Robotic Navigation in Simulated Urban Environments

An investigation on the effect of uncertainty in the observed environment

Supervisor: Ed Manley

## Abstract

Table of Contents

[Abstract 1](#_Toc15378544)

[1. Setting the stage 3](#_Toc15378545)

[1.1 Topic and aim 3](#_Toc15378546)

[1.2 Introduction 3](#_Toc15378547)

[1.3 Literature review 3](#_Toc15378548)

[2. Prerequisites 4](#_Toc15378549)

[2.1 Notation 4](#_Toc15378550)

[2.2 Delivery Robots today 4](#_Toc15378551)

[2.3 Unity – as a simulation engine for research in DRL 5](#_Toc15378552)

[2.4 The environment 6](#_Toc15378553)

[2.4.1 Environment 8](#_Toc15378554)

[2.4.1.1 Walls & Obstacles 9](#_Toc15378555)

[2.4.1.2 Pedestrians 9](#_Toc15378556)

[2.4.1.3 Crowded areas 11](#_Toc15378557)

[2.4.2 Academy 14](#_Toc15378558)

[2.4.3 Agent 16](#_Toc15378559)

[2.4.4 Target 19](#_Toc15378560)

[2.4.5 Tuning of hyperparameters 20](#_Toc15378561)

[Training 21](#_Toc15378562)

[Tensorboard 21](#_Toc15378563)

[Q-learning 21](#_Toc15378564)

[Trust Region Policy Optimisation 21](#_Toc15378565)

[Proximal Policy Optimisation 21](#_Toc15378566)

[Curriculum Learning 21](#_Toc15378567)

[Exploration/Exploitation trade-off 21](#_Toc15378568)

[Insights: 25](#_Toc15378569)

[Stuff for the discussion: 25](#_Toc15378570)

[Thoughts: 25](#_Toc15378571)

[Policy Evaluation 25](#_Toc15378572)

# Setting the stage

## 1.1 Topic and aim

The topic for this dissertation is robotic navigation in simulated urban environments, with the reason being threefold. Firstly, urban environments contain a lot of different interesting challenges, such as static and dynamic obstacles, and the effect of interaction between, along with the relevance and certainty of their position, to mention a few. The mentioned challenges are to be elaborated below, yet these challenges are important for the growing number of *smart cities.* Secondly, movement in urban environments is surrounded by ethical principles, irrespectively of you being a human or a robot. It is a topic to be discussed further below, however, simulated environments present the ability to address ethical concerns in a save manner. *Unity*, introduced doing the programme of Spatial Data Science and Visualisation, provides the software engine to construct such simulated environments. Thirdly, robotic navigation can be done using a variety of methods, of which one is *reinforcement learning* (RL). RL has gained a lot of attention in recent years, mainly because of advances in deep learning, and RL is an active research area within the faculty of CASA, UCL*. Deep reinforcement learning* (DRL), combing the recent advances in deep learning with RL, has enabled super-human performance in certain tasks, yet DRL isn’t the answer for every challenge.

The aim of this dissertation is to explore how the state-of-the-art DRL method for continuous control task, *Proximal Policy Optimisation*, manages some of the challenges present in urban environments. In particular, how uncertainty about the observed environment affects the performance. The simulated environment is a toy model of a real-life urban environment, yet it is constructed with realistic physical settings, possible by using Unity ML-Agents Toolkit by *Unity*, a simulation engine based on state-of-the-art game developing software.

## 1.2 Introduction

## 1.3 Literature review

# Prerequisites

## 2.1 Notation

Words used interchangeably: Sensors/Pedestrians, Robot/Agent

## 2.2 Delivery Robots today

Delivery robots are not a thing of the future, they are already deployed in a few pilot cities, even in London (Nichols, 2019). An increasing number of companies are putting they attention on autonomous delivery robots, to meet increasing customer expectations of companies to ride the technology weaves, to enable low-cost-low-emission products. 2019 has so far been an exciting year in this matter, with three of the biggest players (Merrit, 2019), within the field of autonomous delivery robots (ADR), launching different initiatives taking effect doing 2019.  
Starship, founded by two of the Skype co-founders, newest launch is autonomous delivery of food and beverages at George Mason University, Maryland. The partnership is to accommodate the rising need for smart solutions in a high-paced-high-expectation environment, where nutrition sometimes is overlooked (Nichols, 2019). Starship fleet of ADR’s has over 100.000 logged kilometres and more than 25.000 deliveries under the wheels, and been deployed in cities like London, New York and Washington, DC (Merrit, 2019; Nichols, 2019). Starship’s fleet has gained enough experience to surpass the need for any handholding (Nichols, 2019), which entitles them as the leader in the ADR race.

