Johannes Exenberger, BSc BSc

Modeling Route Choice Behavior in Traffic Networks using Multi-Agent Reinforcement Learning

MASTER'S THESIS

to achieve the university degree of

Master of Science

Master's degree programme: Geospatial Technologies

submitted to

Graz University of Technology

Supervisor

Johannes Scholz, Ass.Prof. Dipl.-Ing. (FH) Dr.techn.

Institute of Geodesy

Graz, October 2020

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Date, Signature

Abstract

Traffic simulations are important for decision-making regarding the optimization of network flows to increase efficiency in real world traffic scenarios. Realistic models of human behavior are an integral part of such simulations, as human decision-making on different levels determines the occurrence and state of traffic in a transportation system. Regarding individual route choice, traffic models often rely on models of behavior that assume rational choice and utility maximization, implying that agents have absolute knowledge of the current network state and the cognitive capabilities to always choose the optimal path within a road network from their source to their destination. Although this concept is widely adopted as an approximation for the prediction of traffic, it does not depict real human behavior, which is a result of past experiences and personal preferences and cannot be described by the concept of rationality alone. Theories of rational choice and utility maximization to describe and predict human behavior have thus been challenged for a long time in scientific research. Furthermore, most choice models used in traffic modeling are static and do not reflect changing behavior of individuals due to learning and adaptation based on feedback from the environment. In multi-agent systems such as traffic systems, this also includes feedback from complex interaction patters where individual decision-making is influenced by the behavior of other agents in the environment. This work presents an approach to model individual route choice behavior in road networks using reinforcement learning, where multiple agents learn simultaneously in a shared environment. Agents choose their paths en-route link by link rather than choosing a predefined route, making decisions based on the current perceived traffic conditions and personal experience from interactions with each other and the environment. A novel approach is presented where an individual agent's state-action space in a network environment is represented as a decision tree rather than a tabular representation of the network, offering agents maximum flexibility in route choice while also reducing learning time.

Kurzfassung

Verkehrssimulationen bilden eine wichtige Grundlage für die Entwicklung von Strategien zur Optimierung von Verkehrsflüssen in Straßennetzwerken. Die Modellierung von menschlichem Verhalten im Straßenverkehr spielt dabei eine wichtige Rolle. Verkehrsmodelle greifen bei der Modellierung von individuellem Verkehrsverhalten wie beispielsweise der Wahl der Route häufig auf klassische Verhaltenstheorien aus den Wirtschaftswissenschaften zurück, welche meist auf der Hypothese von rationalem Verhalten und Nutzenmaximierung basieren. Diese Theorien folgen der Annahme, dass Verkehrsteilnehmer zu jedem Zeitpunkt alle Informationen über den aktuellen Verkehrszustand im gesamten Netzwerk besitzen und auf Basis dieser Informationen stets die optimale Route von ihrem Startpunkt zu ihrem Zielpunkt wählen. Diese Annahme ist jedoch keine realistische Abbildung menschlichen Verhaltens, welchen durch individuelle Erfahrungen und persönliche Vorlieben geprägt wird. Darüber hinaus werden Lern- und Anpassungsprozesse in den meisten Entscheidungsmodellen nicht abgebildet. In komplexen Szenarien mit mehreren gleichzeitig lernenden Agenten sind Interaktionen zwischen den Individuen und die daraus resultierenden Anpassungsprozesse jedoch prägende Faktoren individuellen Verhaltens. Diese Arbeit beschreibt einen Ansatz zur Modellierung von menschlichem Verhalten im Kontext von Routenentscheidungen in Verkehrsnetzwerken mit der Verwendung von Reinforcement Learning. Mehrere Reinforcement Learning Agenten lernen simultan in einem Netzwerk und entwickeln individuelle Routenpräferenzen basierend auf ihren Erfahrungen. Aktionsräume sequentieller Entscheidungsprozesse in Straßennetzwerken werden als Entscheidungsbäume modelliert, welche alle Routen von Startpunkt zu Zielpunkt im Netzwerk enthalten. Dies ermöglicht Reinforcement Learning Agenten eine höhere Flexibilität bei der Routenwahl und führt zu einer Reduktion der Lernzeit.

Acknowledgements

I would like to thank my supervisor Johannes Scholz for his academic support and encouragement during the process of writing this thesis and throughout the course of this master program. I furthermore am grateful to my parents for providing me with the opportunity to receive an academic education and for their constant support.

Contents

List of Figures

List of Algorithms

List of Acronyms and Symbols

Acronyms

ABM	Agent-Based Model
AI	Artificial Intelligence
BPR	Bureau of Public Roads
MAL	Multi-Agent Learning
MARL	Multi-Agent Reinforcement Learning
MAS	Multi-Agent System
MDP	Markov Decision Process
OD	Origin-Destination
RL	Reinforcement Learning
TD	Temporal Difference

Reinforcement Learning

$\mathcal{S}$	Set of states
$\mathcal{A}$	Set of actions
$\mathcal{A}^s$	Set of actions in state s
s, s'	States
a, a'	Actions
t	Time step in episode
S_t	State at time step t

A_t	Action at time step t
R_t	Reward received at time step t (following action A_{t-1} in S_{t-1})
$R_{w,t}$	Weighted reward received at time step t based on congestion level
G_t	Return from time step t onward
$\mathbb{E}$	Estimation
P	Probability
$r(s, a)$	Reward function
π	Policy
$v_\pi(s)$	State-value function of state s following policy π
$q_\pi(s, a)$	Action-value function of state-action pair (s, a) following policy π
π_*	Optimal policy
$v_*(s)$	Optimal state-value function of state s
$q_*(s, a)$	Optimal action-value function of state-action pair (s, a)
$Q(S, A)$	Q-value of state-action pair (S, A)
δ	TD-error
α	Learning rate
γ	Discount factor
τ	Boltzmann temperature
ε	Exploration parameter for epsilon-greedy policy
λ	Risk-sensitivity coefficient
σ	Risk-sensitivity exponent
t_{init}	Start step

Multi-Agent Reinforcement Learning

$\mathcal{N}$	Set of agents

$\mathcal{A}$	Joint action set
$\mathcal{A}_i$	Set of actions available to agent i
$\mathbf{a}$	Joint action

Network Environment

G	Graph
$\mathcal{V}$	Set of vertices/nodes
$\mathcal{E}$	Set of edges/arcs/links
$\mathcal{P}$	Set of all simple paths
$\mathcal{P}_{(o,d)}$	Set of all simple paths for OD pair
$\mathcal{P}_i^k$	Choice set of agent i
v, v'	Nodes
v_s	Node location at state s
e	Edge
p	Path
o_i	Origin node of agent i
d_i	Destination node of agent i
k	Size of path set
$c()$	Cost function
d_e	Demand on edge e
c_e	Capacity of link e
l_e	Level of congestion of edge e

Chapter 1

Introduction

1.1 Reinforcement Learning for Route Choice Modeling: Motivations and Possibilities

In the 20th century, the car became the preferred mode of transportation throughout the world. For many decades, urban and transportation planning therefore strongly focused on individual traffic by car, further marginalizing other transportation alternatives. This dominant status of the car lasts until this day. According to the European Union, more than 80 percent of all passenger kilometers were still traveled by car in 2017 [100]. The persisting high number of individuals choosing the car for their daily mobility needs is causing a variety of problems, with severe impacts on humans and environment. Traffic congestion still is a major concern many cities and commuters are facing daily, especially in big urban agglomerations. This situation has implications for the individual, society and environment alike. People spend several hours in their cars every week because of traffic congestion, a situation that also results in higher stress levels. Furthermore, high emissions due to traffic congestion lead to air pollution, causing problems for the environment and posing a risk to public health. The economic impacts of traffic congestion is severe as well, with billions of euros and dollars estimated to be lost every year in gridlock in the EU and the US [45]. Those problems are prevalent in many urban areas of the world, making new concepts and solutions to increase efficiency in urban road traffic an important and urgent goal. This is a complex issue, as traffic is a dynamic system and the result of many individual choices and processes. For this purpose, traffic modeling is a crucial field that helps to understand the dynamics of road traffic and the causing factors for the occurrence of congestion, providing a foundation for possible measures to prevent it in the future. Models of urban traffic networks can also help to examine traffic patterns and provide solutions for future infrastructure projects with the goal to avoid congestion. As it is the case with any model, traffic

models cannot provide a perfect representation of reality, also having to make certain assumptions and to rely on simplifications of processes. One major difficulty is the modeling of human behavior in traffic situations. Individual decisions can often not be predicted, making the development of a model that represents individual decisions in a population difficult and inaccurate. Additionally, choice models have to rely on socio-economic data, which often has to be acquired with costly and time consuming surveys. Traditional choice models applied for traffic modeling are often derived from economics and game theory, presuming that individuals have perfect knowledge of the current traffic conditions in the network and therefore are able to anticipate the consequences of their decisions exactly, choosing an action rationally to maximize their personal utility. The parameters that define the personal utility can be manifold - regarding the modeling of route choices in traffic networks, each individual normally seeks to minimize his or her personal travel time, therefore choosing the fastest path to travel on. While this simplified view can be a valid approximation of real world processes, it does not depict human behavior and the complex interactions between individuals, where personal utility functions might deviate from one person to another and past experiences and interactions with others play a crucial role in the decision-making process. Road traffic networks can be seen as systems where many individual participants try to maximize their personal utility and where every decision also has an impact on many other individuals in the network. Due to the complexity and dynamic nature of such systems, modeling individual behavior is difficult, as personal experience and ongoing interactions could lead to individual strategies and behavior that cannot be anticipated when designing the model. Most choice models used in traffic modeling are static and do not consider learning and adaptation, providing no way to change decision behavior with increasing experience. Rather than modeling experience in advance, creating simulations where personal experience is directly learned might offer new possibilities for modeling route choice behavior. Artificial intelligence research, mainly the field of machine learning, provide methods that can be used for such purposes. Especially reinforcement learning, one of the three main branches of current machine learning research alongside supervised and unsupervised learning, offers possibilities to be used in simulating individual decision-making processes. In reinforcement learning, artificial agents act and learn based on their experience without the need to explicitly know the dynamics of the environment they are moving in, nor do they have to rely on a supervisor that tells each agent what is or would have been the right choice. Actions are taken based on assumptions of the outcomes of those actions derived from personal experience, making each agent developing individual behavior based on preferences rooted in a subjective perception of the environment and its dynamics. It therefore also offers an opportunity to model traffic without the need for extensive data, with

the possibility to develop meaningful traffic models for networks where no data is available. Planned interventions in the infrastructure are one example, which could be evaluated beforehand. Reinforcement learning thus provides a way for route choice simulation without using explicit choice models, potentially providing a different view on human behavior and decision-making processes. The goal of this thesis is to examine the possibilities of the application of reinforcement learning to model realistic human route choice behavior in the simulation of traffic in road networks, drawing from established behavioral theories of human choice. It provides an approach for creating simulations in traffic networks with multiple agents that interact and find their preferred route from their source to their destination, developing individual behavior by learning form experience. Route decisions are made sequentially link by link, making it possible to adapt to traffic conditions and to avoid single links due to congestion. The main focus of this work is the application of reinforcement learning to model route choice behavior. The traffic simulation is thus implemented at a macroscopic level (according to the definition of macro- and microscopic traffic models), focusing on the routing aspect of each individual agent. It is not the aim of this thesis to provide a microscopic traffic simulation. Individual driving behavior like lane changing, acceleration etc. is therefore not considered and the application of the present reinforcement learning framework in microscopic traffic simulations is left for further research. Contributions are made in the field of reinforcement learning for traffic modeling, providing a new approach to let agents choose their path in the network en-route rather than relying on predefined paths. In this approach, network environments in the Markov decision process are modeled as decision trees, leading to increased learning speed by assuring that agents reach their destination node in every episode of the simulation and thus offering better scalability, while still providing maximum flexibility by letting agents choose their route link after link. To the authors best knowledge, this approach has never been proposed before in the literature. Additionally, this work proposes the integration of behavioral concepts such as risk awareness in combination with reinforcement learning to model human behavior in traffic networks more accurately. The thesis is structured as follows: The next section of this introductory chapter gives an overview over the current body of literature regarding the application of multi-agent systems including multi-agent reinforcement learning for traffic modeling and simulation. Chapter 2 covers the fundamental theoretical background in traffic modeling, human choice modeling and reinforcement learning needed for this thesis. Chapter 3 explains the reinforcement learning framework for route choice modeling in multi-agent traffic systems developed in this thesis. Chapter 4 describes the test case of the framework and offers a discussion and interpretation of the results. Finally, chapter 5 provides a conclusion of the work and potential possibilities for future research based on the results of this thesis.

1.2 Related Work

There is an abundant body of literature concerning agent-based systems for traffic simulation and transportation engineering, spanning applications for all of the four stages of the classic transport model: trip generation, trip distribution, mode choice and traffic assignment [74]. So far, agent-based systems and more precisely multi-agent systems have been used in various segments of transportation engineering such as traffic modeling and simulation, modeling behavior and decision-making processes of drivers, as well as traffic management and control. A comprehensive review of applications in those domains is given in [16], showing that research is not only restricted to road traffic, but considers air traffic and railway transportation as well. Another summary of current research in agent-based traffic simulations is given by Bazzan and Klügl [7], providing an overview of different applications where an agent-based approach is used in traffic simulation such as modeling travel demand, route choice or traffic flows, as well as providing agent-based traffic management solutions. A multi-agent framework for intersection control is presented in [34, 35]. [67] describes a system with multiple agents that learn lane selection strategies based on reinforcement learning. One important part of traffic simulations, which is also the focus of this work, is route choice modeling. As described previously, traditional decision theories assuming rationality are often used to simulate behavior in traffic networks, although this approach offers only a very simplified view on decision-making processes in real world traffic situations. This problem has already been addressed in the literature, resulting in various contributions of alternative route choice models that offer a more accurate description of the complex decision-making processes in road networks, ranging from game theoretic considerations to empirical studies in traffic psychology. [12] offers an overview of many choice models applied in the domain of transportation modeling. Approaches to incorporate prospect theory, probably the most influential behavioral decision theory, in traffic models are reviewed in [60]. Most route choice models applied in traffic modeling are static, not considering the effect of learning on decisions in repeated traffic scenarios. [17] offers an approach to model learning in route choice scenarios. The development of systems using intelligent, adaptive agents that are able to change their route based on current traffic conditions in the network is an interesting field of research regarding potential possibilities for traffic simulations. Applications in this domain has already been studied extensively. [6] describe a traffic simulation using agents that are able to re-route and choose an alternative path based on the perceived traffic conditions. The initial path is the shortest path from their origin to their destination considering travel times with no link demands. When experiencing congestion, agents are able to compute a new shortest path from their present node based on three alternative travel time estimations. Additionally, adaptive traffic lights are implemented in the model that choose between two strategies to

minimize waiting time. In [91], the authors investigate the effects of varying degrees of cooperation between agents and uncertainty of travel times on network flows in a congestion game using the Braess network, showing that cooperation leads to higher network efficiency. They propose the use of cognitive agents that choose their route and level of cooperation based on experience. Given the requirements of such traffic models - agents that learn from experience, ongoing agent-agent and agent-environment interactions - reinforcement learning is a concept well suited for this task. The potential of reinforcement learning for general routing tasks was recognized by researchers already several decades ago and much work has already been done in this domain. An adaptation of the famous Q-learning reinforcement learning algorithm suitable for optimization of packet routing tasks in networks with dynamic link demands and network topologies called Q-routing was proposed in [20]. This algorithm was developed further in [27], proposing predictive Q-routing, an algorithm that avoids routing packets on congested links and focuses on ongoing learning due to the non-stationarity of the network environment where links can quickly switch between states of congestion and non-congestion. Regarding reinforcement learning for route choice in traffic modeling, different approaches can be distinguished. Various works use reinforcement learning for route choice in traffic networks modeled as a single state multi-armed bandit problem. This is a simple reinforcement learning task where an agent learns to take the optimal action from an action set. After taking an action, the agent receives a reward, updates the reward expectations and tries again, without the environment transitioning to a new state [103]. Such decision processes are therefor not sequential and can not be modeled as a Markov decision process. In such scenarios, there is only one single state that represents the origin node. Agents can choose a route to their destination from a set of pre-computed paths, immediately receiving the cumulative reward for all links traveled, computing the reward based on current link demands. Thus also in this scenario, the decisions of all other agents are influencing the reward. After the agent receives the reward, a new episode starts. This approach is used for instance in [4, 56, 73]. [73] compares a stateless variant of Q-learning with several specific multi-armed bandit algorithms to evaluate performance for route choice tasks. [56] use the bandit approach for modeling route choice based on k-shortest paths for a multi-agent traffic scenario in combination with adaptive traffic lights that try to minimize waiting queues using the microscopic traffic simulation software SUMO [61]. [4] uses multiple selfish Q-learning agents that choose their route based on a set of k-shortest paths to model behavior in traffic networks. This is combined with a central authority that provides route recommendations based on a genetic algorithm trying to bring the network closer to the system optimum. The multi-armed bandit approach is an efficient way to implement a simple reinforcement learning problem for route choice modeling that provides

fast learning and is suitable also for bigger networks. As a drawback, the agents' flexibility in route choice is restricted and the travel times used in pre-computation of routes does not necessarily reflect the real travel times when links are showing high demands. Furthermore, adaptation and re-routing of the currently traveled path is not possible. Some of those drawbacks are also discussed in [6, 8]. As an alternative approach, route choice in networks can be modeled as a Markov decision process with state transitions as shown in [5, 38, 39], where Q-learning is used for individual agents to model a multi-agent traffic network scenario where every agent learns to choose the optimal path. The state-action space represents the whole network and links are chosen sequentially at every node in the network, called *edge based Q-learning* in [38]. The nodes in the network thus represent the states of the Markov decision process, the outgoing links from the current node represent the actions. This approach lets the agent travel every possible path in the network and choose links on the fly based on the current perceived traffic situation. As a downside, this approach can not guarantee that agents reach their destination in a time justifiable for simulations. This problem is addressed in [38, 39] by restricting the number of time steps per episode. If the maximum number of steps is reached, the episode terminates even if not all agents arrived at their destination. This could impair the learning process of agents, leading to slower progress in learning. Another problem is scalability: with increasing network size and length of paths that leads to more actions agents have to take en-route, the probability for them to reach their destination decreases, especially in the initial learning phase where choosing good actions is essentially based on trial and error. Reinforcement learning has also been used for direct optimization of network flows and autonomous driving. A deep reinforcement learning framework that manages to optimize congested traffic scenarios with only a few autonomous vehicles that control flows merely through adaptation of their own mode of driving and without any initial knowledge of prevalent traffic models is shown in [118].

Chapter 2

Preliminaries

2.1 Basic Concepts of Traffic Modeling

Modeling traffic has been an integral part of transportation engineering and planning for several decades already, with early applications dating back to the 1950s [65]. Traffic or transportation modeling is concerned with the analysis of all individual parts that constitute complex traffic systems, with the goal to create models that describe real world traffic phenomena. Thereby transportation modeling forms an essential foundation for creating simulations of real world traffic systems, which are needed for the analysis of traffic problems and have an important role in supporting traffic planning and decision-making processes regarding various aspects of transportation systems. It is a broad field of research, making it only possible to scratch the surface by describing the basic concepts in this chapter. The aim of this section is to give a brief overview of the different parts of traffic modeling, to explain the context this thesis is embedded in and to point out the specific areas within the research domain where contributions are made in this work. For a more detailed view on the individual parts of traffic modeling, the interested reader is referred to the work of Ortúzar and Willumsen [74].

