Reward Algorithm

Gail R. Berger

November 24, 2009

Introduction

0.1 Rescorla and Wagner

0.1.1 Reward Learning Model

Many psychologists, neuroscientists, and computer science specialists have developed equations that reflect probabilities underlying reward processes. Rescorla & Wagner [4] were one of the earliest professionals to develop a conceptualization of reward learning as emanating from changes in the associative strength and valuation of a rewarding stimulus and rewarding task. As cited by Rescorla & Wagner [4] their model was strongly influenced by Hullian mathematical model [2], which posited that changes in response probability for a given trial ($\triangle p_n$) emanated from a learning rate parameter (β) times the learning asymptote (λ) less the probability of a response in a given trial (p_n).

$$\triangle p_n = \beta(\lambda \text{ - } p_n)$$

They expanded the Hullian Model by focusing on the interplay of rate parameters relating to stimulus salience and learning, the associative strength inherently in the US and the US's influence on the CS. Accordingly, changes in the associative strength and valuation of a rewarding stimulus and rewarding task, ($\triangle\ V_{AX}$) emanate from the magnifying power of learning rate parameters associated with stimulus salience ($\alpha_{A \text{ or } X}$) and learning (β_n) associated with the unconditioned stimulus (US) and conditioning contexts times the asymptotic level of the US's associative strength (λ_n) less the immediate value of the associative strength of specific contextual stimuli (A and X) associated with the rewarding US stimulus (V_{AX}).

The difference between (λ) and (V_{AX}) reflected CS's ability for taking on US's rewarding qualities and ability for predicting later US's occurrence.

i.e. the smaller the difference between λ_n and V_{AX}, the greater the CS's associative strength. Conversely, the larger the difference, the smaller the associative strength , and the smaller the CS rigor for predicting later US occurrence. The accuracy of the conditioned response (CR) (and evidence of accelerated learning) was reflected in reduced distance between both values λ_n and V_{AX}. The greater the CS's associative strength, the greater its ability at predicting later US occurrence. In this indirect sense the reduced distance between λ_n and V_{AX} is also reflective of enhanced probability of US occurrence.

With nonreinforcement, Rescorla & Wagner [4] noted that there was a reduction in US's, λ_n, availability and this reduced the organism's motivation to work toward obtaining reward. Accordingly, with the reduction in US's availability, the CS, V_{AX}, lost its associative strength in response to the reduction of pairing between the two.

According to Rescorla & Wagner many trials with many different conditioned contextual stimuli (including the CS), AX, may be reflected accordingly.

$$\triangle V_A = \alpha_A \beta_1 (\lambda_1 - V_{AX}) \text{ in trial 1.}$$

$$\triangle V_X = \alpha_X \beta_1 (\lambda_1 - V_{AX}) \text{ in trial 1.}$$

$$\triangle V_A = \alpha_A \beta_2 (\lambda_2 - V_{AX}) \text{ in trial 2.}$$

$$\triangle V_X = \alpha_X \beta_2 (\lambda_2 - V_{AX}) \text{ in trial 2.}$$

If we take this analysis one step further, we may conclude, that the value of task reward learning, in total, can be summarized as follows.

$$\triangle V_{AX} = \sum_{i=1}^{n} \alpha_{AX} \beta_n (\lambda_n - V_{AX})$$

Whereby, reward task learning is the sum of stimulus attributes (i.e. US's and associated CS and contextual stimuli salience and learning rate parameters) times the asymptomic difference between US and CS values.

Rescorla and Wagner [4, p.74-75] identified three central notions underlying their model. The first, as noted above, modified and elaborated on Hullian theory, which addressed key points of a learning parameter (β), asymptote of learning (λ), and probability of response (p_n) [2]. The second central

notion suggested that "organisms only learn when events violate their expectations." Accordingly, expectations generated by the US and associated CS and conditioning contexts are "only modified when consequent events disagree with the composite expectation." The third central notion proposed that conditioning processes are dependent on the associative strength of all stimuli occurring during a trial with the US. Conditioned response (CR) changes to a CS are limited to the maximal generated UR, which a particular US can elicit. Rescorla & Wagner [4, p.75] concluded that their model describes "learning curves for strength of association, not response probability" (as suggested in central notions 1 and 2).

