

Submitted in part fulfilment for the degree of BEng

Applying Coevolutionary techniques to Competitive self-play Reinforcement Learning to better solve non-stationary problems

Tom Clark

Click or tap to enter a date.

Supervisor: Dan Franks

Acknowledgements

Click or tap here to enter text.

Click or tap here to enter text.

Table of Contents

Table of Contents

[Executive Summary 6](#_Toc100571105)

[1 Introduction 1](#_Toc100571106)

[2 Literature Review 2](#_Toc100571107)

[3 Methodology 5](#_Toc100571108)

[3.1 The Minimal Substrate 5](#_Toc100571109)

[3.2 Objective fitness and subjective fitness 6](#_Toc100571110)

[3.3 Experimental Setup 7](#_Toc100571111)

[4 Experiments and Results 11](#_Toc100571112)

[4.1 Experiment 1 – Loss of gradient 12](#_Toc100571113)

[4.2 Experiment 2 – Focusing 16](#_Toc100571114)

[4.3 Experiment 3 – Relativism 18](#_Toc100571115)

[5 Conclusion 21](#_Toc100571116)

[Appendix A 22](#_Toc100571117)

[Appendix B 23](#_Toc100571118)

[Bibliography 24](#_Toc100571119)

# Executive Summary

ABSTRACT HERE

# Introduction

Reinforcement Learning (RL) is a method of machine learning that maximises an agent’s objective function’s performance within an environment, by learning the optimal parameters for the agent within that environment.[1] The environment includes everything outside the agent and is usually presented in the form of a Markov Decision Process (MDP). An MDP takes 5 inputs; the finite state space S, the finite action space A, the transition function T : S x A x S -> [0,1], the reward function R : S x A x S -> R and the discount factor γ ∈ [0,1].[2] An MDP attempts to find a policy π, a function that fully defines the action of an agent whilst in a specific state, that maximises the reward/fitness function.

However, most MDPs used in RL find a stationary policy, a policy that is independent of time. This is a key weakness when they are applied to a non-stationary environment; an environment which changes over time, because now the expected value for each stationary policy can vary.[3] One such example of a non-stationary problem is the simple two player game of Rock-Paper-Scissors, played in a repeated format. Here, two agents will play against one another and attempt to maximise the number of times they win. An ideal policy would learn their opponent’s preference and attempt to counter that by changing its own policy, which prompts the opponent to adjust accordingly, introducing non-stationarity due to changing behaviours.[4]

One such area that where non-stationarity might be introduced into problems is in competitive multi-agent environments. In a single agent environment, there is only a single agent operating in that environment, whereas in a multi-agent environment, multiple agents may be operating and competing within the environment. A popular method of training agents in a competitive multi-agent environment is through competitive self-play RL.

Competitive self-play RL is a type of multi agent RL (MARL) that has an agent play against another agent or agents, or a copy of itself, and judging its fitness off its performances in the matches. The environment for each “match” would be the opposing agent or agents and any other variables[5]. As described above, non-stationarity is evident as agents build up a type of “memory” as they play against other agents and attempt one-up each other.

There exists a separate and much older area of machine learning known as Competitive Coevolution, hereby referred to as Coevolution, that has remained largely ignored by RL research. Coevolution is a type of Evolutionary Algorithm (EA) which was first introduced in 1990 [6] that tasks individuals within a population of agents to compete against each other, thereby measuring their performance with respect to the rest of the population.[7] As in other EAs, agents from the population will then be selected, mutated and cross-bred to produce new “offspring” agents for the next generation. The general aim of Coevolution is to create an “arms-race”, whereby agents in a population will progressively one-up the others, producing an adaptive gradient that is difficult to engineer manually.[8]

The similarities between Coevolutionary algorithms and competitive self-play RL are self-evident; each creates an optimal agent through iterative generations and evaluates agents by having them compete against other agents. Coevolution however holds an innate advantage over RL in non-stationary environments as Evolutionary Strategies do not rely on stationary transition probabilities, whereas MDP-based RL methods do. [9]

Research has been done into the co-application of MARL and Coevolution with a view to create scalable learning techniques.[9] The paper used Atari games as a benchmark for their Coevolutionary Algorithms and compared them against a Deep-Q Network (DQN) variant which learns with RL. Their results showed that the Coevolutionary algorithms can be superior to the DQN on several of the games and hypothesised that the multi-agent aspect of the problems introduced non-stationarity, which as mentioned above is no issue for Evolutionary Strategies but can cause problems for MDP based solutions like DQNs.

This leads to the conclusion that there are some advantages for Coevolutionary learning over RL in certain environments. However, the method of Coevolution can also introduce more problems than it solves. Here I shall outline 3 major problems that exist in Coevolutionary learning techniques.

