we set the same parameters for all conducted experiments (see Table 5). They are corresponding to the real Car2Go specifications described in Chapter 3.1. Additionally, we fixed the unknown prediction accuracy of the fleets available charging power $\widehat{P}^{fleet}$ to an estimate of modern forecast algorithms performance. The impact of the prediction uncertainty on the results will later be determined in a sensitivity analysis.

Table 5: Simulation Parameters

Parameter	Value
EV battery capacity (Ω)	17.6 kWh
EV charging power (γ)	3.6 kW
EV range	145 km
EV rental tariff	$0.24^{16} \frac{€}{\min}$
EV long distance fee (> 200 km)	$0.29^{16} \frac{€}{\text{km}}$
Industry electricity price (p^{ind})	$0.15^{17} \frac{€}{\text{kWh}}$
Prediction accuracy $\widehat{P}^{fleet}$ week ahead	70%
Prediction accuracy $\widehat{P}^{fleet}$ 30 min ahead	90%

5.2 Integrated Bidding Strategy

Research question 1 examines whether a fleet operator can use a VPP portfolio of EVs to profitably bid on multiple electricity markets simultaneously. In Chapter 4.2, we proposed a central control mechanism that charges the fleet with an integrated bidding strategy. The following section evaluates the results of the controlled charging mechanism within the simulation environment.

Table 6 shows the descriptive statistics of the fleet utilization during a simulation run, with data from June 1, 2016 to January 1, 2018. It can be observed that (a) the volatility of EVs parked at a charging station is remarkably high (large standard deviation), and (b) the fraction of EVs that can be utilized for VPP activities is, with 3.55%, comparatively low. It is apparent that a high uncertainty and the low share of EVs that can possibly generate profits are challenging the economic sustainability of our proposed model. Figure 9 shows that despite a changing rental behavior throughout the day (e.g., rush hour peaks between

---

[16]Rental fees according to the Car2Go pricing scheme. See `https://www.car2go.com/media/data/germany/legal-documents/de-de-pricing-information.pdf`, accessed March 15, 2019.

[17]Average prices of electricity for the industry with an annual consumption of 500 MWh - 2000 MWh in Germany 2017 (BMWi, 2019).

7:00-9:00 and 17:00-19:00), the amount of EVs that can utilized for VPP activities is comparably stable throughout the day.
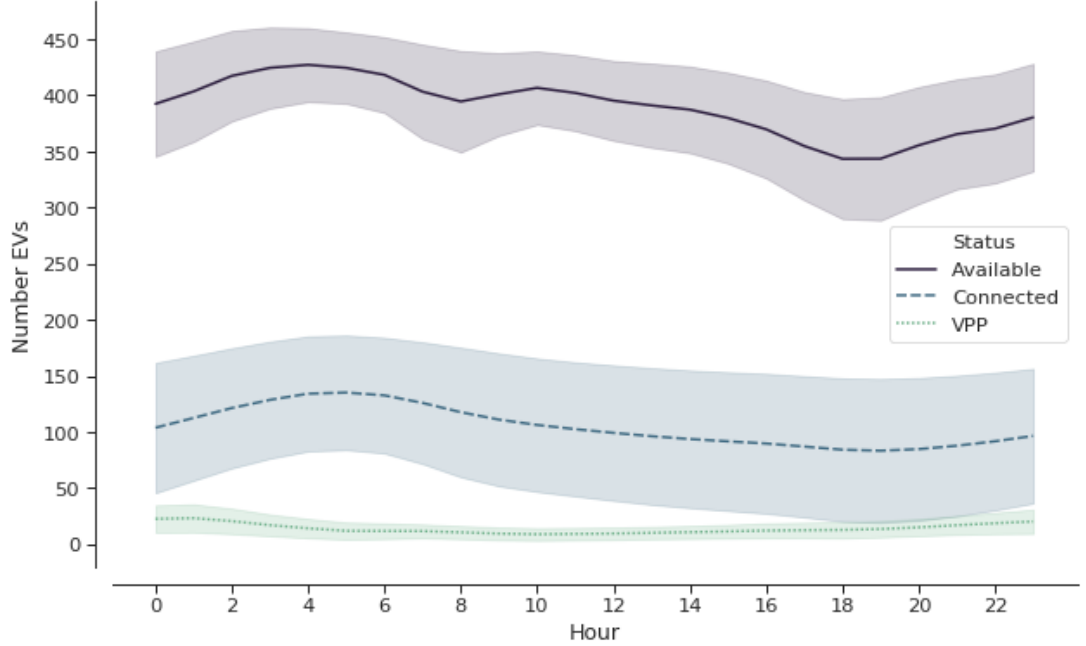

Figure 9: Daily fleet utilization (average, standard deviation) from June 2016 to January 2018. The blue error band is illustrating the large volatility in the amount of EVs that get parked at a charging station. The share of EVs that can be used as a VPP is on average only 3.55% of the fleet's size. Most of the EVs are either not connected to a charging station or are already fully charged.

