Reinforcement Learning Portfolio Optimization of Electric Vehicle Virtual Power Plants

Master Thesis

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

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

April 21, 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.

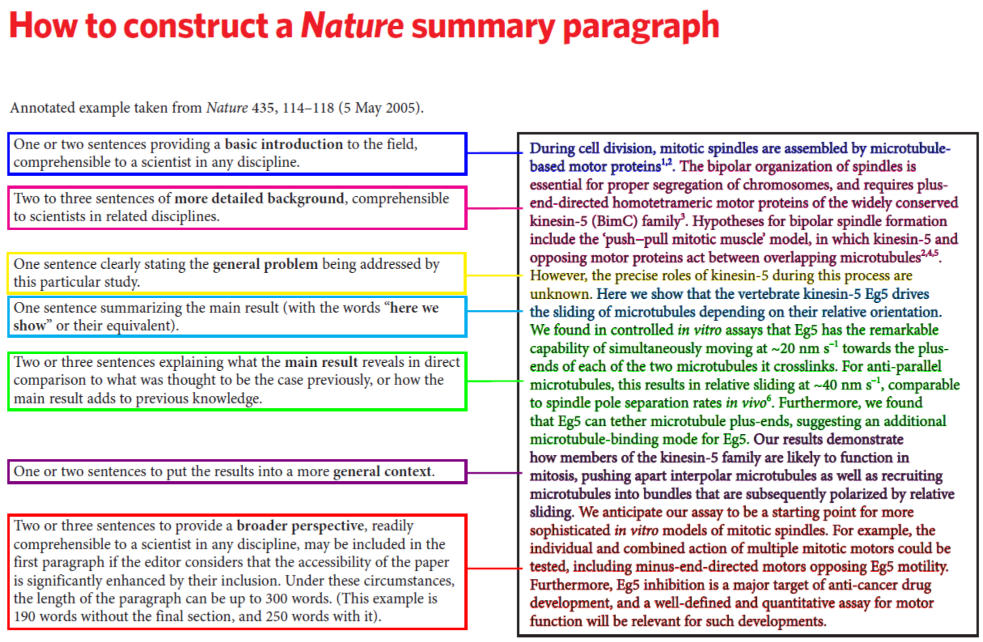
Tobias Richter

Köln, den 01.05.2019

Abstract

This is an abstract

- One or two sentences providing a basic introduction to the field, comprehensible to a scientist in any discipline.
- Two to three sentences of more detailed background, comprehensible to scientists in related disciplines.
- One sentence clearly stating the general problem being addressed by this particular study.
- One sentence summarising the main result (with the words “here we show” or their equivalent).
- Two or three sentences explaining what the main result reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.
- One or two sentences to put the results into a more general context.
- Two or three sentences to provide a broader perspective, readily comprehensible to a scientist in any discipline, may be included in the first paragraph



How to construct a *Nature* summary paragraph

Annotated example taken from *Nature* 435, 114–118 (5 May 2005).

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words “**here we show**” or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline, may be included in the first paragraph if the editor considers that the accessibility of the paper is significantly enhanced by their inclusion. Under these circumstances, the length of the paragraph can be up to 300 words. (This example is 190 words without the final section, and 250 words with it).

During cell division, mitotic spindles are assembled by microtubule-based motor proteins[1,2]. The bipolar organization of spindles is essential for proper segregation of chromosomes, and requires plus-end-directed homotetrameric motor proteins of the widely conserved kinesin-5 (BimC) family[3]. Hypotheses for bipolar spindle formation include the ‘push–pull mitotic muscle’ model, in which kinesin-5 and opposing motor proteins act between overlapping microtubules[2,4,5]. However, the precise roles of kinesin-5 during this process are unknown. Here we show that the vertebrate kinesin-5 Eg5 drives the sliding of microtubules depending on their relative orientation. We found in controlled *in vitro* assays that Eg5 has the remarkable capability of simultaneously moving at ~20 nm s^{-1} towards the plus-ends of each of the two microtubules it crosslinks. For anti-parallel microtubules, this results in relative sliding at ~40 nm s^{-1}, comparable to spindle pole separation rates *in vivo*[6]. Furthermore, we found that Eg5 can tether microtubule plus-ends, suggesting an additional microtubule-binding mode for Eg5. Our results demonstrate how members of the kinesin-5 family are likely to function in mitosis, pushing apart interpolar microtubules as well as recruiting microtubules into bundles that are subsequently polarized by relative sliding. We anticipate our assay to be a starting point for more sophisticated *in vitro* models of mitotic spindles. For example, the individual and combined action of multiple mitotic motors could be tested, including minus-end-directed motors opposing Eg5 motility. Furthermore, Eg5 inhibition is a major target of anti-cancer drug development, and a well-defined and quantitative assay for motor function will be relevant for such developments.

