Learning to Solve the Tragedy of the Commons

The Evolution of Exploration in a Social Dilemma

B. Mintz, F. Fu (Dartmouth College)

AMS Contributed Paper Session on Operations Research, Game Theory, Economics, Information and Control
JMM, Seattle WA, January 2025

Table of Contents

1. Earlier work and Model
2. Results
3. Conclusion

What is R.L.?

Reinforcement Learning is a Machine Learning paradigm inspired by animal psychology that consists of an agent learning the optimal action depending on the state of an environment. The core idea is simple: actions with beneficial consequences will be repeated more.

Multi-Agent R.L.

Most work on R.L. focuses on a single agent learning some task. However in many applications there are multiple individuals, potentially with different objectives.

This further complicates R.L. as the number of agents increases. In particular, the environment becomes less stable, and coordination among agents becomes more difficult, [GD22].

Multi-Agent R.L.

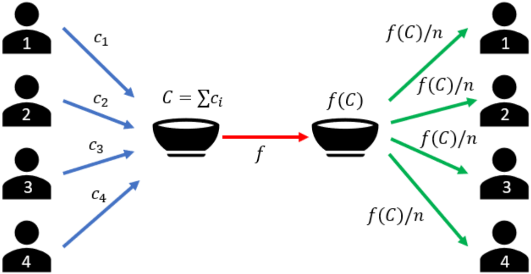
Most work on R.L. focuses on a single agent learning some task. However in many applications there are multiple individuals, potentially with different objectives.

This further complicates R.L. as the number of agents increases. In particular, the environment becomes less stable, and coordination among agents becomes more difficult, [GD22].

The problem we investigate in this work is **Incentive Alignment**: how agents with competing interests learn to work together despite individual incentives to defect.

Public Goods Games

Each of N individuals each contribute some amount c_i to a pot, which is scaled by some function $f(x)$ then distributed evenly among the players.



Free-Rider problem / Tragedy of the Commons: individuals benefit from contributing less, but this hurts the collective.

R.L. in this game

Reinforcement Learning has been used to study / explain human behavioral data through a variety of models, with varying levels of success, [AL04, Cot15, IIO^{+}03].

There has also been some theoretical work, e.g. [NP15, LM19].

Q-Learning

This is a form of temporal differencing, where agents try to learn the value of each action in a given state by keeping a table of approximate values.

Actions are chosen randomly with an exploration rate / temperature T, then the values are updated through

$$Q(s_t, a_t) = (1-\alpha)Q(s_t, a_t) + \alpha\left[r_{t-1} + \gamma \max_a Q(s_{t+1}, a)\right]$$

Q-Learning

This is a form of temporal differencing, where agents try to learn the value of each action in a given state by keeping a table of approximate values.

Actions are chosen randomly with an exploration rate / temperature T, then the values are updated through

$$Q(s_t, a_t) = (1-\alpha)Q(s_t, a_t) + \alpha\left[r_{t-1} + \gamma \max_a Q(s_{t+1}, a)\right]$$

This method has been proven to converge to the optimal policy given "sufficient" updates for each (state, action) pair, and decreasing learning rate, [WD92]. Further, it is shown to have bounded regret, that is, the difference between the actions it chooses and the optimal sequence is bounded, [LP22].

Stochastic Model

Groups often change or consist of agents with different experience levels. In this work, we expand the MARL framework to account for this, studying how the parameters effect pro-social behavior.

Inspired by Evolutionary Game Theory, we investigate population dynamics where agents reproduce and die.

Stochastic Model

Groups often change or consist of agents with different experience levels. In this work, we expand the MARL framework to account for this, studying how the parameters effect pro-social behavior.

Inspired by Evolutionary Game Theory, we investigate population dynamics where agents reproduce and die.

Our model consists of a population of N agents following reinforcement learning: stateless Q-learning with parameters α, γ, and T. Each time step, a group of k individuals is selected to interact, and receives payoffs from the public goods game based on the set of actions they choose. Then each individual has probability r of independently being replaced by another proportional to their fitness (averaged over interactions).

