City Research Online

City, University of London Institutional Repository

Citation: Alonso, E., Fairbank, M. & Mondragon, E. (2012). Conditioning for Least Action. Paper presented at the 11th International Conference on Cognitive Modeling, 13-04-2012 - 15-04-2012, Berlin, Germany.

This is the published version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/5202/

Link to published version:

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

Conditioning for Least Action

Eduardo Alonso (E.Alonso@city.ac.uk)
Department of Computing, Northampton Square
London, EC1V 0HB United Kingdom

Michael Fairbank (abdy934@soi.city.ac.uk)
Department of Computing, Northampton Square
London, EC1V 0HB United Kingdom

Esther Mondragón (e.mondragon@cal-r.org)
Centre for Computational and Animal Learning Research
St Albans AL1 1RQ, United Kingdom

Abstract

It is well known that, in one form or another, the variational Principle of Least Action (PLA) governs Nature. Although traditionally referred to explain physical phenomena, PLA has also been used to account for biological phenomena and even natural selection. However, its value in studying psychological processes has not been fully explored. In this paper we present a computational model, value-gradient learning, based on Pontryagin's Minimum Principle (a version of PLA used in optimal theory), that applies to both classical and operant conditioning.

Keywords: Value-gradient learning; conditioning; behavior systems; bliss point; optimality; principle of least action.

The Principle of Least Action

Of all the possible trajectories a ball thrown into the air can follow why does it follow one in particular, a parabola? Why doesn't it go up, stay a while at its highest point and then fall down? On the one hand, the ball wants to spend a lot much time near the top of its trajectory since this is where the kinetic energy is least and the potential energy is greater. On the other hand, if it spends too much time near the top, it will really need to rush to get up there and get back down and this will take a lot of action. The perfect compromise is a parabolic path. In physical *parlance*, the true dynamical trajectory of the ball is the one that makes the action "least" (actually stationary).

Formally, the action to be minimized is the integral of a function, the Lagrangian, over time. The Lagrangian itself describes completely the dynamics of the system under consideration as the difference between its kinetic energy (the energy due to the motion, how much is "happening") and its potential energy (the energy due to its position or configuration, how much "could happen"). In short, Nature is as lazy as possible: the ball follows a particular trajectory not because of the effect of gravitation *per se*, but because it "minimizes" action. In fact, this condition is equivalent to the Euler-Lagrange equation of motion that encapsulates the Principle of Least Action and that, when transformed into its Hamiltonian form, reflects the symmetries of Nature. These are fundamental concepts upon which modern Physics is based.

The question is, can we export this variational analysis to the study of learning and behaviour?

Optimization in Classical Conditioning

Let's consider acquisition of an eye-blink conditioned response when a light is paired with a mild shock: at first the likelihood of a response to the light is low because of the absence of prior light-shock pairings. There is then a rapid increase in magnitude of the response, which diminishes gradually as training progresses until there are no further increases in the measure of the conditioned response. The shape of the learning curve is typical of that found in many studies of conditioning. How is this pattern of behaviour explained? Why don't animals learn "all" in a single trial? Or learn rapidly at the beginning, then stop and then learn again? In a way, we are facing the same questions as we did when considering ball trajectories. And it is paramount that we answer them since *conditioning is at the basis of most learning phenomena and thus of animal cognition*.

More generally, classical conditioning refers to the type of learning that occurs when pairing two stimuli, typically an originally neutral stimulus (say a tone or a light) and an unconditioned stimulus (US), that is, a stimulus that is biologically relevant to the animal (for instance, food) that elicits an automatic or unconditioned response (UR, for example, salivation). If this pairing is repeated over time, the animal will learn to anticipate the US and start responding to the signal, the neutral stimulus. The neutral stimulus will become a conditioned stimulus (CS) and trigger a response (CR, typically the UR itself).

