Reinforcement learning is an effective strategy to create phenotypic variation and a potential mechanism for the initial evolution of learning.

Jan Teichmann[1], Eduardo Alonso[2], and Mark Broom[1]

[1] Department of Mathematics, City University of London, Northampton Square, London EC1V 0HB jan.teichmann@city.ac.uk Mark.Broom@city.ac.uk
[2] Department of Computer Science, City University of London, Northampton Square, London EC1V 0HB E.Alonso@city.ac.uk

Abstract Evolution leads to animals being well-adapted to their environment, in terms of physical abilities and behaviours. Many animal behaviours, however, are not simply genetically determined, but are a result of learning. Learning is generally assumed to be an adaptation to environmental change, but the relationship between environmental factors, learning, and evolution is complex and not fully understood. Here, the role of reinforcement learning is of increasing interest. We present a general model inspired by evolutionary games which analyses and compares fitness distributions of individuals in a changing environment which either learn or evolve through mutation. We show that reinforcement learning offers a potential mechanism for the initial evolution of learning irrespective of any technical parameters and confirm previous findings of other established models of learning.

1 Introduction

Through evolution, animals are generally very well-adapted to their environment. Phenotypic plasticity allows for suitable adaptations even in the face of changing environments [12]. Thus both physical abilities and behaviours of animals are generally appropriate to their environment. Nevertheless many animal behaviours are not solely genetically determined, though some are, but the response of the animal's learning capabilities. Hence a key question arises: under which conditions is the ability to learn beneficial? From a biological perspective, learning is a mechanism for rapid adaptation (modification) of behaviour during the individual's lifetime and a distinct adaptation to changing environments in particular [7]. The main line of argument is that learning incurs some cost, so that a constant environment should select for a genetically fixed pattern of behaviour over learned behaviour. But the relationship of learning and evolution is complex and an important aspect of learning is environmental predictability (commonly referred to as regularity) [15]. Clearly, there is nothing to learn in an environment which is absolutely unpredictable. So far both factors, environmental change and regularity, have been discussed in the literature as selective

factors in the evolution of learning. A contradiction at first sight, but a solution to the paradox would be that learning is in fact an adaptation to intermediate levels of environmental change [8].

Associative learning is a fundamental cognitive process observed across species (including mollusks, insects, birds and mammals) [11] that affects a wide variety of behaviours ranging from colour recognition [2, 9, 13, 20] and spatial representation [1], to causality judgements [14] and goal-directed behaviour [21]. Of course, animals use other types of learning (e.g. social learning or perceptual learning) and ontogenetic mechanisms (e.g. habituation and phenotypic plasticity) to adapt their behaviour to the environment. Nonetheless, the pervasiveness and relevance of associative learning makes it the ideal candidate to investigate when learning is most effective. We are interested in the initial evolution of learning [3,17] and in particular how reinforcement learning [4,5] could evolve within a population of generally well adapted animals. To answer this question we focus on a simple model of learning where individuals learn to associate events that occur together, for instance two stimuli, a stimulus and a response, or a response and its outcome [10].

2 Model Definition

Our work develops that of [18, 19], which investigated the effects of aversive learning in a changing environment on a predator's diet choice and energy intake. Here we describe fitness distributions of learning individuals in changing environments more generally and compare them with a simplistic mutation process to understand the relationship between evolution and learning.

In our model the learning individual uses Q-learning [22], chosen for the simplicity of its implementation of real-time error-correction learning and as it is increasingly supported by both behavioural and neural data. In Q-learning an individual uses experience following its interactions with the environment to infer optimal decisions. The learning individual utilises an action-value function to build a representation of the environment which describes the expected future payoff following a specific action in a specific state of the environment. It then minimises the error of the function's future payoff prediction building on a growing amount of evidence from past trial-and-error interactions with the environment. These payoff predictions are discounted by a factor γ, representing future uncertainty. The prediction error is modulated by a learning rate α. Finally, the individual translates the action-value function predictions into a decision following a stochastic policy, e.g. Gibb's soft-max policy.

An individual of the mutating population has a genetically determined decision policy chosen randomly from a uniform distribution at the beginning of each generation. The important differences between the two populations are: (i) the mutation process is random and not adaptive, operating on fixed phenotypes and (ii) learning is adaptive but incurs an exploration cost, where suboptimal decisions are made to learn about the environment. Selection is not included as we are only interested in both populations' fitness distributions.

