Learning Complex Group Behaviours in a Multi-Agent Competitive Environment

Kazım Sanlav
Department of Industrial Engineering

Afshan Nabi
Department of Molecular Biology and Genetics

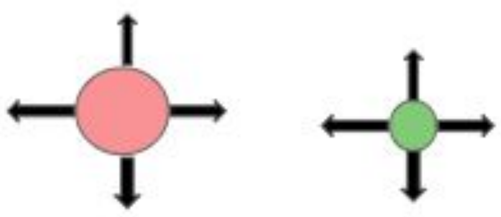
Bilkent University

Introduction

The aim of this project is to use reinforcement learning algorithms to simulate predator-prey dynamics, such as those seen when predator fish and prey fish interact in nature. In order to do so, we have adapted OpenAI's multiagent-particle-envs to simulate two competing groups of fish- a predator group (has a single predator) and a prey group (2 or 10 prey). We used 2 different Q-learning algorithms (Q-Network with Experience Replay & Q-Network with Experience Replay and Fixed Q-Target) and a Monte-Carlo Policy Gradient algorithm to train these predator and prey agents. The predator is rewarded based on the number of times they collide with ("eat") prey. The prey are rewarded negatively for each collision with the prey ("every time they are eaten"). Using such reward functions, the prey group is expected to learn defense strategies and the predator is expected to learn attack strategies.

Simulation Setup

Action Space



State

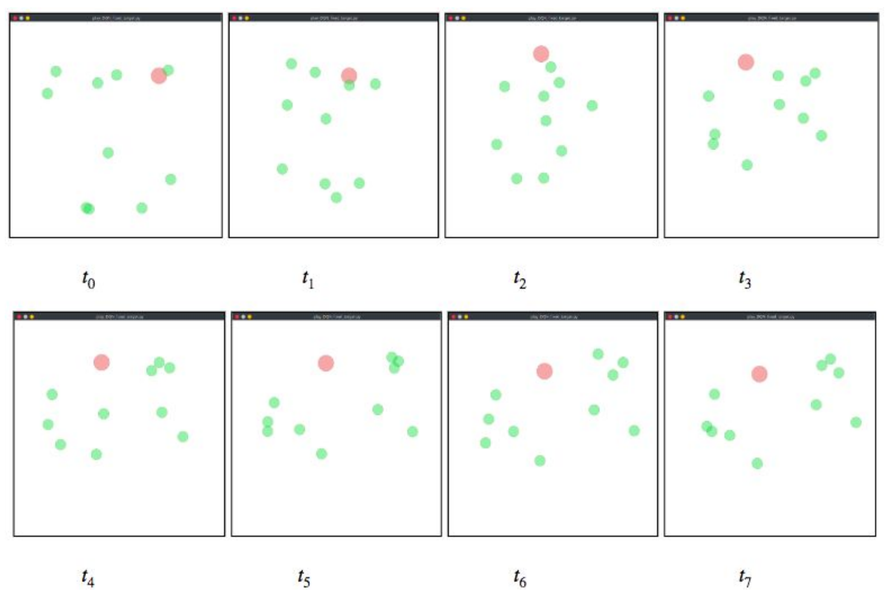
[$state_i$] for agent i is defined as:

$State_i$: [position of agent i , velocity of agent i ,
relative position of other agents,
relative velocity of other agents]

Reward

Predator: +10 for each collision & Boundary Restriction

Prey: -10 for each collision & Boundary Restriction



Methods

Algorithm 1: Q-learning with experience replay

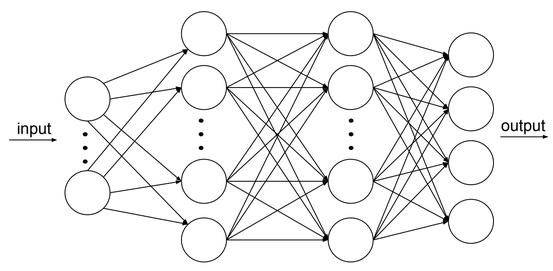
Initialize a replay memory to capacity N
Initialize action-value function Q with random weights θ
For episode 1:M do:
  Initialize environment to get initial state s_1
  For step 1:T do:
    With probability ε select random action a_t
    Otherwise select $a_t = argmax_a\, Q(s_t, a_t; \theta)$
    Execute action a_t and observe reward r_{t+1} , s_{t+1}
    Store experience $s_t, a_t, r_{t+1}, s_{t+1}$ in replay memory
    Set $s_{t+1} \leftarrow s_t$
    Sample a batch $s_j, a_j, r_{j+1}, s_{j+1}$ from replay memory
    Set target $Y_j = r_j + \gamma\, max_{a'}\, Q(s_{j+1}, a'; \theta)$
    Perform gradient descent step on $(Y_j - Q(s_j, a_j; \theta))^2$
  End For
End For

