Reinforcement Learning, Nurturing, and Evolving Risk Neutrality in a Simulated Population.

Kevin Robb and Dr. Dean Hougen; Robotics, Evolution, Adaptation, and Learning Laboratory (REAL Lab); School of Computer Science, Gallogly College of Engineering, University of Oklahoma; Summer 2018

Goals:

- **Discover under what conditions** a reinforcement learning population will evolve greater risk aversion or greater risk neutrality.
- Understand the role of **nurturing**.
- **Determine trends** in the data that reflect a broader insight to machine learning theory.
- **Apply to physical world** to describe environments where a population will diverge into two distinct groups based on minor differences.
- **Contribute to evolution of machine learning** by demonstrating ways to guide evolution of risk evaluation.

Definitions with Context:

Reinforcement Learning (RL): machine learning method in which an agent learns an action **policy** by acting then receiving a numerical **reward** value based on its action. This **reward** tells the agent whether it has done a good job or a bad job, and **the agent will use this to improve its policy**.

State: a shorthand for a certain choice and its possible reward values. Denoting a state as B[0-220] means option B will return either 0 or 220, chosen at random. A[100] has only one possible reward.

Risk Aversion (RA): a preference for **safe alternatives**, even when a risky option may have a higher expected turnout. RA agents **ignore minor statistical benefits and prefer a guaranteed reward**, so they will choose A>(B=C).

Risk Neutrality (RN): a **lack of preference** for a certain option based on its riskiness. RN agents put **main priority on expected value rather than relative safety**. Ideally, a RN agent in this three-choice environment will be able to recognize that B>A>C.

Learning Parameter (L): measurement of "**riskiness**," in the range 0 to 1. **Each agent has its own L**, constant during its life; L is only changed between generations. This L value is used to recalculate estimations after each trial: newEstimate = reward * L + currentEstimate * (1 - L)

- A **high value of L puts more weight on recency,** representing **risk aversion**. Reward 0 drops an estimated value very low, making it unlikely to be chosen again for a while.
- A **low value of L puts more weight on previous experiences,** and represents **risk neutrality**. Reward 0 preserves current estimations, and allows gradual progression towards the true mean.

Nurturing Period: a time to freely explore and learn the true mean values of all options *before* fitness matters. Choices during this time do not affect fitness. This **gives nurtured agents an advantage.**

Table 1: Default Configuration Parameters for Environment
(values not mentioned in analysis are assumed to be the defaults from this table)

Parameter	Value
Initial Learning Parameter	0.5
Mutation Standard Deviation	0.05
Number of Agents	50
Number of Trials	500
Number of Nurturing Trials	150
Number of Generations	500
Tournament Size	2
Certain State	A[100]
Uncertain States (Symmetric)	B[0-220] and C[0-180]
Uncertain States (Asymmetric)	B[0-220] and C[0-150]

Experimental Setup:

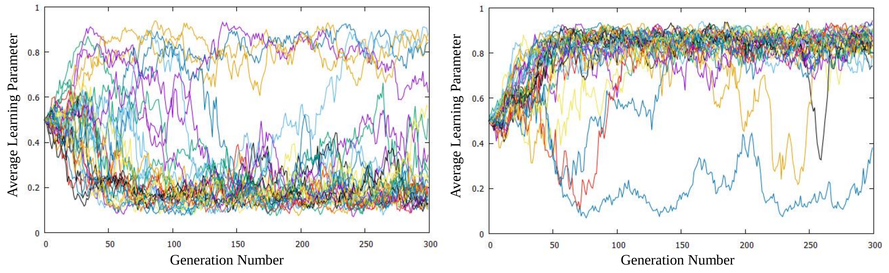
Nurturing and non-nurturing simulations are run completely separately, with the same parameters, and analyzed afterwards. There is no interaction between the populations.

500 Generations
- 500 Trials each. First 150 trials are the **nurturing period.**
  - Each agent makes a **choice**, recalculates **estimations**, and increments **aggregate fitness**.
- Data written at end, and next generation formed.
  - 2 random agents chosen for **tournament**. Agent with higher fitness is **selected**.
    - Higher fitness agents are more likely to "produce offspring," but low fitness agents are not excluded.
    - In the long run, bad genes are removed from the gene pool.
    - A higher tournament size allows risk-seeking agents to take over the population. (avoided)
  - Occurs 50 times to form a new population. Agents can be selected more than once.
  - Selected agents undergo **mutation to their L values**.
    - Mutations are chosen from a **Gaussian distribution** of mean 0.
    - A mutation (positive or negative) is added to each L value.
  - New agents are formed with these L values, and **all estimations reset to 100**.
    - Allowing estimations to carry over would constitute the evolution of instincts, rather than learning, and is not part of this study.

Progression:

Two-state environment, with results opposite of expected. Non-nurturing is risk neutral, and nurturing is risk averse. Graphed: average L as a function of generation number. 30 complete runs of the simulation. 300 generations, 200 trials each.

