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, Philipp Artur Kienscherf

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

February 11, 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.

author

Köln, den xx.xx.20xx

Contents

List of Figures

List of Tables

List of Abbreviations

DSO Distribution System Operator

DP Dynamic Programming

EV Electric Vehicle

GP Genetic Programming

GCRM German Control Reserve Market

MAW Mean Asymmetric Weighted Objective Function

MDP Markov Decision Process

PDF Probability Density Function

RL Reinforcement Learning

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
$\Pr\{X = x\}$	probability that a random variable X takes on the value x
$\mathbb{E}[X]$	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, β	step-size parameters
γ	discount-rate parameter
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_*

1 Introduction (10%)

1.1 Research Motivation

- (Lopes et al., 2011)

1.2 Research Question

1.3 Relevance

2 Related Literature (10%)

2.1 Smart Charging and Balancing the Electric Grid with EV Fleets

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

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

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

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

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

& Kempton, 2007).

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

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

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

In a similar study, the authors showed that carsharing companies can participate in day-ahead markets for arbitrage purposes (Kahlen et al., 2018). In the paper, the authors used a sinusoidal time-series model to predict the available trading capacity. Another central problem for carsharing providers is that committed trades, which can not be fulfilled, result in substantial penalties from the system operator or electricity exchange. In other words, fleet operators have to avoid buying any amount of electricity, which they can't be sure to charge with available EVs at the delivery time. To address this issue, the authors developed a mean asymmetric weighted (MAW) objective function. They used it for their time-series based prediction model, to penalize committing an EV to VPP when it would have been rented out otherwise. Because of the two issues mentioned above, Kahlen et al. (2018) could only make very conservative estimations and commitments of overall available trading capacity, resulting in a high amount of foregone 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.

2.2 Reinforcement Learning in Smart Grids

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 3.2.

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

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

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

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

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

The approach takes microgrid autonomy and electricity prices into special consideration.

It should be noted that advanced RL methods and techniques are not the only solutions for problems in the smart grid, often basic algorithms and heuristics provide satisfactory results. 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 (Vázquez-Canteli & Nagy, 2019).

3 Theoretical Background (10%)

3.1 Electricity Markets

3.1.1 Balancing Market

3.1.2 Spot Market

3.2 Reinforcement Learning

The following chapter will give an overview about the most important RL ideas and concepts and will introduce the corresponding mathematical formulations. (Vázquez-Canteli & Nagy, 2019).

3.2.1 Notation

The input to the network $x \in \mathbb{R}^D$ is fed to the first residual layer to get the activation $y = x + \sigma(wx + b) \in \mathbb{R}^D$ with $w \in \mathbb{R}^{D \times D}$, and $b \in \mathbb{R}^D$ the weights and bias of the layer.

3.2.2 Elements of Reinforcement Learning

Policy

- Agent behavior at a given time
- Mapping states to actions
- Function or Lookup table
- Sufficient to determine behavior
- Policies may be stochastic, give probabilities for each action

Reward signal

- Goal of the RL problem
- Numeric signal the environment sends to the agent
- Agents objective is to maximize the reward signal on the long run
- Reward signal primary reason to change the policy: Low reward following an action of the policy may result in changing the policy to select another action
- Rewards determine the immediate desirability of a state
- Reward signals can be stochastic functions of the state and the actions

Value function

- Value of a state is the total amount of reward an agent can expect to accumulate over the future, starting from that state
- Values indicate the long-term desirability of states, taking future states and their rewards into account.
- We seek actions cause states of highest value, because these actions obtain the greatest amount of reward in int long run.
- Values must be estimated and re-estimated over the agents lifetime.
- Efficiently estimating values is the most important component.

Model of the environment

- Model allows inferences to be made about how the environment will behave. E.g., Given state and action the model predicts the next state and next reward.
- Model-based methods are used for **Planning**: Deciding on a course of action by considering possible future situations before they happen.
  - **Control**: Model-free methods are simpler methods, what are explicitly trial-and-error learners

Planning vs. Control

On-Policy vs. Off-Policy

- On-Policy: $\$\epsilon\$$-greedy policy

Exploitation-Exploration Trade-off

3.2.3 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 1 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 an optimal