Contents

List of Figures

List of Tables

List of Abbreviations

ANN	Artificial Neural Network
DP	Dynamic Programming
DSO	Distribution System Operator
DDQN	Double Deep Q-Networks
EPEX	European Power Exchange
EV	Electric Vehicle
GCRM	German Control Reserve Market
GP	Genetic Programming
MAW	Mean Asymmetric Weighted Objective Function
MC	Monte Carlo
ML	Machine Learning
MDP	Markov Decision Process
PDF	Probability Density Function
RES	Renewable Energy Sources
RL	Reinforcement Learning
TD	Temporal-Difference
TSO	Transmission System Operator
V2G	Vehicle-to-Grid
VPP	Virtual Power Plant

Summary of Notation

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

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

$v_*(s)$	value of state s under the optimal policy
$q_\pi(s, a)$	value of taking action a in state s under policy π
$q_*(s, a)$	value of taking action a in state s under the optimal policy
V, V_t	array estimates of state-value function v_π or v_*
Q, Q_t	array estimates of action-value function q_π or q_*

d	dimensionality—the number of components of $\mathbf{w}$
$\mathbf{w}$	d-vector of weights underlying an approximate value function
$\hat{v}(s, \mathbf{w})$	approximate value of state s given weight vector $\mathbf{w}$
$\mu(s)$	on-policy distribution over states
$\overline{\mathrm{VE}}$	mean square value error

1 Introduction

1.1 Research Motivation

The global climate change is one of the most substantial challenges of our time. Carbon emissions need to be reduced and the shift to sustainable energy sources is inevitable. But the adaption of renewable energy is a complex matter: Solar and wind energy is intermittent and hard to integrate into the electrical power grid. Sustainable electricity production is dependent on the weather conditions, under- and oversupplies occur and are destabilizing the grid. Virtual Power Plants (VPP) play an important role in stabilizing the grid (?, ?). VPPs aggregate distributed power sources to consume and produce electricity when it is needed. At the same time, carsharing companies operate large, centrally managed fleets of Electric Vehicles (EV) in major cities around the world. These EV fleets can be turned into VPPs by using their batteries as combined electricity storage (RES?). In this way, EV fleets can offer balancing services to the power grid or trade electricity on the open markets for arbitrage purposes. Carsharing companies can charge the fleet (buy electricity) and discharge the fleet (sell electricity) when market conditions are favorable.

However, renting out EVs to customers is considerably more lucrative than using their batteries for trading electricity (?, ?). By making EVs available to be used as a VPP, carsharing companies compromise customer mobility and potentially the profitability of the fleet. Knowing how many EVs will be available for VPP usage in a future point of time is critical for a successful trading strategy. Accurate forecasts of rental demand help carsharing operators to determine the amount of electricity that they can trade on the market. EV fleet operators can also participate on multiple electricity markets simultaneously. They can take advantage of distinctive market properties, like auction mechanisms and lead times, to optimize their bidding strategy and reduce risks. Still the ultimate risk remains that the fleet 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 at the same time 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 learned using a *Reinforcement Learning* (RL) approach. A RL agent can adapt to changing rental demands and market conditions. It learns from historical data, the observed environment and realized profits to adjust its trading strategy dynamically. The following tasks are performed by the agent in real-time: 1) *Allocation of plugged in EVs to an idle or a VPP state*, 2) *Learn the optimal VPP portfolio composition* and 3) *Place bids* *and asks on corresponding electricity markets with an integrated trading strategy.*

We show that..

(?, ?)

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 simultanously? How does an integrated bidding strategy look like, which considers this case?*

2. *Can a reinforcement learning 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 (?, ?, ?). It contributes to the body of Design Science in Information Systems (?, ?) and draw upon work, which has been done in a multitude of research areas: Virtual Power Plants in smart electricity markets (?, ?), fleet management of (electric) carsharing as a new way of sustainable mobility (?, ?, ?), and advanced RL techniques for the smart grid (?, ?). We specifically build on research that has been carried out by ? (?, ?). In their papers, the authors concentrate on trading electricity on one market at a time. As proposed by the authors, we will 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 EVs batteries as storage option to the markets with an integrated bidding strategy.

