# Tongji University

MAURIZIO VASSALLO
Student ID:
Academic Tutor: a

# A Collision Avoidance System Using Reinforcement Learning

# Contents

# 1 Introduction

This report is drawn up after the development of the thesis project at Tongji University (同济大学) for the Double Degree project Politong.

The thesis project aims to develop a machine learning algorithm that allows an autonomous vehicle to avoid obstacles.
The research will focus on Reinforcement Learning methods.

Nowadays it is possible to hear more and more about Autonomous Driving Vehicles. According to research firm, autonomous vehicles will match or exceed human safety by late 2020s and fulfill all mobility needs in 2040s to 2060s. Optimists predict that by 2030, autonomous vehicles will be sufficiently reliable, affordable and common to displace most human driving, providing huge savings and benefits. However, there are good reasons to be skeptical. There is considerable uncertainty concerning autonomous vehicle development, benefits and costs, travel impacts, and consumer demand. Considerable progress is needed before autonomous vehicles can operate reliably in mixed urban traffic, heavy rain and snow, unpaved and unmapped roads, and where wireless access is unreliable[1].

The autonomous cars(also known as a self-driving cars or a driverless cars) are a vehicle that are capable of sensing its environment and navigating without human input The ability of autonomous vehicles to operate without human intervention depends on their level of technological sophistication, in accordance with the current six-degree autonomy scale proposed by the International Society of Automotive Engineers (**SAE**)[2]. There are 6 levels of driving automation, from level 0 (no automation) to level 5 (full unlimited automation); intermidiate levels (1 to 3) are considered semi-autonomous[3, 4].
Currently we are between level 2 and 3, so even if the current technology is behind the famous Level 5 of driving automation, there is lot of work to make it happen.

The idea behind these cars is quite simple: outfit the vehicles with sensors that can track all the objects around and make the cars understand the world around them. Autonomous vehicles are driven using technology such as GPS, odometry, radars, laser lights and other devices[4]. These sensors themselves do not make the car 'smart', what make it autonomous are the big computers inside and the algorithms they are running. Usually these softwares run neural networks: these take as input the sensor recordings, elaborate them and output some values like steering angle, accelerate, brake or other important values. Even if the idea behind these technological innovative vehicles is simple the implementation is not: not enough hardware computation, not enough training data, problems with handle different weather conditions(fog, rain, snow, etc.), the current regolation remains in a nascent stage.

Even if this tecnology is not diffused yet, there are many potential benefits that autonomous vehicles could introduce in our society:

- **Transportation Safety**. The most notable predicted benefit of autonomous vehicle technology is a substantial reduction in the human and economic toll of traffic accidents. ndeed, impairment, distractions, and fatigue alone account for over 50% of all fatal crashes. The use of autonomous vehicles could significantly reduce the incidence of such crashes, as vehicles with no human operators are never drunk, distracted, fatigued, or otherwise susceptible to human failings.

- **Access to Transportation**. Another important potential benefit of autonomous vehicle technology is increased mobility for populations currently unable or not permitted to operate traditional vehicles. These populations include older citizens, the disabled, people too young to drive, and others without a driver's license.

- **Traffic Congestion and Land Use**. Autonomous vehicles could reduce congestion and change the way in which cities are planned. Most car are moving only for 5% of their lives, for the 95% they are parked[5]! For this reason lot of space is dedicated to parking lots; the same space that could be used for different purposes (green spaces, etc.).

- **Energy and Emissions**. Autonomous vehicle technology has the potential to reduce both energy consumption and pollution thanks to efficiencies gained through smoother acceleration and deceleration and increased roadway capacity.[6]

To do: find a way to link this part with teh following one!!

# 2 Reinforcement Learning

## 2.1 Introduction

Humans and animals learn through a process of trial and error. This process is based on a reward mechanism that provide a response to our behaviors. The goal of this process is to incentivize the repetitions of actions which trigger positive rewards and disincentivize the repetition of actions which trigger negative ones. Inspired by how animals and humans learn, **Reinforcement Learning** is built around the idea trial and error from an interaction with the enviroment

Reinforcement Learning tries to solve the problem in which a decision-maker, Agent, can perform some action inside a world, called Enviroment. The agent senses the enviroment through the state; for each state the agent has to perform an action. These actions result in an effect: they change the agent's state and give it a feedback, the Reward. The reward is a value which indicated if the action performd is good, then the reward is positive, or it is bad, the reward is negative. Given these rewards the agent understand what action to perform in a given state to get a positive reward.
This cycle of state-action-reward is repeted many time. The goal of the agent is to find a policy such that the actions perform lead to the maximum reward possible. In particual the agent wants to maximize the cumulative reward, the sum of rewards over a period of time. The difference between maximizing the instant reward and cumulative reward is an important distinction because: choosing an action which lead to the maximum next reward could in the next state bring to an end state (game over); instead maximizing the cumulative reward means thinking in a long-term horizon, so that the next states should all have a positive reward (even if the action chosen in one state is not the one that the maximum reward).

