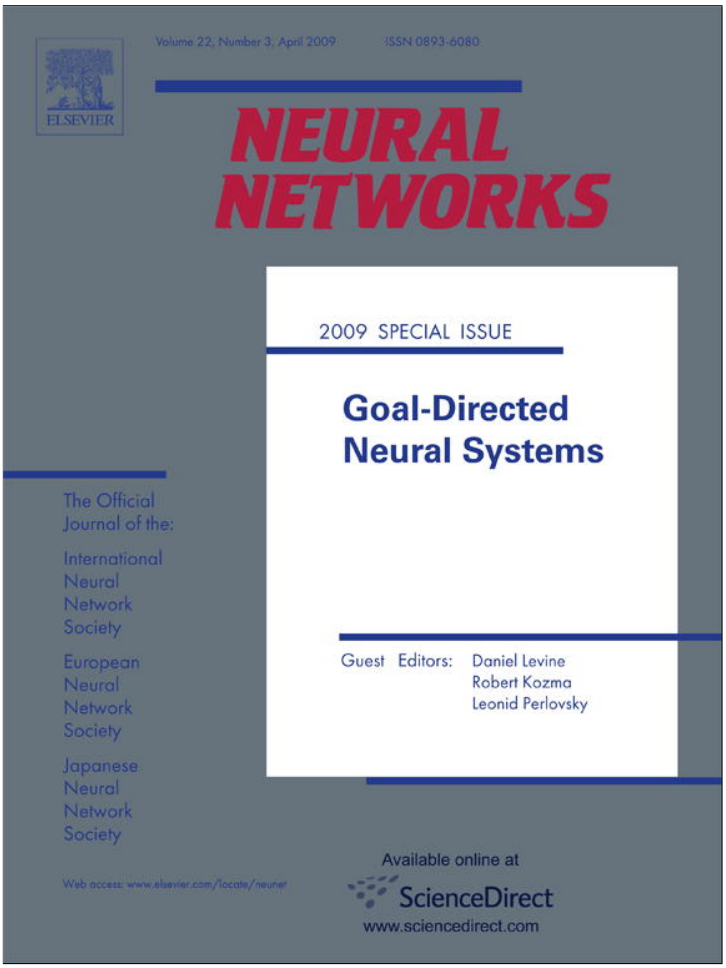

Neural Networks 22 (2009) 229–236

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2009 Special Issue

Adaptive learning via selectionism and Bayesianism, Part II: The sequential case

Jun Zhang

Department of Psychology, University of Michigan, 530 Church Street, Ann Arbor 48109-1043, USA

ARTICLE INFO

Article history:
Received 13 February 2009
Received in revised form 15 March 2009
Accepted 21 March 2009

Keywords:
Sequential decision
Credit-assignment problem
Incentive value
Conditioned reinforcement value
Actor–critic
Reinforcement learning

ABSTRACT

Animals increase or decrease their future tendency of emitting an action based on whether performing such action has, in the past, resulted in positive or negative reinforcement. An analysis in the companion paper [Zhang, J. (2009). Adaptive learning via selectionism and Bayesianism. Part I: Connection between the two. *Neural Networks, 22*(3), 220–228] of such selectionist style of learning reveals a resemblance between its ensemble-level dynamics governing the change of action probability and Bayesian learning where evidence (in this case, reward) is distributively applied to all action alternatives. Here, this equivalence is further explored in solving the temporal credit-assignment problem during the learning of an action sequence ("operant chain"). Naturally emerging are the notion of secondary (conditioned) reinforcement predicting the average reward associated with a stimulus, and the notion of actor–critic architecture involving concurrent learning of both action probability and reward prediction. While both are consistent with solutions provided by contemporary reinforcement learning theory (Sutton & Barto, 1998) for optimizing sequential decision-making under stationary Markov environments, we investigate the effect of action learning on reward prediction when both are carried out concurrently in any on-line scheme.

1. Introduction

Operant conditioning is the learning process in which the frequency of occurrence of a behavior is modified by the consequences of such behavior. Actions that produce a satisfying (discomforting) effect in a particular situation become more (less) likely to be emitted again under such situation in the future. This so-called "Law of Effect", as first suggested by Thorndike (1898), governs operant learning. In most cases, however, the effects of an action (whether reward or punishment) do not immediately follow its execution; a sequence of actions must be completed before a reinforcing or aversive event finally appears as a consequence of the animal having executed the entire action sequence. The fundamental challenge here is how to modify the action probability of intermediate components of an operant sequence in the absence of any accompanying primary reinforcement, the so-called temporal "credit-assignment problem". Empirically, how chaining of actions in a sequence is accomplished based on a terminal, primary reinforcement has been studied extensively in the operant conditioning literature (see e.g., (Catania, 1968; Reynolds, 1968)).

Conceptually, an operant consists of a behavior or action R_i ($i = 1, \ldots, N$) that is (1) drawn from an action repertoire $\mathcal{R} = \{R_1, R_2, \ldots, R_N\}$ with certain probability $\mathrm{p} = (p_1, \ldots, p_N)$, $p_i \geq 0$; $\sum_i p_i = 1$; (2) emitted under a given context called discriminative stimulus S; and (3) resulted in a reinforcer r_i with reward value θ_i. During conditioning, action probabilities p_i ($i = 1, \ldots, N$) increase or decrease, depending on the reinforcement structure $\{\theta_i, i = 1, 2, \ldots, N\}$. As in the companion paper (Zhang, 2009), we only deal with positive reinforcement $\theta_i \geq 0$. After conditioning, the best rewarded action among this set of possible actions will be selected and executed, with probability approaching 1.0, whenever the animal is under environmental context S. The events S–R–r in proper order represent a basic operant unit. Note that the discriminative stimulus or context S itself then acquires a reinforcement value due to its repeated pairing with the primary reinforcer r — it becomes a conditioned reinforcer that predicts the unconditioned, primary reinforcer and hence can be used to reward another operant unit. This is a key step in establishing an operant chain (for a concise review of the underlying psychological principles, see, e.g., (Catania, 1968; Mackintosh, 1983; Reynolds, 1968)). Operant chaining can thus be envisioned as connecting many operant units (segments) into a sequence, as long as the S for the *subsequent* unit has acquired a reinforcing value and can function as an r for the *preceding* unit.

Note that the relationship between S and r in a segment of the operant chain is analogous to first- and higher-order classical conditioning between the two stimuli — their association strength increases as the animal gradually acquires the operant (i.e., executes the correct response). Therefore, the psychological foundation for establishing an arbitrary action sequence is: (a) the *simultaneous* operation of two learning principles (the operant principle for behavioral modification and the Pavlovian principle

E-mail address: junz@umich.edu.