From a business perspective, this thesis is relevant to carsharing companies that operating EV fleets, such as Car2Go or DriveNow. We will show how these companies can increase their profits, using idle EVs as VPPs 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. Further, the DSS will determine optimal capacity-price pairs to place bids on the individual electricity markets. Using an event-based simulation platform, we will estimate the profitability of the proposed methods. This will be done 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. The VPPs can consume excess electricity almost instantly and stabilize the power grid. When integrating more intermittent renewable electricity sources into the grid in the future, such balancing services will become indispensable. 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 prices. ? (?) 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. The VPPs can consume this electricity by charging the EV fleet and the sustainable energy production does not need to be curtailed. EV fleets 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

2.1 Smart Electricity Markets

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

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

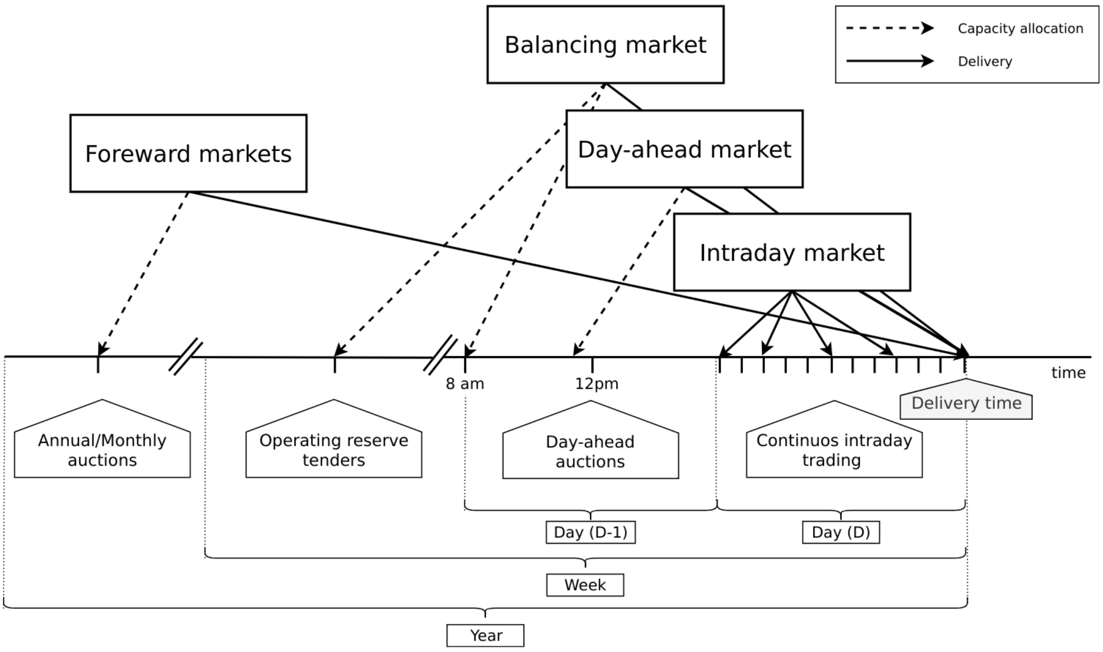


Figure 1: Interaction between electricity markets in relation to capacity allocation

2.2 Electricity Market Theory

2.2.1 Balancing Market

The balancing market is a tool to balance frequency deviations in the power grid. It offers auctions for primary control reserve, secondary control reserve as well as tertiary control reserve (minute reserve), which primarily differ in the required ramp-up times of the participants. As depicted in Figure 1, the balancing market can be seen as the last link in a chain of electricity markets (?, ?). [Explain!] In this study, we will look at the German Control Reserve Market (GCRM), one of the largest frequency regulation markets in the world. However, the presented concepts can be easily transferred to other balancing markets in unbundled energy systems, since the market design is similar (?, ?). Transmission Systems Operators (TSO) procure their required control reserve via tender auctions at the GCRM. The market conducts daily auctions for the three types of control reserve. This thesis focuses on the secondary operating reserve auction, in which participants must be able to supply or absorb a minimum of 1MW of power over a 4-hour interval with a reaction time of 30 seconds.[1] Since EV batteries can absorb energy almost instantly, when they are connected to a charging station, they are suitable to provide such balancing services. Operating reserve providers have to be qualified by the TSO to participate in the market and are able to reliably

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

provide the committed capacity. Although EV fleets are currently not qualified by the GCRM to be used as operating reserve, they could theoretically handle the minimum capacity requirements. Around 220 EVs would need to simultaneously charge at standard 4.6kW charging stations to provide 1MW of downward regulating capacity.