## 2.2 Reinforcement Learning compared to other methods

Machine learning is the science of getting computers to act without being explicitly programmed. Machine Learning methods are appropriate in application settings where people are unable to provide precise specifications for desired program behavior, but where examples of desired behavior are available, or where it is possible to assign a measure of goodness to examples of behavior[7].
There are mainly 3 methods to train Machine Learning algorithms, each with its advantages and disavantages. The 3 methods are:

- Supervised Learning. In this method the algorithm is trained with labeled data; the training examples are of the form ($x_i$,$y_i$), with $x_i$ a *n-dimentional* vector and $y_i$ a scalar (it can represent either a class or a floating value).

The learner (either a classifier or a predictor) tries to find a good mapping fuction that maps an input $x_i$ to its correspoding $y_i$. During this learning process the error between the $y_i$ predicted and the actuall value is calculated and used to make the method learning (usually with Gradient Descent and Back Propagation) and decresing the error over time. A popular example is to classify and image given the raw pixels.

- Unsupervised Learning. The main difference between Unsupervised Learning and Supervised Learning is that in the former method data does not have labels. In this setting, the goal is usually find the relationship in the elements of the dataset. This is done calculating the distance (Euclidian, Hamming, etc.) between the points: if the distance between the points is small, they may share similar characteristics. A popular example is detecting a person purchase preferences analyzing his shopping list with other people shopping's lists.

- Reinforcement Learning. Reinforcement Learning comes into play when examples of desired behavior are not available but where it is possible to score examples of behavior according to some performance criterion[7]. In general, in Reinforcement Learning the goal is to maximized an unknown reward fuction through and trial and error process. A popular example is a car that learns to drive in a given enviroment, with no given prior knowledge given.

Supervised and Reinforcement Learning could be considered similar but there are few key differences, in particual the goal in Supervised Learning is to predict the right label while in Reinforcement Learning the goal is to find an action $x^*$ in other to maximized the reward. For this reason instead of minimizing the error between the predicted class and the real class, the goal is to choose an action that maximize the cumulative reward in a given state.

## 2.3 Application of Reinforcement Learning

Some possible application of RL:

- Self-driving cars. Train a car in a real world can be dangerous (people can be hurt) and expensive (car can hit an object and damage the vehicle). They can be trained in a RL setting: the car is the agent, the world around it the enviroment, it has to take some actions (throttle, steer, etc.), the state are the sensors and the goal is to reach a destination avoiding obstacles. In this setting the agent can get positive or negative rewards depending on the action chosen.

- Games. Popular example are: chess, Go and Dota. In all these games computers were able to beat the champions.

- Finance. Many problems can be formulated as a RL problem, for example the stock trading. Here the action could be: selling, buying, holding and the goal is to maximize the cumulative return over a period of time.
- many other applications[8].

## 2.4 Formal definition

The Reinforcement Learning can be formulated as a Markov Decision Process (MDP), indeed a MDP express the problem of sequential decision-making, where for each state $s$ the dicision maker can choose any action $a$ available in that state $s$. The process respond moving with some probability to the state $s'$ and giving the decision maker a reward $R_a(s, s')$ (read as: 'the reward when in state $s$ the action $a$ is chosen')
The MDP is defined as a tuple of 4 elements (S, A, P, R), where:

- S is a set of states, called the *state space.*
- A is a set of actions, called the *action space.*
- P is the probability from state $s$, at time $t$, of reaching state $s'$, at time $t + 1$ with action $a$:

$$P_a(s, s') = Pr(s_{t+1} = s' | s_t = s, a_t = t)$$

- $R_a(s, s')$ is the immediate reward received after transitioning from state $s$ s to state $s'$, due to action $a$.

The state and action spaces may be finite or infinite.
The MDP is controlled by a sequence of descrete time steps that create a trajectory

$$s_0 \xrightarrow{a_0} s_1 \xrightarrow{a_1} s_2 \xrightarrow{a_2} s_3 \xrightarrow{a_3} \ldots$$

where the states follow the state transition $P_a(s, s')$. The transition function and thereward function are determined only by the current state, and not from the previous states. This property is called Markov property, which is a characterize the MDP and it means that the process is memory-less and that the future states depends only on the current one and not on its history.

The goal of the MDP is to find a good policy for the dicision maker: a function $\pi$ that specifies the action $\pi(s)$ that will be chosen when in state $s$. The policy $\pi$ found will maximize the cumulative reward over a trajectory $\upsilon$:

$$G(\upsilon) = \sum_{t=0}^{\infty} R_{a_t}(s_t, s_{t+1})$$

This return value has the problem that all the rewards contribute in the same weight and this can create some problems for the lack of information. A better return would be to give some more importance to the short-term memories and giving less importance to the one far in the future. This is solved introducing a **discount factor**, denoted with $\gamma$. Then the correct formula is:

$$G(v) = \sum_{t=0}^{\infty} \gamma^t R_{a_t}(s_t, s_{t+1})$$

with value of $\gamma$ satisfying $0 \leq \gamma \leq 1$. A lower discount factor motivates the decision maker to take actions that a close in time instead of actions that a far in the future.