0893-6080/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2009.03.017

for stimulus association); and (b) the *dual* nature of a stimulus, serving a discriminative function (for initiating the action of subsequent segment) and a reinforcing function (for rewarding the action of the preceding segment).

Such psychological insight for solving the temporal credit-assignment problem has now been crystallized as the celebrated reinforcement learning algorithm (e.g., Barto, Sutton, & Watkins, 1990; Montague, Dayan, & Sejnowski, 1996; Samuel, 1959; Sutton, 1988; Watkins, 1989; and numerous more recent work). The critical element is to require the agent to learn reward-predicting values of contexts (discriminative stimuli) associated with intermediate stages, and/or values of all possible actions taken from these stages. In this paper, we will adapt the Bayesian re-formulation of operant reinforcement learning (as developed in the companion paper) to the sequential decision context and show how it naturally leads to the notion of incentive value, namely, the animal's prediction or *expectancy* of average reward which can in turn be effectively used as an internally generated reinforcement signal for modifying action probabilities of non-terminal operant segments. So our contribution here is to connect selectionism and Bayesianism in sequential learning context.

The remaining of the paper is organized as follows. Section 2.1 recalls the linear operator model describing the selectionist style of learning for a single operant segment, and the associated ensemble-level dynamics with an equivalent Bayesian formulation. Section 2.2 analyzes the linear operator model for an operant chain with two segments/stages (without loss of generality), and derives the appropriate reward signal for the non-terminal operant segment. Section 2.3 performs the ensemble-level analysis of the operant chain from two perspectives, one treating it as step-by-step sequential learning, and the other treating it as learning of a unitary, composition action. Section 2.4 establishes the equivalence between these two sets of ensemble-level equations, and obtains their solutions, along with the Bayesian interpretation. Section 2.5 further investigates the notion of incentive value and the effective reinforcement signal of non-terminal segment(s), and explains how incentive value of a state is affected by animal's change in policy (during learning). Section 2.6 discusses the concurrent learning of action probabilities and incentive values in the context of action-critic architecture. Section 2.7 illustrates these concepts in a computer simulation of a simple spatial navigation task. The paper closes with a discussion (Section 3) of the tight coupling of action learning and incentive learning as provided by traditional theories of animal learning as well as modern reinforcement learning
framework.

2. Mathematical analysis

2.1. Background

We briefly review our analysis of operant learning in a linear operator model (Bush & Mosteller, 1955) given by Zhang (2009), as a way of introducing the relevant notations for our subsequent exposition.

Let p_i $(i = 1, 2, \ldots, N)$ denote the probability of the animal emitting any action R_i from an action repertoire (action set) $\mathcal{R} = \{R_1, R_2, \ldots, R_N\}$, and θ_i the reward associated with such an action. The change of action probabilities δp_k, $\forall k$ after the animal executes a particular action R_i is

$$\delta p_k = \epsilon\, \theta_i\, (e_{ik} - p_k), \quad (R_i \text{ having been executed}), \tag{1}$$

with a small >0 as the learning rate parameter, and e_{ik} denotes the Kronecker delta

$$e_{ik} = \begin{cases} 1 & \text{if } k = i \\ 0 & \text{else .} \end{cases} \tag{2}$$

This rule of single-trial operant learning results in an increase in p_i, the probability of selecting the same action R_i due to the recent reward θ_i, and a decrease in p_j $(j \neq i)$, the probabilities of selecting all other actions in the future due to probability normalization.

The above equation describes how action probabilities are modified after a single action is executed (and then the animal is rewarded). During operant conditioning, however, it is possible that any action be selected (and executed). Since the probability of action R_i being selected is p_i, the *average* change of action probabilities Δp_k should be weighted by this factor:

$$\Delta p_k = \sum_i p_i\, \delta p_k.$$

This, along with (1), gives rise to the master equation for operant learning

$$\Delta p_k = \epsilon\, p_k\, (\theta_k - \Theta) \quad k = 1, 2, \ldots, N.$$

Here

$$\Theta = \sum_i p_i\, \theta_i$$

is the average reward the animal receives at any time during learning. The essence of operant conditioning can be viewed as an effective increase or decrease of action probabilities depending on whether the reward for such action is above or below the current value of the average reward Θ (i.e., averaged across the entire action repertoire). It is easily seen that Θ increases monotonously (for sufficiently small ϵ), until it reaches the value of $\theta_{k^*} = \max\{\theta_k, k = 1, 2, \ldots, N\}$ when the animal acquires the operant by performing R_{k^*}, the maximally rewarded action, with probability $p_{k^*} \to 1$. The discrete probability updating Δp_k can be replaced by a continuous version (absorbing ϵ into t)

$$\frac{\mathrm{d}p_k}{\mathrm{d}t} = p_k \left(\theta_k - \sum_i p_i\, \theta_i \right), \tag{3}$$

with analytic solution

$$p_k(t) = \frac{p_k(0)\, \mathrm{e}^{\theta_k t}}{\sum_i p_i(0)\, \mathrm{e}^{\theta_i t}}. \tag{4}$$

Define

$$Z(t) = \sum_i p_i(0)\, \mathrm{e}^{\theta_i t},$$

then

$$\Theta(t) = \sum_i p_i(t)\, \theta_i = \frac{\mathrm{d} \log Z(t)}{\mathrm{d}t}.$$

2.2. Single-trial sequential learning equation

Without loss of generality, assume now that there are two components in an operant sequence A–B, where the first action ("A") could be drawn from the action repertoire $\mathcal{A} = \{A_1, A_2, \ldots, A_N\}$, and the second action ("B") drawn out of *one* of the N repertoires $\mathcal{B}_i = \{B_{1|i}, B_{2|i}, \ldots, B_{M|i}\}$ associated with the state S_i that is reached as a result of chosen action A_i in the first stage. In other words, each action A_i in the first stage (out of N alternatives) leads to a distinct action repertoire $\mathcal{B}_i$ from which the action for the second stage is to be chosen (out of M alternatives). The reward θ_{ij} for this sequence is delivered only after the execution of the second action $B_{j|i}$ (Fig. 1).

Traditionally, in the theory of learning automaton (Narendra & Thathachar, 1989; Thathachar & Ramakrishnan, 1981), this type of sequential problem is treated by a hierarchical system of stochastic

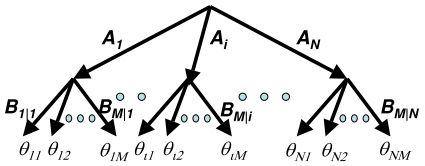

Fig. 1. Schematic illustration of the two-stage sequential choice problem with terminal rewards θ. Actions of the first stage are denoted $A_i, i = 1, \ldots, N$, while actions of the second stage are denoted $B_{j|1}, \ldots, B_{j|N}, j = 1, \ldots, M$ (whose subscripts also indicate the first-stage action they are associated with). The corresponding action probabilities are $p_i(t)$ for the first stage and $p_{j|i}(t)$ for the second stage.

automata (a generalized version of linear operator model), each operating on its own set of action probability and improving according to (a generalized version of) the linear operator model. The ensemble of automata is organized into a tree-like structure, and modification of action probability proceeds at different levels based on the reward signal feeding into all levels of the hierarchy.