The first is the problem of relativism which introduces cycling, whereby the agents may cycle over and over through a set of “optimal” agent types which lack transitivity, e.g. A<B<C, but C<A, leading to a cycle over these 3 types without any progress. This problem breaks the adaptive gradient and is a result of the measure of the agents only being relevant to the other agents in the environment and not against any overarching objective or goal.

Another problem is a loss of gradient called disengagement where one set of the population may dominate all others, thereby rendering it impossible for any meaningful change to be measured. Disengagement (also called disconnection) occurs in Coevolution the Coevolutionary system decouples; when each population no longer “perturbs the other”, which results in the loss of feedback between the populations and the inability to compare populations’ fitness against one another.[10] This may occur in asymmetrical environments, for example it may be easier for prey to evade a predator than it is for the predator to catch the prey, or in symmetrical environments, if one population has learnt techniques that dominate the competition.

Lastly is a loss of focus called focusing, where agents may focus entirely on an opponent’s weakness thereby leading to agents which don’t play to win the game, but to win against that specific opponent, creating weak, non-adaptive agents when challenged by other opponents.[11]

There lies the motivation for this paper; do RL techniques hold advantages over Coevolutionary techniques by avoiding these 3 pitfalls? In order to investigate this, I will be using a coevolutionary paper[11] which both examined and outlined the 3 problems from above in a Coevolutionary context. I will then use the same environment that they introduced, a minimal substrate, but instead apply a RL technique to the agents that compete within the minimal substrate environment. Finally, I will examine the results and see whether they provide evidence that the 3 problems also manifest in RL.

# Methodology

## The Minimal Substrate

In this section I will outline the minimal substrate that will be used in the experiments. This minimal substrate is based on Richard A. Watson and Jordan B. Pollack’s minimal substrate.[11]

The goal of each agent is to maximise a, a scalar value it possesses. This maximisation however will use a fitness function, that returns a value for ’s fitness with respect to a set of other numbers, . contains many other agents all with their own value. I will let count the number of members of that beats:

Eq1

Where = 1 if , 0 otherwise.

Therefore, instead of a trivially easy game of maximising by itself, will be evaluated and increased with respect to all other agents, which will also be increasing.

The second component of the minimal substrate is a second dimension, such that each agent has two scalar values on two dimensions, and . This second dimension is an abstraction of the strengths of each agent, as in more complex implementations of machine learning, agents may possess different strengths and weaknesses. The introduction of this second dimension means that we cannot reduce both agents to a single dimension and compare them, e.g. , as for some opponents, the value of dimension  may be more important that dimension .

As with the original minimal substrate, I will be comparing two agents with regards to whichever dimension is furthest apart.

Eq2:

where

Overall, the goal of each agent is to maximise all its dimensions. There exist, however, potential pitfalls for agents to fall into, such as over-specialising on a single dimension whilst the other remains small, in order to maximise its subjective fitness; its fitness compared to all other agents in the environment. This effect appears in Coevolution, and part of the experiment is to see if it manifests in RL as well.

The original minimal substrate paper also includes a third equation to add intransitive superiority. That is, that in a random game with 3 agents, A, B and C, agent A beats B, B beats C, but C beats A. This creates the scenario where A’s superiority over B doesn’t intransitive, and doesn’t automatically allow it to beat C. This is a key concept in coevolution and can result in algorithmic failure as agents cycle over and over.[13] As with the original minimal substrate, I will be modifying Equation 2 so that the dimension compared between 2 agents will be the dimensions in which the players are closest.

Eq3:

where

I will use the original minimal substrate paper’s example to demonstrate the intransitive superiority introduced by Equation 3 here with 3 agents, a=(1, 6), b=(4, 5), c=(2, 4). a beats b as they are closest in the y dimension, and , b beats c as they are closest in the y dimension and , but c beats a as they are closest in the x dimension and . Overall however, the optimal agent is still one which is maximal in both dimensions and beats all other agents.

## Objective fitness and subjective fitness

I shall be using the terms of objective fitness and subjective fitness as described in the original minimal substrate paper. The objective fitness for an individual would be to have the maximum scalar value possible, across all their dimensions. As such, the best individual possible would have the best objective fitness possible too. The reward of an agent within a MARL environment such as the minimal substrate lines up well with the subjective fitness; it is the fitness of an individual in comparison to all other individuals in the environment. As such, an individual with a very high objective fitness might have a low subjective fitness, if all the other individuals in its environment also have high or even higher objective fitnesses.

This difference between objective and subjective fitness is part of the reason for the loss of gradient problem, as described early and as will be shown in Experiment 1.