In order to explain this type of phenomena, Rescorla and Wagner's model of classical conditioning (Rescorla & Wagner 1972) assumes that learning occurs on a conditioning trial only if the US is surprising. "Surprise" is defined in terms of growth of "associative strength", the strength of the CS's association with the US over trials (V, traditionally measured in terms of number of URs). With each trial there is an increase or jump in associative strength. On early conditioning trials the jumps are large; that is, each trial causes a relatively large increase in associative strength. But the jumps decrease in size as learning progresses until the learning curve approaches its upper limit or asymptote. Once the CS predicts the US, the US is not surprising, and no further learning occurs.

Formally, for a CS s the change in learning on trial n is defined as

$$\Delta V_n(s) = \alpha\beta\big(\lambda - V_{n-1}(total)\big) \quad (1)$$

where α and β represent the salience of the CS and of the US respectively ($0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$), λ is the maximum amount of learning that can occur in that situation at that given trial, and $V_{n-1}(total)$ the cumulative amount of learning up to the previous trial, that is, the sum of associative strengths of all CSs that are present at trial n; in turn, the associative strength of each of the CSs that are present is determined on the last trial on which each CS occurred, ordinarily trial $n-1$. This delta rule is also known as the *error correction rule*: it calculates the prediction error, that is, the difference between the prediction and the actual reward. The result is then used to calculate the new associative strength of the CS as $V_n = V_{n-1} + \Delta V_n$, the *update rule*. Obviously, as the prediction improves the difference in delta is reduced until there is nothing left to be learned.

This deceptively simple theory is nevertheless considered as the most influential model of conditioning. Interestingly, Rescorla and Wagner's rule works pretty much as the Lagrangians in mechanics: during learning we balance what we have learned against what is to be learned so that the total associative strength is at each trial 1 (i.e., it is conserved) and the differences between trials, 0. In terms of optimization, Rescorla and Wagner's model uses equation (1) as a way of minimizing the prediction error between the expected reward and the actual reward – in other words, we apply an optimization principle that maximizes the reward.

Nonetheless, like most models of conditioning (see (Alonso & Schmajuk 2012) for a recent review) Rescorla and Wagner's is limited to classical conditioning: responses are only considered as a way of measuring how animals learn to associate two stimuli but do not form part of conditioning *per se*.

What happens when we study operant (aka instrumental) conditioning and goal-directed behaviour? In other words, what happens when the occurrence of a reinforcer depends on the choices the animal makes? Classical conditioning focuses on how "mental" representations of stimuli are linked whereas operant conditioning deals (mainly) with response-outcome associations. It is agreed though that, at the most general level, their *associative structures* are the same: in both procedures, changes in behavior are considered the result of an association between two concurrent events and explained in terms of operations of a (conceptual) system that consists of nodes among which links can be formed. Notwithstanding the correctness of such analysis, we are showing in the next section that a mere translation of classical conditioning into instrumental terms (for instance, by assuming that instrumental responses take the place of CSs) would impose a series of conditions on optimization that are impossible to meet. To see this point and understand our proposal to "recover" variational principles in the study of learning and behaviour we need to briefly introduce temporal difference, a model that comprises both classical and operant conditioning.

Temporal Difference

Temporal difference (TD) was originally presented as an extension of the Rescorla and Wagner's model in real time (Sutton & Barto 1987). It was argued that time scale invariance over trials should not prevent a model of conditioning from investigating temporal phenomena. Indeed, Rescorla and Wagner's model refers to learning through trials, and thus a number of interesting phenomena are left unexplained (such as second order conditioning). TD adopts Rescorla and Wagner's main psychological premises, namely, cue competition and error correction, but instead of comparing the rewards predicted on consecutive trials, we calculate the change in reward prediction error on every time step t. TD makes predictions over predictions and uses the error to update the old reward prediction and bring it more in line with the animal's moment-to-moment experiences –what is called *bootstrapping*.