Let us first treat the learning of this two-stage operant sequence as a single or *unitary* action, A_i-and-$B_{j|i}$, in the composite repertoire $\mathcal{A} \times \mathcal{B}$. Denote the joint probability $p_{ij} = \text{Prob}\,(A_i \text{ and } B_{j|i})$ of the animal's emitting such a composite action A_i-and-$B_{j|i}$, with $\sum_{i,j} p_{ij} = 1$. At discrete time-step n, applying the single-trial operant rule (1), we have the change of action probability p_{kl} in the composite repertoire as

$$\delta p_{kl}^{(n)} = \epsilon\, \theta_{ij}\, (e_{ik}\, e_{jl} - p_{kl}^{(n)}) \quad k = 1, 2, \ldots, N; l = 1, 2, \ldots, M, \quad (5)$$

where θ_{ij} is the reward for the selected composite action A_i-and-$B_{j|i}$.

On the other hand, we can also write out the change of action probability at each individual level of the hierarchy. Let $p_k = \text{Prob}\,(A_k)$ denote the action probability associated with $\mathcal{A}$, with $\sum_k p_k = 1$. Let $p_{l|k} = \text{Prob}\,(B_{l|k})$ denote the conditional action probability associated with $\mathcal{B}_k$, with $\sum_l p_{l|k} = 1$. According to the single-trial operant rule (1), the change of those action probabilities after the animal emits action A_i and then $B_{j|i}$ at discrete time n is:

$$\delta p_k^{(n)} = a_i\, (e_{ik} - p_k^{(n)}), \quad (6a)$$

$$\delta p_{l|k}^{(n)} = b_{ij}\, e_{ik} (e_{jl} - p_{l|k}^{(n)}), \quad (6b)$$

where the reward values a_i and b_{ij} associated with $\mathcal{A}$ and $\mathcal{B}_k$ are to yet be determined. Since $p_{kl} = p_k \cdot p_{l|k}$,

$$\delta p_{kl}^{(n)} = p_{kl}^{(n+1)} - p_{kl}^{(n)} = p_k^{(n+1)}\, \delta p_{l|k}^{(n)} + p_{l|k}^{(n)}\, \delta p_k^{(n)}. \quad (7)$$

A comparison of (7) with (5) yields

$$a_i = \epsilon\, \theta_{ij}, \quad (8a)$$

$$b_{ij} = \epsilon\, \theta_{ij} / p_i^{(n+1)}. \quad (8b)$$

This is to say, the reward for A_i is simply θ_{ij} delivered stochastically for any fixed i (since it depends on the action $B_{j|i}$ selected at the second stage), while the reward for $B_{j|i}$ is $\theta_{ij}/p_i^{(n+1)}$; here the denominator $p_i^{(n+1)}$ is a factor to equalize the opportunity of learning (i.e., rescaling the learning rate ϵ) since all branches are not reached on every trial. The essence of this learning scheme is that the *same* terminal reward has been used to reinforce action tendencies along all levels of the hierarchy (i.e., terminal and non-terminal stages), and that the updating of action probabilities in the hierarchy proceeds from top-down (i.e., from the beginning component to the terminal one). This kind of reward scheme, which was developed by Thathachar and Ramakrishnan (1981), naturally leads to an absolutely expedient learning algorithm. Alternative algorithms, which adopts non-standard single-trial rules, have been introduced that would considerably speed up learning (Thathachar & Sastry, 1985) or converge to solutions that are optimal under more general definitions of optimality in the hierarchy of learning automata (Thathachar & Sastry, 1987). However, this reward structure (8b) requires action probability to proceed strictly from the beginning or non-terminal component to terminal component; it is not obvious how such off-line, batch-mode learning can be applied to on-line, stage-by-stage learning when the agent transits in a stochastic environment.

2.3. Ensemble-level equations from the two views

The reward rule (8b) for the automata hierarchy was developed based on the formal "equivalence" of two sets of single-trial equations, (5) and (6b), that respectively represent the unitary, composite action approach and the component-based action sequence approach. Now, we investigate the formal equivalence of these two approaches at the level of ensemble-level equations. When the two-stage action sequence A_i-and-$B_{j|i}$ is considered as a single, composite action, the ensemble-level equation associated with single-trial rule (5) is

$$\frac{\mathrm{d}p_{ij}(t)}{\mathrm{d}t} = p_{ij}(t)\, (\theta_{ij} - \Theta(t)), \quad (9)$$

with total average reward

$$\Theta(t) = \sum_{i,j} p_{ij}(t)\, \theta_{ij}. \quad (10)$$

Next, instead of taking the two-component action sequence as a unitary (composite) action being performed and reinforced, we treat the action sequence as being performed sequentially, A_i-and-$B_{j|i}$. The second (terminal) B-stage obviously carries the reward structure (for a fixed i) $\{\theta_{i1}, \theta_{i2}, \ldots, \theta_{iM}\}$. So the ensemble-level equation is

$$\frac{\mathrm{d}p_{j|i}(t)}{\mathrm{d}t} = p_{j|i}(t) \left(\theta_{ij} - \sum_j p_{j|i}(t)\, \theta_{ij}\right).$$

Define

$$\theta_i(t) \equiv \sum_j p_{j|i}(t)\, \theta_{ij} \quad (11)$$

as the average reward for a fixed i, we have

$$\frac{\mathrm{d}p_{j|i}(t)}{\mathrm{d}t} = p_{j|i}(t)\, (\theta_{ij} - \theta_i(t)). \quad (12)$$

As for the first (non-terminal) A-stage, observe from (12) that each action A_i is associated with a unique value θ_i that represents the expected value of primary reward once the action A_i has been executed, we can naturally use the set of quantities $\theta_i, i = 1, \ldots, N$ as the effective reinforcement, at time t, for the action ensemble $\mathcal{A}$. Therefore, the ensemble-level equation for modifying $p_i(t)$ is

$$\frac{\mathrm{d}p_i(t)}{\mathrm{d}t} = p_i(t) \left(\theta_i(t) - \sum_i p_i(t)\, \theta_i(t)\right).$$

Note here, $\theta_i(t)$'s are functions of time, unlike the stationary θ_{ij}'s. Their average is

$$\sum_i p_i(t)\, \theta_i(t) = \sum_i \sum_j p_{j|i}(t)\, p_i(t)\, \theta_{ij} = \sum_{i,j} p_{ij}(t)\, \theta_{ij} = \Theta(t).$$

Thus

$$\frac{\mathrm{d}p_i(t)}{\mathrm{d}t} = p_i(t)\, (\theta_i(t) - \Theta(t)). \quad (13)$$

This is the ensemble-level operant learning equation governing $p_i(t)$, the action probability of the first stage of the A–B operant sequence.

