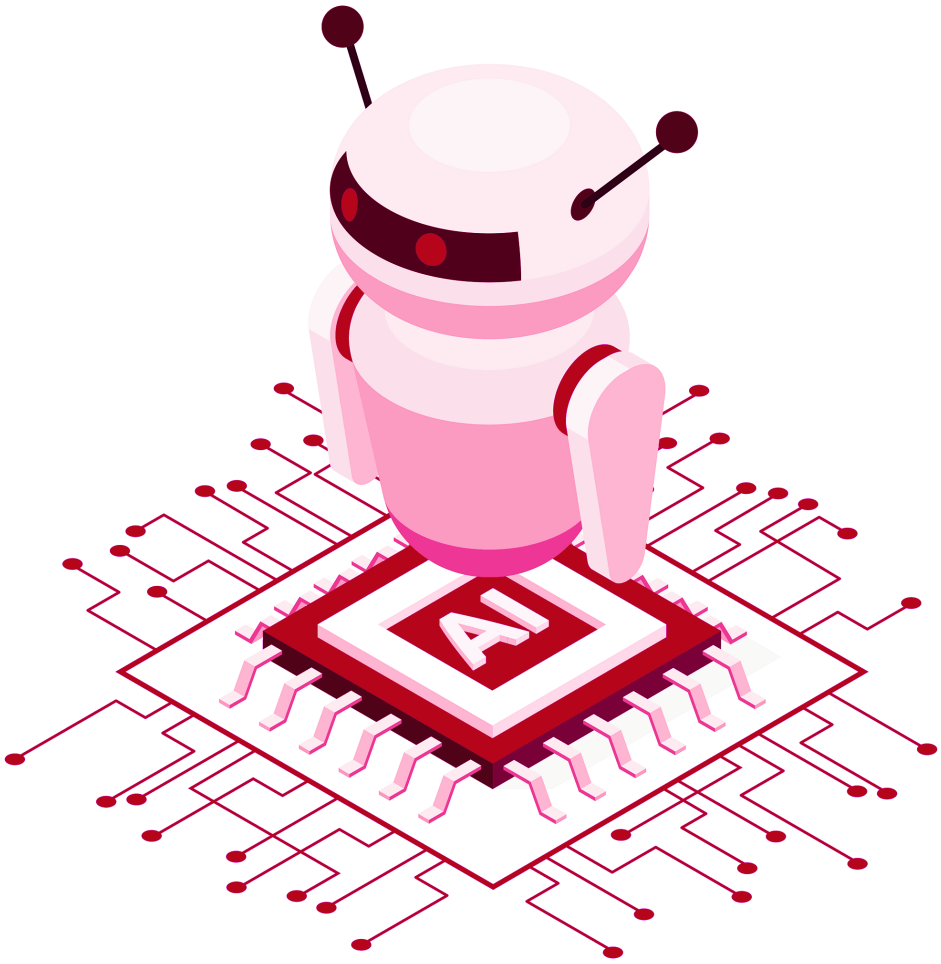

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Solution for Homework 13:

Multi-Agent RL

By:

[Full Name]

[Student Number]

RIML

Spring 2025

Grading

The grading will be based on the following criteria, with a total of 110 points:

Task	Points
Task 1	50
Task 2	50
Clarity and Quality of Code	5
Clarity and Quality of Report	5
Bonus 1	5
Bonus 2	5

Contents

1 Part 1: Game Theory Problems

1.1

Let Player 1 choose strategies (R, S, P) with probabilities (p_R, p_S, p_P), and Player 2 choose strategies (R, S, P) with probabilities (q_R, q_S, q_P).

Step 1: Expected Payoffs for Player 1

The expected payoff for Player 1 when choosing each pure strategy is:

$$\begin{aligned} U_1(R) &= 0 \cdot q_R + 1 \cdot q_S + (-1) \cdot q_P = q_S - q_P, \\ U_1(S) &= (-1) \cdot q_R + 0 \cdot q_S + 1 \cdot q_P = -q_R + q_P, \\ U_1(P) &= 1 \cdot q_R + (-1) \cdot q_S + 0 \cdot q_P = q_R - q_S. \end{aligned}$$

Step 2: Indifference Condition

In a mixed-strategy NE, Player 1 must be indifferent between all strategies they play with positive probability:

$$U_1(R) = U_1(S) = U_1(P).$$

Equating $U_1(R) = U_1(S)$:

$$q_S - q_P = -q_R + q_P \implies q_R + q_S = 2q_P.$$

Equating $U_1(S) = U_1(P)$:

$$-q_R + q_P = q_R - q_S \implies q_S + q_P = 2q_R.$$

Step 3: Solve for Probabilities

Also, probabilities sum to 1:

$$q_R + q_S + q_P = 1.$$

Solving the system:

$$\begin{cases} q_R + q_S = 2q_P \\ q_S + q_P = 2q_R \\ q_R + q_S + q_P = 1 \end{cases}$$

