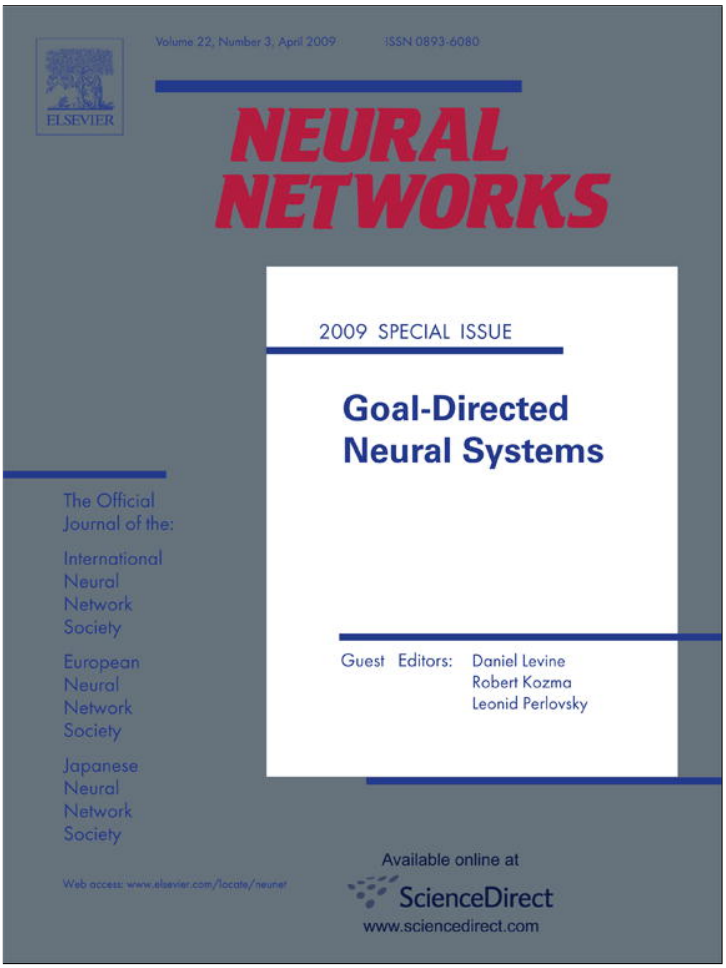

Neural Networks 22 (2009) 220–228

Contents lists available at ScienceDirect

Neural Networks

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

2009 Special Issue

Adaptive learning via selectionism and Bayesianism, Part I: Connection between the two

Jun Zhang

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

ARTICLE INFO

Article history:
Received 12 January 2009
Received in revised form 15 March 2009
Accepted 21 March 2009

Keywords:
Operant conditioning
Law of effect
Linear operator model
Stochastic learning automata
Bayesian update
Gibbs–Boltzman distribution

ABSTRACT

According to the selection-by-consequence characterization of operant learning, individual animals/species increase or decrease their future probability of action choices based on the consequence (i.e., reward or punishment) of the currently selected action (the so-called "Law of Effect"). Under Bayesianism, on the other hand, evidence is evaluated based on likelihood functions so that action probability is modified from *a priori* to *a posteriori* according to the Bayes formula. Viewed as hypothesis testing, a selectionist framework attributes evidence exclusively to the selected, focal hypothesis, whereas a Bayesian framework distributes across all hypotheses the support from a piece of evidence. Here, an intimate connection between the two theoretical frameworks is revealed. Specifically, it is proven that when individuals modify their action choices based on the selectionist's Law of Effect, the learning population, on the ensemble level, evolves according to a Bayesian-like dynamics. The learning equation of the linear operator model [Bush, R. R., & Mosteller, F. (1955). *Stochastic models for learning*, New York: John Wiley and Sons], under ensemble averaging, yields the class of predictive reinforcement learning models (e.g., [Busemeyer, J. R., & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General, 121*, 177–194; 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]).

1. Introduction

Selectionism, as originally used in the Darwinian theory of biological evolution, refers to the mechanism of natural selection through which animal traits are either preserved or become extinct according to their relative fitness. At the level of animal species, the population of organisms carrying a certain trait would increase or decrease depending on their reproductive success which is linked to the fitness of any trait transmitted by their genes. This adaptive process bears resemblance with another dynamic process whereby an animal organism acquires and modifies action skills through operant (also called instrumental) conditioning — the frequency (or tendency) of occurrence of a behavior is modified by the consequences of such behavior performed in the past. The formal parallelism between natural selection during the evolution of species and the process of selection by consequences that characterizes operant learning appeared to be noted by Skinner (1953) first informally in his 1953 book "Science and Human Behavior" and then more seriously (Skinner, 1981, 1984), and was followed up by many others since (see a recent treatment by Hull, Langman, and Glenn (2001), and the references therein). Also noteworthy is another biological system that is now known to employ a similar selection mechanism as that of Darwinian natural selection — the immune system in its reaction to antigens, which operates at the time-scale of an organism's life-span.

Bayesianism, on the other hand, deals with the concept of probability as representing a degree of subjective belief in certain propositions ("hypotheses") about the environment, and with the method of modifying the probability that a hypothesis is true in light of new evidence received from the environment. At the core of the Bayesian framework is the rule ("Bayes formula") to update the degree of belief through an assessment of the likelihood values of such evidence. Though the publication of Thomas Bayes dated back as early as 1763, this approach became enthusiastically embraced by the mainstream statistical community over a decade ago (cf. Barnardo and Smith (1994)) when technical difficulties in its implementation were overcome. By imposing a consistency requirement on belief update, Bayesian analysis enables optimal decision making despite uncertainty about a stochastic environment.

On the surface, the selectionist and the Bayesian frameworks represent two radically different perspectives on how agents (whether an individual organism or animal species) interact with their environment, with the former providing a kind of descriptive, adaptive strategy and the latter prescribing a kind of normative,

E-mail address: junz@umich.edu.

0893-6080/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2009.03.018

rational analysis. To see how the two may be potentially related, we elaborate each perspective below for the case of action modification by an agent interacting with an environment (that is assumed to be stationary for the purpose of simplicity).

1.1. Operant selection revisited

In a target article published by *Behavioral and Brain Sciences*, Hull et al. (2001) defined selection as "repeated cycles of replication, variation, and environmental interaction so structured that environmental interaction causes replications to be differential." These authors, after synthesizing considerable amount of literature, proposed a general account of selection that accommodates the gene-based biological evolution, the reaction of the immune system to antigens, and operant learning as exemplifying processes of selectionism. This general account includes three fundamental elements in a selection system, namely, replication, variation, and environmental interaction, along with the stipulation that "replication must alternate with environmental interaction so that any changes that occur in replication are passed on differentially because of environmental interaction" (Hull et al., 2001, p. 511).

Though not without fierce debate, learning of operant behavior was treated by Hull et al. as a kind of selection-by-consequence process operating on the time-scale of an animal organism's lifespan, in parallel with, per Skinner (1981, 1984), cultural evolution and biological evolution which operate on much longer time-scales. An operant behavior can be defined as an action that affects the environment and that changes over time (in its form and organization) depending on its consequences; changes in operant behavior of a particular individual is called operant learning. "Variation", in the selectionist account of operant learning, involves an operant repertoire made up of interrelated behavioral lineages, with operant lineages originating from what Skinner (1984) called "uncommitted behavior", i.e., behaviors that are inherited and not well organized with respect to the environment. "Environmental interaction" refers to the relationship between the responses and their ensuing consequences — for certain relationships, frequency of responses is increased, called "reinforcement", while for others, their frequency is decreased, called either "extinction" or "punishment". Finally, "replication" is based on modification of neurobiological structures that code for behavior, which is passed along to affect future recurrence of these behaviors. Roughly, the probability of emitting an action in operant selection is mapped to the relative proportion of an animal population carrying a certain trait in natural selection, and the reward/punishment for an action is mapped to the reproductive success in natural selection.