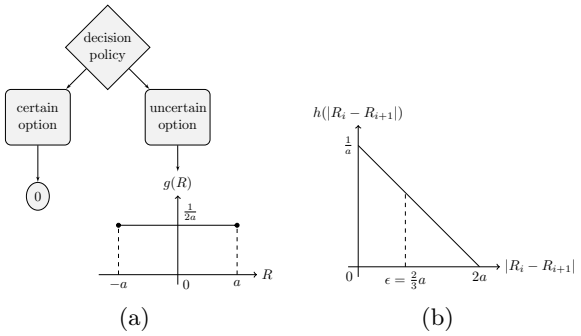


Figure 1. The environment with the choice of a certain and an uncertain option. **(a)** The two options of the environment with the certain option being equal to zero and the uncertain option following a uniform fitness payoff distribution $g(R)$ with limits $-a$ and a as given by Equation (1). (**b**) The distribution of absolute fitness change follows a triangular distribution $h(|R_i - R_{i+1}|)$ with $\epsilon = (2/3)a$ being the average absolute fitness change given the uniform distribution of fitness payoff $g(R)$ with $\beta = 1$.

We define the environment to be stationary and ergodic, consisting of two options, a certain and an uncertain one, as shown in Figure 1. The model parameters are described as follows: (i) R: the fitness payoff following an interaction with the environment, (ii) F: the fitness of an individual at the end of a generation, (iii) $\hat{F}$: the scale-free fitness of an individual, (iv) l: the length of a generation in interactions with the environment, (v) β: the number of environmental changes per generation time, (vi) ϵ: the extent of environmental change per generation time, (vii) α: the learning rate of the learning individual and (viii) γ: the discount rate of future payoffs. The certain option gives a constant fitness payoff $R = 0$ and the uncertain option returns a uniformly distributed fitness payoff $g(R)$ with zero mean. The value of 0 for the fitness of the constant option and the mean of the variable option is chosen for simplicity (it is possible to add an arbitrary constant to both and not qualitatively change our results).

The learning individual cannot draw from any secondary source of information such as a correlation between environmental states which motivates the choice of the uniform distribution $g(R)$. The term regularity refers to the predictability of an environment within models of learning. A perfectly regular environment is one that is constant throughout the lifetime of an individual, with irregularity increasing as the number of changes per lifetime increases. We define irregularity as the expected number of environmental changes in a lifetime, given by β/l. This is perhaps a simplistic definition, and we do not make direct use of it, except that the environment becomes increasingly irregular with greater (smaller) values of β (l). We define the limits $[-a, a]$ of the Uniform distribution $g(R)$ as follows:

$$a = \frac{3}{2\beta}\,\epsilon, \tag{1}$$

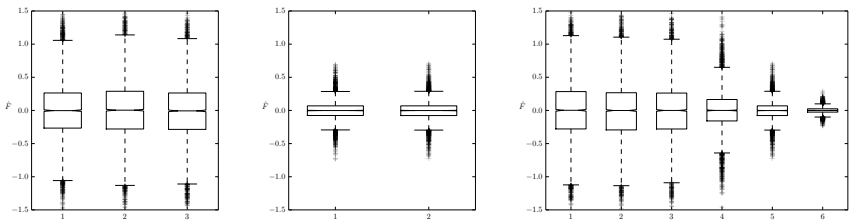
where ϵ is the absolute average fitness change per generation derived from the triangular distribution $h(|R_i - R_{i+1}|)$ of the absolute difference of the uniform fit-

ness payoff $g(R)$ as shown in Figure 1. We assume that an increased frequency of environmental change β results in smoother and less pronounced single changes as reflected in Equation (1). The fitness F of an individual is the sum of the fitness payoffs from its interactions with the environment $F = \sum_{t=1}^{l} R_t$.

3 Results

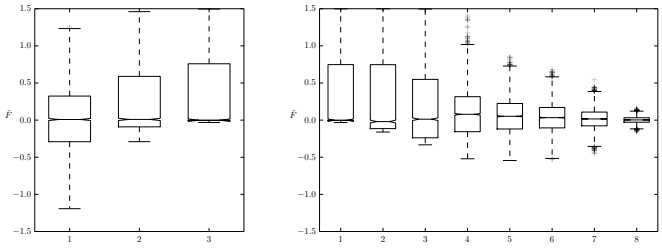
We present the distributions of $n = 5000$ generations interacting with their environment using a scale free variant of the fitness $\hat{F} = \beta F / l\epsilon$, which will allow a more intuitive comparison of the two populations. We will present the results for each population respectively in the form of box-plots and associated Kolmogorov-Smirnov significance tests.