Evolutionary Analysis - Stochastic

To better compare these models, we mainly choose the group size equal to the population size, and the discount rate equal to zero.

Rather than study individuals evolutionary trajectories, we investigate the **fixation probability** among traits, their probability of replacing all individuals with the resident trait. These would allow us to simulate evolutionary trajectories, assuming the mutation rate is low enough that only two types occur at a time.

We simulate the stochastic model until only one trait remains. Repeating this many times allows us to estimate the fixation probability. This is compared the the neutral fixation probability of $1/n$ to see if the selection was positive or negative.

Translation to a Deterministic ODE system

Following [KG12], we consider stateless Q-learning with no discounting ($\gamma = 0$). The equation becomes

$$Q_i(t+1) = Q_i(t) + \alpha[r_i(t) - Q_i(t)] \tag{1}$$

where $r_i(t)$ is the average reward of choosing action i at time t. Under the Boltzmann Mechanism, the probability of choosing action i is

$$x_i(t) = \frac{\exp(Q_i(t))/T}{\sum_i \exp(Q_i(t))/T} \tag{2}$$

Taking the time derivative and rearranging and scaling time by α/T, we find

$$\frac{\dot{x}_i}{x_i} = \left[r_i - \sum_k x_k r_k\right] - T\sum_k x_k \ln\frac{x_i}{x_k} \tag{3}$$

Analytic Model

Let x be the probability of agent i contributing, after some algebra, the previous ODE reduces to

$$\dot{x} = x(1-x)\left(r_1 - r_2 - T\ln\frac{x}{1-x}\right) \tag{4}$$

Where the averaged rewards, when all agents follow the strategy x, are

$$r_C(x) = -1 + \sum_{i=0}^{k-1}\binom{k-1}{i}x^i(1-x)^{k-1-i}f(i+1) \tag{5}$$

$$r_D(x) = \sum_{i=0}^{k-1}\binom{k-1}{i}x^i(1-x)^{k-1-i}f(i) \tag{6}$$

This means the dynamics are governed by jumps $j_i = f(i+1) - f(i)$ in rewards when an additional player contributes

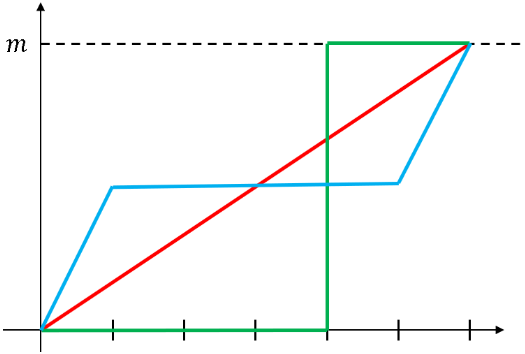
Evolutionary Analysis - Deterministic

In the analytic model, we consider a proxy of fixation probability known as **invasion fitness** $p(m, r) - p(r, r)$ where $p(m, r)$ is the payoff to the mutant if one individual follows the mutant strategy m while the rest follow resident strategy r.

Specifically, we use the payoff once the system reaches an equilibrium, assuming that interaction occurs for long enough this approximates the average payoff received.

The Optimal Reward Function

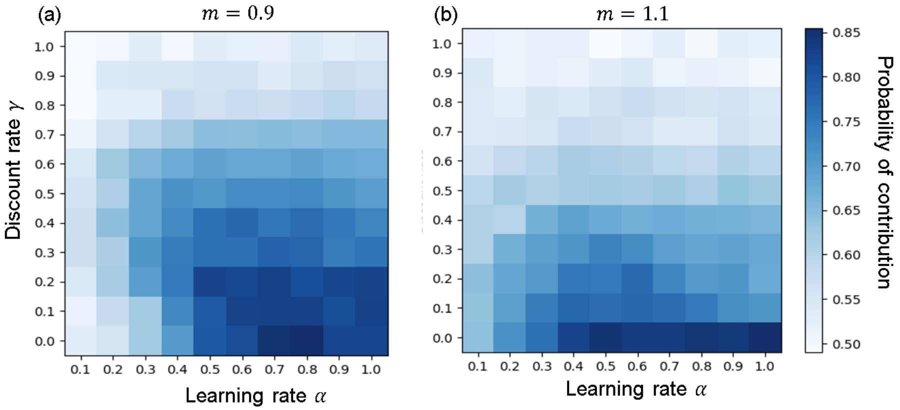
What reward function f leads to the highest levels of contribution?



