Learning and animal behavior:
exploring the dynamics of simple models

Abran M. Steele-Feldman

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

University of Washington

2006

Program Authorized to Offer Degree:
Quantitative Ecology and Resource Management

Table of Contents

List of Figures

List of Tables

Introduction

All living organisms must interact with an external environment and should respond to it in a way that maximizes their probability of reproduction and survival. If an organism can learn, it will be able modify its behavior based on environmental feedback and potentially increase its survival probability. The processes underlying learning and behavior are of interest to researchers in cognitive science, psychology, artificial intelligence (AI), animal behavior, and ecology among others. However, the problems faced by a learning organism in a natural environment are daunting. Due to the complexity of the problems and processes involved, research on learning and behavior often focuses on highly simplified problems.

One way to simplify the learning problem is to treat it as a multi-armed bandit problem. In a multi-armed bandit problem an agent is presented with a set of possible actions and on discrete trials the agent must choose exactly one of these actions. Based on the action chosen the environment returns a reward to the agent, and the process repeats. The paradigmatic ecological example of this problem is a foraging bee. Consider a bee presented with a field of red and yellow flowers that provide nectar rewards. There are two possible actions for the bee, feed on a yellow flower or feed on a red flower, and the bee can only do one of these actions at a time. Each flower provides nectar rewards, and different types of flowers provide nectar rewards with different distributions. Based on the sequence of nectar rewards obtained, a learning bee should modify its behavior and visit the more rewarding flowers more frequently.

Of course the real problem facing the bee is enormously more complex than this caricature, involving complicated sensory association and identification issues. Moreover, in many situations behavioral choices do not take place on discrete trials, but instead must be made continuously in response to changes in the environment. Nonetheless, choice problems can often be abstractly formulated as multi-armed bandit problems, and many models for learning and choice behavior treat them as such (Bush and Mosteller, 1955; Sutton and Barto, 1998; Narendra and Thathachar, 1989; Keaser et al, 2002). Importantly, a common experimental procedure from psychology, the discrete

trials procedure, essentially presents a multi-armed bandit problem, and many mathematical learning models from psychology utilize the bandit framework.

Mathematical learning models were introduced into psychology in the 1950s (e.g. Bush and Mosteller, 1955) and their mathematical properties were examined intensely over the following decades (e.g. Luce, 1959 and Norman, 1972). Models of this sort continue to be proposed as explanations for animal behavior (Lea and Dow, 1984; Weber et al, 2004; Keaser et al, 2002) but in recent years less effort has been devoted to formal mathematical analysis of such models.

The models studied in psychology are closely related to learning automata (LA) (Narendra and Thathachar, 1989) models from artificial intelligence, and in fact many LA models derive from models in psychology. Despite the close similarity and common origins, research in the two disciplines has proceeded largely independently. The AI work on LA models has produced a mathematical formalism for organizing, analyzing, and discussing learning models. On the other hand, researchers focusing on animal behavior have produced a diverse array of models motivated by empirical data, but these models are often presented without a unifying framework to facilitate discussion. As a result, the relationships between models are often opaque, and it is not always apparent whether the mathematical results from one model will apply to others.

This thesis will explore the mathematical dynamics of some simple learning models in the context of several discrete trials foraging experiments. The vocabulary and formalism from LA is used throughout in order to organize the diversity of published models. Using the LA terminology, and some additional definitions, I will prove several analytic results for a special class of LA models, which I call driven estimator learning automata (DELA). DELA type models appear frequently in the literature and have some nice mathematical properties in the context of discrete trial experiments.

The purposes of the thesis are threefold. First, I hope to introduce a framework for thinking about learning models, taken largely from LA, which can be used to organize the models published in psychology and animal behavior. Second, I want to establish some formal analytic results about DELA models and the behavior they predict in some common discrete trials experiments. Finally, I want to numerically evaluate the

performance of some of these models with respect to published data from a sequence of discrete trials foraging experiments using starlings (*Sturnus vulgaris*).

Although a variety of discrete trials experiments are treated, many of the experiments discussed in later chapters focus on risk-sensitive foraging: the way in which animals respond to variability or uncertainty in the rewards provided by the environment. The basic risk sensitivity experiment (BRSE) presents an organism with a choice between two foraging options: a constant option that always provides the same reward and a variable option that provides either a big reward or a small reward stochastically but with the same mean value as the constant option. One of the main analytic results (presented in Chapter 5 and called the hot stove effect) is that, given some constraints on model form, DELA models predict uniformly risk averse behavior: agents using a DELA learning method will prefer the constant food source on the BRSE.

The thesis is structured as follows. The first chapter introduces the LA formalism and then expands this framework to introduce DELA models. Chapter 2 discusses some issues with applying DELA models to living organisms and presents a framework for doing so utilizing three elements: utility functions, estimators, and choice functions. Utilizing the terminology from the first two chapters, Chapter 3 explores model behavior on, and derives predictions for, a simple discrete trials experiment. Chapter 4 then reviews the main results from, and models for, risk sensitivity. Chapter 5 proves that a large class of DELA models will be uniformly risk averse on the BRSE. Chapter 6 explores a model that is capable of generating risk prone behavior and examines the predictions of this model for a set of honeybee foraging experiments. In Chapter 7, I use simulations to evaluate model performance with respect to a larger set of experiments using starlings. Finally, the Conclusion discusses the ramifications of these results and suggests some expansions of the DELA framework.

Chapter 1 Learning automata

A gambler is led to a room containing four different slot machines. Each machine pays out rewards with a different distribution, but these distributions are unknown to the gambler. The gambler is told that she has 1000 free plays to distribute amongst the slot machines as she sees fit. How should the gambler allocate her plays in order to maximize the rewards she receives?

Each morning a starling (*Sturnus vulgaris*) is placed in a box for a three hour experimental session. The box contains two keys and a food dispenser. Each key can be illuminated with either a red or a yellow light, and the bird has been previously trained that pecking at the keys can elicit a food reward from the dispenser. The experiment consists of 30 trials per day over the course of a week. On each trial, both keys light up simultaneously, one red and one yellow; if the bird presses a key while it is illuminated, both lights extinguish and the food dispenser delivers a reward after a short delay. Both keys can take on both colors over the course of the experiment, and the colors of the keys are associated with different reward distributions. This experimental set up is depicted in Figure 1.1 with a possible sequence of choice trials. How will the Starling allocate its choices over the course of a week?

Each of these scenarios represents a *reinforcement learning* problem: an agent learns about stimuli or actions only through the rewards associated with them, and must select future actions on the basis of the rewards previously obtained (Dayan and Abbott, 2001). The first scenario is a special type of reinforcement learning problem, known as the multi-armed bandit problem, so called because slot machines are affectionately known as one-armed bandits. The multi-armed, or *k*-armed, bandit problem has been explored extensively in the statistical and artificial intelligence (AI) literature (Berry and Fristedt, 1985; Sutton and Barto, 1998). If the reward distributions do not change over time, the optimal solution to this problem is to find the slot machine with the largest expected payoff and choose that machine exclusively. There are many ways to do this,

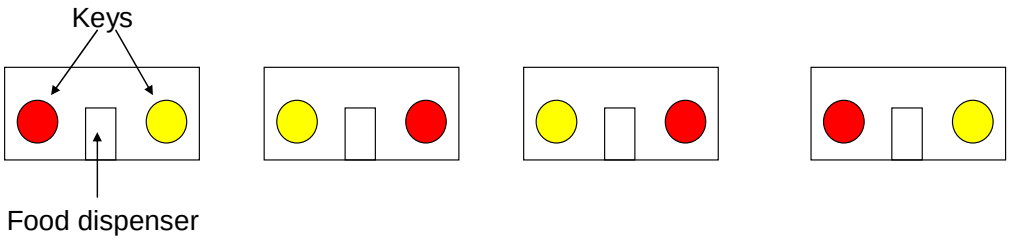


Figure 1.1 A sequence of choice trials from the bird in a box experiment. The box contains two keys and a food dispenser. On choice trials, both keys light up and food is provided when the Starling presses one of the keys. The size of the reward is associated with the color of the key. (After Bateson and Kacelnik, 1996).

however, and much work has focused on deriving solutions to the problem that are efficient or optimal in some way.

The second scenario depicts a discrete trial operant conditioning experiment; a common experimental paradigm in psychology. This 'bird in a box' experiment is largely equivalent to a 2 armed bandit problem: the two colored keys effectively represent different slot machines, and the bird can choose to 'play' one of these machines on each trial. Researchers studying animal behavior are mainly interested in determining or modeling *how* organisms solve such problems rather than looking for the *best* way to solve the problem.

Despite the difference in focus across disciplines, many models for animal behavior utilize the same basic framework as solution methods from AI. Any solution to the multi-armed bandit problem must specify a *decision mechanism* that chooses the next action based on the history of rewards obtained. Learning automata (LA) is a branch of AI that provides a simple framework for formulating decision mechanisms in the context of the multi-armed bandit problem. The LA models have strong connections to classical mathematical learning models from psychology. For example, the learning models of Bush and Mosteller (1958) are basic LA and many recently published models for animal behavior can also be formulated as LA (e.g. Keaser et. al., 2002; Shapiro, 2000; Weber et al, 2004).

The LA formalism seems a natural fit for the bird in the box experiment given the similarity between the bird in a box and the multi-armed bandit problem. However, while the bird in a box experiment is equivalent to a bandit problem in many important ways, the bird in a box is actually faced with a much more complex problem. Moreover,

the bird in a box experiment, while more realistic than the classic multi-armed bandit, is still highly artificial and contrived. In natural environments, the decision problems faced by organisms are dramatically more complex and often fundamentally different in important ways. I will return to these topics later in this chapter, but it will be important to keep in mind the ways in which the bird in a box problem is different from both the classic bandit problem and more natural behavioral choice problems.

Formalizing the problem

Our problem centers upon an agent that must interact with an external environment. During discrete trials the agent can engage in actions, and these actions precipitate a response from the environment in the form of a reward. The environmental response is stochastic, and a given action by an agent will not always result in the same response.

Within LA the problem is often formalized as follows (Narendra and Thathachar 1989). The agent has k available actions, a_i, and taken together these actions compose an action set, $A = \{a_1, a_2, ..., a_k\}$. Each action corresponds to playing on one of the different slot machines; for the bird in the box there are $k = 2$ possible actions: press the red key or press the yellow key. On each discrete trial the agent must choose one and only one of these actions; denote the agent's choice on the nth trial by $a(n) \in A$. Based on the action chosen, the agent receives some reward, $r(n) \in R$, where $r(n)$ is the reward received on the nth trial, and $R \subseteq \mathbb{R}_+$ is the set of possible rewards. Associated with each action, a_i, is a reward distribution, F_i^n, that determines the reward provided in response to the action, i.e.

$$F_i^n(r) = \Pr(r(n) \le r \mid a(n) = a_i, n). \tag{1.1}$$

In general, these distributions can change over time and thus are conditional on the current trial number, n. Taken together, the reward distributions comprise a family of distributions, $\aleph(n) = \{F_1^n, F_2^n, ..., F_k^n\}$.

The LA *environment* can then be summarized by the triple $\langle A, R, \aleph(n) \rangle$. Environments can be categorized based on their reward distributions, $\aleph(n)$. If the reward distributions change over time, the environment is called *transient*; if the distributions do not change over time, the environment is *stationary*, in which case $\aleph(n) = \aleph$, and $F_i^n(r) = F_i(r)$. Often environments are also categorized by the nature of the reward set R. Environments with binary reward sets, $R = \{0,1\}$, are called P-models, those with continuous reward sets on some interval, $R = [c, d]$, are called S-models, and environments where the reward set can take on a finite set of values, $R = \{r_1, r_2, ...r_q\}$, are Q-models. The discrete trial operant conditioning (bird in a box) experiments discussed in later chapters all utilize discrete reward sets, and the focus will thus be on Q-models.

The interaction between the agent and the environment is depicted in Figure 1.2. The agent outputs an action that serves as an input to the environment. Based on the action chosen the environment returns a reward from the appropriate distribution, this reward serves as input to the agent, and the process then repeats. The environment represents the problem that the agent must 'solve'. A solution for this problem, or equivalently a model for the agent's behavior, must specify a decision mechanism for generating the next action.

With this formalism, the bird in a box environment is described by an action set with $k = 2$ actions corresponding to pressing either the red or the yellow key; the reward set and reward distributions are controlled by the experimenter. A trial begins when both keys light up, and ends after the bird presses a key to receive a reward. All that remains is to describe the decision mechanism the bird is using.

Aside: From the bird's perspective

Of course, for a real bird in a box, the preceding formulation is enormously simplified. In order for this description to be useful, at the very least the bird must:

1) Learn that pressing a colored key results in a reward.

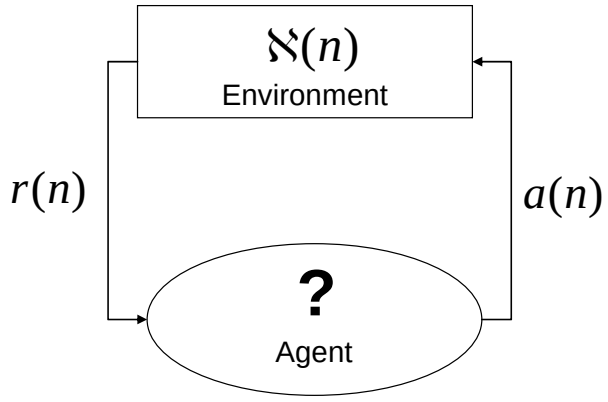


Figure 1.2 The flow of information in a general LA model. The agent chooses an action on each trial. In response the environment returns a reward selected stochastically from the appropriate distribution. The decision mechanism that the agent uses to generate the next action must be specified by the modeler.

2) Learn that the size of the reward received is associated with the color of the key and not, for example, with the location of the key in the box.
3) Treat the simultaneous presentation of two stimuli as a choice between two possible actions.

Presumably these associations were developed previously during training and can be ignored when trying to model behavior during the testing phase of the experiment. Nevertheless, the LA framework largely ignores stimuli and stimuli associations, but for living organisms these constitute essential elements in the problem.

The preceding formulation also neglects to characterize several additional aspects of the bird in a box experiment that might be expected to influence behavior. For example, no mention is made of the energetic state of the bird, nor of the energetic content of the food rewards. If energetic state impacts the decision mechanism that the bird actually uses, and there is evidence that it can (see chapter 4), then our model will necessarily be incomplete. Similarly, the LA environment is formulated in terms of discrete trials, but the bird in a box experiment takes place in continuous time. In transient natural environments, older information is less useful and may get discounted. Since the birds evolved in natural environments, one might expect that the amount of

time between trials could impact the bird's decision mechanism. The passage of time between trials can also impact energetic state by changing the rate of energy intake.

As with all models, modeling the bird in the box experiment as a multi-armed bandit problem abstracts substantially from the reality of the situation. The LA framework formalizes the problem at a very high level, and the hope is that more fundamental processes, such as stimulus association, can be ignored and taken for granted. Similarly, complexities that are unique to living organisms, such as energetic state and the effect of the passage of time, are ignored in the interest of simplicity and analytic tractability. By exploring model behavior with this simple framework and comparing model predictions with experimental results, we can hopefully diagnose deficiencies of these simple models and then search for appropriate embellishments and refinements. With these caveats in mind, let's now return to the project of designing decision mechanisms for the agents in the model.

Building agents

The agent in a LA model is represented by a state, $\mathbf{Q}(n)$, and a learning rule T. The action chosen by the agent, $a(n)$, depends only on the current state $\mathbf{Q}(n)$. After choosing an action and receiving a reward, the learning rule updates the agent's state:

$$\mathbf{Q}(n+1) = T\left(\mathbf{Q}(n), a(n), r(n)\right). \tag{1.2}$$

The next choice is generated from a probability vector, $\mathbf{P}(n) = \{P_1(n), P_2(n), ..., P_k(n)\}$, where $P_i(n) = \Pr(a(n) = a_i)$ denotes the probability of choosing action a_i on trial n. The probability vector is normalized such that

$$\sum_{i=1}^{k} P_i(n) = 1 \tag{1.3}$$

on all trials n. On each trial, the agent chooses an action stochastically with the probabilities given by the probability vector, and the probability vector is generally included as part of the description of the agent's state. Presumably the learning rule should increase the probability of taking an action after receiving a large reward from the action, but decrease the probability after receiving a relatively small reward.

Learning rules can be either *absorbing* or *ergodic*. An absorbing learning rule will eventually settle upon a single action and choose that action repeatedly forever, i.e. $\lim_{n\to\infty} P_i(n) = 1$ for some action a_i and $P_j(n) \to 0$ for all $j \neq i$. Conversely an ergodic learning rule will converge into an asymptotically stable distribution, $\boldsymbol{\pi} = \{\pi_1, \pi_2, ..., \pi_r\}$, where $\lim_{n\to\infty} E\left(P_i(n)\right) = \pi_i$ and $\pi_i > 0$ for more than one i. Note that the convergence of a given learning rule will depend on the environment in which it lives, and a given learning rule could be absorbing in one environment and ergodic in another.