Formally, in the general case the value of a CS at a particular time t is defined as

$$V_t(s_t) = r_{t+1} + \gamma\, r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \dots \quad (2)$$

where the γ parameter takes values between 0 and 1 and acts as a discount factor that causes distant CSs to matter less than immediate ones and r represents the US. If we compare the values at successive steps an interesting relationship emerges, namely,

$$V_t(s_t) = r_{t+1} + \gamma\, V_t(s_{t+1}) \quad (3)$$

This makes sense because $\gamma\, V_t(s_{t+1})$ takes the place of the remaining terms $\gamma\, r_{t+2} + \gamma^2 r_{t+3} + \dots + \gamma^{T-t-1} r_T$, where T refers to the terminal state, i.e., to the end of the trial. This relation describes the simplest TD case, when predictions are carried out one-step ahead. We can generalize it to any number of steps and calculate the delta rule as

$$\Delta V_t(s_t) = \alpha\left(R_t^{(n)} - V_t(s_t)\right) \quad (4)$$

where

$$R_t^{(n)} = r_{t+1} + \gamma\, r_{t+2} + \gamma^2 r_{t+3} + \cdots + \gamma^{n-1} r_{t+n} + \gamma^n\, V_t(s_{t+n}) \quad (5)$$

If we take a number of steps into the calculation we would need to know how much each step contributes towards the return. TD proposes to average the n-steps with a trace λ (not to be mistaken for the US asymptotic value) so that the return is defined by

$$R_t^\lambda = (1-\lambda)\sum_{n=1}^{T-t-1}\lambda^{n-1}R_t^{(n)} + \lambda^{T-t-1}R_t \quad (6)$$

And the new delta rule is

$$\Delta V_t(s_t) = \alpha\left(R_t^\lambda - V_t(s_t)\right) \quad (7)$$

It is easy to see that if λ is set to 0, TD(0), we only use the one-step backup; on the other hand, if λ is set to 1, TD(1), we only learn from the final return like in Rescorla and Wagner's model.

Temporal difference has gained notoriety because there is a strong correlation between its error term and the behaviour of dopamine cells in the brain (Montague, Dayan & Sejnowski 1996, Schultz 2002). Besides, TD focuses unashamedly on optimization and it is unique in that it aims at explaining both classical and instrumental conditioning. In fact TD has become the most successful Reinforcement Learning algorithm, bringing gaps between psychology, neuroscience, machine learning and control, and the new area of neuroeconomics (Glimcher, Camerer, Fehr & Poldrack 2009).

Temporal Difference and Operant Conditioning

Unlike Rescorla and Wagner's model, TD does provide a way (indirect as it might be) of learning how to select actions. The most common idea is to learn a separate value for each action leading out of a state, that is, executed in the presence of a stimulus, rather than for the state (stimulus) itself. An animal is assumed to exist in an environment described by some set of possible states S, where it can perform any actions A. Each time it performs an action a_t in some state s_t, the world enters into a new state $s_{t+1} = f(s_t, a_t)$ and the animal receives a real-valued reward $r_t = r(s_t, a_t)$. With this information, the animal calculates the TD error, typically using the so-called Q-learning rule (Watkins 1989), $\Delta V_t(s_t, a_t) = r_{t+1} + \gamma\, max_{a_{t+1}} V_t(s_{t+1}, a_{t+1}) - V_t(s_t, a_t)$, and updates the value of the state-action pair as $V_{t+1}(s_t, a_t) = V_t(s_t, a_t) + \alpha\, \Delta V_t(s_t, a_t)$. The animal's task is to learn a control policy π, which maximizes the expected sum of rewards.

Under certain conditions Q-learning can be proved to converge to the value function that will yield the optimal policy. Tragically, Q-learning diverges when the state space is too large as it is the case in most biologically relevant problems.