From these equations, we find:

$$q_R = q_S = q_P = \frac{1}{3}.$$

By symmetry, Player 2's mixed strategy is the same, giving Player 1 the same probabilities:

$$p_R = p_S = p_P = \frac{1}{3}.$$

Step 4: Conclusion

Thus, the unique mixed-strategy Nash Equilibrium for the standard Rock-Scissors-Paper game is:

$$\boxed{(p_R, p_S, p_P) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right), \quad (q_R, q_S, q_P) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right).}$$

1.2

Let Player 1 choose strategies (R, S, P) with probabilities (p_R, p_S, p_P), and Player 2 choose strategies (R, S, P) with probabilities (q_R, q_S, q_P).

Step 1: Expected Payoffs for Player 1

The expected payoff for Player 1 when choosing each pure strategy is:

$$\begin{aligned} U_1(R) &= 0 \cdot q_R + 1 \cdot q_S + (-2) \cdot q_P = q_S - 2q_P, \\ U_1(S) &= (-1) \cdot q_R + 0 \cdot q_S + 3 \cdot q_P = -q_R + 3q_P, \\ U_1(P) &= 2 \cdot q_R + (-3) \cdot q_S + 0 \cdot q_P = 2q_R - 3q_S. \end{aligned}$$

Step 2: Indifference Condition

In a mixed-strategy NE, Player 1 must be indifferent between all strategies they play with positive probability:

$$U_1(R) = U_1(S) = U_1(P).$$

Equating $U_1(R) = U_1(S)$:

$$q_S - 2q_P = -q_R + 3q_P \implies q_R + q_S = 5q_P.$$

Equating $U_1(S) = U_1(P)$:

$$-q_R + 3q_P = 2q_R - 3q_S \implies 3q_R - 3q_S - 3q_P = 0 \implies q_R - q_S - q_P = 0 \implies q_R = q_S + q_P.$$

Step 3: Solve for Probabilities

Also, probabilities sum to 1:

$$q_R + q_S + q_P = 1.$$

Substitute $q_R = q_S + q_P$ into the sum:

$$(q_S + q_P) + q_S + q_P = 1 \implies 2q_S + 2q_P = 1 \implies q_S + q_P = \frac{1}{2}.$$

Then $q_R = q_S + q_P = \frac{1}{2}$.

From Step 2, $q_R + q_S = 5q_P \implies \frac{1}{2} + q_S = 5q_P \implies q_S = 5q_P - \frac{1}{2}$.

Also $q_S + q_P = \frac{1}{2} \implies q_S = \frac{1}{2} - q_P$.

Equating the two expressions for q_S:

$$5q_P - \frac{1}{2} = \frac{1}{2} - q_P \implies 6q_P = 1 \implies q_P = \frac{1}{6}.$$

Then:

$$q_S = \frac{1}{2} - \frac{1}{6} = \frac{1}{3}, \quad q_R = \frac{1}{2}.$$

By symmetry, Player 2's mixed strategy is the same:

$$p_R = \frac{1}{2}, \quad p_S = \frac{1}{3}, \quad p_P = \frac{1}{6}.$$

Step 4: Conclusion

Thus, the mixed-strategy Nash Equilibrium for the modified RSP game is:

$$\boxed{(p_R, p_S, p_P) = \left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right), \quad (q_R, q_S, q_P) = \left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right).}$$

2.2

Convergence

Both games' empirical frequencies converge (in the sense of the averages stabilizing) as the number of iterations grows. Initially, there is strong variability, which reduces over time — the moving average (empirical frequency) stabilizes.

Do they converge to the Nash equilibrium?

Yes. The long-run empirical frequencies for Player 1 are very close to the theoretical mixed Nash equilibrium probabilities in both the standard and modified RSP games. The final empirical vectors match the analytic solutions to within a few 10^{-4} (standard) and 10^{-3} (modified), which is consistent with statistical sampling / averaging noise after 10^6 plays.

Why this behavior?

Fictitious play in two-player zero-sum games is known to converge to the Nash equilibrium (the empirical frequencies converge); the observed behavior matches this theory.

Early fluctuations: At the beginning, empirical frequencies are computed from very few samples, so a single action has a large effect. As counts grow, the averaging dampens fluctuations (law of large numbers).