Eqs. (13) and (12) are in the proper form of ensemble-level equations, describing the evolution of $p_i(t)$, action probability associated with the first (non-terminal) stage, and $p_{j|i}(t)$, action probability associated with the second (terminal) stage. Note that

θ_i is interpreted *both* as the average reward associated with all possible actions in the second B-stage (the specific reward for action $B_{j|i}$ is θ_{ij}) *and* as the specific reward associated with the action A_i at the first stage (the average reward for A-stage is Θ). In comparison, the single-trial learning rule (6b) with (8b) used θ_{ij} as the specific reward associated with the A-stage.

2.4. Equivalence and analytic solution

We next show that the two views of the *A–B* operant sequence, i.e., treating it as a unitary, composite action and treating it as two sequentially performed actions A_i-and-$B_{j|i}$, are in fact equivalent at the ensemble-level (their equivalence at the single-trial level was established in Section 2.2 and was used to derive the reward structure for A-stage).

Proposition 1. *The ensemble-level equations (12) and (13) for learning the action sequence A–B by treating θ_{ij} as the reward for the second, terminal stage and $\theta_i(t)$ for the first, non-terminal stage is equivalent to the the ensemble-level (5) treating A–B as a unitary, composite action with θ_{ij} as the reward.*

Proof. We take the time derivative of $p_{j|i}(t)\, p_i(t)$ by invoking the chain rule, and substitute the relations (12) and (13):

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(p_{j|i}(t)\, p_i(t)\right) &= \frac{\mathrm{d}p_{j|i}(t)}{\mathrm{d}t}\, p_i(t) + p_{j|i}(t)\, \frac{\mathrm{d}p_i(t)}{\mathrm{d}t} \\ &= p_i(t)\, p_{j|i}(t)\, (\theta_{ij} - \theta_i(t)) + p_{j|i}(t)\, p_i(t)\, (\theta_i(t) - \Theta(t)) \\ &= p_{ij}(t)\, (\theta_{ij} - \Theta(t)) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\, p_{ij}(t),\end{aligned}$$

where the last step invokes (5), the ensemble-level equation of treating *A–B* as a unitary (composite) action. Hence the relation $p_{ij}(t) = p_{j|i}(t)\, p_i(t)$ holds throughout learning (assuming that it holds at $t = 0$). ◇

The algorithm can be readily shown to be absolutely expedient (Narendra & Thathachar, 1989) and hence to converge in probability as the learning rate approaches zero, see also discussions in Zhang (2009). We first compute the change of the effective reinforcement θ_i (which is the average of the primary reward for the B-stage) during conditioning:

$$\frac{\mathrm{d}\theta_i(t)}{\mathrm{d}t} = \sum_j \frac{\mathrm{d}p_{j|i}(t)}{\mathrm{d}t}\, \theta_{ij} = \sum_j p_{j|i}(t)\, (\theta_{ij} - \theta_i(t))^2 > 0.$$

That $\theta_i(t)$ is monotone increasing but bounded from above by $\max_j \theta_{ij}$ (as easily seen from (11)) guarantees that, as $t \to \infty$, the conditional probabilities

$$p_{j|i}(t) \to e_{jj^*},$$
$$\theta_i(t) \to \theta_{ij^*} \equiv \theta_i^*,$$

in which e denotes the Kronecker delta symbol (2) and

$$j^* = j^*(i) = \operatorname{argmax}_j\{\theta_{ij}, j = 1, 2, \ldots, M\}.$$

As for the A-stage, the change of average reward Θ is

$$\begin{aligned}\frac{\mathrm{d}\Theta(t)}{\mathrm{d}t} &= \sum_i \left(\frac{\mathrm{d}p_i(t)}{\mathrm{d}t}\, \theta_i(t) + p_i(t)\, \frac{\mathrm{d}\theta_i(t)}{\mathrm{d}t}\right) \\ &= \sum_i p_i(t) \left((\theta_i(t) - \Theta(t))^2 + \sum_j p_{j|i}(t)(\theta_{ij} - \theta_i(t))^2\right) > 0.\end{aligned}$$

In the limit of $t \to \infty$,

$$p_i(t) \to e_{ii^*},$$
$$\Theta \to \theta_{i^*}^*,$$

with

$$i^* = \operatorname{argmax}_i\{\theta_i^*, i = 1, 2, \ldots, N\}.$$

Hence, learning in this *A–B* operant sequence will converge to the optimal action sequence with maximal reward

$$p_{ij}(t) \to e_{ii^*}\, e_{jj^*},$$
$$\Theta \to \theta_{i^*j^*} = \theta_{i^*}^*$$

with

$$i^*j^* = \operatorname{argmax}_{ij}\{\theta_{ij}, i, = 1, 2, \ldots, N, j = 1, 2, \ldots, M\}.$$

Following the derivations of solution (4) to the ensemble-level Eq. (3), the ensemble-level equations for the *A–B* operant sequential can be solved analytically. In fact, we have

Corollary 2. *Given the initial conditions, $p_{ij}(0) = p_i(0)p_{j|i}(0)$ with $p_i(0) = \sum_j p_{ij}(0)$, the ensemble-level dynamics are given by*

$$p_{ij}(t) = \frac{p_{ij}(0)\, \mathrm{e}^{\theta_{ij}\, t}}{Z}$$

as the solution to (9), and by

$$p_{j|i}(t) = \frac{p_{j|i}(0)\, \mathrm{e}^{\theta_{ij}\, t}}{Z_i}$$

$$p_i(t) = \frac{p_i(0)\, Z_i}{Z} = \frac{\sum_j p_{ij}(0)\, \mathrm{e}^{\theta_{ij}\, t}}{\sum_{ij} p_{ij}(0)\, \mathrm{e}^{\theta_{ij}\, t}}$$

as the solutions to (12) and (13). Here

$$Z_i(t) = \sum_j p_{j|i}(0)\, \mathrm{e}^{\theta_{ij}\, t}.$$

$$Z(t) = \sum_i p_i(0) Z_i(t) = \sum_{ij} p_{ij}(0)\, \mathrm{e}^{\theta_{ij}\, t}.$$