Table 6: Summary Statistics of the Fleet (n=508)

Statistic	Average	Min	Max	Std
EVs available	389.64	165	496	49.18
EVs connected	61.23	34	290	61.11
VPP EVs	13.84	0	94	9.01

In order to evaluate the performance of the developed model, we defined several "naive" bidding strategies that we benchmark against our model. The strategies are naive in that sense that they are assuming a fixed risk associated with bidding on a specific electricity market. As opposed to the developed RL agent, they do not take information of their environment into account and do not adjust the bidding quantities dynamically. Instead, the controller discounts the predicted amount of available charging power with a fixed risk factor λ (see (19) and (20)). Naturally, the controller estimates a higher risk for bidding on the balancing market week ahead than on the intraday market 30 minutes ahead. We defined the following types of strategies:

1. Risk-averse ($\lambda^{bal} = 0.5$, $\lambda^{intr} = 0.3$)

The controller avoids denying rentals and causing imbalances at all costs. In order to not commit more charging power that it can provide, it places only bids for conservative amounts of electricity on the markets. The risk-averse strategies *Balancing* and *Intraday* are comparable to similar strategies developed by Kahlen et al. (2017, 2018).

2. Risk-seeking ($\lambda^{bal} = 0.2$, $\lambda^{intr} = 0.0$)

   The controller aims to maximize its profits by trading as much electricity on the markets as possible. It strives to fully utilize the VPP and allocate a high percentage of available EVS to charge from the markets. Due to the rental uncertainty and a low estimated risk, the controller is prone to offering more charging power to the markets than it can provide. This may lead to lost rental costs or even imbalances.

3. Full information

   The optimal strategy to solve the controlled charging problem. The controller knows the bidding risks in advance and places the perfect bids on the markets. In other words, it charges the maximal amount of electricity from the markets without having to deny rentals or causing imbalances due to prediction uncertainties.

In Table 7, the simulation results of all tested strategies are listed. As anticipated, the integrated bidding strategies outperform their single market counterparts. With this novel bidding strategy, the controller is able to capitalize on the most favorable market conditions and better utilize the VPP by increasing the share of traded electricity to charge the fleet. The integrated strategies are resulting in 49% - 54% more profits for the fleet than the single market strategies and provide 31% - 61% more balancing power for the grid.

A controller bidding according to the *Integrated (risk-averse)* strategy, pays approximately $0.35\frac{\text{€}}{\text{kWh}}$ less for charging the fleet than the regular industry price, summing up to an profit increase of up to 67040 € over the 1.5 year period. A controller bidding according to the *Integrated (risk-seeking)* strategy, is even more profitable, despite having to account for lost rental profits. On the other side, the controller caused imbalances (highlighted red) which lead to high market penalties or even exclusion from bidding activities. For this reason, imbalances need to be avoided, regardless of potential profits from a higher VPP utilization. We expect that the proposed RL agent learns a bidding strategy, which avoids imbalances while increasing profits at the same time. In the next section we will explore this particular challenge and present results of the RL algorithm.

Table 7: Bidding Strategy Results for a 1.5-year Period

	Balancing (risk-averse)	Intraday (risk-averse)	Integrated (risk-averse)	Integrated (risk-seeking)	Integrated (full information)
VPP utilization (%)	39	47	62	81	71
Energy bought (MWh)	803	985	1292	1681	1473
Energy charged regularly (MWh)	1278	1096	789	400	608
Average electricity price paid ($\frac{€}{kWh}$)	0.128	0.121	0.115	0.111	0.110
No. Lost rentals	0	0	0	1237	0
Lost rental profits (1000 €)	0	0	0	15.47	0
Imbalances (MWh)	0	0	0	1.01	0
Gross profit increase (1000 €)	43.62	45.08	**67.04**	**72.51**	77.36

5.3   Reinforcement Learning Portfolio Optimization

Research question 2 investigates whether an RL agent can optimize the integrated bidding strategy by dynamically adjusting the bidding quantities. The bidding quantities P^{bal}, P^{intr} are based on the evaluated risk associated with bidding on the individual electricity markets. In Chapter 4.3, we introduced an RL approach that learns the risk factors λ^{bal}, λ^{intr} based on its observed environment and received reward signals. The hyperparameters which we used to train the dueling DDQN algorithm and solving the controlled charging problem under uncertainty are presented in Appendix C. The values were determined manually through experimentation for the best results. The speed of convergence was also used as a criterion, since the training environment Google Colaboratory only allows up to 12 hours of computing time.