Algorithm 2: Q-learning with experience replay & fixed Q-target

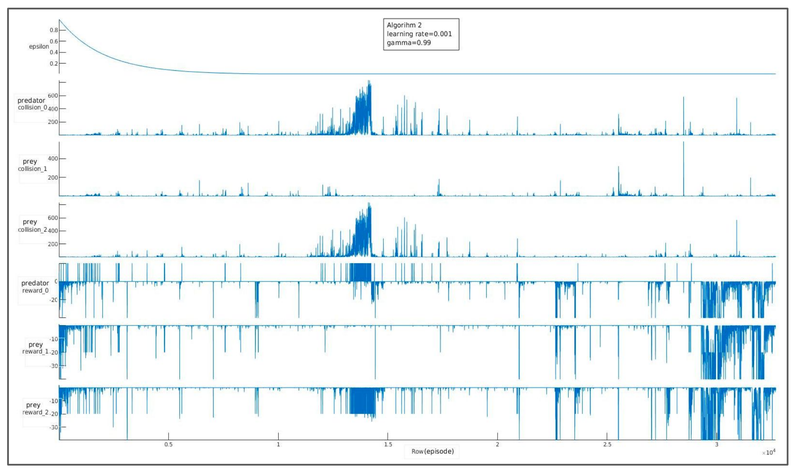
Initialize a replay memory to capacity N
Initialize action-value function Q with random weights θ
Initialize target action-value function $\hat{Q}$ with random weights $\hat{\theta}$
For episode 1:M do:
  Initialize environment to get initial state s_1
  For step 1:T do:
    With probability ε select random action a_t
    Otherwise select $a_t = argmax_a\, Q(s_t, a_t; \theta)$
    Execute action a_t and observe reward r_{t+1} , s_{t+1}
    Store experience $s_t, a_t, r_{t+1}, s_{t+1}$ in replay memory
    Set $s_{t+1} <- s_t$
    Sample a batch $s_j, a_j, r_{j+1}, s_{j+1}$ from replay memory
    Set target $Y_j = r_j + \gamma\, max_{a'}\, \hat{Q}(s_{j+1}, a'; \hat{\theta})$
    Perform gradient descent step on $(Y_j - Q(s_j, a_j; \theta))^2$
    Every C step update $\hat{\theta} = \theta$
  End For
End For

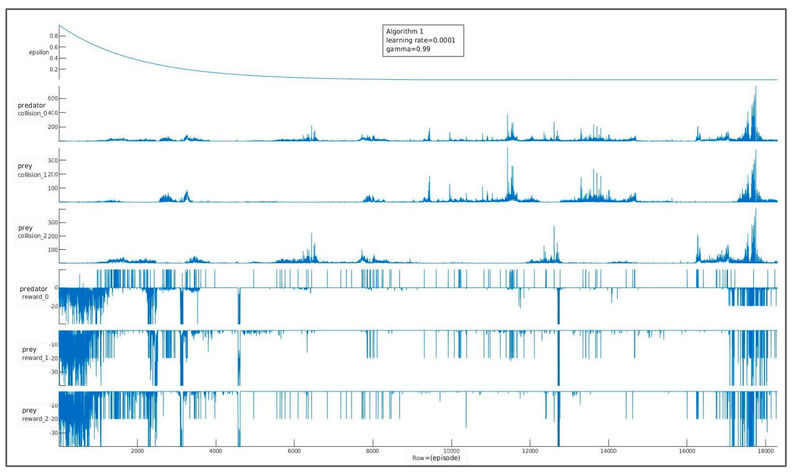
Algorithm 3: Monte-Carlo Policy Gradient

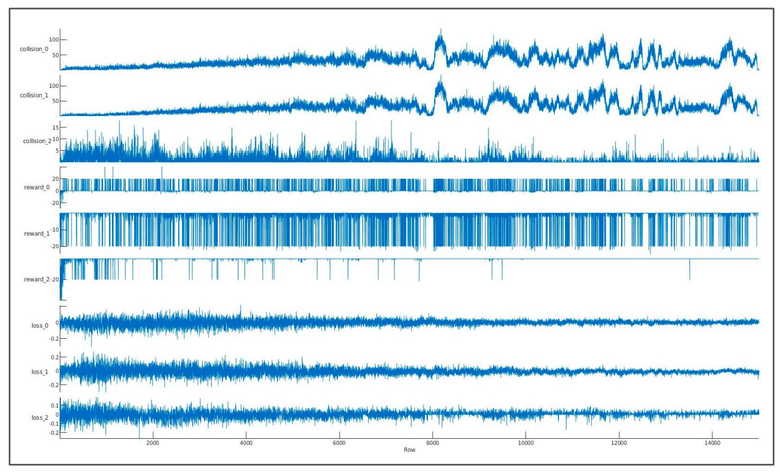
Initialize policy parametrization with random weights θ
For episode 1:M do:
  Generate an episode following π_θ
  For step 1:T do:
    Set $G \leftarrow \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$
    Set $\theta \leftarrow \theta + \alpha \gamma^t G \nabla ln\, \pi(A_t \mid S_t, \theta_t)$
  End For
End For



Results







Conclusion

In conclusion, we note that using reinforcement learning algorithms and simple reward functions for agents, it is possible for the prey agents to discover cooperative strategies that prevent the predator from "eating" them. It is also interesting to note that though the prey agents are the same in the beginning, as training proceeds, the same prey agents learn different strategies. While some are able to successfully evade predators, others are not. Such a concept is similar to the concept of "fitness" observed in real biological systems, where some individuals are able to survive better than others.