Predicting the intensity and location of traffic as it is likely to occur in reality is a multidimensional problem, involving many individual parts that eventually influence and constitute the state of traffic in road networks. The classic model used since the 1960s to describe traffic systems is the four-step model. This model divides the formation of traffic into four sequential parts: trip generation, trip distribution, modal split and trip assignment [74]. The model is shown in figure 2.1. For a holistic traffic model to be useful for further analysis and planning, it thus has to consider all those sub-problems, finding answers to questions of who is going to travel, the source and destination of a trip, the frequency those trips occur and the route that will be taken to travel from point A to point B. Every

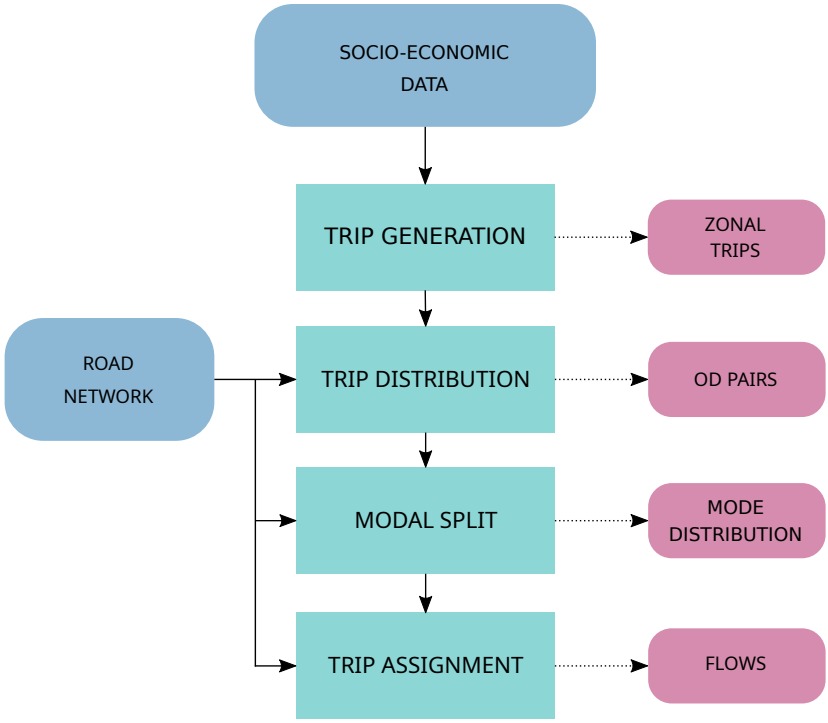

Figure 2.1: The classic four step model of transportation.

step of the four-step model forms its own research domain with an abundant body of literature including modeling approaches for many different scenarios. Due to this complexity, models in research and practice often focus on single steps of the sequence that are of interest rather than the whole system [74]. One of the reasons that makes accurate traffic modeling a difficult task is that it involves human decision-making on multiple levels. Be it the decision of going to the supermarket today or tomorrow, traveling by bus or by car, or either taking the road through the city or the highway, human behavior and decision-making is a major factor to be considered in traffic modeling. How human behavior can be modeled and predicted has been a key question in different scientific domains for many decades, ranging from economics to behavioral psychology (the next section offers a more detailed review of choice modeling). In the following section, the sub-models that together form the four-step model are explained in more detail, based on the explanations given in [74]:

Trip Generation

The first stage of the four-step model is concerned with the factors that lead to the generation or attraction of trips in certain areas. It tries to answer the question where trips are likely to occur and what the purpose of those trips is. This is usually predicted at a zonal level, where the area of interest is divided into several different zones or at household level [65]. A trip generation model gives a statement about the number of trips that are expected to start or end in a certain zone or household, which is influenced by various factors. A variety of parameters define the trips that are undertaken each day, ranging from spatial to socio-economic factors like income and family structure to the attractiveness of potential trip destinations. Those parameters can be approximated statistically from available data [74].

Trip Distribution

While the trip generation step predicts the number of trips that either depart or arrive in a certain area, the trip distribution step generates real trips as pairs of origin and destination zones. Origin and destination can either be different zones, with a trip produced in one zone and attracted to another, or a trip can take place within one zone. The sum of all trip origins and destinations in a zone are corresponding to or approximating the values generated in the previous step, depending on the quality and completeness of the data available [74]. For the generation of origin-destination (OD) pairs, the travel costs have to be considered. Travel costs are a measure for the attractiveness of a trip and can be defined as needed. Often, travel costs represent the free flow travel time [65], but can also be more complex, combining several attributes that contribute to the attractiveness (or the lack thereof) of a certain trip, called the *generalized cost of travel* [74].

Modal Split

The third step of the four-step model is concerned with the mode of transport chosen for a trip. Modeling mode choice is especially important for planning purposes of public transport, as the necessary capacities that affect bus or subway frequencies in the schedule can be anticipated from the demand derived from the model. Analyzing the individual factors that are part in the decision-making process regarding the transportation mode can also help in defining incentives to make more people change to public transport, making mode choice an important factor for reducing car traffic and congestion in urban areas [74].

Trip Assignment

In the final step, the generated origin-destination data is transformed to actual

routes in the network. This is a crucial task, as it is used to predict the demand on specific roads or links in the network, making it an important part of the transportation and infrastructure planning process [74]. Many different considerations fall into the domain of trip assignment. It is a highly diversified field of research with branches specialized in the modeling of specific parts of this domain. One major aspect within trip assignment is route choice modeling. This field of research is concerned with human decision-making in traffic, predicting the routes and single links in the network that are likely to be used. There are many different approaches to model route choice, ranging from deterministic models assuming perfect rational individuals to stochastic models often derived from economics, game theory and psychology. The next section gives a more in-depth overview of current state of the art approaches to route choice modeling. The topic of this thesis is also located within this research domain. Another vital field of research within trip assignment is traffic flow modeling. Flow models are concerned with the description of traffic dynamics, how it flows through road networks and how congestion originates an dissolves and is propagated through the network, taking into consideration the characteristics of the network such as capacities as well as the current demand. Traffic flow can be modeled on several levels, from a general macroscopic view on traffic to fine grained microscopic models. The terminology used in this work follows the definitions of model levels as used in the domain of traffic modeling and simulation. Macroscopic models focus on describing the characteristics of flows and their dynamics, often adopting physical models of flows and wave propagation in liquids [22]. Hence the movement of individual vehicles is not the focus of macroscopic models, but a higher-order description of flow characteristics in terms of vehicle densities and demand [52]. An example of a macroscopic modeling approach can be found in [31]. Microscopic models on the other side focus on individual vehicles, how they move in traffic and interact with others. This includes models for driving behavior such as acceleration and lane changing models as well as modeling human decision-making with detailed route choice models [22, 52]. A third approach are mesoscopic models, which are a hybrid form of macroscopic and microscopic modeling approaches. These models allow a more accurate description of certain properties on the level of individual vehicles - such as route choice - while still describing traffic at a macroscopic level in terms of network capacities and average demand and vehicle density [22, 52]. The methodology of this thesis can be categorized as a mesoscopic approach, as it models route choice at an agent or vehicle level, while traffic is modeled at a macroscopic level.

2.2 Human Behavior and Decision-Making in Traffic Scenarios

2.2.1 Choice Modeling: Theoretical Foundations of Human Behavior and Decision-Making

As already stated in the previous section, predicting human behavior is a crucial part of any reliable traffic model, as human decision-making is involved in every stage of the four-step model, from trip generation to mode and route choice. To predict behavior, models rely on socio-economic data - either from available surveys or from data collected for the study. Transportation planning projects therefore often conduct their own time and cost intensive surveys to acquire data of preferences and behavior of inhabitants living in the areas of interest [74]. But the presence of data alone is not enough to make a good model of future behavior. Beyond a simple replication of the current traffic situation, the difficulty lies in giving valid predictions of behavior in the future under changing conditions [32]. In other words, a model of human behavior has to be able to predict, for example, changes in mode choice when a new bus line is provided or changes in route choice when speed limits on certain roads are adapted. Since the choice sets underlying decision-processes in the context of transportation are usually discrete in nature - such as the decision to travel either by car or bus or the decision to turn left or right at a junction - this section focuses on discrete choice modeling. Most of the literature on choice theory and modeling originates from research in economics and psychology, with interest for the development of scientific models that accurately reflect human behavior strongly increasing in the second half of the 20th century. The growing awareness of the importance to describe and predict human behavior more accurately to generate more precise models is represented in several advances in this field, such as the development of *disaggregate models*, where decision-processes are modeled on the individual level rather than making assumptions for aggregated populations [109]. While economics and psychology are the scientific domains leading research on choice theory, applications in transportation have often been important considerations in the development of discrete choice models [64]. Models of individual behavior in traffic scenarios were primarily applied to describe decision-processes underlying mode choice [12]. Another important issue where discrete choice models are used is the modeling of route choice behavior in traffic networks, which is also the main concern of this thesis. Beside the formulation of specific models of behavior, advances in computation in the last decades provided the possibility of model estimation by simulation, allowing for choice models that incorporate a variety of parameters without the limitations that arise due to the need of simplification [110].

Traditional Economic Models of Behavior

Traditional models of human behavior in economics and social sciences assume that individuals act purely rational to maximize their *utility*. Utility can be defined as a measure of attractiveness that makes an individual prefer one option over another when making a decision [12]. As such, it can be as simple as monetary gains or a more complex function, incorporating several different factors important to the decision maker. In the case of transportation, the utility of a commuter deciding to either use route A or B could for instance be composed from the individual factors travel time, travel cost (tolls) and scenery. Rationality of behavior for utility maximization is based on the assumptions that first, the decision maker has perfect knowledge of the environment and current conditions inside the environment and second, has the computational or cognitive power to correctly assess all available alternatives and identify the optimal action, that is the action that maximizes utility [62, 99]. Additionally, if the outcomes of actions are not deterministic but stochastic, the decision maker knows the probabilities of all possible outcomes and thus is able to choose an action that maximizes the *expected utility* as stated by *expected utility theory* [114]. This rational theory of human behavior has been challenged by scientists for more than half a century already, calling for theories that better reflect complex human decision-making processes [99]. In the following decades, several theories were proposed that led to a different examination of human behavior.

Random Utility Theory

One important foundation for choice modeling is *random utility theory*, a theory first proposed in the field of psychology by Thurstone in 1927 [107]. It was later developed further as reaction to the problem that empirical research in psychology on human behavior did not reflect the results expected from the theory of rational behavior [62]. Subsequently, two main theoretical approaches to random utility theory emerged: constant utility models and random utility models [62]. Both are probabilistic choice models - i.e. the models do not offer a fixed deterministic description of decision-making, allowing the same decision process to result in different choices - but are built on diverging theoretic assumptions of the cause of the observed probabilistic characteristics of behavior. Constant utility models assume a probabilistic decision rule, meaning that the utility of each alternative is constant and that the decision is based on a function that assigns a choice probability to each alternative depending on the utilities [62]. Random utility models on the other hand follow the assumption of a rational decision maker who is consistent in his or her choice selection and follows a deterministic decision rule. The randomness is due to the incapability of the observer describing the decision-

process to accurately identify all factors relevant to the decision maker, such as all variables that constitute the decision maker's utility function, or all alternatives in the choice set. This inaccuracy in the model possibly leads to false predictions, thus making the use of randomized utility functions necessary to account for the errors in the model resulting from unobserved or falsely described variables [12, 63]. In other words, the constant utility approach describes behavior as probabilistic in nature, while the random utility approach states it only appears probabilistic due to the inability do describe it with adequate precision [12]. Random utility models have been very successful in the domain of transportation science - for an explanation of the most commonly used models, see [12, 74]. An early application to travel demand modeling can be found in the work of Domencich and McFadden [32].

Bounded Rationality and Prospect Theory

While random utility theory provides solutions to explain inconsistencies between the theoretical assumption of the rationality of behavior and empirical results in experiments, it still assumes absolute knowledge of the decision maker, as she or he has to accurately assess all alternatives in the decision-process and all possible outcomes to be able to maximize utility as hypothesized by the traditional model. As an alternative to the idea of rational utility maximization, the concept of *bounded rationality* was introduced by Herbert Simon, an early critic of the traditional model [99]. Bounded rationality is based on the assumptions that behavior can only be rational within the boundaries of the decision maker's cognitive capabilities and knowledge and that - rather than trying to find the optimal choice - a decision maker tries to find an alternative that leads to a *satifactory* outcome [99]. While often interpreted as a rational optimization task with constraints defined by limited cognitive capabilities, it is argued that this is a simplified view and that bounded rationality can not be interpreted in terms of utility maximization [93, 94]. Observed shortcomings of utility maximization theories shifted the perspective from prediction of behavior based on normative theories to descriptive behavior models based on empirical psychological research [1]. One influential theory that provides a different perspective on rules of human decision-making under risk is the work on *prospect theory* by Kahneman and Tversky [50]. A revised version called *cumulative prospect theory* was published several years later [113]. Based on psychological choice experiments, prospect theory explains observed decision-making under risk that deviates from expected rational behavior of utility maximization, such as the intransitivity of choices. It states that decision makers are generally *loss averse*, meaning that losses or negative outcomes are weighted more than gains or positive outcomes, that they are *overweighting low probabilities*, making unlikely outcomes to be treated with more importance than justified by their probability, and that

the value of gains or losses is defined from a *reference point* of the decision maker rather than in absolute values, making for instance the same monetary gain less valuable to a person that already possesses the same amount of money than to someone who starts with nothing. Prospect theory had a big impact on how human behavior is modeled. It has been used in various domains, although its application raises difficulties due to open questions regarding formal definitions [3]. Prospect theory has also already been applied to decision-making processes in transportation, especially as an alternative model for route choice (see [60] for an overview). The usefulness of prospect theory as theoretical framework for choice modeling in traffic scenarios was also reviewed critically in the literature [108].

As shown in this section, human behavior and decision-making is still not fully understood and cause of ongoing scientific discussion. The application of theories of choice often proves difficult due to the lack of understanding of the underlying mechanisms that drive human decision-making. This is especially true for the field of transportation, where human behavior plays a role in various different contexts. While progress has been made in the last decades, no single theory is a sufficient general description of behavior. Cherchi [26] reviews utility theory as well as prospect theory and their applications in the domain of transportation, focusing on research in mode choice behavior. The author argues for a combination of theories to provide more adequate models. De Palma et al. [76] reach a similar conclusion, acknowledging prospect theory for being useful to describe several aspects of human behavior in the context of transportation. They moreover make the important distinction between decision under *risk* and decision under *uncertainty*. Decision under risk describes settings of decision-making where the choice of an alternative can lead to different outcomes with certain probabilities of occurrence which are known to the decision maker. Prospect theory is concerned with behavior in such scenarios. The concept of decision under uncertainty also describes decision-making processes where a certain choice can lead to different outcomes, but in this case, the underlying probabilities of occurrence of those outcomes are not known. This distinction is of big importance in traffic scenarios, where uncertainty about the outcome of decisions is prevalent. Furthermore, it raises the question of learning processes and how they affect choice behavior. Theories of behavior are mainly static, not regarding changing behavior as a result of learning processes. Behavior of decision-makers is not assumed to change according to theories of choice, thus choices that diverge from certain theoretical characteristics are often regarded as 'mistakes' not worth to be considered in choice modeling [89]. [36] show that behavior that adapts to feedback indeed does not have to shift towards maximization. They identify the *payoff variability effect* as important factor to describe behavioral adaptation in learning scenarios. This effect states that choice behavior moves

towards random choice with increasing variability of payoffs. Cherchi [26] discusses implications of learning and habit for decision-making in transportation scenarios. The issue of learning in the specific case of route choice behavior is discussed in the next section.

2.2.2 Modeling Route Choice Behavior in Traffic Networks

The theoretical concepts of human choice are applied in various contexts of transportation modeling, such as mode choice or route choice. This thesis focuses on the latter, hence route choice modeling will be described in more detail in this section. A route choice model consists of two main parts: the generation of a *choice set*, that is the alternatives the decision maker can choose from, and the *choice model*, that is the theoretical approach to choice behavior in decision-making processes as described in the previous section.

Choice Set Generation

Although the selection of alternatives in the decision process seems like an easy task, the generation of a choice set is not trivial, as it has to reflect a set of alternatives that form the basis of real human behavior. Those choice sets are formed mainly intuitively by humans with the application of specific heuristics in the brain that are often not known, hence the selection criteria don't follow rules from simple rational thinking [81]. Reasonable choice set generation thus is crucial, as the best fitting choice model will not lead to good results if the alternatives at hand do not resemble the choice set of the decision maker who is modeled. Another important consideration is computational feasibility. Because route choice models are used in traffic simulations, the computational effort of route choice is important for the performance of the simulation. Although generating a choice set with all possible alternative routes that exist in a network might avoid the risk of not considering an important alternative, the computational effort makes it unusable in practice. Hence the size of the choice set is restricted, making it important to select realistic alternatives and omit redundant and unrealistic route options [8]. To achieve this, several approaches to choice set generation were proposed. A thorough review of several algorithms can be found in [81, 83].

The simplest option is the restriction of the choice set to a single alternative, the path exhibiting the lowest overall cost, basically reducing route choice to a single step without the need for complex solutions for choice set generation an choice modeling. This approach is in line with the idea of rational behavior of classical economic decision theory, which assumes a decision-maker with perfect knowledge of the network to choose the optimal alternative. In most cases, the path with the shortest travel time under free flow (that is without delays caused by congestion

effects) is regarded as the optimal path. Modeling traffic flows assigning flow only to shortest paths without regarding the influence of congestion is also called *all-or-nothing assignment* [74]. While being a fast and simple solution, this approach has various downsides. Defining the fastest route from origin to destination as the optimal path for everyone is not realistic, as humans have different preferences that also play a role in route choice. As shown by Bekhor et al. [8], paths that minimize travel time only represented about a third of all commuter paths the researchers collected from a survey where participants were asked to draw their daily commuting path. It is an accepted view in research regarding behavior that human preferences and incentives are much more complex and multidimensional. Of course, this approach can also be applied defining utility functions with more parameters than travel time only, such as preferences of road types. Nevertheless, it relies on the concept of rationality, assuming perfect knowledge and unbounded cognitive capabilities of the decision maker to find the best solution. This is especially true in scenarios where congestion effects on single links apply. To find the optimal path in such scenarios, the driver has to know the current demand on every link in the network to compute the optimal route. A widely-used approach to generate choice sets that better reflect realistic decision-making processes is the application of *k-shortest paths* algorithms. Those algorithms search for the lowest cost path from an origin to a destination in a network and sequentially repeat the search process to find the next shortest paths. Although, as already described, the cost function that defines the attractiveness of a path is dependent on the decision maker and reflects his or her personal utility function and thus doesn't have to depend on distance or travel time only, the term *shortest path* will be used in this context to describe the path with the lowest overall cost from a certain origin to a destination in the network. Although the use of k-shortest paths provides a simple solution to the problem of choice set generation that can easily be adapted to different set sizes, this approach has several drawbacks. The alternatives generated by such algorithms are often very similar and differ only in single links, additionally they might include paths with cycles that do not represent realistic alternatives [81]. Using loopless variants of a k-shortest path algorithm such as the algorithm proposed by Yen [119] prevents the generation of paths with unrealistic cycles. The definition of the cost function that defines a shortest path is a very important factor as well. Beside obvious parameters such as travel time or distance, drivers might value other criteria in their decision-making such as scenery or show a preference for certain types of roads, such as highways. Reducing such multidimensional utility functions to single value representations of preferred paths such as travel time might lead to choice sets that miss alternatives important to the decision maker. The decision maker's utility function thus does not only play a role in the decision-making process itself, but has to be considered in the generation of a choice

set as well. The problem of incorporating travel time in the cost function is also relevant for k-shortest path approaches. Using free flow travel time does not account for potential congestion effects on links that show high traffic intensity and might lead to choice sets with unrealistic routes. This is especially true if bottleneck links are present in the network, links that are constantly showing high demand due to their importance in the network. In cases where the cost of using such bottlenecks is small under free flow, they might be included in a disproportionate number of alternative paths in the choice set generated by k-shortest path algorithms, although the efficiency of such links rapidly declines with rising demand. Figure 2.2 shows a simplified example of this potential problem. As shown in the figure, from all eleven possible paths with origin node 1 and destination node 13 in the network, nine paths include the bottleneck link from node 1 to node 4. Assuming high demand and limited capacity, the efficiency of link $(1, 4)$ would rapidly decrease with increasing demand. Individual drivers could use a faster route either with link $(1, 2)$ or $(1, 3)$, however those would only be included in the set of k-shortest paths if $k > 9$. As an alternative, cost functions could be computed using travel times based on current demands on individual links. This however assumes overall knowledge of the current state of traffic in the network and the current travel times, an assumption that is not realistic for every driver even in the age of ubiquitous access to navigation systems provided by smartphones. Furthermore, the choice set would have to be updated in every episode.

