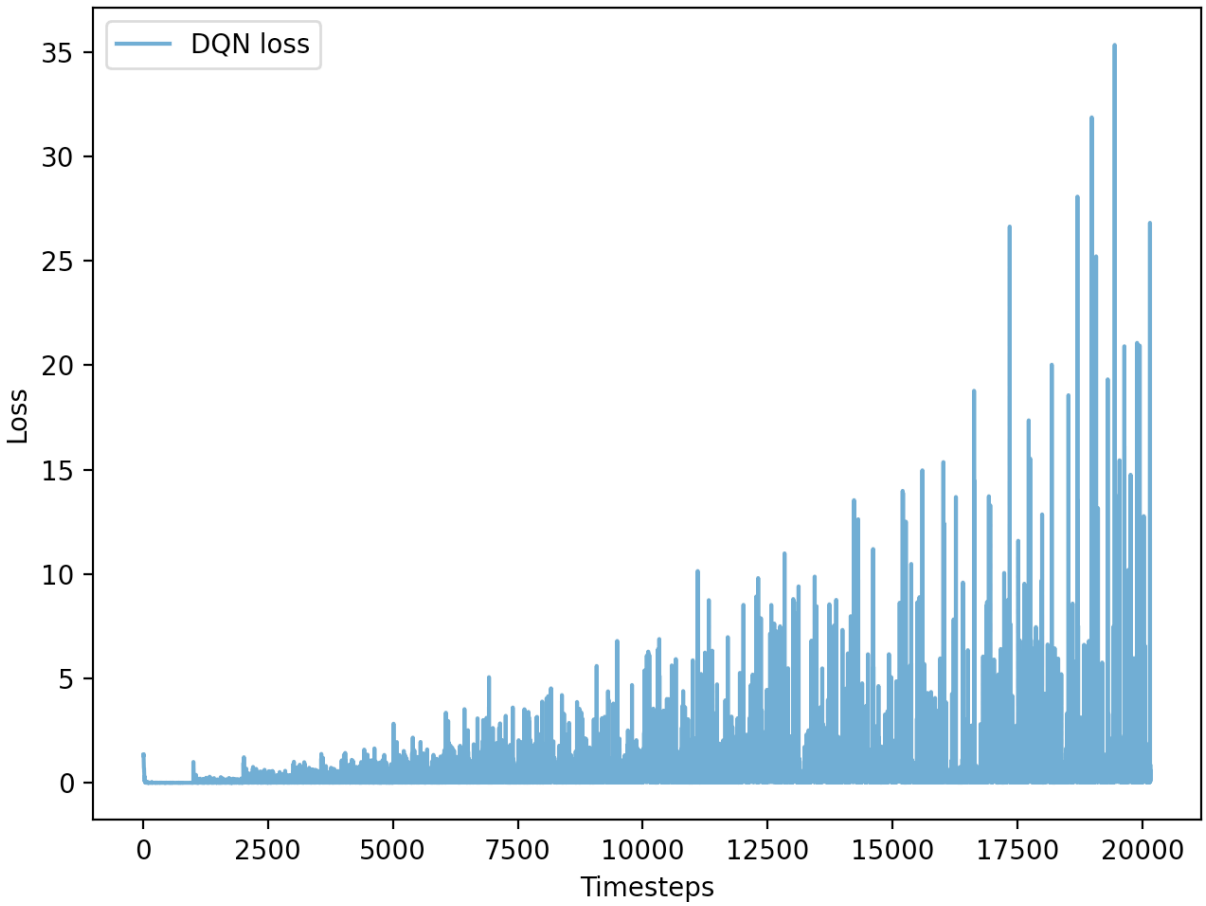

In the loss plot shown we see a linear increase in loss with time. Rewards increase with time as the agent is learning and performance improves, so absolute loss also gets larger. We compare q values from the batch experience with the reward (given an action) plus optimal q value from the network at the next time, thus we are using Adam on a predicted value. Also, in mini batch GD, we effectively use different training data with each iteration and expect loss to be noisy.

The spikes can be caused by a minibatch that contains some outliers. Also, as we are looking at the square errors here, these will make a more significant contribution and appear as increasingly large spikes. If we also have a boolean indicator the game is over, and the agent isn't aware , artificially high future rewards could be predicted, and hence higher losses. As we also update target network with a set frequency, we expect a periodic impact on the loss.