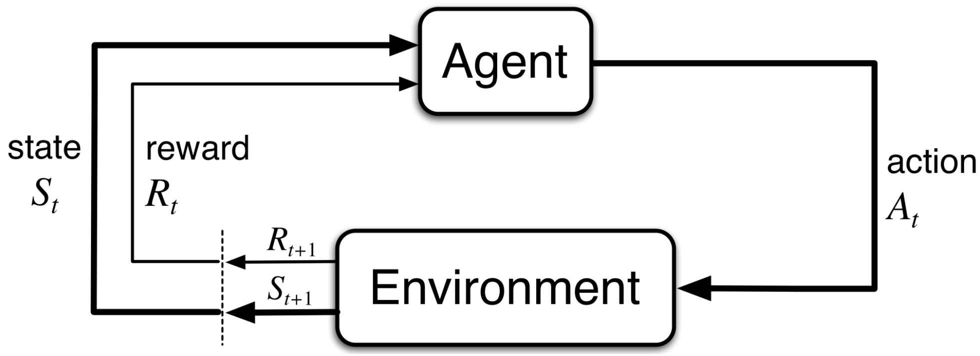

Figure 1: The agent-environment interaction in a Markov decision process (Sutton & Barto, 2018)

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 an action $a \in \mathcal{A}$. In *finite* MDPs, the random variables $\mathcal{R}_t$ 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 following:

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

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 thesis. The dynamics function allows computing the *state-transition probabilities*, another important characteristic of an MDP, as following:

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

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 very 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".

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

As mentioned, there are value function of states and of state-action pairs: 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 (Eq. 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 (Eq. 5).

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

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

An *optimal policy* π_* has a greater (or equal) expected return than all other policies. The *optimal state-value function* is defined as following:

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

The *optimal action-value function* q_* is defined accordingly.

3.2.5 Bellman Equations

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

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi[G_t|S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)\Big[r + \gamma v_\pi(s')\Big], \end{aligned} \tag{7}$$

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 and 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*:

$$\begin{aligned} v_*(s) &= \max_{a\in\mathcal{A}(s)} q_{\pi_*}(s, a) \\ &= \max_a \mathbb{E}_{\pi_*}[R_{t+1} + \gamma G_{t+1}|S_t = s, A_t = a] \\ &= \max_a \mathbb{E}_{\pi_*}[R_{t+1} + \gamma v_*(S_{t+1})|S_t = s, A_t = a] \\ &= \max_a \sum_{s',r} p(s', r|s, a)\Big[r + \gamma v_*(s')\Big] \end{aligned} \tag{8}$$

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 reference to specific policy, it has a unique solution independent from one. It can be seen as a equation system, which can be solved when the dynamics of the environment p are known. Similar Bellman equations to Equations (7) and (8) can also be 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 (8), by using experienced state transitions instead of expected transitions. The most common methods will be explored in the following chapters.

3.2.6 Tabular Methods

Dynamic Programming (DP) is a method to compute optimal policies, the ultimate 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 equation) is found, optimal policies can be easily obtained. Despite the limited utility of DP in real-world settings, it provides the theoretical foundation for all other RL methods. In fact, all of these methods try to achieve the same goal, but without the assumption of a perfect model of the environment (e.g., known *state-transition probabilities*, Eq. 2) and less computational effort.

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 (*policy evaluation*). A sequence of approximated value functions $\{v_k\}$ are updated using the Bellman equation for v_π (Eq. 7) 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 taking the best short-term action $a \in A$ at a given timestep. Alternating between these two steps monotonically improves policies and value functions till they converge to the optimum. This algorithm is called *policy iteration*:

$$\pi_0 \xrightarrow{\mathrm{E}} v_{\pi_0} \xrightarrow{\mathrm{I}} \pi_1 \xrightarrow{\mathrm{E}} v_{\pi_1} \xrightarrow{\mathrm{I}} \pi_2 \xrightarrow{\mathrm{E}} \dots \xrightarrow{\mathrm{I}} \pi_* \xrightarrow{\mathrm{E}} v_*, \tag{9}$$

where $\xrightarrow{\mathrm{E}}$ denotes a *policy evaluation* and $\xrightarrow{\mathrm{I}}$ denotes a *policy improvement* and π_* and v_* are the optimal policy and value function, respectively.

Note that in each iteration of the policy iteration algorithm, a policy evaluation has to 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:

$$\begin{aligned} v_{k+1}(s) &\doteq \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a] \\ &= \max_a \sum_{s',r} p(s',r|s,a)\Big[r + \gamma v_k(s')\Big], \end{aligned} \tag{10}$$

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 (8) 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.