A standard approach to tackle this problem is to introduce a scalar function approximator, $\tilde{V}(s, \vec{w})$ (e.g., a neural network with single output and weight vector). This is called the approximate value function, or the *critic*. The objective of learning is to make this function accurately estimate V^π for all S. We can then define a greedy policy on $\tilde{V}$ as a policy that always considers all possible actions available to it and chooses the one that leads to the state with the highest $\tilde{V}$ value, whilst also taking into account the immediate short terms reward in getting there. The idea is to maximize the cumulative reward by minimizing the error as given by

$$\Delta\vec{w} = \alpha\sum\left(\frac{\partial\tilde{V}_t(s_t, a_t)}{\partial\vec{w}}\right)\left(R^\lambda - \tilde{V}_t(s_t, a_t)\right) \quad (8)$$

TD(λ) and Q-learning can then be used to update $\tilde{V}$ by sampling one trajectory at a time. Variants of these methods have produced some successes in control problems (Tesauro 1994), yet TD algorithms have not been proved to converge in the general case. Why is that?

TD is based on Bellman's Optimality Condition (Bellman 1957): if $\tilde{V} \equiv V^\pi$ for all s in the state space S, where the policy is greedy on $\tilde{V}$, then that greedy policy is globally optimal. The problem is that for the Bellman's condition to be met we need to explore the entire estate space. Even if Bellman's condition is perfectly satisfied along a single trajectory, performance can be extremely far from optimal if Bellman's condition is not satisfied over the neighbouring trajectories too. That is, if the animal tries to avoid Bellman's condition by only exploring a sub-space of the state space there is no guarantee the resulting policy will be locally optimal. This is the *curse of dimensionality* that applies to some degree to all value-based learning algorithms.

Translated into behavioural terms Bellman's condition means that for an animal to find an optimal policy it would need to explore all possible actions at every possible state. Clearly, this is not the way things happen in the natural world. To picture this, imagine Thorndike's cat trying to escape from the puzzle box. Bellman's condition would require that at every single step the cat would have to execute all the actions in its behavioural repertoire (including, for instance, banging its head against the wall). Hence, it is not just that Bellman's condition is computationally intractable. It is psychologically implausible too. Notice that in classical conditioning this problem does not arise since behaviours are reduced to reflexes. It is assumed that animals use a model (their evolutionary history) to "select" actions. In instrumental conditioning, however, any action can occur –at least in principle.

Value-Gradient Learning

In what follows we present a modification of TD, Value-Gradient Learning (VGL), that under certain conditions guarantees optimality. Importantly, VGL is equivalent to a variational principle, Pontryagin's Mimimum Principle (PMP) (Pontryagin, Boltyanskii, Gamkrelidze & Mishchenko 1962) which, in turn, is a version of Hamilton's Principle of Least Action. The main difference between TD and VGL lies on *what* is learned: VGL learns gradients of values as opposed to TD algorithms that learn values. Besides, with regards to *how* learning occurs, VGL follows the gradient ascent on the total reward rather than the gradient descent on the expected reward.

We define the value gradient as

$$G(s) = \frac{\partial V(s,a)}{\partial s} \tag{9}$$

and the *approximate* value gradient as

$$\tilde{G}(s,\vec{w}) = \frac{\partial \tilde{V}(s,a)}{\partial s} \tag{10}$$

Our algorithm is defined by a weight update of the form

$$\Delta\vec{w} = \alpha \sum_t \left(\frac{\partial \tilde{G}}{\partial \vec{w}}\right)_t \left(G'_t - \tilde{G}_t\right) \tag{11}$$

where G'_t is the *target* value gradient defined recursively by

$$G'_t = \left(\frac{Dr}{Ds}\right)_t + \gamma \left(\frac{Df}{Ds}\right)_t \left(\lambda G'_{t+1} + (1-\lambda)\tilde{G}_{t+1}\right) \tag{12}$$

with $G'_t = \vec{0}$ at any terminal state and where $\left(\frac{D}{Ds}\right)$ is shorthand for

$$\left(\frac{D}{Ds}\right) = \frac{\partial}{\partial s} + \frac{\partial \vec{a}}{\partial s}\frac{\partial}{\partial \vec{a}} \tag{13}$$