In addition to using k-shortest paths algorithms to compute shortest paths based on their costs only, heuristics are often applied to achieve a more differentiated set of alternative paths, mainly *link elimination* and *link penalty* [81, 83]. The link elimination approach is based on sequential shortest paths searches. After ever iteration, a number of links that are part of a shortest path are eliminated and cannot be considered in following iterations. This leads to the generation of a set of more differentiated paths. The question which links and how many should be eliminated is not trivial, as some links might be crucial connections for many different paths. In the worst case, this could lead to disconnected networks where no further paths for a certain origin-destination pair exist. Furthermore, the choice of the number of eliminated links has a big impact on the resulting set of alternatives, as single link eliminations again lead to similar routes and the elimination of too many links creates unrealistic alternatives [83]. Link penalty approaches on the other hand don't exclude certain links from the network, but make them a less attractive choice. After every iteration of the shortest path algorithm, the cost of some or all links that are part of the current shortest path are increased. This way, generation of choice sets with highly similar routes is discouraged while some downsides of the link elimination approach such as disconnected networks are avoided [83]. The definition of the cost increase per iteration as well as the choice

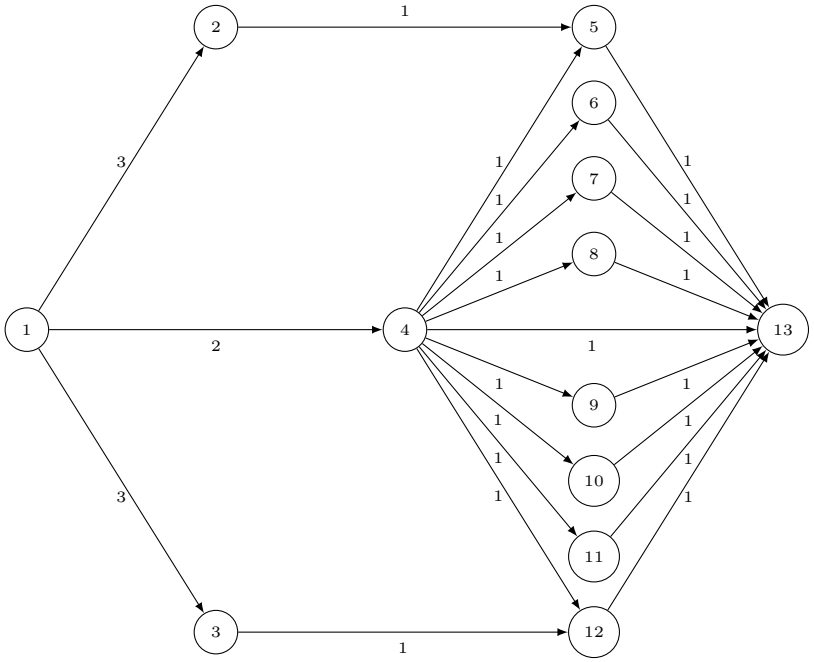

Figure 2.2: An example of a network where a bottleneck link would lead to high costs in a set of k-shortest paths when only considering free flow travel times.

of routes that are penalized still needs thorough consideration and can lead to unintended results [81]. A third approach is *labeling*, proposed in the work of Ben-Akiva et al. [11]. This approach uses a set of different labels to acknowledge the fact that drivers might have a variety of preferences when choosing a route. Route choice behavior is then modeled using a Nested Logit model [83]. So called *simulation approaches* where possible alternatives are generated based on utility functions with parameters sampled from a probability distribution are discussed in [81, 83].

Bekhor et al. [8] provide a comparison of all approaches mentioned above. When validating the generated choice sets with route choice data acquired from a survey, they show that the use of k-shortest paths for choice set generation in combination with link penalty or link elimination only achieves to include an accurate representation of 60% of all real paths from the survey. Allowing small deviations between the generated paths and the paths from the survey data by defining an overlap of 80% as sufficiently accurate, the choice set generated from k-shortest paths covers 80% of all paths. Choosing the path with the least travel time for all origin-destination pairs accounted for only 34% of all routes from the survey. A combination of several algorithms showed the best results, but the

computational effort needed makes this approach unfeasible for many practical applications. This shows the difficulty of generating realistic choice alternatives and the importance of a better understanding of human decision-making and the heuristics applied in the process of intuitive or rational thinking about alternatives in a decision-making process.

Route Choice Modeling

After generating a choice set, a discrete choice model has to be applied to predict route choice behavior. The same theoretical considerations of decision-making and rationality of choice apply as described in the previous section. Widely used route choice models are described in [8, 81, 83]. Behavioral theories of bounded rationality and prospect theory also play a huge role in route choice modeling, leading to active research and discussion of the usefulness of such theories for transportation science. Li and Hensher [60] provide an overview of current literature where prospect theory is applied for modeling route choice behavior and describe the advantages and disadvantages of this behavioral theory for choice modeling in the domain of traffic. They also show that although several works exists that use prospect theory, the majority applies only parts of it - such as risk aversion - without considering the whole theoretical framework. Van de Kaa [47] provides a summary of the crucial concepts of prospect theory and its value for transportation, stating that prospect theory is able to better explain various forms of decision-making in traffic situations than classical utility theory. He also proposes an adaptation of prospect theory for modeling human behavior in traffic scenarios, *extended prospect theory*, incorporating aspects of both theoretical backgrounds (for a more detailed description, see [46]). A more critical review of prospect theory for traffic modeling can be found in the paper of Timmermans [108], arguing that experimental results of prospect theory based on monetary gains and losses cannot be translated easily to the domain of decision-making in traffic scenarios. Other authors also see difficulties in the definition of gains and losses as well as connected reference points (see [60] for a summary). Another issue is the distinction between decision under risk and decision under uncertainty [76] and its implication for behavior modeling in traffic. Several works recognize this potential problem. Using prospect theory to model route choice behavior with known probabilities of variable travel times, Gao et al. [37] for instance acknowledge the fact that uncertainty of probabilities of exact link costs due to variability in travel time could be a more realistic assumption regarding decision-making in traffic.

Learning, Habit and Adaptation in Route Choice Behavior

A crucial problem that is often not addressed in choice modeling is learning and adaption (see also last section). As choice models are generally static, they do not provide adequate answers to behavioral changes due to learning processes. This is true for classical economic theories of utility maximization as well as prospect theory. The concepts of bounded rationality provides a way to describe limited knowledge, but how exactly this affects choice behavior such as exploration of unknown alternatives is another question. This is closely related to the consideration that different decision theories provide adequate approximations of different stages in the learning process of the decision-maker. As argued in [76], concepts from prospect theory reflect behavior in unknown or unfamiliar decision-making scenarios, whereas with more experience, behavior tends to switch towards (expected) utility maximization. The assumption that behavior shifts towards rational decision-making with growing experience is also challenged in the literature. In [68], the authors observe that learning does not lead to more rational and homogeneous route choice behavior in traffic networks. The question how information and experience affects route choice behavior is addressed in the work of Ben-Elia and Shiftan [13, 14], showing that the payoff variability effect is important in route choice adaptation with experience. A learning model for decision-making in the context of route choice is described in [17], identifying *reinforcement learning* as a crucial concept to model learning and incorporate experience in the decision-making process and proposing a Markov model for iterative decision making in day-to-day route choice scenarios. Avinery and Prashker [1] also discuss learning models for route choice behavior, reviewing traditional utility maximization as well as cumulative prospect theory and discussing important concepts of learning such as the payoff variability effect and reinforcement learning.

2.2.3 Rational Choice and Traffic Equilibria in Multi-Agent Scenarios

Although behavioral approaches to choice modeling in transportation focus on describing behavior on the level of the individual, it should not be forgot that decisions in traffic scenarios are always driven by interactions of many decision-makers whose choices have direct impact on others, influencing their decisions as well. Traffic flows in road networks are thus the result of complex interaction patterns. The study of such patterns and the resulting traffic states is an important line of research in transportation modeling, as accurate models of such traffic systems are crucial tools for planning and policy making. Modeling traffic flows as a result of individual behavior is an *agent-based* modeling (ABM) approach. This computational concept describes approaches where individuals are modeled

as artificial (potentially intelligent) autonomous *agents* that interact with each other and their environment, making decisions based on their own perceptions and often showing individual behavior and characteristics [18]. Systems involving multiple individual agents are also called *multi-agent systems* (MAS) [96]. Multi-agent systems have been applied as models for many different domains, including transportation [16, 33]. Such models often have their theoretical roots in game theory. This branch of economics is concerned with the analysis of interactions between individual decision-makers and their cooperative or competitive behavior to maximize their profit. Such scenarios are modeled as *games* with rational *players*, each choosing a *strategy* to maximize his or her *payoff* [75]. The term *player* to describe individual decision-makers is often used in game theory and will be used interchangeably with the term *agent* in this work. Under the assumption of rational behavior, game theory is applicable to many scenarios where the understanding of interactions of individuals with often diverging goals and needs is crucial. It thus has also been used as a tool to describe dynamics in networks and to study the effects of rational decision-making in various traffic scenarios. Such scenarios are part of a special class of games where the price of a resource is dependent on the number of players using it, which are subsumed under the name *congestion games* [84, 105]. Games more specifically concerned with flows in networks are also called *routing games* [87]. Multi-agent scenarios can be very complex, since every single agent or player has to adapt his or her personal strategy to the strategies of the other players to maximize payoff. Such a strategy is called *best response* strategy, which assumes the agent knows about the strategies of all other players. A central interest of game theory is the study of equilibria in such games where no player can further maximize personal payoff and therefore sticks to the current strategy [59]. The concept of such equilibria in non-cooperative games was described in [69], coining the term *Nash equilibrium*. Such equilibria emerging from rational self-centered behavior of each agent are shown to be inefficient, resulting in less payoff when compared to coordinated behavior between the agents [105]. Equilibria and their resulting inefficiencies are a central field of research in *routing games* [87] and are of great interest for traffic engineering as well, given the potential impact of such inefficient behavior on traffic networks. The discrepancy between individual and societal interest in the context of traffic behavior was already described in 1920 in [79], discovering that each agent's rational behavior to minimize travel time can lead to a lower overall network performance. This was later described in more detail by Wardrop in 1952 [115] who showed that this selfish behavior leads to a state of equilibrium within the network where no agent has incentives to change routes, the *user equilibrium* - an idea analogous to the Nash equilibrium. This state of egoistic minimization of every agent's own travel time leads to an overall network flow latency (which corresponds to the average travel time in the

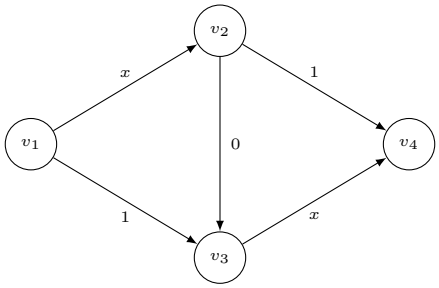

Figure 2.3: The Braess network. Assuming a linear cost function $x = \frac{d_e}{100}$ depending on the current link demand d_e, the Braess paradox occurs when inserting a low cost edge between v_2 and v_3 - every selfish acting agent will reroute it's traffic over $v_1 - v_2 - v_3 - v_4$, which results in a user equilibrium with higher average travel times than without the low cost edge.

network) that is larger than the optimal possible system performance, called the *system optimum*, which Wardrop describes as the network state with minimum average journey times [29, 115]. The term *selfish routing* to describe purely rational behavior in route choice was introduced in the work of Roughgarden [88]. Much of the literature on selfish routing is dealing with the quantification of the inefficiency resulting from selfish routing, a concept first introduced in [51] and later called the *price of anarchy* [77]. Upper bounds for the price of anarchy for different link latency functions were first proven in [85, 86]. The interest in understanding and quantifying inefficiencies associated with traffic equilibria stems from the need to design more efficient road networks and develop better route assignment strategies. The fact that network design plays a crucial role in facilitating efficient network routing was shown by the discovery of the *Braess's paradox*, which describes that in networks with high travel costs, new low-cost edges inserted in the network can lead to an increase in overall network costs, that is to a decrease in network performance due to selfish behavior [21]. The Braess network is shown in figure 2.3 This led to increasing interest in research on inefficiencies in traffic networks caused by selfish behavior [88].

Selfish routing offers a rich theoretical foundation and has been crucial for the investigation of traffic equilibria. It has been applied to approximate network flows in experimental scenarios as well. However, for the purpose of creating a realistic model of human behavior in real world traffic scenarios, the assumptions of rationality and homogeneous behavior as used in game theory do not offer an appropriate and realistic view on individual behavior. The simplistic view of utility maximization as behavioral strategy has been falsified in experiments and raised criticism from many scientific researchers, as already described in the last sections.

Assuming a best response strategy further implies knowledge of the decision-maker about the actions all other players are going to take. In scenarios where the same decision-makers repeatedly play the same game, it could be argued that this knowledge can be acquired through experience by repeatedly playing the same game. In traffic scenarios, this assumption is unlikely, as drivers can never know the decisions of all others in the network. Relying on inference to predict the outcomes of a decision based on past experience will therefore not necessary result in utility maximizing behavior. Beside rationality and utility maximization, the concept of homogeneous behavior is also an unlikely assumption for systems such as traffic networks. In real world scenarios, individuals do not share the same expectations and values. While arriving home after a full day of work can arguably be seen as the main goal everybody shares, some people might find it more desirable to take a route that stays off the highway, even if it means sacrificing a few minutes. Some might value their habits and always take the same route home, while others are more willing to change routes based on the current traffic conditions. Such considerations are part of research in behavioral game theory [24]. The problem of imperfect knowledge in human decision processes and the modeling thereof has also been addressed in game theory for several decades already, leading to different concepts and models of such a *bounded rationality* [89]. A route choice model strictly based on selfish routing thus does not provide an adequate representation of human behavior. This is especially true in dynamic network scenarios with often changing traffic conditions, where the exact travel times can not be known beforehand since they are highly dependent on the decision of all other agents in the network. A lot of work has been done to address this problem, such as studying flow equilibria in networks with selfish routing where travel times are only approximately known [86], called ε - *approximate Nash equilibria* [78] or the incorporation of behavioral theories in game theory to predict human behavior [117], also with the help of artificial intelligence [41]. Using reinforcement learning to let agents learn and adapt to traffic conditions can provide a different perspective on the problem of traffic modeling in networks where link demands are dynamic and agents are constantly adapting and changing their strategies.

2.3 Reinforcement Learning

Reinforcement learning (RL) is one of the major branches in current artificial intelligence (AI) research in addition to the two other fundamental machine learning approaches *supervised learning* and *unsupervised learning*. While initial ideas already date back to the 1950s [90, 103], research on reinforcement learning currently receives new interest across different scientific domains. This section gives a brief introduction to the fundamental concepts of reinforcement learning and provides

an overview of the crucial aspects underlying the learning process of individual agents. There exist a variety of books that offer a deeper understanding of various concepts and methods of reinforcement learning. The book by Sutton and Barto [103] is the standard introductory book and also forms the foundation of much of the explanations and equations given in this section. The mathematical notation used here mostly follows this book. [104] offers another more compact summary, focusing on the mathematical aspects of different reinforcement learning algorithms. The key concept behind reinforcement learning can be summarized as follows: learning how to achieve a goal without initial knowledge or example how to achieve it, doing so merely through interaction with the environment and feedback from the environment. While trying different actions, desirable outcomes are rewarded, making the agent favor those actions. Research shows that his goal-directed learning process based on trial and error and reinforcement signals are fundamental mechanisms in infant and animal learning [90, 103]. Reinforcement learning can be distinguished from other concepts of machine learning in several ways. Unlike supervised and unsupervised learning, reinforcement learning does not need big amounts of preexisting data. While supervised learning relies on training sets of labeled data and unsupervised learning is used to find patterns occurring in unlabeled data, a reinforcement learning agent collects its own data through direct perception of the environment and through interaction with it. In addition, reinforcement learning does not depend on preexisting knowledge. The core of supervised learning methods is learning to choose the right action from a set of examples, a training set. For this purpose, supervised learning agents need training examples of specific situations and information about what to do in this specific scenario. Given a sufficient amount of data, the agent can then find the right action in previously unknown situations by generalizing knowledge gained from a training set. Reinforcement learning on the other hand works by rewarding desired outcomes while letting the agent find a way to maximize the reward [103]. Of course, there are gradations in initial knowledge of reinforcement learning agents as well, mainly distinguishing *model-free* and *model-based* approaches. In the model-free case, the dynamics of the environment are unknown to the agent, whereas in model-based reinforcement learning, the agent is given a priori knowledge about the environment. Reinforcement learning problems can be split into two scenarios, where either a single agent learns in an environment or where multiple agents are simultaneously learning in a shared environment, called *multi-agent reinforcement learning* (MARL). While multi-agent scenarios are based on the principles of single agent reinforcement learning scenarios, some theoretical assumptions of the single agent case no longer hold in multi-agent scenarios. The following sections provide an introduction to both scenarios.

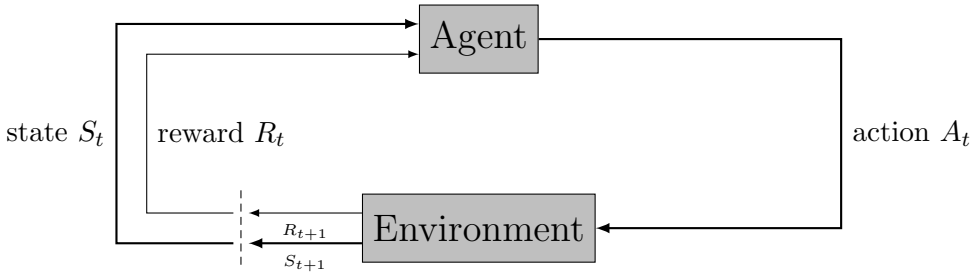

Figure 2.4: Agent-environment interaction, a key concept of reinforcement learning. Figure adapted from [103].

2.3.1 Single Agent Reinforcement Learning

The fundamental process in reinforcement learning is the interaction between agent and environment. The agent perceives the environment and chooses an action based on those perceptions. This decision-making process is what constitutes *behavior* [90, 104]. The agent perceives the environment either through different sensors or through other signals from the environment. How and what aspects of the environment it perceives depends on an agent's function and goal and is an important consideration when designing an agent. The agent thus is finding himself in a *state* $S_t \in \mathcal{S}$ at current time t. Based on this perceived environmental state S_t, the agent takes an *action* A_t. The environment changes corresponding to the action, transitioning to a new state S_{t+1} and returning the *reward* R_{t+1} following the last action to the agent. This fundamental agent-environment interaction is shown in Figure 2.4. This decision-making process, consisting of a sequence of states, actions, rewards and new states is formally described as a *Markov decision process* (MDP), which forms the foundation of almost every reinforcement learning problem [103]. MDPs were first introduced by [10] and are used to model stochastic processes where an action in a certain state is followed by a reward signal and a new state. It is defined as the tuple $(\mathcal{S}, \mathcal{A}, p, r)$, where $\mathcal{S}$ is the set of all states in the environment and $\mathcal{A}$ is the set of all actions available to an agent. $\mathcal{A}^s \subset \mathcal{A}$ denotes the subset of all possible actions in a given state $s \in \mathcal{S}$. In many reinforcement learning problems, $\mathcal{A}^s = \mathcal{A}$, therefore the simple notation $\mathcal{A}$ will be used throughout this introduction to denote the set of possible actions in a state. This work focuses on MDPs where the sets $\mathcal{S}$ and $\mathcal{A}$ are finite, thus called *finite Markov decision processes*. $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the *transition probability* function, which assigns a probability P to every possible transition from a state-action pair (s, a) into a subsequent state $s' \in \mathcal{S}$. The state transition probability is defined as

$$p(s' \mid s, a) = P\{S_{t+1} = s' \mid S_t = s, A_t = a\}. \tag{2.1}$$

For every action, the agent receives a *reward* based on the *reward function* $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. This function computes the reward that can be expected from taking action a in state s and is given by

$$r(s, a) = \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]. \tag{2.2}$$

R_{t+1} is the immediate reward returned to the agent from the environment after taking an action. Since the reward is returned after the action is taken, concluding a single time step of an episode, the notation $t + 1$ is used for rewards returned for actions taken at time step t. Because the reward function only gives an expectation of the reward received following a certain action and makes no predictions about future rewards, it is also called *immediate reward function* [104]. Since a MDP is a sequential decision-making process, every state has to include enough information about past actions and states so that all state transition probabilities from a state $s \in \mathcal{S}$ given any action $a \in \mathcal{A}$ to a subsequent state $s' \in \mathcal{S}$ can be determined only by the information an agent can perceive in the current state s. This necessary condition is called the *Markov property* [103]. Markov decision processes therefore are *memoryless*, since the state description already includes all necessary information to predict the future without having to explicitly store the whole history of past agent-environment interactions.

