Reinforcement Learning

Building a Complete RL System

Mhairi Dunion, Trevor McInroe

9 February 2024

THE UNIVERSITY of EDINBURGH informatics

Lecture Outline

- What is Gymnasium?
- How to implement your own environment?
- How to implement a RL algorithm?
- How to evaluate your results?
- Demonstration

Gymnasium

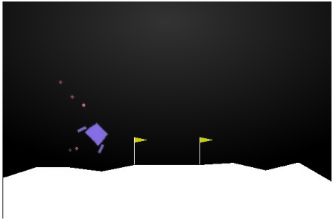
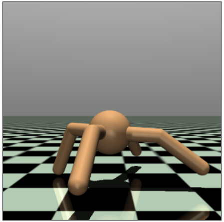
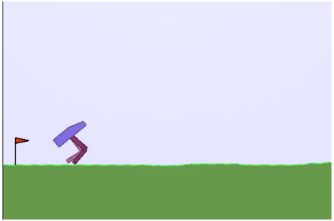
What is Gymnasium? (Towers et al., 2023)

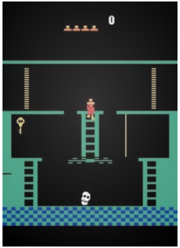
- Open source interface for sequential decision processes
- A fork of OpenAI Gym which was originally developed by OpenAI Research Lab (Brockman et al., 2016)
- Currently maintained by the Farama Foundation
- Collection of RL environments
- Standardised interface for RL environments

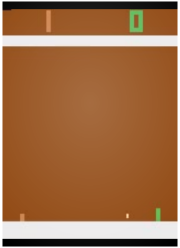
Can be installed with

```
pip install gymnasium as gym
```

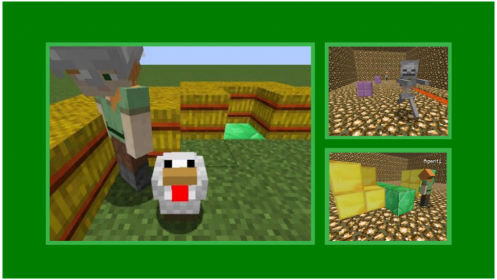
Lots of Interesting Environments! (Towers et al., 2023)



And many more… (Vinyals et al., 2017; Johnson et al., 2016; Kauten, 2018)



Gym Interface

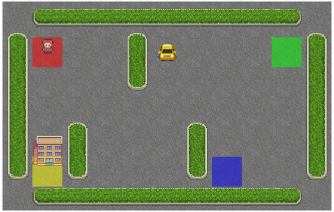
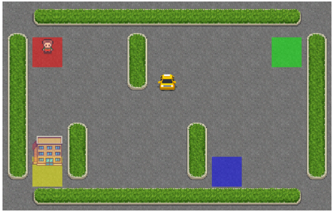
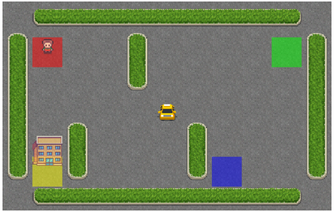
- `gym.make(<env_name>, render_mode=...)` $\longrightarrow$ gym environment
  Create a gym environment
- `env.reset()` $\longrightarrow$ observation, info
  Resets environment for a new episode
- `env.step(action)` $\longrightarrow$ observation, reward, terminated, truncated, info
  Take an action and observe new information
- `env.render()`
  Render a visualisation of the current environmental state
- `env.close()`
  Close created environment