Another two contenders are two well-known giants, Amazon and FedEx, which both revealed their ADR in the first quarter of 2019 aiming at lunching in pilot cities around this time now (FedEx, 2019; Scott 2019).

All three contributions are equally interesting for this project because they appear, as little information is revealed at this stage in the race, to have same specifications (FedEx, 2019; Starship, 2019; Scott, 2019). Them having similar specifications provides guidance on the specifications of the agent, central for the simulations in Unity. For now, it is noted that the ADR’s appears to have a height around 0.5-1.5 metre, a width and depth of 0.5 metre, a total weight (including cargo) of 45-50 kgs and a speed around 5 km/h (FedEx, 2019; Starship, 2019; Scott, 2019).

## 2.3 Unity – as a simulation engine for research in DRL

Unity is best described as being a multi-functional platform, enabling development of everything from complex high resolution multiplayer games to less complex mobile games over to VR/AR applications, and increasingly as a challenging set-up to conduct research on artificial intelligence (AI) (Juliani et al, 2018; OpenAI, 2018; Sadeghi & Levine; 2016).

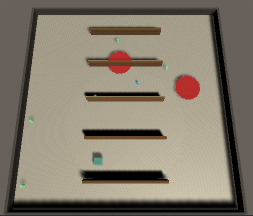
With increasing complexity, sometimes in an exponential manner, comes a need for low latency and distributed computing. Irrespectively of the increasing complexity being a derivative of a high-resolution 3D game consisting of, potentially, thousands of players with widespread interaction patterns, or research on algorithms that roughly needs 18 million video frames to surpass human performance (Hessel et al, 2017), satisfies Unity the need (Juliani, 2019). The underlying engine runs in de-synchronized fashion, supporting simulations at run times at least 100 times faster than real time. This is possible while still maintain physics and frame rendering. A final appealing feature of the computational side is the ability to run concurrent training session, internally in one application as well as externally, enabling the possibility to utilise cloud computing for further increased computational power.

The diversity in the application areas implies that Unity is a serious candidate for modelling the complex dynamics of urban environments. The most appealing feature is the ability to replicate real-life physical complexity, and thereby enabling realistic movement patterns as well as interaction between objects. This is an important factor with an underlying interest in enabling the results to generalise to the real-world, as higher similarity between the environment within the results are obtained and the environment in which the results are deployed, increases the likelihood for generalisation. Another feature of Unity, which makes it appealing for modelling urban dynamics, is the possibility to model complex social interaction. Unity comes with two options to add layers of social interaction to the environment. The *NavMesh* class provides the ability to add AI agents to the environment, useful for spatial queries, as pathfinding. As described in detail later, this class is utilised to model pedestrian behaviour in the constructed urban environment for this paper. NavMesh agents can interact with other NavMesh agents, as well as avoid other moving obstacles, enabling a layer of social interaction with relative ease.   
A more challenging, and perhaps are more interesting way to add social interaction to the environment is by utilising the ML-Agents toolkit, a central part of this paper. The ML-Agents toolkit bridges sophisticated machine learning methods with the graphical interface and complex physical engine of the traditional Unity application, enabling a new setting to push the boundaries for DRL research (Juliani, 2018). The toolkit allows researcher to utilise pre-defined algorithms, based on TensorFlow, or define them themselves, via a Python API. In the light of the NavMesh class, the toolkit puts the control of the interaction in the hands of the researcher, which carries a certain responsibility. It requires the researcher to exhibit a greater understanding of the task and modelling at hand, and so reduces the possibility of headless simulation – limiting the risk of another black box appearance.  
With the introduction of ML-agents, can Unity partly be regarded as the new kid in the class of software’s usable for Agent-Based modelling (ABM). The ML-Agents toolkit makes Unity an appealing contender to the traditional software’s used for ABM, by the fact that the scripting languages are C# and Python. The use of C# and Python implies low latency along with a wide variety of options for further data processing, through open source libraries, and user support, from the enormous communities surrounding the two languages.

Consider the ending of this section, if it should be changed/extended a bit.

### 2.4 The environment

The environment explored in this paper is seen in figure 1, and it is intended to simulate areas of a city which are less trafficated with cars, and more dominated by pedestrians, formally known as pedestrian streets. An example hereof could be the area around Carnaby Street in Soho, London. Simulating a network of pedestrian streets is justified by the current state at which ADR’s are today, i.e. still being in an early stage and facing challenges in the unpredictability of the real world (Nichols, 2019).