Figures 2a and 2b show the main characteristic of the mutation process: as the process is random and not adaptive it is independent of the number of interactions per generation l and of the extent of environmental change per generation ϵ. The fitness distributions are also symmetric with mean zero. Additionally, α and γ do not apply to the mutation process. Figure 2c shows the effects of the frequency of environmental changes β. The fitness distribution is unaffected for $\beta \leq 1$, i.e. when mutations occur more frequently than changes in the environment. If $\beta > 1$ the fitness distribution of the population of mutating individuals becomes increasingly narrow. This is a direct result of the mutation process being non-adaptive and therefore it is less likely that individuals are well suited (or poorly suited) for a number of consecutive environmental states.



(a) (b) (c)

Figure 2. Scale-free fitness distributions of the mutating population, all with $n = 5000$ generations. **(a)** The fitness distribution is independent of ϵ and l. (1) $\epsilon = 0.1$, $\beta = 1$, and $l = 1000$. (2) $\epsilon = 100$, $\beta = 1$, and $l = 1000$. (3) $\epsilon = 10$, $\beta = 1$, and $l = 10$. Distributions are not significantly different using a Kolmogorov-Smirnov test, all with $p > 0.1$. **(b)** Fitness distributions scale equally with β independently of ϵ and l. (1) $\epsilon = 10$, $\beta = 10$, $l = 1000$. (2) $\epsilon = 1$, $\beta = 10$, $l = 10$. Distributions are not significantly different using a Kolmogorov-Smirnov test, all with $p > 0.1$. **(c)** Fitness distributions become narrower with increasing β. Distributions are not significantly different for $\beta \leq 1$ using a Kolmogorov-Smirnov test, all with $p > 0.1$. (1) $\beta = 0.1$, (2) $\beta = 0.5$, (3) $\beta = 1$, (4) $\beta = 2$, (5) $\beta = 10$, (6) $\beta = 100$. In all cases $\epsilon = 1$, $l = 1000$.

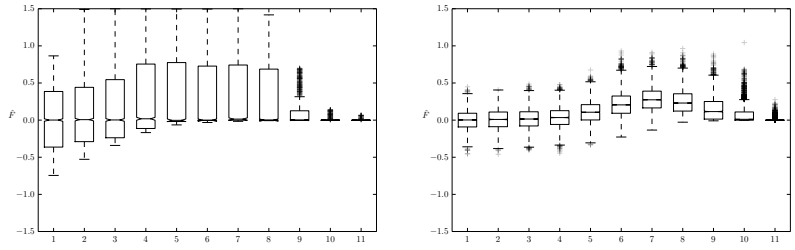


(a) (b)

Figure 3. Scale-free fitness distributions of a learning individual. All cases are with $n = 5000$ generations. **(a)** Learning requires a certain amount of environmental change to occur to be beneficial: (1) $\epsilon = 0.1$, $l = 10$ vs. (2) $\epsilon = 10$, $l = 10$. Learning benefits from longer generation times: (2) vs. (3) $\epsilon = 10$, $l = 1000$. In all cases $\beta = 1$, $\alpha = 0.5$, $\gamma = 0.9$. **(b)** Learning incurs exploration cost and learning benefits diminish in unreliable environments. (1) $\beta = 0.1$, (2) $\beta = 0.5$, (3) $\beta = 1$, (4) $\beta = 2$, (5) $\beta = 3$, (6) $\beta = 5$, (7) $\beta = 10$, (8) $\beta = 100$. In all cases $\epsilon = 1$, $l = 1000$, $\alpha = 0.5$, $\gamma = 0.9$.

The population of learning individuals uses Q-learning to adapt to the current environmental state. This requires exploration which is the sole cost of learning in our model. Additional costs of learning are difficult to quantify and we assume that during the initial evolution of learning these were relatively small [7]. Figure 3a shows that learning requires certain environmental conditions; in particular learning benefits from a changing environment (Figure 3a1 vs. 3a2). The cost and benefits of exploration in the learning population can be seen (Figure 3a1) versus the mutating population (Figure 2a) where the learning behaviour cuts off both tails of the fitness distribution and does not produce the outliers as in the mutation process. Additionally, learning benefits from longer generation times to exploit experience (Figure 3a3).