Proof. By direct verification. ◇

As can be easily verified, these solutions satisfy $p_{ij}(t) = p_i(t)\, p_{j|i}(t)$ and $p_i(t) = \sum_j p_{ij}(t)$. Clearly, $p_{ij}(t), p_i(t), p_{j|i}(t)$ in these forms have Bayesian interpretations. In treating the *A–B* operant sequence as a unitary, composite action, the corresponding Bayesian model consists of a set of $N \times M$ hypotheses, with $\theta_{ij}\, t$ as the log likelihood functions. In treating the *A–B* operant sequence as actions sequentially executed in two stages, the corresponding hierarchical Bayesian model consists of N hypotheses at the top-level of the hierarchy with $\log Z_i(t)$ as the log likelihood function (for hypothesis A_i) and, for each top-level hypothesis, M hypotheses at the bottom-level of the hierarchy with $\theta_{ij}\, t$ as the log likelihood function (for hypothesis $B_{j|i}$). Note that

$$\theta_i(t) = \sum_j p_{j|i}(t)\theta_{ij} = \frac{\sum_j p_{j|i}(0)\mathrm{e}^{\theta_{ij}\, t}}{Z_i(t)} = \frac{\mathrm{d}\log Z_i(t)}{\mathrm{d}t}.$$

Hence

$$\log Z_i(t) = \int_0^t \theta_i(\tau)\mathrm{d}\tau,$$

demonstrating that $\theta_i(t)$ indeed plays the role of effective reinforcement for A-stage at the ensemble-level dynamics.

2.5. Effective reinforcement signal and incentive value

The conceptual development in the sequential treatment of the acquisition of the *A–B* operant sequence is the introduction of an effective reinforcement $\theta_i(t), i = 1, \ldots, N$ to modify $A_1, \ldots, A_N$ in action repertoire $\mathcal{A}$. This quantity $\theta_i(t)$ is associated with S_i, the state-of-the-world arrived by the animal's performing the first action component A_i. The value $\theta_i(t)$ is, on the one hand, the

average primary reinforcement value the animal receives upon performing, with probability $p_{j|i}(t)$, various possible actions $B_{j|i}$ after reaching the state S_i and, on the other hand, used as the reward to reinforce action probabilities $p_i(t)$ associated with A_i. It is this dual nature of $\theta_i(t)$ that makes the sequential learning possible. For this reason, this effective reinforcement $\theta_i(t)$ is also called the secondary (conditioned) reinforcement or *incentive value* associated with an intermediate (non-terminal) context S_i.

Suppose the reward structure is such that $\theta_{ij} = \alpha_i$ for all j (i.e., the terminal reward is independent of the action taken in the second stage). In this situation,

$$\theta_i(t) = \sum_j p_{j|i}(t)\,\alpha_i = \alpha_i,$$

so that the probability p_i is modified according to the reward structure $\{\alpha_i, i = 1, 2, \ldots, N\}$, yet none of the $p_{j|i}$'s is modified since the right-hand side of (12) is zero. Suppose, on the other hand, that the second stage is not the terminal one, but rather the successor state in an environment of a Markov Decision Process (MDP) with finite states; the reward θ_{ij} includes a primary reward α_j, and an effective future reward θ_j discounted with a factor γ ($0 \leq \gamma \leq 1$). Eq. (11) is thus extended to (we omit the t argument)

$$\theta_i = \sum_j p_{j|i}\,(\alpha_j + \gamma\theta_j). \tag{14}$$

This is essentially the Bellman equation for the infinite-horizon decision problems, where θ_i is the value function of the state S_i. Bellman (1957) formulated the stochastic dynamic programming approach to the solution of the Markov decision problems in stochastic environment with stationary, action-dependent transitional probability between states (see Puterman, 1994 for reference on dynamic programming). The goal for the MDP is to seek an optimal policy (for action at each decision epoch) that maximizes the total (discounted or undiscounted) reward. Finite-horizon discrete MDPs can be solved using backward induction (dynamic programming) to recursively evaluate expected rewards at each step. The optimal action in the last decision epoch is determined first, and the average reward is propagated backward in order to calculate the expected reward of the next-to-last decision epoch, and so on.

On-line learning of incentive values θ_i is intimately associated with the process of classical conditioning. Consider the Rescorla–Wagner model (Rescorla & Wagner, 1972) describing the acquisition of association strength of a conditioned stimulus (CS) when it is repeatedly paired with an unconditioned stimulus (US) or reinforcer in classical conditioning experiments. As summarized in Mackintosh (1983), this model successfully accounts for a variety of experimental findings, including the time-course of acquisition and extinction of association strength, the effect of stimulus overshadowing (where the perceptually more salient stimulus acquires greater association strength compared with another less salient stimulus when both are presented as a compound CS), the effect of stimulus blocking (where prior conditioning to a stimulus suppresses the acquisition of association strength for a second stimulus when both are presented as a compound CS), and effects related to post-conditioning manipulation of the US value. In the sequential context, the Rescorla–Wagner rule had been extended as the temporal difference or TD model (Sutton, 1988; Sutton & Barto, 1990). The simplest version, called TD(0) algorithm or AHC, the method of adaptive heuristic critic (Barto, Sutton, & Anderson, 1983), computes the discrepancy in reward estimation over one-step in time to improve reward prediction:

$$\delta\tilde{\theta}_i = \eta\,(\alpha_j + \gamma\tilde{\theta}_j - \tilde{\theta}_i). \tag{15}$$

Here $\tilde{\theta}_i$ with tilde indicates the animal's estimation of θ_i associated with state S_i. The subscript j is the successor state of i, which is contingent upon the animal's behavior at i, and α_j is the primary reward. The term $\alpha_j + \gamma\tilde{\theta}_j$ is the estimated total reward upon entering the state j.

Since action selection is stochastic — the animal enters state j with probability $p_{j|i}$ — the average change of $\tilde{\theta}_i$ is

$$\Delta\tilde{\theta}_i = \eta\left(\sum_j p_{j|i}(\alpha_j + \gamma\tilde{\theta}_j) - \tilde{\theta}_i\right).$$

When $p_{j|i}$ is assumed to be stationary, $\tilde{\theta}_i$ will reach an equilibrium $\tilde{\theta}_i \to \theta_i$ as given by (14) when eventually $\Delta\tilde{\theta}_i = 0$. This is to say, the AHC learning rule (15) will lead to an accurate reward estimate, $\tilde{\theta}_i = \theta_i$, provided that action probability $p_{j|i}$ remains stationary.