*Figure 1 – The environment*  
  
*The environment contains five moving pedestrians (light green), two high-density areas (red), the target (dark green square) and the five static obstacles.*

Developing an environment, formally known as a *Scene* in Unity, to facilitate the possibilities within the ML-agents toolkit requires some basic objects; an actual *environment* to explore, an *academy*, an *agent* and a *target*. With the presence of these objects one gets something similar to the environment seen in figure 2, which on top includes walls to prevent the agent from falling of the surface – yet these are not strictly needed, as to be elaborated on shortly when digging into the set-up of the *agent* component. Figure 2 is an example of how everything could be organised within a scene.

Every application build with Unity is made of *Scenes, GameObject’s, components* and, often at least, *scripts*. An application can contain an arbitrary number of scenes, and the application shown in figure 2 contains one, namely *SensorEnvironment-4.2.* Every scene contains GameObject’s, in which components and scripts are attached, to sustain any form of behaviour imaginable. Components enables the use of all the built-in functionalities in the Unity Engine, and scripts provides the researcher with the option to take full control.

*Figure 2 – A scene in Unity containing the components of the ML-Agents toolkit*A screen shot of a computer

Description automatically generated  
*The scene contains all available elements within the environment.*

One thing to note from figure 2 is that all necessary elements are contained in a *prefab* named *Area\_EnvX*, where X referrers to the current version of the implementation[[1]](#footnote-1). Prefabs are user defined GameObject’s, and the use of prefabs are a neat way of altering similar objects simultaneously. In relation to figure 2, containing all necessary elements in a prefab, is a way to utilise a parallelised set-up, allowing for faster training but more on that later.

#### 2.4.1 Environment

The walkable area is labelled *ground* in figure 2, and it serves two important purposes. Firstly, it defines the extend of the area through its scale, seen in figure 3. The extend of the area is used to ensure that random placing, through scripting, of objects happens within the bounds of the traceable area.   
Secondly, it serves as a container for the objects belonging to this training area[[2]](#footnote-2). Initialising new GameObject’s as children of another GameObject is a way to ensure intended interaction with relevant GameObject’s. It allows the researcher to write generic scripts and not instances specific scripts, which are in general good practice, and especially desirable when working with parallelised set-ups.

*Figure 3 – The ground object  
A screenshot of a cell phone

Description automatically generated*

***Tags***

Every GameObject within the training area is tagged, as seen in the right side of figure 3. Tags is an elegant way to differentiate GameObject’s from each other, especially useful in association with collision detection, collecting observations on the state of the environment and random placing of GameObject’s.

***Layers***

Assigning different GameObject’s to different layers is used to either include or exclude certain GameObject’s from some sort of detection. This is useful in the two-brain set-up, ensuring that one brain handles avoidance of dynamic obstacles and one brain takes care of the general navigation towards the target.

***Static Objects***

To the right of the GameObject’s name is the ability to mark a GameObject as static, which is used in connection with NavMesh agents. Static GameObject’s are part of mesh in which a NavMesh agent can navigate, but this will be described in a bit more detail below.

***Geometry of a GameObject***

Any object having a shape contains a Mesh filter, defining the geometry of the object, and a Mesh renderer, which ensures rendering of the object at the position specified in the transform component. Figure 3 shows that the *ground* element is a plane, having a size of 80x80x1, positioned at (0,0). The height (size in the y direction) of the *ground* element is not as such important, if it is above 0, to sustain the plane rendering.

Within the bounded *ground,* not necessarily as child objects, is six types of objects placed, two of them elaborated in individual sections below, and the other four are *walls, obstacles, pedestrians* and *crowded areas.*

##### 2.4.1.1 Walls & Obstacles

The walls as well as obstacles are, as GameObject’s, identical to the *ground* GameObject, with a different tag along with the obviously different size and position. The walls and obstacles have different materials, to indicate that they represent something different. The walls can be thought of as boundaries for a certain pilot area in a given city, and the obstacles represents buildings. Both walls and obstacles serve as a resetting mechanism upon collision with the agent.

##### 2.4.1.2 Pedestrians

The pedestrian GameObject is a prefab, which is attached to the academy from where it is initialised. The initialisation will be described in more detail in the academy section. Figure 4 shows the components attached to the pedestrian prefab, of which four are interesting to elaborate on. The first two below the transform component serve to define the shape, and as such not interesting.   
The third component is the *collider*, which is the component, together with the *rigidbody* component, that allows for collision detection between the object and another object, with a collider and rigidbody component attached as well. The settings in the collider is not as such interesting, as they are standard settings matching the scale of the GameObject.  
The fourth component is the rigidbody, which is the component that enables the physics engine to talk control of the movement of the GameObject. The *mass* of the GameObject is specified in kilograms and is set equal to a reasonable average value for a male.

