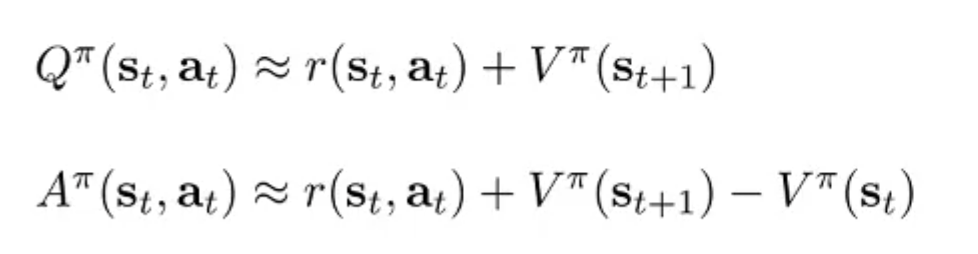
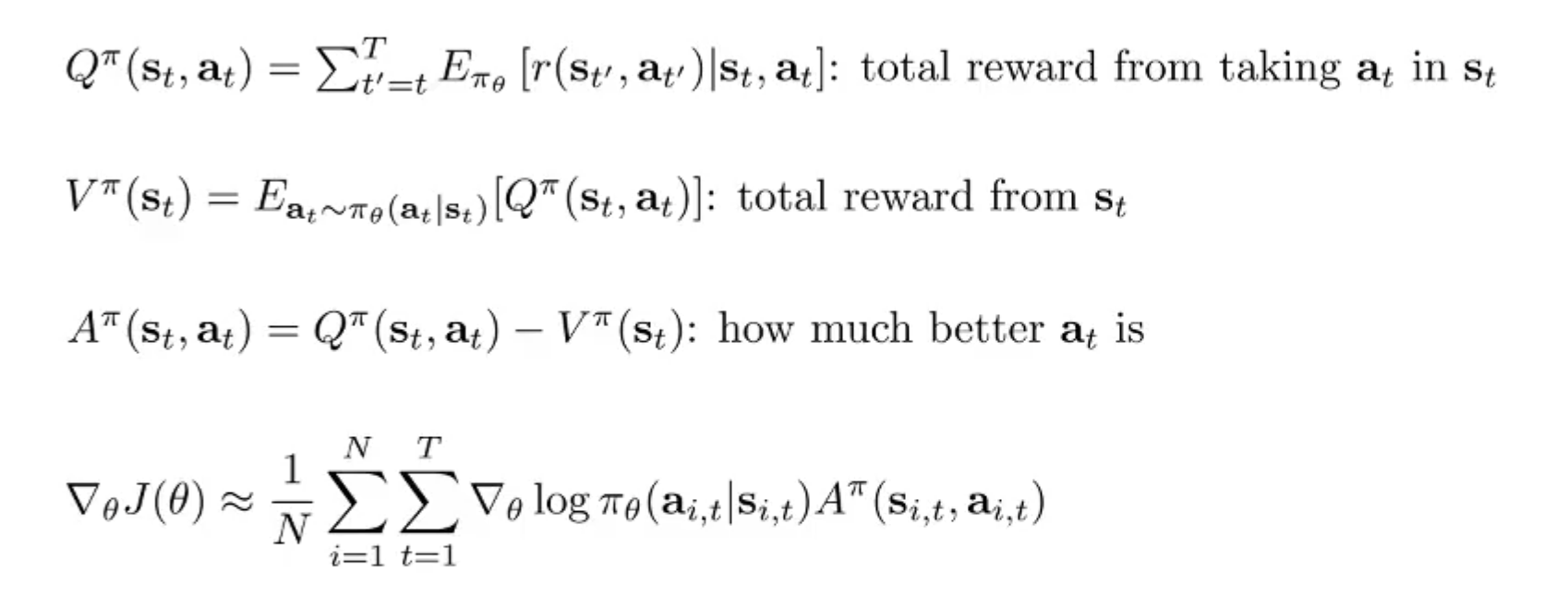
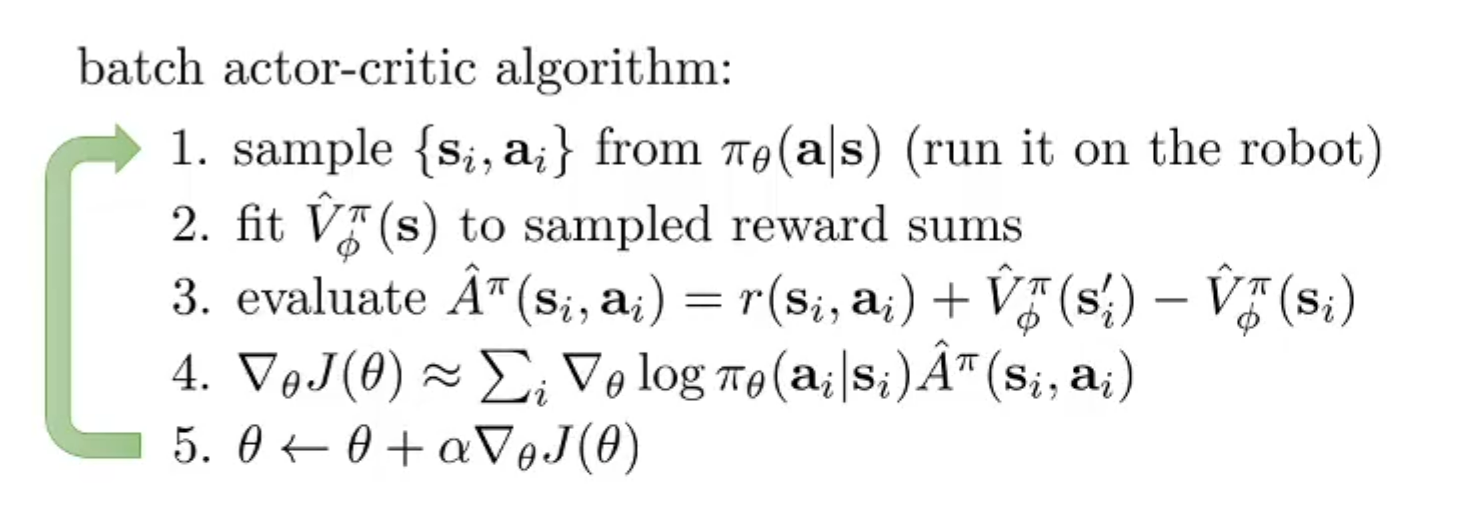
* **Lecture 6**

1. Q，V，A



1. The core of the normal actor-critic approach is introducing a network to evaluate the reward of the state-action pair. Compared to the true reward, this way can lead to lower variance and get good generalization, it is a better choice for training the policy.
2. Actor-critic algorithm



1. At the second step, we can fit V-function by Monte Carlo or Temporal Difference. MC is unbiased but have higher variance. TD is biased but have lower variance, which is more efficient. Moreover, training process only starts after an episode with MC, but can start after a single action with TD. Thus, MC is an offline algo, TD is an online algo.
2. The original policy gradient only has actor network, so it cannot use TD approach, so it is a offline algorithm. However, it also has both on-policy and off-policy implementation.