Gym Example Snippet

Gym control flow

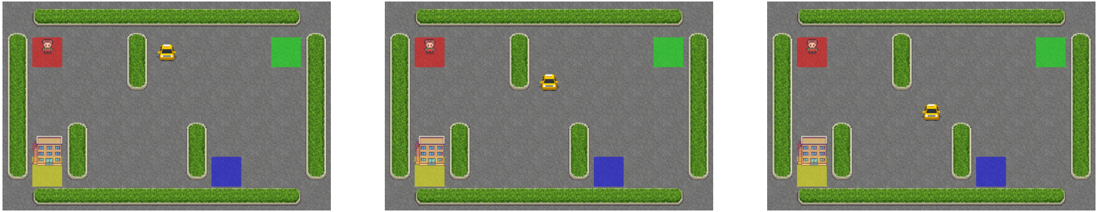
```
env = gym.make('Taxi-v3', render_mode='human')
obs, info = env.reset()
done = False
while not done:
    action = agent.choose_action(obs)
    next_obs, reward, terminated, truncated, info = env.step(action)
    done = terminated or truncated
    obs = next_obs
env.close()
```

Example: Taxi-v3 Environment

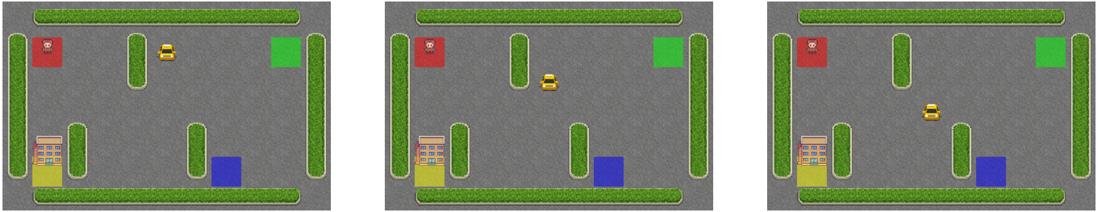
- Gridworld with 5×5 map
- Four designated pick-up and drop-off locations (Red, Green, Yellow and Blue)

Example: Taxi-v3 Environment



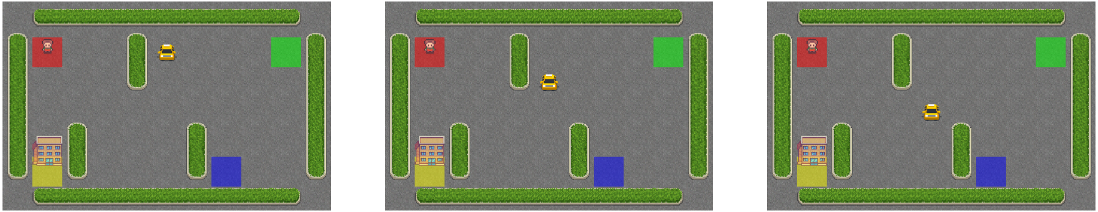
- Gridworld with 5×5 map
- Four designated pick-up and drop-off locations (Red, Green, Yellow and Blue)
- Observations $\in [0, 499]$ including taxi row and col, pass. and dest. index

Example: Taxi-v3 Environment



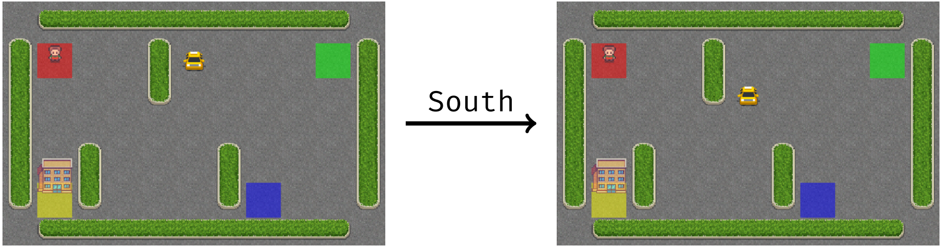
- Gridworld with 5×5 map
- Four designated pick-up and drop-off locations (Red, Green, Yellow and Blue)
- Observations $\in [0, 499]$ including taxi row and col, pass. and dest. index
- Actions: [`South`, `North`, `East`, `West`, `Pick`, `Drop`]

