# OBSTACLE AVOIDANCE USING A-STAR AND DEEP Q-NETWORK REINFORCEMENT LEARNING

**ABHIRAJ CHAUDHARY**

**16BCE1385**

INTERNED AT : INSTITUTE OF SYSTEMS STUDIES AND ANALYSIS , DRDO, NEW DELHI

# INTRODUCTION

Nowadays, use of drones, robots or aerial vehicles have increased by the armed forces or the intelligence services for gathering information, purpose of combat or monitoring various situations or keeping an eye along the line of control. There are times, when some actions need to be taken, without getting in surveillance of the enemy, or neighboring countries using UAVs or aerial vehicles. In the above-mentioned scenario escaping various obstacles put in the path needs to be avoided, so that the target is reached and desired action to be performed without raising any alarms, whether its avoiding enemies' radars, or various infrared equipped gadgets or any other type of obstacle. If a system is in place to find the best path or getting information about location of all obstacles beforehand, it will help the armed forces in executing their actions with more precision and higher efficiency.

The main aim of this project is obstacle avoidance, where obstacle can be in the form of radars or any other type of equipment's. This was achieved using A-Star algorithm, which draws the best possible path through static obstacles for most time saving and efficient actions to be executed and Deep Q-Network Reinforcement Learning, which constantly learns from dynamic obstacles, and after a certain period of time of training, it is able to avoid all dynamic obstacles. Further on visualization of both algorithms was done using Pygame for A- Star approach and Director for Deep Q-Network Reinforcement Learning.

The main challenge in implementing this approach was to the right type of algorithms, which was overcome by insights the project guide and research manuals. Further on choosing the algorithms, main obstacle was to develop the algorithm obstacle avoidance, keeping in mind various scenarios.

Other time consuming issue was, to choose the appropriate way for visualization of the approaches, because the user won't require any type of data, user requires visualization of their object (aircraft, UAV, etc.) avoiding obstacles, so that they are able to have the information before hand and have an insight of the situation. For this, this project uses Pygame which visualizes the path to be taken avoiding static obstacles and Director which helps the object in detecting obstacles and learning to avoid them.