## Experimental Setup

Whilst I shall use the previous equations as defined above, there are several representation and RL setup decisions to be made.

I will be representing the Environment with a custom class, which contains a list of N agents, where the agents are made from another custom-built class. The agents have an array to store their dimension scalars in. When the Environment class has its reset method called, it initialises the N length array of agents, and initialises their dimensions as random integers between 0 and 10.

The Environment class contains methods to evaluate the agents using either Equation 1, 2 or 3 and calculates their reward. When evaluating agents, if the dimensions between two agents are equidistant, I will default to comparing dimension 0. If two agents compare a dimension, and find it is equal, neither will get a bonus to their overall reward. This is to eliminate any possibility of emergent cooperative behaviour.

The Environment class also makes use of OpenAI’s gym module’s Discrete class for representing the observation space and action space. I’ve chosen to represent the observation space, the number of states an agent can be in, as a set of 3 \* number of dimensions many possible states. First, the Environment class calculates how many other agents are greater than the evaluated agent in dimension i, how many are equal in dimension i, less than in dimension i, etc for all dimensions. The state is then the index of whichever one of these numbers is greatest: e.g. if most less than the agent in dimension 1, the state will be 2.

There are two possible directions to take the action space, the set of actions available to agents. Design A uses the original minimal substrate paper’s scalar representation; the scalar in each dimension was represented as a “fixed length binary string and the value they represent is given by the unitation of the string” where the string length was 100. This choice was so that there was a biased mutation, i.e. instead of simply using an integer representation, where mutation may only increase or decrease by a set amount, with no bias either way, whereas by using a binary string representation gives a mutation bias; a string with half 0’s or more is more likely to increase, and a string with half 1’s or more is more likely to decrease. This mutation bias is important in real applications of more complex problems and shall be used for this design. Design A also has 1 + number of dimensions many possible actions: flip a bit at a random index in dimension i or take no action. This design is very similar to the coevolutionary representation in the original Minimal Substrate model and represents individuals mutating randomly but doesn’t consider the nuances of RL.

Design B uses integer scalars, with a maximum value of 100 and minimum of 0 and has 2 \* the number of dimensions many possible actions: increase dimension i by a scalar Z, decrease dimension i by Z etc for all dimensions. Z is a random integer between 1 and 3. The purpose of having Z as a random integer rather than a set integer is to allow agents to “catch up” to agents that are ahead, rather than all increasing by a set value and therefore being uncatchable if both are racing to the maximum scalar value. Attempting to increase a dimension’s scalar over 100, the maximum scalar value, or decrease it below 0 will set it to 100 or 0 respectively and is essentially a “skipped” action. This design better accommodates RL’s policy-based strategy evolution rather than selecting based off of random mutations, as agents have more control over increasing or decreasing their dimensions.

Both these designs use a discrete action space with a pre-defined set of actions, instead of a continuous action space where actions are real valued vectors. A continuous action space implementation would possibly be an even better representation and allow the dimensions’ scalars to change by uncapped amounts, but I have chosen not to take this approach as it will add extra complexity to the whole experiment without providing any extra exploratory value to the experiment, and it would possibly add an extra computational burden can’t be afforded on the hardware I am using. In conclusion, I will only be using Design B. I made this decision because the random nature of evolution doesn’t translate well into policy making; an individual choosing to mutate dimension 0 isn’t making a choice about whether it wants to increase or decrease that dimension, only that it wants to change it. Design B on the other hand provides this choice. I also have created an extension of Design B that is called Design C. Design C is exactly the same, except has 2 populations of individuals and 2 q tables. In Design C, all agents compete against each other, but agents from population 1 share and update q table 1, and also get their actions from it as well. This Design will be used to examine loss of gradient, so it can be better compared to the original paper’s plots.

The Environment contains a step method, that is given a list of N actions in the form of a list of integers, with the ith action being intended for the ith agent. The Environment applies each action to the intended agent, then calculates the rewards with the chosen equation, which is stored in an N length list, so that the reward for the ith agent is the ith item in the rewards list. The same is done for the observations for each agent; they are calculated and stored in an N length list. The Environment then increments a current generation counter, which keeps track of how many times step has been called since the Environment was last reset. If the current generation counter equals the max generation variable, which I have set to 100, then the Environment sets its done flag to True. After the step method is complete, it will return the list of observations, rewards and the done flag. When the reset method is called, it will calculate and return an observation list as well; this is the list of initial states for each Agent before any actions have been taken.

I decided upon a decentralised Q-table approach for the RL part of the program, decentralised meaning that each agent has a separate Q-table. The Q-table calculates and keeps track of the expected rewards for actions at each state, so that the overall policy, the strategy, is improved over time.