A crucial part of reinforcement learning is how an agent chooses its action and how it changes its behavior through learning which better actions to take in future situations. A reinforcement learning agent follows a stochastic policy $\pi : \mathcal{S} \times \mathcal{A} \to [0, 1]$ which assigns to every possible action in a state the probability for taking that action. π can also be deterministic, if for every state s there is only one possible action a. This overview will focus on the stochastic case. In general, an agent tries to maximize the sum of all received rewards, called the *return* denoted as G. The return G_t is computed from the current time step t onward as the sum of all rewards to be received in the future:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}. \tag{2.3}$$

γ is called the *discount factor*, defined as $\gamma \in [0, 1]$. This parameter reduces the value of future rewards, meaning immediate rewards are worth more than rewards that are expected to be received in the far future. There are various reasons for applying a discount factor. Most importantly, the use of a discount factor $\gamma < 1$ in continuous reinforcement learning scenarios where tasks don't terminate has mathematical reasons, since otherwise the total return would accumulate to infinity

[103]. Such continuous tasks include for example ongoing control tasks. This does not apply to episodic tasks, which are reinforcement learning tasks that have a defined terminal state. The scenario in this work where an agent travels from origin to destination in a network is episodic, since it terminates when the agent reaches the destination. For episodic tasks, a value of $\gamma = 1$ is often used. Another reason for using a discount factor is to account for the uncertainty of reward predictions that lie far in the future [23]. Since rewards received in future states cannot be determined with absolute certainty, the *expected* return is maximized, which is an estimation of all future rewards expected to be received when being in state s following a policy π. This estimator is called a *value function*. There are two ways of predicting the future return. One possibility is a statement about how much future reward can be expected starting in state s and following π, called the *state-value function* $v_\pi : \mathcal{S} \to \mathbb{R}$:

$$v_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s] \quad \forall s \in \mathcal{S}. \tag{2.4}$$

In other words, the state-value function evaluates the state s based on the return that can be expected from this state onward, assuming actions are chosen under policy π. It does not evaluate single actions in $\mathcal{A}$ an agent might take. Hence, another way is to estimate the return from a particular state-action pair (s, a), that is the total return that can be expected when taking an action a in state s under the policy π. This is called the *action-value function* $q_\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$:

$$q_\pi(s, a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] \quad \forall s \in \mathcal{S}, a \in \mathcal{A}. \tag{2.5}$$

The difference is that the state-value function evaluates the state, assuming that every next decision which action to take follows π, whereas the action-value function evaluates single actions, making the same assumption for any action a onward which itself does not have to follow π. Looking at the definition of the return G_t in equation 2.3, it is apparent that the state- and action-value functions consider not only immediate rewards, but also all future estimated rewards. This is a very important condition for reinforcement learning and ultimately means that a value function $v_\pi(s)$ depends also on the value function of the next state $v_\pi(s')$. This recursive property is formulated in the *Bellman equation*, which is a fundamental concept in dynamic programming [15, 103]:

$$\begin{aligned} v_\pi(s) &= \mathbb{E}[R_{t+1} + \gamma v_\pi(s') \mid S_t = s] \\ &= \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} p(s' \mid s, a) \Big[r(s, a) + \gamma v_\pi(s') \Big] \quad \forall s \in \mathcal{S}. \end{aligned} \tag{2.6}$$

This shows that the value function $v_\pi(s)$ is actually computing the average of the value functions of all possible next states $v_\pi(s')$, weighting them by their probability of occurrence, which is the combination of the probability for choosing an action a when in s following policy π and the transition probability for s' following s given that action a. When an agent interacts with the environment, visiting states, choosing actions and receiving rewards many times, it will learn what rewards to expect from certain states and actions, leading to more precise estimations of expected rewards which ultimately form the basis for an agent's behavior. A policy that maximizes the return is called an *optimal policy* π_* and is based on the optimal value function $v_*(s)$:

$$v_*(s) = \max_\pi v_\pi(s) \quad \forall s \in \mathcal{S}. \tag{2.7}$$

For a finite MDP, there exists a unique optimal state-value function [103]. The optimal action-value function $q_*(s, a)$ is defined by

$$q_*(s, a) = \max_\pi q_\pi(s, a) \quad \forall s \in \mathcal{S},\ a \in \mathcal{A}. \tag{2.8}$$

From the definition of the optimal value function, the Bellman equation for v_* can be formulated, also called the *Bellman optimality equation* [15, 103]. It describes that $v_*(s)$, which is the expected return in state s under an optimal policy π_*, is the expected return when taking an optimal action, that is an action with the highest value from the action-value function.

$$\begin{aligned} v_*(s) &= \max_{a \in \mathbb{A}} q_{\pi_*(s,a)} \\ &= \max_{a \in \mathbb{A}} \sum_{s' \in \mathbb{S}} p(s' \mid s, a) \Big[r(s, a) + \gamma v_*(s') \Big] \quad \forall s \in \mathcal{S}. \end{aligned} \tag{2.9}$$

Similarly, the Bellman optimality equation can be formulated for the action-value function:

$$q_*(s, a) = \sum_{s' \in \mathcal{S}} p(s' \mid s, a) \Big[r(s, a) + \gamma \max_{a'} q_*(s', a') \Big] \quad \forall s \in \mathcal{S}. \tag{2.10}$$

While v_* gives the optimal state values, q_* directly leads to the optimal policy. If q_* is known, an agent only has to choose those actions with the highest q-value. The optimal policy is therefore

$$\pi_*(s) = \arg\max_a q_*(s,a) \quad \forall s \in \mathcal{S}. \tag{2.11}$$

If the optimal action-value function is known, the optimal policy in this case is thus a *greedy* policy that always chooses the action that maximizes the return and neglects other possible actions. To find those best actions, agents have to be able to try various different actions from which they have no knowledge what return to expect. It is therefore necessary for agents to take risks that might lead to higher rewards. This is called *exploration* and is a vital concept in reinforcement learning, meaning that agents follow a stochastic policy that introduces randomness in the decision-making process. There are a variety of possibilities to permit exploratory behavior in reinforcement learning while still *exploiting* already known actions with high rewards. This *exploration-exploitation dilemma* is one of the crucial problems in designing reinforcement learning tasks [103]. One widely used approach to enable exploration is to make agents follow a ε - *greedy policy*. Following such a policy, the agent will choose the action with the highest estimated return - the greedy action - with a probability of $1-\varepsilon$. With a probability of ε, the agent will choose a random action. Equation 2.12 shows the probability of an action A to be chosen in a state S_t following an ε - greedy policy, given a set of feasible actions $\mathcal{A}^s$ of size n. The value of ε is usually decreased over time to enable convergence towards the optimal policy.

$$P(A \mid S_t) = \begin{cases} 1-\varepsilon+\frac{\varepsilon}{n} & \text{if } A = \arg\max_a q(S_t,a) \\ \frac{\varepsilon}{n} & \text{if otherwise} \end{cases} \tag{2.12}$$

Starting from these fundamental concepts, there are many ways to make agents learn and change their behavior over time to optimize the return. Reinforcement learning research provides various approaches based on different theoretical concepts on how the learning process can be constituted, including algorithms, policy designs, etc. with the goal to maximize performance and to provide a fast learning mechanism for agents. For a more in-depth explanation of these approaches and actual implementations, the reader is referred to [103, 104].

2.3.2 Multi-Agent Reinforcement Learning

Multi-agent systems, where several independent agents interact in a shared environment, have been an active field of research for many years already, emerging from artificial intelligence research to study possible implementations of such complex systems [102]. A fundamental part of multi-agent systems is multi-agent learning (MAL), which is concerned with learning processes in systems where multiple agents interact. It is a complex field spanning various scientific domains from

social, life and computer sciences, involving questions about learning mechanisms in living beings, human interactions and decision-making processes as well as digital representations of such processes and interactions. Much of research lies at the intersection of game theory and artificial intelligence research and has received much attention from both sides [98]. Due to the conceptual diversity of multi-agent systems, ranging from scenarios of competition to cooperative tasks, many different approaches have been studied to solve the question of how to facilitate learning in environments with multiple agents. Stone [102] defines four main classes of multi-agent systems, differentiating whether agents are homogeneous or heterogeneous in their behavior, goals, etc. and whether they communicate with each other or act independently. The application of reinforcement learning is one of the main strategies to make agents learn and develop their behavior in MAS research, commonly known as *multi-agent reinforcement learning* (MARL). Such systems offer many possibilities to model environments with complex agent-environment and agent-agent interactions where behavior is directly learned from experience rather than designed beforehand. This could potentially lead to new behavioral strategies of agents in the system which could not be anticipated in an a priori behavior design [23]. The occurrence of intrinsic innovation in multi-agent reinforcement learning systems has already been observed and is discussed as an important topic for further research [2, 55]. Though the application of single agent strategies to multi-agent systems has shown some success, theoretical understanding of the underlying processes is difficult due to the complexity of such systems, leading to many questions that are still unsolved [120]. There is a rich body of literature with approaches ranging from direct application of single agent reinforcement learning to the multi-agent case to the design of specific MARL algorithms involving game theoretic concepts. For comprehensive summaries of multi-agent reinforcement learning and algorithms see [23, 42, 120].

For multi-agent reinforcement learning, some of the fundamental concepts applying to the single agent case have to be adapted. As studied in game theory, a Markov decision process including more than one agent in an environment can be described as a *stochastic game* [96]. A stochastic game is defined as a tuple $(\mathcal{S}, \mathcal{N}, \mathcal{A}, p, r)$ where $\mathcal{S}$ is a finite set of states, $\mathcal{N}$ is a finite set of players, $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$ is the joint action set where $\mathcal{A}_i$ is the finite set of all actions available to player $i \in \mathcal{N}$. As in the single agent case, for simplicity it is assumed that the actions available to player i are the same in every state, so that $\mathcal{A}_i^s = \mathcal{A}_i$. p is the state transition function shown in equation 2.1 and r is the reward function defined in equation 2.2. Analogous to the single agent case, every agent acts rational to maximize the expected return, following a policy $\pi_i(a_i|s)$. In multi-agent settings, the expected return following a policy is not only depending on every agent's own actions, but on the actions of all other agents in the environment as well, the *joint action*

$\mathbf{a} = (a_i, a_{-i}) \in \mathcal{A}$, where $-i = \mathcal{N} \setminus \{i\}$, that is the set of all agents $\mathcal{N}$ except agent $i \in \mathcal{N}$ [42]. A *joint policy* of all agents is thus given by $\pi(\mathbf{a}|s) = \prod_{j \in \mathcal{N}} \pi_j(a_j|s)$, leading to the joint value function adapted from the Bellman equation (eq. 2.6) [42, 120]:

$$v_{\pi_i}(s) = \sum_{\mathbf{a} \in \mathcal{A}} \pi(\mathbf{a}|s) \sum_{s' \in \mathcal{S}} p(s' \mid s, \mathbf{a}) \Big[r(s, \mathbf{a}) + \gamma v_{\pi_i}(s') \Big] \quad \forall s \in \mathcal{S}. \tag{2.13}$$

In the multi-agent setting, the best action is thus dependent on the action of all other agents in the environment. This leads to a fundamental problem in multi-agent learning: In a Markov decision process, the transition probabilities (eq. 2.1) define the dynamics of the environment, as they tell the agent how the environment is likely to change or behave after a certain action from the agent. In the model-free reinforcement learning case, those dynamics are initially unknown and learned over time from experience, following the assumption that the environment is *stationary*. This means that the environment dynamics remain the same over time, i.e. the probabilities for possible state transitions to occur after a certain action is taken are static, thus following rules that can be observed and learned to later act upon those observations. If there are more agents learning independently in an environment, each following its own behavior and adapting constantly, the environment becomes *non-stationary*. The agent can no longer sufficiently learn the dynamics of the environment, because rewards and state transitions now also depend on the decisions of all other learning agents in the environment. State transitions can no longer be anticipated just from the current state alone, thus the Markov property is lost [54]. Finding an optimal policy in such a constantly changing environment becomes a moving target problem [112]. Possible approaches to solve this problem depend highly on the nature of the multi-agent scenario. [23] define three core scenarios: the *fully cooperative* case, where all agents have a joint goal and share one reward function ($r_1 = r_2 = \cdots = r_i$), trying to maximize the collective payoff. In the *fully competitive* case, the scenario resembles a *zero-sum* game: what one player wins, the other loses [96], resulting in contrary reward functions $r_1 = -r_2$. Though zero-sum games are also possible with more agents, most literature focuses on the two player case [23]. The third scenario are *mixed* tasks, where each agent is purely self-interested and acts to maximize its own reward, called *general-sum* games [96]. This involves cases where agent decisions are completely independent from each other as well as scenarios including congestion games where agents compete for resources, therefore influencing each others rewards. In all scenarios, the problem of a non-stationary environment due to multiple agents populating it can be addressed differently. A major factor regarding non-stationarity is the level of knowledge every agent has about the decisions and actions of others in the environment. This knowledge and adaptation to the behavior of others can be modeled very differently, ranging from complete disregard of the presence of other agents to sophisticated

predictions about possible actions of others [42]. In the simplest case, one agent treats all other agents as part of the environment, perceiving all fluctuations in rewards and state transitions caused by the ongoing learning process of all other agents as dynamics of the environment. Such *independent learners* [28] make use of single agent reinforcement learning algorithms like Q-learning [116] in multi-agent settings, neglecting some of the theoretical shortcomings of this approach. For algorithms used in single agent settings to converge, a stationary environment is required, which means that convergence guarantees in multi-agent settings no longer hold [28, 53, 54]. Regardless of this problem, there are many practical examples where the application of algorithms like Q-learning in MARL have shown good results (see [23, 42, 120] for an overview). Much work has also been done in developing new methods specifically for the multi-agent setting and extending existing single agent algorithms for multi-agent applications [19, 25, 44, 57, 106]. Since optimal actions in single agent Markov decision processes are translated to best response actions in multi-agent scenarios that result in Nash equilibria, many of those concepts for designing MARL algorithms treat the convergence towards such an equilibrium as an essential property. Nevertheless, this strong focus on equilibria in multi-agent learning as it is practiced in game theory has been questioned critically in the literature, especially for complex scenarios or where realistic human decision-making is the objective [53, 97, 98]. Beyond a critical view on the goal of convergence to equilibria, the usefulness of reducing interactions in complex multi-agent systems to game theoretic concepts to find solutions has been challenged altogether [101]. Besides theoretical considerations about the objective of multi-agent reinforcement learning algorithms and the definition of research categories, many other open questions still have to be answered, be it the investigation of multi-agent scenarios with more than two agents (at least in game theory) and the need for better models of human behavior [98] to the need to study interactions between diversified agents, that is agents with different learning processes and objectives [42]. In addition to develop efficient learning schemes for various applications in artificial intelligence, multi-agent reinforcement learning could also help in providing a better understanding of complex systems where exact behavior cannot be anticipated beforehand. A potential field of use is thus the investigation of complex systems by modeling individual behavior and continuous adaptation through interactions and experience. In this context, this thesis examines the potential of multi-agent reinforcement learning to model realistic decision-making and route choice in road networks that can be applied in traffic simulations.

Chapter 3

Reinforcement Learning Framework for Traffic Networks

3.1 The Environment

As described in section 2.3, Markov decision processes with ongoing agent - environment interactions are the foundation of reinforcement learning. The model of the environment is thus a crucial part of reinforcement learning, as information about the quality of a certain action is transferred back to the agent as reward. In this work, the environment is given by a traffic network represented as a graph $G = (\mathcal{V}, \mathcal{E})$, consisting of a finite set of *vertices* or *nodes* $\mathcal{V} \neq \{\}$ and a set of *edges*, *links* or *arcs* $\mathcal{E}$, where each link $e \in \mathcal{E}$ is defined as a tuple of two connected vertices $v \in \mathcal{V}$, so that $e = (v, v')$. Since the main focus in this work are road networks, only *directed* graphs are considered feasible networks, that is graphs where traffic can only flow in the given direction of each link. Every agent $i \in \mathcal{N}$ in the network, with $\mathcal{N}$ denoting the set of all agents corresponding to the definition of a stochastic game in section 2.3.2, is moving from its source node o_i to its destination node d_i. The tuple (o_i, d_i) is called an agent's origin-destination pair or *OD pair*. When moving through the network, an agent travels on a path $p \in \mathcal{P}^k_{(o_i,d_i)} \subset \mathcal{P}$, with $\mathcal{P}$ denoting the set of all simple paths in the network and $\mathcal{P}^k_{(o,d)}$ denoting the subset of paths for a specific origin-destination pair in the network, with the size of the subset determined by the number of paths in the choice set k. A *simple path* in graph theory is defined as a path that has distinct source and destination nodes and no self-loops, so that every node in the network is visited at most once. Especially in larger networks, the number of available paths for every agent has to be restricted as the size of the set of simple paths $\mathcal{P}$ can be too large to compute in reasonable time. Therefore, the size of the choice set is determined by k. Methods to restrict the choice set are discussed in section 2.2.2. For simplicity, the agent specific notation

$\mathcal{P}_i^k = \mathcal{P}_{(o_i,d_i)}^k$ will be used to denote a subset of size k of all simple paths from an agent's origin to its destination. Each path $p \in \mathcal{P}$ consists of a sequence of vertices $p = \{v_1, \cdots, v_n\}$. Every agent knows the vertex v it is located on, as well as all previous nodes it visited so far (for the description of the exact implementation, see section 3.4). In every step of the episode, each agent i chooses an action from the set of possible actions $\mathcal{A}_i^s \in \mathcal{A}$ at current state $s \in \mathcal{S}$. Each action represents a network link $e = (v_s, v_{s'})$ with $v_s, v_{s'} \in p \in \mathcal{P}_i^k$, with the current state node v_s as start and a possible next state node $v_{s'}$ as end. Additionally, the agent perceives the current level of congestion of the link e it chose in the last step and the current level of congestion of all links represented as actions in the current state action set $\mathcal{A}_i^s$. This one-step lookahead follows the assumption that in real world scenarios, congested roads can be spotted early when reaching an intersection, making it possible to react and choose a different road. The congestion level l_e of a link e is determined by the ratio of its current *demand* or *load* d_e, defined as the number of agents currently traveling on that link, and its *capacity* c_e, that is the maximum demand on the link where free traffic flow is possible without leading to congestion. Three congestion levels were defined so that $l_e \in \{0, 1, 2\}$, where level 0 translates to 'not congested', level 1 means 'congested' and level 2 translates to 'heavily congested'. The congestion levels are derived from the ratio of the demand d_e and the capacity c_e as follows:

$$l_e = \begin{cases} 0 & \text{if } \frac{d_e}{c_e} < 1 \\ 1 & \text{if } 1 \geq \frac{d_e}{c_e} < 1.5 \\ 2 & \text{if } \frac{d_e}{c_e} \geq 1.5 \end{cases} \tag{3.1}$$

A state $s \in \mathcal{S}$ in the MDP is thus determined by the current state node v_s from the set of all vertices of all simple paths in the network $v \in p \in \mathcal{P}$, the sequence of all previously visited nodes, the congestion level of the last traveled link and the current congestion levels of all links $e = (v_s, v_{s'})$ that are possible actions. Note that also when considering all simple paths of a network, the set $\{v \in p \in \mathcal{P}\} \neq \mathcal{V}$, as each node $v \in \mathcal{V}$ can be represented as a distinct state multiple times, once for every path $p \in \mathcal{P}$ where $v \in p$. The *cost function* $c : \mathbb{N} \to \mathbb{R}$ represents the current cost for traveling on a link $e \in \mathcal{E}$. In the specific case of this work, the cost function is defined globally as the current travel time on a link e depending on the current demand. The environment observes the demand on each link and computes the corresponding link travel times based on the cost function c. An explanation of the specific cost function used in the experiment is given in section 4.1.1. The reward R_{t+1} returned to the agent i for taking an action A_t in state S_t is computed by

$$R_{i,t+1} = -\,c(d_{e,t}), \tag{3.2}$$

where the current travel time on an edge e with demand d_e is given by the cost function $c(d_e)$ and is returned to the agent as negative reward. The return G of an episode, defined as the sum of all rewards received in an episode (eq. 2.3) is thus the negative value of the sum of all link travel times of the simple path p traveled by an agent i,

$$G_i = -\sum_{e \in p} c(d_e). \tag{3.3}$$

Since many agents are present in the same environment, individual decisions affect the travel times of all other agents as well, making the agent - environment interaction also an agent - agent interaction where agents indirectly influence the decision of others by changing the characteristics of the environment. The MDP can thus be seen as a sequence of *congestion games* at every step of an episode.

