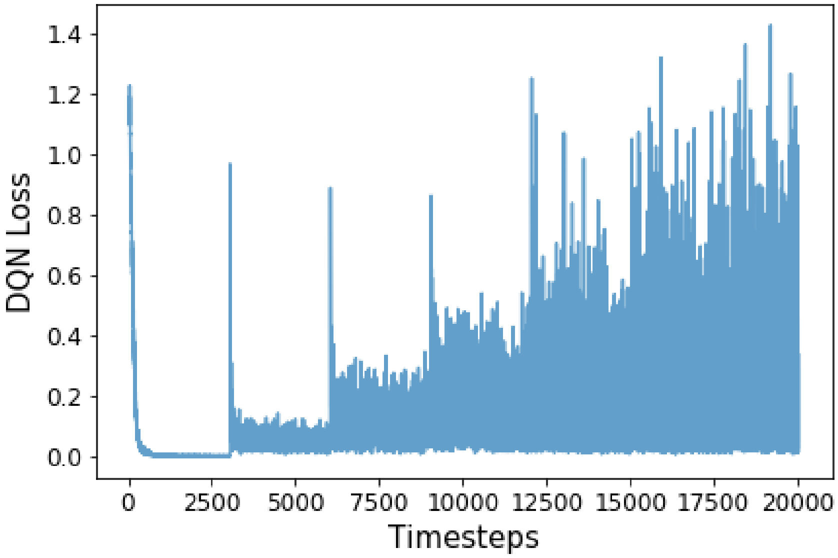

- This behaviour might seem counterintuitive. Nonetheless, in this particular case, this occurs because as the agent gets better, estimating the reward becomes harder (e.g. since the agent is now collecting rewards more frequently, the reward is not just 0). Furthermore, as the agent gets better, the episodes last for longer periods of time, adding variance to the loss calculation. Finally, we continue to have the “moving target” problem due to non-stationarity, which is somewhat mitigated with the Target network, but still isn’t a fully solved problem.

- The “spikes” observed at regular intervals occur because of the hard updates that are done at every *C* steps. The loss is calculated using stationary values, and when this hard update occurs, the network learns how off-target it has really been, resulting in a high loss value.