Furthermore, the imbalance costs β were set to an artificially high value to incentivize the agent to learn to always avoid imbalances. Whenever the agents takes an action that causes imbalances, it will receive a highly negative reward signal, leading to a low estimated Q-value of that chosen action in a specific state.
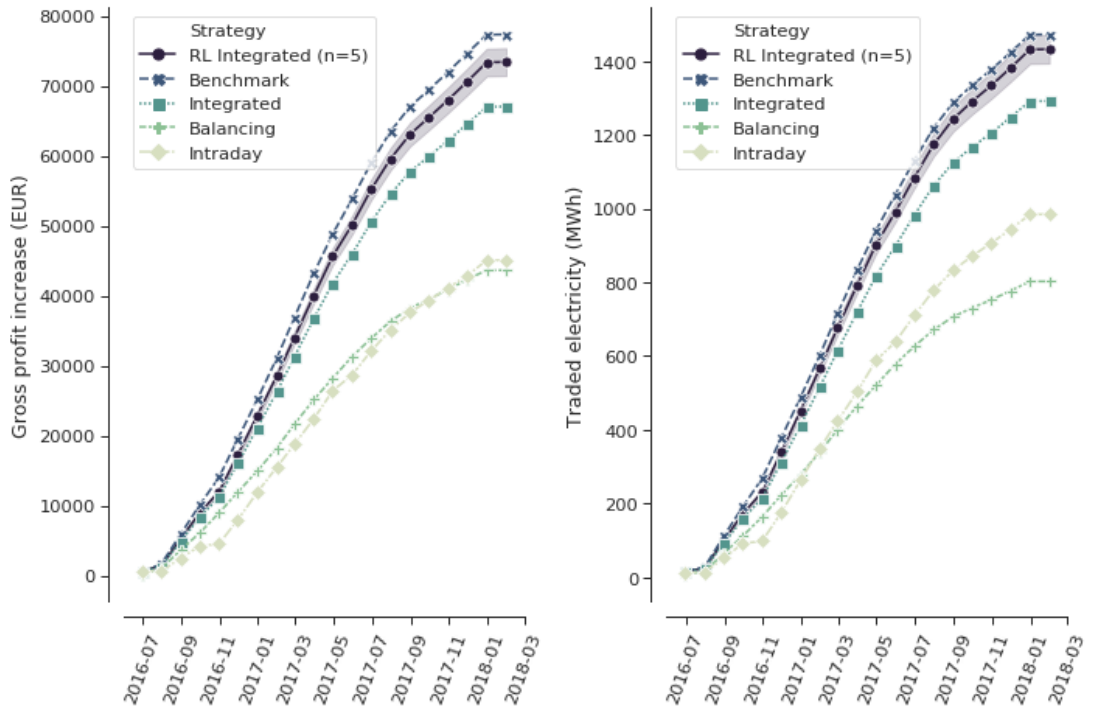
Figure 10: Comparison of gross profits and traded electricity between the proposed optimized integrated strategy and the other three naive charging strategies. The RL algorithm increases the achieved gross profits of the integrated bidding strategy on average by 12% and accomplishes nearly optimal results when compared to the benchmark strategy.

In Figure 10, the performance of the optimized integrated bidding strategy is presented. The proposed RL algorithm increases the gross profits of the fleet

on average (n=5) by approximately 72% - 75% when compared with the naive single market strategies and by approximately 12% when compared with the naive integrated strategy. To reach the optimal benchmark solution the gross profits would only need to be increased by another 3%. The fleet provided a total of 1461 MWh balancing power to the grid and thereby charged the EVs 25.3% cheaper than the regular industry price. In summary, the RL agent managed to optimize the VPP portfolio composition, increased the amount of traded electricity and avoided imbalance at the same time.

In another experiment, we evaluated the performance of the proposed RL algorithm in comparison to other RL algorithms with a simpler architecture. In particular, we were interested what impact modern advances in deep RL have on the ability to quickly learn to improve the agents policy, while still achieving good results after the whole training period. This question is especially relevant for the case, when no prior training for the fleet controller is possible, and the agent has to quickly learn to avoid procuring more energy from the markets that it can charge. Therefore, we removed the notion of imbalance costs and changed the simulation setup to instantly stop the training episode when imbalances occur. In this way, the agents learns to maximize its reward while circumventing imbalances at all costs. The agent achieves a higher reward the longer it trades electricity on the markets without committing to charge more electricity than it can. We compared the DQN algorithm (Mnih et al., 2015) with the Double DQN algorithm (van Hasselt et al., 2016) with and without the dueling architecture (Wang et al., 2015).