More formally, we want to find the values of the jumps j_i that maximize the equilibrium that has the initial strategy of 0.5 in it's basin of attraction.

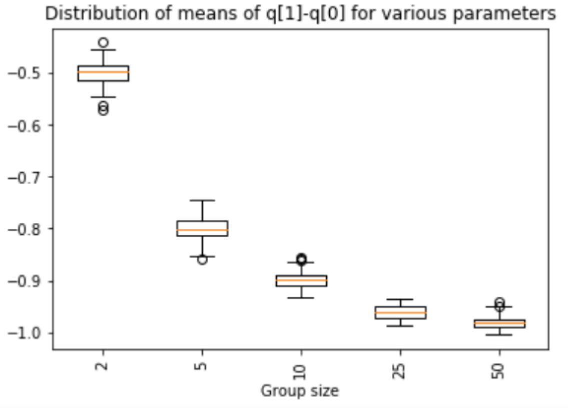
Table of Contents

Effect of learning parameters



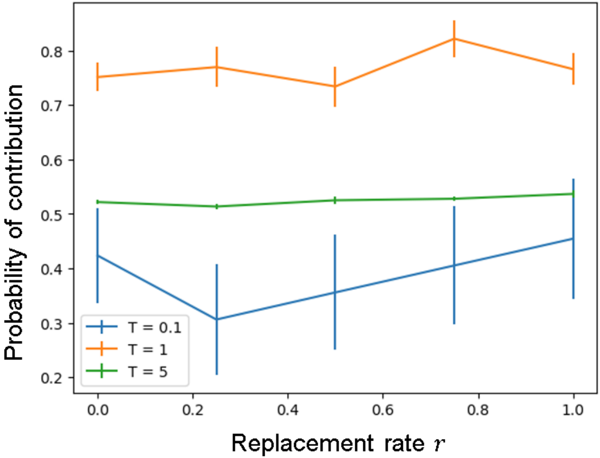
Here the reward functions are linear, $f(x) = mx$.

Group size



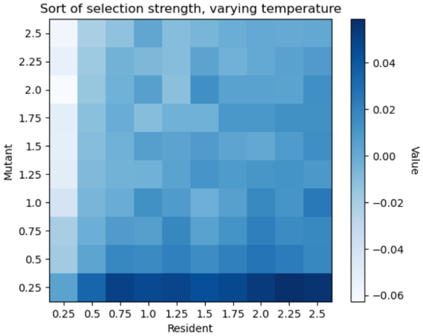
Smaller groups are more cooperative, consistent with the Free-Rider effect. This may be because the jumps shrink if the same reward function of percentage contributing is used.

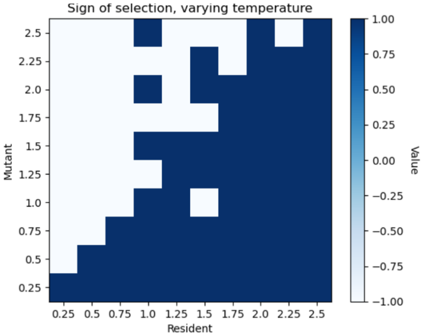
Replacement rate and Temperature



Temperature has a non monotonic effect on contribution levels, and replacement generally increases contribution.

