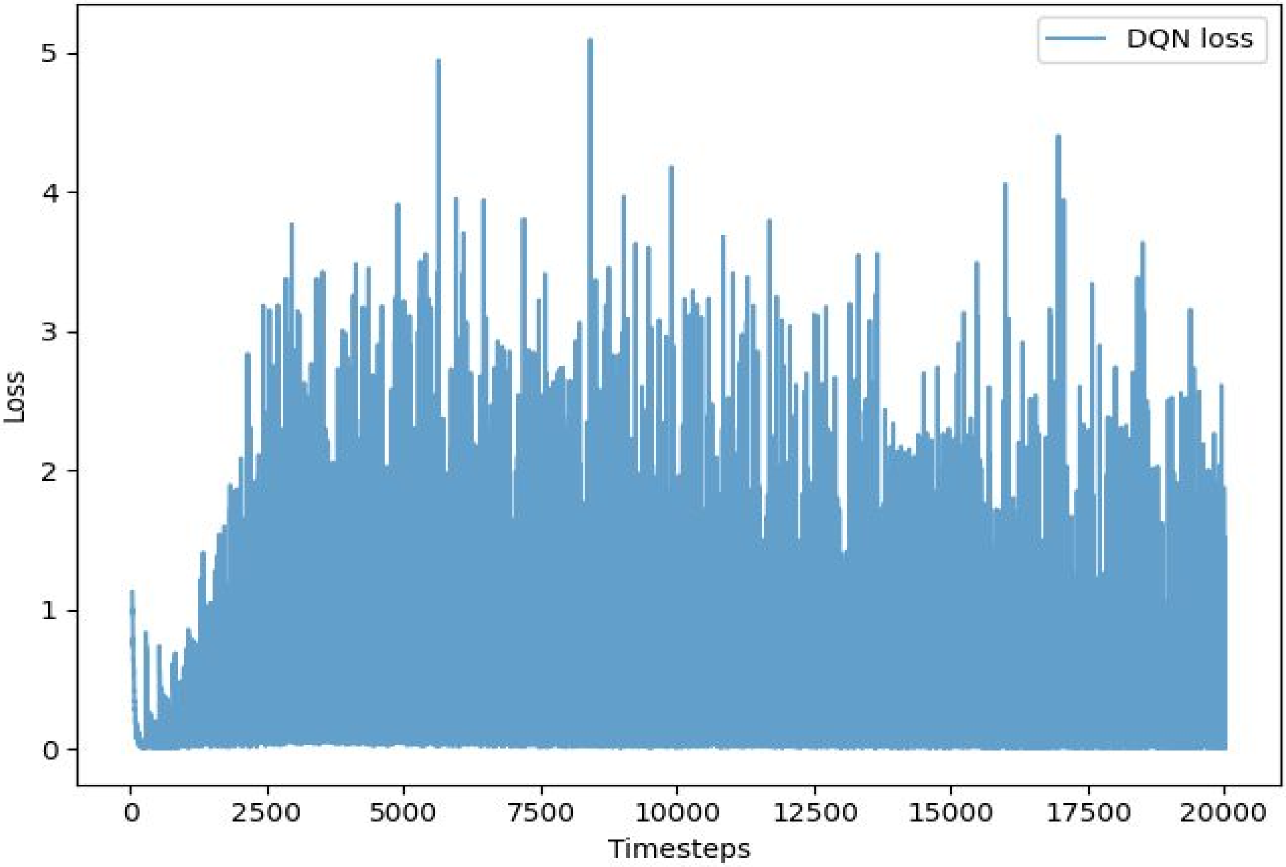
Reinforcement Learning: Deep Q-Learning (DQN) Loss Function in OpenAI CartPole Environment

i) Explain why the loss is not behaving as in typical supervised learning approaches (where we usually see a fairly steady decrease of the loss throughout training)
ii) Provide an explanation for the spikes which are standing out in the plot.

Answer:
In typical supervised learning (SL) approaches, we observe the loss function begin at a large value before decreasing with the timesteps as the agent learns via stochastic gradient descent (SGD). The loss function decreases because SL training examples are assumed to be independently and identically distributed samples from this stationary training distribution, allowing the model to make steady progress towards minimizing its training loss. However, in DQN, the target we optimize towards - approximated by the target network - is **not** stationary as it changes every time we copy over our parameters from the critic network to the target network (`hard_update`). Hence, for every `target_update_freq` timesteps, we expect the loss to periodically spike as the critic network's parameters have not been optimized towards this new target.


DQN loss
Loss
Timesteps