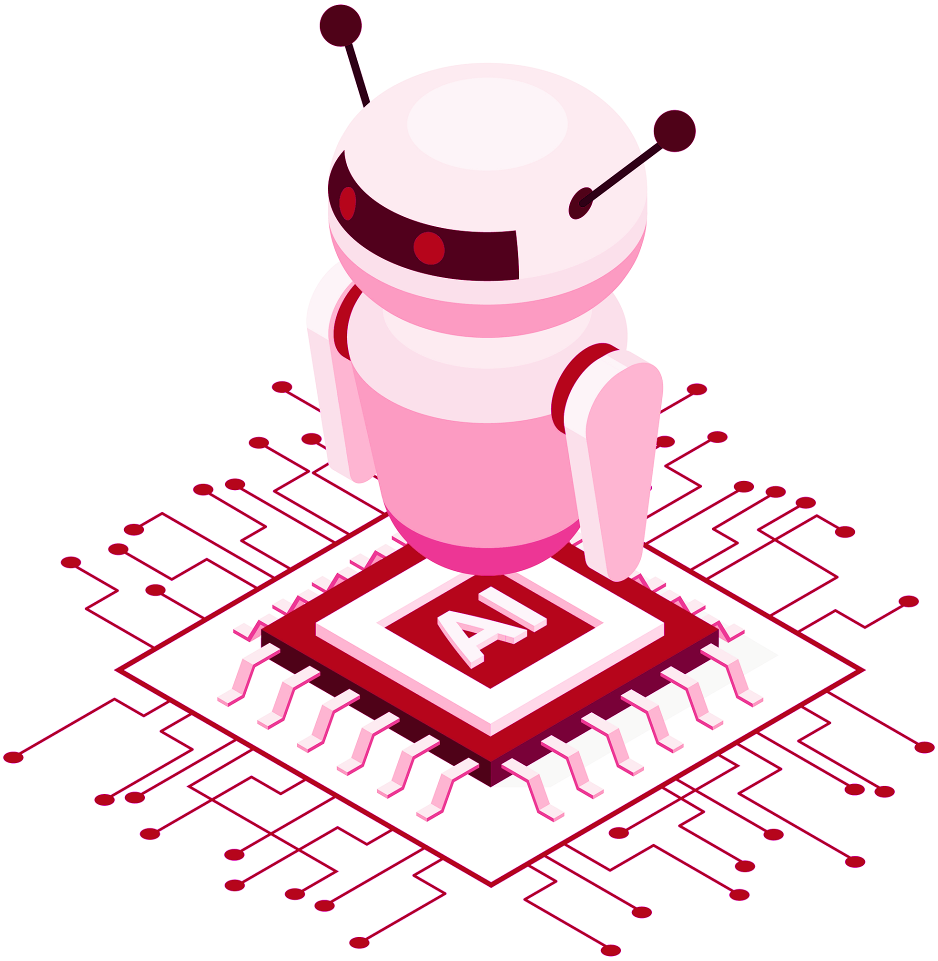

Deep Reinforcement Learning

Professor Mohammad Hossein Rohban

Solution for Homework 13:

Multi-Agent RL

By:

Payam Taebi

400104867

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

Problem 1: Nash Equilibrium (Theory)

1.1 Standard Rock–Scissors–Paper (Question)

Given the standard RSP payoff matrix:

Player 1	Rock	Scissors	Paper
Rock	0, 0	1, -1	-1, 1
Scissors	-1, 1	0, 0	1, -1
Paper	1, -1	-1, 1	0, 0

Task: Analytically derive the mixed-strategy Nash Equilibrium for this game. Show the steps for setting up the indifference equations for Player 1 and solving for Player 2's equilibrium strategy probabilities (q_R, q_S, q_P).

1.1 Standard Rock–Scissors–Paper (Answer)

Let Player 2 play Rock/Scissors/Paper with probabilities (q_R, q_S, q_P) (with $q_R + q_S + q_P = 1$). Player 1's expected payoff from each pure action against (q_R, q_S, q_P) is:

$$\begin{aligned}
u_1(R) &= 0 \cdot q_R + 1 \cdot q_S + (-1) \cdot q_P = q_S - q_P, \\
u_1(S) &= (-1) \cdot q_R + 0 \cdot q_S + 1 \cdot q_P = q_P - q_R, \\
u_1(P) &= 1 \cdot q_R + (-1) \cdot q_S + 0 \cdot q_P = q_R - q_S.
\end{aligned}$$

At a mixed NE, Player 1 must be indifferent among pure strategies: $u_1(R) = u_1(S) = u_1(P)$. Equating,

$$\begin{aligned}
q_S - q_P = q_P - q_R &\Rightarrow q_R + q_S = 2q_P, \\
q_P - q_R = q_R - q_S &\Rightarrow q_P + q_S = 2q_R, \\
q_R + q_S + q_P = 1.
\end{aligned}$$

Solving gives $q_R = q_S = q_P = \frac{1}{3}$. By symmetry, Player 1 also mixes uniformly, and the game value is 0.