Example: Taxi-v3 Environment

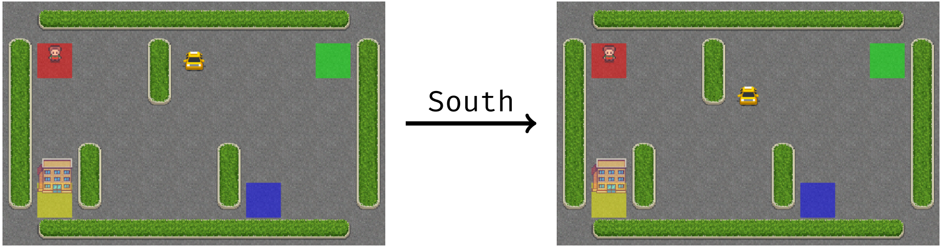


- Goal: Pickup passenger and drop it off at destination
- Reward: $+20$ for successful delivery, -1 at each timestep, -10 for illegal move
- Challenge: navigate gridworld

Taxi Environment Step I



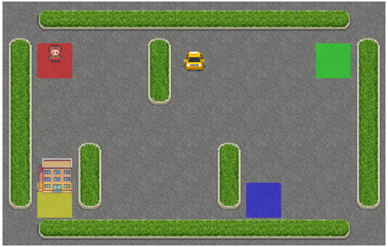
Taxi Environment Step I



$\mathtt{nobs},\ \mathtt{r},\ \mathtt{term},\ \mathtt{trunc},\ \_ = \mathtt{env.step(a)}$:

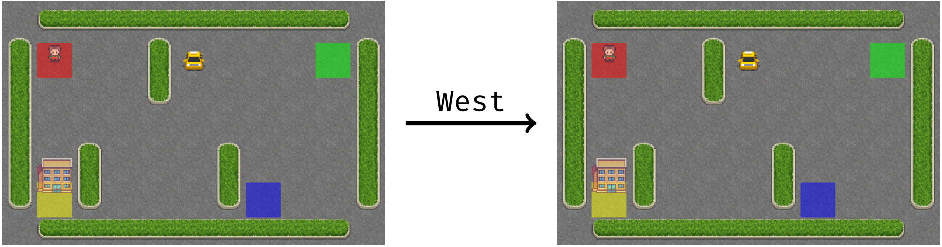
$$o = 42 \xrightarrow{\mathtt{a}=0\ (\mathtt{South})} \langle \mathtt{nobs} = 142,\ \mathtt{r} = -1,\ \mathtt{term} = \mathtt{False},\ \mathtt{trunc} = \mathtt{False} \rangle$$

Taxi Environment Step II



West

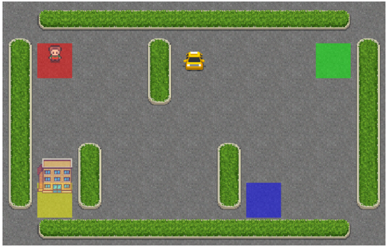
Taxi Environment Step II

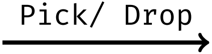


$$\texttt{nobs, r, term, trunc, \_} = \texttt{env.step(a)}:$$

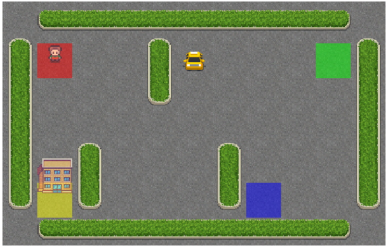
$$\texttt{o} = 45 \xrightarrow{\texttt{a}=3\ \texttt{(West)}} \langle \texttt{nobs} = 45,\ \texttt{r} = -1,\ \texttt{term} = \texttt{False},\ \texttt{trunc} = \texttt{False} \rangle$$

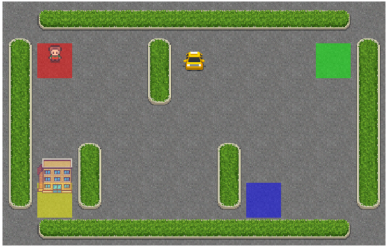
Taxi Environment Step III





Taxi Environment Step III





$$\texttt{nobs, r, term, trunc, \_} = \texttt{env.step(a)}:$$

$$\texttt{o} = 45 \xrightarrow{\texttt{a}=4/5\ \texttt{(Pick/ Drop)}} \langle \texttt{nobs} = 45,\ \texttt{r} = -10,\ \texttt{term} = \texttt{False},\ \texttt{trunc} = \texttt{False} \rangle$$