To implement this, I created a list of length N, which contains N Q-tables, where the ith Q-table belongs to the ith agent. Each Q-table is initialised as a 2d array full of 0’s, with a shape of (3\*number of dimensions, 2\*number of dimensions), 3\*number of dimensions being the observation space, and 2\*number of dimensions being the action space, where the number of dimensions is the number of dimensions the agents have. For the ith agent, the Q-value for action A whilst in state S will be at index [i][S][A]. This is all contained within a Q Network class I created and done at initialisation. I also chose hyperparameters of: 0.1 for the learning rate α, 0.6 for the discount rate γ and 0.2 for the greed value ε.

The Q Network then runs for M iterations. At the start of each iteration it resets the environment and sets the current state to the returned observation. The environment then loops through until the max generation is reached. At each cycle of the loop, a set of N actions is decided for each agent. There is a random ε chance that a random action is chosen, otherwise, so long as the Q values are not all 0 at index [i][S] it will select the index with the highest Q value, with that index corresponding to the chosen action, A. If the Q values are all 0 it picks a random action. This prevents the network having a very strong natural bias to whatever action corresponds to index 0, as a quirk of choosing the maximum Q value is that it defaults to index 0 if all Q values are equal.

The actions are then passed to the Environment’s step method, which returns the observations and rewards, as well as the done flag. The Q values for each agent is then updated using the formula

Where

And returns the index of the maximum Q value of the list of Q values at index [i][observation] in the Q table, for the ith agent. The state is then set to the observation, which will carry over for the next cycle.

My source code can be accessed on GitHub[[1]](#footnote-1).

# Experiments and Results

First, an explanation for how plots were generated and how to read them. Polygon

Description automatically generated

Figure 1 - Plots of Subjective Fitness (Rewards) and Objective Fitness (Scalars) for Design B with only one dimension

Figure 1 shows the plots generated after training has taken place. The rows 1 and 3 plot the average subjective rewards of the individuals against the generations, and rows 2 and 4 plot the average of the 2 scalar values against the generations. The plots are all colour coded and colour synced; in the top left, the green line shows the average reward for individual Z, and the plot below it has a green line that shows the average scalar value for its sole dimension. Each graph shows a maximum of 3 individuals to make it easier to read.

In Design B, the values on the Y axis that are plotted are the average for each individual over all iterations. During the training, the subjective rewards and scalar values of every individual are stored, along with the generation counter. To calculate the average subjective reward, the plot then sums up individual Y’s subjective reward over all iterations when the generation counter equals 0, then divides by the number of iterations, which generates the average reward at generation 0, this is then repeated for every other generation for individual Y, then for every other individual. The same is done to calculate the Objective Fitness from the average scalar values.

Chart, line chart

Description automatically generated

Figure 2 - a Design C plot where agents have only one dimension

In Design C the plots are different, as show in Figure 2. The two lines represent the average of all individuals from each population. The red line would represent population 1, and the green line population 2.

The Y axis values are calculated from the average subjective reward and average scalar value for each population, which in turn is the average of all the average subjective rewards and scalars from every individual belonging to that population.

## Experiment 1 – Loss of gradient

This experiment was performed with Equation 1, and thus all agents have only a single dimension with one scalar. The purpose of this experiment is to show that using the subjective fitness does not necessarily result in the creating the best individuals with a high objective performance, i.e. the highest scalar value.

First, I trained Design B with 25 individuals over 50 iterations.

Polygon

Description automatically generated

Figure 3 – Average Subjective and Objective Fitness over all iterations for Design B with Equation 1

Polygon

Description automatically generated

Figure 4 - Subjective and Objective Fitness in the final iteration for Design B with Equation 1

Figure 3 shows that the average subjective fitness of all individuals eventually decreased to around ~8 or ~6, even though the average scalar value for all individuals climbed to over 80 for all individuals except one. After generation ~80 most individuals reached their highest average scalar value, and after this point the subjective fitness didn’t vary significantly for any agent.

Figure 4 shows only the final iteration. 3 individuals had their subjective fitness fall to 0 or close to 0. In turn, these individuals’ scalar values all stagnated and didn’t increase again significantly. The subjective fitnesses for all the other individuals also decreased over time, showing the loss of gradient as even though their objective fitness increased, they weren’t any better off individually as all the other individuals also increased their objective fitnesses.

Next, I trained Design C for 50 iterations, using 2 populations which had 25 individuals each.

