DeepSeek Highlights:

**Changes from ACER to DQN:**

Replay Memory: Changed from EpisodicReplayMemory to a simpler ReplayMemory that stores individual transitions.

Network Architecture: Replaced the actor-critic network with a simple Q-network (DQN).

Loss Calculation: Implemented the DQN loss function, which uses the Bellman equation to compute the loss.

Target Network: Added a target network to stabilize training.

Action Selection: Changed from sampling actions using a policy to selecting actions using an epsilon-greedy strategy (not explicitly shown here, but can be added).

**Changes from ACER to TD3:**  
Actor Network: The actor network will output a probability distribution over discrete actions (similar to ACER), but we will use a Gumbel-Softmax trick to sample actions during training.

Critic Networks: TD3 uses two Q-networks (critics) and takes the minimum of the two Q-values to reduce overestimation bias.

Target Networks: TD3 uses target networks for both the actor and critics, which are updated using polyak averaging.

Delayed Policy Updates: TD3 updates the policy (actor) less frequently than the Q-functions (critics).

Exploration Noise: Since the action space is discrete, we will use a simple epsilon-greedy strategy for exploration.

**Tricks**

In order for the algorithm to have stable behavior,

the replay buffer should be large enough to contain a wide range of experiences,

but it may not always be good to keep everything.

If you only use the very-most recent data, you will overfit to that and things will break;

if you use too much experience, you may slow down your learning. This may take some tuning to get right.

SpinningUp DDPG implementation uses a trick to improve exploration at the start of training.

For a fixed number of steps at the beginning (set with the start\_steps keyword argument),

the agent takes actions which are sampled from a uniform random distribution over valid actions.

After that, it returns to normal DDPG exploration.

**Novelty Features**

To facilitate getting higher-quality training data, you may reduce the scale of the noise over the course of training.

(We do not do this in SpinningUp implementation, and keep noise scale fixed throughout.)