At the core of operant selection is Thorndike's (1898) *Law of Effect*, which states that responses that produce a satisfying (or dissatisfying) effect in a particular situation become more (or less) likely to occur again in that situation. Such an adaptive learning rule has been extensively investigated in animal learning literature, which historically led to the class of linear operator models (Bush & Mosteller, 1955; Norman, 1972) in the mathematical psychology community and subsequently the theory of stochastic learning automata (Narendra & Thathachar, 1974, 1989) in the machine learning community. The key feature of such models is that, for any single behavioral trial, probability for an action will be increased (or decreased) as long as the action selected on that trial brings about a reward (or penalty). Even though not necessarily the best or worst among all possible actions in the action repertoire, the selected action will absorb the consequence of either reinforcement or punishment. Stated alternatively, the feedback from the environment directly affects the selected action and only indirectly, through normalization of action probability, affects other non-selected actions.

Conceptually, an operant consists of an observable action R_i ($i = 1, \ldots, N$) that (a) is drawn with a certain probability p_i from a set $\mathcal{R} = \{R_1, \ldots, R_N\}$, called "action repertoire," and is emitted under a given context S, called "discriminative stimulus;" and (b) results in a reinforcer r_i being delivered to the animal. The sequence of events S–R–r in this order represents a basic operant unit. Note that the reinforcer r_i is delivered after an action R_i, while the stimulus S precedes and accompanies that action. The discriminative stimulus S merely sets the occasion for operant learning, though after learning, it acts as if in control of the occurrence of a particular action. Since actions are ultimately emitted out of the animal's action repertoire $\mathcal{R}$, action probability p_i is the proper variable to study.

Operant conditioning embodies a wide range of reinforcement learning situations in which the animal's response tendency is modified by experience. This probabilistic nature of operant responses is in contrast with the instinctual, reflexive responses in classical (i.e., Pavlovian) conditioning that can always be "elicited" by the unconditioned stimulus. The linear operator model, though overly simplistic, characterizes operant learning as

$$p_i^{\text{new}} = p_i^{\text{old}} + \vartheta(1 - p_i^{\text{old}}) \quad (0 < \vartheta < 1), \tag{1}$$

where ϑ is a quantity in proportional to the reward value obtained by the learning agent, and i is the selected action. Linear operator models formed the basis of much sophisticated analyses, the so-called theory of stochastic learning automata. Modern approaches in machine learning (i.e., temporal-difference based methods) often treat operant conditioning and Pavlovian condition in a unified fashion, resulting in the form of predictive reinforcement learning models and the actor–critic architecture with concurrent learning of both action probability and reward prediction (reviewed in Sutton and Barto (1998)). However, such treatment, by making reinforcement contingent upon reward comparison, seems to have deviated from a strict interpretation of selectionism (Thorndike's Law of Effect, claiming to amplify any selected action so long as it is rewarded).

1.2. Bayesian belief update revisited

If we view action selection in learning an operant behavior as an act of hypothesis testing, the set of allowable actions in the action repertoire as the set of hypotheses, and the reward or penalty following the performance of an action as evidence obtained from the environment, then the Bayesian framework also provides a formula to update choice (or action) probability according to evidence gathered on individual trials. This is done through the use of likelihood functions, which evaluate evidence from the perspectives of different hypotheses—the likelihood value associated with each hypothesis accounts for the same evidence but with differential strength. Belief about a hypothesis prior to receiving evidence ("prior probability") is modified by the likelihood functions and becomes updated ("posterior probability") upon the receipt of evidence by the learning agent. Since beliefs about all competing hypotheses are being updated at the same time, a single piece of evidence will modify the entire probability distribution—feedback as a consequence of interaction with environment is to be distributed across all possible hypotheses to achieve adaptivity in action selection.

Formally, let the set of hypotheses be denoted as a discrete set $\{1, 2, \ldots, N\}$, where the ith hypothesis states that action R_i is the best action to perform, and the degree of belief of it being true as p_i, ($i = 1, \ldots, N$). After receiving evidence e, and given likelihood functions $l_i(e)$ which describe how e might be generated in the environment, Bayesian analysis prescribes an update of belief to be

$$p_i^{\text{new}} = \frac{p_i^{\text{old}} l_i(e)}{\sum_i p_i^{\text{old}} l_i(e)}. \tag{2}$$

We take (1) and (2) to be representatives of the distinct styles of adaptive computation afforded by selectionist and Bayesian formulations, respectively, though undoubtedly they are grossly simplified versions. Their sharp difference, nevertheless, on how action probability (or belief) is to be modified based on environmental feedback is striking: the selectionist formulation (linear operator model) attributes reward exclusively to the selected action (focal hypothesis), whereas Bayesian formulation distributes evidence across all hypotheses. Given such a difference, one wonders whether there can be any formal connection between these two styles of adaptive computation? Are there any circumstances under which the selectionist computation can be viewed as normative and rationally based? We intend to provide answers to these questions by demonstrating an intimate connection between the operant reinforcement learning dynamics and the Bayesian learning dynamics.

Specifically, it will be shown that when individuals modify their action probabilities based on the selectionist rule, i.e., the Law of Effect as exemplified by (1), the learning population, on the ensemble level, will evolve according to a Bayesian dynamics as exemplified by (2). Since ensemble averaging, in the limit of infinitesimal learning rate, is equivalent to time averaging, this means that selectionism, despite its seemingly arbitrariness in trial-by-trial action selection and exclusiveness in causal attribution, nevertheless is optimal in the Bayesian sense when incremental modification occurs at small steps. The Law of Effect, we argue, actually implements (albeit inefficiently and approximately) the Bayesian inference scheme so long as there is an ensemble of learning agents each faithfully carrying out the trial-and-error type instrumental learning — the accuracy of such approximation depends solely on the learning rate.

The remaining of the paper is organized as follows. Section 2.1 recalls a simple version of the linear operator model describing how the action probability of a learning agent changes due to differential reinforcement. Section 2.2 then analyzes an ensemble of such learning agents, and shows how linear operator model turns into a predictive reinforcement learning model involving reward comparison. Section 2.3 studies the continuous limit of the ensemble-level dynamics by solving a first-order ordinary differential equation. Section 2.4 compares the solution of the ensemble-level equation to the Bayes formula, and identifies the likelihood function in terms of reward. Section 2.5 further investigates the formal equivalence between the selectionist and Bayesian formulation by elaborating the meaning of ensemble-level analysis. A computer simulation is reported that demonstrates the difference between the two formulations, and the condition under which they become identical. Section 3 closes with a discussion of the implications of this work in relation to an experiment on adaptive learning. In the sequel (Zhang, 2009), we will show how this equivalence, when applied to the acquisition of an action sequence (“operant chain”), naturally leads to the notion of an internally generated conditional (or secondary) reinforcement signal that serves to predict terminal (or primary) reward, the cornerstone for solving sequential credit-assignment problem.