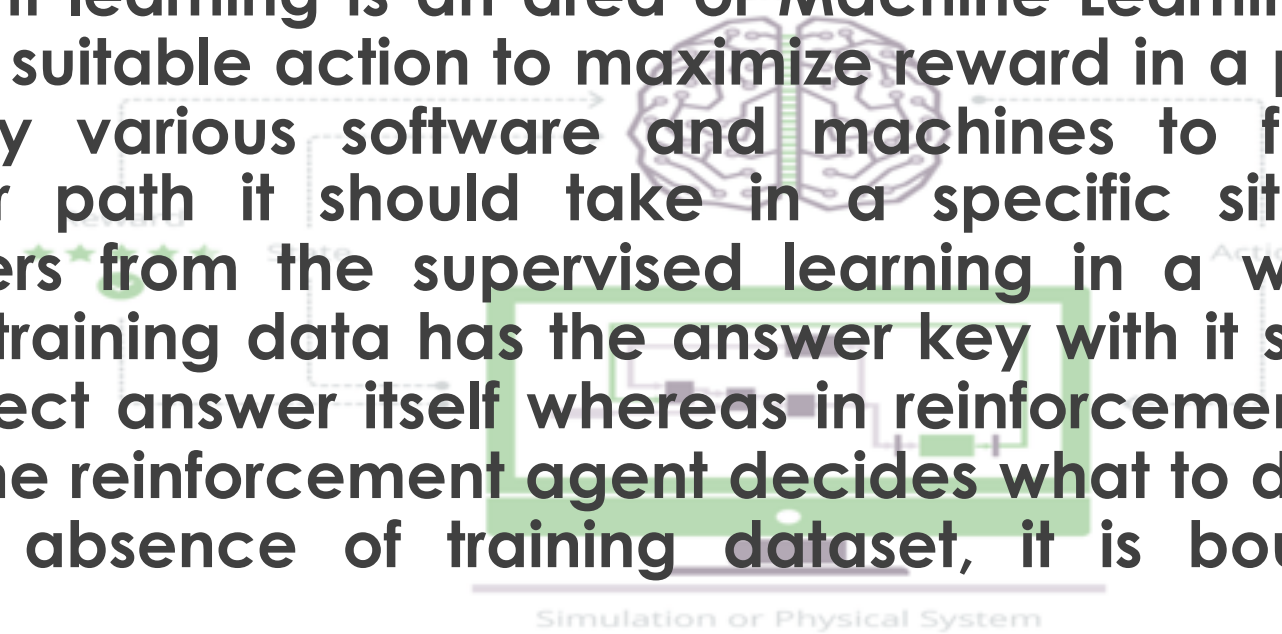
# PRE-REQUISITE KNOWLEDGE

# A-STAR

A-Star is one of the most popular methods for finding the shortest path between two locations in a mapped area. A* was developed in 1968 to combine heuristic approaches like Best-First-Search (BFS) and formal approaches like Dijsktra's algorithm. The defining characteristics of the A* algorithm are the building of a "closed list" to record areas already evaluated, a "fringe list" to record areas adjacent to those already evaluated, and the calculation of distances travelled from the "start point" with estimated distances to the "goal point". It is an informed search algorithm, or a best-first search, meaning that it is formulated in terms of weighted graphs: starting from a specific starting node of a graph, it aims to find a path to the given goal node having the smallest cost (least distance travelled, shortest time, etc.). It does this by maintaining a tree of paths originating at the start node and extending those paths one edge at a time until its termination criterion is satisfied.

# REINFORCEMENT LEARNING

Reinforcement learning is an area of Machine Learning. Reinforcement. It is about taking suitable action to maximize reward in a particular situation. It is employed by various software and machines to find the best possible behaviour or path it should take in a specific situation. Reinforcement learning differs from the supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of training dataset, it is bound to learn from its experience.



# Q LEARNING

Q-learning is a model-free reinforcement learning algorithm. The goal of Q-learning is to learn a policy, which tells an agent what action to take under what circumstances. It does not require a model of the environment, and it can handle problems with stochastic transitions and rewards, without requiring adaptations. For any finite Markov decision process (FMDP), Q-learning finds a policy that is optimal in the sense that it maximizes the expected value of the total reward over any and all successive steps, starting from the current state. Q-learning can identify an optimal action-selection policy for any given FMDP, given infinite exploration time and a partly-random policy. "Q" names the function that returns the reward used to provide the reinforcement and can be said to stand for the "quality" of an action taken in a given state. The ‘q’ in q-learning stands for quality. Quality in this case represents how useful a given action is in gaining some future reward.

# DEEP Q LEARNING

- Q-learning is a simple yet quite powerful algorithm to know the expected reward for each state for our agent. This helps the agent figure out exactly which action to perform. In situation with a large environment and numerous actions to be taken from Q-Learning is not that feasible. In deep Q-learning, we use a neural network to approximate the Q-value function. The state is given as the input and the Q-value of all possible actions is generated as the output.
- Steps involved in reinforcement learning using deep Q-learning networks (DQNs) are:
- All the past experience is stored by the user in memory
- The next action is determined by the maximum output of the Q-network
- The loss function is mean squared error of the predicted Q-value and the target Q-value – Q*. Hence basically a regression problem. Actual or target values are not known as a reinforcement learning problem is dealt with. Going back to the Q-value update equation derived from the Bellman equation :

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

# A-STAR ALGORITHM FOR OBSTACLE AVOIDANCE

PART -1

# ALGORITHM

PART 1

A-Star works by making a lowest-cost path tree from the start node to the target node. A-Star is better than other alternative algorithms because for each node, A-Star uses a function f(n) that gives an estimate of the total cost of a path using that node. Therefore, A-Star is a heuristic function, which differs from an algorithm in that a heuristic is more of an estimate and is not necessarily provably correct. A-Star expands paths that are already less expensive by using this function:

*f(n) = g(n) + h(n),*

*f(n) = total estimated cost of path through node n*

*g(n) = cost so far to reach node n*

*h(n) = estimated cost from n to goal. This is the heuristic part of the cost function*

Using a good heuristic is important in determining the performance of A-Star. The value of h(n) would ideally equal the exact cost of reaching the destination. This is, however, not possible because path is not known. Nonetheless, choose a method that will give us the exact value some of the time, such as when traveling in a straight line with no obstacles. This will result in a perfect performance of A-Star in such a case. Primary aim is to select a function h(n) that is less than the cost of reaching our goal. This will allow h to work accurately, if we select a value of h that is greater, it will lead to a faster but less accurate performance. Thus, it is usually the case that we choose an h(n) that is less than the real cost.

