0.1 Environment System Construction

Environment is defined in four layers, with three 2-dimensional Gaussian distillation layers and one uniform distribution layer. For Gaussian distributed layers, 300, 250 and 200 MSs follows a 10, 7 and 6 meters standard deviation respectively, and their mean are randomly distributed. Also, following 300 MSs follow a uniform distribution. After adding them all together. For an unbiased comparison environemnt, the random seed is fixed. All the following plots are running on the same environment.

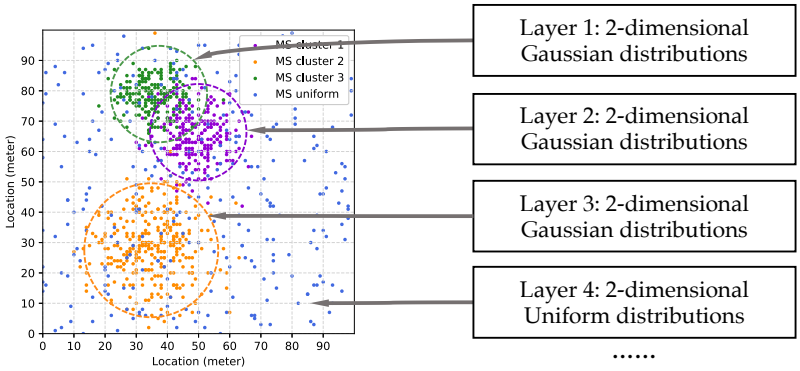

Figure 1: Environment construction

0.2 Communication System Construction

Signal to Interference Plus Noise Ratio and Reference Signal Received Power, which are used for evaluating the potential signal quality between Airborne-BSs and users' device. And a threshold (40dB) is proposed to distinguish the connection states.

$$\mathrm{RSRP}_{n,u} = \frac{\mathrm{EIRP}}{L_s} = \frac{\mathrm{EIRP}\ c^2}{(4\pi f_c d)^2} \tag{1}$$

The RSRP for the link between user u and Airborne-BS n is calculated according to the EIRP, and free space path loss is given in Equation **??**.

$$\mathrm{SINR}_{n,u} = \frac{\mathrm{RSRP}_{n,u}}{N + \sum_{\forall i \neq n} \mathrm{RSRP}_{i,u}} \tag{2}$$

Signal to Interference plus Noise Ratio (SINR) is given in Equation 2, where N is the noise power in Watts.

0.3 Training and Results

The Q-Learning, SARSA, DQN are firstly tested in our environment. In this interim report, the SARSA and Q-Learning is evaluated, and DQN will be implemented in the following experiment. During the experiment, SARSA and Q-Leaning are tested in an even parameter space with the leaning rate $\alpha = 0.1$, the discount factor $\lambda = 0.9$ and the greedy policy $\epsilon = 0.9$. The leaning includes 100 episode, where 2000 steps are in either of it. For each step, every Airborne-BS can do a single action, including 'east', 'west', 'south', 'north' and 'stay'. When beginning a new episode, the Airborne-BSs are settled back to the initial position. The initial points of the Mobile Stations are set with a fixed random seed for different algorithm. And the Airborne-BSs' initial position are evenly distributed in the four corners and sides of the environment.

Algorithm 1: DQN implementation

Initialization
for every episode j **do**
  $s_1 \leftarrow$ random
  for Every iteration t **do**
    for Every drone δ **do**
      $a_t \leftarrow \max_a Q(\phi(s_t), a; \theta)$ with probability ϵ select a random action
      $r_t, x_{t+1} \leftarrow$ based on the action a_t
      $\phi_{t+1} = \phi(s_{t+1}) \leftarrow s_{t+1} = (s_t, a_t, x_{t+1})$
      $D \leftarrow$ Add data to the dataset $D + (\phi_t, a_t, r_t, \phi_{t+1})$
      $eva Q_{t+1} \leftarrow$ evaQ learn from data selected in D
      $s_t \leftarrow s_{t+1}$
    end for
    $Q \leftarrow$ fine-tune the parameters from $eva Q$ for every n steps
  end for
end for