In Figure 3b we see that learning is directly affected by β compared to the population of mutating individuals which is unaffected for $\beta \leq 1$ (Figure 2c). The effect of β on the fitness distribution of the learning individuals is not linear, and there are multiple underlying factors. In environments with very rare changes an increasing majority of the population benefits from learning (Fig 3b1). At first, an increasing frequency of environmental change increases the fraction of individuals benefiting less from learning with the fitness distribution developing a more pronounced tail of learning individuals having negative relative fitness (Figure 3b1-5). Nevertheless, environmental change benefits learning at the same time with the median of the population increasing (Figure 3b4). Secondly, a further increase of β results in the cost of consecutive exploration and consequent errors outweighing this initial benefit and the distribution increasingly aligns with the fitness distribution of the mutating population. Finally, learning does not provide any benefits in highly irregular environments interfering with any possibility of exploitation. Therefore the only difference in the fitness distributions is the shorter tails for the learning individuals, which is the result of continuous exploration (Figure 3b8 vs. Figure 2c6).



(a) (b)

Figure 4. Scale-free fitness distributions of the learning individuals showing the effects of the extent of environmental change (ϵ). Learning benefits from a certain extent of environmental change but too severe changes incur a high cost of mistakes during the required exploration. Additionally, there is a combined effect of regularity and change. (1) $\epsilon = 0.1$, (2) $\epsilon = 0.5$, (3) $\epsilon = 1$, (4) $\epsilon = 2$, (5) $\epsilon = 5$, (6) $\epsilon = 10$, (7) $\epsilon = 20$, (8) $\epsilon = 50$, (9) $\epsilon = 100$, (10) $\epsilon = 500$, and (11) $\epsilon = 1000$. All cases are with $n = 5000$ generations, $l = 1000$, $\alpha = 0.5$, and $\gamma = 0.9$. **(a)** $\beta = 1$. **(b)** $\beta = 10$.

Figure 4 shows the combined effect of ϵ and β. We have already discussed the individual effects of β relating to Figure 3b. For ϵ we can see that in environments with small values of absolute environmental change throughout a generation the fitness distribution of the learning individuals aligns with the fitness distribution of the population of mutating individuals. A learning individual prioritises continuous exploration if the environmental change is small which is a consequence of learning being an adaptation to changing environments. The shorter tails in the fitness distribution of the learning population compared to the mutating population are the result of this continuous exploration as discussed previously (Figure 4a1 vs. Figure 2a and Figure 4b1 vs. Figure 2b). An increase of ϵ has beneficial effects for learning individuals as learning requires a certain extent of environmental change to exploit. A further increase of ϵ makes mistakes during exploration more expensive which can potentially neutralise the benefits of exploiting beneficial states of the environment. The important difference between a severe extent of environmental change (ϵ) and an irregular environment (β) is that mistakes in the case of ϵ are extremely aversive and stop any further costly exploration. This is why the fitness distribution of learning individuals in violently changing environments loses the negative tail compared to rapidly changing environments (Figure 4a11 vs. Figure 3b8). Combining frequency and extent of environmental change shows that there is a specific combination of these factors which hugely benefits the learning individuals (Figure 4b7).

Our results also show that the fitness distribution of the learning population is independent of the learning rate α and the discount factor γ, within a meaningful range (clearly α cannot be too low; $\alpha = 0$ corresponds to no learning at all). This result should not be misinterpreted: there are specific values of α and γ best suited for achieving optimality in a specific state of the environment. But within a changing environment the distribution of fitness is independent of the specific choice of α and γ.

4 Discussion

In this paper we look at the fitness distribution of individuals using reinforcement learning, i.e. Q-learning, in a changing environment. Our model confirms previous findings: (i) learning requires environmental change and longer generation times to be beneficial, (ii) learning is optimal for specific combinations of regularity and size of environmental change, and (iii) regularity is the key environmental factor which impacts whether learning is advantageous. In addition, we show that (iv) the fitness distribution of learning individuals in changing environments is independent of the learning rate α and the discount factor γ.

We do not present an evolutionary theory of learning in itself, but show that a simple reinforcement strategy (increasingly backed by experimental studies of neural correlates) is beneficial for a vast range of environmental parameters. In particular, the fact that the success of reinforcement learning is independent of technical parameters of learning α and γ, is a new and reassuring insight. These are technical parameters which allow the tuning of over-fitting and the extent of exploration for a specific learning task and have great importance in the field of computing. But in a biological context of changing environments these technical learning parameters become negligible. This significantly reduces the complexity of the initial evolution of reinforcement learning in comparison to the combinatorial explosion of the parameter space in models of connectionism [6].

Our model does not include interactions between individuals. Nevertheless, it reproduces many widely accepted theories of learning in the context of change and regularity [17]. In addition, it has been widely acknowledged that the benefit of learning is the ability to adapt to a changing environment faster than the time scale on which evolution operates [7]. This is an important evolutionary dynamical argument, but is distinct from those of our paper; here the benefits of learning lie in exploitation ability rather than adaptation speed itself.