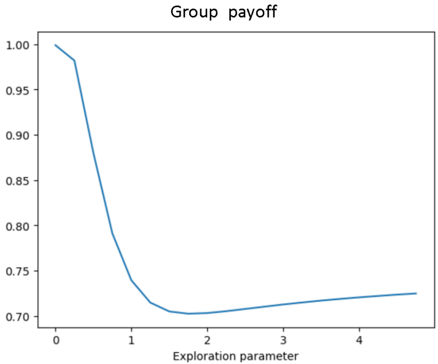
Invasion Dynamics - Stochastic

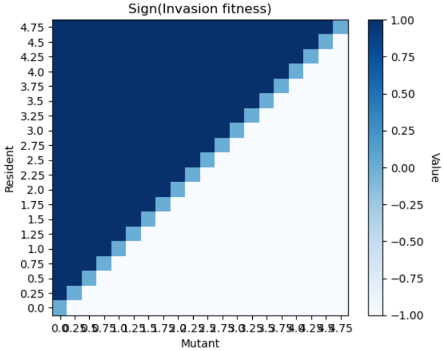

Sort of selection strength, varying temperature


Sign of selection, varying temperature

10,000 runs, $r = 0.01$, $n = 3$, rewards $= [0, 0, 0, 10]$. This took 14 hours to run.

Invasion Dynamics - Analytic

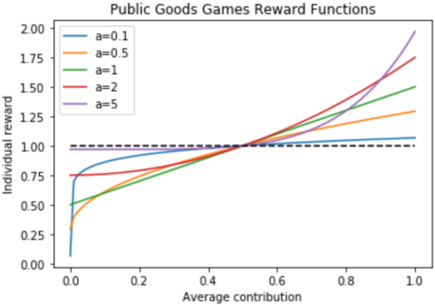


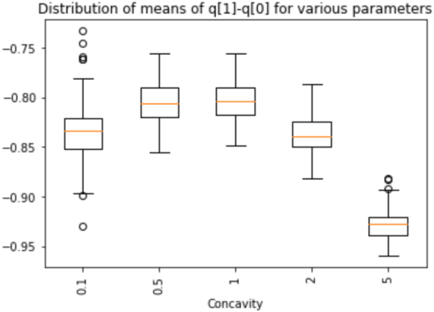


By varying the reward function, we found cases where the temperature evolved up, down, and to a stable value. The above plot uses rewards $[1, 1, 1, 1, 5]$. Since the system was solved numerically, there was also some nuance here.

Optimal Concavity

Here $f(C)/N = C^a + (1 - t^a)$.





Contribution levels are optimal for an intermediate level of concavity in the reward function. The threshold had little effect.

Optimal Reward Function

We seek to maximize the function $F(x, j_1, ..., j_n) = x$ subject to the constraints that ensure x is an equilibrium, that the $\sum j_i \leq m$ and that each j_i is non-negative:

$$\begin{aligned} 0 &= G(...) = \left[\sum_{i=0}^{n} j_i \binom{n}{i} x^i (1-x)^{n-i}\right] - T \ln \frac{x}{1-x} \\ 0 &= H(...) = \sum j_i - m + s^2 \\ 0 &= I^i(...) = j_i - t_i^2 \qquad \forall i \end{aligned}$$

here s and the t_i are auxiliary variables to allow for inequality constraints.

These are functions in the variables $x, j_1, ..., j_n, s, t_1, ..., t_n$.

Lagrange Multipliers

The equations for each component of the gradient are:

$$1 = F_x = \lambda G_x + \mu H_x + \sum_{k=0}^{n} \nu_k I_x^k = \lambda G_x \tag{7}$$

$$0 = F_{j_i} = \lambda G_{j_i} + \mu H_{j_i} + \sum_{k=0}^{n} \nu_k I_{j_i}^k = \lambda G_{j_i} + \mu + \nu_i \qquad (0 \leq i \leq n) \tag{8}$$

$$0 = F_s = \lambda G_s + \mu H_s + \sum_{k=0}^{n} \nu_k I_s^k = 2\mu s \tag{9}$$

$$0 = F_{t_i} = \lambda G_{t_i} + \mu H_{t_i} + \sum_{k=0}^{n} \nu_k I_{t_i}^k = -2\nu_i t_i \qquad (0 \leq i \leq n) \tag{10}$$

We solve this by cases. Note the first equations guarantees that $\lambda \neq 0$.