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

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

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

types of spot markets exist, the day-ahead market and the intraday market. In this research, we consider the European Power Exchange (EPEX Spot) as it is the largest electricity market in Europe, with a total trading volume of approximately 567TWh in 2018[3], but most electronic spot markets in western economies work with similar market mechanisms.

In Germany, the most important spot market is the day-ahead market with a trading volume of over 234TWh in 2018[3]. Participants place asks and bids for hourly contracts of the following day on the *EPEX Spot Day-ahead Auction* market until the market closes at 12pm on the day before delivery (see Figure 1). The day-ahead market plays an essential role in integrating volatile renewable energy sources (RES) into the power system (?, ?). Generators forecast the expected generation capacity for the next day and sell those quantities on the market (?, ?). After the market closes, the participants have the opportunity to trade the difference between the day-ahead forecast and the more accurate intraday forecast on the intraday market (?, ?). 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 (?, ?).

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

Purchasing electricity on the continuous intraday market is attractive for EV fleets with uncertain mobility demand. Due to the intradays market's short time before delivery, EV fleet operators can rely on highly accurate forecasts of available battery capacity to charge, before submitting an order to buy. In this way,

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

they can reliably charge at a potentially lower price at the intraday market than the regular industry tariff. In an integrated bidding strategy, EV fleet operators can, similarly to RES generators, balance out forecast errors of available battery capacity on the intraday market. Trades on the intraday market can complement bids that have been committed to other markets earlier (e.g., to the secondary operating reserve market).

2.3 EV Fleet Control in the Smart Grid

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

? (?) 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, ? (?) investigated the effect of smart charging on public charging stations in California. Controlling for arrival and departure times, the authors presented beneficial results for the distribution system operator (DSO) and the owners of EVs. Their approach resulted in a price reduction in energy bills and a peak load reduction. An extension of the smart charging concept is Vehicle-to-Grid (V2G). When equipped with V2G devices, EVs can discharge their batteries back into the grid. Existing research has focused on this technology in respect to grid stabilization effects and arbitrage possibilities. For instance, ? (?) showed that the usage of EVs can decrease average consumer electricity prices. Excess EV battery capacity can be used to charge in off-peak hours and discharge in peak hours, when the prices are higher. These arbitrage possibilities reverse welfare effects of generators and increase the overall welfare and consumer surplus. ? (?) 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.

? (?) investigated energy arbitrage profitability with V2G in the light of battery depreciation effects in the US. The results of their study indicate that large-scale use of EV batteries for grid storage does not yield enough monetary benefits to incentivize EV owners to participate in V2G activities. Considering battery depreciation cost, the authors arrived at an annual profit of only 6$ - 72$ per EV. ? (?) 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. ? (?) used EV fleets to offer balancing services to the grid. Evaluating the impact of V2G in their model, the authors conclude that V2G would only be profitable if reserve power prices were twice as high. Considering the results from the studies mentioned above, our research does not include V2G, since only marginal profits are expected.

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

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

Previous studies often assume that car owners or households can directly trade on electricity markets. In reality, this is not possible due to the minimum capacity requirements of the markets, requirements that single EVs do not meet. For example, the German Control Reserve Market (GCRM) has a minimum trading capacity of 1MW to 5MW, depending on the specific market. In order to reach the minimum capacity, over 200 EVs would need to be connected to the grid via a standard 4.6kW charging station at the same time. ? (?) introduced the notion of electricity brokers, aggregators that act on behalf of a group of individuals or households to participate in electricity markets. ? (?) and ? (?) successfully showed that electricity brokers can overcome the capacity issues by aggregating EV batteries. In addition to electricity brokers, we apply the concept of Virtual Power Plants (VPPs). VPPs are flexible portfolios of distributed energy resources, which are presented with a single load profile to the system operator, making them eligible for market participation and ancillary service provisioning (?, ?). Hence, VPPs allow providing regulation capacity to the market 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 charge and the availability of EVs, an intelligent agent decides in real-time which vehicles provide the capacity.

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

? (?) showed that it is possible to use free float carsharing fleets as VPPs to profitably offer balancing services to the grid. In their study, the authors compared cases from three different cities across Europe and the US. They used an event-based simulation, bootstrapped with real-world carsharing and secondary operating reserve market data from the respective cities. A central dilemma within their research was to decide whether an EV should be committed to a VPP or free for rent. Since rental profits are considerably higher than profits from electricity trading, it is crucial not to allocate an EV to a VPP when it could have been rented out otherwise. To deal with the asymmetric payoff, ? 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 (?, ?). In the paper, the authors used a sinusoidal time-series model to predict the available trading capacity. Another central problem for carsharing providers is that committed trades, which can not be fulfilled, result in substantial penalties from the system operator or electricity exchange. In other words, fleet operators have to avoid buying any amount of electricity, which they can't be sure to charge with available EVs at the delivery time. To address this issue, the authors developed a mean asymmetric weighted (MAW) objective function. They used it for their time-series based prediction model, to penalize committing an EV to VPP when it would have been rented out otherwise. Because of the two issues mentioned above, ? (?) 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. ? (?) 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 (?, ?). This research followed ? (?), who proposed to work on a combination of multiple markets in the future.