*Figure 4 – The pedestrian prefab*  
A screenshot of a cell phone

Description automatically generated

The *drag* is a force working in the opposite direction to the movement of the object, specifying at what pace the movement of the object is decreased. The value for the drag of the pedestrian prefab is calculated using (1), to ensure realistic behaviour in the simulation, because the default value is zero which is not in accordance with realistic behaviour.

Where is density of the fluid that the object passes through, air in this case here, is the speed (m/s) at which the object moves, is the drag coefficient (unit less) and A is the cross sectional area related to the movement, which is the area of the object normal to the direction of the movement. Table 1 shows the values used, and the calculated drag.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table 1. Drag** |  | Value | Unit |
| Density (air), rho: |  | 1.225 | Kg/m^3 |
| Speed, v: |  | 1 | m/s |
| Drag Coefficient, Cd\*: |  | 1.3 | Unitless |
| Cross sectional area, A: |  | 3.141593 | m^2 |
| *Average of human body in upright position, and at the same time the coefficient of a short cylinder.* | | |  |
|  |
| Drag: |  | 2.501493 |  |

The angular drag is how much a rotation is slowed down, and it is kept at standard value, because the default value is within a realistic order.

The rigidbody is marked as being kinematic, which implies that the object isn’t influenced by any forces. Why have the rigidbody attached then? Because it ensures better collision properties having both the collider and the rigidbody attached to an object, and all movement is handled by the NavMesh agent component below.

The fifth component is the Nav Mesh Agent component, which is what turns an empty GameObject into a Nav Mesh agent. At least initially, is every parameter herein kept default, except speed and angular speed. The speed of the agent is set equal to 1 m/s, the same as the ML agent, which is roughly equal to a speed of 3.5 km/t – the average relaxed walking speed.  
The angular speed is set equal to 150, a bit above the default value, as it is equal to the same angular speed used in the ML agent.

The final component is a script, written to ensure that the pedestrian walks around the training area continuously. The script has two public variables, the radius and frequency. The radius is the distance from the agent, that a new target point is draw within. One target point is drawn within the frequency specified, and there should therefore strike a balance between the radius and the time, such that the agent has time to travel that distance within the specified time.

##### 2.4.1.3 Crowded areas

The crowded areas are a prefab as well and is initialised from the academy object. The crowded area object represents an area with a lot of potential challenging obstacles, which could be a public square or an area on a pedestrian street, where some entertainment unfolds etc. It is an area where the agent can pass through, but it will require a significant amount of additional effort, which implies that it potentially is beneficial for the agent to avoid interaction. How to learn the agent that, is a real challenge as will be clear later, and different things are tried to learn the agent. For now, just focus on the fact that the crowded area can attain different densities, and later it will be specified how the densities are used to shape the reward function, which ultimately affects the learning process. The crowded are is represented as a circle, which as a certain *density*, see figure 5.

*Figure 5 – The crowded area prefab*A screenshot of a cell phone

Description automatically generated

The radius is set through the academy and the density is drawn from three possible values, and it changes every time a new episode is started, along with the location of the crowded area.

The density values that a crowded area can possess are based on empirical population density estimates for 2019 from London, see table 3 and figure 6 & 7.

|  |  |
| --- | --- |
| **Table 3 - Population densities, London.** | |
| Max Ward: | 0.0289 |
| Max Boroughs: | 0.0164 |
| Overall: | 0.0058 |

*Figure 6 – Population density, Boroughs, London  
A close up of a piece of paper

Description automatically generated*

*"Contains National Statistics data © Crown copyright and database right [2015]" and "Contains Ordnance Survey data © Crown copyright and database right [2015]"*

Figure 6 and 7 show the population density in London borough and ward level respectively. It is no surprise that the densest areas are close to the city centre, and the densest ward is located within Westminster borough, namely Church Street.

*Figure 7 – Population density, Wards, London   
A close up of a map

Description automatically generated  
"Contains National Statistics data © Crown copyright and database right [2015]" and "Contains Ordnance Survey data © Crown copyright and database right [2015]"*

#### 2.4.2 Academy

So far has none of the outlined parts of the environment been specific to the ML-Agents toolkit, yet that changes at this point. The academy is one of the three cornerstones of the toolkit, and it serves to bridge actions and observations from the actual environment to the TensorFlow-based models in Python. It is actual a GameObject referred to as the *brain* that serves as the bridge, but one cannot have a brain without an academy. The content of the academy object is seen in figure 8.

