i) In supervised we have a stationary ground truth from which we are trying to achieve. This is not the case in DQN as the target is non-stationary. As the agent improves, the target values are getting better so it finds it harder to estimate the reward, hence the increase in loss. This is also affected by the fact that the episodes last longer as the reward increases, creating higher variance which also stops the loss from decreasing.

ii) The spikes occur at the timesteps of the target network updates. This is caused by the critic network still ‘chasing’ the target network, but since this has just been updated, the loss will be calculated between two largely different values, creating the spike.