Implement your RL Agent

Recap: SARSA

On-Policy TD Control: Sarsa

$\longrightarrow$ learn q_π and improve π while following π

Updates: $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\right]$

Exploration: ϵ-greedy policy π

On-Policy TD Control: Sarsa

$\longrightarrow$ learn q_π and improve π while following π

Updates: $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\right]$

Exploration: ϵ-greedy policy π

Initialize $Q(s, a), \forall s \in \mathcal{S}, a \in \mathcal{A}(s)$, arbitrarily, and $Q(\textit{terminal-state}, \cdot) = 0$
Repeat (for each episode):
  Initialize S
  Choose A from S using policy derived from Q (e.g., ε-greedy)
  Repeat (for each step of episode):
    Take action A, observe R, S'
    Choose A' from S' using policy derived from Q (e.g., ε-greedy)
    $Q(S, A) \leftarrow Q(S, A) + \alpha[R + \gamma Q(S', A') - Q(S, A)]$
    $S \leftarrow S'$; $A \leftarrow A'$;
  until S is terminal

SARSA Agent Class Structure

- `__init__` Initialise agent and Q-table as dictionary mapping `(obs, act) -> q-val`
- `act`: ϵ-greedy policy
- `learn`: Update Q-table given new experience

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\right]$$

- `schedule_hyperparameters`: Update hyperparameters given training progress

And now in Code ... act

Epsilon-greedy Action Selection

```
def act(self, obs):
    act_vals = [self.q_table[(obs, act)] for act in range(self.
   n_acts)]
    max_val = max(act_vals)
    max_acts = [idx for idx, act_val in enumerate(act_vals) if
   act_val == max_val]

    if random.random() < self.epsilon:
        return random.randint(0, self.n_acts - 1)
    else:
        return random.choice(max_acts)
```

And now in Code ... learn

SARSA Q-Update

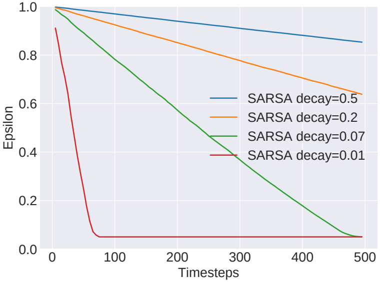
```python
def learn(self, obs, action, reward, n_obs, n_action, done):
    target_value = reward + self.gamma * (1 - done) * self.
  q_table[(n_obs, n_action)]
    self.q_table[(obs, action)] += self.alpha * (
        target_value - self.q_table[(obs, action)]
    )
    return self.q_table[(obs, action)]
```

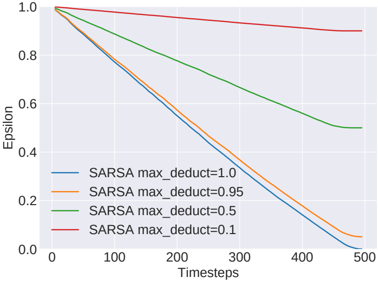
$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t) \right]$$

And now in Code ... schedule_hyperparameters

SARSA ϵ-Scheduling

```
def schedule_hyperparameters(self, timestep, max_timestep):
    max_deduct, decay = 0.95, 0.07
    self.epsilon = 1.0 - (min(1.0, timestep/(decay *
 max_timestep))) * max_deduct
```





Evaluate your Results

Why do We Evaluate in the First Place?

- It gives our approach credibility
- Empirical evaluation is no proof, but can give strong indication about the strengths and limitations of an approach (when done right!)

Why do We Evaluate in the First Place?

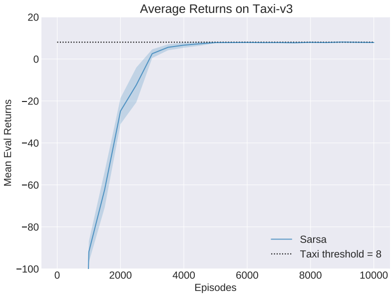
- It gives our approach credibility
- Empirical evaluation is no proof, but can give strong indication about the strengths and limitations of an approach (when done right!)