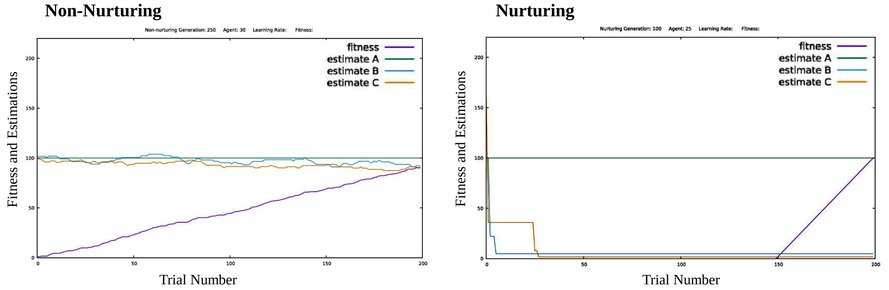
Non-Nurturing **Nurturing**



Issues: **inconsistent**, **random**, contradicts RL theory

Three-state environment implemented at this point. Tracking specific data.
Graphed: **fitness** and **estimations** as functions of trial number. 200 trials. One graph represents a single agent during a single generation.
The flat segment of the fitness line (purple) is the **nurturing period.** Ideally the estimations will each converge to approximate their real mean values. This does not happen very successfully yet.
Significant spikes and drops mean the agent has a high L and is very risk averse. If the lines experience minimal change, the agent has a low L and is very risk neutral. Non-nurturing (left) is still risk neutral, and nurturing (right) is risk averse.

Non-Nurturing **Nurturing**



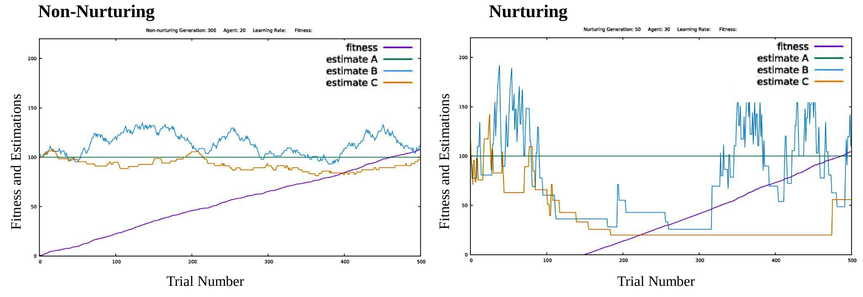
Issues: still contradicts RL theory, cannot distinguish between risky options

Continued at top of next column.

Continued Progression:

In the physical world, nurtured organisms (e.g., humans) live well beyond their nurturing period. Increased number of trials to 500, but left 150 nurturing trials, so now an agent is **nurtured for 30% of their life**, rather than 75%. This change caused nurtured agents to also evolve risk neutrality, and both became more successful.

Non-Nurturing **Nurturing**



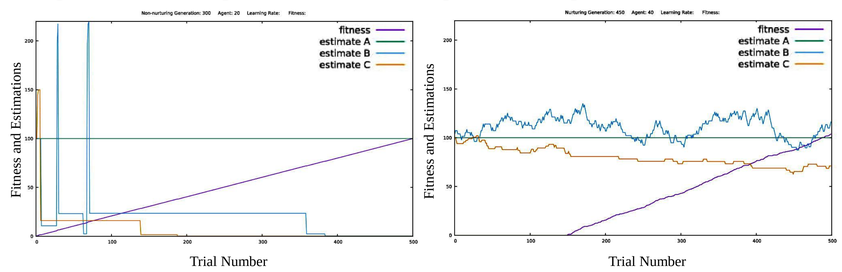
Asymmetry introduced: Using B[0-220] and C[0-150], the average of B's and C's expected values no longer equals A's, but is **lower**.
- If this was higher, both cases would of course become RN.
- Because it is lower, **being risky is bad unless an agent can tell the difference between B and C.**
- Nurturing allows agents to learn to choose B and avoid C (evolving RN)
- Without nurturing, agents avoid both B and C, sticking with safe A (evolving RA).

This behavior meets our initial query, but the type of environment that cause this behavior was unexpected.

Results:

- When **negatively-skewed asymmetry** is introduced, **non-nurtured agents** are only able to **tell that riskiness is bad on average** (Figure 1), while **nurtured agents** are **able to distinguish** between the two uncertain options and choose the better (Figure 2).
- The **non-nurtured population evolves risk aversion** and is closely gathered around a mean fitness of 100.
- The **nurtured population evolves risk neutrality** and is very spread out around a mean fitness greater than 100.
- The **fitness landscape has two optima;** which one is settled depends on mutation rate (MR).
  - Low L: global optimum, but small basin of attraction (hard to find; high MR or nurturing)
  - High L: local optimum, but large basin of attraction (easy to find, low MR)

Figure 1: Non-Nurturing **Figure 2: Nurturing**



Footnotes:

This project was successful in demonstrating that the type of environment affects what type of risk evaluation will evolve in an artificially intelligent system. An interesting follow-up to this experiment would be evolving instincts, such as the estimations (to some extent), or changing the environment throughout the simulation to see how that affects the population's reliance on learning.

The non-nurturing case evolving RA upholds the concepts proven in the 2002 paper written by Yael Niv et al., which defined the fundamentals of RL. In the future, I will continue working with risk evaluation and the evolution of learning in Dr. Hougen's REAL Lab.

I was able to undertake this research through the FYRE program at OU because of NSF grant #1726889.