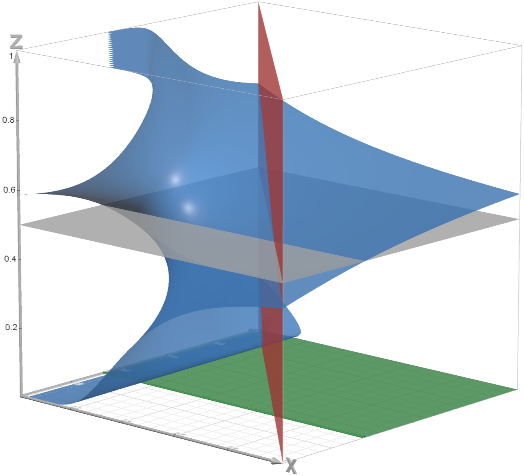
If $\mu = 0$, then any of the j_i partial derivative equations become $0 = \lambda G_{j_i} + \nu_i$. If $\nu_i = 0$, then $0 = \lambda G_i = \lambda \binom{n}{i} x^i (1-x)^{n-i}$. But as above, $\lambda \neq 0$, and x is not zero or one (these are trivially always solutions), so this solution is nonsensical. So $\nu_i \neq 0$, then $t_i = 0$, so $j_i = 0$ for all i. This too is nonsensical. Thus, $\mu \neq 0$, so $s = 0$, that is $\sum j_i = m$.

We may now handle the equations from the t_i partial derivative equations case wise. If any $\nu_i \neq 0$, then $t_i = 0$, so the corresponding $j_i = 0$. In particular, if all ν_i are nonzero, then all j_i are zero, which is again nonsensical. Further, if only one ν_k is zero, then only that j_k is nonzero, and since they all sum to m, it must be that $j_k = m$. There are then a multitude of cases where some subset of the j_i are nonzero.

Consider $n = 3$ and the case $\nu_1 = \nu_2 = 0$, but $\nu_0 \neq 0$, so $t_0 = 0$ and therefore $j_0 = 0$. By rearranging the equations from the j_i partial derivatives, we find $-\mu = \lambda G_{j_1} = \lambda G_{j_2}$. Since λ is nonzero we can divide by it to obtain $2x(1-x) = G_{j_1} = G_{j_2} = x^2$. This is solved when $x = 0$ or $2/3$. Similarly, solving for the other pairs yields a different x. In general, with larger n, once finds x is an intersection of some subset of these binomial terms $\binom{n}{i} x^i (1-x)^{n-i}$.

However, this doesn't match the solutions observed by playing around with the functions.

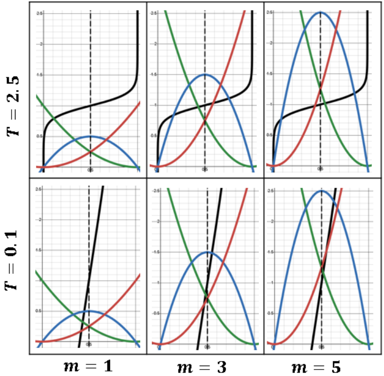
Equilibria surface ($m = 3$, $T = 0.1$)

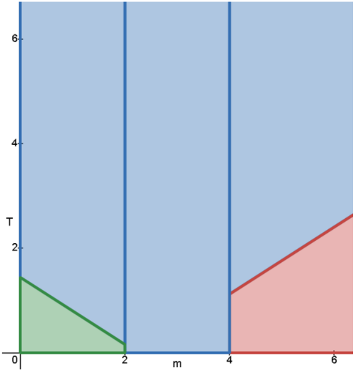


https://www.desmos.com/3d/zjlyuwootj

Classification over m, T-space

Just considering the vertex cases $[j_0, j_1, j_2] = [1, 0, 0], [0, 1, 0], [0, 0, 1]$, we can classify when each is optimal:





Intermediate values seem to only do better in the region in the bottom right of $2 < m < 4$.

Optimal Reward Functions - Stochastic

The intermediate reward functions seem to do optimally when the initial selection gradient is just barely above zero. This is fine in the deterministic setting, but poses a challenge in the stochastic model, since this means the initial strategy is a repellor of the learning dynamics, so could result in defection.