2.4 Reinforcement Learning Controlled EV Charging

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

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

? (?) facilitated learned EV fleet charging behavior to optimally purchase electricity on the day-ahead market. Similarly to ? (?), 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, ? (?) 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). ? (?) 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.

? (?) proposed a multi-agent approach for residential demand response. The authors investigated a setting in which 9 EVs were connected to the same transformer. The RL agents learned to charge at minimal costs, without overloading the transformer. ? (?) utilized *W-Learning* to simultaneously learn multiple policies (i.e., objectives such as ensuring minimum battery charged or ensuring charging at low costs). ? (?) 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. ? (?) 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.

? (?) 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 ? (?), the authors employed standard *Q-Learning* and the ϵ-greedy approach for action selection. ? (?) 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 (?, ?). 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.

? (?) 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. ? (?) 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. ? (?) proposed a centralized controller for managing V2G activities in multiple microgrids. The proposed method considers mobility and electricity demands of microgrids, as well as SoC of the EVs. The authors formulated a MDP with discrete state and action spaces and use standard *Q-Learning* with ϵ-greedy policy to derive an optimal charging policy. The approach takes microgrid autonomy and electricity prices into special consideration.

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

2.5 Reinforcement Learning Theory

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

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

2.5.1 Markov Decision Processes

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

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

[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/)

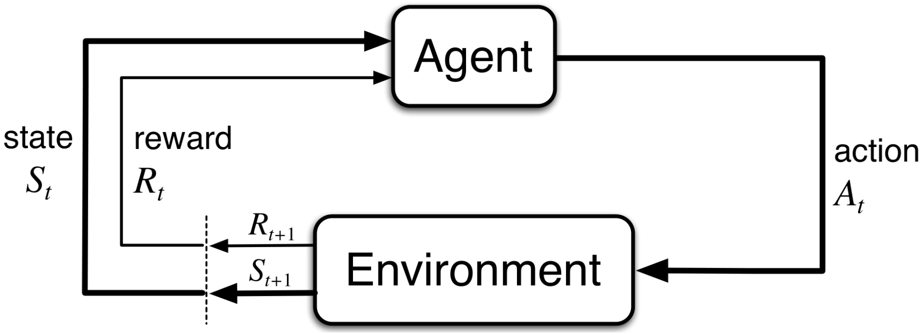


Figure 2: The agent-environment interaction in a Markov decision process (?, ?) [4]

an optimal choice of actions. In each discrete timestep $t = 0, 1, 2, ..., T$ the RL agent interacts with the environment, which is perceived by the agent as a representation, called *state*, $S_t \in \mathcal{S}$. Based on the state, the agents selects an *action*, $A_t \in \mathcal{A}$, and receives a numerical *reward* signal, $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$, in the next timestep. Actions influence immediate rewards and successive states, and consequently also influence future rewards. The agent has to continuously make a trade-off between immediate rewards and delayed rewards to achieve its long-term goal.

The *dynamics* of a MDP are defined by the probability that a state $s' \in \mathcal{S}$ and a reward $r \in \mathcal{R}$ occurs, given the preceding state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$. In *finite* MDPs, the random variables R_t and S_t have well-defined probability density functions (PDF), which are solely dependent on the previous state and action. Consequently, it is possible to define ($\doteq$) the *dynamics* of the MDP as follows:

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

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

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

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

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

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

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

2.5.2 Policies and Value Functions

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

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

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

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

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

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

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

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

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

2.5.3 Bellman Equations

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

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

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

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

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

2.5.4 Dynamic Programming

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

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

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

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

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

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

2.5.5 Temporal-Difference Learning

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

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

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

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

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

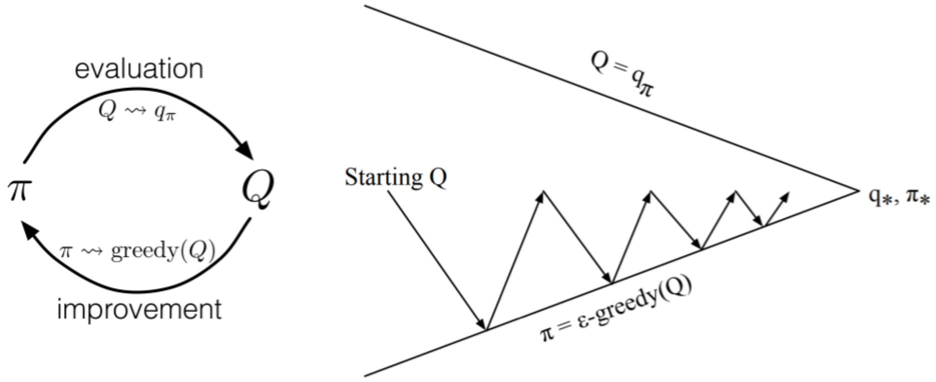

evaluation
$Q \rightsquigarrow q_\pi$
π
Q
$\pi \rightsquigarrow \text{greedy}(Q)$
improvement
$Q = q_\pi$
Starting Q
q_*, π_*
$\pi = \varepsilon\text{-greedy}(Q)$

Figure 3: On-policy control with Sarsa (?, ?). [5]

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

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

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

[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/)

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

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

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

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

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

2.5.6 Approximation Methods

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

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

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

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

In practice, deep artificial neural networks networks (ANNs) have shown great success as function approximators, which coined the term *Deep Reinforcement Learning* (?, ?, ?). 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 $\mathbf{w} \in \mathbb{R}^{d \times d}$ (?, ?). Advancements in the field of *Deep Learning* facilitated remarkable performance improvements in RL applications. Despite that, the RL theory is mostly limited to tabular and linear approximation methods. Refer to ? (?) for a comprehensive review of deep learning methods.

[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/)

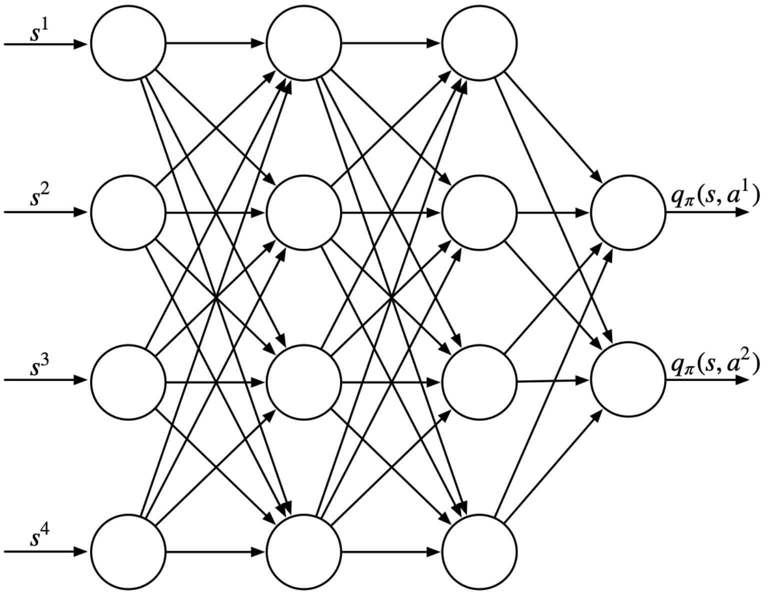


Figure 4: A sample ANN consisting of four input nodes, two fully connected hidden layers and two output nodes. When approximating the action-value function $q_\pi(s, a)$ the number of input nodes equals the size of the state space and the number of output nodes the size of the action space. The learned 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. In the research there are many more topics that were not covered here. *Eligibility traces* offer a way to more general learning and faster convergence rates. Almost any TD method can be extended to use eligibility traces, a popular methods is called Watkins's Q(λ) (?, ?). *Fitted-Q Iteration* (?, ?) 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 (?, ?) 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* (?, ?) is an multi-agent approach that aims to solve these difficulties.

3 Empirical Setting

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

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

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

3.1 Electronic Vehicle Fleet Data

The Car2Go dataset consists of GPS data of around 500 Smart ED3 Fortwo vehicles in Stuttgart. These subcompact cars are equipped with a 17.6kWh battery and a standard 3.6kW on-board charger. They fully charge in about six to seven hours and can reach a maximum driving distance of 145km according to the manufacturer. When equipped with an additional 22kW fast charger the charging time reduces to about an hour.