If the heuristic function *h* is admissible, meaning that it never overestimates the actual minimal cost of reaching the goal, then A-Star is itself admissible (or optimal) if a closed set is not used. If a closed set is used, then *h* must also be monotonic (or consistent) for A-Star to be optimal. This means that for any pair of adjacent nodes x and *y*, where *d(x,y)* denotes the length of the edge between them, :

$$h(x) \leq d(x, y) + h(y)$$

This ensures that for any path *X* from the initial node to x:

$$L(X) + h(x) \leq L(X) + d(x, y) + h(y) = L(Y) + h(y)$$

where *L* is a function that denotes the length of a path, and *Y* is the path *X* extended to include *y*. In other words, it is impossible to decrease (total distance so far + estimated remaining distance) by extending a path to include a neighbouring node. Monotonicity implies admissibility when the heuristic estimate at any goal node itself is zero, since (letting $P = (f, v_1, v_2, \ldots, v_n, g)$ be a shortest path from any node f to the nearest goal g):

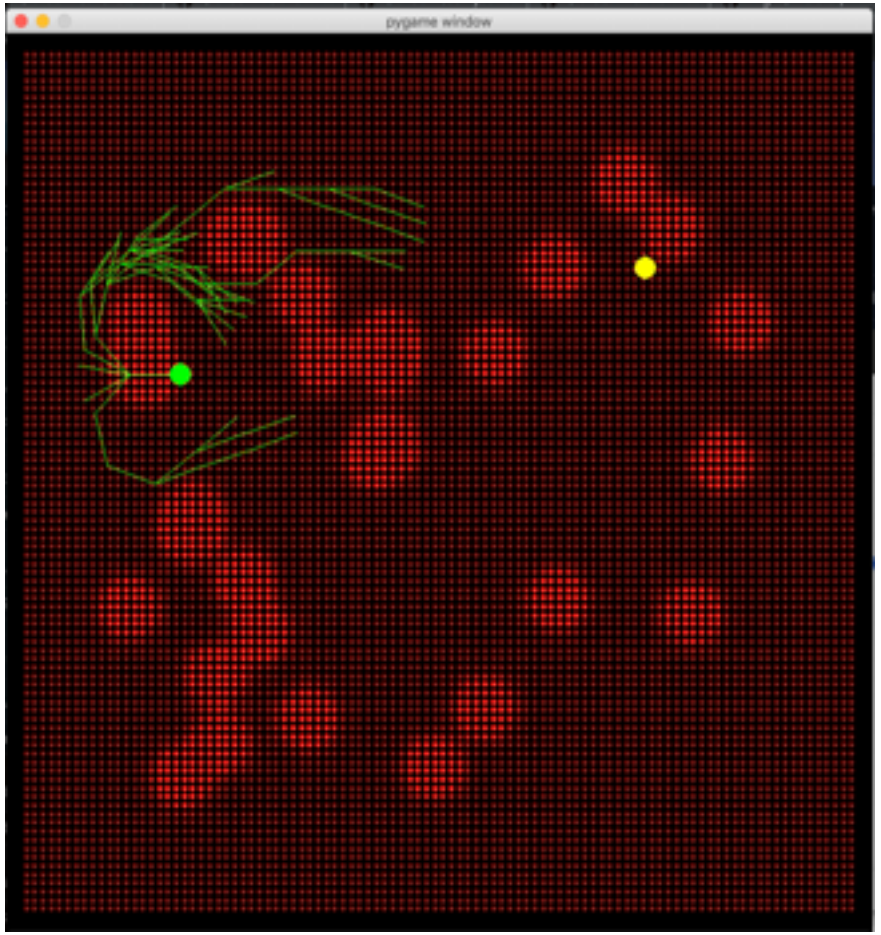
$$h(f) \leq d(f, v_1) + h(v_1) \leq d(f, v_1) + d(v_1, v_2) + h(v_2) \leq \ldots \leq L(P) + h(g) = L(P)$$