In transient environments the reward probabilities can change, and an action that was once bad (low expected reward) can suddenly become good (high expected reward). So ergodic learning rules where $\pi_i > 0$ for all i are preferable in transient environments. Transience is the rule in natural environments, so ergodic learning rules might expected for living organisms. Indeed Maynard Smith (1984), suggested that ergodicity would be an essential characteristic of evolved learning rules. However, optimal learning rules for the *k*-armed bandit problem in a stationary environment are absorbing (Narendra and Thathachar, 1989). It seems likely then that organisms would perform sub-optimally when confronted with a stationary *k*-armed bandit problem.

Representing the agent's decision mechanism in terms of a state, $\mathbf{Q}(n)$, a learning rule T, and a probability vector $\mathbf{P}(n)$, there are still many possible types of decision mechanisms. Here I would like to discuss three general approaches for implementing decision mechanisms.

Classical learning automata

Classical learning automata (CLA), formulate the learning rule as a linear operation directly on the probability vector:

$$\mathbf{P}(n+1) = T(\mathbf{P}(n), a(n), r(n)). \tag{1.4}$$

For CLA, the agent's state is identical to the probability vector, $\mathbf{Q}(n) = \{\mathbf{P}(n)\}$, and the learning rule must maintain the normalization of the probability vector as expressed in Equation (1.3).

Within mathematical psychology these models are also known as linear models (Norman, 1972). One of the simplest, and earliest, models of this sort is the linear reward penalty (LRP) model (Narendra and Thathachar, 1989). Introduced by Bush and Mosteller (1955) as the alpha-model and also known as the fractional adjustment model (March, 1996), this model is most easily applied to environments with binary rewards (P-model environments) and two actions. Within such an environment, the learning rule can be implemented as a simple linear operation on the probabilities. If $a(n) = a_1$ the learning rule is

$$\begin{aligned} P_1(n+1) &= (1-m)r(n) + mP_1(n) \\ P_2(n+1) &= (1-m)(1-r(n)) + mP_2(n) \end{aligned}, \tag{1.5}$$

but if $a(n) = a_2$ the learning rule is

$$\begin{aligned} P_1(n+1) &= (1-m)(1-r(n)) + mP_1(n) \\ P_2(n+1) &= (1-m)r(n) + mP_2(n) \end{aligned}. \tag{1.6}$$

Here m is a parameter, the memory coefficient, that controls the rate of learning; smaller values of m correspond to faster learning.

CLA models have been widely used in psychology as models for animal learning, and their mathematical properties have been studied thoroughly (e.g. Norman, 1972; Part II of Bush and Estes, 1959). Models of this sort have proved somewhat effective in fitting experimental results from psychology, but they have also been criticized on theoretical and empirical grounds (e.g. Gallistel, 1990) and it has been difficult to fit data from diverse experiments while using a single memory coefficient (Killeen, 1985; Lea and Dow, 1985). From an AI perspective, convergence to the 'correct' solution (i.e. exclusive choice of the action with the highest expected reward) is slow (Vasilakos and Papadimitriou 1995). Although the probability of converging to an incorrect solution can be made arbitrarily small for CLA models operating in stationary environments, within AI other types of LA perform better and are proposed more frequently (Narendra and Thathachar, 1974).

Estimator learning automata

Estimator learning automata (ELA) were introduced into the LA literature by Thathachar and Sastry (1985). Their innovation was to introduce a vector of estimators, $\hat{\mathbf{r}}(n) = \{\hat{r}_1(n), \hat{r}_2(n), ..., \hat{r}_k(n)\}$, where $\hat{r}_i(n) \in \mathbb{R}^+$ represents the agent's estimate of the expected value of action a_i based on its experiences up to trial n. The agent's state is expanded to include these estimates, $\mathbf{Q}(n) = \{\mathbf{P}(n), \hat{\mathbf{r}}(n)\}$, and the learning rule for an ELA model is broken into two steps:

$$\begin{aligned} \hat{\mathbf{r}}(n+1) &= T_1\left(\hat{\mathbf{r}}(n), a(n), r(n)\right) \\ \mathbf{P}(n+1) &= T_2(\mathbf{P}(n), \hat{\mathbf{r}}(n+1)) \end{aligned}. \tag{1.7}$$

The ELA first updates the estimators $\hat{r}_i(t)$ based on the outcome of the trial and then updates the probability vector based on the updated estimators.

An ELA model must specify how to derive the estimates, usually by using some type of *averaging function*. For example, Vasilakos and Papadimitriou (1995) propose using a time window moving average (TWMA) to generate the estimators:

$$\hat{r}_i(n) = \frac{\left[\begin{array}{l}\text{The total reward received} \\ \text{from } a_i \text{ over the last } W \text{ times} \\ \text{it was selected}\end{array}\right]}{W} \tag{1.8}$$

where W is a parameter that defines the length of the time window. The next chapter introduces several other averaging functions.

In the LA literature, ELA models generally perform quite well, providing rapid convergence to 'good' choice probabilities (Vasilakos and Papadimitriou ,1995). However, with increased performance comes increased computational demands (both the probabilities and the estimates must be updated at each time step) and memory demands (both the vector of estimators and the vector of probabilities must be stored in memory).

Driven estimator leaning automata

Define driven ELA (DELA) models as a subset of possible ELA models. For a DELA model, the agent's state consists solely of its estimators, i.e. $\mathbf{Q}(n) = \{\hat{\mathbf{r}}(n)\}$, and the model dynamics are 'driven' by these estimators. Rather than maintain and update a probability vector, the vector is computed at each time step with a choice function, $\mathcal{C} : \mathbb{R}^k \to [0,1]^k$, as

$$\mathbf{P}(n) = \mathcal{C}(\hat{\mathbf{r}}(n)). \tag{1.9}$$

This choice function must generate a normalized probability vector over the action set, A, as in Equation (1.3) above. With this formalism, the dynamics of the learning process are determined through the evolution of the estimator vector and its interaction with the choice function. Since the probabilities are computed at each time step, the learning rule need only update the estimators:

$$\hat{\mathbf{r}}(n+1) = T\left(\hat{\mathbf{r}}(n), a(n), r(n)\right). \tag{1.10}$$

Models of this sort are not frequently proposed in the AI literature, but they are quite common in psychology and animal behavior. As defined above, DELA models are analogous to *v-scale* (Luce, 1959) and *additive models* (Norman, 1972) from mathematical psychology. For example Luce's (1959) simple beta-model can be represented as a DELA model with the learning rule

$$\hat{r}_i(n+1) = \begin{cases} \beta_{r(n)}\hat{r}_i(n) & \text{if } a(n) = a_i \\ \hat{r}_i(n) & \text{else} \end{cases} \tag{1.11}$$

and choice function

$$P_i(n) = \mathcal{C}\left(\hat{\mathbf{r}}(n)\right) = \frac{\hat{r}_i(n)}{\sum_{j=1}^{k} \hat{r}_j(n)}. \tag{1.12}$$

Here $\beta_{r(n)}$ is a parameter that depends on the value of the reward received.

DELA models appear frequently in the animal behavior literature. Published models that are effectively DELA models include those considered by Lea and Dow (1984); the honeybee foraging models in Keaser et al. (2002),Dayan and Abbott (2001),

and Shapiro et al (2001); Herrnstein type matching models as in Herrnstein (1961) and March (1996). Moreover, other models that are not always equivalent to DELA models can be treated as DELA models in certain experimental contexts (see Chapters 2 and 6 for details).

Discussion

The differences between the model types are summarized in Table 1.1. Note that CLA models are equivalent to DELA models where the choice function is the identity function and the estimator vector is the probability vector; thus CLA are a subset of possible DELA models. DELA models require less memory than ELA models but can be more computationally taxing due to the introduction of the choice function which often involves a costly normalization operation. ELA models and many CLA models often manage to avoid the normalization operation by beginning with a normalized probability vector and updating this vector using a learning rule that maintains normalization. This is only possible, however, due to an assumption fundamental to the LA framework: the action set A is the same on all trials.

Natural choice problems are generally characterized by a changing set of feasible actions, $\tilde{A}(t) \subseteq A$, that depend on the current state of the environment. Many actions will only be useful given some set of sensory stimuli; for example the action 'attack a prey item' is only meaningful if there is a prey item accessible that can be attacked. In such cases, the feasibility of certain actions can depend on the occurrence of events external to the agent, and the presence or absence of a given event can alter the composition of the feasible action set.

In such an environment, ELA and CLA type models will also have to renormalize the probability vector whenever the composition of the feasible action set changes. Thus in more natural environments, the computational advantage of updating the probabilities using a normalized learning rule largely disappears. Indeed if probabilities are continuously being re-normalized, the value of maintaining and updating a probability vector can be questioned. A dynamic feasible action set implies an estimator based model in a natural way (a similar argument was made by Luce, 1959).

Table 1.1 Different types of learning automata models. Classical LA represent the agent's state with the probability vector alone, and update this vector directly with the newest sample. In addition to the probability vector, estimator LA introduce a vector of estimates, and update this vector with the results from the newest sample. The updated estimator vector is then used to update the probability vector. Finally, driven estimator LA maintain only the vector of estimates, and then compute the probability vector using a choice function.

	State	Updating Rules	Generate Choice
Classical	$\{\mathbf{P}(n)\}$	$\mathbf{P}(n+1) = T(\mathbf{P}(n), a(n), r(n))$	$\mathbf{P}(n)$
Estimator	$\{\mathbf{P}(n), \hat{\mathbf{r}}(n)\}$	$\hat{\mathbf{r}}(n+1) = T_1\left(\hat{\mathbf{r}}(n), a(n), r(n)\right)$ $\mathbf{P}(n+1) = T_2(\mathbf{P}(n), \hat{\mathbf{r}}(n+1))$	$\mathbf{P}(n)$
Driven Estimator	$\{\hat{\mathbf{r}}(n)\}$	$\hat{\mathbf{r}}(n+1) = T\left(\hat{\mathbf{r}}(n), a(n), r(n)\right)$	$\mathbf{P}(n) = \mathcal{C}(\hat{\mathbf{r}}(n))$

A DELA model seems more appropriate in these natural environments. Since the probability of choosing any given action will depend on the composition of the set of available actions, it makes sense to track the values associated with the actions independently and then only compute probabilities when confronted with a particular action set. The choice is modeled with a choice function that computes probabilities and then chooses an action based on these normalized probabilities. Presumably a choice is in reality determined by some underlying dynamic process and is only normalized because one action eventually gets chosen. Thus we should think of the choice function as a mathematical abstract that is attempting to encapsulate the dynamics of some more complex underlying process.

The choice function abstraction simplifies the mathematical analysis of these models, and the following chapters explore the dynamics of DELA models in the context of some simple bird in a box type experiments. The diversity of published DELA models (see above) is substantial. The next chapter introduces a framework for analyzing different DELA models in terms of a utility function, estimators, and a choice function. This utility-estimator-choice structure will then be used throughout the rest of the thesis to guide the discussion and analysis of model behavior.

Chapter 2 The utility-estimator-choice framework

I claim that, in the face of the multi-armed bandit problem, every model for learning and decision making must specify at least three different pieces or elements. First, a utility function defines how the agent measures or values individual rewards. Second, the history of obtained rewards is used to compute an estimator of the expected reward from each action. Finally, a choice function selects the next action based on the values of the estimators. Of course the complexity of each element can vary widely between different models, and many models will augment this simple framework with additional elements. Nonetheless, these three elements are essential for analyzing learning models in the context of the multi-armed bandit problem. This chapter presents some definitions and mathematical characteristics for each element that will be useful for analyzing the behavior of DELA models.

Utility functions and choice functions are fairly straightforward and they are discussed in more detail below. However, there are many ways to compute estimators of the reward expected from an action; for example, in Equation (1.8) the estimators are updated recursively after each trial based on the rewards obtained. More abstractly, estimators can be defined in terms of an *averaging function* that maps from the full sequence of actions chosen and rewards obtained to an estimate of the reward expected from each action. We can think of the outcome of each trial as a sample composed of two values, $X(n) = \{a(n), r(n)\}$, and after taking n samples the agent has access to a sampling history, $\mathbf{H}(n) = \{X(1), X(2), ..., X(n-1)\}$. The sampling history contains all the information available to the agent, and the averaging function computes estimators from this history. Of course implementing a learning model in this way would be horribly inefficient, implying that the entire sampling history is maintained in memory. I do not propose here that living organisms have such memories. The averaging function is simply an abstract and useful construct for analyzing model behavior, and one that is hopefully intuitive to the statisticians in the audience. In the following I will often speak of averaging functions and estimators interchangeably and thus will speak of the utility-

averaging-choice framework in addition to the utility-estimator-choice framework. Nonetheless, it will be important to keep in mind that there are more efficient ways of deriving estimators.

In total, then, three functions are needed to implement a DELA model: a utility function, an averaging function, and a choice function. The utility function defines how individual rewards are measured, the averaging function combines these individual rewards into estimators of the reward expected from each action, and the choice function selects the next action based on these estimators. The flow of information in a DELA model is shown in Figure 2.1 for an environment with two possible actions. The utility-averaging-choice framework will be used throughout the rest of this thesis, and hopefully provides an intuitive way to think about decision mechanisms (indeed, while this thesis was in preparation, Yechiam and Busemeyer (2005) proposed an almost identical framework for analyzing learning models). Each element of the framework will impact the resulting model behavior, and all the elements need to be examined together when analyzing model behavior.

The utility function- measuring individual rewards

It is important to differentiate between the *physical rewards* as measured by the experimenter, $r'(n)$, and the *subjective rewards* as measured by the organism, $r(n)$. The experimenter presumably measures the rewards using physical units that are related to the rate of energy intake provided. These are the physical rewards. The organism must measure the value of the physical rewards, but it may be using different units than the experimenter or attending to different aspects of the rewards. These are the subjective rewards. A model must specify a utility function, $\mathcal{U}(.)$, which determines the relationship between the physical rewards and the subjective rewards:

$$r(n) = \mathcal{U}\left(r'(n)\right). \tag{2.1}$$

The utility function determines how organisms measure individual rewards, and non-linear utility function can generate interesting behavior. The notion of a utility function

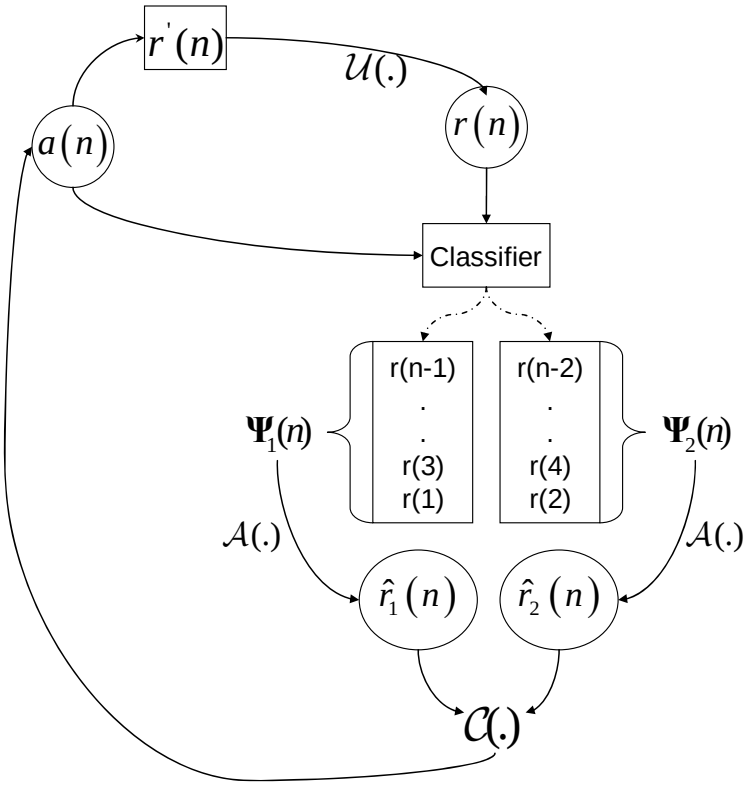


Figure 2.1 Flow of information in a DELA model. The depicted environment has two actions. The action chosen on trial n, $a(n)$, results in a physical reward from the environment, $r'(n)$. The utility function, $\mathcal{U}(.)$, converts the physical reward into the subjective reward $r(n)$. Based on the action selected, a classifier appends this newly measured reward to the appropriate memory vector, Ψ_i. The averaging function, $\mathcal{A}(.)$, then converts each memory vector into an estimator, $\hat{r}_i$. Finally, the choice function, $\mathcal{C}(.)$, selects an action based on the estimators and the process repeats.

is familiar in economics where it is the basis of expected utility theory (Von Neumann and Morgenstern, 1947). For a living organism, the utility function is expected to be an increasing function of the rate of energy intake provided by the physical reward, and will be assumed to be non-negative, i.e. $\mathcal{U} :\rightarrow \mathbb{R}_+$.

In general, physical rewards can be described by multiple characteristics. For example, the experiments discussed in the next chapter characterize rewards by both a waiting time (the delay to food delivery) and a reward size (number of food items provided). When rewards have multiple characteristics each reward can be represented as a vector with the appropriate number of dimensions, but these multivariate physical rewards can greatly complicate model dynamics. For simplicity the models discussed

hereafter will assume in such cases that the organism uses a univariate subjective reward, $r(n) \in \mathbb{R}_+$, and thus that the utility function converts a vector-valued reward into a single positive real value. Some authors have suggested that organisms do not use absolute utilities, but instead use relative utilities that depend on the organism's energetic state (Marsh et al 2004, Pompilio and Kacelnik 2005). The following discussion is confined only to constant utilities such that $\mathcal{U}\left(r'(n)\right)$ depends only on the reward provided and not on the trial number or the internal state of the agent.