[7] `https://www.car2go.com`

[8] `https://www.drive-now.com`

[9] `https://procom-energy.de`

Table 1: Sample Raw Car2Go Data in Stuttgart

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

Table 2: Sample Processed Car2Go Trip Data in Stuttgart

Number Plate	Trip	Start Time	Start Latitude	Start Longitude	Start SoC (%)
S-GO2471	1	24.12.2017 20:10	9.19121	48.6890	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

In Table 1 the raw data is displayed, as we have obtained it by Car2Go. The dataset contains spatio-temporal attributes, such as timestamp, coordinates, and the address of the EVs in 5 minute intervals. Additionally, status attributes of the interior and exterior are given (not displayed). Especially relevant for our research is the state of charge (SoC, in %) and information whether the EV is plugged into a charging station. Note that the data only contain EVs that are *available for rent*, i.e., they are not currently rented out by a customer. EVs which are parked at a charging station are also not available until they have charged up to approximately 70% SoC. For further analysis, individual trips had to be reconstructed using the GPS data of the cars, moreover we inferred trip distances and trips prices based on Car2Go specifications. See Appendix **??** for a detailed listing of the executed preprocessing steps.

3.2 Balancing Market Data

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

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

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

Product	Capacity Price	Energy Price	Payment	Offered	Accepted
NEG-HT	0	1.1	TSO to bidder	5	5
NEG-HT	10.73	251	TSO to bidder	15	15
NEG-HT	200.3	564	TSO to bidder	22	22
...	...	...	...	...	...

Continued on next page

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 on 15-minute intervals in secondary operating reserve auctions will be possible in future energy markets. As mentioned in Chapter 2.2.1, the market design of the GCRM secondary operating reserve tender was adjusted in 2017. Daily tenders with 4-hour bidding intervals were introduced in favor of weekly tenders with only two time segments. This change represents the trend by the TSOs to change the market design in order to better include RES into the operating reserve markets (?, ?). 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 ? (?), we define $\overline{p}^c$ $(\frac{€}{\mathrm{MW}})$ as the capacity price of the bid that was just barely accepted, whereas $\overline{p}^e$ $(\frac{€}{\mathrm{MWh}})$ is the highest energy price that was payed for activated control reserve during that interval. For every 15-minute interval within the given tender period of one week, the activated control reserve in that interval was matched with the accepted bids in that tender period. At the point where supply, i.e., offered capacity of bids, met demand, i.e., activated control reserve, the critical price $\overline{p}^e$ was determined.

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

13MW for 15 minutes (0.25h), instead of the submitted 15MW, since in total only 18MW of control capacity was activated, which was partly fulfilled by the 5MW of the first bidder.

3.3 Spot Market Data

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

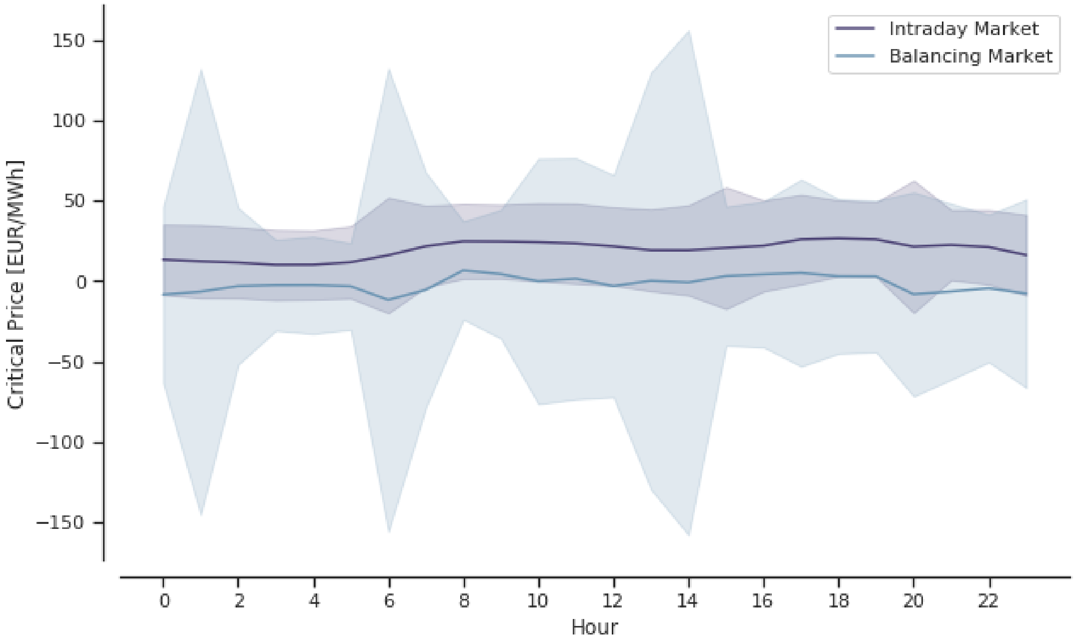


Figure 5: Daily critical electricity market prices (average, standard deviation) from June 6, 2016 to January 1, 2018. Highly volatile prices, e.g., between 12:00 and 14:00, illustrate the benefits of an integrated trading strategy, which considers trading on both markets depending on the market conditions.