Considering the effects of selection the mutating population in our model has a constant relative arithmetic mean fitness of zero. The environmental changes only affect the fitness variability of the mutating population in a symmetric fashion. As the arithmetic payoff of both options in our model are equal, selection would increase long-term fitness of the mutating population by discarding the uncertain option from the action space of the mutation process in order to reduce fitness variability [16, 18]. This provides an alternative interpretation of why learning is a distinct adaptation to changing environments alongside the cost argument: a simplistic mutation process cannot exploit environmental change without the introduction of increased fitness variability at the same time.

Taking selection into account, our results show that reinforcement learning is a promising starting point for the initial evolution of learning. The key environmental factor here is regularity. If selection cannot discard the uncertain option from the action space of the mutational process, learning is always beneficial as it has lower fitness variability even in extremely irregular environments when compared to the mutating population. If selection can in fact discard the uncertain option in the case of the mutating population then learning becomes disadvantageous in highly irregular environments.

Acknowledgements

This research was supported by a research studentship to Jan Teichmann provided by City, University of London.

References

1. Albasser, M.M., Dumont, J.R., Amin, E., Holmes, J.D., Horne, M.R., Pearce, J.M., Aggleton, J.P.: Association rules for rat spatial learning: The importance of the hippocampus for binding item identity with item location. Hippocampus 23, 1162–1178 (2013)
2. Carew, T.J., Hawkins, R.D., Kandel, E.R.: Differential classical conditioning of a defensive withdrawal reflex in aplysia californica.. Science 219, 397–400 (1983)
3. Chalmers, D.J.: The evolution of learning: An experiment in genetic connectionism. In: Proceedings of the 1990 connectionist models summer school. pp. 81–90 (1990)
4. Dayan, P., Daw, N.D.: Decision theory, reinforcement learning, and the brain. Cognitive, Affective, & Behavioral Neuroscience 8, 429–453 (2008)
5. Doya, K.: Reinforcement learning: Computational theory and biological mechanisms. Human Frontiers Science Program Journal 1, 30–40 (2007)
6. Hinton, G.E., Nowlan, S.J.: How learning can guide evolution. Complex systems 1(3), 495–502 (1987)
7. Johnston, T.D.: The selective costs and benefits of learning: an evolutionary analysis. Adv. Stud. Behav 12, 65–106 (1982)
8. Johnston, T.D., Turvey, M.T.: A sketch of an ecological metatheory for theories of learning. The psychology of learning and motivation 14, 147–205 (1981)
9. Leimar, O., Enquist, M., Tullberg, B.S.: Evolutionary stability of aposematic coloration and prey unprofitability: A theoretical analysis. American Naturalists 128, 469–490 (1986), http://www.jstor.org/stable/2461331
10. Mackintosh, N.J.: The psychology of animal learning. Academic Press (1974)
11. Macphail, E.M.: Brain and intelligence in vertebrates. Claredon (1982)
12. Pigliucci, M.: Phenotypic plasticity: beyond nature and nurture. JHU Press (2001)
13. Ruxton, G.D., Sherratt, T.N., Speed, M.P.: Avoiding Attack: The Evolutionary Ecology of Crypsis, Warning Signals and Mimicry. Oxford University Press (2004)
14. Shanks, D.R.: The psychology of associative learning. Cambridge University Press (1995)
15. Staddon, J.E., Simmelhag, V.L.: The" supersitition" experiment: A reexamination of its implications for the principles of adaptive behavior. Psychological Review 78, 3–43 (1971)
16. Starrfelt, J., Kokko, H.: Bet-hedging—a triple trade-off between means, variances and correlations. Biological Reviews 87(3), 742–755 (2012)
17. Stephens, D.W.: Change, regularity, and value in the evolution of animal learning. Behavioral Ecology 2, 77–89 (1991)
18. Teichmann, J.: Models of aposematism and the role of aversive learning. Ph.D. thesis, City, University of London (2014)
19. Teichmann, J., Broom, M., Alonso, E.: The application of temporal difference learning in optimal diet models. Journal of theoretical biology 340, 11–16 (2014)
20. Teichmann, J., Broom, M., Alonso, E.: The evolutionary dynamic of aposematism: a numerical analysis of co-evolution in finite populations. Math. Model. Nat. Phenom. 9, 148–164 (2014)
21. Valentin, V.V., Dickinson, A., O'Doherty, J.P.: Determining the neural substrates of goal-directed learning in the human brain. The Journal of Neuroscience 27, 4019–4026 (2007)
22. Watkins, C., Dayan, P.: Q-learning. Machine learning 8, 279–292 (1992)