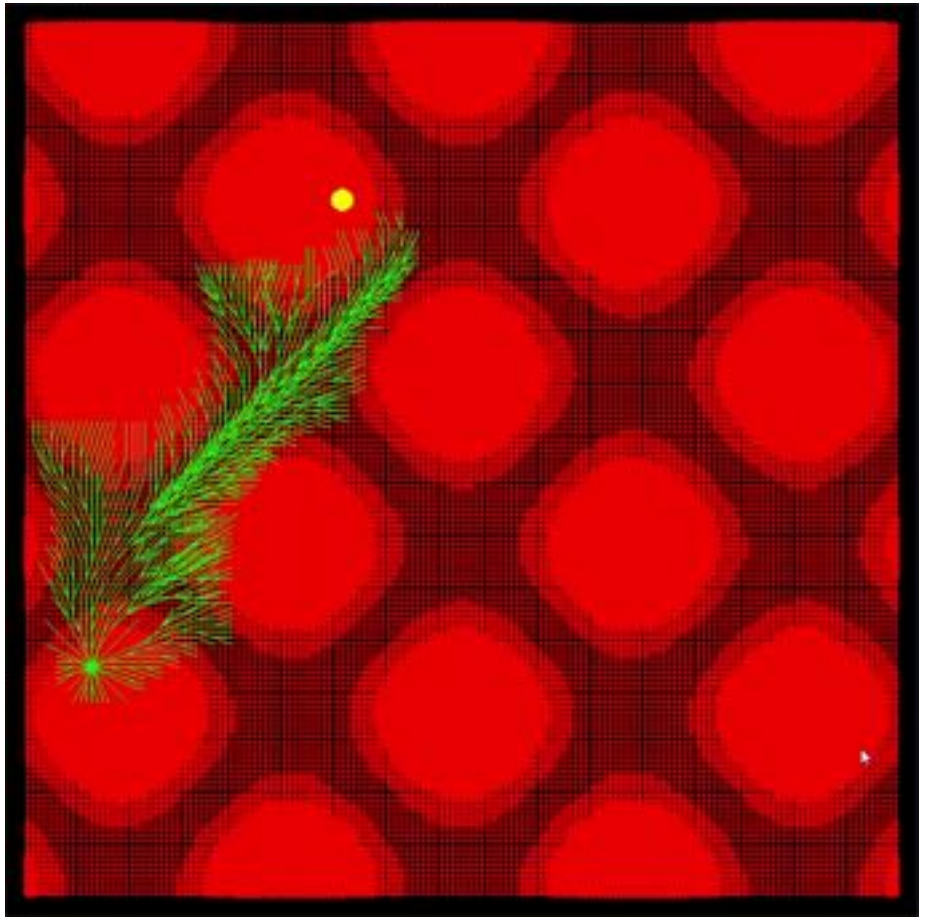
A-Star is also optimally efficient for any heuristic *h*, meaning that no optimal algorithm employing the same heuristic will expand fewer nodes than A-Star, except when there are multiple partial solutions where *h* exactly predicts the cost of the optimal path. A-Star is admissible and, in some circumstances, considers fewer nodes than any other admissible search algorithm with the same heuristic. This is because A-Star uses an "optimistic" estimate of the cost of a path through every node that it considers—optimistic in that the true cost of a path through that node to the goal will be at least as great as the estimate. But, critically, as far as A-Star "knows", that optimistic estimate might be achievable. To prove the admissibility of A-Star, the solution path returned by the algorithm is used as follows:
*When A* terminates its search, it has found a path whose actual cost is lower than the estimated cost of any path through any open node. But since those estimates are optimistic, A* can safely ignore those nodes. In other words, A* will never overlook the possibility of a lower-cost path and so is admissible.*

# IMPLEMENTATION

Implementation starts with defining a map with obstacles, start position and the target. Locations of above are assigned to various nodes, like on a map, following which these nodes are mapped to coordinates on screen. Object starts from a starting position and starts exploring various nodes, using A-Star algorithm. Exploration is represented using green colour and optimal path is displayed using yellow colour between start and target, avoiding all the obstacles. For all visualization of the process, Pygame (library of Python) is used.

# OUTPUT



# RESULT

- Result from obstacle avoidance using A-Star algorithm is a starting position and a target for an object is set. Following which obstacles (which can be radars, high mountains and etc) are determined in path between target and starting position. It can be clearly seen that algorithm has tried all possible nodes, but has rejected those with obstacles or rather a high cost. Then algorithm starts working and finds the most optimal path to the target, escaping from all the obstacles one by one. In the end, a yellow line is shown to represent the final path i.e. without colliding into any obstacles and shortest possible distance from start to target, hence obstacle avoidance is achieved.