Small residual jitter: Small residual fluctuations remain because players best-respond to empirical frequencies deterministically, and the process is stochastic only due to tie-breaking or initial choices. After many iterations, the empirical frequencies fluctuate around the NE with amplitude roughly $O(1/N)$, where N is the effective sample size. Deterministic tie-breaking can also introduce tiny periodicities, but they vanish asymptotically.

2 3.2

Small ϵ (e.g., 0.01, 0.1)

- Behavior is almost identical to standard fictitious play.
- Frequencies converge close to the NE $[0.5, 0.25, 0.25]$.
- Occasional random moves create small fluctuations but do not derail convergence.

Moderate ϵ (e.g., 0.3)

- Convergence is slower.
- The strategy distribution oscillates around the NE but does not settle as tightly.
- The equilibrium is still "visible" in the long-run averages, but the noise floor is higher.

Large ϵ (e.g., 0.5)

- Exploration dominates.
- Frequencies no longer converge cleanly to NE; instead, they hover around a mixture between random play and NE.
- This means each action probability is biased toward uniform distribution $(1/3, 1/3, 1/3)$, diluting the equilibrium signal.

3 4.2

Instantaneous strategy: Oscillates a lot, reflecting which actions currently have the highest positive regret. It does not converge to the Nash Equilibrium directly.

Average strategy: Smoothly converges to the NE. This matches the theoretical property of Regret Matching: average strategies form a coarse correlated equilibrium, and in two-player zero-sum games, this equals the Nash Equilibrium.

Theoretical Explanation: Convergence of Regret Matching

1. Definitions and Goal. Fix a player i. Let the sequence of joint action profiles over T rounds be $(a^1, \ldots, a^T)$, where $a^t = (a_i^t, a_{-i}^t)$. Define the *external regret* for not having played a fixed action $s \in A_i$ as

$$R_T(s) = \sum_{t=1}^{T} \big(u_i(s, a_{-i}^t) - u_i(a_i^t, a_{-i}^t)\big).$$

The average per-period regret is $R_T(s)/T$. A no-regret algorithm guarantees

$$\max_s \frac{R_T(s)}{T} \longrightarrow 0 \quad \text{as } T \to \infty.$$

2. Vanishing regret implies a coarse correlated equilibrium. Define the empirical distribution over joint actions:

$$\hat{\pi}_T(a) = \frac{1}{T}\#\{t : a^t = a\}.$$

Then

$$\frac{1}{T}R_T(s) = \mathbb{E}_{\hat{\pi}_T}[u_i(s, a_{-i})] - \mathbb{E}_{\hat{\pi}_T}[u_i(a_i, a_{-i})].$$

If $\max_s R_T(s)/T \le \varepsilon_T \to 0$, we have

$$\mathbb{E}_{\hat{\pi}_T}[u_i(a_i, a_{-i})] \ge \mathbb{E}_{\hat{\pi}_T}[u_i(s, a_{-i})] - \varepsilon_T, \quad \forall s,$$

which is the definition of an ε_T-coarse correlated equilibrium (CCE).

3. From CCE to Nash in two-player zero-sum games. In two-player zero-sum games, any CCE achieves the minimax value for both players. Thus, the *marginal* or *average* strategy of each player under repeated play is an approximate Nash equilibrium. The approximation error vanishes as regrets vanish.

4. Why Regret Matching guarantees this outcome. Regret Matching (Hart & Mas-Colell, 2000) updates play probabilities proportionally to positive cumulative regrets. It can be shown that maximum external regret grows sublinearly in T, ensuring

$$\max_s \frac{R_T(s)}{T} \longrightarrow 0.$$

Hence, by points 2 and 3, the time-average strategy converges to the Nash equilibrium.

5. Intuition. - Regret measures how much better a fixed alternate action would have done in hindsight. - If all regrets vanish, no player has a systematic incentive to deviate. - In two-player zero-sum games, "no systematic incentive to deviate" implies that the empirical average of play achieves the minimax value, which is exactly the Nash equilibrium.

6. Summary. Vanishing external regret $\Rightarrow$ empirical distribution is an approximate CCE (Hart & Mas-Colell). In two-player zero-sum games, any CCE corresponds to a Nash equilibrium in expected payoff. Therefore, Regret Matching's *average strategy* converges to a Nash equilibrium.

4 Part 2: Implementing MADDPG/IDDPG

1. In our training loop, the `DDPGLoss` module utilizes `target_policies` to estimate the value of the next state. Explain clearly why employing these slowly-updating target networks, rather than the main policy networks (which change rapidly), is essential for ensuring the stability of the DDPG algorithm. (Hint: Consider what might happen if the critic tried to optimize toward a continuously moving target.)