2. Mathematical analysis

In this section, we revisit the learning of a basic operant unit. In this case, the stimulus factor S, which sets the context for operant learning, is completely hidden from the learning dynamics. It is the probability distribution of the set of all allowable actions under S that is under consideration, and it is the competitive reinforcement among them that governs their modification. Reformulating the linear operator model (Bush & Mosteller, 1955) and examining the “average” learning trajectory of an ensemble of such animal organisms allows us to propose a Bayesian interpretation of the ensemble dynamics. The effects of different environmental contexts (different S's) will be studied in Zhang (2009) under the context of acquisition of an operant sequence.

2.1. Single-trial modification of action probability

Suppose that under environmental context S, an animal randomly executes an action R_i out of its action repertoire $\mathcal{R} = \{R_1, R_2, \ldots, R_N\}$. The conditional probability is denoted $p_i = \text{Prob}(R_i|S)$. After the action R_i, the animal receives a reward θ_i in return (assuming, for simplicity, that all $\theta_i \geq 0$ and that they are not identical). This, according to Law of Effect, results in an increase in p_i, the probability or tendency of the animal's executing the same action again. We follow the standard model in mathematical learning theory (see Atkinson, Bower, and Crothers (1965), Bush and Mosteller (1955), Estes (1950), and Norman (1972)) to describe this change $\delta p_i^{(n)}$ at discrete time step n of the learning process:

$$\delta p_i^{(n)} = \epsilon\theta_i(1 - p_i^{(n)}),$$

or equivalently

$$p_i^{(n+1)} = (1 - \epsilon\theta_i)p_i^{(n)} + \epsilon\theta_i = p_i^{(n)} + \epsilon\theta_i(1 - p_i^{(n)}).$$

Here, the small constant $\epsilon > 0$ reflects the learning rate, and is explicitly written out for convenience. Note that the reward values θ_i can also be stochastic, and we assume that $\epsilon\theta_i < 1$. On the other hand, because of normalization of probability

$$\sum_k p_k^{(n+1)} = \sum_k p_k^{(n)} = 1,$$

the increase of p_i due to reinforcing (with θ_i) the animal's spontaneous execution of R_i results in an effective decrease of the probability of the same animal executing other actions R_j ($j \neq i$):

$$\delta p_j^{(n)} = -\epsilon\theta_i p_j^{(n)} \quad (j \neq i),$$

or equivalently

$$p_j^{(n+1)} = p_j^{(n+1)} - \epsilon\theta_i p_j^{(n)}.$$

Compactly written, in terms of components of the action probability vector $\mathbf{p}^{(n)} = (p_1^{(n)}, \ldots, p_N^{(n)})$, the animal's executing action R_i at step n gives rise to

$$\delta p_k^{(n)} = \epsilon\theta_i(e_{ik} - p_k^{(n)}) \quad k = 1, \ldots, N \tag{3}$$

with the Kronecker delta

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

Collectively for all components of action probability, (3) can be written as

$$\delta\mathbf{p}_i^{(n)} = \epsilon\theta_i(\mathbf{e}_i - \mathbf{p}_i^{(n)})$$

where

$$\delta\mathbf{p}_i^{(n)} = (\delta p_1^{(n)}, \ldots, \delta p_i^{(n)}, \ldots, \delta p_N^{(n)})$$

is the change of action probability $\mathbf{p}^{(n)}$ after action R_i is performed. The single-trial learning rule (3) reflects a Markov transition of action probability values, from $\mathbf{p}^{(n)}$ to $\mathbf{p}^{(n+1)}$, with absorbing states at the corners $\mathbf{e}_i$ (i.e., the unit vector with the kth component e_{ik}) of the simplex $\mathcal{S}$ of all allowable action probability

$$\mathcal{S} = \left\{ \mathbf{p} : p_k \geq 0; \ \sum_k p_k = 1 \right\}.$$

This is well-known from the theory of stochastic learning automata (see Narendra and Thathachar (1974)). The learning rule (3) correspond to their S-model with linear reward-inaction (L_{R-I}) scheme.

2.2. Ensemble-level modification of action probability

The operant learning rule describes, on a single-trial basis, the change of action probability $\delta \mathbf{p}_i^{(n)}$ when a given action R_i is executed at time n. During operant conditioning, any actions in the action repertoire is possible. Since the probability of action R_i being executed at time n is $p_i^{(n)}$, the *average* change of $\mathbf{p}^{(n)}$, (average in the sense of an ensemble of such learning systems), should be calculated after applying this weighting factor. Component-wise, it is[1]

$$\begin{aligned}\Delta p_k^{(n)} &= \sum_i p_i^{(n)} \delta p_k^{(n)} = \sum_i p_i^{(n)} \epsilon \theta_i (e_{ik} - p_k^{(n)}) \\ &= \epsilon p_k^{(n)} \theta_k - \epsilon p_k^{(n)} \sum_i p_i^{(n)} \theta_i.\end{aligned}$$

Introducing the expected reward (averaged under action probability $\mathbf{p}^{(n)}$) at step n

$$\Theta^{(n)} = \sum_k p_k^{(n)} \theta_k, \tag{4}$$

we obtain the equation governing time evolution of $\mathbf{p}^{(n)}$ (on the ensemble level)

$$\Delta p_k^{(n)} = \epsilon p_k^{(n)} (\theta_k - \Theta^{(n)}), \quad k = 1, 2, \ldots, N. \tag{5}$$

Note that in the above calculation of $\Delta p_k^{(n)}$, two sources of contribution to a change in $p_k^{(n)}$ have been taken into account:

(1) an active increase, in the amount $\epsilon p_k^{(n)} \theta_k (1 - p_k^{(n)})$, due to reinforcement when action R_k is executed;
(2) passive decreases, in the amount $-\epsilon \sum_{j \neq k} p_j^{(n)} \theta_j p_k^{(n)}$, due to probability normalization when actions R_j ($j \neq k$) are executed.

These two contributions sum to exactly give the expression in Eq. (5). The overall dynamics of learning, according to (5), is easily understood: at each learning step n, those actions R_k that yield above-average rewards will increase their probability of emission, whereas those actions that yield below-average rewards will decrease their probability of emission:

$$\begin{aligned}\Delta p_k^{(n)} &> 0 \quad \text{if } \theta_k > \Theta^{(n)}, \\ \Delta p_k^{(n)} &< 0 \quad \text{if } \theta_k < \Theta^{(n)}.\end{aligned}$$

The consequence of such learning is to promote those actions with higher rewards and to weed out those with lower rewards, where the expected (average) reward is used as a standard to be compared to. The expected reward Θ itself would increase because action probability will be redistributed in favor of better-rewarded actions. Formally,

$$\begin{aligned}\Delta \Theta^{(n)} &= \sum_k \Delta p_k^{(n)} \theta_k = \sum_k \epsilon p_k^{(n)} (\theta_k - \Theta^{(n)}) \theta_k \\ &= \epsilon \left(\sum_k p_k^{(n)} (\theta_k - \Theta^{(n)})^2 \right) = \epsilon \sigma^{(n)} > 0,\end{aligned}$$

where

$$\sigma^{(n)} = \sum_k p_k^{(n)} (\theta_k - \Theta^{(n)})^2 \tag{6}$$

is the $\mathbf{p}^{(n)}$-weighted (i.e., ensemble averaged) reward variance. This shows that as learning proceeds (i.e., as n increases), the average reward Θ increases monotonically. The size of its increase $\Delta \Theta^{(n)}$ is proportional to the variance of reward distribution $\sigma^{(n)}$ at step n. Thus, the learning equation (5)effectively achieves a stochastic *ascend* of the average reward Θ. In the language of stochastic processes (see e.g., Karlin and Taylor (1975)), the stochastic variable $\Theta^{(n)}$ can be described as a "sub-martingale" process with respect to the stochastic action selection and modification as learning progresses.