How to do it *right*?

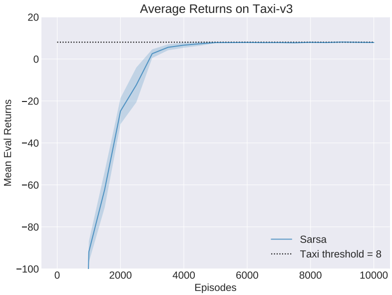
What to Evaluate?

Evaluation Returns

- Plot mean returns over multiple runs
- Visualise standard deviation or confidence interval



What to Evaluate?



Evaluation Returns

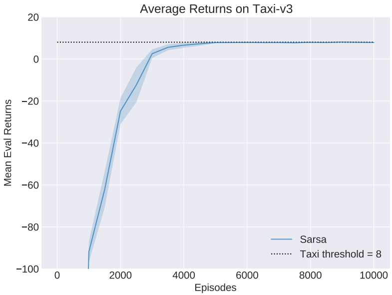
- Plot mean returns over multiple runs
- Visualise standard deviation or confidence interval

Which returns do we plot?

What to Evaluate?

Evaluation Returns

- Plot mean returns over multiple runs
- Visualise standard deviation or confidence interval

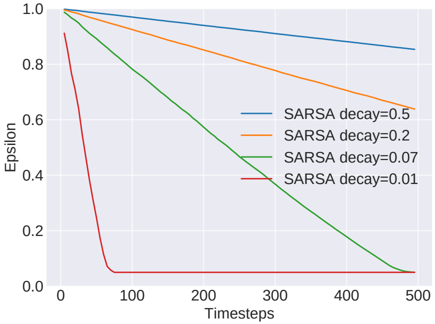


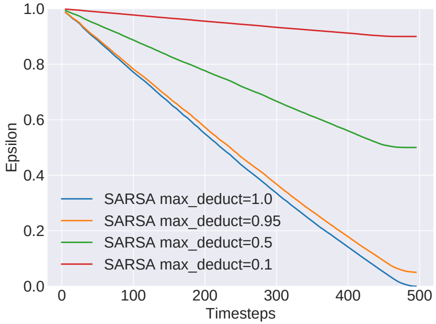
Which returns do we plot?

- Execute multiple evaluation runs with $\epsilon = 0$ at fixed intervals
- Evaluation does not involve any learning!

Hyperparameters

- Track hyperparameters, here ϵ-decay
- Try various values in a grid- or random-search and find good configuration