The averaging function- combining multiple rewards

The utility function specifies how the organism measures individual reward values, but the averaging function combines these samples into estimators of the reward expected from each action. The averaging function converts the sampling history into a vector of estimators,

$$\hat{\mathbf{r}}(n) = \mathcal{A}(\mathbf{H}(n)), \tag{2.2}$$

where $\mathcal{A} : (R \times A)^n \to \mathbb{R}_+^k$ is the averaging function. Intuitively, the expected rewards from an action should only depend on the samples obtained from that action. That is, if $\mathbf{H}_i(n) \subset \mathbf{H}(n)$ denotes the subset of the sampling history obtained from action a_i, we might expect that $\hat{r}_i(n) = \mathcal{A}(\mathbf{H}_i(n))$, and thus that the associated estimator, $\hat{r}_i(n)$, does not depend on the samples obtained from other actions. Averaging functions of this sort will be called *separable* averaging functions (Yechiam and Busemeyer, 2005 refer to these as interference models). Separable averaging functions have simpler dynamics than non-separable averaging functions, and the following chapters only consider DELA models with separable averaging functions. Many published learning models are not separable in this sense; for example most CLA models, such as the linear reward penalty model of Bush and Mosteller (1958), are not separable. Thus separability is an important characteristic distinguishing CLA from DELA models as discussed hereafter.

Averaging functions can also be distinguished by the length of the memory implied. Define a finite memory averaging function as one that, given enough samples from the environment, only depends on a subset of the full sampling history. That is, a separable averaging function has finite memory if it can be expressed as

$$\hat{r}_i(n) = \mathcal{A}(\mathbf{\Psi}_i(n)) \tag{2.3}$$

for some $\mathbf{\Psi}_i(n) \subset \mathbf{H}_i(n)$, given that n is greater than some critical n^*. I will refer to $\mathbf{\Psi}_i(n)$ as the *memory vector* for action a_i. If each memory vector has the same maximum length, W, call it a length-W memory. In other words, with a separable length-W memory averaging function, the estimator $\hat{r}_i(n)$ depends only on the last W rewards obtained from action a_i.

Different types of averaging functions make sense in different types of environments. In stationary environments where the reward distributions do not change, non-finite memories make sense because each sample contributes information about the current reward distribution. However, in transient environments, older samples provide less information about current reward distributions, and the organism must be able to adapt to changes in the distributions. Finite memory averaging functions are better suited for these transient environments.

While laboratory experiments often present organisms with stationary environments, natural environments are likely to be transient. Thus natural decision mechanisms are probably adapted for transient environments, and we might expect organisms to have finite memory averaging functions, or at least averaging functions that weight more recent samples more highly. The following four averaging functions will be mentioned frequently in following chapters. All are separable, but the memory lengths vary.

Example: Long term average (LTA):

The long term average,

$$\hat{r}_i(n) = \frac{\text{sum of all rewards from action } a_i}{\text{total \# of times action } a_i \text{ was chosen}}, \tag{2.4}$$

does not have finite memory. Many published models assume a LTA function by default.

Example: Time window moving average (TWMA):

A TWMA, with memory length W,

$$\hat{r}_i(n) = \frac{\begin{array}{c}\text{sum of the last } W \text{ rewards} \\ \text{obtained from action } a_i\end{array}}{W}, \tag{2.5}$$

is a length-W memory averaging function. This type of averaging function has been used in the learning automata literature (Vasilakos and Papadimitriou 1995).

Example: Exponentially weighted moving average (EWMA):

A EWMA averaging function is defined by recursively updating the previous value of the estimator:

$$\hat{r}_i(n+1) = \begin{cases} (1-m)r(n) + m\hat{r}_i(n) & \text{if } a(n) = a_i \\ \hat{r}_i(\text{n}) & \text{else} \end{cases}. \tag{2.6}$$

Here m is a parameter, the memory coefficient, that controls the rate at which the estimate changes; with smaller values of m estimates change more quickly and newer samples are weighted more strongly. Thus smaller values of m correspond to shorter memory lengths.

EWMA based models are extremely common as models for learning. Within reinforcement learning, these models are also known as exponential recency-weighted averages (Sutton and Barto, 1998). Note that Equation (2.6) above is quite similar to the linear reward penalty updating rule from the previous chapter (Equation(1.5)), the difference being that the LRP rule updates the probabilities directly, whereas here the estimators are updated. Similarly, the Rescorla-Wagner learning rule uses a EWMA updating rule, but is thought of as updating the association strength between an action and a stimulus (see Couvillon and Bitterman, 1991). EWMA models are pervasive in psychology, so much so that Lea and Dow (1984) invoke the EWMA updating procedure as the distinguishing characteristic of the “common model” for animal learning.

Although EWMA averaging functions are not actually finite, in certain contexts they can be approximated to arbitrary precision by length-W memory averaging functions. For example, assume that there exists some largest reward, $r_{\max} < \infty$, then the EWMA averaging function can be approximated by a length-W averaging function with

$$W = \frac{\log\left(\frac{\varepsilon}{r_{\max}}\right)}{\log(m)} \tag{2.7}$$

where ε denotes the desired degree of accuracy and can be made as small as necessary. So in many contexts EWMA functions can be treated as if they have finite memory (see Chapter 6 for a more formal discussion).

Example: Generalized linear combination (GLC)

Let $m_j \in [0,1]$ $(j = 1,2,...W)$ be weights with $\sum_{j=1}^{W} m_j = 1$, and let $\psi_{i,j}$ be the jth value in the memory vector $\boldsymbol{\Psi}_i = \{\psi_{i,1}, \psi_{i,2}, ..., \psi_{i,W}\}$. Then the GLC averaging function is given by

$$\hat{r}_i(n) = \sum_{j=1}^{W} m_j \psi_{i,j} \,. \tag{2.8}$$

The weights define the GLC, and each of the previous finite memory averaging functions is a special case of the GLC averaging function. For example, the TWMA uses $m_j = \frac{1}{W}$.

The choice function- making a decision

The choice function defines how the next action is selected based on the current value of the estimators. Recall that the choice is modeled with a probability vector, $\mathbf{P}(n) = \{P_1(n), P_2(n), ...P_k(n)\}$, where $P_i(n) = \Pr(a(n) = a_i) \in [0,1]$ represents the probability of choosing action a_i on the nth trial. The probability vector must be normalized, and is generated by the choice function, $\mathcal{C} : \mathbb{R}_+^k \to [0,1]^k$, as

$$\mathbf{P}(n) = \mathcal{C}(\hat{\mathbf{r}}(n)). \tag{2.9}$$

The choice function determines the magnitude of preferences developed in stationary environments. Strong preferences will develop from choice functions that produce a high probability of choosing the action with the currently largest estimate.

For example, the *greedy choice function* (Sutton and Barto, 1998) always chooses the action with the currently largest estimate and is a step function:

$$P_i(n) = \mathcal{C}(\hat{\mathbf{r}}(n)) = \begin{cases} 1 & \text{if } \hat{r}_i = \max\limits_{j=1,\ldots,k} \left(\hat{r}_j\right) \\ 0 & \text{else} \end{cases}. \tag{2.10}$$

At the other extreme is the *indifferent choice function* that attributes equal weight to each action, $P_i(n) = \frac{1}{k}$, independent of the value of the estimators. More reasonable choice functions will lead to intermediate levels of preference. I will often refer to the *choosiness* of a choice function: more choosy (choosier) choice functions lead to stronger preferences. The greedy choice function is the choosiest, and the indifferent choice function is the least choosy.

As with the averaging functions, different environments call for different choice functions. In stationary environments, the optimal behavior is to pick the action with the largest expected reward and choose it exclusively, suggesting some version of the greedy choice function. In transient environments, however, reward distributions can change such that a formerly poor distribution begins to provide the largest expected reward. In order to detect and exploit improvements in the reward distributions, choice functions in transient environments should always maintain at least a small probability of choosing each action, even actions with currently low expected rewards. One might expect that choice functions that evolved in transient environments would be less choosy than those from stationary environments.

The following chapters are mainly concerned with problems that present $k = 2$ possible actions. When there are only two possible actions, the probability vector can be summarized by a single value, $\phi(n) = P_1(n)$, and the choice function expressed as

$$\phi(n) = \mathcal{C}\left(\hat{r}_1(n), \hat{r}_2(n)\right). \tag{2.11}$$

Note that $P_2(n) = 1 - \phi(n)$, and here, $\mathcal{C} : \mathbb{R}_+^2 \to [0,1]$. In the following chapters, I will say that a choice function is *proper* if it meets the following four conditions:

- C.0: $0 \le \mathcal{C}(x, y) \le 1$ for all $x, y \in \mathbb{R}_+$
- C.1: $\mathcal{C}(x, y) = \frac{1}{2}$ if $x = y$
- C.2: $\mathcal{C}(x, y) = 1 - \mathcal{C}(y, x)$
- C.3: $\mathcal{C}(x_1, y) \ge \mathcal{C}(x_2, y)$ if $x_1 > x_2$, and $\mathcal{C}(x, y_1) \le \mathcal{C}(x, y_2)$ if $y_1 > y_2$.

Condition C.3 says that $\mathcal{C}$ is a monotonically increasing function in its first argument; if the inequality is strict then $\mathcal{C}$ is strictly increasing on x. These conditions follow mainly from characteristics of probabilities and hopefully correspond to the intuitive notion of a choice function.

It will be helpful to define an additional characteristic of choice functions in analogue to the traditional notion of convexity. Recall that a function, $f(x)$, is convex on an interval $[j,k]$ if and only if for every $x_1, x_2 \in [j,k]$

$$\frac{f(x_1) + f(x_2)}{2} \ge f\left(\frac{x_1 + x_2}{2}\right). \tag{2.12}$$

Analogously, I will say that a choice function, $\mathcal{C}(x, y)$, is *reflectively convex* (R-convex) on an interval $[j,k]$ if, for every $x_1, x_2 \in [j,k]$ and $\bar{x} = \frac{x_1 + x_2}{2}$,

$$\frac{\mathcal{C}(x_1, \bar{x}) + \mathcal{C}(x_2, \bar{x})}{2} \ge \mathcal{C}\left(\frac{x_1 + x_2}{2}, \bar{x}\right), \tag{2.13}$$

or equivalently, using condition C.1,

$$\frac{\mathcal{C}(x_1, \bar{x}) + \mathcal{C}(x_2, \bar{x})}{2} \ge \frac{1}{2}. \tag{2.14}$$

Similarly, say that a choice function is *reflectively concave* (R-concave) if

$$\frac{\mathcal{C}(x_1, \bar{x}) + \mathcal{C}(x_2, \bar{x})}{2} \le \frac{1}{2}. \tag{2.15}$$

In the special case when equality holds on the interval (i.e. both R-concave and R-convex) say that the function is *reflectively affine* (R-affine).

The R-convexity of a function will have important ramifications in future chapters. All of the functions we will be considering are R-concave. By rearranging Equation (2.15), we see that an R-concave function has the special property where

$$\mathcal{C}(x_1, \overline{x}) \leq 1 - \mathcal{C}(x_2, \overline{x}), \tag{2.16}$$

or, using condition C.2,

$$\mathcal{C}(x_1, \overline{x}) \leq \mathcal{C}(\overline{x}, x_2). \tag{2.17}$$

For a strictly R-concave function,

$$\mathcal{C}(x_1, \overline{x}) < \mathcal{C}(\overline{x}, x_2), \tag{2.18}$$

and for an R-affine function

$$\mathcal{C}(x_1, \overline{x}) = \mathcal{C}(\overline{x}, x_2). \tag{2.19}$$

These properties will be used to derive the results in Chapters 3 and 5.

Difference-based vs. ratio-based

There is some debate in psychology about whether organisms choose between options by computing ratios or differences. I will say that a choice function is *ratio-based*, or *odds-based*, if there exists some function $\mathcal{C}^* : \mathbb{R}_+ \to (0,1)$ such that

$$\mathcal{C}(x, y) = \mathcal{C}^*\left(z\right) \tag{2.20}$$

for $z = \frac{x}{y}$. Similarly I will say that a choice function is *difference-based* if there exists some function $\mathcal{C}^{**} : \mathbb{R} \to (0,1)$ such that

$$\mathcal{C}(x, y) = \mathcal{C}^{**}\left(z\right) \tag{2.21}$$

for $z = x - y$. Ratio-based choice functions are consistent with some fundamental psychological results such as Weber's law, and are proposed more frequently (Fantino and Goldshmidt, 2000). There is still an ongoing debate, however, and some experimental results are better explained by difference-based choice functions (Fantino and Goldshmidt, 2000; Savastano and Fantino, 1996), while other results match the predictions of ratio-based choice functions (Gibbon and Fairhurst, 1994; Mazur, 2002).

Analogs of Conditions C.1-C.3 can be formulated for both types of choice functions (Condition C.0 is trivial). For a ratio-based choice function, the three conditions become

- C.1*: $\mathcal{C}^*(1) = \frac{1}{2}$
- C.2*: $\mathcal{C}^*(z) = 1 - \mathcal{C}^*\left(\frac{1}{z}\right)$
- C.3*: $\mathcal{C}^*(z_1) \geq \mathcal{C}^*(z_2)$ if $z_1 > z_2$.

Importantly, all strictly increasing ratio-based choice functions are also strictly R-concave. To see this, note that

$$\frac{x-i}{x} < \frac{x}{x+i} \tag{2.22}$$

for all $0 < i \leq x$ and $x \in \mathbb{R}_+$. Thus if $\mathcal{C}^*$ is strictly increasing,

$$\mathcal{C}(x-i, x) = \mathcal{C}^*\left(\frac{x-i}{x}\right) < \mathcal{C}^*\left(\frac{x}{x+i}\right) = \mathcal{C}(x, x+i). \tag{2.23}$$

Since $\frac{x-i+x+i}{2} = x$, this is equivalent to strict R-concavity by Equation (2.17).

For a difference-based choice function, the three analogous conditions are

- C.1**: $\mathcal{C}^{**}(0) = \frac{1}{2}$
- C.2**: $\mathcal{C}^{**}(z) = 1 - \mathcal{C}^{**}(-z)$
- C.3**: $\mathcal{C}^{**}(z_1) \geq \mathcal{C}^{**}(z_2)$ if $z_1 > z_2$.

Difference-based choice functions are all R-affine since

$$\mathcal{C}(x-i, x) = \mathcal{C}^{**}((x-i) - x) = \mathcal{C}^{**}(-i) = \mathcal{C}^{**}(x - (x+i)) = \mathcal{C}(x, x+i). \tag{2.24}$$

Example: Matching choice function

Perhaps the most straightforward choice function, the matching choice function is

$$\mathcal{C}(x, y) = \frac{x}{x+y}. \tag{2.25}$$

This choice function is strongly associated with the matching law (Herrnstein, 1961), and is often assumed by default (e.g. Keaser et al, 2002). Clearly the matching choice

function is proper (i.e. it meets conditions C.0-C.3); the matching choice function is also ratio-based, with

$$\mathcal{C}^*(z) = \frac{1}{1+z^{-1}} \tag{2.26}$$

for $z = \frac{x}{y}$.

Example: Boltzmann choice function

Often used in reinforcement learning, the Boltzmann choice function (Sutton and Barto, 1998) has its origins in statistical mechanics:

$$\mathcal{C}(x,y) = \frac{e^{\frac{x}{\gamma}}}{e^{\frac{x}{\gamma}} + e^{\frac{y}{\gamma}}}. \tag{2.27}$$

The parameter γ controls the magnitude of the preferences that develop, and is analogous to the temperature of a physical system. As $\gamma \to \infty$, each action will be equally likely to be chosen, while as $\gamma \to 0$, the action with the largest estimate will be chosen with probability approaching one. Smaller values of γ lead to choosier choice functions (Figure 2.2).

The Boltzmann choice function is proper and is difference-based:

$$\mathcal{C}^{**}(x,y) = \frac{1}{1+e^{\frac{z}{\gamma}}} \tag{2.28}$$

for $z = y - x$.