It has been proved that any greedy trajectory satisfying $\tilde{G}_t = G'_t$ for all t must be locally extremal, and often optimal (Fairbank, Prokhorov & Alonso 2012). This local optimality condition needs satisfying only over a single trajectory, whereas for TD the corresponding optimal condition (Bellman's) needs satisfying over the whole state space. It is easy to see that our demonstration is based on PMP. Unlike Bellman's condition, PMP states the *necessary* conditions for a trajectory to be (locally) optimal and thus it can be considered as a version of Bellman's Optimality Principle, if localized down to considering the current trajectory only. As a consequence, VGL can lead to increased efficiency. Moreover, it must be noticed that if we apply PMP to all the trajectories we "recover" global optimality.

Value-Gradient Learning and Temporal Difference

If we compare equation (11) against its TD equivalent (8), we see that they are analogous except for the introduction of the model in equation (12). More specifically, the definition of the target gradient G' is the full derivative with respect to s of the "λ-Return" which is the target used in the TD(λ) weight update. This may give the wrong impression that VGL(λ) is just a differentiated form of TD(λ). Contrarily, they differ in a fundamental way: in VGL, if the weight update is at a fixed point at every time step along a trajectory generated by a greedy policy, for any lambda, (i.e., if the learning objective $\tilde{G}_t = G'_t$ is met for all t along the trajectory), then that trajectory is locally extremal, and often locally optimal (Fairbank *et al.* 2012). This contrasts to TD methods in that it is possible for the TD weight update to be at a fixed point at every time step along a trajectory generated by a greedy policy, without the trajectory being optimal. This is because for Bellman's condition to apply, the TD weight updates' objective needs satisfying over all of weight space, and hence lots of stochastic exploration is needed. Contrarily, VGL methods have a much lesser requirement for exploration. What we mean by this is that provided the VGL learning algorithm makes progress towards achieving $\tilde{G}_t = G'_t$ all along a greedy trajectory, then provided the trajectory remains greedy, it will make progress in *bending* itself towards a locally optimal shape, and this will happen without the need for any stochastic exploration. In comparison to VGL, the failure of TD without any exploration in a deterministic environment is dramatic and common, even when the value function is perfectly learned along a single trajectory.

The main insight is that it is not enough to use the derivatives of the values. This is what the Jacobi-Hamiltonian-Bellman equations do in extending the Bellman condition to continuous state spaces. Unfortunately, such derivation does not exploit fully the information contained in gradient values. We can't just consider the change in V over the particular step s along the trajectory. This is like "dotting" $\frac{\partial V}{\partial s}$ with Δs, which is approximately equal to the TD error in equation (4), once you add in r and include a discount factor.

In VGL, it is the *sideways* components of $\frac{\partial V}{\partial s}$ that are important, those that are *not parallel* to Δs. Such components are used in the calculation of G', in the terms $\frac{Df}{Ds}$ and $\frac{Dr}{Ds}$ in particular. That is, you have to know these terms, that constitute the model function, in order to calculate a target value gradient, and you need a target in order to do a weight update.

In addition, the model function is relevant to the greedy policy. Using a first order expansion of the greedy policy gives

$$\pi(s,\vec{w}) = \arg\max_a \left(r(s,a) + \tilde{V}(f(s,a),\vec{w})\right) \tag{14}$$

$$\approx \arg\max_a \left(r(s,a) + \tilde{V}(s,\vec{w}) + \left(\frac{\partial \tilde{V}(s,\vec{w})}{\partial s}\right)^T (f(s,a) - s)\right)$$

$$\approx \arg\max_a \left(r(s,a) + \left(\frac{\partial \tilde{V}(s,\vec{w})}{\partial s}\right)^T (f(s,a) - s)\right)$$

Hence the greedy policy depends on the value gradient but not on the values themselves. This is critical since changing $\frac{\partial \tilde{V}}{\partial s}$ will immediately affect the greedy policy; by

moving it towards its correct target we will steer the trajectory in the correct (locally optimal) direction: TD's paradigm "exploration *vs.* exploitation" becomes "exploration *and* exploitation" or, in other words, exploration comes for free when we combine "greedy" and "gradient" in VGL.

