Universidad de La Sabana Felix Mohr & Sergio Amortegui
Facultad de Ingenería 2025.2

Fundamentos de Inteligencia Artificial

- Sheet 11-

Environment & Rules. Use Python with only `numpy`. Do not use global random state; all randomness must come from a provided `numpy.random.RandomState` named `rng`. Unless explicitly stated otherwise, functions that sample must accept a seed or `rng`. Keep the function signatures from the provided `.py` templates unchanged, as they will be tested automatically on Gradescope. You will only see public test outputs.

Files to implement. `gpi/*.py` Do not rename files or functions.

Exercise 1 (7 Points)

1. Implement the abstract evaluator `TrialBasedPolicyEvaluator` and the concrete policy evaluators `FirstVisitMCEvaluator` and `ADPEvaluator`. Make sure that the `step` function returns a dictionary with relevant statistics of the step, e.g., the length of the generated trial.

   Hint: For debugging, it is recommended to use a deterministic MDP (for example the `LakeMDP` with 100% success chance) and to compute the actual state values using the already implemented `LinearEquationEvaluator`. The state values you learn with your method should coincide with the true values already after one trial.

2. Take the standard Lake MDP and the optimal policy and try to learn the correct state values for the policy with the four techniques (ADP / First-Visit with exploring starts enabled/disabled). For each of these runs, create a `pandas` DataFrame with one column for each (non-terminal) state and one row for each iteration (generated trial). In cell i, j store the absolute error of your estimation on $v(s_j)$ after iteration i. Store these dataframes in CSV files for 10^5 iterations.

   Hint: Use GPI without an improver component but just the evaluator component.

3. (Practice - Not Gradable) Then create 4 plots in which you show how the errors in the state value estimates for the 11 non-terminal states evolve over time.

Exercise 2 (3 Points)

1. Implement the `StandardTrialInterfaceBasedPolicyImprover`. Unknown q-values may be initialized with 0.

   Hint: The trial interface only serves to tell the improver which actions are possible in all states, since the q-table does not necessarily contain keys for all of those.

2. (Practice - Not Gradable) Apply GPI with this improver and with ADP/FVMC using exploring starts for 10^5 steps. Create two plots: one showing the lengths of the trials in each round (for both algorithms, i.e., two curves) and one showing the number of states with still sub-optimal actions (also for both algorithms, i.e., two curves).

Exercise 3 (Practice: Theoretical Exercises) These problems are not graded on Gradescope but are important for consolidating your theoretical understanding of the material. Write your answers by hand or in LaTeX.

1. Show that
$$\sum_{t=T}^{\infty} \gamma^t = \frac{\gamma^T}{1-\gamma}.$$
Hint: Use the fact that
$$\sum_{i=0}^{n-1} a^i = \frac{1-a^n}{1-a} \quad \text{for all } a \in [0,1[\quad \text{(property of the geometric series).}$$

2. Show that
$$\left| \sum_{t=T}^{\infty} \gamma^t r(S_t) \right| < \varepsilon \quad \text{for any } \varepsilon > 0 \quad \text{if } T > \log_\gamma \left(\frac{\varepsilon(1-\gamma)}{\max |r|} \right),$$
where $\max |r|$ is the highest absolute value of any observable reward.
Hint: Use the triangular inequality, and also observe that taking the logarithm of values < 1 yields negative values, which affects the direction of the inequality symbol. Also use the result of the previous exercise.

3. Suppose that we want that this whole accumulated utility after T is less than 10^{-4}, and we have $\gamma = 0.9$ and $\max |r| = 1$. After how many iterations may we cut the trial?