Example: Modified Boltzmann choice function

Equation (2.27) can be modified as

$$\mathcal{C}(x,y) = \frac{e^{\frac{1}{\gamma}\frac{x}{x+y}}}{e^{\frac{1}{\gamma}\frac{x}{x+y}} + e^{\frac{1}{\gamma}\frac{y}{x+y}}} \tag{2.29}$$

This modification makes the parameter γ scale independent, and smaller γ values lead to choosier choice functions (Figure 2.2). The modified Boltzmann is proper, but, unlike the traditional Boltzmann function, it is ratio-based:

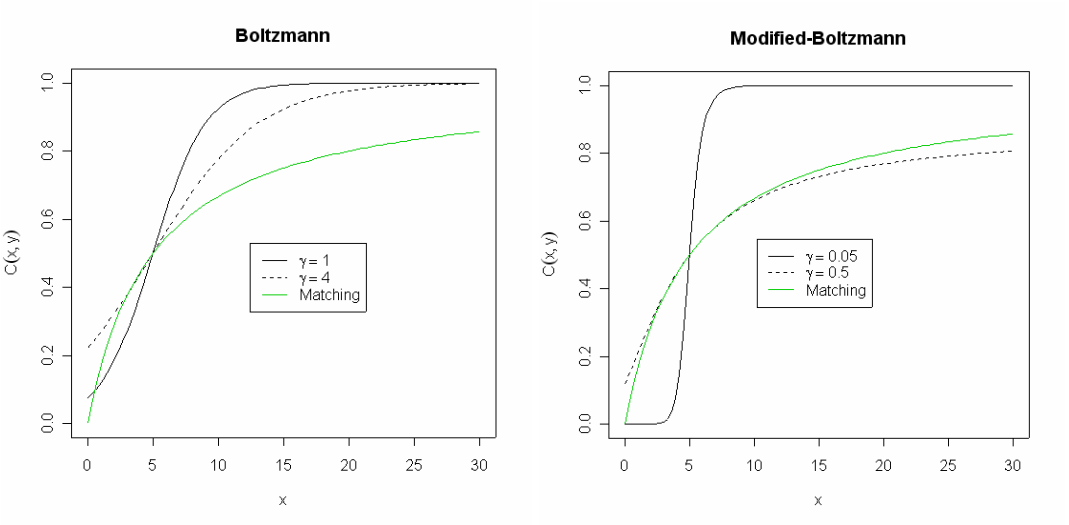


Figure 2.2. Graphs of the Boltzmann and modified Boltzmann choice functions. Shows the probability of choosing action x as a function of its estimate and for $y = 5$. For comparison, the matching choice function is also depicted in green. Note that smaller values of γ are associated with steeper (choosier) choice functions.

$$C^*(z) = \frac{1}{1 + e^{\frac{g(z)}{\gamma}}} \quad g(z) = \frac{1}{1+z} - \frac{1}{1+z^{-1}} \tag{2.30}$$

for $z = \frac{x}{y}$.

Example: Gaussian choice function

Not all decision mechanisms can be conveniently expressed in terms of a probability vector. The Gaussian choice function generates k normally distributed random variables, $Y_i \sim N(\hat{r}_i, \gamma^2 \hat{r}_i^2)$, each time a choice must be made. The mean of each distribution is given by the current value of the associated estimator, and the parameter γ determines the coefficient of variation of these distributions. On each choice trial, the agent chooses the action associated with the largest Y_i. When there are only two options, the Gaussian choice function simplifies to

$$C(x, y) = \Phi\left(\frac{x - y}{\gamma\sqrt{x^2 + y^2}}\right). \tag{2.31}$$

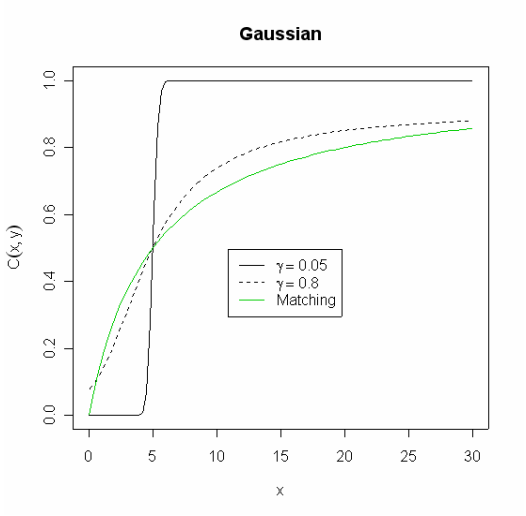


Figure 2.3 The Gaussian choice function. All symbols as in Figure 2.2. Choosiness increases as γ decreases.

where $\Phi(.)$ represents the cumulative distribution function for a standard Gaussian random variable. As with the other choice functions, smaller values of γ are associated with stronger preferences and increased choosiness (Figure 2.3), and $\gamma = 0$ gives the greedy choice function. Gaussian choice functions have been proposed in connection with considerations of Weber's law (Bateson and Kacelnik, 1995). This choice function is ratio-based, with the equivalent representation

$$\mathcal{C}^*(z) = \Phi\left(\frac{1}{\gamma}\left(\frac{1}{\sqrt{1+z^{-2}}} - \frac{1}{\sqrt{1+z^2}}\right)\right) \tag{2.32}$$

for $z = \frac{x}{y}$.

Discussion

Figure 2.1 shows the flow of information in a DELA model confronted by an environment with two actions. Individual samples are measured using the utility function and the averaging function converts these samples into estimates of the expected value for each action. The agent's state is represented by the vector of estimates $\hat{\mathbf{r}}(n)$, but if the averaging function is separable and has length-W, the agent's state can equivalently be represented by the memory vectors $\mathbf{\Psi}_i(n)$, with $\hat{r}_i(n) = \mathcal{A}(\mathbf{\Psi}_i(n))$. Note that the

classifier unit is needed to assign each measurement to the appropriate memory set, effectively performing the job of stimulus association.

The next chapter applies the framework to a simple bird in the box experiment conducted by Bateson and Kacelnik (1995). The experimental design allows for an analytic derivation of predictions for several simple models. Interestingly, the experimental results are numerically inconsistent with the predictions of many simple decision models, including the model proposed by the authors.

Chapter 3 A simple experiment

When organisms choose between foraging options providing qualitatively different types of rewards, for example apples and oranges, it is not immediately clear how they should or do value these different reward types. Much experimental work within psychology and economics has explored how different types of rewards are valued. One way to explore how organisms attribute value is to offer a repeated choice between two rewards and then slowly modify one of the rewards until the organism is indifferent between the two choices. For example, we could start by providing a choice between a single apple and five oranges and then slowly increase the number of apples provided until we reach an *indifference point*: the point at which both options are chosen with equal probability.

Within psychology, this experimental approach is known as the *titration procedure* and it has been used in a variety of experimental contexts, but most frequently it is used to evaluate how utility functions respond to different reward characteristics (Lea, 1976; Mazur 1984, 1986a, 2000, 2005; Cardinal et. al. 2002). The titration procedure presents a choice between a *standard* option, for which the reward schedule is held constant, and an *adjusting* option for which the reward schedule is slowly adjusted over the course of the experiment based on the subject's choices. For example, one of the experiments conducted by Mazur (1984) presented pigeons with a choice between a standard option delivering a small amount of food after a 10 second delay and an adjusting option delivering a large amount of food after an adjusting delay. The goal of the experimenter was to determine the delay that would cause the birds to be indifferent between the two options.

Although such experiments usually use disparate reward types, effectively comparing apples and oranges, several experiments have also used the procedure when the rewards are otherwise identical, thus comparing apples with apples (Lea, 1976; Mazur 1984, 1986a, 1986b; Mazur et al. 1985; Bateson and Kacelnik, 1995). For example, in one of the experiment conducted by Bateson and Kacelnik (1995) the constant option

always provided three food pellets while the adjusting option provided an adjusting number of food pellets. Presumably the indifference point should be obtained when the adjusting option provides three food pellets, but surprisingly the indifference point was obtained with a reward size of more than three pellets. Indeed most of the titration experiments offering equivalent reward types displayed a *titration bias*: the value of the adjusting option at the indifference point was significantly larger than the corresponding value of the standard (Mazur 1986a and Lea 1976 are exceptions). The titration bias is puzzling and currently remains unexplained.

In this chapter I explore the Bateson and Kacelnik (1995) experiments in detail and show how the observed titration bias will emerge if the organisms are using a ratio-based choice function. This fact potentially has substantial ramifications for the interpretation of titration results. I also show that the model proposed by Bateson and Kacelnik (BK) is numerically inconsistent with the results of their experiment. Specifically, their model cannot fit the results from the first stage of the experiment, wherein reward distributions were held stationary, and simultaneously fit the results from the second stage of the experiment wherein the reward distributions were transient. One explanation for this inconsistency, that the starlings respond differently in stationary and transient environments, is discussed briefly; this hypothesis is discussed further in Chapter 7.

Experimental procedure

The experiment utilized a discrete trials procedure and presented each bird with a choice between two colored keys representing different feeding options. Trials were organized into block of ten, and there were two types of trials: forced trials and choice trials. Forced trials forced the bird to experience each of the options, while choice trials asked it to choose between them (i.e. express their preferences). On forced trials, only one of the keys was illuminated, and if the bird pressed the key while it was illuminated, a reward was provided. On choice trials both keys were illuminated simultaneously, but only the first key pressed delivered a reward. Each block consisted of 8 forced trials

followed by 2 choice trials. The key color in the forced trials was chosen quasi-randomly so that each key was illuminated on four of the forced trials in each block but the order of presentation was random; this procedure ensured that the birds received equal exposure to each option during the forced trials. All trials were separated by an inter-trial interval, and the length of this interval was varied between birds, acting as an additional treatment, ranging from 9.3 seconds to 132 seconds.

There were two stages to the experiment. In the first stage, one of the key colors delivered a reward of 3 food pellets, while the other delivered a reward of 9 food pellets. The environment was held stationary during the first stage of the experiment and reward sizes remained unchanged throughout. Eventually all of the birds chose the better key, i.e. the key delivering 9 pellets, more frequently. A bird completed stage one when it selected the better key in 70% or more of the choice trials on each of two consecutive days. There were 12 blocks, or 24 choice trials, per day, so stage one ended for a bird when it chose the better key 17 or more times each day for two consecutive days.

There were two main treatments in the second stage of the experiment. For half the birds, the standard key always provided a 9 pellet reward and for the other half it always provided a 3 pellet reward. Under both treatments the adjusting reward changed between blocks according to a titration rule, and thus the environment was not stationary. If a bird chose the adjusting key on both of the choice trials in a block, the adjusting reward was decreased by one, whereas if the bird selected the standard key twice, the adjusting reward was increased by one. If the bird chose each key once, the adjusting reward remained unchanged. Negative reward values were not allowed and rounded up to zero. The titration procedure was designed to eventually oscillate around a single value, presumably the value of the standard. Different birds experienced different numbers of blocks during stage two (not all birds completed 12 blocks per day), but all the birds except one completed at least 270 blocks over the course of stage 2. The data from this lone laggard bird were discarded, so the results for stage two include data from eleven birds in all.

Table 3.1 Results from stage 2 of the experiment. The Bias is the difference between the mean of the adjusting option and the value of the standard reward. The percent bias is the bias divided by the size of the standard reward. Note that the mean of the adjusting option is larger than the value of the standard (positive bias) and that the percent bias is similar in both treatments. Note also that the CV is slightly higher in the 3 pellet treatment than in the 9 pellet treatment.

Treatment	Standard Reward	Adjusting Mean	Bias	Percent Bias	Adjusting CV
Small	3 pellets	3.97	0.97	32%	.583
Large	9 pellets	12.10	3.10	34%	.494

Experimental results

All twelve birds completed the first stage of the experiment. The birds took 4.25 days ($SD = \pm 3.25$) on average to complete the first stage. The number of days required to meet the stopping criterion (two consecutive days with 70% or more choices of the larger key) was the only data reported from stage one.

BK report a variety of results from stage two, but the main results for our purposes are presented in Table 3.1 and Figure 3.1. Table 3.1 shows the mean and coefficient of variation (CV) of the adjusting option during stage two of the experiment. Note that both treatments show a titration bias: the mean value of the adjusting option is approximately 33% larger than the value of the standard. The CV was similar between treatments, but was slightly larger for the 3 pellet treatment than it was in the 9 pellet treatment.

Figure 3.1, taken directly from BK (1995), shows the probability of making correct choices on both of the choice trials in a block; I will refer to this probability as the *potential*. A choice is correct if the key with the currently larger reward value is chosen; when the value of the adjusting key is less than the value of the standard key, it is correct to choose the standard key, but it is correct to choose the adjusting key when its associated value is larger than the standard. The top row of Figure 1 shows the values predicted by BK's model (discussed below) in both experimental treatments, and the bottom two graphs show the empirical results averaged across all of the birds in the experiment. Note that both sets of graphs are somewhat asymmetric and the potential increases more quickly to the left of the standard (i.e. where the adjusting option is less

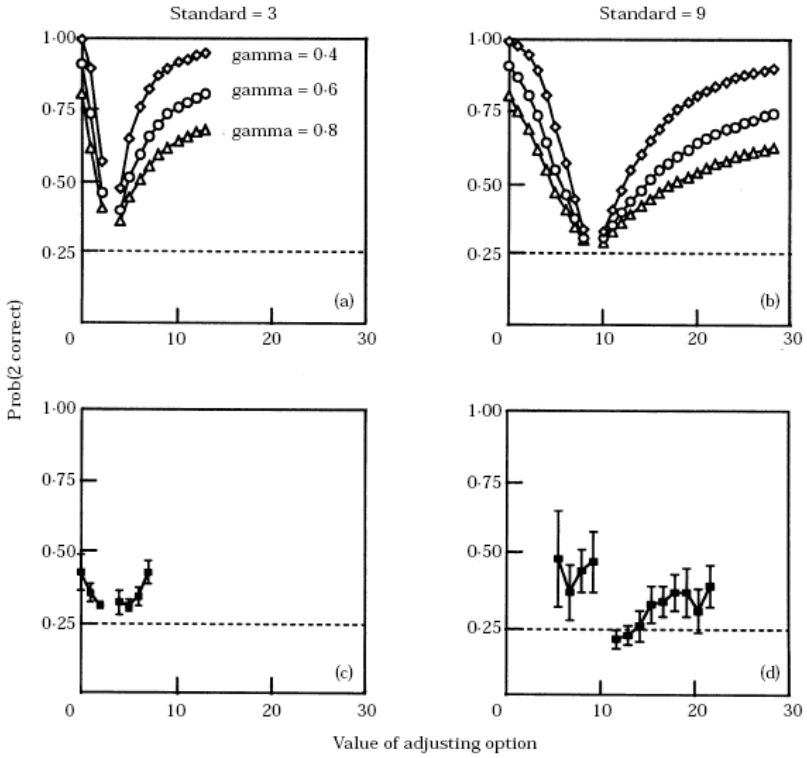


Figure 3.1 The potential, i.e. the probability of making 2 correct choices, as a function of the current value of the adjusting option. The top row depicts the predictions from Bateson and Kacelnik's model for several values of the parameter γ. The bottom row shows the empirical results averaged across all 8 birds. The left column shows the 3 pellet treatment, and the right shows the 9 pellet treatment. (Reprinted from Bateson and Kacelnik 1995 with permission from Elsevier)

than the standard). For the empirical data, this asymmetry is most visible in the 9 pellet treatment.

Modeling the experiment

The action set has $k = 2$ actions, and in the second stage $A = \{a_S, a_J\}$, where a_S corresponds to pressing the standard key and a_J corresponds to pressing the adjusting key. The reward values are characterized by the number of food pellets provided, which must be non-negative, so $R' = \mathbb{Z}^+$. The standard option always provides the same reward, r_S', with $r_S' = 3$ or 9 depending on the treatment. Both reward distributions are stationary during stage one, but during stage two the reward provided by the adjusting

option can change between blocks N. The reward provided in any given block, $r_J^{'}(N)$, is a random variable whose distribution is determined by the agent's responses on the choice trials.

Given this simple environment, many decision mechanisms are reasonable: there are many possible combinations of utility functions, averaging functions, and choice functions. For simplicity, assume here that the utility function is the identity function and thus that agent's subjective rewards are identical to the physical rewards, $r(n) = r^{'}(n) \in \mathbb{Z}^{+}$. The utility function can then be ignored, and model behavior is dictated by the averaging function, the choice function, and the environment. Given certain simple averaging functions, model predictions can be derived in closed form.

If the agent's estimates of current reward value depend only on the samples obtained in the current block, model dynamics are much easier to analyze. Due to the forced trials, the experimental structure ensures that the bird will always be exposed to at least four samples from each option in each block. If an averaging function has a memory length of four (the agent remembers the last 4 rewards obtained from each action) or less, the value of the estimators during the choice trials will only depend on the samples obtained during the current block. As a result, the agent's estimate during the choice trials depends only on the value of the adjusting key in the current block and not on the sequence of previous values. While slightly unrealistic, this assumption will allow us to derive analytic predictions, and BK also make this assumption in order to analyze their model.

A time window moving average (TWMA) with a memory length $W \leq 4$ satisfies this assumption. Since the TWMA is unbiased, on the choice trials the estimators will be identically equal to the current value of the titrating option: So for a TWMA with $W \leq 4$

$$\hat{r}_S(n) = r_S \quad \textit{and} \quad \hat{r}_J(n) = r_J(N). \tag{2.33}$$

As a result of Equation (2.33), model behavior is determined only by the current reward values, r_S and $r_J(N)$.

Deriving predictions for stage one

In stage one, the bird has a choice between two actions, $A = \{a_1, a_2\}$, where action a_1 corresponds to choosing the better key providing 9 food pellets per trial and action a_2 corresponds to choosing the key providing 3 food pellets. To complete stage one, each bird must choose the better key on more than 70% of the trials (at least 17 times in 24 trials) on each of two consecutive days. Define *daily success* as choosing the better key at least 17 times in a day. The number of days needed to complete the experiment, D, is then equivalent to the number of days needed before obtaining two consecutive daily successes. Let θ be the probability of daily success on any given day. Then D is distributed as a generalized geometric random variable with order 2 and success probability θ.