We iterated over all rewards functions by enumerating possible combinations of jumps. While the theoretical results often led to substantial cooperation, they were also often outperformed by nearby functions.

Another Game and periodic exploration functions

We also explored this framework in the Prisoner's Dilemma. It yielded similar results, but was constrained to simpler interactions with just two agents, so we focused on the Public Goods Game.

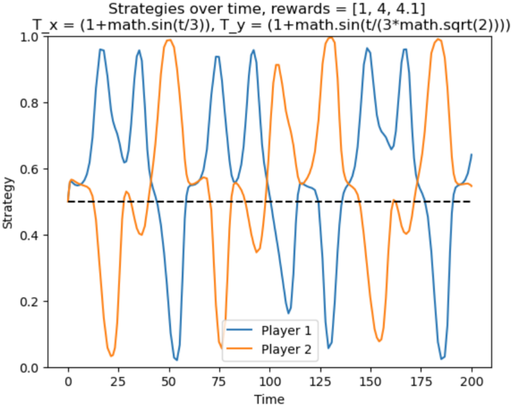


Table of Contents

1. Earlier work and Model
2. Results
3. Conclusion

Summary

- Reinforcement Learning is a powerful Machine Learning framework, but is not sufficiently understood, especially when multiple agents are learning simultaneously.
- Through stochastic simulations and analysis of a deterministic system of ODE's, we investigate group cooperation in a social dilemma.
- A variety of selection effects on the exploration rate can occur depending on the reward function. It can evolve up, down, or to intermediate levels.
- We theoretically and experimentally determine the reward functions leading to optimal levels of contribution, characterizing this with respect to the strength of the dilemma and randomness of the agents.

Next Steps

- Add a number of rounds, or continuation probability, parameter to tune group stability.
- Investigate other games, or allow continuous contribution levels in the Public Goods Game.
- Find a criterion for when exploration rates will evolve up or down, or some other possibility.
- Incorporate states into the model (this is a bit subtle, since the groups change frequently, making the environment unstable).
- Use more sophisticated learning methods, such as Frequency-Adjusted Q-learning or WoLF.
- Extend the analysis to groups with heterogeneous exploration rates.
- Optimize the reward function with more players.
- Compare the levels of cooperation to populations where agents don't learn (T=0).

What questions do you have?

My website above has these slides.

References I

Jasmina Arifovic and John Ledyard, *Scaling up learning models in public good games*, Journal of Public Economic Theory **6** (2004), no. 2, 203–238.

Chenna Reddy Cotla, *Learning in repeated public goods games-a meta analysis*, Available at SSRN 3241779 (2015).

Sven Gronauer and Klaus Diepold, *Multi-agent deep reinforcement learning: a survey*, Artificial Intelligence Review (2022), 1–49.

Atsushi Iwasaki, Shuichi Imura, Sobei H Oda, Itsuo Hatono, and Kanji Ueda, *Does reinforcement learning simulate threshold public goods games?: a comparison with subject experiments*, IEICE TRANSACTIONS on Information and Systems **86** (2003), no. 8, 1335–1343.

Ardeshir Kianercy and Aram Galstyan, *Dynamics of boltzmann q learning in two-player two-action games*, Physical Review E—Statistical, Nonlinear, and Soft Matter Physics **85** (2012), no. 4, 041145.

References II

Olof Leimar and John M McNamara, *Learning leads to bounded rationality and the evolution of cognitive bias in public goods games*, Scientific reports **9** (2019), no. 1, 16319.

Stefanos Leonardos and Georgios Piliouras, *Exploration-exploitation in multi-agent learning: Catastrophe theory meets game theory*, Artificial Intelligence **304** (2022), 103653.

Heinrich H Nax and Matjaž Perc, *Directional learning and the provisioning of public goods*, Scientific reports **5** (2015), no. 1, 1–6.

Christopher JCH Watkins and Peter Dayan, *Q-learning*, Machine learning **8** (1992), 279–292.