However, when action probability is allowed to update simultaneously during learning, the animal's calculation of the incentive value associated with a state ought to also take into account of the status of action learning in that state — actual average reward will be higher if the state is relatively better mastered in terms of action selection (i.e., the animal knows what to do) compared with a state that is relatively unlearned (i.e., the animal does not know what to do). The modification of action probability itself will lead to a change in the average reward associated with that state; therefore the animal's expectation of reward, i.e., the incentive value, should have included a consideration of this experience of action learning and its effect on reward. Formally, suppose that in state S_i the animal performs an action indexed by j to transit to state S_j and receives reward β_j, with single-trial learning

$$\delta p_{k|i} = \beta_j\,(e_{jk} - p_{k|i}) \quad \forall k. \tag{16}$$

The change of the value of $\tilde{\theta}_i$ due to change of $p_{k|i}$ can be computed:

$$\delta\tilde{\theta}_i = \sum_k \delta p_{k|i}\,(\alpha_k + \gamma\,\tilde{\theta}_k) = \beta_j \sum_k (e_{jk} - p_{k|i})\,(\alpha_k + \gamma\,\tilde{\theta}_k)$$
$$= \beta_j\left(\alpha_j + \gamma\,\tilde{\theta}_j - \sum_k p_{k|i}\,(\alpha_k + \gamma\,\tilde{\theta}_k)\right)$$

or, using (14),

$$\delta\tilde{\theta}_i = \beta_j\left(\alpha_j + \gamma\,\tilde{\theta}_j - \tilde{\theta}_i\right). \tag{17}$$

This describes how the estimated incentive value $\tilde{\theta}_i$ would have been affected by the updating of action probabilities $p_{k|i}$. Comparing (17) with the AHC rule (15), it can be seen that the two are formally identical if and only if $\beta_j = \text{const}$ — the latter happens when the reward for each action is non-discriminatory, i.e., the animal is exploring the environment without improving its action selection.

2.6. Simultaneous learning of the actor and the critic

In the now popular "actor–critic" architecture (Barto et al., 1983), both the incentive value and action probability undergo modification at each learning step. This kind of learning architecture can be viewed as the integration of the classical conditioning principle (the learning of the "critic") with the operant conditioning principle (the learning of the "actor"); it provides a potential framework for unifying animal learning theories. Here, we explore the consequences of combining action learning with incentive value learning on a step-by-step basis in such an architecture. Define the inconsistency of estimate of incentive values

$$\rho_i \equiv \tilde{\theta}_i - \sum_k p_{k|i}\,(\alpha_k + \gamma\,\tilde{\theta}_k).$$

At a given learning step (when the animal is in state S_i), modification of action probability $\delta p_{k|i}$ is according to (16), and modification of incentive value $\delta\tilde{\theta}_i$ is according to (17). Note that $\delta\tilde{\theta}_j = 0$ for $j \neq i$, since only reward estimate at i is modified when

i is visited. After the simultaneous action and incentive learning (for index i), the change of ρ_i resulting from a change in $p_{k|i}$ and a change in $\tilde{\theta}_i$ is

$$\delta\rho_i = \delta\tilde{\theta}_i - \sum_k \delta p_{k|i}\,(\alpha_k + \gamma\,\tilde{\theta}_k).$$

Substituting (16) and (17), it is easy to show that

$$\delta\rho_i = -\beta_j\,\rho_i,$$

or

$$\delta(\rho_i)^2 = -2\beta_j\,(\rho_i)^2 < 0. \tag{18}$$

Eq. (18) says that, when the actor and the critic learn simultaneously according to (16) and (17) respectively, the discrepancy in reward estimate will decrease.

There can be two ways (18) may be satisfied:

(1) When $\beta_j = \eta$ independent of j, incentive learning follows the AHC rule (15), while action learning uses a non-discriminatory reward (independent of action chosen):

$$\begin{cases}\delta p_{k|i} = \eta\,(e_{jk} - p_{k|i}),\\ \delta\tilde{\theta}_i = \eta\,(\alpha_j + \gamma\,\tilde{\theta}_j - \tilde{\theta}_i);\end{cases}$$

(2) When $\beta_j = \epsilon\,(\alpha_j + \gamma\,\tilde{\theta}_j)$ as a function of $\tilde{\theta}_j$, action learning is guided by the learned incentive values, while incentive learning adopts a rule slightly different from the vanilla AHC rule:

$$\begin{cases}\delta p_{k|i} = \epsilon\,(\alpha_j + \gamma\,\tilde{\theta}_j)\,(e_{jk} - p_{k|i}),\\ \delta\tilde{\theta}_i = \epsilon\,(\alpha_j + \gamma\,\tilde{\theta}_j)\,(\alpha_j + \gamma\,\tilde{\theta}_j - \tilde{\theta}_i).\end{cases}$$

In the first case, the action probability vector for each state will fluctuate and diffuse away from whatever the starting value, though the average change or drift is zero — the agent is exploring the environment without being committed to a direction of change. In the second case, action probabilities will improve according to the estimated total reward (primary reward plus the incentive value) — the agent is making use of the learned (or partially learned) incentive values to adjust its future action tendency. As for the concurrent incentive learning, our proposed modification here is to take into account of the effect of action modification on reward estimation. When the learning of the actor and of the learning of the critic are coordinated in such fashion, the inconsistency in reward estimates resulting from the uncertainty of the environment (that the animal does not have control of) and that resulting from the randomness of action selection (that the animal has control of) will be separated. Of course, though $(\rho_i)^2$ will decrease whenever state S_i is visited, it may *increase* when the animal visits any other state that is one-step reachable from S_i. A more rigorous analysis is needed to prove the online convergence of such simultaneous learning scheme.

2.7. An example: Spatial navigation task

We simulate a typical sequential decision task, the spatial navigation task, where the animal seeks a primary (terminal) reward in its environment in the presence of barriers. This route-finding task is adapted from Barto et al. (1990) and Dayan (1992) where the methods of TD learning were used to solve for the shortest path to terminal reward. A 12×8 grid represents a region of space, with each intersection of the grid lines representing a spatial location. In addition to a goal location, the region contains a C-shaped barrier that the navigating agent (animal) cannot cross over. The animal is allowed to move from a location to one of its neighbors and, if the goal is reached, rewarded (whereby a trial is terminated). As in Dayan (1992), we restrict the animal to have the same starting location (7, 4) each trial (the lower left corner is chosen as the origin of the Cartesian coordinate system). The task is to find the shortest route to the goal location $\mathtt{G} = (3, 6)$.