A geometric random variable represents the number of trials needed to obtain a single success from a sequence of independent Bernoulli random variables. Analogously, a generalized geometric random variable of order j represents the number of trials needed to obtain j consecutive successes from a sequence of independent Bernoulli random variables. The expectation and variance for the generalized geometric distribution was derived by Philippou et al (1983):

$$\bar{D} = E(D) = \frac{(1-\theta^j)}{(1-\theta)\theta^j}, \tag{2.34}$$

and

$$\sigma^2(D) = Var(D) = \frac{1-(2j+1)(1-\theta)\theta^j - \theta^{2j+1}}{(1-\theta)^2\theta^{2j}} \tag{2.35}$$

where j is the order of the distribution and θ is the success probability.

From the previous chapter, the choice function can be written as

$$\phi(n) = \mathcal{C}\left(\hat{r}_1(n), \hat{r}_2(n)\right) \tag{2.36}$$

where $\phi(n)$ represent the probability of making the correct choice, choosing the larger key, on each trial. With the assumed averaging function, $\hat{r}_1(n) = 9$ and $\hat{r}_2(n) = 3$ on all of the choice trials in stage one. So $\phi(n) = \phi = \mathcal{C}\left(\hat{r}_1(n), \hat{r}_2(n)\right)$ is constant across the entire

first stage, and the number of correct choices each day, Z, is binomially distributed: the number of successes (choices of the larger key) in 24 trials. The probability of a daily success (17 or more correct choice) is then given by:

$$\theta = \Pr(Z \geq 17) = \sum_{z=17}^{24} \binom{24}{z} \phi^z (1-\phi)^{24-z}. \quad (2.37)$$

By substituting Equation (2.37) into Equation (2.34), we can use compute $\bar{D}$ and $\sigma^2(D)$ with the appropriate choice function.

Graphing $\bar{D}$ and $\sigma^2(D)$ as a function of ϕ (Figure 3.2), shows that values of ϕ in the neighborhood of 0.7 appear to match both of the empirical results well. The matching choice function gives $\phi = \frac{9}{9+3} = 0.75$, slightly larger than expected value of 0.7 depicted on these plots.

Deriving predictions for stage two

Equation (2.33) allows stage 2 of the experiment to be analyzed as a Markov process. The states of the system is represented by the value of the adjusting key, $r_J(N)$, in the current block N, and the state transition probabilities depend only on the current state. Using Equation (2.33) and the definition of the choice function in Equation (2.36), the probability of choosing the adjusting option can be written in terms of the current value of the adjusting key on block N as

$$\phi(N) = P_J(N) = C(r_J(N), r_S). \quad (2.38)$$

Let $\phi_i, \{i = 0,1,2,...\}$, be the probability of choosing the adjusting option given that the value of the adjusting option on the current block equals i, i.e.

$$\phi_i = C(i, r_S) = \Pr(a(n) = a_J \mid r_J(N) = i). \quad (2.39)$$

The adjusting reward only changes when the bird chooses the same option on both of the choice trials in a block. When the adjusting reward is equal to i, the probability of choosing the adjusting option on both choice trials in a block is ϕ_i^2. Due to normalization, $P_S(n) = 1 - P_J(n) = 1 - \phi_i$, and the probability of choosing the standard option on both choice trials in a single block is $(1-\phi_i)^2$.

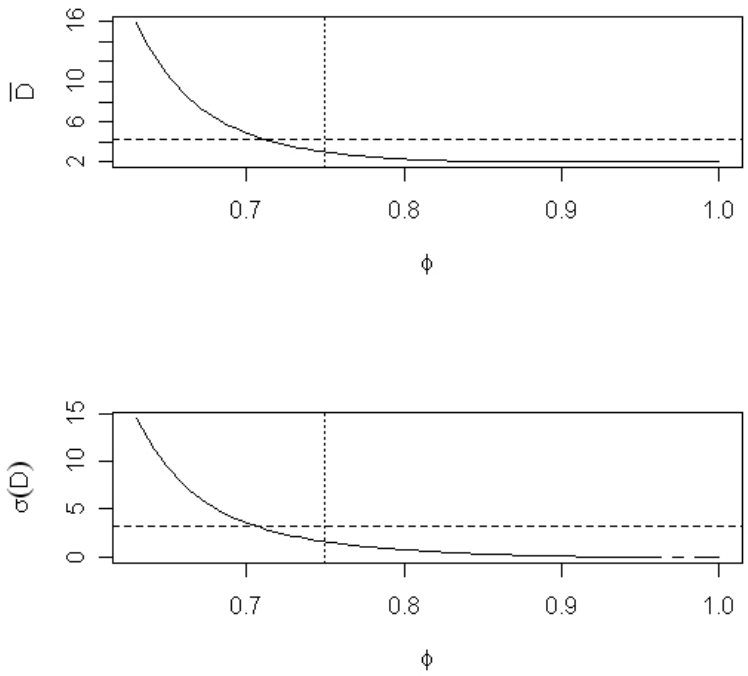


Figure 3.2- Predicted number of days needed to complete stage 1 graphed as a function of ϕ, the probability of choosing the larger option on each choice trial. The top graph shows the mean and the bottom the standard deviation. In both graphs the horizontal dashed line shows the empirical value. The vertical dotted line shows the choice probability predicted by the matching choice function, i.e. $\frac{9}{3+9}$.

So $\{r_J(N), N \in \mathbb{Z}\}$ is a discrete time Markov birth-death process with countable states, $i \in \{0,1,2,...\}$, and one step transition probabilities, $p_{ij} = \Pr(r_J(N) = j \mid r_J(N-1) = i)$, given by

$$p_{ij} = \begin{cases} \phi_i^2 & \text{if } j = i-1 \\ 1-\left(1-\phi_i\right)^2 - \phi_i^2 & \text{if } j = i \\ \left(1-\phi_i\right)^2 & \text{if } j = i+1 \\ 0 & \text{else} \end{cases} \tag{2.40}$$

for $i > 0$, and

$$p_{0j} = \begin{cases} (1-\phi_0)^2 & \text{if } j = 1 \\ 1-(1-\phi_0)^2 & \text{if } j = 0 \\ 0 & \text{else} \end{cases} . \tag{2.41}$$

Markov birth-death processes arise frequently in a variety of disciplines and have been studied extensively (for a review see Zikun and Xiangqun, 1980).

Using Equation (2.40) and the flow equations for the Markov process, the asymptotic mean and variance of the adjusting reward can be derived (Gallager, 1996; Howard, 1971). The mean is

$$\overline{r}_J = E(r_J) = \frac{\sum_{i=0}^{\infty} i\alpha_i}{\sum_{i=0}^{\infty} \alpha_i}, \tag{2.42}$$

and the variance is

$$\sigma^2(r_J) = Var(r_J) = \frac{\sum_{i=0}^{\infty} (i-\overline{r}_J)^2 \alpha_i}{\sum_{i=0}^{\infty} \alpha_i} \tag{2.43}$$

where

$$\alpha_0 = 1 \text{ and } \alpha_i = \prod_{j=0}^{i-1} \left(\frac{1-\phi_j}{\phi_{j+1}} \right)^2. \tag{2.44}$$

Note that the sum in (2.42) will only converge if

$$\sum_{i=0}^{\infty} \alpha_i < \infty, \tag{2.45}$$

otherwise the value of the adjusting option will wander off to infinity and the titrating procedure will not stabilize. A sufficient condition for (2.45) to hold is $(1-\phi_i)^2 < \phi_i^2$, or equivalently $\phi_i > \frac{1}{2}$, for all sufficiently large i (Gallager, 1996). Models that don't converge predict an infinite expected value for the titrating reward, $\overline{r}_J = \infty$, but all strictly increasing choice functions will necessarily converge. In the actual experiments, the titrations stabilized for all but one bird.

Using these mathematical results, we can now compute predictions for different models in both stages of the experiment.

Testing choice functions

BK propose a model based on scalar expectancy theory (Gibbon et al, 1984) which assumes that agents recall values from memory with some error, and that the

variance in the recalled value is proportional to the real value in memory. Individual samples are represented by normal distributions with constant coefficients of variation. When confronted with a choice, the agent takes a sample at random from the distributions associated with each action and chooses the action with the largest sample. This model is not formulated in terms of an averaging function and a choice function per se. Nonetheless the model BK test is mathematically equivalent, in this experiment, to a DELA model using a TWMA with $W = 4$ and the Gaussian choice function:

$$\phi(n) = \mathcal{C}\left(r_J(n), r_s(n)\right) = \Phi\left(\frac{r_J(n) - r_S(n)}{\gamma\sqrt{r_S^2(n) + r_J^2(n)}}\right). \tag{2.46}$$

Here Φ is the cumulative distribution function of a standard Gaussian random variable, γ is a model parameter, and as $\gamma \to 0$ this choice function becomes choosier, leading to more extreme preferences.

Figure 3.3 presents the predictions of BK's model on both stages of the experiment; Figure 3.4 and Figure 3.5 show predictions for the modified Boltzmann and Boltzmann choice functions respectively. In all cases, the mean number of days needed to complete the first stage was computed using Equations (2.34) and (2.37), while the predictions for stage two were computed with Equations (2.42) -(2.44) and the appropriate choice function equations (Chapter 2). The range of γ was limited by the predictions in stage one and only γ values predicting $2.1 < \bar{D} < 10.0$ were considered. The dashed lines on each graph show the empirical values from the experiment. If all the dashed lines cross the associated solid lines at the same, or similar, γ value, the model fits the data well. For all three choice functions, smaller values of γ lead to choosier (steeper) choice functions, and as is clear from the figures, choosier choice functions are needed to fit the results from stage one and less choosy choice functions are needed to fit the results from stage two.

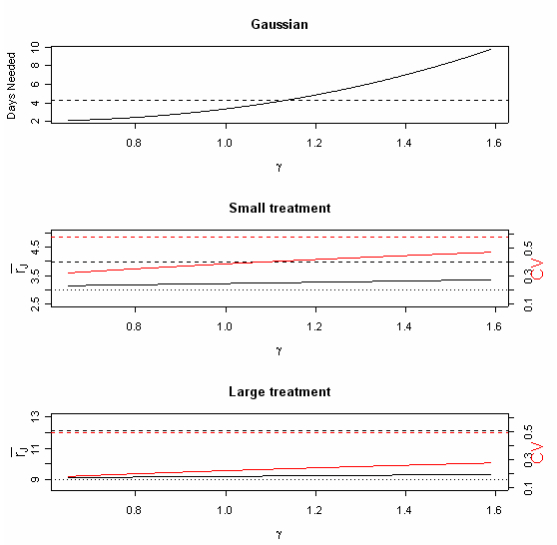


Figure 3.3 Predictions from the Gaussian model. The dashed lines depict the empirical values and solid lines model predictions. The top graph shows the predicted mean number of days needed to complete stage one. The middle graph depicts the predicted mean value (in black) and CV (in red) for the adjusting option in the small treatment, and the bottom graph depicts the same in the large treatment. The black dotted line in the bottom two graphs shows the value of the standard option in each treatment.

Clearly none of the models displayed in Figure 3.3-Figure 3.5 produce a good quantitative fit over the parameter ranges tested. Although each model can fit the data from stage one using an appropriate γ value, none of the models can come close to matching the numerical values for $\overline{r}_J$ on either treatment. Only the Boltzmann choice function is able to reproduce the empirical CV and only in the smaller treatment. All three models appear unable to reproduce the quantitative results from stage two over the parameter ranges tested, let alone fit the data from both stages with the same parameter value.

However, the models do replicate several qualitative aspects of the data. All three models predict some titration bias: the Gaussian and the modified Boltzmann choice functions predict a substantial titration bias in both treatments while the Boltzmann choice function predicts a smaller bias in the 3 pellet treatment and no bias in the 9 pellet treatment. Moreover, all three models predict larger CV values in the 3 pellet treatment than in the 9 pellet treatment.

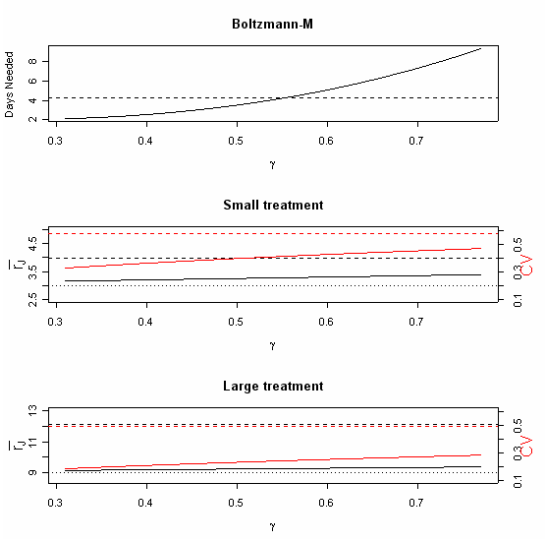


Figure 3.4 Predictions from the modified Boltzmann choice function. All symbols are as in Figure 3.3.

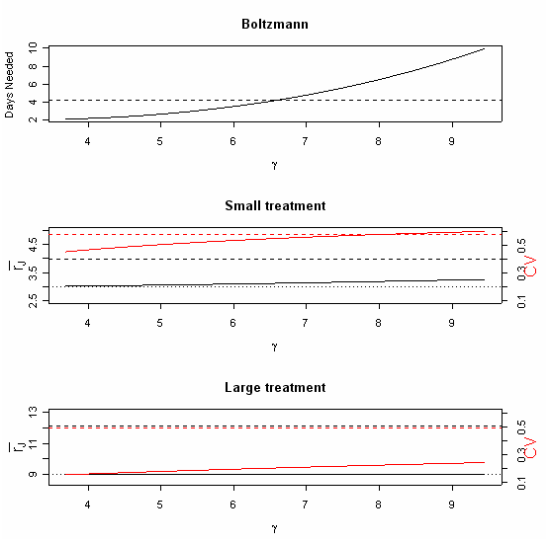


Figure 3.5 Predictions from the Boltzmann choice function. All symbols are as in Figure 3.3.

Analytic results

The Boltzmann choice function predicts less of a titration bias than the other two choice functions over the parameter ranges depicted in Figure 3.3-Figure 3.5. The source of this discrepancy is made clearer by examining the potential, V, which is equal to the probability of making correct choices on both choice trials in a block (Figure 3.6). When the value of the adjusting option is less than the value of the standard option, i.e. $r_J(n) = r_S - l$ for $(l = 0,1,...,r_S)$, the potential, V_{-l}, is equal to the probability of choosing the standard option twice

$$V_{-l} = \left(1 - \mathcal{C}\left(r_S - l, r_S\right)\right)^2. \tag{2.47}$$

When the adjusting option is larger than the standard, the potential is equal to the probability of choosing the adjusting option twice

$$V_{+l} = \mathcal{C}(r_S + l, r_S)^2. \tag{2.48}$$

The experimental results in Figure 3.1 show an asymmetric potential with $V_{-l} > V_{+l}$. A model will predict this asymmetric potential if

$$\left(1 - \mathcal{C}\left(r_S - l, r_S\right)\right)^2 > \mathcal{C}(r_S + l, r_S)^2. \tag{2.49}$$

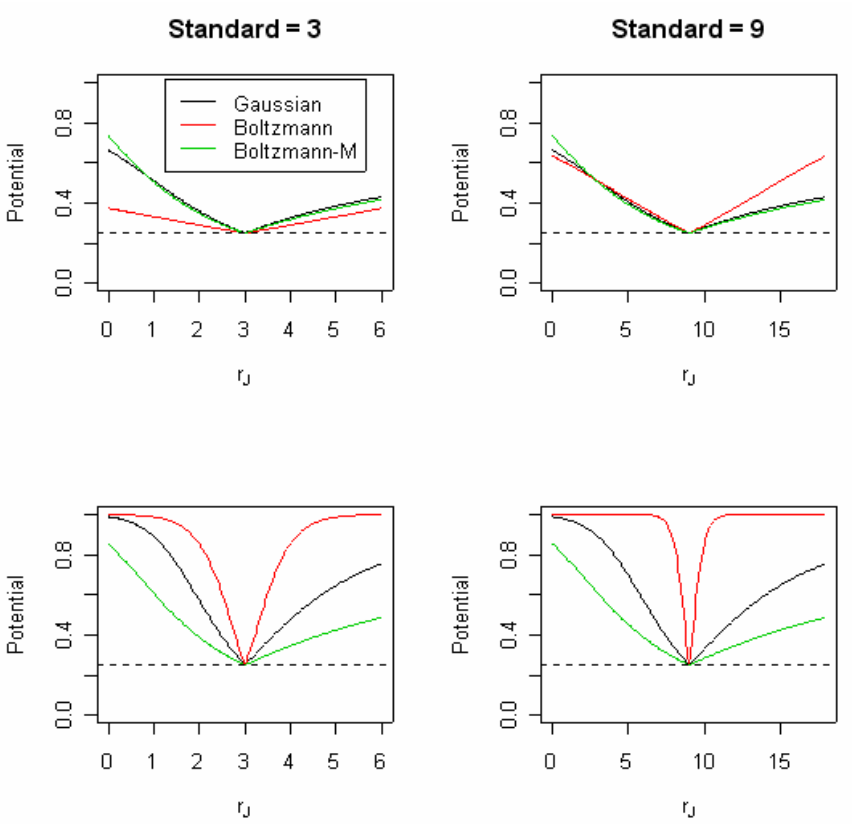


Figure 3.6 The potential as a function of the value of the adjusting option. The top graphs show model predictions using the γ values suggested in stage one of the experiment (Gaussian =1.12, Boltzmann = 6.6, Boltzmann-M = 0.56). These values were estimated from Figure 3.3-Figure 3.5. The bottom graph shows predictions with $\gamma = 0.4$ for all models. Note that the Gaussian and modified Boltzmann functions are asymmetric about the value of the standard, increasing more quickly to the left, while the Boltzmann choice function is symmetric.