The ensemble-level learning equation (5) can be cast as

$$\Delta p_k = p_k \left((1 - p_k)\theta_k - \sum_{i \neq k} p_i \theta_i \right) = p_k (1 - p_k)(\theta_k - \theta_k'), \tag{7}$$

where

$$\theta_k' = \frac{\sum_{i \neq k} p_i \theta_i}{1 - p_k} = \frac{\sum_{i \neq k} p_i \theta_i}{\sum_{i \neq k} p_i} \tag{8}$$

is the average reward *excluding* action R_k (i.e., when R_k is no longer an option). Therefore, $\Delta p_k > 0$ (or < 0) if and only if performing action R_k yields more (or less) reward than not performing action R_k.

A learning algorithm that ensures $\Delta \Theta^{(n)} > 0$ for each step of learning is said to be "absolutely expedient" (Narendra & Thathachar, 1989). Absolute expediency is a property of general associative reinforcement learning algorithms (Williams, 1992). In the current case, even when the learning rate is non-constant $\epsilon = \epsilon(\mathbf{p})$, the algorithm (3) is absolutely expedient (Narendra & Thathachar, 1989). Nevertheless, it should be kept in mind that an individual learning automaton (agent) may sometimes converge to non-optimal solutions, though the probability of such errors can be made arbitrarily small when the learning rate approaches zero (called "ϵ-optimality"). This distinction between single-trial and ensemble-level dynamics will be discussed further in Section 2.5.

In short, operant learning (on the ensemble level) is characterized by the simultaneous modification of (a) action probability and (b) average reward (see Fig. 1). On the one hand, action probability increases or decreases depending on whether the reward received is above or below the average reward; on the other hand, average reward increases as a result of applying such rules of modifying action probabilities. The repeated application of such learning scheme leads to gradual improvement of the fitness (adaptiveness) of animal behavior on a population level: action probabilities will be redistributed more and more towards better-rewarded actions while average reward increases steadily. Therefore, the operant learning principle, when operating at the ensemble level, augments certain action tendencies while diminishing others (according to their consequences), even though at the single-trial level, it appears that any chosen action results in an augmentation due to the positivity of a reward. The ensemble-level equation (5) resembles an entirely different class of learning models based on reinforcement prediction and comparison, such as the adaptive decision model for human choice performance (Busemeyer & Myung, 1992), and the predictive Hebbian learning for modeling choice behavior of insects (Montague, Dayan, Person, & Sejnowski, 1995) and primates (Montague, Dayan, & Sejnowski, 1996) — all these models have in common a process of trial-by-trial reward estimation and prediction, a procedure of comparing the actual reward received with the predicted reward value, and using their discrepancy as the reinforcement signal to drive learning.

2.3. Ensemble-level dynamics: Continuous limit

The ensemble-level dynamics introduced in the last subsection describes the (average) change of action probability when the "sampling" of actions is governed by the action probability itself. The ensemble averaging operation applied to the linear operator model has the following two interpretations. First, the population

[1] Because we are calculating average change of $p_k^{(n)}$, we may replace the stochastic reward variables θ_k by their means $\hat{\theta}_k$ in the equations. For simplicity, we retain the notation θ_k here and below to represent the mean reinforcement value associated with action R_k.

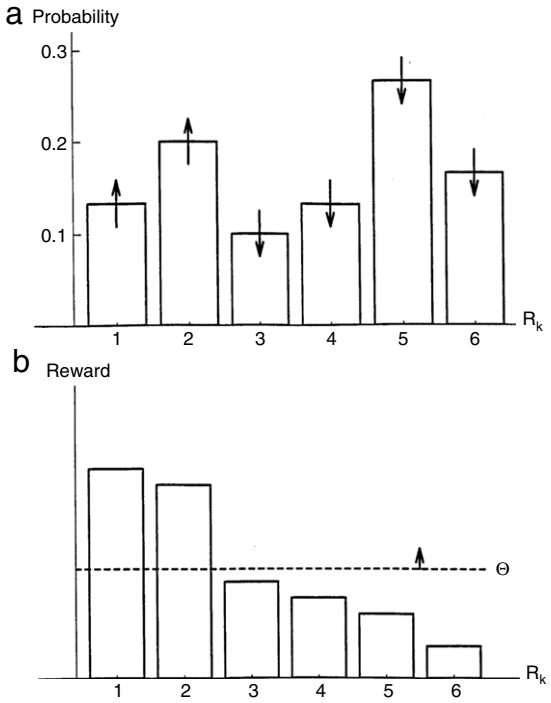

Fig. 1. Schematic illustration of interaction between action probability and average reward, for an action repertoire with $N = 6$. (a). Action probabilities $p_k^{(n)}$ are indicated by the height of the bar. The arrows indicate whether the average change $\Delta p_k^{(n)}$ is greater or smaller than 0. (b). The reward values θ_k associated with action R_k, which remain fixed throughout learning, are indicated by the height of the bar. The dotted line represents the value of $\Theta^{(n)}$, the reward average at step n. Depending on whether $\theta_k > \Theta^{(n)}$ or $\theta_k < \Theta^{(n)}$, action probabilities $p_k^{(n)}$ will increase or decrease (arrows in (a)). As a result, $\Delta\Theta^{(n)} > 0$.

of learning agents is assumed to be homogeneous, with each agent starting with the same initial action probability and undergoing action modification according to the linear operator model — the only difference between agents is that different actions may be selected which is probabilistically prescribed by the then-current action probability value itself. Second, we imagine that any single agent samples its various actions using this action probability (such that some actions are carried out by more agents than others) while modifying the value of action probability every time any action is carried out.[2] When the learning rate ϵ is small, then the ensemble averaging operation can be approximated by the time averaging operation, such that ensemble dynamics is to be described by some continuous dynamics. Making the identification $t = n\epsilon$ and $dt = \epsilon\,\delta n = \epsilon$ for $\delta n = 1$, the trajectory of single-trial evolution of action probability can be described by the Ito stochastic differential equation

$$\mathrm{d}p_k = p_k\left(\theta_k - \sum_i p_i\theta_i\right)\mathrm{d}t + o\left(\epsilon\mathrm{d}\mathbf{W}\right)$$

where $\mathbf{W}$ is the Wiener noise (diffusion) process (see Gardiner (1985)). The corresponding Langevin equation has been used to study the convergence properties of single-trial trajectories (Phansalkar & Thathachar, 1995; Thathachar & Sastry, 1985). In the limit $\epsilon \to 0$, the diffusion term may be discarded to result in an ordinary differential equation which is the continuous equivalent of the discrete version (5)[3]:

$$\frac{\mathrm{d}p_k}{\mathrm{d}t} = p_k\left(\theta_k - \sum_i p_i\theta_i\right). \tag{9}$$

The right-hand side of (9) defines a flow field in the probability simplex $\mathcal{S}$, with fixed points at $\mathbf{p} = \mathbf{e}_l$, the unit vector (with the kth component e_{lk}) at one of its corners. It is interesting to note that the system of Eq. (9) is also known as the "replica dynamics" in population biology (see, e.g., Hofbauer and Sigmund (1998)).