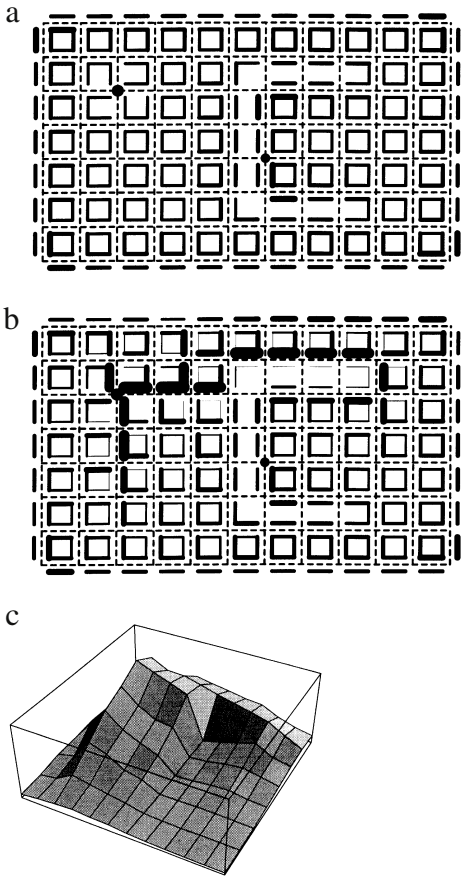

Fig. 2. Pattern of action probabilities (in (a) and (b)) and conditioned reinforcement value or reward prediction (in c) in a spatial navigation task. The navigating agent is positioned at an intersection of an 8×12 grid (indicated by the dotted lines), and may move to a neighboring intersection except as prohibited by the C-shaped barrier. When standing at any grid location and facing towards one of its neighbors, the probability of taking that action is indicated by the thickness of a filled line drawn to the right of the dotted grid line. The C-shaped barrier is, therefore, represented by the absence of such filled lines connecting respective locations. The starting location and the goal location are indicated by a filled circle. (a) Before learning starts. (b) After 500 trials of learning. (c) The conditioned reinforcement values are represented by the height of the landscape after 500 trials.

In Fig. 2, we represent action probability associated with each location of the grid (drawn as dotted lines) by the *thickness* of a solid line toward the *right* of the grid line when the animal is at a given intersection. Standing at any intersection and facing the next location that is one-step reachable, the solid line on the right-hand side graphically represents the probability that the animal will move to such location ("right-of-way"). Fig. 2 *a* gives the initial action probabilities, which are set to be all equal (unless blocked by the barrier in the environment). The C-shaped barrier is reflected as the absence (zero action probability) of solid lines connecting the locations at the two sides of the barrier.

During training, the animal applies the rules of on-line modification of action probability and reward prediction. After being placed in the initial starting position, the sequence of events for a trial is as follows: (1) When positioned at a location $\mathtt{X}$, the animal randomly selects an action according to $p(\mathtt{Y}|\mathtt{X})$; (2) The

animal moves to the new position Y, as a consequence of executing the chosen action; (3) The reinforcement received, which can be either θ_{Y} (the conditioned reinforcement value if $\mathrm{Y} \neq \mathrm{G}$) or $r_{\mathrm{G}} = 1.0$ (the primary reward if $\mathrm{Y} = \mathrm{G}$), serves to modify both the probability of action selection $\{p(\mathrm{Y}'|\mathrm{X}) : \mathrm{Y}'$ one-step reachable from $\mathrm{X}\}$ and the value of conditioned reinforcement $\tilde{\theta}_{\mathrm{X}}$ at the former position X:

$$\delta p(\mathrm{Y}'|\mathrm{X}) = \epsilon\, \tilde{\theta}_{\mathrm{Y}}\, (e_{\mathrm{YY}'} - p(\mathrm{Y}'|\mathrm{X})),$$
$$\delta \tilde{\theta}_{\mathrm{X}} = \epsilon\, \tilde{\theta}_{\mathrm{Y}}\, (\gamma \tilde{\theta}_{\mathrm{Y}} - \tilde{\theta}_{\mathrm{X}}).$$

Note that the incentive learning rule used is a modification of the AHC rule (the proportionality term is, instead of a constant η, the variable $\epsilon\, \tilde{\theta}_{\mathrm{Y}}$). These procedures are repeated for the new position Y until the animal reaches the goal position G. The learning rate parameter is $\epsilon = 0.1$, with $\gamma = 0.9$ for this infinite-horizon problem. After 500 trials of learning, the optimal path is demonstrated by the spatial pattern of action probabilities (Fig. 2b). The spatial map of the incentive values after learning is given in (Fig. 2c). This simulation is based on a lookup table representation for incentive values and for action probabilities; state aggregation methods (e.g. Singh, Jaakkola, & Jordan, 1995) will be needed to efficiently deal with increasing number of states.

3. Discussion

Operant learning is characterized by the interaction between the animal and its environment. On the one hand, the behavior of an organism has operated on the environment and modified the state-of-the-world; on the other hand, the consequence of such operation (in the form of reward or punishment) will modify the organism's behavior and determine the likelihood of its being repeated on subsequent occasions. This operant (or instrumental) conditioning process, which follows Thorndike's "Law of Effect", should be distinguished from the process of classical conditioning, in which the presentation of one stimulus (CS) together or in close proximity with another reinforcing stimulus (US) results in the former's acquisition of the meaning of the latter, according to Pavlov's "Principle of Substitution". It was Skinner (1937) who first drew a clear distinction, on operational grounds, between these two forms of associative learning, i.e., according to the rules for the experimenter's delivery of reinforcers — for classical conditioning, it is the contingency between CS and a reinforcer (US) regardless of the subject's behavior, while for instrumental conditioning, it is the relationship between the subject's own behavior and the occurrence of the reinforcer. Of course, Skinner's operational definition does not imply separable learning processes involved, even though response classes are thought to be somewhat differentiated across these two types of conditioning (e.g., visceral/glandular versus skeletal/somatic, without versus with sensory feedback, involuntary versus voluntary). A variety of so-called "two-factor" theories propose that operant conditioning involves both an association between the discriminative stimulus and instrumental response and an association between discriminative stimulus and reinforcer (Rescorla & Solomon, 1967). There is still considerable debate as in what sense the discriminative stimulus finally became a classical CS in an instrumental experiment (see Mackintosh, 1983).

Despite the details of underlying psychological mechanisms, it is now well established that acquisition of an operant chain (based on a terminal reward) is possible as long as the discriminative stimuli for intermediate stages acquire conditioned reinforcement values. Our analysis recapitulated such a theme: the dual role played by such stimulus — reinforcer for the previous component and instigator for the upcoming component — enables successive operant components to be "glued" to form a chain. Since the acquired (conditioned) reinforcement values depend on their correlation with the primary reinforcement (terminal reward), their magnitudes are necessarily graded and decrease the farther away from the terminal component. It is then the second-(or higher-)order Pavlovian conditioning mechanisms that help establish a value for the discriminative stimulus, as long as the primary reward to the conditioned reinforcer is always maintained. In this connection, we can view operant learning as animals' natural intelligence for solving the "credit-assignment problem" by simple principles such as incrementally adjusting action probabilities based on the effect of such actions (the Law of Effect) coupled with incremental establishment of reward-predicting values of environmental cues (the Principle of Stimulus Substitution). The interplay between stimulus-reinforcer and response-reinforcer relationships during operant conditioning has long been recognized (see Jenkins, 1977). Our view is that, along with the acquisition of response operant (through reinforcement), the classical (Pavlovian) conditioning between the stimulus and the reinforcer develops in parallel in strength over time.