Rearranging Equation (2.49) using condition C.2, gives

$$\mathcal{C}(r_S - l, r_S) < \mathcal{C}(r_S, r_S + l) \tag{2.50}$$

which is equivalent to the definition of strict R-concavity in Equation (2.18). Any strictly R-concave choice function will thus produce an asymmetric potential, but R-affine functions will not. So difference-based choice functions, like the Boltzmann, will not produce asymmetric potentials, but ratio-based choice functions, like the Gaussian and modified Boltzmann, will. Thus the empirically observed asymmetric potential is more consistent with a ratio-based choice function than with a difference-based choice function.

In general a potential asymmetry will produce a titration bias. Think of the state of the system (shown on the x-axis in Figure 3.6) as undergoing a random walk. When $r_J = r_S$, the system is in an effective equilibrium where the probability of taking a step to the left or the right (the potential) is equal. Away from equilibrium, the potential is equivalent to the probability that the system will take a step back towards equilibrium. To the right of equilibrium, the probability of stepping back towards equilibrium, V_{+1}, increases more slowly than does the probability to the left, V_{-1}. So the system will tend to wander further to the right than it will to the left because on the right there is less potential 'pushing' the system back towards equilibrium. Any choice function that produces this asymmetric potential will predict a larger expected value for the adjusting option and thus a titration bias.

Define the titration bias as $\Delta_J = \overline{r}_J - r_S$. Appendix 1 proves that all strictly R-concave choice functions predict a titration bias with $\Delta_J > 0$. However, Appendix 1 also shows that R-affine choice functions predict a titration bias, $\Delta_J > 0$, but that more extreme values for γ are needed for the bias to become substantial. This bias is not a result of asymmetry in the potential but is instead a result of bounding the system at zero. Since the system is allowed to wander infinitely far to the right but only r_S steps to the left (bounded at zero), the titration is inherently biased. However, the magnitude of this effect is small for R-affine choice functions with steep slopes, and the difference-based Boltzmann choice function can only generate a substantial titration bias with relatively large γ values, in contrast to the small values needed to fit the results from the first stage of the experiment.

Adding a utility function

Taken together, the asymmetric potential and substantial titration bias suggest strongly that the starlings use ratio-based choice functions. Nonetheless, despite the better performance by ratio-based choice functions, none of the simple models produced a good quantitative fit to the empirical data. Especially in the 9 pellet treatment the

empirical titration bias was substantially larger than the bias predicted by any of the models over the parameter ranges tested. By embellishing these simple models, for example by adding a non-linear utility function, model fit can be improved. In fact, a non-linear utility function is able to induce a more asymmetric potential, and thus a more extreme titration bias, even for R-affine choice functions.

In another paper, BK (1996) conducted a similar titration experiment but the titration failed to stabilize, and the value of the adjusting option increased continually. The reward sizes involved were quite large, and BK suggested that the starlings were unable to accurately measure such large rewards and thus effectively did not differentiate between the adjusting option and the standard option. In this vein, consider an agent that cannot count above a certain number. Able to count only up to some highest value, the agent will treat all larger rewards as having the same value. This can be represented with the utility function

$$\mathcal{U}(r') = \min(r', \beta). \tag{2.51}$$

Here $\beta \in \mathbb{Z}$ is the parameter determining how high the agent can count. This simplistic utility function is actually able to improve the model performance in the 9 pellet treatment.

Figure 3.7 shows model predictions on the 9 pellet treatment using the utility function with $\beta = 10$. The addition of the new utility function substantially improves model fit for all three models. The ratio-based choice functions can fit the results from the 9 pellet treatment and the results from stage 1 with similar γ values (compare Figure 3.7 with the stage one predictions in Figure 3.3 and Figure 3.4). The Boltzmann choice function also displays a substantial titration bias but it is not as large as the other two models. Note that this utility function with $\beta = 10$ will have no impact on model predictions during stage one and only minimal impact on model predictions in the 3 pellet treatment, so the utility has a net positive impact on model performance overall.

As shown in Figure 3.8, the utility function increases the titration bias in the 9 pellet treatment because it reduces the probability of a correct choice when the system is to the right of equilibrium ($r_J > 9$). With a smaller potential to the right of equilibrium,

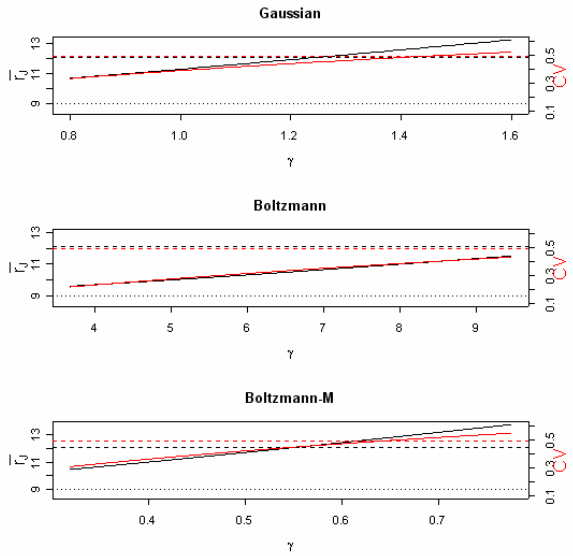


Figure 3.7 Model predictions in the larger 9-pellet treatment using the non-linear utility function with $\beta = 10$.

the system can more easily move in that direction. Note that values of $\beta < 10$ will lead, in the large treatment, to violations of the condition in Equation (2.45). In such cases, $\mathcal{U}(r_S^{'} + j) = \mathcal{U}(r_S^{'})$ for all positive integers j, both actions will be chosen with equal probability, and the system will undergo a truly random walk to the right of the value of the standard option.

This utility function, while highly contrived, presents a simple way to improve model predictions, and many concave utility functions will have similar effects. The take away point is that increasing the asymmetry in the potential will increase the titration bias. Moreover, reducing the choosiness of the choice function (larger values for γ) will reduce the slope of the potential while simultaneously increasing the titration bias and the CV of the titrating value. However, less choosy choice function will also reduce the choice probability during stage one and thus increase the number of days needed to complete the stage. Interestingly, increasing the length of the agent's memory is a natural way to reduce the effective choosiness of the choice function during stage two, while not substantially impacting the time needed to complete stage one.

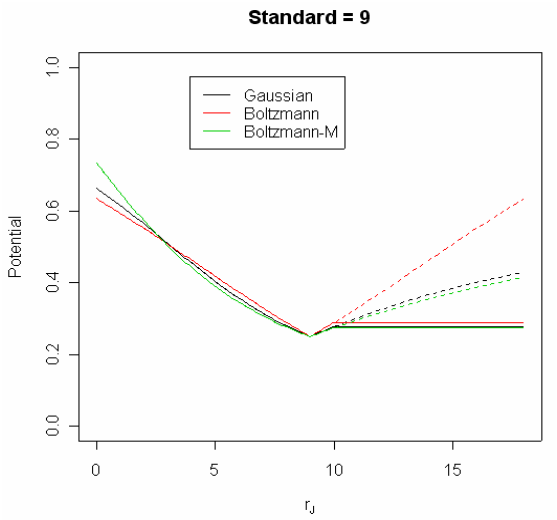


Figure 3.8 The potential with a non-linear utility function. The utility function is shown with $\beta = 10$ and the same γ values as in the top row of Figure 3.6. The dashed lines show the values with a linear utility function (equivalent to $\beta = \infty$).

If the memory is longer than four, behavior will depend on the value of the adjusting option in previous trials and will lag behind the current value during stage two. This could reduce the slope of the potential and increase the CV of the titrating value. Moreover, if the averaging function maintains or increases the asymmetry in the slope of the potential, the titration bias will persist or grow. In stage one this modification will only increase the predicted time to completion slightly, for once the memory is 'full' model predictions are as derived above.

Unfortunately, increasing the memory length also increases the complexity of model dynamics, and the preceding proofs lose validity. In lieu of analytic formulas, at this point I must resort to simulation to obtain model predictions. Chapter 7 conducts simulations for these experiments in the context of a larger analysis of DELA model performance.

Discussion

The observed asymmetry in the potential functions, as well as the observed titration bias, is more consistent with ratio-based choice functions than with difference-based choice functions. These results do not rule out difference-based choice functions since non-linear utility functions can lead to similar phenomena. Nonetheless, on the

whole these results suggest strongly that the starlings are using ratio-based choice functions. If so, the Boltzmann choice function, despite its prevalence in the reinforcement learning and AI literature, is probably not a good model for the way organisms make decisions. A more complete analysis of model performance, presented in Chapter 7, will further strengthen this view.

If organisms do indeed use ratio-based choice functions, results derived with the titration procedure must be reevaluated because this procedure will always lead to biased indifference points. Interestingly, this bias is a result of the way the titration is conducted. The titration procedure treated here, as well as most other published procedures, utilizes an arithmetic step size, adding or subtracting one unit from the value of the adjusting option. However, there are other ways to conduct the titration; Lea (1976) utilized an alternative approach with a geometric step size: titration proceeded by either multiplying or dividing the value of the adjusting option by 1.5. Interestingly, Lea (1976) did not find a titration bias with this procedure, instead finding indifference points almost exactly equal to the value of the standard option. Geometric adjustment makes more sense for a ratio-based choice function, and perhaps the titration procedure would be more accurate if geometric adjustments were always used.

One additional aspect of this experimental results deserves comment, namely the discrepancy between the results from stage one and stage two. For the simple models considered here, choosier choice function were needed to fit the results from stage one while less choosy functions were needed to fit the results from stage two. Interestingly, the environment was stationary during stage one but transient during stage two. Less choosy choice functions will lead to more exploration/sampling of sub-optimal options. In transient environments where reward distributions can change unexpectedly, exploration/sampling is more valuable, and reduced choosiness makes intuitive sense. In stationary environments, there is less uncertainty and thus a premium on exploitation and choosier choice functions. So the discrepancy between experimental stages could be explained if the starlings are able to recognize environmental transience and respond differently to transient environments than to stationary ones.

Chapter 7 will further explore this hypothesis using data from several additional experiments with starlings. Most of these other experiments presented the birds with a special type of multi-armed bandit problem, the basic risk sensitivity experiment, which is functionally equivalent to a one-armed bandit problem. In the next chapter I introduce the basic risk sensitivity experiment, and review some theories and experimental results from risk-sensitive foraging theory. After proving some results about the risk sensitive behavior of ratio-based DELA models in Chapters 5 and 6, Chapter 7 will return to the issue of model selection and hopefully shed some more light on behavioral differences in the face of transient environments.

Chapter 4 Risk sensitivity

Consider an organism faced with a choice between two food sources providing the same mean amount of food but with different variances. One option always provides the same reward, while the other provides either a large or a small reward with equal probability. If the organism is offered a repeated choice between these two options, which will it prefer?

An organism that displays a preference when faced with the preceding problem is said to exhibit *risk sensitivity*. Here risk refers to the uncertainty/variability/variance in the rewards. If the organism prefers the constant or certain option, it is called *risk averse*, and if it prefers the variable or uncertain option, it is called *risk prone*. Risk sensitivity in this sense is of great interest in economics, psychology, and animal foraging (Bateson and Kacelnik, 1996; Weber et al., 2004).

Experimentalists have focused on documenting the risk preferences that organisms or individuals express while theorists have attempted to provide explanations, justifications, or models for these behaviors. This chapter will explore the main experimental results, and some proposed models, from the animal behavior literature.

The formal problem

The basic risk sensitivity experiment (BRSE) in animal foraging presents an organism with a choice between two foraging options providing food rewards. The constant (or certain) option always delivers the same reward, $r_c^{'}$; the variable (or stochastic) option delivers either a large reward, $r_+^{'}$, or a small reward, $r_-^{'}$, with probability p_+ and p_- respectively. All rewards are non-negative, $r_i^{'} \in \mathbb{R}_+$, and the rewards are selected so that both options provide the same mean reward:

$$r_c^{'} = \overline{r}_v^{'} = p_+ r_+^{'} + p_- r_-^{'} \tag{3.1}$$

where $\bar{r}_v^{'}$ is the mean value of the rewards from the variable option. The action set has two possible actions, $A = \{a_c, a_v\}$, corresponding to foraging on the constant or variable option respectively, and the reward set contains three possible values: $R^{'} = \{r_c^{'}, r_+^{'}, r_-^{'}\}$. In general the reward distributions are held constant and the BRSE environment is stationary. From a LA perspective, the BRSE presents a Q-model one armed bandit problem.

Many foraging experiments characterize rewards with two dimensions: a reward size, s, and a waiting time or delay to the reward, w. For example, the experiment in Chapter 3 delivered rewards five seconds after the bird pressed key, so $w = 5$, and the reward size was equivalent to the number of pellets provided. In such cases the physical rewards are represented as bivariate vectors and in general

$$\mathbf{r}_+^{'} = \{s_+, w_+\} \quad \mathbf{r}_-^{'} = \{s_-, w_-\} \quad \mathbf{r}_c^{'} = \{s_c, w_c\} \tag{3.2}$$

for the BRSE. However, variance is usually introduced in only one dimension, either into the waiting time or the reward size but not both. If variance is introduced into the reward size, the waiting times are identical and the reward sizes are chosen so that the mean values are equal (Figure 4.1-a). If variance is introduced into the waiting times, the reward sizes are identical and the mean waiting times are equal (Figure 4.1-b). As we shall see, organisms appear to respond differently depending on whether variability is introduced into the reward sizes or the waiting times.

An organism's preference is usually summarized by the asymptotic proportion of choices for the variable option, θ. Let N_n be the total number of choices for the variable option on the nth trial of an experiment. Then

$$\theta = \lim_{n \to \infty} \frac{N_n}{n}, \tag{3.3}$$

and if $\theta > .5$, the organism displays risk prone behavior, while $\theta < .5$ denotes risk averse behavior. Of course most experiments are only run for a finite number of trials, in which case the limit in Equation (3.3) is not infinite. Nonetheless, Equation (3.3) is useful for deriving model predictions. Ideally, the preference value predicted by a model, $\tilde{\theta}$, should be compared quantitatively with the empirically observed value, θ, but it is fairly

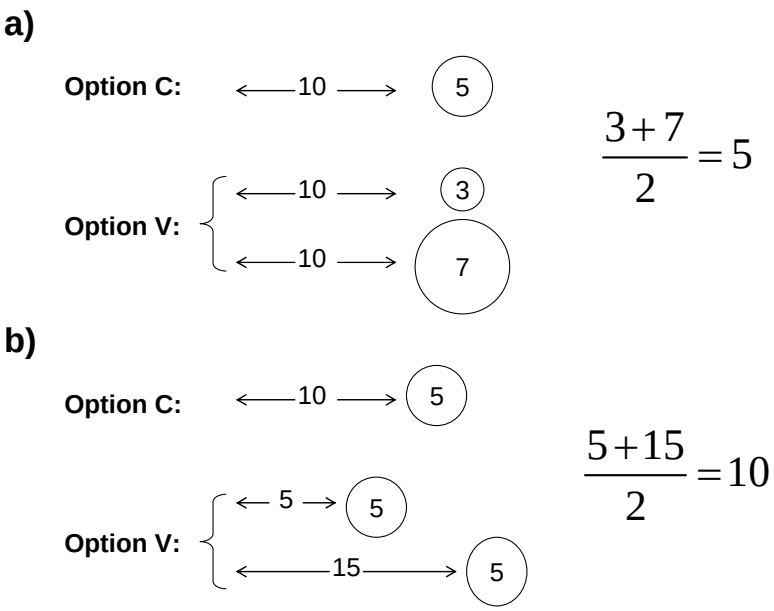


Figure 4.1- Two possible experimental schedules for the BRSE. The top figure shows a schedule with variability in amount, while the bottom figure shows a schedule with variability in the waiting time. In both cases, the variable option and the constant option provide the same mean reward and waiting times.

common (e.g. Bateson and Kacelnik 1995b) to instead focus on correctly predicting the direction of risk sensitivity while ignoring the magnitude. Perhaps due to the complexity of the experimental subjects (usually birds), model assessment is often limited to determining whether a model appropriately predicts $\theta > .5$ or $\theta < .5$ in concert with the observed experimental results.