3.2 Agent Learning and Behavior

The reinforcement learning framework consists of a network environment where several individual learning agents try to find the fastest route from their *origin* or *start node* o_i to their *destination node* d_i. The setting resembles a congestion game where travel times on single links are computed by a link cost function and increase with the number of agent using it. The scenario can thus be seen as a competition for scarce resources. OD pairs assigned to agents are restricted to pairs of origin and destination nodes where $o_i \neq d_i$. Furthermore, no cycles within a path are allowed, which means every vertex of the network can only be visited once in each trip. Paths that meet those two preconditions - paths have distinct source and destination nodes and every node is just visited once - are called *simple paths*. Q-Learning [116] is used as the learning framework for individual agents. This algorithm is very well studied and has been used for a variety of different reinforcement learning tasks in single agent as well as multi-agent settings. This algorithm updates the Q-value for a certain state action pair at every step of the episode, incorporating the received reward R_{t+1} and a predicted return based on a greedy action taken in the next state S_{t+1} [103]. In Q-Learning, the corresponding action values are updated as follows:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)]. \tag{3.4}$$

The Q value represents the expected reward from taking an action A in a certain state S, called a *state-action pair* (S_t, A_t) and can be seen as the experience an agent collects over time (measured in steps t of an episode). In the specific case of

this work, the actions are the set of possible route choices from a network node v the agent is currently located on. Thus, for every state s there is a different action set $\mathcal{A}^s$. Since the task is modeled as a Markov decision process as described in section 2.3, the state description has to include any information necessary to predict state transitions. As already stated, only simple paths are considered valid paths, rendering already visited nodes invalid action choices. The definition of a state therefore has to include not only the current node location, but also the sequence of all actions taken in previous steps. The parameter $\alpha \in [0, 1]$ is the *learning rate*. The learning rate controls the influence of the most recent experience on the overall evaluation of a certain state-action pair, represented as the Q-value. The higher the learning rate, the more influence the outcome of the most recent decision has on the Q-value. In other words, a higher α value emphasizes an agent's short-term memory whereas a lower α value makes an agent value past experiences more, emphasizing a long-term memory. In the case of $\alpha = 1$, the Q-value represents only the most recent experience, overwriting any previous experiences. $\gamma \in [0, 1]$ is the discount factor as described in section 2.3.1. The future reward is predicted based on the Q-value of the next state under a strictly greedy policy, notated as $\max_a Q(S_{t+1}, a)$. Q-Learning is thus an *off-policy* reinforcement learning algorithm, since it evaluates states based on the return that can be expected following a greedy policy, although the Q-learning agent is actually following another policy, making decisions that diverge from a strictly greedy policy. For a more thorough explanation of the Q-Learning algorithm and also a discussion about performance see [103]. The reward R returned from the environment at each step is the negative travel time, as shown in equation 3.2. It is assumed that agents are generally congestion averse, trying to minimize time spent on congested links. As suggested in [43], commuter satisfaction decreases with increased time spent in congestion, leading to a perceived longer travel time. The negative reward returned from the environment corresponding to the travel time is thus increased if time was spent in congestion, based on the severity of the congestion of the link given by the congestion level l_e as defined in equation 3.1. The weighted negative reward representing the *perceived* travel time and denoted as $R_{w,t}$ is computed as follows:

$$R_{w,t} = \begin{cases} R_t & \text{if } l_e = 0 \\ R_t \cdot 1.25 & \text{if } l_e = 1 \\ R_t \cdot 1.5 & \text{if } l_e = 2 \end{cases} \tag{3.5}$$

To prevent unrealistic congestion scenarios due to every agent departing at the same time, the initial time step of an agent defined as $t_{init} \in \mathbb{N}$ is chosen randomly. The interval of possible time steps for an agent to depart has to be defined according to the characteristics of the network and the simulation needs.

Humans act and learn differently, and those differences in individual behavior should be also reflected in reinforcement learning agents when trying to simulate human route choice behavior. For the decision-making process, two important factors are affecting the resulting decision: First, past experiences as well as current perception of the environment. Second, the weighting of those individual perceptions during the decision-making process. While agents are collecting different experiences over time and thus all have a subjective view of the environment that forms the basis for their further actions, the question how those experiences are exactly influencing future decisions also has to be considered when trying to model individual behavior. The parameters α and γ and the exploration parameters used in the multi-agent reinforcement learning scenario can be sampled from a probability distribution to emphasize differences in human characteristics within a population. The choice of the method and parameter values to enable exploration can be used as an abstract measure of how 'adventurous' an agent's behavior is, while differences in the influence of past experience on decision-making can be modeled with the α parameter. γ can be used to differentiate between agents that consider mainly short-term consequences of their actions and agents that act more farsighted. As the previous chapters have shown, human behavior however is complex and the question how artificial intelligence can be used to model it is part of ongoing research. This is of course also true in traffic scenarios [58]. Trying to condense diversity in behavior to only a handful of variables can thus only be a mere approximation. Another factor to consider is individually different perception of variances in travel time. This is addressed in this work by the introduction of risk-sensitivity in the algorithm, as described in the next section.

3.3 Risk-Sensitive Decision-Making

As discussed in section 2.2, empirical evidence indicates that human behavior cannot accurately be described by the (expected) utility maximization approach of classical economic theories of decision-making. Furthermore, learning and adaptation are problems not considered in the majority of models of behavior, which are mostly static. Reinforcement learning not only has shown major advances in artificial intelligence research, but also offers an opportunity to model realistic learning processes of individuals based on experience. This is also due to the fact that the computational concept of reinforcement learning is based on theories of learning and behavior from psychology and cognitive sciences and thus resembles strategies of learning observed in humans and animals [71]. Incorporating this concept in simulations concerned with the implications of decision-making thus has the potential to generate more accurate models. However, to better replicate human behavior, concepts of theories of decision-making have to be integrated in

the learning process. Behavioral theories suggest that risk is an important factor for decision-making. Many works regarding behavior modeling acknowledge the implications of risk on decisions, such as avoiding risk and the preference of choices with safer outcomes (see section 2.2). Risk-aware reinforcement learning can thus be a concept to further enhance the performance of such models in predicting decisions. Apart from realistically modeling human behavior, there are several reinforcement learning and control tasks that require risk assessment as part of the action decision process. This includes scenarios where the outcomes of decisions could potentially have severe implications. Such implications could include safety risks in tasks where humans are involved or the risk of high cost due to potential damage of the agent [49]. In reinforcement learning, evaluations of the quality of states and actions are usually based on estimates of rewards to be expected defined as the mean value of past rewards (in the case of a decreasing learning rate α, it is a weighted average). In many cases, rewards received for a certain action are dependent on processes in the environment the agent cannot perceive or control, which are observed as random fluctuations in the reward signal. If the reward signal shows high variance, the mean value is not an appropriate criteria to estimate the quality of a decision. This is especially true in scenarios where the distribution of rewards is an important factor the decision-making, such as cases where risk has to be avoided. Several approaches to risk-sensitive reinforcement learning have been proposed in the literature, ranging from methods that strictly avoid risk by evaluating worst-case scenarios to approaches including transformation of the return (utility) based on expected utility theory. A more detailed overview can be found in [66].

Another approach evaluates the risk of an action based on the *TD error* (temporal difference error). The TD error is the difference between the *estimated* return from a state and the updated estimate after the reward for a certain action was received. TD errors are an important aspect of temporal difference learning algorithms [103]. Equation 3.6 shows the computation of the TD error, with $V(S_t)$ representing the value estimate of a state S at time step t:

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t). \tag{3.6}$$

After an action was taken, the new estimate is the received reward R_{t+1} plus the discounted estimate of the new state S_{t+1}. The TD error is also included in the Q-value update of the Q-learning algorithm (eq. 3.4), computing the difference between the expected return from a state-action pair and the updated expectation of the return after receiving the reward:

$$\delta_t = R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(A_t). \tag{3.7}$$

The concept of TD errors in reinforcement learning is supported by the hypothesis of *reward prediction error* described in [92], where observed dopamine neuron activity in laboratory experiments showed patterns corresponding to TD errors. Risk-sensitive reinforcement learning algorithms based on this concept are proposed in [66, 72, 95]. [66] propose a risk-sensitive version of TD-learning and Q-learning algorithms, where the temporal difference value is transformed based on an asymmetrical piecewise linear weighting function depending on whether the TD error is positive or negative. This approach weights outcomes that are worse than expected differently to those that are better. A slightly modified version of this TD-learning algorithm is used in [72], also transforming the TD difference value based on the TD error. The temporal difference update is multiplied by a coefficient $0 \leq \eta \leq 1$ depending the TD error δ:

$$V(S_t) = \begin{cases} V(S_t) + \eta^+ \delta_t & \text{if } \delta_t > 0 \\ V(S_t) + \eta^- \delta_t & \text{otherwise} \end{cases} \tag{3.8}$$

In this approach, the learning rate α is completely substituted by η. The model was compared to fMRI (functional magnetic resonance imaging) images of brain activity where participants had to choose between actions with sure or uncertain outcomes, showing the benefit of including risk awareness in reinforcement learning models that reflect human behavior. A third similar approach for a risk-sensitive variant of the Q-learning algorithm incorporating prospect theory is proposed in [95]. Instead of using a piecewise linear function, the authors use an asymmetrical piecewise nonlinear function as proposed by prospect theory to model utilities for gains and losses, where a positive TD error represents a gain and a negative TD error a loss. The algorithm performance is compared to fMRI data, also showing correlation between neural activity and the risk-sensitive Q-learning algorithm. The nonlinear transformation function $u(x)$ where $x = \delta_t$ was defined as follows:

$$u(x) = \begin{cases} k^+ x^l & \text{if } x \geq 0 \\ k^- x^l & \text{otherwise} \end{cases} \tag{3.9}$$

Different values for k were used (k^+, k^-), depending on whether the TD error is positive or negative. Several combinations of k and l were discussed in the paper. The idea of variable learning rates was also discussed in a different research context. In [19], the authors propose a multi-agent reinforcement learning algorithm WoLF (Win or Learn Fast) based on game theory, using a variable learning rate to make agents learn faster from bad experiences. In [80], the authors propose a reinforcement learning framework that integrates cumulative prospect theory for risk-sensitive control applications.

The reinforcement-learning framework proposed in this work implements risk-sensitivity based on TD prediction errors as described in [66, 72]. TD errors are computed to evaluate outcomes in comparison to predictions based on past experienced returns. TD updates are then transformed based on an asymmetrical piecewise nonlinear function following the approach used in [95]. The function is defined as:

$$Q(S_t, A_t) = \begin{cases} Q(S_t, A_t) + \alpha[\lambda\, \delta_t^{\sigma}] & \text{if } \delta_t \geq 0 \\ Q(S_t, A_t) + \alpha[(1+\lambda)\, \delta_t^{\sigma}] & \text{otherwise} \end{cases} \tag{3.10}$$

Parameter $\lambda \in [0, 1]$ is the *transformation parameter* for the TD error δ_t. The exponent $\sigma \in (0, 1)$ shapes the function so that it is *concave* for gains and *convex* for losses and is thus an approximation of the utility function proposed by prospect theory. The definition of a reference point for gains and losses was identified as a crucial difficulty in the application of prospect theory in various works. In this approach, the reference point is the agent's TD prediction before choosing an action, following the reward prediction hypothesis. Receiving a higher reward, resulting in a more optimistic new expectation of the total return and a *positive* TD error, is thus perceived as a *gain*, whereas a low reward and a worse new expectation of the return results in a *negative* TD error and is perceived as a *loss*. Both parameters η and σ can be sampled from a probability distribution to simulate different levels of risk aversion and different levels of travel time variability across agents.

Behavior under risk plays a crucial role also in classic reinforcement learning tasks that are not explicitly risk-sensitive. The *exploration - exploitation dilemma* in reinforcement learning, which describes the trade-off between the exploitation of actions that are known to result in high rewards and the exploration of new actions and strategies that could potentially lead to even better rewards, but also to worse outcomes is a key difficulty. Reinforcement learning agents have to take risky decisions from time to time to be able to explore the state space and find the best actions. Otherwise, convergence would not be guaranteed. This exploration can also result in actions with less attractive or even poor outcomes. When risk-sensitivity is a key concept, the exploration scheme thus has to be considered as well. The proposed transformation function that weights good or bad outcomes has to be combined with a exploration scheme that takes those different state-action values into account. The widely used ε - *greedy* exploration scheme described in section 2.3.1 assigns a probability of 1 - ε to the action with the highest predicted return. All other actions are chosen with a probability of ε divided by the number of possible actions. Differences in the expected return between those actions have no influence on their probability of being selected. The ε - greedy policy is thus not practical to use in combination with the proposed temporal difference weighting, as selection probabilities of actions aside from the greedy action are not influenced by

the different risk evaluations of those actions. Therefore, the *Boltzmann* or *softmax* exploration scheme is used, which is also widely applied in reinforcement learning research [48, 103]. The selection probability of an action $a \in \mathcal{A}^s$ is computed based on the Boltzmann distribution:

$$P(a \mid s) = \frac{e^{Q(a)/\tau}}{\sum_{a' \in \mathcal{A}^s} e^{Q(a')/\tau}} \tag{3.11}$$

$\tau > 0$ represents the *temperature* parameter. A higher the temperature τ results in more exploratory behavior as the probabilities of all actions align with $\tau \to \infty$ [48]. The advantage of the Boltzmann exploration method is that it assigns selection probabilities to all possible actions $a \in \mathcal{A}^s$ based on their relative value [104]. This means that unlike the ε - greedy strategy, which assigns a high probability to the best action and the same smaller probability to all other actions, the Boltzmann exploration scheme assigns different probabilities to all actions based on their expected return. As the evaluation of risk of an action is integrated in the action value by shifting the Q-value according to the difference in prediction and real outcome of an action, the Boltzmann scheme is useful as it takes this into consideration by differentiating between single actions and their expected returns during exploration.

3.4 Modeling Road Networks as MDP Environments

The agent behavior model is based on the assumption that, as opposed to the concept of selfish routing, agents have no knowledge of the network and the traffic conditions and do not know the shortest path from their origin to their destination. Agents therefore have to rely on trial and error to find their way in the beginning of the learning phase. Finding a way to the desired destination node might thus be a time intensive task. This is particularly true in larger networks with many hundreds or thousands of nodes, making scalability of the reinforcement learning framework a concern. Aside from scalability issues, letting agents simply roam the network to find their destinations can lead to slow learning, especially in the initial phase of the simulation, as agents have to explore many different paths to find routes that lead to their destination. While this might be less significant in single agent scenarios, this approach is impractical for multi-agent tasks with many thousands of agents learning simultaneously. As described in section 1.2, the potential problem that agents could roam around the network without ever finding their destination is addressed in the literature by simply terminating the episode after a predefined number of time steps, even with agents still traveling

through the network. However, this solution does not solve the problem of slow learning and does not help agents reaching their destination faster. This problem can be solved by applying heuristics to the route finding process. As described in section 2.2.2, a possible solution is the restriction of possible paths to a set of routes computed by a k-shortest path algorithm. From this choice set, a predefined route is selected. This approach is used widely in the literature (see section 1.2), however it reduces the MDP to a stateless multi-armed bandit problem. Furthermore, it strictly reduces the number of possible routes to a set of already fixed paths, thus restricting an agent's set of decisions, and eliminates the possibility of choosing individual links en-route. To allow agents to navigate through the network by choosing individual links sequentially and reduce simulation time compared to pure trial and error wayfinding, a different approach is proposed in this work. The main requirement to the model is the possibility for agents to select links in the network sequentially, thus building their path on the fly and with the ability to adapt the route to current congestion patters. To prevent agents from traveling through the network without ever finding their destination or reaching dead ends, the agent's state-action space representation of the network environment has to ensure that agents will reach their destination node while still allowing on the fly route decisions. Each agent's representation of the network environment is thus modeled as a decision tree including all paths the agent can travel on to reach the destination. This approach guarantees that the agent will arrive at the desired destination, independent on the actions chosen along the way. The number of paths in the set of possible routes can be chosen freely. Depending on the network size and application, either all simple paths or a set of k-shortest paths in the network are computed in advance for every agent's OD pair using the all simple path or shortest path algorithm implemented in the python *networkx* library [40]. The set of paths is of course not limited to those options and choice sets can be generated by any method. From those paths, a decision tree is generated with the start node as root and the destination node as leave nodes. An example for such a decision tree for the Braess network is shown in figure 3.1.

As a result, every agent perceives the shared environment as a decision tree, navigating through the tree to find the best path rather than traveling though the whole network. Conceptually, MDPs can be described as trees, consisting of sequences of states and actions. Such a representation is used for example for *backup diagrams* in [103] to visualize reinforcement learning algorithms. Modeling network environments as tree structures is thus a consequent extension of this concept. This approach has several advantages compared to tabular network representations. It ensures that all actions taken by the agent are leading toward its destination, regardless the size of the choice set. This is especially useful in larger networks and big choice sets, making learning progress faster. Furthermore, a state is sufficiently

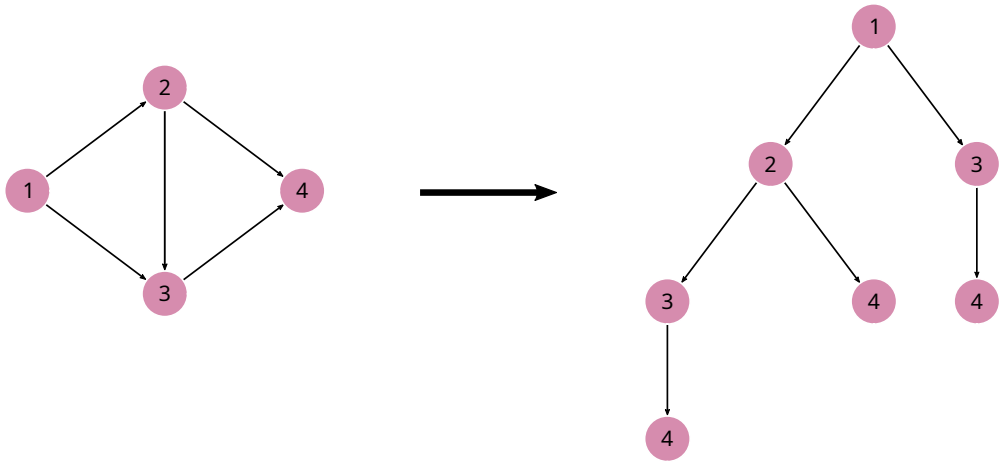

Figure 3.1: An example of an agent decision tree using the Braess network (left). The network is transformed into a decision tree including all possible action sequences that bring an agent to the destination node. Here, the state transitions are deterministic, but the same concept can be applied to MDPs with stochastic environment dynamics.

described as an agent's current location inside the tree. By computing the set of paths in a preprocessing step, no further computation time is needed during the simulation for agents to find possible routes. Computing choice sets of a large network can of course be time intensive, but the computation has to be done only once in the preprocessing step. Network and path data can later be shared to facilitate ongoing research and validation of research results.