In Figure 11 the average learning performances of the tested RL approaches are displayed. The experiment shows that the dueling DDQN algorithm learns the fastest and shows a large increase in mean reward per action after roughly 60 episodes of training, which equals about 227 days of simulation time. The dueling DDQN algorithm shows the largest reward increase and highest reward per action after the whole training period, which makes it the best algorithm to solve the charging problem. Despite that, it still has a larger mean absolute error than the DDQN algorithm, indicating that it is more likely to cause imbalances with the dueling architecture than without. None of the algorithms determined a policy that never caused imbalances after training on the full 1.5 years of simulation time, which is about one hour of computing time. In other words, without prior training on existing data the RL agent would need more than one and a half years to learn to avoid imbalances. A possible explanation is the problem of learning from long delayed rewards, first discussed by Watkins (1989). Long delayed rewards increase the difficulty of RL problems, since the agent needs to connect occurring decision outcomes to specific actions way back in the past. In
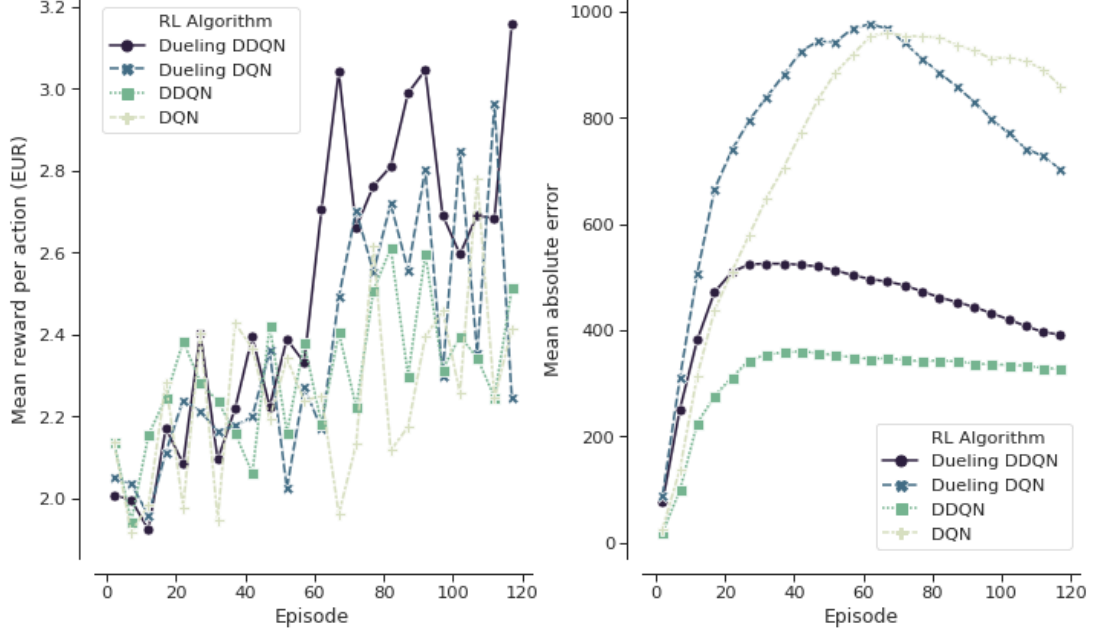
Figure 11: Comparison of the learning performance between the proposed RL algorithm and the other three simpler algorithms, averaged over 5 training attempts. Each training period is performed in 1.5 years of simulation time with real world data. The dueling DDQN algorithm (dark blue line) learns faster, and achieves better end-results than prior algorithms.

the case of the presented controlled fleet charging problem, this effect is especially pronounced because a negative reward signal, caused by imbalances, can occur up to 672 timesteps (one week) after the agent decided on the bidding quantity for the balancing market. In contrast, the reward signals from trading on the intraday market occur almost immediately after 2 timesteps (30 minutes).

In summary, both experiments show that our approach is able to learn a profit-maximizing bidding strategy under varying circumstances, without using any a priori information about the EV rental patterns. The proposed control mechanism improves existing approaches, and the RL agent can successfully optimize the VPP portfolio strategy by estimating the risks that are associated with bidding on the markets.

5.4   Sensitivity Analysis

The ability to accurately forecast the available fleet charging power plays an important role in determining the optimal bidding quantity to submit to the markets. If the fleet controller is certain about the number of connected EVs that it can use for VPP activities in the future, it can aggressively trade the available charging power on the markets, without being concerned about turning away customers or facing the risk of not being able to charge the committed amount of