Chart, line chart

Description automatically generated

Figure 5 - Average Subjective and Objective Fitness over all iterations and individuals, for two populations using Design C with Equation 1

Figure 5 again shows that after the average scalar value hit its maximum at around generation 80, the subjective reward did not vary significantly.

Finally, I ran Design B again for 50 iterations but with only 2 individuals.

Chart, line chart

Description automatically generated

Figure 6 - Average Subjective and Objective Fitness over all iterations for Design B using Equation 1 and only 2 individuals

Figure 6 shows a different result, here the teal individual dominated the red individual, and teal climbed to ~80 for its scalar value whereas red maxed at ~40. The subjective rewards graph also shows that the reward for teal was 1 – the reward for red, and vice versa. Taking this plot at face-value gives the wrong conclusion however, and instead plots of single iterations are required.

Chart, line chart

Description automatically generated

Figure 7 - Subjective and Objective Fitness of the penultimate iteration for Design B using Equation 1 and only 2 individuals

Chart, line chart

Description automatically generated

Figure 8 - Subjective and Objective Fitness of the final iteration for Design B using Equation 1 and only 2 individuals

Figure 7 shows the penultimate iteration, and Figure 8 shows the final iteration. From these two plots, the true result can be found: whichever individual started with the higher initialisation ended up dominating the other. This is because there is no subjective reward gradient in this 2-agent environment, either the individual is winning or it is losing; there is no way to tell how much it is winning by or losing by, so therefore the losing individuals never learn how to come back from losing, whilst winning individuals continue to reward increasing their scalar value.

To conclude Experiment 1, Figures 3, 4 and 5 have shown that a loss of gradient can occur through the subjective fitnesses of those individuals or populations of individuals not improving significantly after a certain point, because even though the individuals themselves have a higher objective fitness, their scalar value, they are not any better in comparison to the other agents in the environment, as all the other agents have also improved proportionally. This is known as the Red Queen Effect[13], and has been previously documented in Coevolutionary research.

Additionally, a slightly different type of loss of gradient can be seen in Figures 7 and 8, one caused by disengagement.

Disengagement is discussed in the original minimal substrate paper, which states the effect is only seen with low population sizes or low sample sizes (the number of individuals an individual is compared against). An extreme form of disengagement can be seen in Figures 7 and 8, as one individual dominates the other one, resulting in the dominant individual being unbeatable and providing no opportunity for the dominated individual to learn to beat it.

## Experiment 2 – Focussing

In this experiment I will be using Equation 2. The purpose of this experiment is to demonstrate the problem of a loss of focus and over focusing on weaknesses in opponents.

First, I ran Design B with 25 agents for 50 iterations, where every agent had 2 dimensions with a maximum value of 100. The objective fitness here is the sum of both dimensions.

Diagram

Description automatically generated with medium confidence

Figure 9 - Average Subjective and Objective Fitness over all iterations for Design B with Equation 2

Next, I repeated the previous experiment by using Design B with 25 agents for 50 iterations, but this time every agent had 10 dimensions where each dimension has a max scalar value of 50, so that the max scalar value can be reached in earlier generations. The objective fitness is the sum of all 10 dimensions.

Diagram, shape

Description automatically generated

Figure - Average Subjective and Objective Fitness over all iterations for Design B with Equation 2, where agents have 10 dimensions

Diagram

Description automatically generated

Figure - Subjective Fitness and Scalar Values in the final iteration for Design B with Equation 2, where agents have 10 dimensions

In both Figures 9 and 10, every individual failed to even come close to the maximum possible objective fitness values of 200 and 500 respectively. The issue here is that only a single dimension is relevant in comparisons, and only a single dimension can be increased at a time. The consequences of which can be observed in Figure 11.

In Figure 11, most individuals will increase a single dimension to the maximum value of 50 in an attempt to capitalise on every other individual’s weakness; that the most distinct dimensions are the ones compared, which allows individuals which increase a single dimension to the maximum to use that dimension as the comparing dimension, ensuring victory. This leads to very brittle and non-generalist individuals which wouldn’t compete well against other individuals trained in other environments with different individuals. Some individuals do attempt to generalise and increase their other dimensions as well, and these individuals are the ones which maintain a high subjective fitness.

One key difference between this implementation and the original minimal substrate paper’s implementation is that the original had a mutation bias that causes dimensions to naturally drift towards half of the maximum scalar value. This is missing from my implementation as RL doesn’t have mutation biases, as agents develop and learn differently. As such, individuals do not “forget” previously learned skills (such as a dimension being increased to the maximum, but then drifting back to the midpoint of 25). Despite this difference however, the issue of focusing can still be observed.