# DEEP Q-NETWORK REINFORCEMENT LEARNING AGENT FOR OBSTACLE AVOIDANCE

PART 2

# ALGORITHM

PART 2

- A Markov chain is a mathematical model that experiences transition of states with probabilistic rules. A Markov Decision Process (MDP) is an extension of the Markov chain and it is used to model more complex environments. In this extension, add the possibility to make a choice at every state which is called an action. Also add a reward which is a feedback from the environment for going from one state to another taking one action. This is called a stochastic environment (random), in the sense that for one same action taken in the same state, there may be different results.
- The reward is the feedback from the environment that tells us how well something is doing. Goal is to maximize the total reward.

$$R_t = r_{t+1} + r_{t+2} + r_{t+3} + \ldots$$

- This is the total reward starting at some point t in time. The total reward is the sum of all the immediate rewards we get for taking actions in the environment. Defining reward leads to a few problems that sum can potentially go to infinity, which doesn't make sense since aim is to maximize it or are accounting as much for future rewards as for immediate rewards, hence use a decreasing factor for future rewards.

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots$$

- Setting γ=1 takes us back to the first expression where every reward is equally important. Setting γ=0 results in only looking for the immediate reward (always acting for the optimal next step). Setting γ between 0 and 1 is a compromise to look more for immediate reward but still account for future rewards.

$$\begin{aligned} R_t &= r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots \\ &= r_{t+1} + \gamma(r_{t+2} + \gamma r_{t+3} + \dots) \\ &= r_{t+1} + \gamma R_{t+1} \end{aligned}$$

- A policy is a function that tells what action to take when in a certain state. This function is usually denoted *π(s,a)* and yields the probability of taking action *a* in state *s*. Aim is to find the policy that maximizes the reward function.

$$\sum_a \pi(s,a) = 1$$

- Below is the expected immediate reward for going from state *s* to state *s'* through action *a*.

$$R^a_{ss'} = \mathbb{E}\left[r_{t+1} \quad | \quad s_t = s, a_t = a, s_{t+1} = s'\right]$$

- Below is the transition probability of going from state *s* to state *s'* through action *a*.

$$P^a_{ss'} = \mathbb{P}\left[s_{t+1} = s' \quad | \quad s_t = s, a_t = a\right]$$

- The state value function, and the action value function. These functions are a way to measure the "value", or how good some state is, or how good some action is, respectively

$$V^\pi(s) = \mathbb{E}\left[R_t \quad | \quad s_t = s\right]$$

- The value of a state is the expected total reward we can get starting from that state. It depends on the policy which tells how to take decisions.

$$Q^{\pi}(s,a) = \mathbb{E}\big[R_t \quad | \quad s_t = s, a_t = a\big]$$

- The value of an action taken in some state is the expected total reward we can get, starting from that state and taking that action. It also depends on the policy.
- The expected operator is linear.

$$\begin{aligned} V^{\pi}(s) &= \mathbb{E}\big[R_t \quad | \quad s_t = s\big] \\ &= \mathbb{E}\big[r_{t+1} + \gamma R_{t+1} \quad | \quad s_t = s\big] \\ &= \mathbb{E}\big[r_{t+1} \quad | \quad s_t = s\big] + \gamma \mathbb{E}\big[R_{t+1} \quad | \quad s_t = s\big] \\ &= \sum_a \pi(s,a) \sum_{s'} P^a_{ss'} R^a_{ss'} + \gamma \sum_a \pi(s,a) \sum_{s'} P^a_{ss'} V^{\pi}(s') \end{aligned}$$

- Below, is expansion of the action value function.

$$\begin{aligned} Q^{\pi}(s,a) &= \mathbb{E}\big[R_t \quad | \quad s_t = s, a_t = a\big] \\ &= \mathbb{E}\big[r_{t+1} + \gamma R_{t+1} \quad | \quad s_t = s, a_t = a\big] \\ &= \mathbb{E}\big[r_{t+1} \quad | \quad s_t = s, a_t = a\big] + \gamma \mathbb{E}\big[R_{t+1} \quad | \quad s_t = s, a_t = a\big] \\ &= \sum_{s'} P^a_{ss'} R^a_{ss'} + \gamma \sum_{s'} P^a_{ss'} V^{\pi}(s') \end{aligned}$$

- This form of the Q-Value is very generic. It handles stochastic environments

$$Q^{\pi}(s_t, a_t) = r(s_t, a_t) + \gamma V^{\pi}(s_{t+1})$$

- In a greedy policy context, below is relation between the state value and the action value functions.