In a network environment representation as seen in figure 3.1, every node of the network is represented as a state, actions in the MDP correspond to links in the network. State transitions are thus deterministic, as every action is always followed by the same state. The only feedback agents receive about the current traffic conditions in the network is the reward signal after every step, corresponding to the travel time on the used link. As travel time in such a multi-agent scenario is depending on the actions of all other agents in the environment, the reward signal an agent receives will show fluctuations, with every agent constantly adapting to the new conditions. This leads to a feedback of seemingly random rewards. To provide the agent with more information about the current traffic conditions, the agent observes the current level of congestion of the link it travels on as well as of all links that represent next actions, as described in section 3.1. A node in the

tree is thus modeled as a set of 3×3 matrices for every link in the action set $\mathcal{A}^s$. Every cell of the matrix represents a different state, depending on the traffic conditions experienced on the previously used link and the currently perceived traffic conditions of the next link, i.e. action. The traffic condition refers to the congestion level l_e of a link as defined in equation 3.1. Every node is thus represented as 9^n states, where n describes the number of possible actions (i.e. links) at the current node, a state consisting of all individual Q-values of the next actions depending on their current congestion level and the congestion in the previously used link. Figure 3.2 shows a visualization of such an MDP based on the Braess network. The actual algorithm is shown on page 59. Providing agents with knowledge about the traffic conditions on the current link and a one-step lookahead is a realistic assumption for human behavior in route choice scenarios and offers more flexibility of choice for agents, as they are able to make decisions based on knowledge about the current network state. However, it also introduces non-stationarity. An action can lead to different states depending on link congestion, which means that state transitions in the environment become stochastic. As link congestion levels in multi-agent scenarios with constantly adapting agents cannot be predicted with sufficient precision, the MDP becomes non-stationary, possibly interfering with convergence. However, it can be argued that convergence to the optimal policy anyway is an unrealistic assumption regarding the prediction of human behavior. The topic of non-stationarity in Markov decision processes in multi-agent reinforcement learning is already discussed in section 2.3.2.

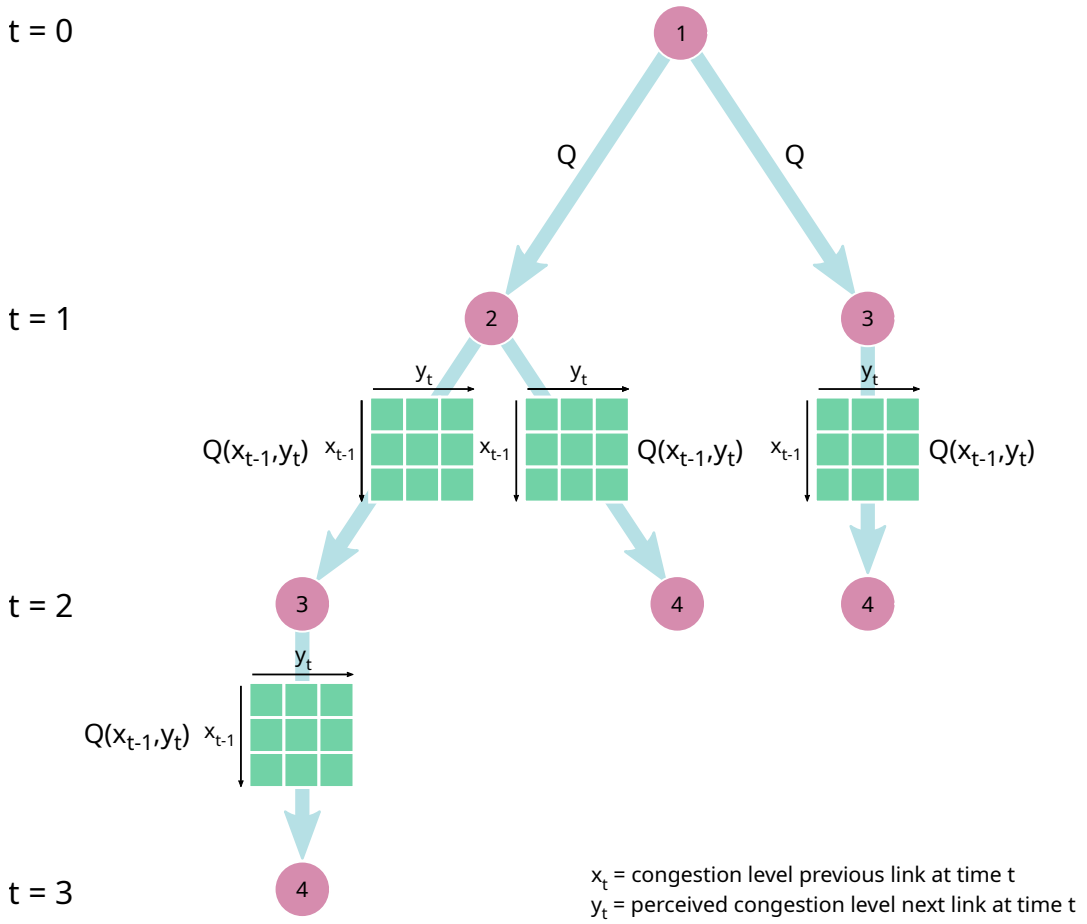

Figure 3.2: Decision tree representation of traffic network with stochastic transition probabilities. Agents observe the traffic conditions of the previous used link as well as the traffic conditions on all links that are possible actions.

Algorithm 1: Q-Learning for Route Choice

Input: Set of agents $\mathcal{N}$ with parameters $\tau > 0,\ \alpha, \gamma, \lambda \in [0, 1],\ \sigma \in (0, 1)$, $t_{init} \in \mathbb{N}$.
Input: Network environment as graph $G = (\mathcal{V}, \mathcal{E})$.

Build decision tree and initialize $Q(s, a) = 0 \quad \forall s \in \mathcal{S}_i \subset \mathcal{S}, a \in \mathcal{A}_i \subset \mathcal{A}$, $\forall i \in \mathcal{N}$
for *n in number of episodes* **do**
  Initialize set of agents not in terminal state $\mathcal{N}_a = \mathcal{N}$.
  while $\mathcal{N}_a \neq \{\}$ **do**
    for $i \in \mathcal{N}_a$ **do**
      if *current time step* $t \geq t_{i,init}$ **then**
        take action $A = e \in \mathcal{E}$ from current state S based on Boltzmann policy.
      else
        do nothing.
    update link demands d and link congestion levels $l \in \{0, 1, 2\}$.
    compute current link travel times from cost function $c(d, c)$.
    for $i \in \mathcal{N}_a$ **do**
      observe travel time as negative reward R and link congestion l_e on current link.
      observe next state S'
      compute weighted reward R_w based on link congestion l_e.
      if $l_e = 0$ **then**
        $R_w = R$
      else if $l_e = 1$ **then**
        $R_w = R \cdot 1.25$
      else if $l_e = 2$ **then**
        $R_w = R \cdot 1.5$
      compute TD error: $\delta = R_w + \gamma \max_a Q(S', a) - Q(S, A)$.
      update Q-value:
      if $\delta \geq 0$ **then**
        $Q(S, A) \leftarrow Q(S, A) + \alpha[\lambda\, \delta^{\sigma}]$
      else
        $Q(S, A) \leftarrow Q(S, A) + \alpha[(1 + \lambda)\, \delta^{\sigma}]$
      if *S' is terminal state* **then**
        $\mathcal{N}_a = \mathcal{N}_a \setminus i$.

Chapter 4

Experiment

4.1 Experiment Setup

4.1.1 Road Network

The Sioux Falls Road Network was used as a simulation environment. It is a well studied network and has been used in many simulation scenarios, including reinforcement learning for traffic modeling (see for example [4]). The network, shown in figure 4.1, consists of 24 nodes and 76 links. The network data was derived from [111], with some changes applied to the original data from the source to simplify the simulation scenario. As Sioux Falls is a representation of a highway network, the capacities c_e of a link e was set to $c_e = 1800\, cars/h$ for all links in the network, following the average capacity values recommended in [70]. As link cost function, the widely applied BPR cost function from the Bureau of Public Roads [82] was used, which is defined as:

$$c(e) = t_e\left(1 + \alpha\left(\frac{d_e}{c_e}\right)^{\beta}\right), \tag{4.1}$$

where t_e is the free flow travel time of an edge e, d_e is the current demand and c_e the capacity on that edge. The parameters α and β were chosen according to the 'traditional' values $\alpha = 0.15$ and $\beta = 4$. The same cost function was applied to all links in the network. To simulate commuter traffic, it was assumed that 80 percent of all OD flows originate in the city center and arrive at a peripheral node of the network, and 20 percent of all flows start at a peripheral node and arrive at a central node. The four nodes with the ids [10, 11, 14, 15] were defined as central nodes, the five nodes with the ids [1, 2, 7, 13, 20] were considered peripheral, resulting in a set of 40 possible OD pairs. From these constraints and the given distribution of flows, the actual amount of flows on those OD pairs was then created randomly.

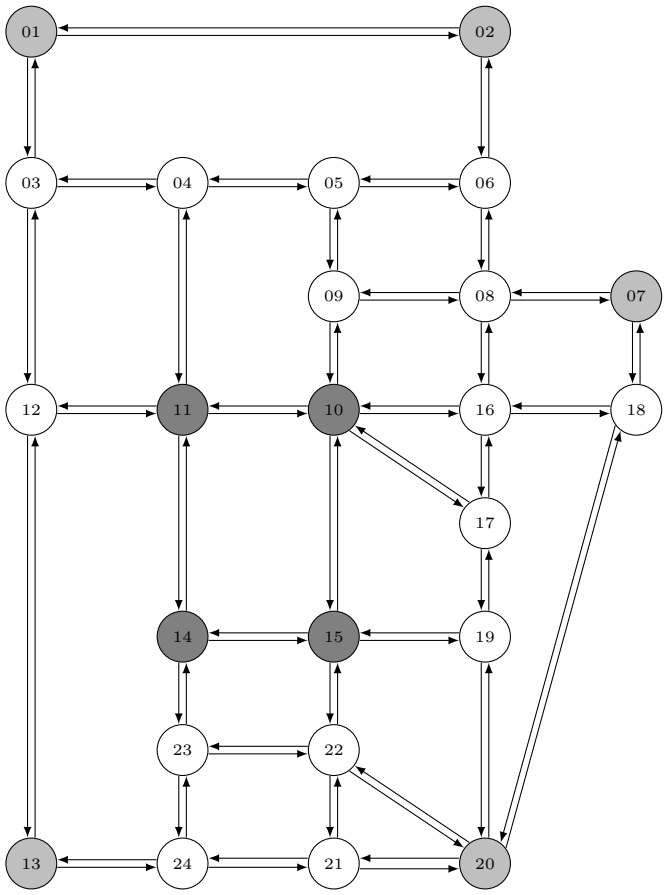

Figure 4.1: The Sioux Falls network. Nodes in darkgray are central nodes, while those in lightgray are peripheral. White nodes are transit nodes that are neither origins nor destinations of OD pairs in the simulation.

The choice set of each individual agents was defined as the set of k-shortest paths in the network, with $k = 1000$. To the authors best knowledge, the size of the choice set thus exceeds that of most approaches described in the literature in this domain. The paths were computed using the python *networkx* library [40] using Yen's k-shortest path algorithm [119].

4.1.2 Agent Parameters

The traffic simulation in the experiment was conducted using 50000 individual reinforcement learning agents, using the risk-sensitive and congestion-averse variant of the Q-learning algorithm described in section 3.2. As proposed in this section,

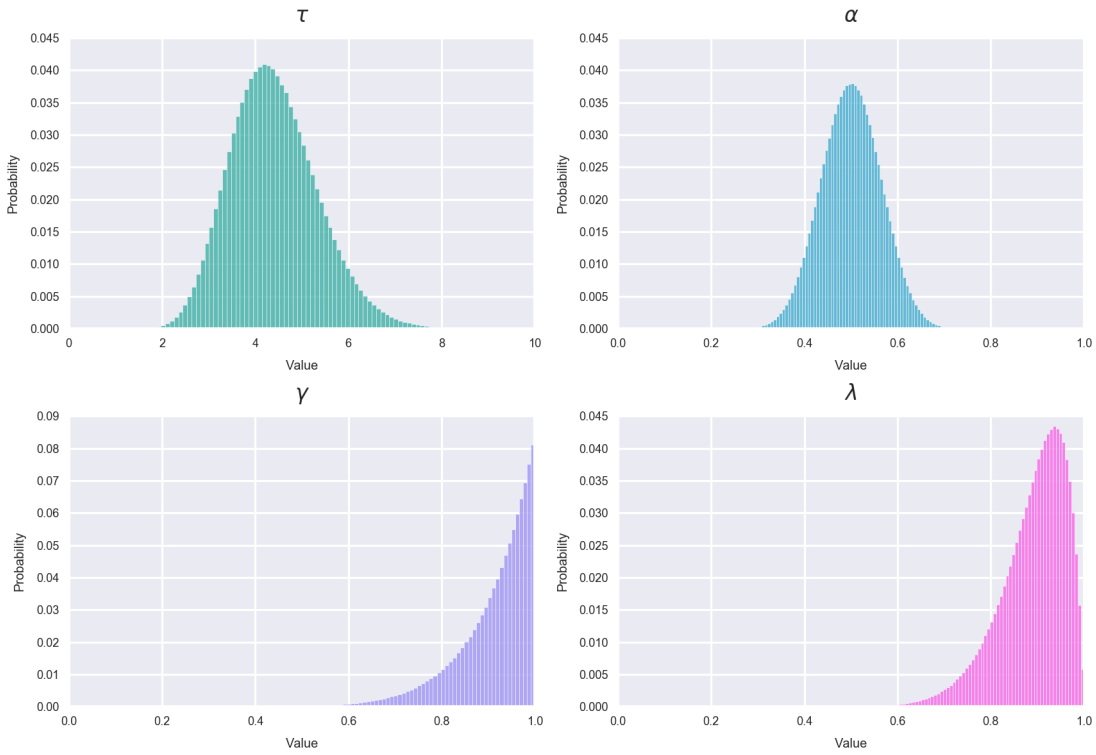

Figure 4.2: Value distributions for agent parameters used in the simulation. τ was sampled from a Gamma(11,0.3) probability distribution, α from a Beta(30,30) distribution, γ from a Beta(10,1) distribution and λ from a Beta(16,2) distribution.

Q-learning parameters can be used to model differences in behavior within a population. Individual values for each agent were thus sampled from different probability distributions to account for diversity in behavior and learning. The shapes of the parameter distributions were determined according to the defined interval of the parameters as well as their suitability for reinforcement learning and their assumed plausibility. The learning rate α and the discount factor γ were sampled from a beta distribution, as both parameters are defined in the interval $[0, 1]$. The learning rate α was sampled from a Beta distribution using the shape parameters $(30, 30)$, the values for the discount factor γ were sampled from a Beta$(10, 1)$ distribution. The initial values of the Boltzmann temperature τ were sampled from a Gamma$(11, 0.3)$ distribution. Additionally, a temperature decay rate was applied to reduce exploratory behavior over time. Every agent's temperature value was updated each episode by $\tau_e = \tau_{e-1} \cdot 0.999$. Individual risk-sensitivity was modeled by sampling the risk coefficient λ from a Beta$(16, 2)$ distribution. For the exponent σ, a fixed value of $\sigma = 0.8$ was assumed. In

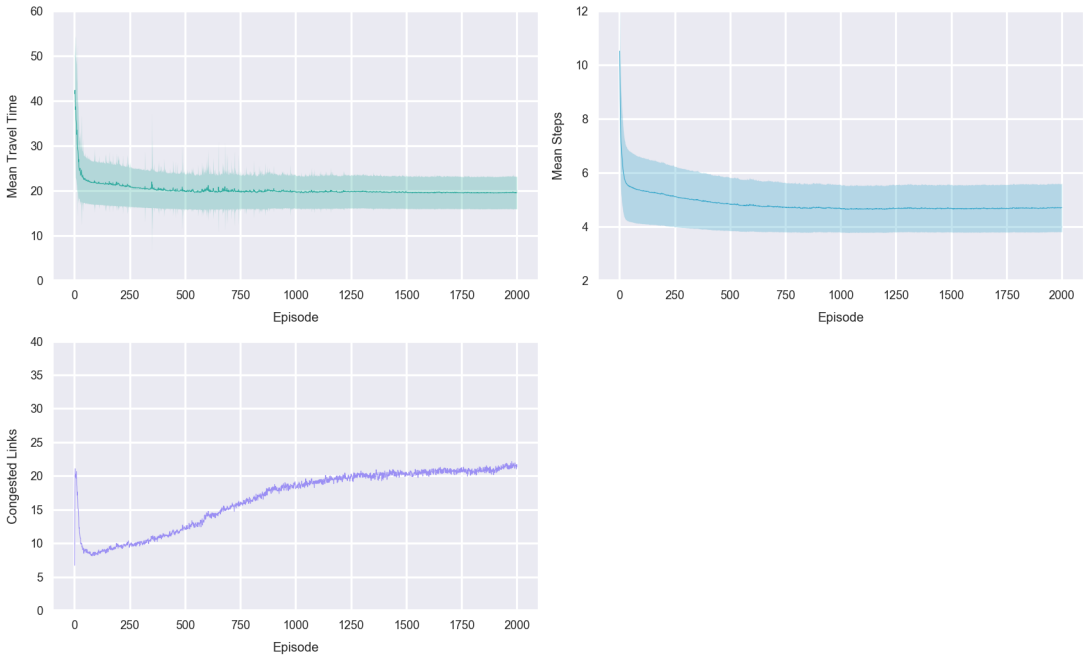

Figure 4.3: Simulation results showing mean travel time, mean number of steps and number of congested links. The results are averaged from 100 individual simulation runs.

every episode, the departure time t_{init} was chosen randomly for each agent, with $t_{init} \in [1, 3]$. Figure 4.2 shows the distribution of sampled values for each parameter in the test case, taken from the parameter samples of all simulation runs.

