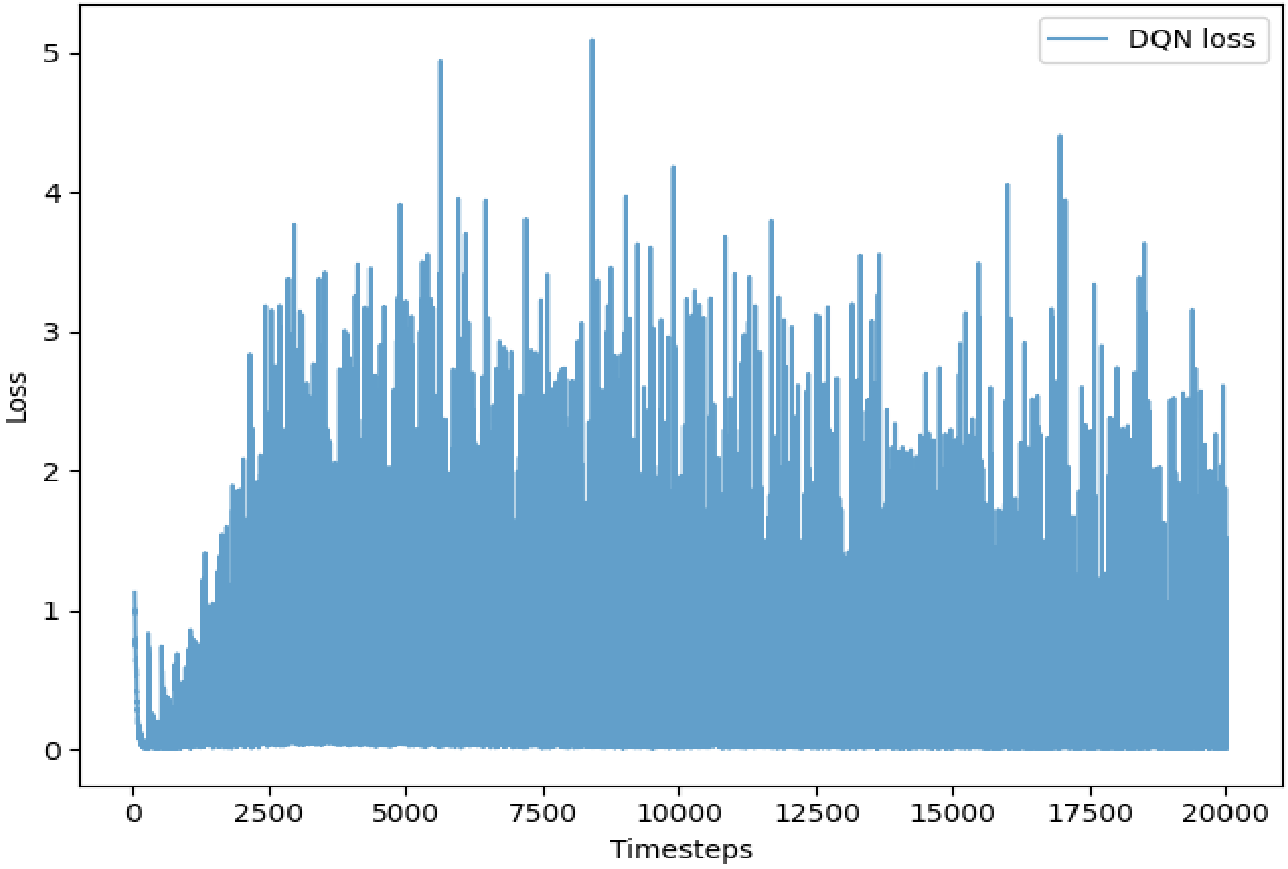
Reinforcement Learning: Deep Q-Learning (DQN) Loss Function in OpenAI CartPole Environment

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 supervised learning (SL) approaches, a function is taught using a single stationary target distribution where training samples are assumed to be independent and identically distributed (iid). This leads to a steady decrease of the loss during training as the optimizer approaches a local minimum in the loss landscape. However, in DQN, the target we optimize towards -given by the target network- is **not** stationary as it changes every time we perform a `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. Additionally, another contributing factor preventing smooth declines of the loss is that the training samples in DQN originate from the replay buffer where the iid assumption may be violated.


DQN loss
Loss
Timesteps