SARSA Gridsearch over Learning Rate α for Taxi-v3

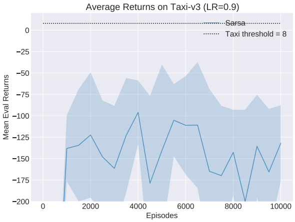


Figure 1: $\alpha = 0.9$

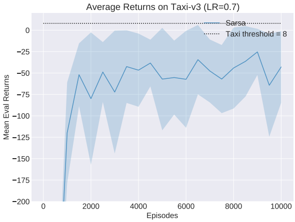


Figure 3: $\alpha = 0.7$

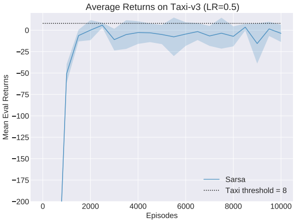


Figure 5: $\alpha = 0.5$

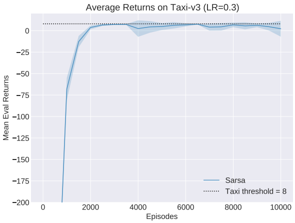


Figure 2: $\alpha = 0.3$

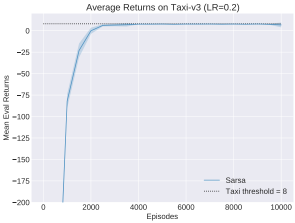


Figure 4: $\alpha = 0.2$

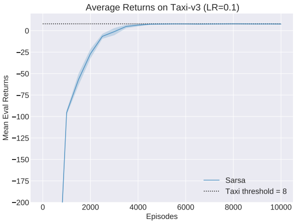


Figure 6: $\alpha = 0.1$

SARSA Learning Rate α Gridsearch Overview

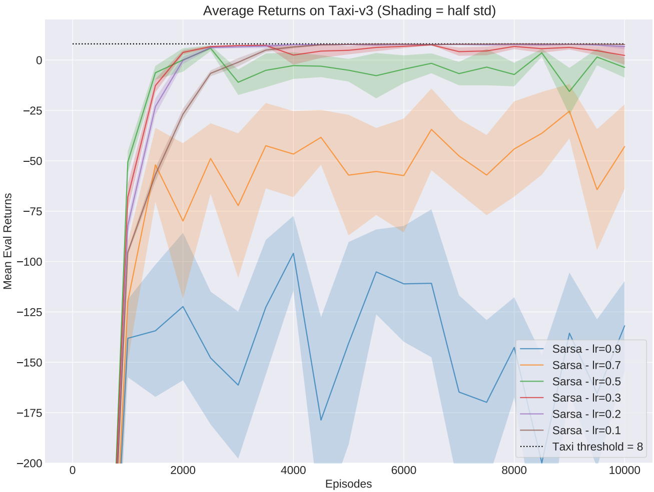


Figure 7: Gridsearch overview over learning rate α with half standard deviation as shading

Common Pitfalls (1)

”But it worked last time!”

Common Pitfalls (1)

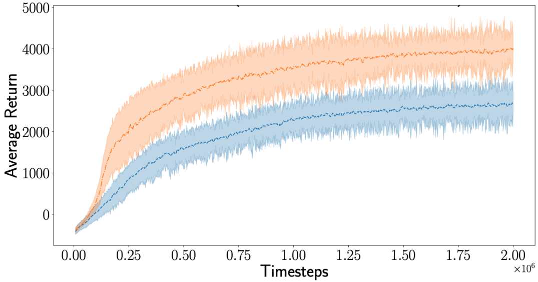
”But it worked last time!”

- It’s not enough to make it work once!
- Meaningful evaluation achieves consistent performance over multiple randomised runs
- Most RL algorithms have random components (e.g. ϵ-greedy policies)

Common Pifalls (2) (Henderson, 2018; Colas et al., 2019)

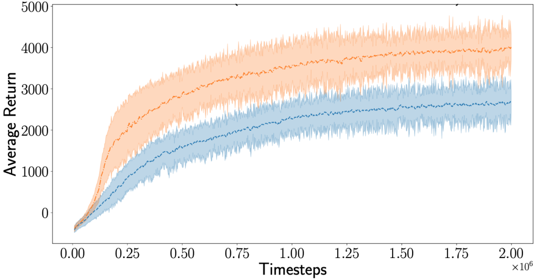
Is plotting the mean return, even with confidence interval, enough?

Is plotting the mean return, even with confidence interval, enough?



Common Pifalls (2) (Henderson, 2018; Colas et al., 2019)

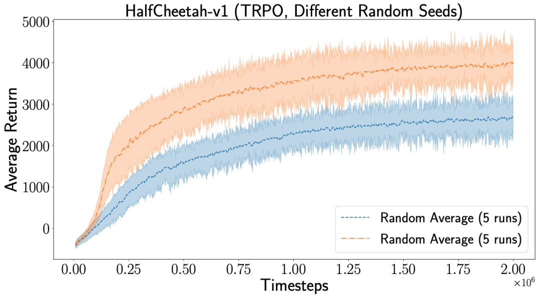
Is plotting the mean return, even with confidence interval, enough?



Which one is better?