The brain is what makes determines the action, based on the observed observations in the current state, and the academy facilitates the training of the brain. There are two types of brains, learning brains and player trains. Learning brains learns the policy based on the neural network, implemented in TensorFlow and is to be elaborated on in the following section. The player brains allow the researcher to test before invoking the learning brain, by giving the researcher the option to control the agent with keys on the keypad.   
The brains can take in *vector observations* as well as *visual observations*, and it outputs an action vector. The size of both the observation vectors and the action vector can be specified by the researcher, and it is highly problem specific.  
The learning brain used this case takes in vector observations, as visual observations requires more computational power than available for this paper. The size of the vector is 126 and the size is determined by specifications in the agent. Describe partial/full observability.  
When the control checkbox is ticked, the agent learns otherwise is uses a pre-trained model.

*Figure 8 – The academy*  
A screenshot of a cell phone

Description automatically generated

It is currently possible to add more brains to a single agent; however, it is not possible to train multiple brains on a single agent – yet. Training multiple brains on an agent would present some interesting possibilities, which will be discussed at the end of this paper.

The academy has, in this case here, only a single component attached. The main purpose, besides acting as a container for the brain, of the academy is to initialise and alter the environment. This is revealed by taking a look at the code underlying the academy script, which except for some helper functions contains just two methods; InitialiseAcademy() and AcademyReset().  
AcademyReset(), for this paper, is only used if *curriculum learning* is utilised, and the purpose is to add increased complexity to the environment at specific times. More on that in the coming section. One could alternatively change location of certain objects from the academy and thereby either using the option to specify a maximum number of steps call *Done* to reset, but that is not how it is done in this paper. Changing location of certain objects is done through the agent script, which is described shortly.

Before going over the configuration specifications, it is natural to touch the reset parameters as they are related to the above paragraph. The reset parameters are the variables that changes when the environment is being reset, and they are a major part of the agent script, to be described. In this paper here, they take input from the specified curriculum.

Below the reset parameters are the two prefabs, *pedestrian* and *crowded area*, attached for initiation of the environment.

The final thing in the academy script is the configuration parameters. The width and the height determine the size of the application window when training is done outside the editor, which implies the need for creating a build first – Training is described next.

The quality level is the quality of the camera input, if visual observations are provided to the brain, and is so not relevant for this paper.

The time scale is speed at which the simulation is carried out – 1 is real time and 100 is 100 times faster than real time. The actual level of time scale should not as such affect the performance of the training, only the training time, yet it can because some physics calculations gets inaccurate with a time scale of 100. This should only be relevant if one has objects that travels at high speed, which isn’t the case in this paper.

The final configuration parameter is the targeted frame rate, which is the rate at which Unity aim as rendering the frames, which shouldn’t be altered unless one is using visual inputs.

#### 2.4.3 Agent

The agent object is by far the most complex, in terms of the number of components and methods contained in the attached script, object of all the object revisited in this section. The content of the agent object is seen figure 8.

Other than the mesh components as well as the rigidbody and collider component, which are previously described, contains the agent a ray perception component along with two custom scripts; one to draw trails and one to hold the necessary methods needed for any agent, to leverage the ML-Agents toolkit.

The ray perception component enables the agent to cast rays in direction and length specified by the researcher, which are used to observe the environment in the specified direction. Ray casting is used to collect observations for the agent, about the state of the environment.

Any ML agent needs an agent script, to contain the needed methods, just as the academy needed an academy script. An agent script comes with some default variables and options, which are listed above the grey line in figure 8.

*Figure 8 – The content of the agent object  
A screenshot of a cell phone

Description automatically generated*

***Brain***

An agent needs a brain to control the movement of the agent, and it is the same brain as specified in the academy, at least doing training. Certain methods are needed to control the movements, more specifically; CollectObservations, AgentAction and MoveAgent. To have a closer look at all methods, the interested reader should open the agent script using ones preferred text editor.

The CollectObservations method entitles the agent with vision, and here the agent has 180 degrees sensor vision, in steps of 10 degrees, spanning in front and with a length of 50 meters. These observations are what makes up the vector observations, mentioned in the description of the brains. The agent is provided with five tags to recognise, and so it casts a ray for each of the tags to recognise, for each of the degrees specified. Furthermore, it keeps track of distance to the objects, and if an object has been missed. The implication is that the resulting observation vector has the dimension , which for the specific case here means that the observation vector is .