“Hall of fame” solutions to increase generalisation are present in reinforcement learning[12] as well as coevolution and using a hall of fame that is used in calculating the reward of individuals, where the hall of fame contains individuals from previous and different testing cycles, along with a sufficient number of iterations and a high enough environment maximum generation number or low enough max scalar value, I believe that the Q-network would likely be able to produce an individual that reaches the maximum value in all dimensions.

## Experiment 3 – Relativism

In this experiment I will be using Equation 3. The purpose of this experiment is to demonstrate the problem of relativism.

First, I ran Design B with 25 agents for 50 iterations, where every agent had 2 dimensions with a maximum value of 100. The objective fitness here is again the sum of both dimensions.

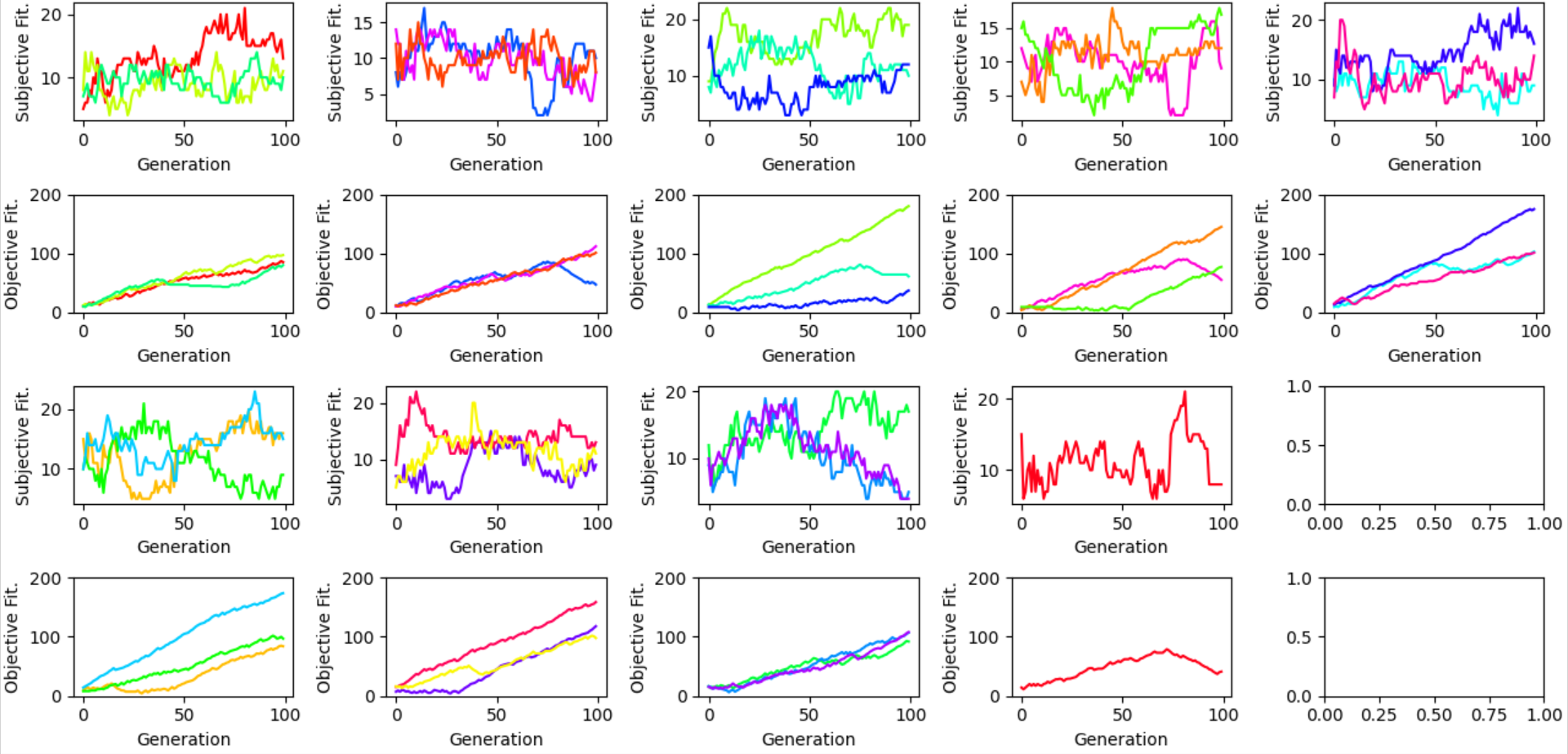


Figure 12 - Subjective and Objective Fitness of the final iteration for Design B with Equation 3

Most individuals increased their objective fitness over time in a gradual curve, top right’s blue individual is a good example of one such individual which increased both dimensions to high scalar values and was both subjectively and objectively a good individual. Other individuals, such as row 3/4 column 4’s red individual began to decrease its objective fitness around generation 60 and its subjective fitness spiked massively to 20. Several other individuals also expressed this behaviour of decreasing their objective fitness and being rewarded with a higher subjective fitness. This is a good example of both loss of focus and relativism as defined by the original minimal substrate paper, these individuals have focused on the weaknesses of other individuals (failing to increase both scalar values), and as the reward of the individuals is defined in comparison to other individuals, they have introduced variation that has lowered the overall standard of play and subverted the game, whilst also guaranteeing that those individuals that have decreased their scalar values will maintain low scalar values and not seek to improve.