Common Pifalls (2) (Henderson, 2018; Colas et al., 2019)

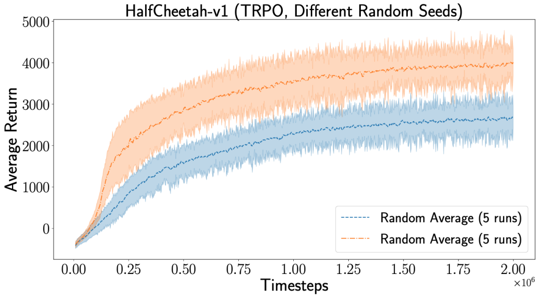
Is plotting the mean return, even with confidence interval, enough?



Which one is better? It's actually the same method!

Common Pifalls (2) (Henderson, 2018; Colas et al., 2019)

Is plotting the mean return, even with confidence interval, enough?



Which one is better? It's actually the same method!

Apparently, it's not enough! $\longrightarrow$ Statistical hypothesis testing (Colas et al., 2019) and effective statistical evaluation (Agarwal et al., 2021)

Common Pitfalls (3) (Bishop, 2019)

”Why should I use those random seeds? `random` already delivers random values!”

Common Pitfalls (3) (Bishop, 2019)

”Why should I use those random seeds? `random` already delivers random values!”

- Our goal with empirical evaluations is to make meaningful claims about the implemented approach and achieve **reproducible** performance
- Random seeds allow us to fixate random behaviour
- Reproducibility is key for meaningful research

Common Pitfalls (3) (Bishop, 2019)

”Why should I use those random seeds? `random` already delivers random values!”

- Our goal with empirical evaluations is to make meaningful claims about the implemented approach and achieve **reproducible** performance
- Random seeds allow us to fixate random behaviour
- Reproducibility is key for meaningful research

But NEVER choose/ tune your random seeds!

Common Pitfalls (3) (Bishop, 2019)

"Why should I use those random seeds? `random` already delivers random values!"

- Our goal with empirical evaluations is to make meaningful claims about the implemented approach and achieve **reproducible** performance
- Random seeds allow us to fixate random behaviour
- Reproducibility is key for meaningful research

But NEVER choose/ tune your random seeds!

Rein in the four horsemen of irreproducibility

Dorothy Bishop describes how threats to reproducibility, recognized but unaddressed for decades, might finally be brought under control.

Dorothy Bishop

Demonstration

All code is available at `https://github.com/uoe-agents/Building-a-Complete-RL-System_Demonstration`

References

Agarwal, R., Schwarzer, M., Castro, P. S., Courville, A. C., and Bellemare, M. (2021). Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34.

Bishop, D. (2019). Rein in the Four Horsemen of Irreproducibility. *Nature*, 568(7753).

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). OpenAI Gym.

Colas, C., Sigaud, O., and Oudeyer, P.-Y. (2019). A Hitchhiker's Guide to Statistical Comparisons of Reinforcement Learning Algorithms. *arXiv preprint arXiv:1904.06979*.

Henderson, P. (2018). *Reproducibility and Reusability in Deep Reinforcement Learning*. PhD thesis, McGill University Libraries.

Johnson, M., Hofmann, K., Hutton, T., and Bignell, D. (2016). The Malmo Platform for Artificial Intelligence Experimentation. In *IJCAI*, pages 4246–4247.

Kauten, C. (2018). Super Mario Bros for OpenAI Gym. GitHub.

Towers, M., Terry, J. K., Kwiatkowski, A., Balis, J. U., Cola, G. d., Deleu, T., Goulão, M., Kallinteris, A., KG, A., Krimmel, M., Perez-Vicente, R., Pierré, A., Schulhoff, S., Tai, J. J., Shen, A. T. J., and Younis, O. G. (2023). Gymnasium.

Vinyals, O., Ewalds, T., Bartunov, S., Georgiev, P., Vezhnevets, A. S., Yeo, M., Makhzani, A., Küttler, H., Agapiou, J., Schrittwieser, J., et al. (2017). Starcraft II: A new Challenge for Reinforcement Learning. *arXiv preprint arXiv:1708.04782*.

Any questions about this lecture or the demonstration?