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.

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

where $\mathfrak{T}$ is the set of all trades in a bidding interval. *Example:* The critical unit price of the trades listed in Appendix **??** is $\overline{p}^u = 51.00 \frac{€}{\mathrm{MWh}}$ (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. For a detailed description of the intraday continuous market see Chapter 2.2.2.

4 Conclusion

Integrating volatile renewable energy sources into the electricity system imposes challenges on the electricity grid. In order to ensure grid stability and avoid blackouts, balancing power is needed 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 (?, ?). 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 via continuous trades on the intraday market, where participants procure or sell energy to self-balance their portfolios (?, ?). 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. For a sustainable business model fleet operators also need to balance VPP activities with their primary offering, customer mobility. Denying customer rentals to ensure that the fleet can fulfill the market commitments, results in opportunity costs of lost rentals that compromise the profitability of the fleet.

In the following chapter, we will summarize the conducted research of this thesis to address the aforementioned challenges. Furthermore, the contribution of this work is outlined and the main results are presented and discussed. Finally, we list specific limitations of this study and give insights on further areas of research.

4.1 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 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, like previous studies did (?, ?, ?), our proposed strategy aims to maximize profits by procuring electricity from the multiple markets to the greatest extent possible without causing imbalances. Forth, we proposed the usage of a 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 a risk factor, depending available and predicted fleet and market information. Therefore, we formulated a MDP that is designed to work in previously unknown environments and uncertain conditions. The RL agent is designed with the focus on real-world applicability, generalizability and fast convergence rates. We expect the proposed approach to work with a variety of different EVs (e.g., electronic bikes) and independent of the fleets geographical location.

We evaluated the proposed method with real-world carsharing data from Car2go in Stuttgart and German electricity market data from June 2017 till January 2018. The results show that our approach would increase the gross profits of the fleet by roughly 78 000 € over the 1.5 year period. Free float carsharing has inherently uncertain demand patterns, since cars can be freely parked and picked up within the city area.

At the same time the VPP activities have an environmental impact by reducing CO_2 through more efficient use of renewables.

- What we have shown/found (include key numbers)
  - 1. RQ
    - ∗ Integrated improves existing approaches
    - ∗ Mitigate risk & improve profits through exploitation of different market setups
  - 2. RQ
    - ∗ RL Algorithm that
    - ∗ Online learning algorithm (difference? – real data?)
    - ∗ RL can learn to dynamically adjust bidding quantities by learning risk associated with bidding on each market. (What are the risks?)
    - ∗ RL Architecture/Advancements matters
    - ∗ RL takes a long time to learn
    - ∗ Charging infrastructure & Advancement in battery technology/charging technology matters

Discuss:

- In free float carsharing, mobility demand patterns are extremely uncertain. → lower bound.
- Compare to other studies!
  - Fleet Charging
    - No uncertainty
    - Only one market
    - No sensitivity on accuracy prediction (We found very important)
  - RL
  - Other approaches (VPP, stochastic) (?, ?) (no imbalance costs!) (?, ?)
- Discuss results?
  - Gross profits small - Mention before?
  - Policy
  - Investment
  - V2G?
  - Market Setup

4.2 Limitations

- Model:
  - Bidding Mechanism: one week ahead, always accepted
  - Policy & Regulation: EVs not allowed to provide balancing power, minimum bidding quantities 1MW.
  - Markets: Fleet is a price-taker, what about larger fleets? Simulate market influence
  - Deny rentals only in the same market period (More deny, less imbalance)
- RL: See (?, ?) conclusion for limitations.
  - Training time in real time. Generalization to other cities?

4.3 Future Research

- Model:
  - Investigate modern/current market design, that changed their bidding mechanisms to to better integrate renewable energy generators.
    - Daily/Day-ahead tenders with 4 hour market periods.

    Mischpreisverfahren

    i.e. daily w/ 4h slots. German "Mischpreisverfahren"
- RL: Long-delayed rewards, different reward structure, memory based
- Prediction Algorithms improvement, reference to sensitivity analysis