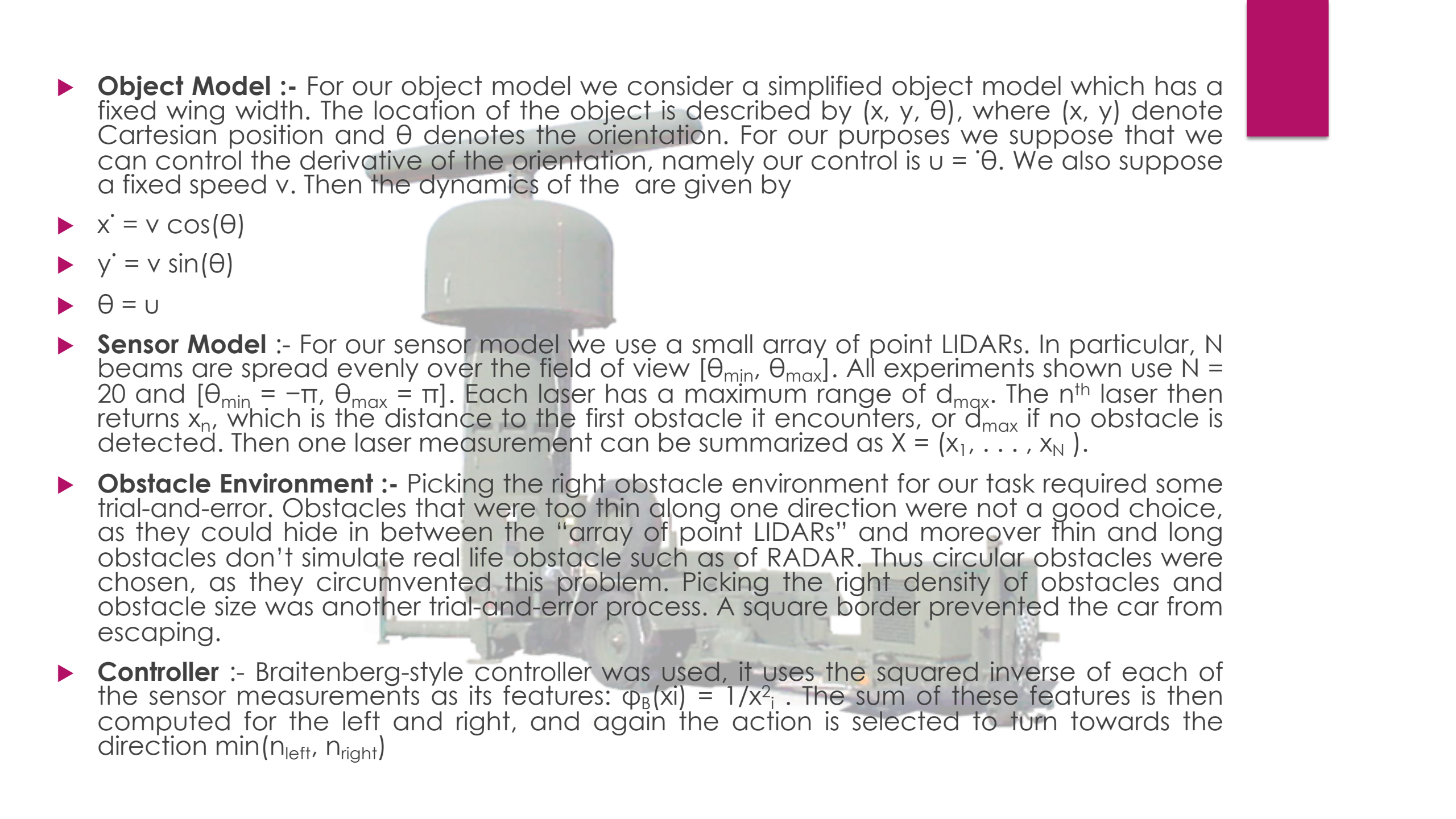
$$V(s_t) = \max_a Q(s_t, a)$$

- Q-Value of a (state, action) pair in a deterministic environment, following a greedy policy.

$$Q(s_t, a_t) = r(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1})$$

- And this is the Bellman equation in the Q-Learning. It tells that the value of an action *a* in some state *s* is the immediate reward you get for taking that action, to which add the maximum expected reward which can be received in the next state.