The agent chooses an action, based on the observations about the current state of the environment. The action/-’s is chosen by the brain, and facilitated to the agent through the AgentAction method, in which the action signal is translated to actual movement via the MoveAgent method. One example is, that the agent is moved forward with a constant speed, and the action of the agent is degrees to rotate, to sustain the desired navigation around the environment. This is how it is done in this paper, and the level of the constant speed is subject to change by the researcher via the **speed** variable, below the grey line.

***Camera/Render textures***

The agent script contains the option to specify camera/-s and/or render textures, if the brain attached to the agent uses visual observations.

***Max step***

As with the academy, the option to specify a maximum number of steps is present. The agent will be reset when the number of steps surpasses the specified number, which is useful to break unfavourable movement patterns, as is shown later.

***Reset on Done***

Another way to reset the agent is to call *Done*at some point, which usually is after colliding with an object in the environment.

Both *max step* and *reset on Done* is used in this paper. Collisions are detected using the OnCollisionEnter method, and if the agent collides with either a wall, a static obstacle or a pedestrian, the agent is rewarded with a large negative reward (-1) and Done is called, to reset the agent, and start a new episode. If the agent collides with the target, a large positive reward (1) is rewarded and *Done* is called as well.

On reset, not only the position of the agent resets, but the position of the target and the sensor clouds, along with the associated density, changes as well. Therefore, is the option to specify the levels and number of possible **densities** available under the agent script.

***Decisions***

Decisions can be done either at a specific interval or on demand, and this paper here uses decision at a specific interval. Choosing the actual decision frequency is done in the exploration/exploitation section.

***Draw trails***

The agent script contains the option to enable drawing, and specify the number of trails drawn, which serves to visualise changes in learning patterns doing training. A custom implementation is used, because the default implementation *trail renderer* does not consider the resetting of the agent, and so draws the jump from where one episode ends to the start position of the agent, which minimises the information provided by drawing the trails.

***Name of file***

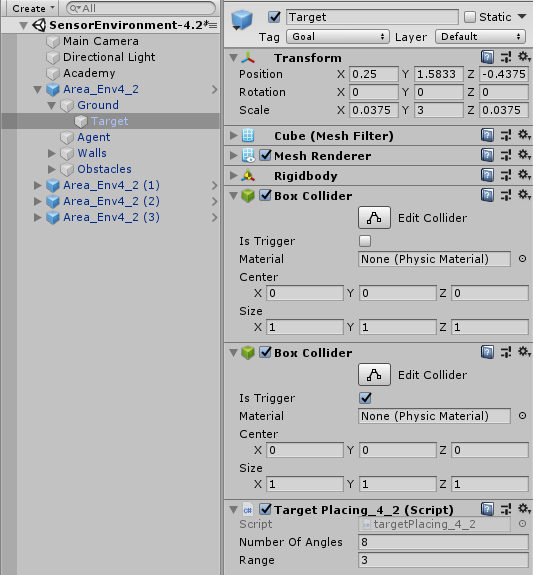
If desired, by checking ***verbose***, the agent collects additional information, compared to the information provided via TensorBoard and writes it to specified files. The additional information provides deeper insights into the progress of the training and highlights potential shortcomings. The additional information is information on the number of collisions with pedestrians and crowded areas, the steps used to locate the goal and the steps taken in the crowded area. However, this can be changed to suite the environment and the need of the researcher.

***Second brain***

The agent script is provided with an option to add a second brain, and potentially many more, which is useful to investigate the effect of separating tasks on individual brains. This is investigated at the later point.

#### 2.4.4 Target

The target is an important element of the environment, as it is the sparse positive reward signal, which the agent searches for. In this paper, the target is a cube randomly positioned in the environment. As irrelevant as the geometry of the target is, as relevant is the size, relative to the environment. The relative size of the target affects how easily the target is located and can be thought of as the difference in locating a building on a street compared to a specific brick on a specific building, at that same street. The size of the target is fixed here, at a relative size of 3.75%, chosen arbitrary and the effect of the relative size of the target could be a topic for future investigation. 3.75% is equivalent to size of 3 meters.   
The content of the target is seen in figure 9.

*Figure 9 – the content of the target object  
*

The target contains, besides the custom target script, only familiar components. The target script has two box colliders .. Check if that is even necessary

The custom target script serves two purposes; randomly setting a new position of the target and check for collision with static objects in the environment. The public parameters of the script are input to the method used to check for collisions. By default, 4 rays are casted with a length of 3, covering all sides.

The target is not allowed to be located within another static object in the environment, to prevent conflicting collisions. Having the target located separately from other static objects, simulates the idea about locating a position outside a building for delivery by the ADR.