0.2 Sutton and Barto

0.2.1 An Interpretation of Rescorla & Wagner

Sutton and Barto's [6] conceptualization of Rescorla & Wagner's model centered on Rescorla & Wagner's central notion number 2, namely that learning occurs in situations where events violate expectations. Accordingly, the US (λ) represents "the actual US level on a trial" [6, p.519](as opposed to Rescorla & Wagner's US's maximal associative strength) and $\bar{V}$ as representing the expected or predicted value (in contrast to Rescorla & Wagner's CS's maximal associative strength it could attain in its temporal, proximal association with the US).

Where the Rescorla & Wagner model examines the relationship between the US-CS, Sutton & Barto's [6, p.502] conceptualization of Rescorla & Wagner's $\lambda - V_{AX}$ centered on the discrepancy between λ - $\bar{V}$, "where(by) λ represents the actual US level on a trial and $\bar{V}$ represents the expected or predicted level".

Accordingly, Sutton & Barto's [6] interpretation evolved in the following manner.

$$\triangle V = (\text{Level of } \underline{\text{US}} \text{ Processing}) \times (\text{Level of } \underline{\text{CS}} \text{ Processing})$$

and

$$\triangle V = \text{Reinforcement} \times \text{Eligibility}$$

Where the change in reinforcement value evolves from the US's ability for reinforcement and the eligibility or accessibility for reinforcement it grants to the CS in response to their temporal, proximal and associative relationship. As cited above Rescorla & Wagner's model is as follows.

$$\triangle V = \alpha\beta(\lambda - V_{AX})$$

Sutton & Barto [6] interpreted the above as follows.

$$\triangle V_i = \beta\,(\lambda - \bar{V}) \;\times\; \alpha_i X_i$$

Their reinterpretation can be further delineated as follows.

$$\triangle V_i = \underbrace{\beta\,(\lambda - \bar{V}\,)}_{\text{REINFORCEMENT}} \times \underbrace{\alpha_i\, X_i}_{\text{ELIGIBILITY}}$$

According to Sutton and Barto's [6] interpretation of Rescorla & Wagner's [4] , $\beta\ (\lambda - \bar{V})$ was associated with the US. β was a positive constant and $\lambda - \bar{V}$ reflected the difference and discrepancy between expected US occurrence less its actual occurrence. The salience rate parameter (α) was associated with CS's salience. In contrast, Rescorla & Wagner associated both rate parameters for stimulus salience (α) and learning (β) with both the US & CS. Therefore one major difference between both models stemmed from their respective interpretation of rate parameters or constants

A second major difference between the two centers on their respective conceptualization of V_{AX} and $\alpha_i X_i$. Rescorla & Wagner's V_{AX} represented the CS and the entire conditioning context and learning space. Sutton & Barto's $\alpha_i X_i$ represented the *eligibility* granted to the CS by the US in response to its temporal and proximal association. According to Sutton & Barto [6, p.505] *stimulus eligibility trace* offered and is evidenced in an organism's or subject's attention, perceived stimulus salience, concept generalization, perceived contrast, stimulus learning (memory) traces, and other aspects of CS representations.

A third major difference between both models stems from their respective interpretation of $(\lambda - V_{AX})$ and $(\lambda - \bar{V})$. Rescorla & Wagner associated $(\lambda - V_{AX})$ with the US and CS, respectively. They monitored the extent by which the CS takes on US qualities and the evidentiary difference between the two. The smaller the difference they asserted, the larger the associative strength and the larger the CS rigor for predicting later US occurrence and vice versa. The accuracy of the conditioned response (CR) and evidence of accelerated learning are thus reflected in reduced distance between both values.