electricity. In our previous experiments, we assumed a fixed prediction accuracy that we set to an estimate of what modern mobility demand forecasts algorithms can achieve. In order to test the robustness of the results and their dependence on the prediction accuracy, we conducted a sensitivity analysis. Therefore, we tested the previously introduced RL approach with increasing levels of prediction accuracy, from 50% to 100% accurate forecasts, 7 days and 30 minutes ahead.
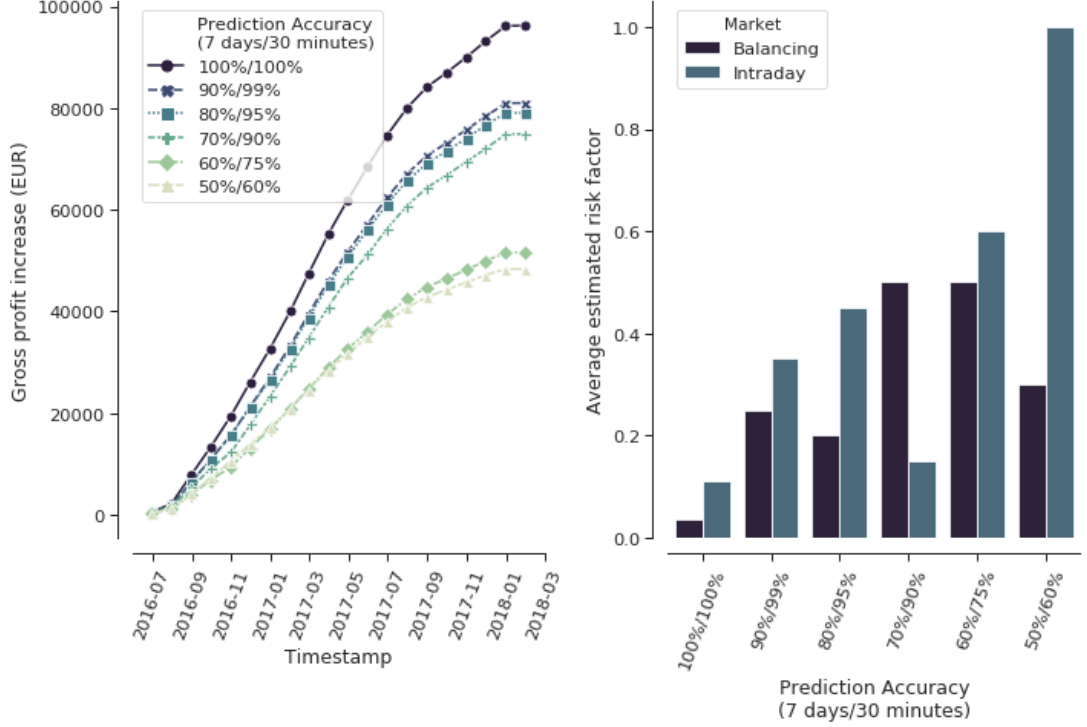

Figure 12: Sensitivity analysis of the forecasting algorithms prediction accuracy. With increasing accuracy the estimated bidding risks decrease, while the realized gross profits grow considerably.

Figure 12 depicts the results of the performed sensitivity analysis. The left plot shows the effect of the prediction accuracy on the total gross profit increase, whereas the right plot shows the effect on the learned risk factors of the RL agent. Intuitively, the realized profit increases with rising accuracy of the forecasts, while the estimated risk factors decrease with more accurate forecasts. Interestingly, the RL agent does not always estimate higher risks for bidding on the balancing market than on the intraday market, despite lower accuracy levels for predicting the available charging power 7 days ahead than 30 minutes ahead. This result indicates that the RL performance underlies some variations, that could not be fully eliminated in the conducted experiments. We hypothesize that the agent learns to avoid imbalances first, and only later learns to optimize the portfolio by fine-tuning the risk factors of both markets. We are confident that the agent's limited amount of training steps is the reason of these variations in the learning

success and expect to achieve more robust results with an increased training time.

Also remarkable is the magnitude of the prediction accuracy's effect on the profit increase (see Figure 12, left plot). After 1.5 years of simulation time, an RL agent that can rely on perfect predictions, with 100% accuracy, generates almost twice as much profit from trading electricity than an agent that can only rely on predictions with 50% and 60% accuracy, 7 days and 30 minutes ahead respectively. It is notable that the prediction accuracy has a larger effect (99.13% profit increase between lowest and highest accuracy) on the realized profit than the type of bidding strategy (73.10% profit increase between worst and best strategy).

6   Conclusion

Integrating volatile RES into the electricity system imposes challenges on the electricity grid. In order to ensure grid stability and avoid blackouts, balancing power is necessary to match electricity supply and demand. Balancing power can be provided by VPPs that generate or consume energy within a short period of time and offer these services on the electricity markets (Pudjianto et al., 2007). EV fleet operators can utilize idle vehicles to form VPPs and offer available EV battery capacity as balancing power to the markets. The fleet can offer balancing services directly via tender auctions on the balancing market or indirectly via continuous trades on the intraday market, where participants procure or sell energy to self-balance their portfolios (Pape et al., 2016). Both markets have complementary properties in terms of price levels and lead times to delivery, which motivates the creation of a VPP portfolio to profitably participate in both markets and extend the business model of the fleet.