Following the derivation of (7), Eq. (9) may be re-written as

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\log\frac{p_k}{1-p_k}\right) = \theta_k - \theta_k',$$

where θ_k' is given by (8). The log-odds formulation has the clear interpretation that action probability is to be reinforced if and only if it yields more reward when performed than when not performed.

We now remove the assumption that the average reinforcement value θ_k associated with action R_k is stationary, and allow it to be a function of time t. Under this less restrictive assumption about the environment, $\theta_k = \theta_k(t)$, we can prove

Proposition 1. *Given reward functions $\theta_k(t)$, $k = 1, \ldots, N$, the solution for (9) is:*

$$p_k(t) = \frac{p_k(0)e^{\phi_k(t)}}{\sum_i p_i(0)e^{\phi_i(t)}} \tag{10}$$

where $p_k(0)$ is the action probability prior to learning, i.e., the initial bias for action selection, and

$$\phi_k(t) = \int_0^t \theta_k(\tau)\mathrm{d}\tau$$

with $\phi_k(0) = 0$. For stationary rewards $\theta_k(t) = \theta_k$, the solution becomes:

$$p_k(t) = \frac{p_k(0)e^{\theta_k t}}{\sum_i p_i(0)e^{\theta_i t}}. \tag{11}$$

Proof. First, Eq. (9) can be cast into the following form:

$$\frac{\mathrm{d}\log p_k}{\mathrm{d}t} - \theta_k(t) = -\sum_i p_i(t)\theta_i(t) \tag{12}$$

where the right-hand side is a function of t independent of index k. As a first-order ODE, the homogeneous equation corresponding to (12) is

$$\frac{\mathrm{d}\log\tilde{p}_k}{\mathrm{d}t} - \theta_k(t) = 0,$$

whose solution is immediately found to be

$$\tilde{p}_k(t) = \tilde{p}_k(0)e^{\phi_k(t)}.$$

Note that, for the non-homogeneous Eq. (12), the right-hand side is a function of t irrespective of subscript k. This prompts us to try the following closed-form solution:

$$p_k(t) = \frac{\tilde{p}_k(t)}{Z(t)} = \frac{1}{Z(t)}p_k(0)e^{\phi_k(t)} \tag{13}$$

[2] This describes the scenario called "on-policy control" in reinforcement learning (Sutton & Barto, 1998), i.e., when the policy being modified is also the policy being implemented for action selection in learning.

[3] In the generic absolutely expedient models where $\epsilon = \epsilon(\mathrm{p})$, the learning-rate factor cannot be simply absorbed into t. However, the direction of such flow is still given by the following equation.

with the function $Z(t)$ yet to be determined subject to $Z(0) = 1$. Since

$$\frac{\mathrm{d}\log p_k}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(\log p_k(0) + \phi_k(t) - \log Z(t)) = \theta_k(t) - \frac{\mathrm{d}\log Z(t)}{\mathrm{d}t},$$

we have, from (9),

$$\frac{\mathrm{d}\log Z(t)}{\mathrm{d}t} = \sum_i p_i(t)\theta_i(t) = \sum_i \frac{1}{Z(t)} p_i(0) e^{\phi_i(t)}\theta_i(t)$$

where (13) is substituted in the last step. Therefore,

$$\frac{\mathrm{d}Z}{\mathrm{d}t} = \sum_i p_i(0) e^{\phi_i(t)}\theta_i(t).$$

Integrating both sides from 0 to t:

$$Z(t) - Z(0) = \sum_i p_i(0)(e^{\phi_i(t)} - 1) = \sum_i p_i(0) e^{\phi_i(t)} - 1,$$

we obtain

$$Z(t) = \sum_i p_i(0) e^{\phi_i(t)}.$$

With $Z(t)$, the solution $p_k(t)$ is readily obtained and expressed as Eq. (10). ◇

The solution (10) can be interpreted as a normalized exponential dynamic using cumulative reward functions ϕ_k (or equivalently θ_k): the numerator indicates that an action tendency for R_k will increase exponentially due to continuous, positive reinforcement θ_k; the normalizing denominator indicates a competition between various action tendencies ($k = 1, \ldots, N$) that are all being concurrently reinforced but at different rates. The evolution of action probability $p_k(t)$ reflects both a reinforcement effect and a competition effect. The result of this competitive, reinforcement learning is determined by the reward structure in terms of θ_k's. We mention in passing that an affine transformation of the reward structure

$$\theta_k \to a\theta_k + b$$

does not affect the ensemble-level dynamics. The constant b does not affect (11), whereas the constant $a > 0$ provides the scaling of time for learning (akin to the temperature parameter in a Gibbs–Boltzmann distribution).

We note that the quantity $Z(t)$ introduced in the proof is analogous to the thermal-dynamical partition function, and where the probability p_k is said to obey the Gibbs–Boltzmann distribution. Under stationary reward (i.e., θ_k's are constants), $Z(t)$ can also be view as the moment-generating function of a discrete probability distribution. The partition (moment-generating) function $Z(t)$ can be used to calculate a variety of useful quantities, including the mean expected reward $\Theta(t)$

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

and the variance $\sigma(t)$ of p_k-distributed reward values

$$\sigma(t) = \frac{\mathrm{d}^2\log Z(t)}{\mathrm{d}t^2} = \sum_i p_i(t)(\theta_i - \Theta(t))^2.$$

These two quantities satisfy

$$\frac{\mathrm{d}\Theta(t)}{\mathrm{d}t} = \sigma(t),$$

which is the continuous-time version of Eq. (6).

The continuous dynamics (9) can be extended from the discrete response case (where actions and rewards are indexed by $k = 1, 2, \ldots, N$) to a continuum of actions (where action probability $p(\mathbf{x}, t)$ and reward $\theta(\mathbf{x}, t)$ are specified by continuous variable $\mathbf{x}$) in the action repertoire:

$$\frac{\partial p(\mathbf{x}, t)}{\partial t} = p(\mathbf{x}, t)\left(\theta(\mathbf{x}, t) - \int p(\mathbf{x}, t)\theta(\mathbf{x}, t)\mathrm{d}\mathbf{x}\right). \quad (14)$$

This formulation is useful whenever a continuum of actions is examined, for instance, during the acquisition of motor skills where movement parameters (such as force or timing) are being fine tuned. The average and variance of reward can be analogously defined:

$$\Theta(t) = \int p(\mathbf{x}, t)\theta(\mathbf{x}, t)\mathrm{d}\mathbf{x},$$

$$\sigma(t) = \int p(\mathbf{x}, t)(\theta(\mathbf{x}, t) - \Theta(t))^2\mathrm{d}\mathbf{x}.$$

The solution of (14) can be expressed using the partition function

$$Z(t) = \int p(\mathbf{x}, 0) e^{\phi(\mathbf{x}, t)}\mathrm{d}\mathbf{x},$$

with

$$\phi(\mathbf{x}, t) = \int_0^t \theta(\mathbf{x}, \tau)\mathrm{d}\tau.$$

2.4. Bayesian interpretation

The evolutionary trajectory of $p_k(t)$ given by Eq. (10) has an interesting Bayesian interpretation. Recall the Bayes formula (2) which specifies the relationship between the prior and posterior probabilities (that a hypothesis is true before and after obtaining evidence) and the likelihood function (that such evidence is produced under the various hypotheses). Comparing with (10), we can interpret $p_k(0)$ as the prior probability (action probability before learning), $p_k(t)$ as the posterior probability (action probability after learning), and $\phi_k(t)$ as the log likelihood function:

$$\log l_k = \phi_k(t).$$

This is to say, the animal can be viewed as actively and constantly testing and revising its belief about a set of hypotheses, by emitting actions and modifying action tendencies by treating reward for actions as evidence.