Sutton & Barto [6] interpreted Rescorla & Wagner's $(\lambda - V_{AX})$ as being a discrepancy "between expected and actual US events $(\lambda - \bar{V})$. (Sutton & Barto [6, p.502] suggested that) they (Rescorla & Wagner) denoted this discrepancy $\lambda - \bar{V}$ where, λ represents the actual US level on the trial and $\bar{V}$ represents the expected or predicted level. The predicted level, $\bar{V}$, is a composite or total prediction depending upon the associative strength of all

the CSs present in the trial. . . If training is continued with the same CSs, then their composite prediction, $\bar{V}$, should approach λ.".

Sutton & Barto [6, p.503]assessed a weakness in the Rescorla & Wagner's model and referenced its lack of representation of second order-conditioning (i.e. stimuli that are in temporal proximity with the CS). However, Rescorla & Wagner incorporated all stimuli associated with the conditioning context and the CS into their V_{AX} value and, as a result, did indeed account for instances of second order conditioning.

Furthermore, Sutton & Barto questioned the negative value attained from $\lambda - \bar{V}$. When experimentally implemented the negative value derived from this formula seemed to decrease with CS prediction accuracy. As $\bar{V}$ approached its valuation to λ, it was never really possible for $\bar{V}$ to reach numerical valuation of zero or 1. In addition they cited that when the US (λ) was not available during a trial, the CS ($\bar{V}$) became a negative number. But this assertion is weakened when one recalls that the formula, $\lambda - V_{AX}$, suggested the relationship between the US and the CS and the nature of the associative strength (however manifested) between both, not the presence or absence of either in any specific trial.

0.2.2 Time Derivative Model of Pavlovian Conditioning

As Rescorla & Wagner, Sutton & Barto were influenced by Hull approach to reward learning theory. Like Hull [3], they believed that reward learning analysis should include an approach to equation development that emphasizes probability and prediction. They believed in the importance of a theoretical reference to the organism's or subject's perception of the interval between CS offset and US onset (or the temporal, proximity between the US-CS) [5, 6]. Hull called this interval *stimulus trace*; Sutton & Barto called it *eligibility trace*. Instead of monitoring task-related parameters, Sutton & Barto reasoned that a representation and objectification of a theoretical external reward learning state more accurately depicted processes underlying the later generation of task-related behavior.

Sutton & Barto [5] initially sought to clarify on their reward learning model when they developed their $\dot{Y}$ ("Y dot") theory of Pavlovian reinforcement. Accordingly, Y represented the perceived associative strengths of all stimuli (e.g. US, CS, and stimuli in the conditioning context) present during the reward learning task. $\dot{Y}$ represented the perceived change in time interval. These relationships are noted below.

$$\dot{Y}(t) = Y(t) - Y(t - \triangle t)$$

Whereby, the change in associative strength in all stimuli at time (t) is the sum of associative strengths of all stimuli less the difference between the associative strength of stimuli at one point in time and less the time the change occurred. The difference, $Y(t - \triangle t)$, yields the time that the task stimuli exerted some influence at mediating future reward learning response or the change in learning response. Succinctly stated, the change in associative strength is the difference between the total associative strength of all stimuli at time (t) less the impact of recent learning on the total.

When Sutton & Barto [6] integrated both equations, i.e. the $\dot{Y}$ theory of reinforcement with the $\bar{X}$ eligibility trace, the following time derivative model was developed.

$$\triangle V_i = \beta \dot{Y} \ \times \ \alpha_i X_i$$

Whereby the organism's or subject's change in reinforcement value evolves from US's actual occurrence evidenced in β and the changes in its associative strength $(\dot{Y})$ as well as the interaction between the CS's (α_i) and the eligibility trace afforded to it by the US's X_i. Unlike the Rescorla & Wagner Model, the Sutton & Barto model [6] did not examine the relationship between the US and CS as well as the US relationship with the conditioning context (V_{AX}). It did not, as well, monitor the rate of learning or salience. But due to its emphasis on predictability, the Sutton & Barto model seeks to represent the external state of the the reinforcement environment, monitors theoretical expected outcomes and a task's trial's actual outcome, and monitors the changes in state responses over progressive time periods.