However, there are certain risks associated with this business model extension. EVs can only be allocated to a VPP portfolio if they are connected to a charging station and have sufficient free battery capacity available, information which is unknown to the fleet at the time of market commitment. This uncertainty makes it difficult for fleet operators to estimate the size of the VPP and calculate the optimal bidding quantities. Moreover, if fleet operators offer more balancing power than they can provide, they face high imbalance penalties from the markets. In order to operate a sustainable business, fleet operators also need to balance VPP activities with their primary offering, customer mobility. If the fleet is denying customer rentals to ensure that it can fulfill the market commitments, it must account for opportunity costs of lost rentals that may be higher than potential profits of offering the cars' batteries as balancing power.

In the following chapter, we will summarize the conducted research of this thesis to address the aforementioned challenges. We will outline the contribution of this work and present and discuss the main results of our research. Finally, we will list the limitations of this study and give insights on further areas of research.

6.1   Contribution

The core contributions of this thesis are the following: First, we conceptualized a DSS for controlled EV charging under uncertainty. The DSS constitutes the core of a business model extension for fleet operators to manage a VPP portfolio of EV batteries. A controlled charging problem has been mathematically formulated

and a control mechanism introduced that aims to solve the proposed problem. Second, we developed an event-based simulation platform that facilitates fleet management research with real-world data. We evaluated various baseline bidding strategies within the simulation platform and tested out the behavior of intelligent agents in the context of controlled EV charging. Third, we proposed a novel integrated bidding strategy that offers balancing services of a VPP portfolio to multiple electricity markets simultaneously. Instead of submitting only conservative amounts of battery capacity to a single market as previous studies did (Kahlen et al., 2017, 2018), our proposed strategy aims to maximize profits by procuring electricity from the multiple markets to the greatest extent possible without causing imbalances. Fourth, we proposed the usage of an RL agent that learns to optimize the composition of the VPP portfolio and the associated risks of bidding on the markets. The agent dynamically determines risk factors, dependent on the predicted fleet information and market conditions. Therefore, we formulated an MDP that is designed for agents to work in previously unknown environments and uncertain conditions. The RL approach was developed with the focus on real-world applicability, fast convergence rates, and generalizability. We expect the proposed approach to work in a variety of different settings, with different kinds of EVs (e.g., electric bikes), independent of the geographical location of the fleet.

6.2   Main Findings

The proposed method was evaluated in a series of experiments with real-world carsharing data from Car2go in Stuttgart and German electricity market data from June 2017 until January 2018. The results show that the developed method improves existing approaches and would increase the gross profits of the fleet by roughly 75000 € over a 1.5-year period. We found that the integrated bidding strategy generates 49% - 54% more profits compared to the naive benchmark strategies. Fleet controllers following the integrated bidding strategy can mitigate risks and increase profits by exploiting market properties of both markets. The fleet would also contribute to the integration of RES by providing around 1500 MWh of balancing power to stabilize the electrical power grid. Since the inherently uncertain demand patterns of free float carsharing are imposing additional challenges on securely providing balancing services, better results are expected with other kinds of EV fleets.

Furthermore, we showed that the proposed RL agent could optimize the integrated bidding strategy under uncertainty by roughly 12%, missing the optimal result of a bidding strategy with full information available by only 3%. The agent was able to successfully estimate the bidding risk, avoid imbalances, and

keep lost rentals to a minimum. Additionally, we tested and compared the performance of various modern RL algorithms and found that recent advances in deep RL do improve the robustness and convergence rates in real-world applications, such as controlled EV charging. Despite these improvements, all tested RL approaches needed more than 1.5 years of simulation time to learn to avoid imbalance penalties, which makes an RL approach unsuitable to deploy in unseen fleet environments without prior training data.

A sensitivity analysis showed that the prediction accuracy of available fleet charging power has a large impact on the fleet profitability. If the forecasts about available EVs for VPP use in the future are very uncertain, the RL agent estimates high bidding risks and can only offer conservative amounts of balancing power on the markets. Surprisingly, the effect of prediction accuracy on the total gross profit is more pronounced than the choice of bidding strategy, which emphasizes the need for highly accurate mobility demand forecasting algorithms.

Our results compare well to other studies. While Kahlen et al. (2017, 2018) reported an annual profit increase of up to 121 € (163 $) per EV, the experiment conducted in this thesis yielded an annual increase of 98 € per EV from charging cheaper than the regular industry price. However, in their model, Kahlen et al. (2017, 2018) did not take imbalance cost into account, which we found to be a determining factor in the bidding process. When examining an RL approach of EV fleet charging schedules, Vandael et al. (2015) reported that after 20 to 30 days of training, the algorithm converged to a near-optimal solution. In our experiments, convergence was reached only after training more than 1.5 years of simulation time. In contrast to their study, in which the fleet participated in the day-ahead market, our research focused on balancing markets, which restrict participants to self-balance their portfolio on other markets and charge high imbalance penalties instead. Therefore, our proposed RL agent had to learn never to cause imbalances, which can occur up to one week after the agent decides on a risk factor for a single bid. Such a long-delayed reward is traditionally challenging for RL algorithms, which explains the lower convergence rate of our approach. Chis et al. (2016) investigated an RL approach to reduce charging costs of an individual EV and reported cost savings of 10% to 50%. While the authors considered a fixed driving schedule of a single EV, our research considered charging a whole EV fleet with previously unknown mobility patterns. Despite the additional uncertainty, we could achieve a cost reduction of charging the fleet by 25%.