VGL is an extension of well-known methods in adaptive dynamic programming, Dual Heuristic Programming and Generalized Dual Heuristic Programming in particular, that have been proved to be successful in solving complex tasks such as autopilot landing, power system control, simple control benchmark problems such as "pole balancing", and many others (Wang, Zhang & Liu 2009). From a psychological point of view, VGL(1) is equivalent to Rescorla and Wagner's model. However, where as Rescorla and Wagner's model only considers classical conditioning VGL works for instrumental tasks –VGL, so to speak, is Rescorla and Wagner's model applied to operant conditioning.

Value-Gradient Learning and Pontryagin's Minimum Principle

In the next section we state how Hamilton's principle (aka the Principle of Least Action) and VGL apply to learning and behaviour. But first we need to be more precise about the relationship between VGL and Pontryagin's Minimum Principle.

As defined by Pontryagin, the Hamiltonian of a control system is a function of four variables: $\mathcal{H}(s,p,a,t) = \mathcal{L}(s,a,t) + p_t^T f(s,a,t)$ where $p_t = -\frac{\partial \mathcal{H}}{\partial s}$ is a costate interpreted as a Lagrange multiplier: If the state given by the function represents constraints in the minimization problem, the costate represents the cost of violating those constraints. In other words, p is the rate of change of the Hamiltonian as a function of the constraint. For example, in Lagrangian mechanics, the force on a particle $F = -\nabla V$ can be interpreted as p determining the change in action (transfer of potential to kinetic) following a variation in the particle's constrained trajectory. In economics, the optimal profit is calculated according to a constrained space of actions, where p is the increase in the value of the objective function due to the relaxation of a given constraint –the marginal cost of a constraint, called the shadow price.

Intuitively, the constraint f can be thought of as competing with the desired function to pull the system to its minimum or maximum (or to a steady state). And the Lagrange multiplier p can be thought of as measure of how hard f has to pull in order to make those forces balance out in the constraint surface.

Pontryagin's Minimum Principle (PMP) states that $\mathcal{H}(s_t^*, a_t^*, p_t^*, t) \leq \mathcal{H}(s_t^*, a_t, p_t^*, t)$ with the associated conditions for a maximum, namely, $p_t = -\frac{\partial \mathcal{H}}{\partial s}$, $s_t = \frac{\partial \mathcal{H}}{\partial p}$, and $\frac{\partial \mathcal{H}}{\partial a} = 0$. How is this related to VGL? Taking $\mathcal{H} = \mathcal{L} + pf$, we can make it correspond to VGL as follows: $\mathcal{L}$ is the quantity to be maximized (or minimized), that is, the cummulative reward; the constraints are defined following the model of the world, f and r (henceforth, f for short); and p is G' (obviously $\tilde{G}$ if the trajectory is optimal, that, is if $G' = \tilde{G}$). Hence we can express VGL in Hamiltonian form as $\mathcal{H} = r + G'f$. In fact, our re-formulation of PMP is somehow simpler, since PMP's conditions are reduced to two, namely, the costate and the max function that defines the greedy policy. At the end of the day, PMP can be described as $a^*(s,p) = \arg\min_a \mathcal{H}(s,p,a)$, which is a form of the greedy policy, and the adjoint equation $p_t \cong \frac{\partial V}{\partial s}(s_t,t)^T \in \Re^n$, our gradient.