4.2 Results

The duration of one simulation was set to 2000 sequential episodes. In total, 100 simulation runs were performed using the Google Cloud Computation Engine. The data gained from the individual simulation runs was averaged to generate the final results presented in this section. The results can be seen in figure 4.3, showing the mean travel time in the network, the average steps or links needed to reach the destination from the origin as well as the number of congestions occuring in the network per episode. The light areas in the mean travel time and mean steps diagrams show the average standard deviation, which was computed as the square root of the mean variance in each episode over all simulations. The mean travel time of all agents in the network decreases from 41.6 minutes in episode 1 to 19.6 minutes in episode 2000. The number of steps needed to reach the destination

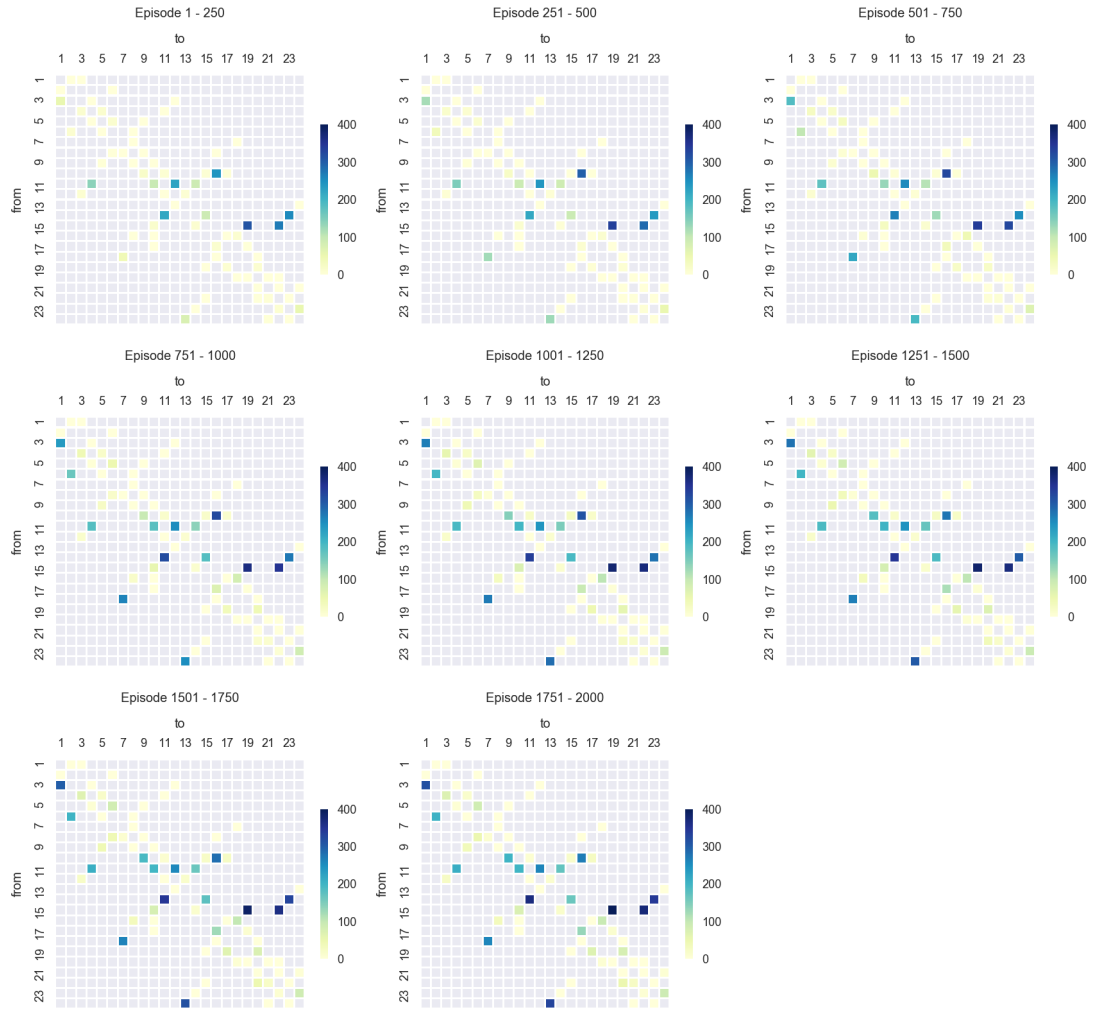

Figure 4.4: OD Matrices showing average patterns of link congestion occurrence. The color indicates the total count of link congestion states that occur on average within the given episode interval. Blue indicates links where congestion occurs more frequently.

decreases from 10.5 steps in episode 1 to 4.7 in episode 2000. The data indicates a fast learning rate in the initial phase of the simulation. The average number of link congestions occuring in each episode shows a different pattern. Starting at an average of 6.7 congested links in episode 1 followed by a peak at around 21 and a quick decline to the initial values, the graph again shows an increase in link congestions, resulting in 21.3 link congestions on average in episode 2000. Figure 4.4 shows the OD matrix of the Sioux Falls network, visualizing the congestion

patterns at different stages of the simulation. For every link in the network, the number of time steps t of each episode was counted where the link is in a state of congestion. The average values from all simulations were then summed for intervals of 250 episodes to give a better impression of the congestion patterns, as the comparison of single episodes is not a meaningful statement due to the stochastic nature of the route choice process. A state of congestion is given if the demand on a link e exceeds its capacity, that is where the current congestion level $l_e > 0$ as given in equation 3.1. It can be seen that the link demand patterns in the network are changing over time, indicating adaptation of the agents to the traffic conditions.

4.3 Discussion

As seen in the average travel time and mean number of steps per episode in figure 4.3, agents show a steep learning curve in the initial phase of the simulation, with average travel time in the network decreasing significantly as agents discover more efficient paths in the network. This is also reflected in the data of average time steps needed per episode, as the average number of steps or links used to reach the destination node in an episode decreases simultaneously, indicating the use of shorter paths with less links. For ever OD pair, the choice set was defined as the set of k-shortest loopless paths with $k = 1000$. thus exceeding the choice set size in most of current literature in this domain. The approach presented using decision trees as state-action space representations shows promising results, with a fast learning rate in the initial period of the simulation. It can thus be a good alternative for current approaches in the domain of multi-agent reinforcement learning for traffic networks. The performance of this approach indicates that it is feasible also for simulations of bigger road networks and with more agents learning simultaneously, something that has to be tested in the future. The number of links in the network experiencing congestion rapidly decreases after an initial peak. This can be attributed to the fact that in the first few episodes, agents have to explore the network to find efficient paths, resulting in low rewards. As many agents choose similar paths, they experience congestion, learning that it leads to longer travel times. The initial phase of the learning process thus reflects adaptation to general network conditions, such as path lengths. The increase of congested links over time seems surprising and contradictory to the principle of congestion aversion implemented in the algorithm. However, rising link demands are also a consequence of the learning process. As agents learn about shorter paths, congestion starts to occur more often, as efficient routes with shorter travel times are more frequently used. The maximum link demand in each episode is shown in figure 4.5 for all links where congestion occurs on average in more than 200 time

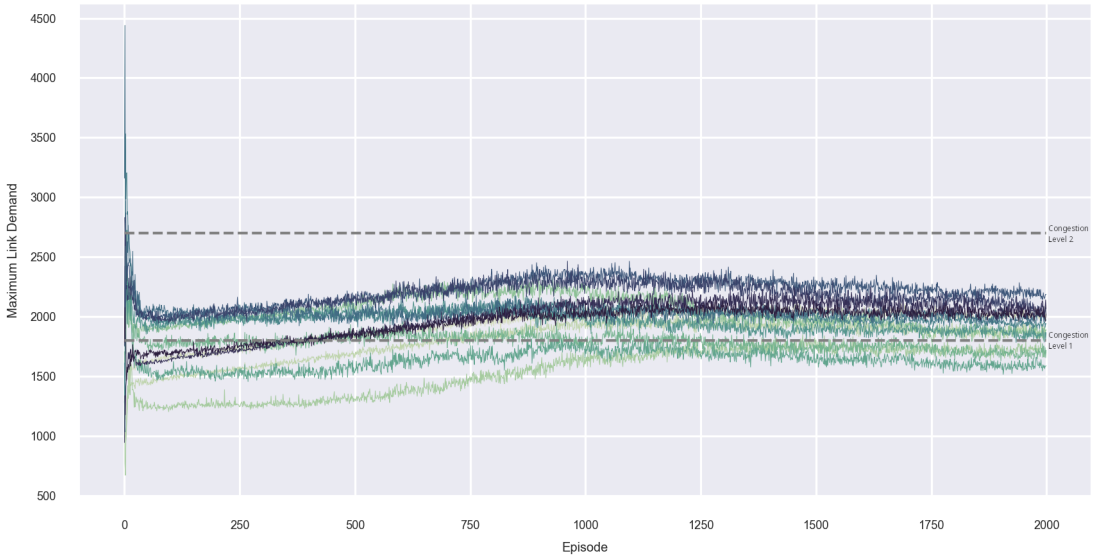

Figure 4.5: Maximum link demands for all links where congestion occurs on average at least 200 times in an interval of 250 episodes. Thresholds of link demands for congestion levels 1 and 2 are shown as dashed lines.

steps in an interval of 250 episodes, as seen in figure 4.4. The graph also shows the defined thresholds of congestion levels as formulated in equation 3.1. It is apparent that agents try to find a balance between reaching the maximum demand while keeping the travel cost as low as possible. Since the BPR cost function used in this experiment is a polynomial function, link travel times show a nonlinear increase with increasing demand. Figure 4.5 shows that the agents try to solve this optimization problem, keeping the link demand as close to the maximum capacity as possible, not exceeding a certain threshold. Congestion aversion was implemented according to equation 3.5, applying a weighting to the perceived travel time depending on the congestion level on a link. The influence of this weighting is also visible in figure 4.5, as maximum link demands never exceed the threshold set for congestion level 2. This shows the influence of the definition of the parameters for the behavior of individual agents, making it necessary to invest further research in how the single parameters could be shaped to represent realistic human behavior and deviations across populations.

Figure 4.6 shows the mean travel times in more detail, for each of the possible OD pairs in the test case. The left column shows mean travel times for OD pairs starting at a central node and ending at a peripheral node, the right column shows the results for OD pairs starting at a peripheral node. As suspected, agents traveling from a peripheral node to the center need less time, as they experience

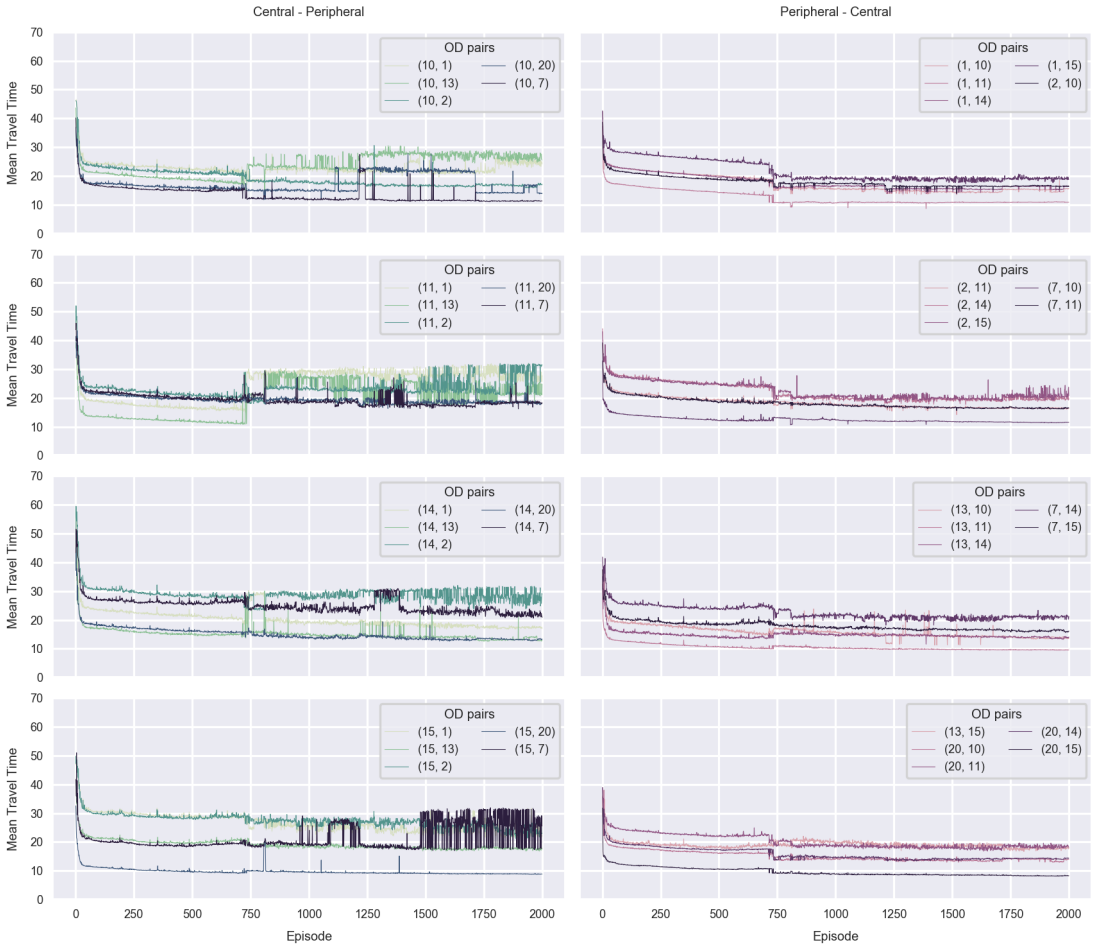

Figure 4.6: Average travel times for all 40 possible origin-destination pairs in the experiment. The left column shows travel times for OD pairs from a central node to a peripheral node, the right column travel times for OD pairs from a peripheral node to a central node.

less congestion due to the fact that flows with the origin at a peripheral node account for only 20 percent of all flows. Looking at the OD flows originating at a central node, after a phase of steady decrease in travel time, a disruption is visible, with a abrupt increase in travel time. This disruption can be observed in many individual simulation runs, often occuring around the mark of the 750th episode. After a steady decrease in travel time follows a sharp increase. This results in more volatile behavior of the agents and stronger fluctuations of travel times. One possible explanation is that agents learn about efficient paths, increasing the probability of choosing such a path. At the same time, exploratory behavior is

decreasing with ongoing simulation duration, leading to increasing probability of greedy actions. This leads to higher demand on certain links used by many agents, resulting in congestion and longer travel times. The fluctuations in mean travel time for some OD pairs also reveals the non-stationarity of the MDP. While trying to find efficient routes, agents experience random fluctuations due to the decisions of all other agents as well as the non-fixed starting time of the commute. Experiencing a sharp increase in travel time results in a large negative TD-error and thus a Q-value update towards a lower expectation of return, making agents prefer other actions. If differences in the Q-values of those actions are small and the policy followed converges towards the greedy policy because of a decrease in exploration, the action with the best expected return can thus change frequently due to the non-stationary environment. This results in volatile behavior of the agents. A diminishing learning rate over time - as done in many reinforcement learning tasks - could solve this problem, with the downside of not allowing continuing adaptation to environmental changes. The network used as an environment also has a big impact on this behavior. The free flow travel times of the links in the Sioux Falls network used in this experiment are very similar for all links, thus resulting in less differentiation between actions and thus to more fluctuation. The results indicate that agents adapt to the conditions in the network, balancing between congestion and detours. Fast learning indicates that the decision-tree approach developed in this work could be applied to larger networks. Volatile behavior is visible in some scenarios, indicating the influence of the non-stationarity of the MDP. The definition of the parameters for learning, risk-sensitivity and congestion aversion are crucial for the performance, and further research is important in this direction.

Chapter 5

Conclusion

The accurate description and prediction of human behavior still is a difficult problem. It is a question spanning various scientific backgrounds, with research predominantly in the fields of economics and psychology. As new technologies emerge, leading to research with the possibility to simulate and understand systems of increasing complexity, modeling human behavior is as important as ever. Research in artificial intelligence also led to continuing interest in the structure human decision-making and learning to incorporate knowledge about human behavior in the development of new artificial agents. The question how human beings make decisions plays an important role in many fields. While huge amounts of data generated daily help to describe human behavior ex post, correct predictions of individual decisions still is difficult. This is especially true if the decision-making context is changing constantly, such as the emergence of new choice opportunities that are not pictured in the data. Furthermore, in scenarios where this data is not available, research still has to rely on models of human choice. Many research domains model complex systems where constant interaction and adaptation of individuals occur as agent-based models, trying to simulate dynamics in such multi-agent systems by modeling entities with distinct behavioral characteristics. Such models also enable learning and adaptation, leading to behavior originating from the agent's direct interactions and observations in the environment. This potentially leads to behavioral patterns that could not be anticipated a priori, and thus can help in understanding such patterns of interaction. In traffic modeling, human behavior is a major factor for the accuracy of traffic simulations. Decision-making models are part of almost every consideration made in traffic models, spanning decisions from the choice of the destination to the choice of the mode of transportation and the choice of the actual route taken. Many choice models used in traffic modeling are derived from classical utility maximization models, while much of research has shown that those concepts fail to really describe behavioral characteristics of human beings. Additionally, the majority of traffic models assume traffic equilibria and

static behavior, where agents are not able to learn from experience and the traffic conditions in the network. Reinforcement learning is seen as a possible approach for learning agents in traffic simulations. This thesis proposed such an approach for simulating route choice behavior of humans in traffic networks, incorporating multi-agent reinforcement learning based on the Q-learning algorithm [116] and findings of behavioral and cognitive research to provide an idea how human behavior in traffic could be approximated with the help of artificial intelligence. A new method was proposed to model road networks as decision trees to be used as reinforcement learning environments. This enables individual agents to choose links in the network en-route, making it possible to adapt to current traffic conditions in the network. It furthermore facilitates learning, as the tree structure guarantees that every agent will arrive at its destination in each episode. This work suggests that reinforcement learning can be used to better reflect human decision-making in traffic by incorporating behavioral research. Sensitivity to the risk of potentially undesired outcomes is included in the learning process, as well as the assumption that humans generally show an aversion towards congestion in commuter traffic. For the test case, algorithm parameters were sampled from Beta and Gamma distributions to indicate diverging behavioral characteristics within a population. As shown in the results, agents achieve a significant reduction in travel time. Learning progress is fast in the initial period of the simulation, showing the advantage of modeling networks as decision trees. The changes in congestion patterns indicate that agents try to avoid congestion, while accepting congested links if a detour leads to long delays. The choice of parameters is identified as a crucial factor for the simulation results.

The following aspects were identified as possible lines of further research. First, while the approach presented introduces aspects of behavioral theories to provide a more accurate description based on current scientific research, it is still based on travel time as the major factor of the utility function. While this is a useful approach to evaluate the performance of the reinforcement learning framework, more aspects have to be included to model human behavior. In future research, an extension of the reward function to picture more diversified utility functions is thus important. This can be achieved in the selection processes of the choice set, as well as with the use of network data providing additional information that can be used in the reward function. Furthermore, agents can be modeled with different levels of knowledge of the network, resulting in more diverse behavior. New agents could be added during the simulation to analyze disturbances in traffic caused by agents with a lack of knowledge about the network. Second, applying the framework for route choice to bigger networks. While the Sioux Falls network is a widely used test network, it does not represent a realistic road network. Results gained from simulations in bigger networks can help to develop the presented approach

further. Third is closely related to the last point and concerns possible approaches for validation and calibration based on real network flow data. Sampling algorithm parameters from a probability distribution is used to represent diversity within the population, but the probability distributions used in the present experiment are based on assumptions. Using data from real networks could be used to calibrate and validate the reinforcement learning route choice framework. A fourth line for further research considers the integration of the RL route choice framework with existing microscopic traffic simulators. The traffic simulator SUMO [61] as an example offers a Python API to access the simulation. This could be used to combine the presented reinforcement learning approach for route choice behavior with microscopic traffic models to achieve more accurate traffic predictions. This offers many possibilities, potentially integrating reinforcement learning models also in other sub-parts of the traffic simulation that rely on the description of human behavior, such as lane change models. Finally, it is important to further integrate risk-sensitivity in the learning process of RL agents. Modeling the risk linked with certain decisions is crucial for modeling realistic behavior. In addition to concepts reviewed and used in this thesis, ongoing research in reinforcement learning shows promising advances in this field, with new developments to incorporate reward distributions in reinforcement learning algorithms. One recent development is *distributional reinforcement learning* [9], where the agent learns whole reward distributions instead of averages. It was shown that distributional reinforcement learning leads to better results than traditional reinforcement learning algorithms. Recent research on brain activity furthermore suggest that learning probability distributions of outcomes for actions is also present in the human brain, encouraging the potential of this reinforcement learning approach [30]. The applicability of distributional reinforcement learning in non-stationary multi-agent scenarios is a question to be answered in the future.

Bibliography

[1] E. Avineri and J. N. Prashker, "Sensitivity to travel time variability: Travelers' learning perspective," *Transportation Research Part C: Emerging Technologies*, vol. 13, no. 2, pp. 157–183, 2005.

[2] B. Baker, I. Kanitscheider, T. Markov, Y. Wu, G. Powell, B. McGrew, and I. Mordatch, "Emergent tool use from multi-agent autocurricula," *arXiv: 1909.07528 [cs, stat]*, 2019.

[3] N. C. Barberis, "Thirty years of prospect theory in economics: A review and assessment," *Journal of Economic Perspectives*, vol. 27, no. 1, pp. 173–196, 2013.

[4] A. L. C. Bazzan, "Aligning individual and collective welfare in complex socio-technical systems by combining metaheuristics and reinforcement learning," *Engineering Applications of Artificial Intelligence*, vol. 79, pp. 23–33, 2019.

[5] A. L. C. Bazzan and R. Grunitzki, "A multiagent reinforcement learning approach to en-route trip building," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 5288–5295.

[6] A. L. C. Bazzan and F. Klügl, "Re-routing agents in an abstract traffic scenario," in *Advances in Artificial Intelligence - SBIA 2008*, G. Zaverucha and A. L. da Costa, Eds., Berlin, Heidelberg, 2008, pp. 63–72.