6.3   Limitations and Future Work

The conducted research has certain limitations related to the modeled electricity markets. First, we assume that future price information of the markets is available to the fleet controller. The controller exploits this information to optimally place bids that always get accepted by the markets. Although highly accurate forecasting algorithms exist (Avci et al., 2018), we eliminated the remaining uncertainty. In reality, the markets may or may not accept the offered balancing services, which can compromise the profit of the fleet when it has to charge at the regular industry price instead.

Second, at a regulatory level, the electricity markets are currently not easily accessible to single EVs, EV fleets, or small VPPs. For example, the GCRM only allows actors to participate in the market, which can provide a quantity of at least 1 MW of balancing power over a 4-hour period. Since it is barely possible to overcome this constraint with a typical EV fleet consisting of 500 EVs with unknown rental patterns and the existing charging infrastructure, we ignored the constraint in our model. We propose changes in the current market design to give equal access to DER, such as EVs. In order to phase out conventional power plants and increase the share of renewables in the energy mix, balancing power needs to be provided from other resources. Instead of 4 to 8-hour segments, the markets should introduce bidding slots of 15 minutes or less. Moreover, the time between auction and physical delivery of balancing power should be reduced as far as possible. Smaller time frames mitigate forecasting errors, which are a major obstacle for renewable energy generators to efficiently offer their capacities to the markets. Future research should investigate how these market design changes and new modifications in the pricing structure (e.g., the German "Mischpreisverfahren") affect the profitability of using EV batteries as VPPs.

Finally, we see a limitation in the learning rate of the developed RL approach that concerns the applicability of such an algorithm in a real-world setting without previously available training data. Future work should test and evaluate RL control algorithms in physical systems or real appliances in smart electricity markets. It should be noted that despite the fast-paced development of RL approaches and its success in many research areas, it is not always the best solution for all problem types (Vázquez-Canteli & Nagy, 2019). Future research should investigate how RL approaches in the field of fleet control and VPP optimization compare against classical stochastic programming methods, for example from Pandžić et al. (2013). We would also like to emphasize the need for highly accurate mobility demand forecasting algorithms. Our results showed that the accuracy of such algorithms has a high influence on the effectiveness of the fleet control operation.

A   EPEX Spot Intraday Continuous Data

Table A.1: 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	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:53:26	8031391	59.00	10000	TenneT	TenneT	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:53:26	8031390	58.90	10000	TenneT	TenneT	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:53:15	8031389	52.30	7000	50Hertz	50Hertz	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:53:13	8031386	59.00	500	TenneT	TenneT	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:53:13	8031387	51.00	3600	Amprion	Amprion	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:53:13	8031388	52.00	1400	Amprion	Amprion	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:53:02	8031385	58.90	11000	TenneT	TenneT	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:52:38	8031380	60.00	10000	Amprion	Amprion	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:52:38	8031381	57.50	8000	Amprion	Amprion	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:52:38	8031382	58.00	2000	Amprion	Amprion	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:52:38	8031383	58.90	4000	TenneT	TenneT	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:52:38	8031384	60.00	4000	Amprion	Amprion	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:52:27	8031379	52.30	8000	50Hertz	50Hertz	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:51:33	8031378	66.00	5000	TransnetBW	TransnetBW	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:51:28	8031377	54.00	8000	Amprion	Amprion	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:51:24	8031376	54.00	7000	TenneT	TenneT	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:49:34	8031375	51.00	4000	TenneT	TenneT	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:49:26	8031374	54.00	5000	50Hertz	50Hertz	Quarter	07:15 - 07:30	2017-12-04
2017-12-04 06:49:23	8031373	55.10	8000	50Hertz	50Hertz	Quarter	07:15 - 07:30	2017-12-04

B   Car2Go Data Preprocessing

The raw Car2Go dataset contained about 63 million data points (7.82 GB) and has been processed to only contain about 1.25 thousand data points (100 MB), while still compromising all necessary information. The following preprocessing steps have been taken to prepare the data for further analysis:

1. Determine rental trips

   Rental trips were inferred based on changing GPS coordinates between two data points of the same EV (see Table 1, 4th to 5th row). Since the GPS accuracy of industry sensors is approximately 5 meters under open sky, and worse near buildings, bridges, and trees [18], measurement errors occur. To reduce the number of falsely identified trips, a decreased resolution of a 10-meter radius was calculated. Note that we hereby assume that customers do not undertake trips, which begin and end at the same location. In total, 1246040 customer rentals were determined.