Let's recapitulate and see what happens with traditional value-based approaches: if there is no model, the Hamiltonian will not be constrained, thus it will be left to try all possible actions, not just those which "follow" the constraints. Indeed: without f, $\mathcal{H} = r + V'f$ reduces to $\mathcal{H} = r + V'$ –the old V formula.

Value-Gradient Learning and Behaviour Systems

To summarize, we have restored optimality. If we learn the gradient of the value function by choosing greedy actions that follow the full model of the system, Pontryagin's Minimum Principle applies and the trajectory so built is guaranteed to be locally optimal, that is, to minimize the error and to maximize the reward. This analysis begs the question: How does VGL apply to the study of behaviour?

At the end of the day, animals are behaviour *systems* – sets of behaviours that are organized around biological functions and goals like feeding (Timberlake 1983), defence (Fanselow 1994) or sex (Domjan 1994). When such systems are free to act as they please, their preferred or optimal distribution of activities defines a behavioural bliss point (BBP) or baseline level of activity. In dynamic terms the BBP is a natural, steady and stable, attractor.

This view encapsulates the behavioural regulation theory and generalizes the concept of homeostasis and negative feedback from physiology to psychology. Physiological homeostasis keeps parameters such as body temperature close to an optimal or ideal level. This level is "defended" in that deviations from the target temperature trigger compensatory physiological mechanisms that return the system to its homeostatic levels. In behavioural systems, what is defended is the organism's BBP against instrumental contingencies that create disturbances to which the system adapts. Other metaphors are possible: At the end of the day, the bliss point represents an equilibrium in a population of behaviours –pretty much as the equilibrium observed in the number of different types of ants in a colony or between competing (prey-predator) species in an environment.

More specifically, Staddon's model (Staddon 1979) explains operant behaviour in terms of time constraints and feedback constraints, the reinforcement schedule to which the animal is subjected. Starting from a BBP, the animal finds the optimal equilibrium between instrumental and contingent responses –the one that minimizes the cost involved. Instrumental conditioning procedures are seen as response constraints that disrupt the free choice of behaviour and prevent the organism from returning to the BBP. The organisms achieve a contingent optimization by

approaching its bliss point under the constraints of the instrumental conditioning procedure. Put it this way, the analysis of operant behaviour is an optimal control problem and thus we should be able to express it in terms of VGL: $\mathcal{L}$, the Lagrangian, is defined as the cost to be minimized, f are the time and feedback constraints and p, the multiplier or conjugate momentum, is now explicitly represented as G'. Not surprisingly, this formulation matches Staddon's term by term (see Appendix A, Staddon 1979).

Let's recapitulate, VGL's G value would be the gradient of the cost associated with a departure from a given distribution of actions. If the cost of a given distribution is represented as V, then $G(s) = \frac{\partial V(s,a)}{\partial s}$ represents the change in cost as we change the distribution –where s represents the distribution and a represents a given set of responses (both instrumental and contingent). G' re-acts against the constraints to minimize the cost.

What are the advantages of using VGL? Firstly, VGL tells us exactly which form the multiplier must have. In particular, G' must be defined according to $\frac{Dr}{Ds}$ and $\frac{Df}{Ds}$: the former tells us how the rate of contingent responses (r) changes as the distribution of responses changes and the latter how the constraints themselves change. These two quantities define the change of cost that we minimize and give us the optimal distribution.

