**Appendix D: Hyperparameters used in experiments**

|  |  |  |
| --- | --- | --- |
| **Hyperparameter** | **Values** | **Comment** |
| batch\_size | 128 | Number of transitions sampled from replay buffer per training batch. |
| MEMORY\_CAPACITY | 200,000 | Maximum capacity of the experience replay buffer. |
| learning\_rate | 0.01 | Step size used by the optimizer for updating the weights. |
| gamma | 0.99 | Discount factor for computing the present value of future rewards. |
| episodes | 2,000 | Total number of episodes for the training process. |
| tau | 0.001 | Soft update rate for the target network parameters. |
| eps | 1.19209e-07 | Epsilon is used to prevent division by zero or very small values in numerical operations. |
| epsilon\_start | 1.0 | Initial value of the exploration rate for the epsilon-greedy strategy. |
| epsilon\_final | 0.01 | Final exploration rate at the end of training. |
| epsilon\_decay | 0.995 | The decay rate of the exploration rate over time. |
| alpha | 0.6 - 0.7 | Exponent of prioritization in Prioritized Experience Replay, determining the importance of TD error. |
| beta | 0.4 - 0.5 | Exponent of importance sampling in Prioritized Experience Replay, starting from this value and annealing to 1 |
| update\_target | 1,000 | Number of steps between updates of the target network weights. |
| max\_grad\_norm | 0.5 - 1 | Maximum gradient norm for gradient clipping. |
| reward\_scaling | 1.0 | Scaling factor for rewards to control the magnitude and impact on the gradient scale. |
| learning\_starts | 1,000 | A number of timesteps before learning starts to initially populate the replay buffer. |