0.2.3 Reinforcement Learning

When Sutton & Barto [1] developed their own model for Reinforcement Learning, their theoretical framework included references to *discrete time* (t_1), i.e. time intervals or separate events units, *payoffs*, i.e. incentive for producing certain actions (a_n), *state* information (s_t), i.e. external state representations "of complex world models and memories of past sensations and behaviors" $(s_{_{t-1}})$ (p.548), and finally *optimality*. *Optimality* reflected the tendency for current payoffs to retain their full valuation in response to their immediacy (r_t). Future payoffs $(r_{n+1}$ or $R)$ tend to increasingly reduce their valuation and are subject to the effects of incremental discounting (γ^n and $\gamma^n r_{n+1}$). This discount rate is facilitated by the Present Value of future rewards [7]. This is evidenced in the characterization below. Whereby, more immediate returns (r_1) are fully valued, but those in the future ($\gamma^2 r_3$ and $\gamma^n r_{n+1}$) are experienced as decreasing in valuation in response

to increasing discounting, where $0 \leq \gamma < 1$ [1, p.552].

$$r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots \gamma^n r_{n+1} \cdots .$$

Barto,Sutton, &, Watkins [1] indicated that the state of any system is the sum of its past and current state. This can be depicted accordingly.

$$S_{sy} = \sum_{i=1}^{n} s_{t-1} + s_t \ldots \ y, \text{ where } s_{t+1} = y, sy \in S \mid s_t = s.$$

However, Sutton & Barto [1, 551] asserted that the system state in total (S_{sy}), including future input, will determine probabilities and the basis for characterizing action, irrespective of how that state system evolved and emerged. The organism or subject *observes* this state system and performs an action, which precipitates later system delivery of a return or payoff. This action triggers a transition to a new external state. The organism or subject *observes* this new state engages in another new action and receives another new payoff and transition to another state.

$$Prob\{s_{t+1} = y | s_0, a_0, s_1, a_1, \ldots, s_t = s, a_t = a\}$$

$$= Prob\{s_{t+1} = y | s_t = s, a_t = a\} P_{sy}(a)$$

The future probability for transitioning from one initial state (s_t) to a future (reward) state (y_t) evolves from the initial basal state and subsequent action and then later sequentially generated states and actions. Each state (s_t) is in a discrete time period (t_n) and is elicited from a specific action (a_n), seeks to obtain a payoff or return or primary reinforcement (r_t). The task-related goal is to select actions $(a_0, a_1, a_2, \ldots, a_n)$ with associated previous and subsequent states $(s_0, s_1, s_2, \ldots, s_n)$ that can increase the probability (Prob$\{s_{t+1}\}$) for obtaining a total and cumulative amount of payoff or (future) return (r_{n+1}). The outcome of this behavioral sequence is a transition to a new state (from state s to state y) having a probability of $P_{sy}(a)$ as a result of the agent's action (a_i). This process forms a Markov Chain with transition state probabilities where $P_{sy} = P_{sy}(\pi(s))$.

Each action that elicits a later payoff is based on a decision or policy (π), which at a certain previous point in time had been generated from a theoretical perceptual and spatial analysis of the state of the task and reward-related perceptual stimuli (i.e. its temporal proximity, intensity, valence,

meaningfulness, etc.). The policy (π) is a "mapping from each state and action..." [7, p.68]

During each trial, the expected (E) future return (r_{t+1}) evolves from the state (s_t) and action (a_t) at a discrete time period (t). The expected value of the future return (r_{t+1} or R) is associated with state and action information. This can be summarized as follows.