Perhaps more importantly, VGL does not only give a solution to an optimization problem –in this case, the optimal distribution of responses under certain constraints. Of course, it does if we assume that such functions are perfectly known; yet, VGL is also a learning algorithm and as such serves a mechanistic agenda as well as an equilibrium agenda. VGL allows us to calculate how the animal is adapting to the optimal distribution when the constraints are a *moving* target, solving the so-called "teleological conundrum": of course, animals do not know what the reinforcement schedule would be or the corresponding optimal response ratio –and yet they adapt to the optimal solution and they do so in an optimal way. Perhaps an analogy may clarify this point: Physicists found it puzzling that particles behaved as if they knew what the future would be. Traditionally, the movement of particles was interpreted in terms of global symmetries and thus it was difficult to explain how particles abided by the Principle of Least Action locally, when constraints appeared and disappeared as the system interacted with "unexpected" forces. Surely, the symmetries were broken in such cases; and yet, Nature seemed to account for them so as to comply with global symmetries –"as if nothing had happened", symmetry was restored. We know that the answer lies in *gauge* symmetries: Indeed, at each step, deviations are counter-balanced so as to bring the system back (or as close as possible) to the original symmetry. In terms of cognition, this is precisely what VGL does.

Conclusion

This paper does not present quantitative predictions or new results. It presents a formal model that integrates current theories of conditioning with fundamental principles of Nature. Our main assumption is that learning and behaviour, conditioning more in particular, follow the same variational principles as any other natural phenomena: they must make a functional of some sort of extremal. In that we follow Peter Killeen's program (Killeen 1992). We have shown that Temporal Difference is an inadequate model of optimal behaviour and proposed a new model, Value-Gradient Learning, equivalent to Pontryagin's Minimum Principle –in turn, a version of Hamilton's Principle of Least Action, that may serve as a model of both classical and operant conditioning.

References

Alonso, E., & Schmajuk, N. (2012). (Eds.), Special Issue on Computational Models of Classical Conditioning, *Learning & Behavior*, *No.: 3*.

Bellman, R. (1957). *Dynamic Programming*. Princeton, NJ: Princeton University Press.

Domjan, M. (1994). Formulation of a behavior system for sexual conditioning. *Psychonomic Bulletin & Review*, *1*, 421-428.

Fairbank, M., Prokhorov, D., & Alonso, E. (2012). Approximating Optimal Control with Value Gradient Learning. In J. Si, A.G. Barto, W.B. Powell & D. Wunsch (Eds.), *Handbook of Learning and Approximate Dynamic Programming*, *Vol. 2*. Wiley-IEEE Press.

Fanselow, M. S. (1994). Neural organization of the defensive behavior system responsible for fear. *Psychonomic Bulletin & Review*, *1*, 429-438.

Glimcher, P.W., Camerer, C.F., Fehr, E., & Poldrack, R.A. (2009). *Neuroeconomics: Decision Making and the Brain*. San Diego, CA: Academic Press.

Killeen, P.R. (1992). Mechanics of the animate. *Journal of the experimental Analysis of Behavior*, *57*, 429-463.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience, 16*, 1936-1947.

Pontryagin, L.S., Boltyanskii, V.G., Gamkrelidze, R.V., & Mishchenko, E.F. (1962). *The Mathematical Theory of Optimal Processes*. New York, NJ: Gordon and Breach Science Publishers.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning: II. Theory and research* (pp. 64-99). New York: Appleton-Century-Crofts.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, *36*, 241-263.

Staddon, J.E. (1979). Operant behavior as adaptation to constraint. *Journal of Experimental Psychology: General*, *108*, 48-67.

Sutton, R. S., & Barto, A. G. (1987). A temporal difference model of classical conditioning. In *Proceedings of the Seventh Annual Conference of the Cognitive Science Society* (pp. 355-378). Erlbaum.

Tesauro, G.J. (1994). TD-Gammon, a self-teaching backgammon program, achieves master level play. *Neural Computation*, *6*, 215-219.

Timberlake, W. (1983). Rats' responses to a moving object related to food or water: A behavior-systems analysis. *Animal Learning & Behavior*, *11*, 309-320.

Wang, F.-Y., Zhang, H., & Liu, D. (2009). Adaptive dynamic programming: An introduction, *IEEE Computational Intelligence Magazine*, *May*, 39–47.

Watkins, C.J.C.H. (1989). *Learning from Delayed Rewards*. Ph.D. Thesis, Cambridge University.