### 2.4.5 Tuning of hyperparameters

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 2. Environment configuration (200.000 steps)** | | | | | | | | |
| Speed | Decision Interval | Time Scale | Training time | Deviation\* | ACR: Mean\*\* | ACR: Std. dev. | AEL: Mean | AEL: Std. Dev. |
| 2 | 5 | 60 | 1650 | 22.2% | 0.663 | 0.65 | 40.32 | 19.41 |
| 1 | 5 | 60 | 1500 | 11.1% | 0.920 | 0.59 | 39.55 | 15.59 |
| 1 | 3 | 60 | 1350 | 0.0% | 0.916 | 0.80 | 101.31 | 198.75 |
| 1 | 7 | 60 | 1750 | 29.6% | 0.896 | 0.63 | 34.07 | 15.86 |
| 1 | 10 | 60 | 1800 | 33.3% | 0.812  **Re-create in Tex** | 0.56 | 26.22 | 9.32 |
| 1 | 5 | 100 | 1350 | 0.0% | 0.908 | 0.62 | 43.03 | 19.58 |
| 1 | 5 | 20 | 2020 | 49.6% | 0.908 | 0.57 | 44.02 | 21.15 |
| \*: Relative to the fastest, \*\*: Mean of converged path | | | | |  |  |  |  |

### Training

hej

### Tensorboard

Inspired by (Juliani, 2018b)

## Q-learning

## Trust Region Policy Optimisation

## Proximal Policy Optimisation

## Curriculum Learning

Specifying curriculums in Unity is done in a json file, seen in figure 3, and contains five different parameters

# Exploration/Exploitation trade-off

Coming of configuring the set-up, it is time to explore the environment, address the challenges and hopefully solve the environment.  
When is the environment *solved*? The environment is solved when the agent can cope with the different type of challenges that the environment processes. As mentioned in section X, the environment of this paper processes four main challenges; difficult areas, static obstacles, crowded areas and dynamic obstacles.

The first relevant question to address is, *what is the look of learning*?   
Learning comes with many looks, and a wide range of relevant statistics from training is visualised via TensorBoard. For a description of how each statistic expresses whether learning has occurred, see section X. The most intuitive statistics is the average cumulative reward (ACR) and the average length of an episode (ALE), examples are seen from figure 4.  
The ACR is expected to converge to the maximum obtainable level of reward within a single episode, when learning has taken place and the agent has solved the environment. The ALE should converge to the minimum number of steps needed to obtain the reward. The speed of which the converge takes place is a direct consequence of the complexity of the environment and the degree of assistance provided to the agent, as will be clear doing this section.

*Figure 3 – The simplest version of the environment.*

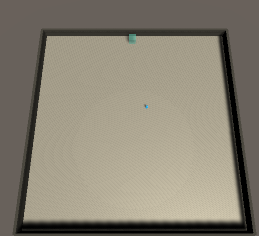
   
*The simple environment contains only a target (darker green square) and the agent (light blue).*

Figure 3 shows the simplest possible environment, namely without any obstacles. The target is here in the upper middle of the environment, yet it changes position every episode. Learning in this environment is straight forward, and this is seen from figure 4. Figure 4 shows how ACR and ALE converges rapidly to their optimal levels, which are obtained after around 75.000 steps in this case.

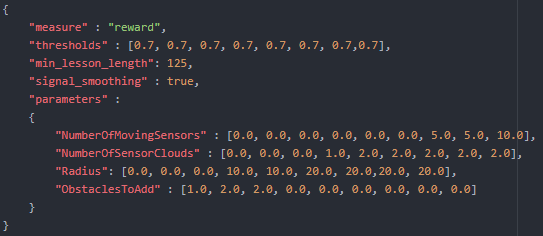
*Figure 4 – ACR and ALE for the environment from fig.3*  
A screenshot of a cell phone

Description automatically generated *The average cumulative reward and average episode length obtained from learning in the simple environment from figure 3.*

The shapes of ACR and ALE seen in figure 4 are ideals, and far from trivial to obtain with the slightest degree of complexity present. The road to meaningful learning requires careful design of the aid provided to the agent to ensure generalisation and so providing the agent with the ability to handle unseen environments, which is one of the aims of this paper. Figure 5 shows the same output, as in figure 4, for the far more complex environment seen in figure 1.