   Target networks provide a *slowly moving reference* for the critic's learning target. In DDPG, the critic's loss depends on the estimated value of the next state, which is produced by the actor (policy) network. If we were to use the *current*, rapidly-updating actor to generate these targets, then every time the critic updates, the target itself would also change drastically. This creates a "moving target" problem: the critic would constantly chase a value that shifts at every gradient step, leading to instability and potential divergence.

   By instead employing *target networks*—which are updated only slowly (e.g., by Polyak averaging)—the target values evolve smoothly. This stabilizes the critic's learning because the temporal-difference targets remain relatively consistent across successive updates. Consequently, the actor can be trained on a critic that provides a more reliable estimate of the value function, ensuring overall stability of the DDPG training process.

2. (bonus) Consider the training plot shown in Figure 1, which resulted from modifying a single scalar hyper-parameter in the training script.

   (a) Describe the issue with the learning process depicted in the plot.

   (b) Identify which hyper-parameter you believe was changed, and explain the role of this parameter within the MADDPG algorithm.

1. In Figure 1, the rewards for all three agents initially fluctuate strongly and only improve very slowly. Even after many iterations, the reward curves show high variance and no smooth or steady convergence. This indicates that the training process is unstable and the agents struggle to consistently improve their policies.

2. A likely cause is that the *Polyak averaging coefficient* τ (used for the soft updates of the target networks) was set too **high**. In MADDPG, τ controls how quickly the target networks track the main networks:

$$\theta^{\text{target}} \leftarrow \tau\theta^{\text{main}} + (1-\tau)\theta^{\text{target}}.$$

   If τ is too large, the target networks change almost as fast as the main networks. This removes the stabilizing effect of slowly moving targets and causes the critic to chase a moving objective, which produces the observed instability in the learning curves.

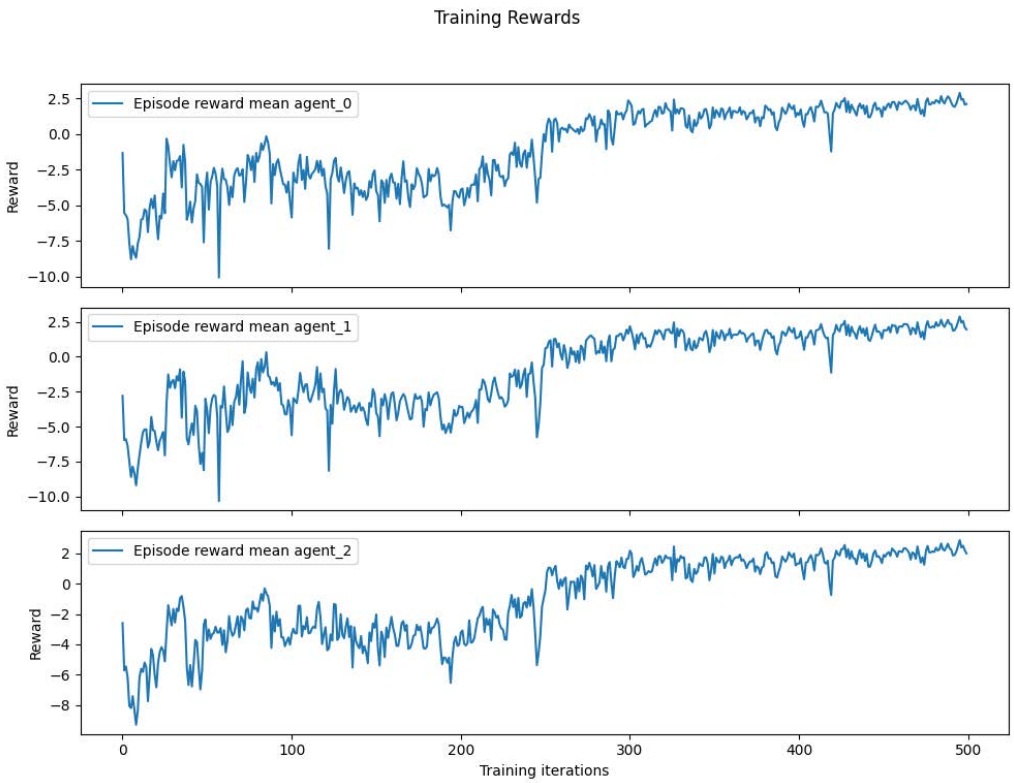

Figure 1: Agents performance after modifying a scalar hyper-parameter.

References

[1] Cover image designed by freepik

[2] Ryan Lowe, Yi I. Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. arXiv:1706.02275