2. Determine trip distances

   The trip distances were inferred based on the difference in SoC level before and after the rental and the average energy consumption per kilometer. Although the real distance that can be covered with a certain amount of energy depends on driving style and traffic situation, we are confident that the deviation will be small, since all trips take place within the same urban space.

3. Determine rental profits

   The rental profits were inferred based on the trip duration and the estimated trip distance. Car2Go has a minutely pricing scheme with an additional long distance fee that applies when the customer drives more than a specified range. The rental profits form an important piece of information for our model since they determine the opportunity cost ρ the fleet controller has to account for if it denies the rental in favor of VPP activities.

4. Clean data
   - *Service trips*: 3308 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

---

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

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*: 52 EVs were removed that showed incorrect charging behavior. These EVs reported a battery level increase of more than 20% between trips or during trips, while not being located at a charging station.

C   DDQN Hyperparameters

Table C.1: Hyperparameters for the Dueling Network DDQN Algorithm.

Hyperparameter	Value
Target network update interval	0.01
Optimizer	Adam (Kingma & Ba, 2014)
Batch size	32
Learning rate	10^{-3}
Replay buffer size	10^5
Warm-up steps	1000
Training steps	10^5
Discount factor γ	0.99
Reward normalization	r $\in$ [-1, 1]
Gradient clipping	Gradient in [-1, 1] for output layer
Network architecture	See Appendix D
ϵ-greedy annealing	Linear over 10^5 steps
ϵ-greedy minimum value	0.1

D   DDQN Network Definition

Python

```python
from keras.models import Sequential
from keras.layers import Dense, Activation, Flatten

# Define sequential neural network
model = Sequential()
model.add(Flatten(nb_states))
model.add(Dense(16))
model.add(Activation("relu"))
model.add(Dense(16))
model.add(Activation("relu"))
model.add(Dense(16))
model.add(Activation("relu"))
model.add(Dense(nb_actions))
model.add(Activation("linear"))

# Print summary
model.summary()
```

Console output

```
_____
Layer (type)                 Output Shape              Param #
=================================================================
flatten_2 (Flatten)          (None, 1)                 0
_____
dense_6 (Dense)              (None, 16)                32
_____
activation_5 (Activation)    (None, 16)                0
_____
dense_7 (Dense)              (None, 16)                272
_____
activation_6 (Activation)    (None, 16)                0
_____
dense_8 (Dense)              (None, 16)                272
_____
activation_7 (Activation)    (None, 16)                0
_____
dense_9 (Dense)              (None, 121)               2057
_____
activation_8 (Activation)    (None, 121)               0
=================================================================
Total params: 2,633
Trainable params: 2,633
Non-trainable params: 0
```

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... Zheng, X. (2016). TensorFlow: A System for Large-Scale Machine Learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*.

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

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike Adaptive Elements That Can Solve Difficult Learning Control Problems. *IEEE transactions on systems, man, and cybernetics*.

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

Bichler, M., Gupta, A., & Ketter, W. (2010). Designing Smart Markets. *Information Systems Research*, 688-699. Retrieved from `https://doi.org/10.1287/isre.1100.0316` doi: 10.1287/isre.1100.0316

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

BMWi. (2019). *Prices of Electricity for the Industry in Germany From 2008 to 2017* [Website]. Statista - The Statistics Portal. Retrieved from `https://www.statista.com/statistics/595803/electricity-industry-price-germany/` (Accessed April 10, 2019)

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

Hevner, March, Park, & Ram. (2004). Design Science in Information Systems Research. *MIS Quarterly*, 75. Retrieved from `https://doi.org/10.2307/25148625` doi: 10.2307/25148625

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

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

Ketter, W., Peters, M., Collins, J., & Gupta, A. (2016). A Multiagent Competitive Gaming Platform To Address Societal Challenges. *MIS Quarterly*, 447-

460. Retrieved from `https://doi.org/10.25300/misq/2016/40.2.09` doi: 10.25300/misq/2016/40.2.09

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

Kingma, D. P., & Ba, J. (2014). Adam: a Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

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

Nair, V., & Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th international conference on machine learning (icml-10)* (pp. 807–814).

Nguyen, H. T., & Le, L. B. (2016). Bidding Strategy for Virtual Power Plant With Intraday Demand Response Exchange Market Using Stochastic Programming. In *2016 ieee international conference on sustainable energy technologies (icset).* Retrieved from `https://doi.org/10.1109/icset.2016.7811763` doi: 10.1109/icset.2016.7811763

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-Tschansky, 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. (1996). Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding. In *Advances in neural information processing systems.*

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. (2010). Double Q-learning. In *Advances in neural information processing systems.*

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