For the RL task I first made the environment which would generate random coefficients between -30 and 30 while initializing the initial position of the point to be randomly chosen between 0 to 100. These metrics were arbitrarily chosen and can be tweaked later. Every step of an episode moves in the specified direction by the given magnitude. The reward is calculated as the rise/fall of the value of the quadratic function defined by the coefficients.

The model used is a simple Linear model with 5 layers each with tanh activation except the last layer which does not have any activation. Previously relu was used but it led to the values exploding which i believe is because it neglects all negative values. Thus the numbers keep adding up until they lead to extremely large values causing overflows. Tanh solved this issue so it was kept as is. I still don't know why the model is not working as intended but suspect there is an error in the environment as pointed out by Vedant Shah.