$$E[r_{t+1}|s_t a_t] = R(s_t, a_t)$$

A selected policy (π) is often valued (V) and associated with reward (r) and the reward state (s) and can be summarized as $V^{\pi}(s)$. $V^{\pi}(s)$ can lead to future outcomes for future reward (r_{t+1}) and associated state (s_{t+1} or state y). This expected future return state ($E_{r_{t+1}}(s)$) is also discounted (γ) as a bird in hand (reward) is worth more than $2, 3, 4, \ldots, n$ in a bush. Accordingly, the value (V) of a policy or strategy (π) that is associated with the reward state (s) is the expected policy's (E_{π}) sum of discounted, total future returns (r_{t+1}).

$$V^{\pi}(s) = E_{\pi}[\sum_{t=0}^{\infty} \gamma^t r_{t+1}|s_t = s]$$

And the value (V) of the policy or strategy for experiencing the reward state ($V^{\pi}(s)$) is the expected policy's (E_{π}) future return and discounted sum of discounted future returns (r_{t+1}).

$$V^{\pi}(s) = E_{\pi}\{\ r_{t+1} + \gamma \sum_{t=0}^{\infty} \gamma^t r_{t+1}|s_t = s\}$$

As with the Bellman Equation the value [7, p.68-69]of a policy is a state, which is associated with all policies and respective reward states and actions, all probabilities and associated with actions facilitating transitions from states s to y, future returns associated with actions that allow transitioning from one state (s) to another (y), and the discounted expected policies (γE_{π}) sum of discounted future returns ($\gamma^t r_{t+1}$).

$$V^{\pi}(s) = \sum_{a} \pi(s, a) \sum_{y} P^{a}_{sy} \left[R^{a}_{sy} + \gamma E_{\pi}\{\sum_{t=0}^{\infty} \gamma^t r_{t+2}|s_{t+1} = y\}\right]$$

The value of the policy or strategy is a state, which is associated with all policies and reward states and actions, all probabilities and actions facilitating transitions from states s and y, future returns associated with actions that transition from one state (s) to another (y), and the value of the discounted valued policies (γV^{π}) that are associated with future returns states.

$$V^{\pi}(s) = \sum_{a} \pi(s,a) \sum_{y} P^{a}_{sy} \left[R^{a}_{sy} + \gamma V^{\pi}(y) \right]$$

0.2.4 Temporal Difference Prediction (λ)

As noted in Section 0.2.1, the formula, $\lambda - \bar{V}$, reflected the difference and discrepancy between an expected US occurrence (λ) less its actual outcome ($\bar{V}$). This difference fundamentally represented the learning process and the ability for predicting a future occurrence for reward.

According to Sutton & Barto [6, 7] prediction can be expressed through two different models, the Monte Carlo Method and the Temporal Difference Model. The Monte Carlo Method is based on previous reward state information observed during a completed sequence of trials within an episode. Knowledge of the reward and state information is therefore known prior to subsequent problem analysis. Accordingly, the Monte Carlo Method uses this model to base its analysis and future prediction. In order to calculate a prediction, some prior knowledge of the state and return is needed before problem analysis. The valued state at time, t, can be estimated from prior knowledge of the valued state from an earlier episode, $s_{_{t-1}}$, a step-state parameter, α, and the total future return R_t, and time, t, less the value of the state at time, t.

$$V(s_t) \leftarrow V(s_{t-1}) + \alpha[R_t - V(s_t)]$$

The Temporal Difference Model [7] does not require prior knowledge of the environmental model. It bases its estimate on a previous trial within a specific task or problem. By analyzing what was learned in the prior trial, one can generalize, estimate, and predict what will later happen. As such, preliminary knowledge of the environment is not necessary in the Temporal Difference Method's prediction calculation. The valued state at time, $V(s_t)$, can be calculated from a small sample (e.g. single trial reward-state information) representing the valued state, $s^{\frac{1}{n}}$, a step-state parameter, α, the assessed future return, $r_{_{t+1}}$, and the discounted value of the future state,

$\gamma V(s_{_{t+1}})$, less the value of the state at time, $V(s_t)$. $V(s_{_{t+1}})$ reflects the end of the task or the problem.