Another important notion in MDP and Reinforcement Learning is the value fuction. While the return $G(v)$ gives the reward over a trajectory it does not tell much about how good are the single states. The value fuction does exactly this, it estimates how good is for the decision maker to be in a given state. The notion of "how good" is defined in terms of future rewards that can be expected in terms of expected return.
The value fuction $V_\pi(s)$ can be formally defined as:

$$V_\pi(s) = \mathbb{E}_\pi(R_t|s_t = s) = \mathbb{E}(\sum_{t=0}^{\infty} \gamma^t R_{a_t}(s_t, s_{t+1})|s_t = s)$$

The expected return when starting at state $s$ and following policy $\pi$.

Similary, another notion: the action-value fuction, the expected return from state $s$ with an initial action $a$:

$$Q_\pi(s, a) = \mathbb{E}_\pi(R_t|s_t = s, a_t = a) = \mathbb{E}(\sum_{t=0}^{\infty} \gamma^t R_{a_t}(s_t, s_{t+1})|s_t = s, a_t = a)$$

The value fuction and the action-value fuction are called V-fuction and Q-fuction. They are related and satisfy a particular relationship, used in many Reinforcement Learning contests, that for any policy $\pi$ and state $s$, the following condition holds:

$$V_\pi(s) = \mathbb{E}_\pi(Q_\pi(s, a))$$

The V-fuction can be decomposed in 2 terms:

$$V_\pi(s) = \mathbb{E}_\pi(R_t|s_t = s) = \underbrace{\mathbb{E}_\pi(R_{t+1}|s_t = s)}_{\text{immidiate reward}} + \underbrace{\mathbb{E}_\pi(\gamma V_\pi(s_{t+1}|s_t = s))}_{\text{discounted value of next state}}$$

This is the Bellman Equation that defines the value fuction recursively, enabling the estimations of the next states. Similary it is possible to write Bellman equation for the Q-fuction :

$$Q_\pi(s, a) = \mathbb{E}_\pi(R_t|s_t = s, a_t = a) = \mathbb{E}_\pi[R_t + \gamma Q_\pi(s_{t+1}, a_{t+1})|s_t = s, a_t = a]$$

In this way the V-fuction nad Q-fuction are updated with the values of the successive states without the need to know the trajectory to the end

## 2.5 Q-Learning

This part concerns the st

### 2.5.1 Dataset

m movements

# 3 Conclusion

This internship was a formative experience that allowed me to discover many new topics but also to understand how a company works. Moreover, it allowed me to understand that there is a substantial difference between the university environment and the corporate environment: university, as mentioned by some professors during the courses, has the task of giving a general knowledge of the topics and the real task is to give a way of thinking and analysing a problem to find a solution; indeed, during this period in the company I realized that computer knowledge was not always essential but it was more important to understand what was needed to be done and how to do it.

I believe that this internship was very useful because it allowed me to deepen and learn about new topics and gave me an idea to understand if the tasks assigned could be addressed in a future university career or in a possible job. This is also an experience that allows to grow both personally and professionally, not to mention that it is a way to enrich the Curriculum Vitae.

I am fully satisfied with the internship as AI researcher and developer and I am satisfied with the atmosphere of serenity and professionalism that I lived in the company thanks to the kind and extremely helpful colleagues.

# References


[1] Todd Litman. 2021. *Autonomous Vehicle Implementation Predictions.*
https://www.vtpi.org/avip.pdf

[2] SAE. 2016. *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles.*
https://www.sae.org/standards/content/j3016_201806/

[3] Monika Stoma, Agnieszka Dudziak, Jacek Caban and Paweł Drozdzie. 2021. *The Future of Autonomous Vehicles in the Opinion of Automotive Market Users.*
https://www.mdpi.com/1996-1073/14/16/4777/pdf

[4] Piotr CZECH, Katarzyna TUROŃ, Jacek BARCIK. 2018. *AUTONOMOUS VEHICLES: BASIC ISSUES.*
https://doi.org/10.20858/sjsutst.2018.100.2

[5] Paul Barter. 2013. *"Cars are parked 95% of the time". Let's check!.*
https://www.reinventingparking.org/2013/02/cars-are-parked-95-of-time-lets-check.html

[6] Jeremy A. Carp. 2018. *AUTONOMOUS VEHICLES: PROBLEMS AND PRINCIPLES FOR FUTURE REGULATION.*
https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=1048&context=jlpa

[7] ANDREW G. BARTO and THOMAS G. DIETTERICH. *Reinforcement Learning and its Relationship to Supervised Learning.*
http://web.engr.orst.edu/~tgd/publications/Barto-Dietterich-03.pdf

[8] Yuxi Li. 2019. *REINFORCEMENT LEARNING APPLICATIONS.*
https://arxiv.org/pdf/1908.06973.pdf