1.2 Modified Rock–Scissors–Paper (Question)

Consider the modified RSP game where the stakes are higher:

Player 1	Rock	Scissors	Paper
Rock	0, 0	1, -1	-2, 2
Scissors	-1, 1	0, 0	3, -3
Paper	2, -2	-3, 3	0, 0

Task: Derive the mixed-strategy Nash Equilibrium for this modified game.

1.2 Modified Rock–Scissors–Paper (Answer)

Let Player 2 mix with (q_R, q_S, q_P), $q_R + q_S + q_P = 1$. Player 1's expected payoffs are

$$\begin{aligned}
u_1(R) &= q_S - 2q_P, \\
u_1(S) &= -q_R + 3q_P, \\
u_1(P) &= 2q_R - 3q_S.
\end{aligned}$$

Indifference $u_1(R) = u_1(S) = u_1(P)$ gives

$$\begin{aligned} q_R + q_S &= 5q_P, \\ q_P + q_S &= q_R, \\ q_R + q_S + q_P &= 1. \end{aligned}$$

Solving yields $(q_R, q_S, q_P) = \left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right)$. By symmetry, Player 1 uses the same mix and the value is 0.

Problem 2: Learning by Observation — Fictitious Play (Analysis)

2.2 Analysis (Question)

1. Run the simulation for $1{,}000{,}000$ iterations on both the *standard* and *modified* RSP games.
2. Generate two plots (one per game) with action-frequency trajectories and horizontal lines at the theoretical NE.
3. Analyze: Do the frequencies converge? If so, to the NE?

2.2 Analysis (Answer)

Final empirical frequencies (after 10^6 iterations):

- **Standard RSP** (NE $= (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$)
  P1: $[0.334,\ 0.332631,\ 0.333369]$; P2: $[0.333002,\ 0.33372,\ 0.333278]$.
- **Modified RSP** (NE $= (\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$)
  P1: $[0.500082,\ 0.332963,\ 0.166955]$; P2: $[0.49924,\ 0.334497,\ 0.166263]$.

In both games, the *empirical frequencies* converge to the theoretical NE. Instantaneous best responses may keep cycling, but the running averages settle near the NE, with residual fluctuations diminishing over time.

Problem 3: Fictitious Play with Exploration (Analysis)

3.2 Analysis (Question)

1. Run `simulate_epsilon_greedy_fp` on the *modified* RSP for 10^6 iterations with $\epsilon \in \{0.01, 0.1, 0.3\}$.
2. Plot the results for each ϵ.
3. Analyze: How does ϵ affect learning dynamics? Does the strategy converge to the NE? If not, to what? Discuss the impact of exploration.

3.2 Analysis (Answer)

With a *constant* exploration rate $\epsilon > 0$, one might expect averages to shift toward the uniform policy. However, in this zero-sum game the NE has *full support*, so each pure action is payoff-equivalent at equilibrium beliefs. The exploit component (played with probability $1 - \epsilon$) compensates for uniform exploration, yielding long-run averages that stay at the same NE, provided feasibility holds (here, $\epsilon \leq$

$3 \min_i p_i^\star = 0.5$). Empirically, for $\epsilon \in \{0.01, 0.1, 0.3\}$ the empirical frequencies converge to the NE; larger ϵ increases variance (noisier curves) but does not bias the limit within this range.

Problem 4: Learning from "What If" — Regret Matching (Analysis)

4.2 Analysis (Question)

1. Run regret matching on the *modified* RSP for 10^6 iterations.
2. Produce a figure with two subplots: (i) instantaneous strategy of P1 over time; (ii) average strategy of P1 with NE lines.
3. Analyze: Compare the two plots. Which one converges to the NE? (Bonus: Why is this the expected theoretical outcome?)

4.2 Analysis (Answer)

Final strategies (your run):
Instantaneous (last step): $[0.5891,\ 0.2683,\ 0.1426]$; Average (up to 10^6): $[0.5011,\ 0.3336,\ 0.1653]$.

The *instantaneous* strategy continues to oscillate and does not settle at the NE, while the *average* strategy converges tightly to the NE $(\frac{1}{2}, \frac{1}{3}, \frac{1}{6})$. This matches theory: Regret Matching ensures vanishing external regret; when both players minimize regret, the time-averaged joint play approaches the set of correlated equilibria. In two-player zero-sum games, this implies the marginals converge to minimax/Nash strategies, so averages converge to NE whereas instantaneous strategies may keep fluctuating.

