Deep Reinforcement Learning (Sp25)

Instructor: Dr. Mohammad Hossein Rohban

Quiz Solutions - Lecture 20

Solution By: Behnia Soleymani

Q1

Random Network Distillation (RND) estimates novelty by:

A) Calculating the probability density of the current state under a model trained on past states.
B) Measuring the prediction error between a trained network and a fixed, randomly initialized target network on the current state.
C) Counting the number of times a state's hash code has been observed previously.
D) Training a classifier to distinguish the current state from a buffer of past states and using the classifier's confidence.

Correct Answer: B

Explanation: RND uses two networks: a fixed, random target network (f_ϕ) and a trained predictor network (f_θ). The exploration bonus is the prediction error $B(s_t) = ||f_\theta(s_t) - f_\phi(s_t)||^2$.

Q2

A key distinction between intrinsic motivation methods like RND or pseudo-counts and the Go-Explore strategy is that:

A) Go-Explore does not use neural networks, while RND and pseudo-counts do.
B) Intrinsic motivation methods add bonuses, while Go-Explore explicitly remembers and returns to promising states using an archive.
C) Go-Explore is only applicable to tasks with dense rewards, unlike RND.
D) Intrinsic motivation methods require simulator state resetting, while Go-Explore does not.

Correct Answer: B

Explanation: Intrinsic motivation methods modify the reward signal with a bonus $B(s, a)$ based on novelty or uncertainty. Go-Explore maintains an archive mapping state representations (cells) to high-performing states and trajectories, and explicitly returns to these archived states to explore further. Go-Explore heavily relies on state resetting.

Q3

The primary reason methods like pseudo-counts (via density or hashing) are needed instead of direct state counting in environments like Atari games is:

A) Direct counting is too slow computationally.
B) The state space is continuous or extremely large, making exact state revisits highly improbable.
C) Storing direct counts requires excessive memory.
D) Density models provide more accurate reward signals than counts.

Correct Answer: B
Explanation: These methods are needed in complex environments where the sheer size or continuous nature of the state space makes visiting the exact same state twice improbable. Direct counting becomes infeasible in such scenarios.

Q4

A key difficulty of Robustification (Phase 2) of Go-Explore is:

A) Building the initial archive takes too long.
B) Finding suitable cell representations for the state space.
C) Translating specific archived trajectories into a generally reliable policy without state-resetting.
D) The high computational cost of imitation learning algorithms.

Correct Answer: C
Explanation: Phase 2 (Robustification) aims to train a robust policy that can reach high-reward states *without* relying on state-resetting. A key challenge is effectively "translating a specific, high-performing trajectory from the archive into a generally reliable policy".

Q5

In Random Network Distillation (RND), what is the fundamental difference between the "target network" and the "predictor network"?

A) The target network is much deeper than the predictor network.
B) The predictor network is trained, while the target network's weights remain fixed after random initialization.
C) The target network predicts the next state, while the predictor network outputs the bonus.
D) The predictor network uses hashing, while the target network uses density estimation.

Correct Answer: B
Explanation: RND employs a "fixed, randomly initialized target network f_ϕ" and a "predictor network f_θ trained to mimic the target's output". The predictor f_θ is trained to minimize prediction error, while f_ϕ remains unchanged.

Q6

Compare and contrast using **Density Estimation (Pseudo-Counts)** versus **Hashing (Discrete Pseudo-Counts)** for quantifying novelty. What is a potential advantage and disadvantage of each?

Answer: Both methods aim to generalize state visit counts.

- **Density Estimation:** Learns a model $p_\theta(s)$ of visited state probability. Novelty corresponds to low probability, converted to a low pseudo-count $\hat{N}(s_t)$.
  - *Advantage:* Can potentially capture fine-grained similarities in continuous or large state spaces.
  - *Disadvantage:* Requires training potentially complex/computationally expensive density models and ensuring accurate density estimation.
- **Hashing:** Maps state s to a discrete code $\phi(s)$ and counts code occurrences $N(\text{code})$. Novelty corresponds to a low count for the code.
  - *Advantage:* Conceptually simpler; counting is computationally cheap.
  - *Disadvantage:* Effectiveness heavily depends on the quality of the hash function $\phi(s)$ correctly grouping similar states; poor hashing leads to poor novelty signals.