The simultaneous modification of action probability and of conditioned reinforcement value (incentive value) lies at the heart of reinforcement learning. In the "actor–critic" architecture (Barto, 1995; Barto et al., 1990), reward prediction and action selection are conducted separately. The temporal difference methods allow the agent to improve the accuracy of prediction (with learning driven by the reduction of prediction error), whether that being predicted is the value of the states (V function) or action-values (Q function). In the case of the Q-learning (Watkins, 1989), acquisition of action-value does not depend on what policy the agent follows during learning (or even whether the action policy changes during learning). However, for learning the incentive values of states, the action policy matters — since the incentive values of states are defined with respect to a given stationary policy. Here, we explicitly consider the effect of action learning on reward prediction (of states). Our proposed rule of learning reward prediction (which incorporated Rescorla–Wagner rule as a special case) explicitly takes into account the effect of action probability modification on the change of conditioned reinforcement value.

The tight coupling between action modification and incentive acquisition in operant reinforcement learning (which arises naturally from our formulation) need to be further investigated in terms of neurophysiological mechanisms in structures like basal ganglia that is known to play an important role both in action planning and sequencing (e.g., Aldridge, Berridge, Herman, & Zimmer, 1993; Graybiel, Aosaki, Flaherty, & Kimura, 1994) and in mediating reinforcement and reward prediction (e.g. Robbins & Everitt, 1992; Schultz, 1992). The dopaminergic circuits in basal'ganglia has been thought to mediate neural computation of prediction error (Montague et al., 1996), and the motivational attribution called incentive salience (Berridge, 2007; Berridge & Robinson, 1998) that modulates learned reward values (Berridge, Zhang, & Aldridge, 2008; Tindell, Berridge, Zhang, Pecina, & Aldridge, 2005) . Future physiological investigation would provide further details of the interaction in basal ganglia between the motor generation/selection aspect and reward/incentive aspect of the operant reinforcement learning.

Acknowledgements

The author thanks Min Chang for the discussion and performing the simulation reported in Section 2.7. This work was first presented in abstract form in Zhang and Chang (1996) at the 29th Annual Meeting of the Society for Mathematical Psychology, August 2-4, 1996, University of North Carolina at Chapel Hill.

References

Aldridge, J. W., Berridge, K. C., Herman, M., & Zimmer, L. (1993). Neuronal coding of serial order: Syntax of grooming in the neostriatum. *Psychological Science*, *4*, 391–395.

Barto, A. G. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 215–232). Cambridge: MIT Press.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuron-like elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-13*, 835–846.

Barto, A. G., Sutton, R. S., & Watkins, C. J. C. H. (1990). Learning and sequential decision making. In M. Gabriel, & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 539–602). Cambridge: MIT Press.

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

Berridge, K. C. (2007). The debate over dopamine's role in reward: The case for incentive salience. *Psychopharmacology (Berl)*, *191*, 391–431.

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews*, *28*, 309–369.

Berridge, K. C., Zhang, J., & Aldridge, J. W. (2008). Computing motivation: Incentive salience boosts of drug or appetite states. *Behaviourial and Brain Sciences*, *31*, 440–441.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning*. New York: John Wiley and Sons.

Catania, A. C. (1968). *Contemporary research in operant behavior*. Glenview, Illinois: Scott, Foresman and Company.

Dayan, P. (1992). The convergence of TD(λ) for general λ. *Machine Learning*, *8*, 341–362.

Graybiel, A. M., Aosaki, T., Flaherty, A. W., & Kimura, M. (1994). The basal ganglia and adaptive motor control. *Science*, *265*, 1826–1831.

Jenkins, H. (1977). Sensitivity of different response systems to stimulus-reinforcer and response-reinforcer relations. In H. Davis, & H. M. B. Hurwitz (Eds.), *Operant-Pavlovian interactions* (pp. 47–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mackintosh, N. J. (1983). *Conditioning and associative learning*. Oxford: Clarendon.

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.

Narendra, K. S., & Thathachar, M. A. L. (1989). *Learning automata: An introduction*. NJ: Prentice-Hall.

Puterman, M. L. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. New York: John Wiley & Sons.

Rescorla, R. A., & Solomon, R. L. (1967). Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning. *Psychological Review*, *74*, 151–182.

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

Reynolds, G. S. (1968). *A primer of operant conditioning*. Glenview, Illinois: Scott, Foresman and Company.

Robbins, T. W., & Everitt, B. (1992). Functions of dopamine in the dorsal and ventral striatum. *Seminars in the Neurosciences*, *4*, 119–127.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. In *IBM Journal of Research and Development* (pp. 210–229). Reprinted in E.A. Feigenbaum and J. Feldman (Eds.), *Computers and thought*. (McGraw-Hill, 1963.

Schultz, W. (1992). Activity of dopamine neurons in the behaving primate. *Seminars in the Neurosciences*, *4*, 129–138.

Singh, S. P., Jaakkola, T., & Jordan, M. I. (1995). Reinforcement learning with soft state aggregation. In G. Tesauro, D. S. Touretzky, & T. K. Leen (Eds.), *Advances in neural information processing systems*: *Vol. 7* (pp. 361–368). Cambridge, MA: MIT Press.

Skinner, B. F. (1937). Two types of conditioned reflex: A reply to Konorski and Miller. *Journal of Genetic Psychology*, *16*, 272–279.

Sutton, R. S. (1988). Learning to predict by the method of temporal difference. *Machine Learning*, *3*, 9–44.

Sutton, R. S., & Barto, A. G. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel, & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497–537). Cambridge: MIT Press.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge: MIT Press.

Thathachar, M. A. L., & Ramakrishnan, K. R. (1981). A hierarchical system of learning automata. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-11*, 236–241.

Thathachar, M. A. L., & Sastry, P. S. (1985). A new approach to the design of reinforcement schemes for learning automata. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-15*, 168–175.

Thathachar, M. A. L., & Sastry, P. S. (1987). A hierachical system of learning automata that can learn globally optimal path. *Information Sciences*, *37*, 143–166.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Monograph*, *2*(8).

Tindell, A. J., Berridge, K. C., Zhang, J., Pecina, S., & Aldridge, J. W. (2005). Ventral pallidal neurons code incentive motivation: Amplification by mesolimbic sensitization and amphetamine. *European Journal of Neuroscience*, *22*, 2617–2634.

Watkins, C.J.C.H. 1989. Learning from delayed reward. Ph.D. Thesis, University of Cambridge, England.

Zhang, J. (2009). Adaptive learning via selectionism and Bayesianism. Part I: Connection between the two. *Neural Networks*, *22*(3), 220–228.

Zhang, J., & Chang, M. (1996). A model of operant reinforcement learning. *Journal of Mathematical Psychology*, *40*, 370.