To further investigate this, I loaded the Q table generated from Figure 12’s Q network into a new network which I then tested in an environment where all the parameters were identical, but instead ran for 300 generations, and for only a single “iteration” (as there is no additional training being done to the Q network).

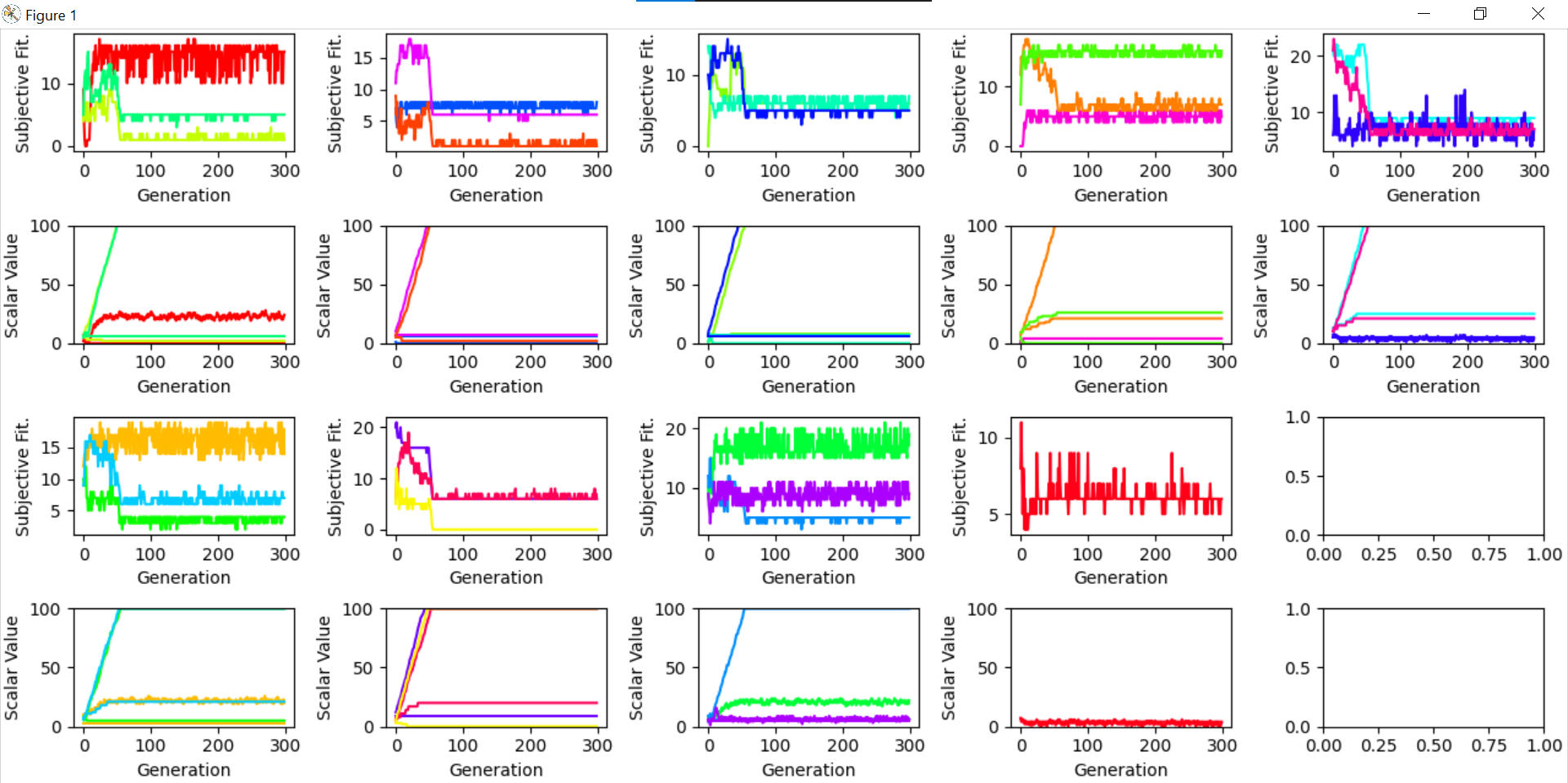


Figure - Subjective Fitness and Scalar Values using Figure 12's Q network tested for 300 generations