# ENVIRONMENT

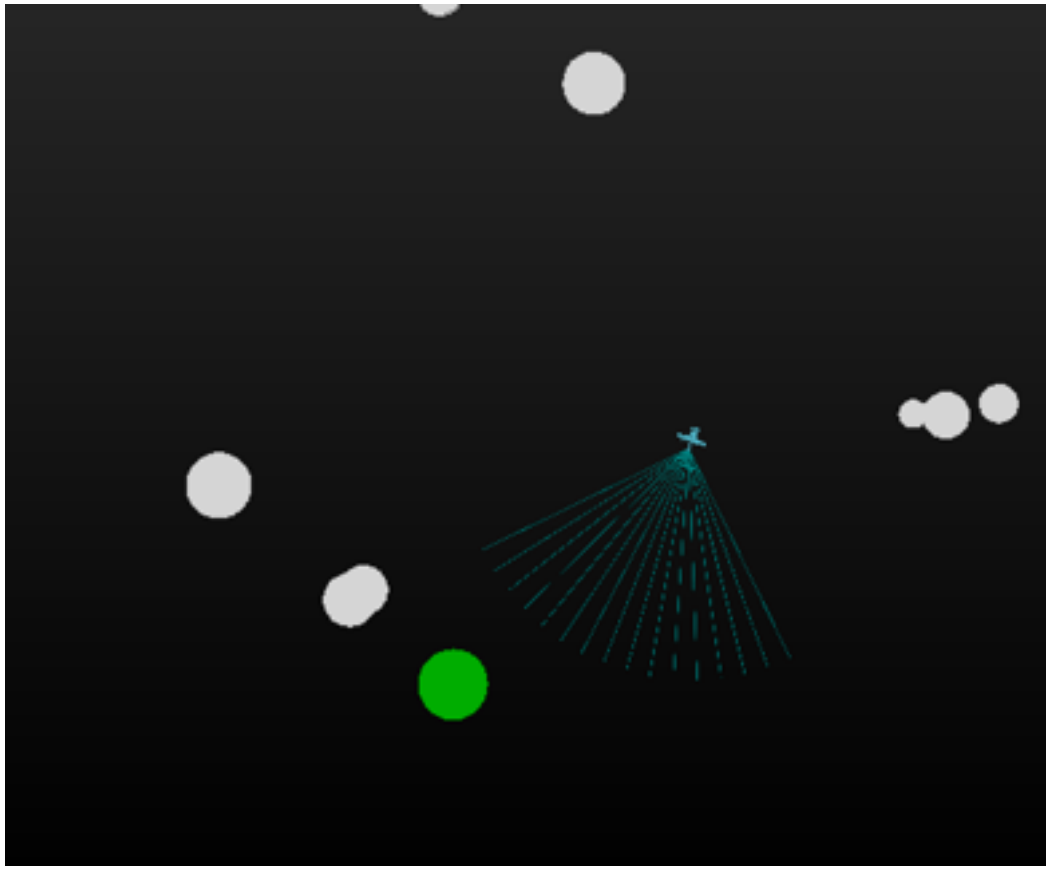


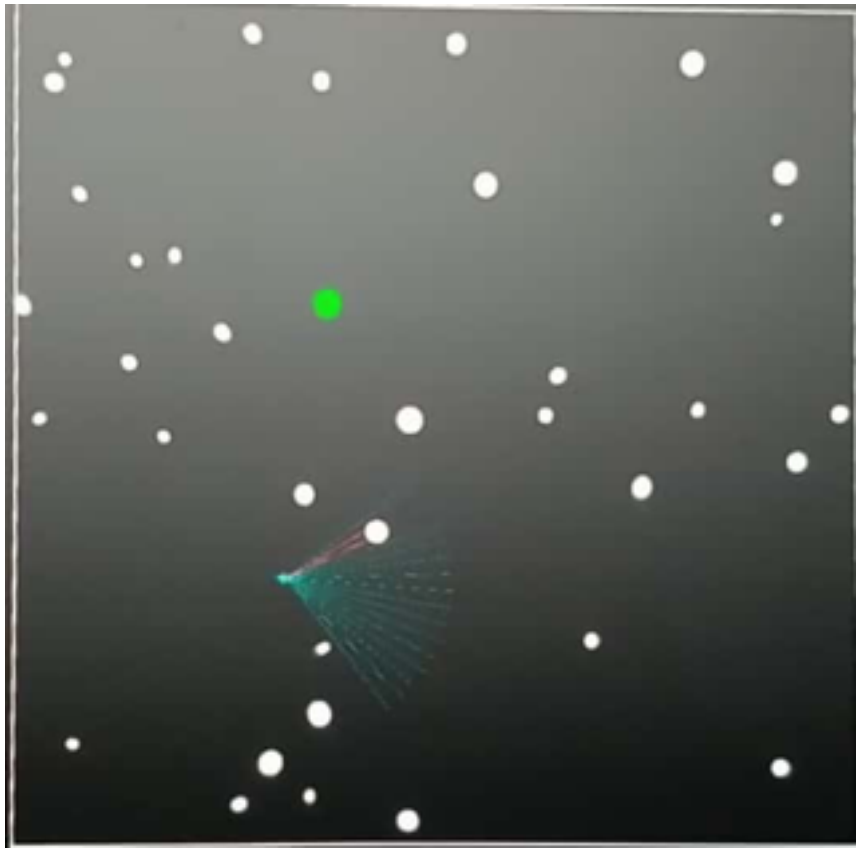
- **Object Model :-** For our object model we consider a simplified object model which has a fixed wing width. The location of the object is described by (x, y, θ), where (x, y) denote Cartesian position and θ denotes the orientation. For our purposes we suppose that we can control the derivative of the orientation, namely our control is $u = \dot{\theta}$. We also suppose a fixed speed v. Then the dynamics of the  are given by
- $\dot{x} = v \cos(\theta)$
- $\dot{y} = v \sin(\theta)$
- $\theta = u$
- **Sensor Model** :- For our sensor model we use a small array of point LIDARs. In particular, N beams are spread evenly over the field of view $[\theta_{min}, \theta_{max}]$. All experiments shown use N = 20 and $[\theta_{min} = -\pi, \theta_{max} = \pi]$. Each laser has a maximum range of $d_{max}$. The $n^{th}$ laser then returns $x_n$, which is the distance to the first obstacle it encounters, or $d_{max}$ if no obstacle is detected. Then one laser measurement can be summarized as $X = (x_1, \ldots, x_N)$.
- **Obstacle Environment :-** Picking the right obstacle environment for our task required some trial-and-error. Obstacles that were too thin along one direction were not a good choice, as they could hide in between the "array of point LIDARs" and moreover thin and long obstacles don't simulate real life obstacle such as of RADAR. Thus circular obstacles were chosen, as they circumvented this problem. Picking the right density of obstacles and obstacle size was another trial-and-error process. A square border prevented the car from escaping.
- **Controller** :- Braitenberg-style controller was used, it uses the squared inverse of each of the sensor measurements as its features: $\varphi_B(xi) = 1/x^2_i$ . The sum of these features is then computed for the left and right, and again the action is selected to turn towards the direction $\min(n_{left}, n_{right})$