*Figure 5 – ACR and ALE for full environment from fig. 1.*

The first step towards improving learning is to introduce a curriculum, for the agent to learn from, to gradually add complexity to the environment, as the agent learn to cope with novel challenges. An interesting aspect is how the curriculum should be designed, in order to facilitate optimal learning of the agent, and this is almost an area for research itself. The aim of this paper is not to shed light on optimal construction of curriculums, and the curriculum used here is chosen such that it addresses three of the four main challenges equally, namely the static obstacles, the crowded areas and the dynamic obstacles. The difficult areas a separate task, which will be addressed shortly.  
The initial curriculum used in this paper is seen from figure 6, and the reader should revisit section X.Y for a full discussion of each of the parameters.

*Figure 6 – The initial curriculum used to facilitate improved navigation of the robot.*

Starting out with the parameters specified in the curriculum, it is noticed that there is at least one parameter for each of the obstacles for the robot to avoid. The curriculum is designed such that the obstacles are introduced relative to their degree of complexity. The first challenge is the static obstacles, then the semi-static crowded areas and finally the dynamic pedestrians. Each type of obstacle is introduced over three lessons, and the same threshold to be passed is required. The threshold is at this point not fixed, as it should be set based on the two parameters of the environment, which will be addressed below. Signal smoothing and the minimum number of episodes, in which the threshold is surpassed, are here chosen with the same objective, so sustain generalisation. Signal smoothing is enabled to ensure robustness of the signal. The minimum number of episodes are chosen such that the number ensures robustness, yet not to such a degree that it becomes a burden. The minimum number of episodes should be chosen with the threshold in mind. Too high a threshold compared with too high a requirement on the robustness of the signal, will likely prevent any progress, depending on the complexity of the environment. Choosing too low a threshold and too low a level of the robustness, can imply that progress occurs too soon. A common misconception when working with reinforcement learning is the amount of data needed to sustain meaningful learning, the planning fallacy as of (Irpan, 2018). Figure 7 shows ACR and ALE, along with the changes in lessons[[3]](#footnote-3), for a training session with the initial curriculum.

Figure 7 - *ACR and ALE for first training session using curriculum learning*

As discussed in section X, is the ability of the agent to generalise to unseen environments one of the

Figure 4 illustrates the aim at the end of this section, however, the challenges of the environment will blur these graphs

## Insights:

*Randomization of rewards usually poses a problem for the agent - it doesn't know what to do because it doesn't know what reward to expect.*

*.. the agent can't learn a perfect model of how observations correspond to rewards (i.e., they're "noisy")*.

Source: <https://github.com/Unity-Technologies/ml-agents/issues/1680>

## Stuff for the discussion:

As RL is nothing more than a function mapping, either exact or approximative, it follows sort of trivially that randomness makes the mapping blurry, which is the reason why the agents has a hard time dealing with the sensor clouds.

## Thoughts:

If it turns out that, no matter the actions taken, it seems impossible for the agent to learn avoiding the sensor clouds, presumably because of the randomness in the penalties of the steps (as derivative of above), one idea can be to let the penalty be fixed but use the density to change timescale/speed.

## Policy Evaluation

References:

Juliani, A. et al. (2018a). Unity: A General Platform for Intelligent Agents.

Juliani et al. (2018b), *Using TensorBoard to Observe Training*. Online resource, url: <https://github.com/Unity-Technologies/ml-agents/blob/master/docs/Using-Tensorboard.md>. Visited: 14/06/2019.

Irpan, Alex. (2018). Deep Reinforcement Learning Doesn’t Work Yet. Online resource, URL: <https://www.alexirpan.com/2018/02/14/rl-hard.html>. Visited: 23/07/2019.

Merrit, Tom. (2019). Top 5 Delivery Robots. TechRepublic, URL: <https://www.techrepublic.com/article/top-5-delivery-robots/>.

Nichols, Greg. (2019). Delivery Wars: Amazon’s new delivery robot vs. Starship’s college munchie robot. ZDNet, URL: <https://www.zdnet.com/article/amazons-new-delivery-robot-vs-starships-college-munchie-robot/>.

FedEx. (2019). Delivering the Future: FedEx Unveils Autonomous Delivery Robot. FedEx newsroom, URL: <https://about.van.fedex.com/newsroom/thefuturefedex/>.

Starship. (2019). The Self-Driving Delivery Robot. Starship/About, URL: <https://www.starship.xyz/business/>.

Scott, Sean. (2019). Meet Scout. The Amazon blog; dayone, URL: <https://www.starship.xyz/business/>.

1. It is good code practice to create new versions every time major changes are implemented, to minimise the risk of lost functionality and malfunctioning. [↑](#footnote-ref-1)
2. Made up of the *Area* prefab. [↑](#footnote-ref-2)
3. Each lesson is an entity in the arrays of the parameters in the curriculum. [↑](#footnote-ref-3)