[7] ——, "A review on agent-based technology for traffic and transportation," *The Knowledge Engineering Review*, vol. 29, no. 3, pp. 375–403, 2014.

[8] S. Bekhor, M. E. Ben-Akiva, and M. S. Ramming, "Evaluation of choice set generation algorithms for route choice models," *Annals of Operations Research*, vol. 144, no. 1, pp. 235–247, 2006.

[9] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, Sydney, NSW, Australia, 2017, pp. 449–458.

[10] R. Bellman, "A markovian decision process," *Indiana University Mathematics Journal*, vol. 6, no. 4, pp. 679–684, 1957.

[11] M. Ben-Akiva, M. J. Bergman, A. J. Daly, and R. Ramaswamy, "Modelling inter urban route choice behavior," in *Proceedings of the Ninth International Symposium on Transportation and Traffic Theory*, J. Volmuller and R. Hamerslag, Eds., Utrecht, The Netherlands, 1984, pp. 299–330.

[12] M. E. Ben-Akiva and S. R. Lerman, *Discrete choice analysis: theory and application to travel demand*, 9. Cambridge, Mass, 1985.

[13] E. Ben-Elia, I. Erev, and Y. Shiftan, "The combined effect of information and experience on drivers' route-choice behavior," *Transportation*, vol. 35, no. 2, pp. 165–177, 2008.

[14] E. Ben-Elia and Y. Shiftan, "Which road do i take? a learning-based model of route-choice behavior with real-time information," *Transportation Research Part A: Policy and Practice*, vol. 44, no. 4, pp. 249–264, 2010.

[15] D. P. Bertsekas, *Abstract dynamic programming*. 2018.

[16] Bo Chen and H. H. Cheng, "A review of the applications of agent technology in traffic and transportation systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 2, pp. 485–497, 2010.

[17] E. A. I. Bogers, M. Bierlaire, and S. P. Hoogendoorn, "Modeling learning in route choice," *Transportation Research Record*, vol. 2014, no. 1, pp. 1–8, 2007.

[18] E. Bonabeau, "Agent-based modeling: Methods and techniques for simulating human systems," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 7280–7287, Supplement 3 2002.

[19] M. Bowling and M. Veloso, "Multiagent learning using a variable learning rate," *Artificial Intelligence*, vol. 136, no. 2, pp. 215–250, 2002.

[20] J. A. Boyan and M. L. Littman, "Packet routing in dynamically changing networks: A reinforcement learning approach," in *Advances in Neural Information Processing Systems 6*, J. D. Cowan, G. Tesauro, and J. Alspector, Eds., 1994, pp. 671–678.

[21] D. Braess, "Über ein Paradoxon aus der Verkehrsplanung," *Unternehmensforschung*, vol. 12, no. 1, pp. 258–268, 1968.

[22] W. Burghout, "Hybrid microscopic-mesoscopic traffic simulation," Ph.D. dissertation, Royal Institute of Technology, Deptartment of Infrastructure, Stockholm, 2004.

[23] L. Buşoniu, R. Babuška, and B. De Schutter, "Multi-agent reinforcement learning: An overview," in *Innovations in Multi-Agent Systems and Applications - 1*, D. Srinivasan and L. C. Jain, Eds., ed. by J. Kacprzyk, vol. 310, Berlin, Heidelberg, 2010, pp. 183–221.

[24] C. Camerer, *Behavioral game theory: experiments in strategic interaction*. New York, N.Y. : Princeton, N.J, 2003.

[25] A. C. Chapman, D. S. Leslie, A. Rogers, and N. R. Jennings, "Convergent learning algorithms for unknown reward games," *SIAM Journal on Control and Optimization*, vol. 51, no. 4, pp. 3154–3180, 2013.

[26] E. Cherchi, "Modelling individual preferences, state of the art, recent advances and future directions," in *Travel behaviour research in an evolving world: selected papers from the 12th International Conference on Travel Behaviour Research ; [Jaipur, Rajasthan, India, December 13-18, 2009]*, Jaipur, India, 2009.

[27] S. P. M. Choi and D.-Y. Yeung, "Predictive q-routing: A memory-based reinforcement learning approach to adaptive traffic control," in *Advances in Neural Information Processing Systems 8*, D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, Eds., 1996, pp. 945–951.

[28] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proceedings 15th National Conference on Artificial Intelligence (AAAI-98)*, 1998, pp. 746–752.

[29] J. R. Correa, A. S. Schulz, and N. E. Stier-Moses, "Selfish routing in capacitated networks," *Mathematics of Operations Research*, vol. 29, no. 4, pp. 961–976, 2004.

[30] W. Dabney, Z. Kurth-Nelson, N. Uchida, C. K. Starkweather, D. Hassabis, R. Munos, and M. Botvinick, "A distributional code for value in dopamine-based reinforcement learning," *Nature*, vol. 577, no. 7792, pp. 671–675, 2020.

[31] C. Daganzo, *Fundamentals of transportation and traffic operations*, 1st ed. Oxford ; New York, 1997.

[32] T. A. Domencich and D. McFadden, *Urban travel demand: a behavioral analysis: a Charles River Associates research study*, in collab. with C. R. Associates, 93. Amsterdam : New York, 1975.

[33] A. Dorri, S. S. Kanhere, and R. Jurdak, "Multi-agent systems: A survey," *IEEE Access*, vol. 6, pp. 28 573–28 593, 2018.

[34] K. Dresner and P. Stone, "Multiagent traffic management: A reservation-based intersection control mechanism," in *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2*, New York, New York, 2004, pp. 530–537.

[35] ——, "Multiagent traffic management: Opportunities for multiagent learning," in *Learning and Adaption in Multi-Agent Systems*, K. Tuyls, P. J. Hoen, K. Verbeeck, and S. Sen, Eds., Berlin, Heidelberg, 2006, pp. 129–138.

[36] I. Erev and G. Barron, "On adaptation, maximization, and reinforcement learning among cognitive strategies.," *Psychological Review*, vol. 112, no. 4, pp. 912–931, 2005.

[37] S. Gao, E. Frejinger, and M. Ben-Akiva, "Adaptive route choices in risky traffic networks: A prospect theory approach," *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 5, pp. 727–740, 2010.

[38] R. Grunitzki and A. L. C. Bazzan, "Comparing two multiagent reinforcement learning approaches for the traffic assignment problem," in *2017 Brazilian Conference on Intelligent Systems (BRACIS)*, 2017, pp. 139–144.

[39] R. Grunitzki, G. d. O. Ramos, and A. L. C. Bazzan, "Individual versus difference rewards on reinforcement learning for route choice," in *2014 Brazilian Conference on Intelligent Systems*, 2014, pp. 253–258.

[40] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, G. Varoquaux, T. Vaught, and J. Millman, Eds., 2008, pp. 11–16.

[41] J. S. Hartford, J. R. Wright, and K. Leyton-Brown, "Deep learning for predicting human strategic behavior," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds., 2016, pp. 2424–2432.

[42] P. Hernandez-Leal, M. Kaisers, T. Baarslag, and E. M. de Cote, "A survey of learning in multiagent environments: Dealing with non-stationarity," *arXiv:1707.09183 [cs]*, 2019.

[43] C. D. Higgins, M. N. Sweet, and P. S. Kanaroglou, "All minutes are not equal: Travel time and the effects of congestion on commute satisfaction in canadian cities," *Transportation*, vol. 45, no. 5, pp. 1249–1268, 2018.

[44] J. Hu and M. P. Wellman, "Nash q-learning for general-sum stochastic games," *Journal of Machine Learning Research*, vol. 2003, no. 4, pp. 1039–1069, 2003.

[45] INRIX, *Global traffic scorecard 2019*, 2020. [Online]. Available: `https://inrix.com/scorecard/` (visited on 08/20/2020).

[46] E. J. van de Kaa, "Extended prospect theory: Findings on choice behaviour from economics and the behavioural sciences and their relevance for travel behaviour," Ph.D. dissertation, TU Delft, 2008.

[47] ——, "Prospect theory and choice behaviour strategies: Review and synthesis of concepts from social and transport sciences," *European Journal of Transport and Infrastructure Research*, vol. 10, no. 4, 2010.

[48] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.

[49] G. Kahn, A. Villaflor, V. Pong, P. Abbeel, and S. Levine, "Uncertainty-aware reinforcement learning for collision avoidance," *arXiv:1702.01182 [cs]*, 2017.

[50] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica*, vol. 47, no. 2, pp. 263–292, 1979.

[51] E. Koutsoupias and C. Papadimitriou, "Worst-case equilibria," in *Proceedings of the 16th Annual Conference on Theoretical Aspects of Computer Science*, Trier, Germany, 1999, pp. 404–413.

[52] S. Krauß, "Microscopic modeling of traffic flow: Investigation of collision free vehicle dynamics," Ph.D. dissertation, Universität zu Köln, 1998.

[53] M. Lanctot, V. Zambaldi, A. Gruslys, A. Lazaridou, K. Tuyls, J. Perolat, D. Silver, and T. Graepel, "A unified game-theoretic approach to multiagent reinforcement learning," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, pp. 4190–4203.

[54] G. Laurent, L. Matignon, and N. Fort-Piat, "The world of independent learners is not markovian.," *KES Journal*, vol. 15, pp. 55–64, 2011.

[55] J. Z. Leibo, E. Hughes, M. Lanctot, and T. Graepel, "Autocurricula and the emergence of innovation from social interaction: A manifesto for multi-agent intelligence research," *arXiv:1903.00742 [cs, q-bio]*, 2019.

[56] L. L. Lemos and A. L. C. Bazzan, "Combining adaptation at supply and demand levels in microscopic traffic simulation: A multiagent learning approach," *Transportation Research Procedia*, vol. 37, pp. 465–472, 2019.

[57] D. S. Leslie and E. J. Collins, "Individual q-learning in normal form games," *SIAM Journal on Control and Optimization*, vol. 44, no. 2, pp. 495–514, 2005.

[58] G. Leu, N. J. Curtis, and H. Abbass, "Modeling and evolving human behaviors and emotions in road traffic networks," *Procedia - Social and Behavioral Sciences*, vol. 54, pp. 999–1009, 2012.

[59] K. Leyton-Brown and Y. Shoham, *Essentials of game theory: a concise, multidisciplinary introduction*, 3. San Rafael, California, 2008.

[60] Z. Li and D. Hensher, "Prospect theoretic contributions in understanding traveller behaviour: A review and some comments," *Transport Reviews*, vol. 31, no. 1, pp. 97–115, 2011.

[61] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wiessner, "Microscopic traffic simulation using SUMO," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2575–2582.

[62] R. D. Luce and P. Suppes, "Preference, utility, and subjective probability," in *Handbook of Mathematical Psychology*, R. D. Luce, R. R. Bush, and E. Galanter, Eds., vol. 3, New York, 1965, pp. 249–410.

[63] C. F. Manski, "The structure of random utility models," *Theory and Decision*, vol. 8, no. 3, pp. 229–254, 1977.

[64] ——, "Structural models for discrete data: The analysis of discrete choice," *Sociological Methodology*, vol. 12, pp. 58–109, 1981.

[65] M. G. McNally, "The four-step model," in *Handbook of Transport Modelling*, D. A. Hensher and K. J. Button, Eds., 2nd ed., 2007, pp. 35–53.

[66] O. Mihatsch and R. Neuneier, "Risk-sensitive reinforcement learning," *Machine Learning*, vol. 49, no. 2, pp. 267–290, 2002.

[67] D. E. Moriarty and P. Langley, "Learning cooperative lane selection strategies for highways," in *Proceedings 15th National Conference on Artificial Intelligence (AAAI-98)*, 1998, pp. 684–691.

[68] S. Nakayama, R. Kitamura, and S. Fujii, "Drivers' route choice rules and network behavior: Do drivers become rational and homogeneous through learning?" *Transportation Research Record*, vol. 1752, no. 1, pp. 62–68, 2001.

[69] J. Nash Jr., "Non-cooperative games," *Annals of Mathematics*, vol. 54, no. 2, pp. 286–295, 1951.

[70] E. National Academies of Sciences and Medicine (U.S.) and Transportation Research Board, *Highway capacity manual: a guide for multimodal mobility analysis.* 2016.

[71] Y. Niv, "Reinforcement learning in the brain," *Journal of Mathematical Psychology*, vol. 53, no. 3, pp. 139–154, 2009.

[72] Y. Niv, J. A. Edlund, P. Dayan, and J. P. O'Doherty, "Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain," *Journal of Neuroscience*, vol. 32, no. 2, pp. 551–562, 2012.

[73] T. B. F. de Oliveira, A. L. C. Bazzan, B. C. da Silva, and R. Grunitzki, "Comparing multi-armed bandit algorithms and q-learning for multiagent action selection: A case study in route choice," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–8.

[74] J. d. D. Ortúzar and L. G. Willumsen, *Modelling transport*, 3rd ed. Chichester New York, 2001.

[75] M. J. Osborne and A. Rubinstein, *A course in game theory*. Cambridge, Mass, 1994.

[76] A. de Palma, M. Ben-Akiva, D. Brownstone, C. Holt, T. Magnac, D. McFadden, P. Moffatt, N. Picard, K. Train, P. Wakker, and J. Walker, "Risk, uncertainty and discrete choice models," *Marketing Letters*, vol. 19, no. 3, pp. 269–285, 2008.

[77] C. Papadimitriou, "Algorithms, games, and the internet," in *Proceedings of the Thirty-third Annual ACM Symposium on Theory of Computing*, New York, NY, USA, 2001, pp. 749–753.

[78] C. H. Papadimitriou, "The complexity of finding nash equilibria," in *Algorithmic Game Theory*, N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, Eds., New York, 2007, pp. 461–486.

[79] A. C. Pigou, *The Economics of Welfare*. London, 1920.

[80] L. A. Prashanth, C. Jie, M. Fu, S. Marcus, and C. Szepesvári, "Cumulative prospect theory meets reinforcement learning: Prediction and control," *arXiv:1506.02632 [cs, math]*, 2016.

[81] C. G. Prato, "Route choice modeling: Past, present and future research directions," *Journal of Choice Modelling*, vol. 2, no. 1, pp. 65–100, 2009.

[82] U. B. of Public Roads, *Traffic Assignment Manual for Application with a Large, High Speed Computer*. 1964.

[83] M. S. Ramming, "Network knowledge and route choice," Thesis, Massachusetts Institute of Technology, 2002.

[84] R. W. Rosenthal, "A class of games possessing pure-strategy nash equilibria," *International Journal of Game Theory*, vol. 2, no. 1, pp. 65–67, 1973.

[85] T. Roughgarden, "Selfish routing," Ph.D. dissertation, Cornell University, 2002.

[86] ——, *Selfish Routing and The Price Of Anarchy*. Cambridge, Massachusetts, 2005.

[87] ——, "Routing games," in *Algorithmic Game Theory*, N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, Eds., New York, 2007, pp. 461–486.

[88] T. Roughgarden and E. Tardos, "How bad is selfish routing?" In *Proceedings 41st Annual Symposium on Foundations of Computer Science*, Redondo Beach, CA, USA, 2000, pp. 93–102.

[89] A. Rubinstein, *Modeling bounded rationality*. Cambridge, Mass, 1998.

[90] S. J. Russell and P. Norvig, *Artificial intelligence: a modern approach*, Third Edition, Pearson New International Edition. Harlow, 2014.

[91] J. Scholz and R. L. Church, "Shortest paths from a group perspective — a note on selfish routing games with cognitive agents," *ISPRS International Journal of Geo-Information*, vol. 7, no. 345, p. 16, 2018.

[92] W. Schultz, P. Dayan, and P. R. Montague, "A neural substrate of prediction and reward," *Science*, vol. 275, no. 5306, pp. 1593–1599, 1997.

[93] R. Selten, "Bounded rationality," *Journal of Institutional and Theoretical Economics (JITE) / Zeitschrift für die gesamte Staatswissenschaft*, vol. 146, no. 4, pp. 649–658, 1990.

[94] ——, "What is bounded rationality?" In *Bounded Rationality: The Adaptive Toolbox*, G. Gigerenzer and R. Selten, Eds., Cambridge, Mass., 2001, pp. 13–36.

[95] Y. Shen, M. J. Tobia, T. Sommer, and K. Obermayer, "Risk-sensitive reinforcement learning," *Neural Computation*, vol. 26, no. 7, pp. 1298–1328, 2014.

[96] Y. Shoham and K. Leyton-Brown, *Multiagent systems: algorithmic, game-theoretic, and logical foundations*. Cambridge ; New York, 2009.

[97] Y. Shoham, R. Powers, and T. Grenager, "Multi-agent reinforcement learning: A critical survey," Computer Science Department, Stanford University, Technical Report, 2003.

[98] ——, "If multi-agent learning is the answer, what is the question?" *Artificial Intelligence*, vol. 171, no. 7, pp. 365–377, 2007.

[99] H. A. Simon, "A behavioral model of rational choice," *The Quarterly Journal of Economics*, vol. 69, no. 1, pp. 99–118, 1955.

[100] Statistical Office of the European Communities (EUROSTAT), *Energy, transport and environment statistics: 2019 edition*. Luxembourg, 2019.

[101] P. Stone, "Multiagent learning is not the answer. it is the question," *Artificial Intelligence*, vol. 171, no. 7, pp. 402–405, 2007.

[102] P. Stone and M. Veloso, "Multiagent systems: A survey from a machine learning perspective," *Autonomous Robots*, vol. 8, no. 3, pp. 345–383, 2000.

[103] R. S. Sutton and A. G. Barto, *Reinforcement learning: an introduction*, Second Edition. Cambridge, Massachusetts, 2018.

[104] C. Szepesvári, *Algorithms for Reinforcement Learning*. San Rafael, California, 2010.

[105] É. Tardos and T. Wexler, "Network formation games and the potential function method," in *Algorithmic Game Theory*, N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, Eds., New York, 2007, pp. 487–516.

[106] G. Tesauro, "Extending q-learning to general adaptive multi-agent systems," in *Advances in Neural Information Processing Systems 16 (NIPS 2003)*, S. Thrun, L. K. Saul, and B. Schölkopf, Eds., 2004, pp. 871–878.

[107] L. L. Thurstone, "A law of comparative judgment," *Psychological Review*, vol. 34, no. 4, pp. 273–286, 1927.

[108] H. Timmermans, "On the (ir)relevance of prospect theory in modelling uncertainty in travel decisions," *European Journal of Transport and Infrastructure Research*, vol. 10, no. 4, 2010.

[109] K. Train, *Qualitative choice analysis: theory, econometrics, and an application to automobile demand*, 10. Cambridge, Mass, 1986.

[110] ——, *Discrete choice methods with simulation*, 2nd ed. Cambridge ; New York, 2009.

[111] Transportation Networks for Research Core Team, *Transportation networks for research github repository*. [Online]. Available: `https://github.com/bstabler/TransportationNetworks` (visited on 06/09/2020).

[112] K. Tuyls and G. Weiss, "Multiagent learning: Basics, challenges, and prospects," *AI Magazine*, vol. 33, no. 3, pp. 41–52, 2012.

[113] A. Tversky and D. Kahneman, "Advances in prospect theory: Cumulative representation of uncertainty," *Journal of Risk and Uncertainty*, vol. 5, no. 4, pp. 297–323, 1992.

[114] J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton, N.J., 1944.

[115] J. G. Wardrop, "Some theoretical aspects of road traffic research.," *Proceedings of the Institution of Civil Engineers*, vol. 1, no. 3, pp. 325–362, 1952.

[116] C. Watkins, "Learning from delayed rewards," Ph.D. dissertation, University of Cambridge, 1989.

[117] J. R. Wright and K. Leyton-Brown, "Beyond equilibrium: Predicting human behavior in normal-form games," in *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[118] C. Wu, A. Kreidieh, K. Parvate, E. Vinitsky, and A. M. Bayen, "Flow: A modular learning framework for autonomy in traffic," *arXiv:1710.05465 [cs]*, 2019.

[119] J. Y. Yen, "Finding the k shortest loopless paths in a network," *Management Science*, vol. 17, no. 11, pp. 712–716, 1971.

[120] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *arXiv:1911.10635 [cs, stat]*, 2019.