The probability ratio of any two actions R_i and R_j during learning is:

$$\frac{p_i(t)}{p_j(t)} = \frac{p_i(0)}{p_j(0)} e^{\phi_i(t) - \phi_j(t)}.$$

When rewards are constant (i.e., a stationary environment) $\theta(t) = \theta_k$, $\phi_k(t) = \theta_k t$, the above ratio will, depending on the relative difference in reward magnitudes, increase or decrease exponentially as a function of time. Therefore, the learning algorithm makes very fine discriminations against minute differences of reward values and will eventually acquire the maximally rewarded action.

The functions $\phi_k(t)$ represent the "total" accumulated rewards for each action R_k up to time t. Denote

$$\phi_i(t|s) = \int_s^t \theta_i(\tau)\mathrm{d}\tau = \phi_i(t) - \phi_i(s).$$

Then we have, for all $0 \le s \le t$,

$$p_k(t) = \frac{p_k(s) e^{\phi_k(t|s)}}{\sum_i p_i(s) e^{\phi_i(t|s)}}.$$

This implies that the functions $[\phi_1(t|s), \ldots, \phi_N(t|s)]$ "translates" the action probability values $\mathbf{p}(s) = [p_1(s), \ldots, p_N(s)]$ at time s

to action probability values $\mathbf{p}(t) = [p_1(t), \ldots, p_N(t)]$ at a later time t.

Since the ensemble-level dynamics is invariant under $\phi_k \to \phi_k + b(t)$, we can define the functions ψ_k

$$\psi_k(t|s) = \phi_k(t|s) - \sum_i p_i(s)\phi_i(t|s)$$

which have zero-expectation (with respect to $\mathbf{p}(s)$). We have

Corollary 2. *For all $0 \le s \le t$ and with $\mathbf{p}(s)$ fixed, the set of functions $\psi_k(t|s)$ and $\mathbf{p}(t) = [p_1(t), \ldots, p_N(t)]$ are in one-to-one correspondence. In particular,*

$$p_k(t) = \frac{p_k(s)e^{\psi_k(t|s)}}{\sum_i p_i(s)e^{\psi_i(t|s)}}$$

$$\Longleftrightarrow \psi_k(t|s) = \log\frac{p_k(t)}{p_k(s)} - \sum_i p_i(s)\log\frac{p_i(t)}{p_i(s)}.$$

Proof. The first equation is a slight modification of (10), while the second equation is then obtained by direct substitution. ◇

Therefore, with any fixed reference point $\mathbf{p}(s)$, the action probability $p_k(t)$ and the set of $\psi_k(t|s)$'s can be mutually inferred. See Zhang and Hasto (2006) for a more general discussion of this type of probability representation as used in Bayesian formulation.

Recall the Kullback–Leibler divergence (cross-entropy)

$$\mathrm{K}(\mathbf{p}(s), \mathbf{p}(t)) = \sum_i p_i(s)\log\frac{p_i(s)}{p_i(t)},$$

which characterizes the asymmetric distance between $\mathbf{p}(s)$ and $\mathbf{p}(t)$. It can be shown, by direct calculation, that the KL measure in the current case is positive

$$\mathrm{K} > 0,$$

strictly increasing

$$\frac{\mathrm{dK}}{\mathrm{d}t} = \Theta(t) - \Theta(s) > 0,$$

and strictly convex

$$\frac{\mathrm{d}^2\mathrm{K}}{\mathrm{d}t^2} = \frac{\mathrm{d}\Theta}{\mathrm{d}t} = \sigma(t) > 0.$$

This shows that learning achieves a global ascend on the KL measure and the dynamics accelerates as time progresses.

To summarize: the ensemble-level operant learning equation, when cast in the form of (5), presents a clear picture of how the change of action probability is gauged by the current, average reward, and how the change of average reward, in turn, is related to the change of action probability. In the continuous limit (9), the analytic solution offers additional insight into the dynamics of this selectionist's learning and its connection to Bayesianism.

2.5. Relationship between single-trial and ensemble-level learning

It is important to note the difference and connection between the single-trial operant rule (3) and the ensemble-level equation (5). In our formulation, the learning agent is conceptualized by a probability vector that summarizes its action tendency at any given point, and the ensemble is made up by a multitude of such agents. Consider an ensemble of agents all with the same starting action probability vector $\mathbf{p}^{(n)}$, at learning step n. Update of action probability of all such agents is a stochastic process that splits the initial value $\mathbf{p}^{(n)}$ into N possible values $\mathbf{p} + \delta\mathbf{p}_i^{(n)}$ $(i = 1, 2, \ldots, N)$ with probabilities $p_i^{(n)}$, depending on the action chosen by the individual agent. As long as the learning rate ϵ is non-diminishing, the evolution of the ensemble (more accurately, of the probability density function over $\mathbf{p}$) involves both a first-moment drift process, which describes the mean (ensemble-average) of action probability change $\Delta\mathbf{p}^{(n)}$ as given by (5), and a second-moment diffusion process, which is related to the covariance matrix associated with the N vectors $\delta\mathbf{p}_i^{(n)}$. When Eq. (5) is used in place of (3), the drift factor of action probability update is captured while the diffusion factor is omitted.

Individual learning agents operating on (3) will eventually converge (with probability 1) to one of the absorbing states at the corners of the probability simplex $\mathcal{S}$, though it is also known that they sometimes converge to non-optimal solutions so long as ϵ is non-vanishing (Lakshmivarahan & Thathachar, 1973, 1976). The upper bound of the probability of such non-optimal performance has been estimated (Norman, 1968). Since it is the accumulation of diffusion that may lead the single-trial Markov chain of action probability to be absorbed by non-optimal states, we now examine the effects of its omission when we turn to the ensemble-level description (5) which solely captures the drift process. For any vector $\mathbf{p}$ in the probability simplex $\mathcal{S}$, we can define the converging (absorbing) probability $\Gamma_k(\mathbf{p})$ as the probability that the learning agent, when starting at the initial value $\mathbf{p}$ and being updated according to (3) step-by-step and repeatedly, will eventually evolve into $\mathbf{e}_k$, the kth vertex of $\mathcal{S}$. The Markov nature of single-trial updating implies that

$$\sum_i p_i\Gamma_k(\mathbf{p} + \delta\mathbf{p}_i) = \Gamma_k(\mathbf{p}).$$

It follows that

$$\Gamma_l(\mathbf{p} + \Delta\mathbf{p}) - \Gamma_l(\mathbf{p}) = o(\epsilon^2).$$

Since $\Delta\mathbf{p}$ is on the first order of ϵ, the number of step N_T needed for $\mathbf{p}$ to drift across some finite distance T is $N_T \sim T/\epsilon$. The cumulative change in absorbing probability is $N_T \cdot \epsilon^2 \sim \epsilon \cdot T$, which approaches zero as $\epsilon \to 0$. This is to say, absorbing probability is almost (as the learning rate approaches 0) conserved when action probability is updated according to the ensemble-level rule (5) where the diffusion factor has been omitted from the single-trial rule (3). In fact, since diffusion is on the order of ϵ^2, the total deviation of the trajectory of single-trial Markov process from the trajectory of ensemble-level equation is also $\epsilon \cdot T \to 0$. The above arguments on the absorbing probability and on the behavior of single-trial trajectory can be stated more rigorously using the language of weak convergence introduced into the study of stochastic learning automata (Phansalkar & Thathachar, 1995; Thathachar & Sastry, 1985): the trajectory of a discrete, stochastic process converges in probability, at all points of the trajectory, to the trajectory of a deterministic ordinary differential equation (ODE), so long as ϵ is small. In this current setting, the deterministic ensemble-level dynamics (5) is given by (10).