$$V(s_t) \leftarrow V(s_t^{\frac{1}{n}}) + \alpha[r_{t+1} + \gamma V(s_{t+1}^n) - V(s_t)]$$

The Monte Carlo Method [7] is an off-policy method, which when implemented, compromises ongoing task-mediated exploration and exploitation due to analysis that draws from prior representations. Sutton & Barto developed a TD on-policy method, the Sarsa On-Policy Method, which allows for intermittent exploration, problem-solving, and on-line evaluation. Like the Temporal Difference Model, the Sarsa On-Policy Method does not require prior knowledge of the environmental model, bases its analysis on an element or sample within the task or problem, generalizes those results, etc. Accordingly, the on-line valued state and action for attaining reward at time, $Q(s_t, a_t)$, can be derived from a small sample representing the valued state and action (from a single trial) or $s^{\frac{1}{n}}$ or $a^{\frac{1}{n}}$, a step-state parameter (α) the assessed future return ($r_{_{t+1}}$) and the discounted value of selecting future action and the resultant state ($\gamma Q(s_{_{t+1}}, a_{_{t+1}})$) less the actual state and action at time ($Q(s_t, a_t)$). $Q(s_{_{t+1}}, a_{_{t+1}})$ reflects the end of the task or the problem. Therefore, $(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1})$ suggest a quintuple of intervals, which transition from one state-action to another. Manifestation of Q^π is suggested in the interaction between s_t and a_t.

$$Q(s_t, a_t) \leftarrow Q(s_t^{\frac{1}{n}}, a_t^{\frac{1}{n}}) + \alpha[r_{t+1} + \gamma Q(s_{t+1}^n, a_{t+1}^n) - Q(s_t, a_t)]$$

According to Sutton & Barto [7], TD methods represent a *policy* (π) independent of the value function (V). The TD's policy underlies later action selection. The *estimated value function* (γV) "criticizes ...actions made..." and, as such, provides feedback for a selected action's accuracy (and the underlying policy's accuracy) at matching to and satisfying task requirements. This is reflected in a temporal difference error (δ). This scalar signal is the only manifested output representing the interaction between π and its estimated value function, γV. When $\delta \leq 0$ and the TD error is negative, the selected action is judged inaccurate, and should be abandoned. If $\delta > 0$ and the TD error is positive, the selected action is considered accurate and should be utilized. As such, the TD error can evaluate and is a manifestation of the new state.

$$\delta = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$$

Such that the temporal difference error (δ) at time (t) is the future return and the discounted value of the future state less the value of the state at time t.

The temporal difference model may be a useful tool for understanding the nature and magnitude of emotion. Subsequent discussion will examine and analyze existing models, which seek to explain sensori-emotional appraisal and emotion. Later discussion will seek to integrate previously discussed concepts into a model, which can be used for understanding the role of reward and the impact of its disruption on the chronic stress response.

References

[1] Andrew G. Barto, Richard S. Sutton, and Christopher J. Watkins. Learning and sequential decision making. In M. Gabriel and J.W. Moore, editors, *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, pages 539–602. MIT Press, Cambridge, Massachusetts, 1990.

[2] Robert R. Bush and Frederick Mosteller. *Stochastic Models for Learning*. Wiley, New York, New York, 1955.

[3] Clark L. Hull. *Principles of Behavior*. Appleton-Century-Crofts, New York, New York, 1943.

[4] Robert A. Rescorla and Allan R. Wagner. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black and W.F. Prokasy, editors, *Classical Conditioning II: Current Research and Theory*, pages 64–99. Appleton-Century-Crofts, New York, New York, 1972.

[5] Richard S. Sutton and Andrew G. Barto. Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review*, 88(2):135–170, 1981.

[6] Richard S. Sutton and Andrew G. Barto. Time-derivative models of pavlovian reinforcement. In M. Gabriel and J.W. Moore, editors, *Learning and Computational Neuroscience: Foundations of Adapative Networks*, pages 497–537. MIT Press, Cambridge, Massachusetts, 1990.

[7] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, Massachusetts, 1998.