The BRSE can be modified and complicated in many ways. For example, the organism can be presented with more than two options, options that deliver more than two rewards, or options that deliver rewards with different mean values. In all cases the behavior of interest is the development of preferences based on the variance, not the mean, of the rewards provided. Although here the problem is formulated as a foraging problem with food rewards, the basic framework is easily applied to rewards other than food, and many of the models discussed below have their roots in economics and psychology. The following brief survey of experimental work and published models, however, will focus on the animal behavior literature. See Weber et al. (2004) for a recent review of similar experimental results from humans. An extensive review of the experimental work from animal behavior can be found in Kacelnik and Bateson (1996), and Bateson (2002) reviews some more recent experimental results. Here I summarize

some of the major experimental commonalities, especially those discussed in Kacelnik and Bateson (1996).

Main experimental results

The direction of risk sensitivity generally depends on which dimension is variable. Organisms tend to be risk averse with respect to variability in gains or benefits but risk prone with respect to variability in costs or punishments (Weber et al., 2004; March, 1996). In foraging experiments with bivariate rewards, $r_i^{'} = \{s_i, w_i\}$, the waiting time is effectively a cost but the size represents a gain. Consistent with the more general pattern of risk aversion for gains and risk preference for losses, foraging organisms tend to be strongly risk prone when variance is introduced into the waiting time but weakly risk averse or indifferent when reward sizes are variable (Kacelnik and Bateson, 1996).

The pattern of strong risk preference for variability in waiting time but weak risk aversion for variability in amount is pervasive but not universal. There is also some evidence that an organism's risk preferences can change as a function of its energetic state. In a classic experiment, yellow-eyed juncos (*Junco phaeonotus*) were risk averse while on positive energy budgets but risk prone under negative energy budgets (Caraco et al. 1990). This so-called energy budget rule (Stephens and Krebs, 1986) has been replicated in several other studies, but there is some evidence that it only applies to smaller organisms and that larger animals do not display such transitions (Kacelnik and Bateson, 1996).

The energy budget rule and the asymmetry in risk preferences are probably the two most pervasive and well documented results from the literature. However, the utility-estimator-choice modeling framework presented in Chapter 2 does not incorporate any energetic considerations, so energy budget effects cannot emerge. Thus the following discussion will focus on the asymmetry in risk preferences and models that can generate these asymmetries.

From the perspective of basic optimal foraging theory (Stephens and Krebs, 1986), risk sensitivity is a puzzling phenomenon. Most classical foraging models assume

that organisms try to maximize their long term rate of energy intake. However, in the BRSE both options provide equivalent long term rates of energy intake, and classical foraging models predict risk indifference. The overwhelming experimental evidence for risk sensitivity has motivated the formulation of a variety of models trying to explain how and why organisms display risk sensitive behavior.

Most models generate risk sensitive behavior either through the utility function or through the dynamics of the learning process. Utility based models utilize a non-linear utility function, $\mathcal{U}$, to generate the observed risk sensitive behavior. Learning based models exploit the fact that simple learning models, including many of those discussed in Chapter 1, can produce risk sensitive behavior. Learning based models are discussed in the next chapter, and the remainder of this chapter reviews some of the utility based models.

Utility based models for risk sensitivity

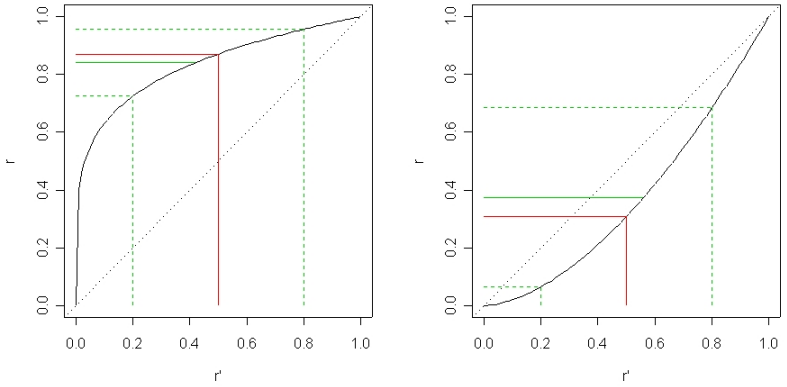
Non-linear utility functions are perhaps the most common explanations for risk sensitive behavior. The BRSE arises when an organism is presented with two options providing the same mean physical rewards but with different variances. However, the means are only equivalent from the experimenter's perspective, and Equation (3.1) refers to the physical rewards, r_i', as measured by the experimenter. If the utility function, $\mathcal{U}$, is non-linear then the mean subjective rewards might be unequal,

$$r_c \neq \overline{r}_v = E(r_v) = p_+ r_+ + p_- r_- , \tag{3.4}$$

despite the fact that $r_c' = \overline{r}_v'$. Assuming that the organism's choice function depends on the long term average of the subjective rewards from each option, the inequality in Equation (3.4) can lead to risk sensitivity and $E(r_v) > r_c$ suggests risk prone behavior.

Risk sensitivity from non-linear currencies is generally a direct result of Jensen's inequality which states that, if $\mathcal{U}(r')$ is a convex function and r' is a random variable,

$$E\left(\mathcal{U}(r')\right) \geq \mathcal{U}\left(E(r')\right). \tag{3.5}$$



$$\frac{r_+ + r_-}{2} < r_c \qquad \frac{r_+ + r_-}{2} > r_c$$

Figure 4.2- Risk sensitivity as a result of a non-linear utility function. The left graph depict a concave, and risk averse, utility function while the right graph shows a convex, and risk prone utility function. The solid black line shows the utility function, and the dotted line shows the linear $r = r'$ line. The physical rewards depicted are $r_+' = 0.8, r_-' = 0.4$ (shown as dashed green lines), and $r_D' = 0.6$ (shown in red). The mean value of the utilities is shown as a solid green line.

The inequality is strict if $\mathcal{U}(r')$ is strictly convex. On the BRSE, $r_c' = \overline{r}_v' = E(r_v')$, and Equation (3.5) is equivalent to

$$E\left(\mathcal{U}(r_v')\right) \geq \mathcal{U}\left(r_c'\right) \rightarrow E\left(r_v\right) \geq r_c . \tag{3.6}$$

So by Jensen's inequality, strictly convex utility functions will produce risk prone behavior, but strictly concave utility functions will produce risk averse behavior. This effect is depicted in Figure 4.2.

Non-linear utility functions are pervasive in the economic literature as models for risk sensitive behavior. In economics, expected utility theory (Von Neumann and Morgenstern, 1947) and prospect theory (Kahneman and Tversky, 1979) both utilize non-linear utility functions. These models can also be applied directly to animal foraging experiments, but the focus in the foraging literature is often slightly different due to the bivariate rewards (waiting time and reward size). Within foraging theory, models specify a utility function, $\mathcal{U}: \mathbb{R}_+^2 \rightarrow \mathbb{R}_+$, that can generate the experimentally observed asymmetry in the response to variance in waiting time and variance in reward size.

The *expectation of ratios* (EOR) or short term rate maximization model (Bateson and Kacelnik 1996, Harder and Real 1987), for example, proposes that organisms use the utility function

$$r = \mathcal{U}(s, w) = \frac{s}{w}. \tag{3.7}$$

This utility is linear with respect to s, so for a BRSE with variance in reward size ($s_c = \bar{s}_v = p_+ s_+ + p_- s_-$ and $w_c = w_+ = w_-$), this utility predicts risk indifference:

$$\bar{r}_v = \frac{\left(p_+ s_+ + p_- s_-\right)}{w_c} = \frac{s_c}{w_c} = r_c. \tag{3.8}$$

However, this utility is strictly convex with respect to w. Thus for a BRSE with variance in waiting time ($w_c = \bar{w}_v = w_+ p_+ + w_- p_-$ and $r_c = r_+ = r_-$), the EOR utility predicts risk prone behavior as a direct result of Jensen's inequality. So the EOR successfully predicts the observed risk prone behavior for variability in waiting time, but cannot generate risk averse behavior for variability in reward size.

The EOR utility has the unattractive feature that it goes to ∞ as $w \to 0$. The *hyperbolic-decay utility function* (Mazur, 2001; Mazur, 1984) eliminates this unwanted feature and introduces a free parameter, $\alpha > 0$:

$$r = \mathcal{U}(s, w) = \frac{s}{1 + \alpha w}. \tag{3.9}$$

This utility function is finite for all values of w, and is hence more realistic in practice than the EOR utility. This utility function remains strictly convex with respect to w, so it will still predict risk prone behavior for variability in waiting time. It also remains linear with respect to reward size and thus predicts risk-indifference for variability in amount.

If there is a positive correlation between the reward size and the waiting time in the denominator of Equation (3.7), then the EOR utility further predicts risk aversion for variability in amount (Caraco et al.,1992). The EOR utility can be modified to introduce this correlation by assuming that the organism also includes the time needed to consume the reward in the denominator. Consider the following modification of the EOR utility:

$$r = \mathcal{U}(s, w) = \frac{s}{\alpha s + w}. \tag{3.10}$$

Here $\alpha > 0$ denotes the time needed to consume one unit of food, and this utility function assumes that the time needed to consume a reward increases linearly with the size of the reward. Importantly, this utility is concave with respect to s and so by Jensen's inequality predicts risk aversion for variability in amount. Of course any utility function that is concave with respect to s and convex with respect to w will produce the observed asymmetry in risk preferences. Many different currencies have been proposed but few have the proper concavity and convexity properties, as most authors have focused on biologically realistic and intuitive currencies (see K&B, 1996 for a review).

Models with non-linear currencies display what I will call *apparent risk sensitivity*: what appears to the experimenter as a preference based on variance is, from the organism's perspective, simply a preference based on means. When the proper units (subjective rewards) are used, the behavior can be seen as a choice based on the average values. In contrast, *genuine risk sensitivity* involves a preference between two options even when the means are equal from the organism's perspective, i.e. $\overline{r_v} = r_c$. In this sense, utility based models are unable to produce genuine risk sensitivity by definition; genuine risk sensitivity must be produced by some other aspect of the decision mechanism.

Within the utility-averaging-choice structure, models need to additionally specify averaging and choice functions in order to generate quantitative model predictions. The averaging function specifies how the organism combines multiple rewards from a food source into an estimate of the source's overall value, and the choice function translates these estimates into probabilities of action or choice. In order to generate genuine risk sensitivity, appropriate averaging and choice functions are also needed.

Recall that most utility based models assume that the organism's decision is based on a comparison of the expected value of the rewards, i.e. $E(r_v)$ vs. r_c. We can think of these models as implicitly assuming that the organism is using the long term averaging function (LTA, Chapter 2) to estimate expected reward values. After sufficient samples from the variable option, the estimate derived with a long term average, $\hat{r}_v$, will necessarily converge to the expected value $E(r_v)$ by the weak law of large numbers.

Any model using a LTA, a proper choice function, and a strictly concave utility function, will display risk averse behavior. However, this is apparent risk sensitivity. As we shall see in the next Chapter, other averaging functions are able to generate genuinely risk sensitive behavior even when $\overline{r}_v = r_c$.

Chapter 5 The hot stove effect

While utility based models are the most common explanations for risk sensitive behavior, models emphasizing the role of learning have also been proposed. Experiential learning models, as exemplified by the learning rule for classical LA or by the averaging function in a DELA model, can generate risk aversion on the BRSE simply because of the dynamics of the sampling process. Several authors have noted this tendency towards risk aversion, and it has been called the *hot stove effect* (Denrell and March, 2001), in honor of the Mark Twain quote:

> We should be careful to get out of an experience only the wisdom that is in it – and stop there; lest we be like the cat that sits down on a hot stove lid. She will never sit down on a hot stove lid again—and that is well; but she also will never sit down on a cold one.

An unpleasant experience associated with an action will, for a learning organism, generally decrease the probability of choosing that action in the future. However, an experiential learner will only learn about the returns from an action by engaging in the action. The cat, traumatized by one bad experience, will likely never sit down on the stove again and thus can't learn that sitting on a *warm* stove lid might be quite pleasant. The unpleasant experience effectively scares the cat off, and the cat does not learn about the potential for good returns from sitting on the stove.

In the context of the BRSE, the hot stove effect can lead to risk aversion because of runs of bad luck. After a run of bad luck, i.e. several small rewards from the variable option, many learning agents will underestimate the 'true' expected value of the variable option relative to the certain option and reduce their probability of choosing the variable option accordingly. When the agent is overestimating the value of the variable option, i.e. after a run of good luck, the probability of selecting the variable option will increase. Since an overestimate increases the probability of selecting the variable option, this type of estimation error will be corrected quickly, but an underestimate will tend to persist because it reduces the probability of selecting the variable option (Denrell, 2005). Thus the organism will spend more time underestimating the value of the variable option, and will tend to be risk averse.

March (1996) used computer simulations to demonstrate that three simple decision mechanisms would generate genuinely risk averse behavior on the BRSE. March's first model, the fractional adjustment model, is a version of the Bush and Mosteller (1958) linear reward penalty (LRP) model (discussed in Chapter 1). The other two models are equivalent to DELA models; both use the matching choice function, but the second uses a long term average (LTA) while the third uses an exponentially weighted moving average (EWMA). The LRP and EWMA models displayed risk aversion throughout the model simulations. The LTA model, although eventually converging to indifference, showed risk aversion in the short term. Thus all three models could generate short term risk aversion, but the LTA model failed to generate any long term risk preference.

Other authors have proposed similar learning models as explanations for risk aversion. Weber et al. (2004) use a version of the LRP model and compared the results to human choice data, while Keaser et al. (2002) test a model equivalent to March's EWMA model using data from bumblebees. Also with data from bumblebees, Shapiro et al. (2000) simulate the EWMA based model proposed by Couvillon and Bitterman (1991). All of these studies used computer simulations in order to demonstrate the risk sensitive characteristics of the models.

While simulation is the most common method for obtaining model predictions, Denrell (2005) and Niv et al. (2004) were able to show analytically that certain types of learning models will always produce risk aversion. Denrell considered an S-model environment with positive and negative rewards that is quite different from the BRSE, while Niv et al. focused only on models using EWMA averaging functions. The remainder of this chapter presents a proof for the BRSE environment, demonstrating that a specific class of DELA models will always produce risk averse behavior. All three proofs have some fundamental similarities, but utilize different formalisms.

A general DELA model for the BRSE

Consider a DELA model using a separable, length-W averaging function. On the BRSE, the agent's state is represented by two memory vectors, $\mathbf{\Psi}_v(n)$ and $\mathbf{\Psi}_c(n)$, each of length W, with $\mathbf{\Psi}_i(n) \in \mathbb{R}_+^W$ for $i \in \{c, v\}$. Recall that v denotes the variable option, c the constant option, and n the trial number. An averaging function, $\mathcal{A}: \mathbb{R}_+^W \to \mathbb{R}_+$, converts each memory vector into an estimate of the expected value for the associated food source

$$\hat{r}_i(n) = \mathcal{A}\left(\mathbf{\Psi}_i(n)\right) \quad i \in \{c, v\}. \tag{4.1}$$

Let $\phi(n) \in [0,1]$ denote the probability of selecting the variable option on the next trial, $\phi(n) = P_v(n) = \Pr(a(n) = a_v)$. This probability is generated by a choice function, $\mathcal{C}: \mathbb{R}_+^2 \to \mathbb{R}_+$, that acts on the estimates:

$$\phi(n) = \mathcal{C}(\hat{r}_v(n), \hat{r}_c(n)). \tag{4.2}$$

Since some action must be chosen on each trial, the probability of choosing the constant option is equal to $1 - \phi(n)$. The agent's state is determined by its memory vectors and the model dynamics are driven by the action of the averaging and choice functions on these memory vectors.

For a separable length-W averaging function, the memory vector $\mathbf{\Psi}_i(n)$ records the values of the last W rewards obtained from action a_i as of the nth sample. The variable option provides two different rewards and one possible realization for the memory vector might be

$$\mathbf{\Psi}_v(n) = \{r_+, r_-, r_-, r_+, r_-, r_-\}.$$

In this example $W = 6$, and the rightmost value represents the most recent reward. When a new sample is obtained from the variable option, the organism 'forgets' the first element on the list and appends the new sample. Thus the memory functions as a finite length queue that is updated by eliminating the leftmost element and appending a new value on the right. Importantly, the memory vectors only change when the associated

actions are chosen. With this updating rule, the organism's memory for a food option can only change when that food option is chosen or experienced.

On the BRSE, after W samples from the constant option the associated memory vector will be $\mathbf{\Psi}_c(n) = \{r_c, r_c, r_c, ..., r_c\}$ and will not change thereafter, so this memory vector is effectively constant after enough samples are taken. The memory for the variable option can be represented as a length-W binary sequence, with a one representing r_+ and a zero representing r_-. There are thus 2^W possible states for the variable memory vector and associated with each state is a unique integer, $\Psi(n) \in \{0,1,2,3,...2^W - 1\}$, such that $\Psi(n)$ is the integer corresponding to the binary sequence $\mathbf{\Psi}_v(n)$. We can think of the organism's state as defined completely by the integer $\Psi(n)$ and rewrite the averaging function as $\hat{r}_v(n) = \mathcal{A}\left(\Psi(n)\right)$.