Here we numerically simulated sample paths of single-trial operant learning for a three-choice situation (Fig. 2). Each simulation started with an action probability $\mathbf{p} = (0.1, 0.3, 0.6)$. The stochastic reward associated with each action was normally distributed, with mean values $(\theta_1, \theta_2, \theta_3) = (0.05, 0.03, 0.02)$, and standard deviation 0.006. There the maximally rewarded action R_1 had lowest probability initially ($p_1 = 0.1$), and the learning-rate parameter ϵ was absorbed into the reward values. At each step, action was randomly chosen according to $\mathbf{p} = (p_1, p_2, p_3)$; then $\mathbf{p}$ was updated based on the reward for that selected action using (3). Three sample paths were generated, with the associated curves for average reward Θ plotted as well. After about 200 learning trials, Θ reached an asymptotic value of 0.05, which was the value of maximal reward associated with action R_1.

It should be noted all simulated trajectories did not end up in vertex **A**, the maximally rewarded action. As mentioned earlier,

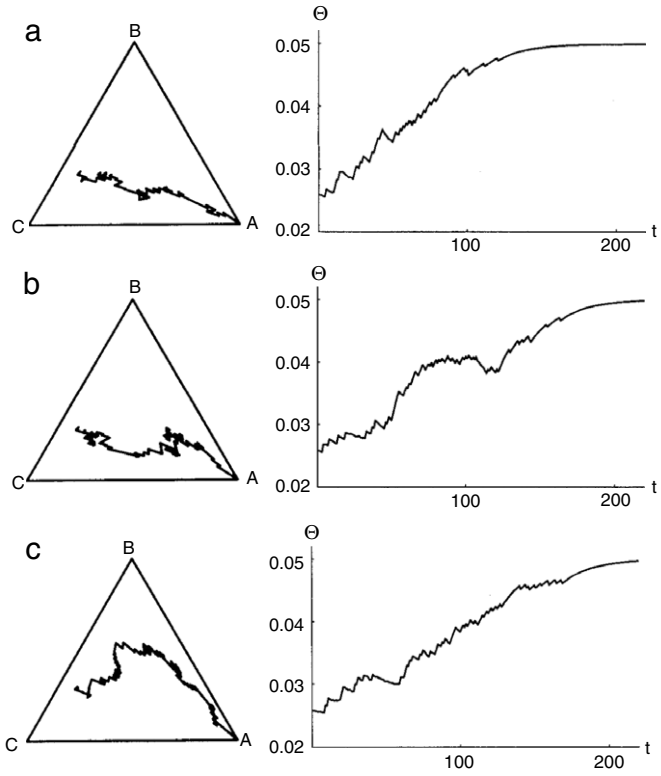

Fig. 2. Single-trial learning trajectory of action probability (on the left) and the growth of average reward Θ (on the right) for $N = 3$. The action probability $\mathbf{p} = (p_1, p_2, p_3)$ is represented by a point in the triangle representing the probability simplex $\mathcal{S} = \{\mathbf{p} : p_1 + p_2 + p_3 = 1\}$ with corners $\mathbf{A} = (1, 0, 0)$, $\mathbf{B} = (0, 1, 0)$, and $\mathbf{C} = (0, 0, 1)$. Stochastic reward values are used with means $(\theta_1, \theta_2, \theta_3) = (0.05, 0.03, 0.02)$. Three separate runs were depicted in (a)–(c). Each run starts with the same initial position: at $t = 0$, $(p_1, p_2, p_3) = (0.1, 0.3, 0.6)$, with $\Theta = 0.1 \cdot 0.05 + 0.3 \cdot 0.03 + 0.6 \cdot 0.02 = 0.026$. Due to the stochastic nature of the value of the reward and the action actually selected on each trial, the step changes in the trajectory of p_k's vary considerably. The average reward Θ increases despite trial-by-trial fluctuations.

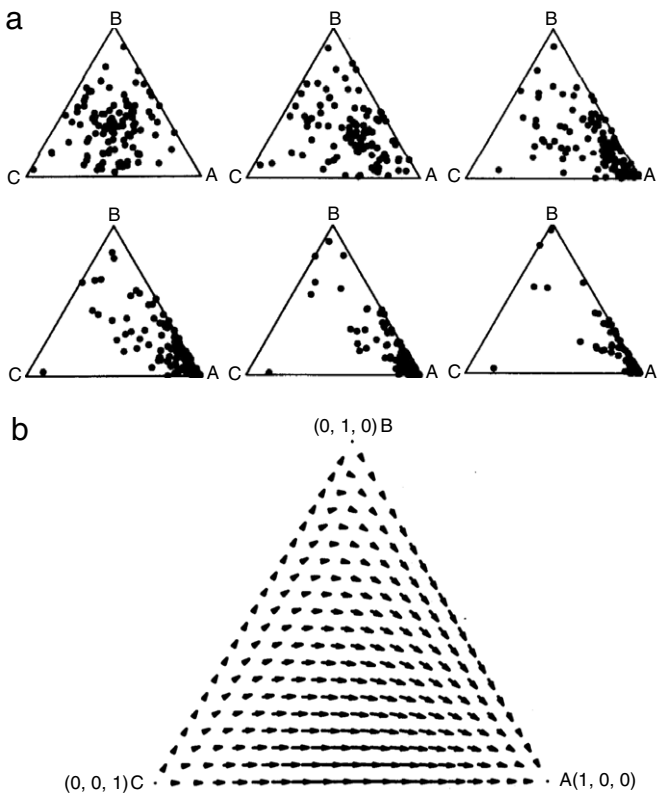

Fig. 3. Diagrams of (a) the evolution of a population of agents undergoing operant conditioning and (b) the associated mean flow field. The same reward values as in Fig. 2 were used. In (a), initial action probabilities ($t = 0$) are randomly generated and evenly distributed on the probability simplex. Gradually, the distribution of the population density shifts towards vertex **A**, which corresponds to the maximally rewarded action. The six panels correspond to $t = 0$, $t = 30$, $t = 60$, $t = 90$, $t = 120$, and $t = 150$. Each dot represents a learning agent, with a total of 100 dots in the simulation. In (b), the flow field reflects the evolution of the ensemble-level equation in which arrows indicate the direction of average change of action probability $\Delta\mathbf{p}$.

there is a non-zero probability of convergence to vertex **B** or **C**. To appreciate such occurrence, we looked at the behavior of an ensemble of such learning agents (Fig. 3(a)), where each dot represented a learning agent (animal organism). We assumed their starting positions (initial states) were randomly distributed within the probability simplex, and took "snapshots" of their positions as time evolved. While this population became redistributed, with the center of mass shifted towards vertex **A**, there were a few animals that got stuck at and eventually were absorbed by vertex **B**. It is also apparent that the population tended to cluster along the line **AB**, which corresponded to $p_3 = 0$; this is predicted by the dynamics of the ensemble-level evolution: action options drop out one by one, from the least favorable one to the next, and so on (see Fig. 1). For reader's information, we also plotted the trajectory of the ensemble-level dynamics (11) in the current case (Fig. 3(b)).

