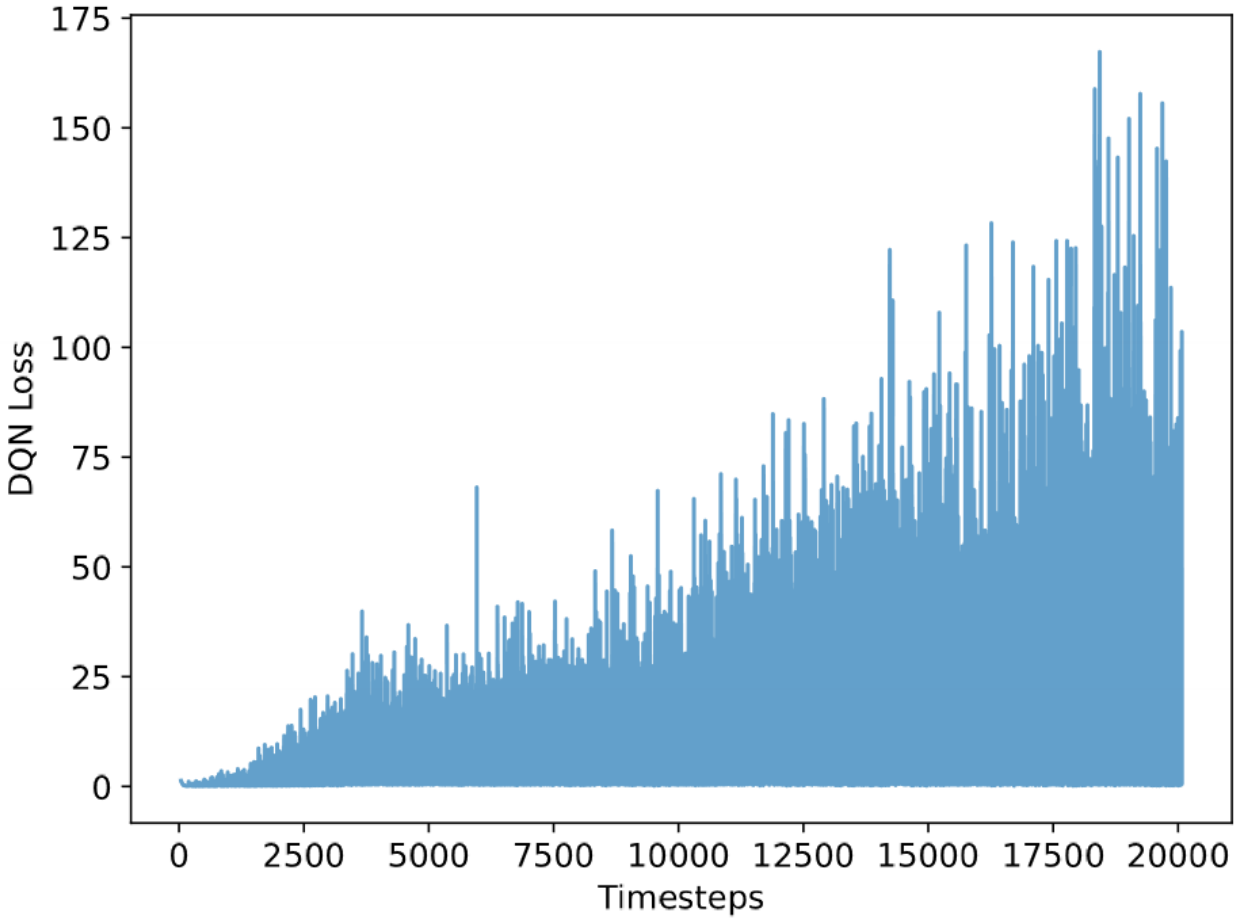

The DQN algorithm optimises (minimises) the loss against a target network, but at each update (when the update counter reaches the target update frequency) the parameters of the target network change, so the loss is optimised against the new target network. Thus, the updated parameters do not minimise the loss of the new network, causing a spike in loss. And using MSE, significantly increases the spike loss due to the squared error. Then, the loss decreases until the next update causes another spike. At each update, the loss magnitude increases because the model improves, and takes longer to fail, so the episodes become longer. Therefore, there is higher variance in the losses at later timesteps because the model keeps improving. The training data remain the same in supervised learning, in contrast to DQN training where a batch of changing data is sampled from an experience replay buffer to update the network parameters and minimise loss.