2 Part 2: Implementing MADDPG/IDDPG

1. **Why use slowly-updating target networks?**

   DDPG/MADDPG uses bootstrapping. The critic's regression target is

   $$y = r + \gamma(1 - \text{done})(1 - \text{terminated})\, Q_{\theta^-}\big(s',\, \mu_{\phi^-}(s')\big),$$

   where (θ^-, ϕ^-) are the *target* critic/actor parameters. If we used the online networks (θ, ϕ) instead, every gradient step would change the very target the critic is trying to fit, creating a *moving target*. This tight feedback (off-policy data + function approximation + bootstrapping) typically yields large oscillations or divergence.

   Target networks are updated by slow Polyak averaging,

   $$\theta^- \leftarrow (1-\tau)\,\theta^- + \tau\,\theta, \qquad \phi^- \leftarrow (1-\tau)\,\phi^- + \tau\,\phi, \quad \text{with } \tau \ll 1 \text{ (e.g., } 0.005\text{)},$$

   which makes y change slowly and thus *stabilizes* critic learning. A steadier critic then provides a smoother gradient for the actor as well.

2. **(bonus) Interpreting Fig. 1**

   (a) **Issue.** The learning curves exhibit high variance and occasional sharp drops early on, followed by gradual improvement. This indicates *noisy, unstable early learning* rather than clean monotonic improvement.

   (b) **Likely hyper-parameter and its role.** The pattern is most consistent with an increased *exploration noise* magnitude (e.g., a larger σ_{init} and/or slower annealing in the additive Gaussian noise). In MADDPG this noise is added to actors' actions during data collection to encourage exploration; making it larger or decay more slowly yields more stochastic actions, higher return variance, slower convergence, and occasional reward crashes even mid-training.

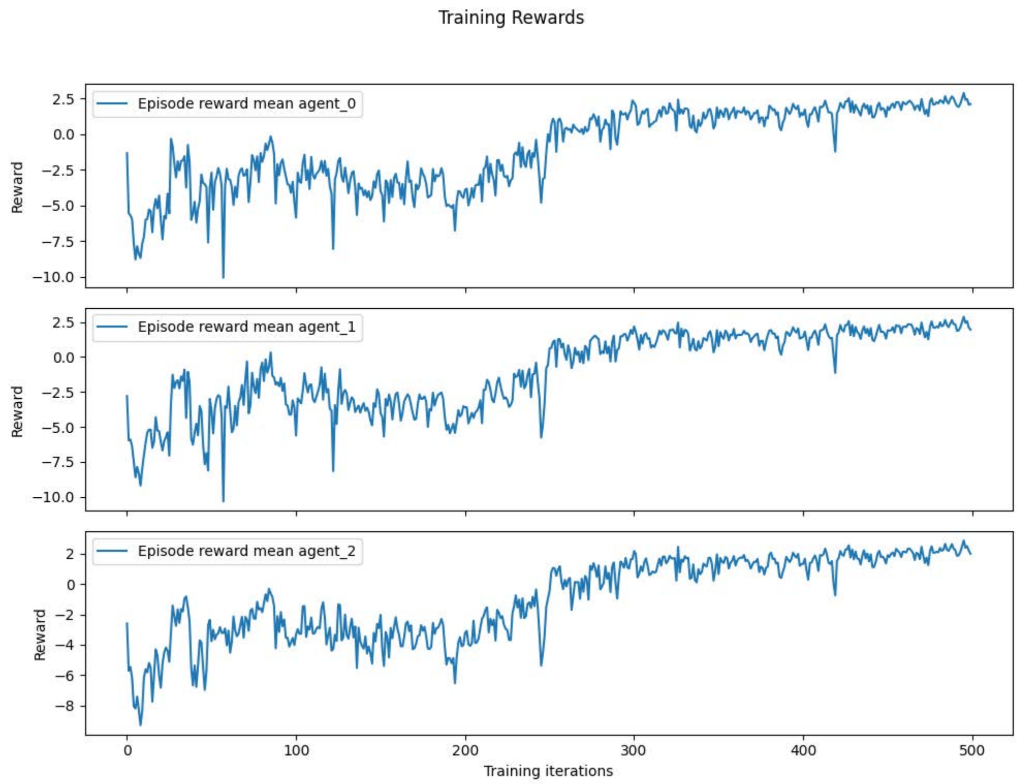

Figure 1: Agents performance after modifying a scalar hyper-parameter.

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