3. Discussion

Since the first generation of linear operator model (Bush & Mosteller, 1955), there have been tremendous advances in the formal characterization of adaptive computation (both Pavlovian and instrumental conditioning) in animal learning. Modern theories of reinforcement learning, e.g., Temporal Difference (TD) learning (Sutton, 1988; Sutton & Barto, 1990), emphasize the need of prediction and the use of the discrepancy between predicted and actual reward to drive learning. TD learning is a natural extension of the Rescola–Wagner rule for learning incentive values of a conditioned stimulus (Rescorla & Wagner, 1972) — in which the difference between (a) an actually delivered reward and (b) the internally generated expectancy of reward governs the adjustments of connection weights between stimulus-units and reward-predicting units (summarized in Barto (1995)). In the actor–critic architecture where there is now abundant neurophysiological support, TD learning is applied not only to the acquisition of incentive value of a stimulus ("V-function"), but also to the adjustment of action values ("Q-functions") needed for modifying a policy. Our current analysis shows that, when it comes to modifying action probability, we can, instead of implementing it as predictive learning, use the selectionist framework (*a la* Bush–Mosteller) and still achieve the same dynamics so long as the learning rate is small. In other words, action probability can be shaped not necessarily by comparing relative merits against each other but possibly through the differential magnitude of reinforcement achieved by each action when emitted. One may argue that predictive learning is crucial for solving sequential decision problems (e.g., through Watkins's (1989) Q-learning). It will be shown, in the companion paper to follow (Zhang, 2009), that applying the selectionist framework in the sequential setting naturally leads to the notion of conditioned reinforcement values for the intermediate states, and that action sequencing may be achieved so long as the incentive values of those states are learnt and used as reinforcers for preceding states and for non-terminal actions in a chain.

The linear operator model we analyzed here as exemplifying the selectionist-style operant learning is obviously overly simplistic.

Nevertheless, our simulated trajectories (Fig. 2) are in qualitative agreement with the learning trajectories of human subjects as reported by Busemeyer and Myung (1992) in their Exp. 1. There, subjects were given three alternatives to choose from, each resulting in a stochastic payoff with different means. The learning trajectories of the subjects were modeled by an adaptive network based on a reinforcement comparison rule. We have shown here by simulation, as well as by mathematical analysis, that successive application of operant reinforcement learning rule (3) will, with a small enough learning-rate parameter, result in the same effects as the reinforcement comparison rule (5) used by Busemeyer and Myung (1992). Since ensemble averaging is equivalent to time averaging at the limit $\epsilon \to 0$, the ensemble dynamics for operant learning may also describe the performance of a single learning agent. So our revelation on the connection between the selectionist and Bayesian frameworks, if anything, suggests that at behavioral level, it is rather difficult (theoretically impossible if the learning rate is infinitesimal) to distinguish the class of linear operator model from the class of reinforcement comparison rule that lie at the heart of modern reinforcement learning. On a philosophical level, selectionism, though widely recognized as an inefficient and wasteful algorithm, may well be Nature's way of faithfully and robustly implementing a rational scheme of adaptive computation when adjustment is incremental and each step of the iteration ("replication" in Hull et al.'s (2001) characterization for selection) is of a sufficiently small step size.

Acknowledgements

The author thanks Min Chang for contributing some crucial insights and for assistance in computer programming and graphics. This manuscript, and its sequel, were based on work first presented in the abstract form in Zhang and Chang (1996) to the 29th Annual Meeting of the Society for Mathematical Psychology, August 2–4, 1996, University of North Carolina at Chapel Hill.

References

Atkinson, R. C., Bower, G. H., & Crothers, E. J. (1965). *An introduction to mathematical learning theory*. New York: John Wiley and Sons.

Barnardo, A. F. M., & Smith, J. (1994). *Bayesian theory*. UK: John Wiley and Sons.

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.

Busemeyer, J. R., & Myung, I. J. (1992). An adaptive approach to human decision making: Learning theory, decision theory, and human performance. *Journal of Experimental Psychology: General*, *121*, 177–194.

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

Estes, W. (1950). Towards a statistical theory of learning. *Psychological Review*, *57*, 94–107.

Gardiner, C. W. (1985). *Handbook of stochastic methods for physics, chemistry and the natural sciences* (2nd ed.). Springer-Verlag.

Hofbauer, J., & Sigmund, K. (1998). *Evolutionary games and population dynamics*. Cambridge University Press.

Hull, D. L., Langman, R. E., & Glenn, S. S. (2001). A general account of selection: Biology, immunology, and behavior. *Behavioral and Brain Sciences*, *24*, 511–573.

Karlin, S., & Taylor, H. M. (1975). *A first course in stochastic processes*. San Diego: Academic Press.

Lakshmivarahan, S., & Thathachar, M. A. L. (1973). Absolutely expedient learning algorithms for stochastic automata. *IEEE Transactions on System, Man & Cybernetics*, *SMC-3*, 281–286.

Lakshmivarahan, S., & Thathachar, M. A. L. (1976). Absolutely expediency of Q- and S-model learning algorithms. *IEEE Transactions on System, Man & Cybernetics*, *SMC-6*, 222–226.

Montague, P. R., Dayan, P., Person, C., & Sejnowski, T. J. (1995). Bee foraging in uncertain environment using predictive hebbian learning. *Nature*, *377*, 725–728.

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. (1974). Learning automata—A survey. *IEEE Transaction on Systems, Man and Cybernetics*, *SMC-4*, 323–334.

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

Norman, M. F. (1968). Some convergence theorems for stochastic learning models with distance diminishing operators. *Journal of Mathematical Psychology*, *5*, 61–101.

Norman, M. F. (1972). *Markov process and learning models*. New York: Academic Press.

Phansalkar, V. V., & Thathachar, M. A. L. (1995). Local and global optimization algorithms for generalized learning automata. *Neural Computation*, *7*, 950–973.

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.

Skinner, B. F. (1953). *Science and human behavior*. Free Press.

Skinner, B. F. (1981). Selection by consequences. *Science*, *213*, 501–504.

Skinner, B. F. (1984). Selection by consequences. *Behavioral and Brain Sciences*, *7*, 477–510.

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

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Monograph*, *2*, No.8.

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

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*, 229–256.

Zhang, J. (2009). Adaptive learning via selectionism and Bayesianism. Part II: The sequential case. *Neural Networks*, *22*(3), 229–236.

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

Zhang, J., & Hasto, P. (2006). Statistical manifold as an affine space: A functional equation approach. *Journal of Mathematical Psychology*, *50*, 60–65.