# IMPLEMENTATION

- Deep Q-Network (DQN) reinforcement learning agent is set which navigates an object in a simulator to a target waypoint while avoiding obstacles. For this project assumption a map of the environment a priori is not taken but rather rely on sensor (in this case a LIDAR) to inform control decisions. Software stack is integrated with a visualization tool called Director. Director code is used for visualization and raycasting. The simulation environment is written in Python. This implementation has a main class called Simulator which combines together several other modules, e.g. moving_object, Sensor, World, Controller, net and etc, to construct a working simulation class. One nice feature of the Director application is that it uses Visualization Toolkit as the underlying graphics engine, and thus raycast calls are actually in C++ and thus very efficient. A timestep of dt = 0.05 is used, and scipy.integrate is used to for integrating forward the dynamics.

# OUTPUT





# RESULT

- Result from using Deep Q-Network (DQN) Reinforcement learning agent for obstacle avoidance is a moving object is seen with rays protruding from the object, which are actually to detect obstacles and learn from. Then a starting position and target is defined, following which object starts moving through clutter of obstacles. Then training and learning of object starts, when it collides with an obstacle it learns from it , so that next time it doesn't collide. After a few iterations of learning, object starts detecting obstacle (rays turn red when an obstacle is in course) and it diverts from the course and keeps looking forward for the target

# CONCLUSION

A-Star and Deep Q-Learning algorithms were successfully studied and implemented for obstacle avoidance. Both the mentioned techniques were thoroughly studied and implementation of obstacle avoidance is working efficiently using both mentioned above and their success is measured by 'Object being able to detect obstacle and take action to prevent itself from colliding'. Visualization of obstacle avoidance was done in both A-Star and Deep Q-learning using Pygame and Director respectively. Hence successful obstacle avoiddnce has been achieved.