Row 1/2 column 2’s orange individual and row 3/4 column 4’s red individual both had interesting results. Orange increased one scalar value to the maximum as soon as possible and left the other scalar value near its starting value. Red on the other hand kept both scalars low but adjusted them many times. Objectively, orange is much better as its objective fitness would be over 100, whereas red would be less than ~30, however subjectively, red performed far better and hovered around ~6 on average, whereas orange dropped and stayed at 4.

Subjectively, the best individual of all is row 3/4 column 3’s green individual. This individual maintained an average subjective fitness of ~19, meaning that on average it beat 19/24 of the other individuals, whilst its objective fitness was around ~26 [[2]](#footnote-2). This individual subverted the game and maintained low scalar values in order to consistently beat other individuals. This again an example of a loss of focus and relativism taking place.

One issue for individuals with high scalar values is that when compared against other individuals at the maximum scalar value cap, if both individuals are at the cap in the same dimension, then both individuals will receive no reward.

Another issue for many of the individuals is that they increased one of their scalar values to the maximum, but upon reaching the maximum value they kept attempting to increase it, counting as a skipped action, and providing no change to their scalars and no improvement overall, whilst the other individuals were free to adjust and change. This is unintended behaviour that could be a result of the individuals’ observation of the environment, as they count how many other individuals are below or above them in either dimension, and their state becomes whichever count is highest. It is possible that this behaviour would not occur if the observation was instead related to the reward of an individual; how many other individuals it beat last “turn” for example, but it introduces further complexity as now the observation space is a value between 0 and 24 which could lead to insufficient data collected for each state and lead it to make poor decisions. Turning the observation space into a continuous space is also an option but would require a massive fundamental change in the network’s architecture.

# Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Pellentesque

# Bibliography

[1] J. Oh *et al.*, “Discovering Reinforcement Learning Algorithms,” *arXiv [cs.LG]*, 17-Jul-2020.

[2] V. Francois-Lavet, P. Henderson, R. Islam, M. G. Bellemare, and J. Pineau, “An Introduction to Deep Reinforcement Learning,” *arXiv [cs.LG]*, 30-Nov-2018.

[3] E. Lecarpentier and E. Rachelson, “Non-Stationary Markov Decision Processes a Worst-Case Approach using Model-Based Reinforcement Learning.”

[4] G. Papoudakis, F. Christianos, A. Rahman, and S. V. Albrecht, “Dealing with Non-Stationarity in Multi-Agent Deep Reinforcement Learning,” *arXiv [cs.LG]*, 11-Jun-2019.

[5] S. Devlin, “Potential-Based Reward Shaping for Knowledge-Based, Multi-Agent Reinforcement Learning.pdf,” 2013.

[6] W. Daniel HILLIS, “Co-evolving parasites improve simulated evolution as an optimization procedure,” *Physica D*, vol. 42, no. 1–3, pp. 228–234, 1990.

[7] E. De Jong E Popovici A Bucci R Wiegand, “Coevolutionary Principles.pdf,” 2012.

[8] T. G. Tan, J. Teo, and H. K. Lau, “Competitive Coevolution with K-Random Opponents for Pareto Multiobjective Optimization,” *Third International Conference on Natural Computation (ICNC 2007)*. 2007.

[9] D. Klijn and A. E. Eiben, “A coevolutionary approach to deep multi-agent reinforcement learning,” *arXiv [cs.NE]*, 12-Apr-2021.

[10] J. Cartlidge and S. Bullock, “Combating coevolutionary disengagement by reducing parasite virulence,” *Evol. Comput.*, vol. 12, no. 2, pp. 193–222, Summer 2004.

[11] R. A. Watson and J. B. Pollack, “Coevolutionary Dynamics in a Minimal Substrate,” 2002.

[12] O. Vinyals *et al.*, “Grandmaster level in StarCraft II using multi-agent reinforcement learning,” *Nature*, vol. 575, no. 7782, pp. 350–354, Nov. 2019.

[13] D. Cliff and G. F. Miller, “Tracking the red queen: Measurements of adaptive progress in co-evolutionary simulations,” *Advances in Artificial Life*. pp. 200–218, 1995.

1. https://github.com/TomClarkJorvik/Tom-Clark-CS-BEng-Project [↑](#footnote-ref-1)
2. See the 22nd individual in logB2\_2.txt [↑](#footnote-ref-2)