- Learning from actual experience, Requires no propior knowledge of environment/model but can still achieve optimal behaviour
- Method to **learn** from partial returns (as opposed to averaging over complete returns from episode, as in MC, and opposed to **compute** value function in DP when full knowledge of MDP available)
- When model not available it is useful to estimate *action* values q_*. In DP, because of full knowledge of the environment, it was possible to just look on step ahead, and chose the action that leads the reward and next state. Without a model (model-free) that is not possible, the vale of each action needs to be estimated. Consequently, the algorithm needs to visit every *state-action* pair, as opposed to just the state. That is a problem, since not every *state-action* pair will be visited when following a deterministic policy, as in DP. Hence, with following a policy π the algorithm will only observe returns for one of the actions. (Since a deterministic policy $\pi(a|s)$ returns exactly one action, given the current state.) *Exploration* needs to be maintained. In order to evaluate the value function for *state-action* pairs q_π, continual exploration needs to be ensured. One way to achieve exploration is to only consider *stochastic* policies. Stochastic policies have a non-zero probability of selecting all actions in each state.
- On-policy (\$$\varepsilon$ \$-greedy) vs off-policy (control and exploration policies)

$$V(S_t) \leftarrow V(S_t) + \alpha\big[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\big] \tag{11}$$

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

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

3.2.7 Approximation Methods

On-policy Prediction with Approximation

1. Linear Methods
2. Non-linear Methods: ANN

(On-policy Control with Approximation)

4 Empirical Setting / Data (10%)

4.1 Carsharing Fleets of Electric Vehicles

4.1.1 Raw Data

The dataset consists of 500 EVs in Stuttgart. As displayed in Table 1, the data contain spatio-temporal attributes, such as timestamp, coordinates, and address of the EVs. Additionally, status attributes of the interior and exterior are given, the relative state of charge and information whether the EV is plugged into one of the 200 charging stations in Stuttgart.

Table 1: Raw Car2Go Trip Data from Stuttgart

Number Plate	Latitude	Longitude	Street	Zip Code	Engine Type
S-GO2471	9.19121	48.68895	Parkplatz Flughafen	70692	electric
S-GO2471	9.15922	48.78848	Salzmannweg 3	70192	electric
S-GO2471	9.17496	48.74928	Felix-Dahn-Str.45	70597	electric
S-GO2471	9.17496	48.74928	Felix-Dahn-Str.45	70597	electric
S-GO2471	9.17496	48.74928	Felix-Dahn-Str.45	70597	electric
Number Plate	**Interior**	**Exterior**	**Timestamp**	**Charging**	**State of Charge**
S-GO2471	good	good	22.12.2017 20:10	no	94
S-GO2471	good	good	24.12.2017 23:05	no	72
S-GO2471	good	good	26.12.2017 00:40	yes	81
S-GO2471	good	good	26.12.2017 00:45	yes	83
S-GO2471	good	good	26.12.2017 00:50	yes	84

4.1.2 Preprocessing Steps

4.2 Electricity Markets Data

4.2.1 Secondary Operating Reserve Market

4.2.2 Intraday Continuous Spot Market

5 Model: FleetRL (20%)

5.1 Information Assumptions

5.2 Mobility Demand & Clearing Price Prediction

5.3 Reinforcement Learning Approach

5.4 Bidding Strategy

5.5 Dispatch Heuristic / Algorithm

6 Evaluation (30%)

6.1 Event-based Simulation

6.2 Benchmark: Ad-hoc Strategies

6.3 FleetRL

6.4 Sensitivity Analysis: Prediction Accuracy

6.5 Sensitivity Analysis: Infrastructure Changes

6.6 Sensitivity Analysis: Bidding Strategy

7 Discussion (5%)

7.1 Generalizability

7.2 Future Electricity Landscape

7.3 Limitations

8 Conclusion (5%)

8.1 Contribution

8.2 Future Research

References

Bellman, R. E. (1957). *Dynamic Programming.* Courier Dover Publications.

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

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

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, 12). 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

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

Kahlen, M., Ketter, W., & Gupta, A. (2017). Fleetpower: Creating Virtual Power Plants in Sustainable Smart Electricity Markets.

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.

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

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

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

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

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

Mashhour, E., & Moghaddas-Tafreshi, S. M. (2011). 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

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

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

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

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

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

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

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

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

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