Using this representation, the memory or state updating operation can be represented as an arithmetic operation

$$\mu\left(\Psi(n), r(n)\right) = (2\Psi(n) + r^*(n)) \bmod 2^{W-1}, \tag{4.3}$$

where $r^*(n) = 1$ if $r(n) = r_+$ and $r^*(n) = 0$ if $r(n) = r_-$. Here $\mu(\Psi, r)$ is a left bit-shift operation, shifting the bits in the associated binary sequence one place to the left and appending the new sample; the modulus operation is needed because of the finite memory. The full learning rule is then

$$\Psi(n+1) = \begin{cases} \mu(\Psi(n), r(n)) & \text{if} \quad a(n) = a_V \\ \Psi(n) & \text{else} \end{cases}, \tag{4.4}$$

and after each trial the organism's state is updated according to Equation (4.4). The next action is then selected using the averaging and choice functions as in Equations (4.1) and (4.2). Since $\mathbf{\Psi}_c(n)$ is constant after W samples from action a_c, the associated estimate $\hat{r}_c(n) = \mathcal{A}\left(\mathbf{\Psi}_c(n)\right)$ will also be constant and hereafter will be denoted as $\hat{r}_c$. In what follows I will often suppress the choice function's dependence on $\hat{r}_c$ and express the choice function as $\phi(n) = \mathcal{C}(\hat{r}_v(n))$, or equivalently $\phi(n) = \mathcal{C}\left(\mathcal{A}\left(\Psi(n)\right)\right)$.

Model dynamics as a semi-Markov process

Note, from Equation (4.4), that state transitions only occur when the variable option is chosen. Say that the system occupies state i when $\Psi(n) = i$, and let the random variables $S_1 < S_2 < S_3 < ...$ denote the trials on which state transitions occur, i.e. trials when the variable option is chosen. Let $Z_\eta, \eta \in \mathbb{Z}$, denote the state entered on trial S_η, i.e. $Z_\eta = \Psi(S_\eta) \in \{0,1,2,...2^{W-1}-1\}$. The stochastic process $\{\Psi(n), n \in \mathbb{Z}\}$ is then a discrete time *semi-Markov process* with a finite state space (Howard, 1971; Gallager, 1996). The sequence $\{Z_\eta, \eta \in \mathbb{Z}\}$ is called the *embedded Markov chain* for this semi-Markov process. For a semi-Markov process, the number of trials between state transitions in the embedded chain is a random variable.

The transition probabilities, $P_{ij} = \Pr(Z_\eta = j \mid Z_{\eta-1} = i)$, for the embedded chain are given by

$$P_{ij} = \begin{cases} p_+ & \text{if } j = \mu(i, r_+) \\ p_- & \text{if } j = \mu(i, r_-) \\ 0 & \text{else} \end{cases}. \tag{4.5}$$

for $i, j \in \{0,1,...,2^W - 1\}$. The transition probabilities for the embedded chain are determined solely by the variable option's reward probabilities, and the embedded Markov chain defined by Equation (4.5) is irreducible and positive recurrent. Hence there exists a stationary distribution for the embedded chain, and this distribution has steady state probabilities, π_i^e, that represent the long term probability of finding the system in the state i. Importantly, π_i^e is equivalent to the probability of obtaining the binary string associated with the integer i. In other words, if we let k_i be the number of ones in the binary string corresponding to the integer i, then the stationary probability of finding the system in state i is

$$\pi_i^e = p_+^{k_i} p_-^{W-k_i}. \tag{4.6}$$

However, the number of trials spent in state i before moving to another state is a random variable.

Define the holding time, T_η, as the number of trials needed before moving to another state, i.e. the interval between state transitions: $T_\eta = S_\eta - S_{\eta-1}$. The holding time can be thought of as the number of trials needed to obtain a single success (choose the variable option). The probability of choosing the variable option on the nth trial, $\phi(n) = \mathcal{C}\left(\mathcal{A}\left(\Psi(n)\right), \hat{r}_c\right)$, depends only of the current state, $\Psi(n)$. The holding time in each state is thus a geometrically distributed random variable with success probability $\phi_i = \mathcal{C}(\mathcal{A}(i), \hat{r}_c)$ and distribution

$$\Pr(T_\eta = t \mid Z_{\eta-1} = i) = (1-\phi_i)^{t-1}\phi_i . \tag{4.7}$$

If we assume that $\phi_i > 0$ for all $i \in \{0,1,...2^W - 1\}$, then the expected holding time in state i, $\overline{T_i}$, is the expected value of a geometric random variable

$$\overline{T_i} = E(T_\eta \mid Z_{\eta-1} = i) = \frac{1}{\phi_i} = \frac{1}{\mathcal{C}(\mathcal{A}(i), \hat{r}_c)} . \tag{4.8}$$

Recall that we want to predict θ, the expected proportion of choices for the variable option, as defined in Equation (3.3). Since N_n denotes the cumulative number of choices for the variable option as of the nth trial, it also represents the total number of state transitions for the semi-Markov process $\{\Psi(n), n \in \mathbb{Z}\}$ as of trial n. One of the most important results for semi-Markov processes (see e.g. Howard 1971) is that

$$\theta = \lim_{n\to\infty} \frac{N_n}{n} = \frac{1}{\sum_{i=0}^{2^W-1} \pi_i^e \overline{T_i}} , \tag{4.9}$$

where the π_i^e are the stationary probabilities for the embedded chain. Using Equation (4.8), and suppressing the dependence on $\hat{r}_c$ this becomes

$$\theta = \frac{1}{\sum_{i=0}^{2^W-1} \pi_i^e \frac{1}{\mathcal{C}(\mathcal{A}(i))}} . \tag{4.10}$$

The sum in the denominator is the expected value of a function, $1/\mathcal{C}(\mathcal{A}(.))$ of a random variable, $\Psi(n)$, whose asymptotic distribution is given by the embedded Markov chain, i.e.

$$\theta = \frac{1}{E_e\left(\frac{1}{\mathcal{C}\left(\mathcal{A}(\Psi(n))\right)}\right)}, \tag{4.11}$$

where E_e denotes the expected value with respect to the embedded chain.

The right hand side of Equation (4.11) is the harmonic mean (expected value), $H_e(.)$, of a function of a random variable with respect to the embedded Markov chain, so we can rewrite it as

$$\theta = H_e\left(\mathcal{C}\left(\mathcal{A}(\Psi(n))\right)\right). \tag{4.12}$$

Equation (4.12) says that the expected proportion of choices for the variable option is equal to the harmonic mean (harmonic expected value) of the probability of choosing the variable option in each memory state $i \in \{0,1,...2^W - 1\}$. Note from Equation (4.10) that states i with a small probability of selecting the variable option, $\mathcal{C}(\mathcal{A}(i)) \approx 0$, will substantially decrease the resulting value of θ. When the memory is in these 'bad' states (small ϕ_i), the organism will be less likely to choose the variable option, leading to longer holding times. The 'bad' memory states will persist for a longer period of time, thus generating the hot stove effect identified by Denrell and March. The organism effectively under samples the 'good' memory states and spends more time in the 'bad' memory states. As one might expect, this effect can generate risk aversion.

Consequences for risk sensitivity

To prove that the hot stove effect produces risk aversion we will need some additional restrictions on the choice and averaging functions. In addition to the definitions of proper and R-concave choice functions (Chapter 2); we will also need to

define two characteristics of the averaging functions, $\mathcal{A}:\mathbb{R}_+^W \to \mathbb{R}_+$. First, say that an averaging function is *unbiased* for a given BRSE if

$$\hat{r}_c = \mathcal{A}(\boldsymbol{\Psi}_c) = r_c \quad \text{and} \quad E_e\left(\mathcal{A}\left(\Psi(n)\right)\right) = \overline{r}_v \tag{4.13}$$

where again $E_e\left(\mathcal{A}\left(\Psi(n)\right)\right) = \sum_{i=0}^{2^W-1} \pi_i^e \mathcal{A}(i)$ is the expected value with respect to the embedded Markov chain. An averaging function is unbiased if the expected value of the averaging function across memory states is equal to the expected value of the associated reward distribution. Second, define a *commutative* averaging function as one for which there exist 2^{W-1} memory states i with 2^{W-1} associated but distinct memory states i' and non-negative constants, $K_i \in [0,1]$, such that

$$\mathcal{A}(i) = \overline{r}_v(1-K_i) \quad \text{and} \quad \mathcal{A}(i') = \overline{r}_v(1+K_i). \tag{4.14}$$

While this definition of a commutative averaging function may seem odd, all generalized linear combination (GLC, Chapter 2) averaging functions are commutative, and thus EWMAs and TWMAs are also commutative. The key point is that commutative averaging functions are symmetric about the mean value. I refer to this as commutative since, for a GLC averaging function, the binary sequence corresponding to i' is obtained by exchanging all of the ones and zeros in the binary sequence corresponding to i.

With these definitions in hand, we can proceed with the proof. From Equation (4.12), a DELA model will display risk prone behavior if

$$\theta = H_e\left(\mathcal{C}\left(\mathcal{A}(\Psi(n)), \hat{r}_c\right)\right) > \frac{1}{2}. \tag{4.15}$$

By condition $C.1$ from Chapter 2, $\mathcal{C}(\hat{r}_c, \hat{r}_c) = \frac{1}{2}$ for a proper choice function, and thus with an unbiased averaging function, $\mathcal{C}\left(E_e\left(\mathcal{A}(\Psi(n))\right), \hat{r}_c\right) = \frac{1}{2}$. Equation (4.15) can then be rewritten as

$$\theta = H_e\left(\mathcal{C}\left(\mathcal{A}(\Psi(n))\right)\right) > \mathcal{C}\left(E_e\left(\mathcal{A}(\Psi(n))\right)\right), \tag{4.16}$$

where the second choice function argument has been suppressed. Note the similarity between Equation (4.16) and Jensen's inequality. Jensen's inequality says that, for all convex functions $\mathcal{C}$ and random variables X, $E(\mathcal{C}(X)) \geq \mathcal{C}(E(X))$; Jensen's inequality

expresses a relationship between the arithmetic mean of a function of a random variable, $E(\mathcal{C}(X))$, and the function of the arithmetic mean, $\mathcal{C}(E(X))$. Analogously, we can write Equation (4.16) as

$$H_e\left(\mathcal{C}(X)\right) > \mathcal{C}\left(E_e(X)\right), \tag{4.17}$$

where $X = \mathcal{A}(\Psi(n))$ (a function of a random variable is also a random variable). Equation (4.17) replaces the arithmetic mean (expected value) on the left hand side of Jensen's inequality with the harmonic mean (expected value) and makes the inequality strict.

We can think of convexity as being defined by a relationship between the arithmetic mean of a function and a function of the arithmetic mean. Expanding on this notion, Niculescu (2003) introduced a generalized notion of convexity, MN convexity. If a continuous function, $\mathcal{C}$, is MN convex, then for any random variable X,

$$N\left(\mathcal{C}(X)\right) \geq \mathcal{C}\left(M(X)\right), \tag{4.18}$$

where M and N are generalized means. Using Niculescu's terminology, Equation (4.17) is the condition for strict AH convexity (arithmetic-harmonic). Thus only strictly AH convex choice functions can generate risk prone behavior. Importantly, the well known power means inequality (Bullen, 2003) says that

$$E(\mathcal{C}(X)) \geq H(\mathcal{C}(X)), \tag{4.19}$$

with equality if and only if $\mathcal{C}(x_1) = \mathcal{C}(x_2)$ for all x_1, x_2 in the domain of X. As a result, the condition for risk-prone behavior, Equation (4.16), is stronger than Jensen's inequality, and choice functions satisfying Jensen's inequality will not necessarily produce risk-prone behavior.

In the special case when the rewards are equiprobable, $p_+ = p_-$, Appendix 2 proves that all models combining R-concave choice functions with unbiased and commutative averaging functions will generate risk averse behavior. This proof applies to the models presented in Keaser et al. (2004), Niv et al. (2002), Bateson and Kacelnik (1995), and the last model in March et al. (1996). Denrell (2005) presents a similar result for a general model with an R-affine choice function and generalized linear combination

averaging function but deals with continuous reward distributions. Niv et al. (2002) also present a similar result but for a EWMA averaging function and concave choice function.

Forced trials

The BRSE, as treated above, presents the agent with a continuously repeated sequence of choice trials offering two different options. However, most experiments instead use a block procedure that presents a sequence of forced trials followed by a sequence of choice trials. For example, the startling experiment discussed in Chapter 3 utilized blocks of 8 forced trials followed by two choice trials. The introduction of forced trials should reduce the impact of the hot stove effect, as the agent will obtain information about the variable option independent of its actions on the choice trials.

To see how forced trials modify the dynamics, consider an experiment where each block consists of only two trials: one forced trial followed by one choice trial. On each forced trial the agent can only sample from the variable option, and on the choice trial both options are presented. In this case, the expected proportion of choices for the variable option is

$$\theta = E_e\left(\mathcal{C}\left(\mathcal{A}(\Psi(n))\right)\right), \tag{4.20}$$

and with an unbiased averaging function the condition for risk prone behavior is equivalent to Jensen's inequality. This experiment represents one extreme and the experiment with continuous choice trials represents another. Intermediate experimental schedules will produce intermediary behavior and the predicted proportion of choices for the variable option, θ, will be bounded

$$E_e\left(\mathcal{C}\left(\mathcal{A}(\Psi(n)), r_c\right)\right) \geq \theta \geq H_e\left(\mathcal{C}\left(\mathcal{A}(\Psi(n)), r_c\right)\right) \tag{4.21}$$

for these other schedules.

In practice, most experimental schedules fall between the two extremes of only choice trials and alternating forced/choice trials. Moreover, the rewards distributions from the forced trials are often modified so that the organism experiences the 'true' reward probabilities in each block (see Chapter 7). I cannot present analytic results for

these more complex experimental designs, but it stands to reason that decreasing the number of consecutive choice trials should decrease exhibited risk aversion. As the number of consecutive choice trials increases, the proportion of choices for the variable option should decrease.

Discussion

This chapter showed how the hot stove effect will generate risk aversion for a wide class of DELA models confronted by a BRSE with equiprobable rewards. Specifically, any DELA model with a proper R-concave choice function and an unbiased and commutative averaging function will display risk aversion. This phenomenon arises due to the dynamics of the experiential learning process: since the agent only gains information by choosing the variable option, runs of bad luck have a disproportionate impact.

Experiments with forced trials eliminate this effect by providing the agents with information independent of their choices. As the number of consecutive choice trials decreases, the impact of the hot stove effect will also decrease and the proportion of choices for the variable option should increase. This prediction can be tested experimentally by varying the number of choice trials in a block and observing the impact on risk sensitive preferences. Moreover, some of the ambiguity in the risk sensitive foraging literature could conceivably be due to differences in methodology. A meta-analysis of published experiments examining the relationship between risk aversion and the number of consecutive choice trials in a block might be useful in this respect. If organisms use something like a DELA model, the magnitude of expressed risk preferences will depend on the experimental structure, and experimental results should be reevaluated in light of this finding.

If a model, faced with continuous choice trials, is to generate genuinely risk-prone behavior, it must violate one of the assumptions used to prove the hot stove effect. The next chapter presents a DELA model that can generate risk prone behavior. This model is able to do so by using a biased and non-commutative averaging function.

Chapter 6 A risk prone model

Couvillon and Bitterman (1991) introduced a simple model for how honeybees make choices. Inspired by EWMA type models from mathematical psychology, this model is equivalent to a DELA model, albeit one that violates some of the assumptions in the previous chapter. Importantly, this model can, with appropriate parameter settings, generate genuinely risk-prone behavior, largely as a result of its novel averaging function. For most parameter settings the averaging function is not unbiased or commutative, and thus this model is not subject to the hot stove effect proof in Chapter 5.

Couvillon and Bitterman applied their model to a series of foraging experiments with honeybees and found that the model was able to fit the results satisfactorily. More recently, Shapiro (2000) and Shapiro et al. (2001) applied the model to a series of risk sensitivity experiments with honeybees and similarly found that the model could fit the results well. All three studies simulated model dynamics on a computer and obtained the best fitting parameters using a factorial search procedure.

In this chapter, I will derive some analytic results for the Couvillon and Bitterman (CB) model and discuss how this model can generate risk prone behavior. To derive these results, I will treat the model dynamics as an iterated function system with probabilities (Barnsley, 1988). The next section discusses the risk sensitivity experiments to which the CB model has been applied, and the following section presents the CB model. Finally I will introduce the iterated function system formalism and use it to derive some formulas for computing model predictions.

The experiments

The experiments examined in Shapiro (2000) and Shapiro et al. (2001) are mainly modified versions of the BRSE (Table 6.1). Without going into the details of the experimental protocol, it suffices to say that the experiments utilized a discrete trials procedure presenting bumblebees with a choice between two foraging options providing

Table 6.1 Reward schedules in each Honeybee experiment. The first 8 experiments come from Shapiro (2000), while the last 5 come from Shapiro et al. (2001). For options providing more than one possible reward, both reward values are reported along with the probability of the larger reward. All of the experiments utilized a block structure and the number of blocks is presented for each experiment. Two of the experiments reversed the reward distributions midway through the experiment, and the number of blocks pre and post reversal are then reported. Note that experiments 2001-4A and B did not have any forced trials, while all of the other experiments had one choice trial followed by one forced trial each block.

Experiment	Constant Option		Variable Option			Reverse?	# Blocks
	Conc (%)	Amount (ml)	Conc (%)	Amount (ml)	Prob Big		
Prelim-C	40	10	20	10	1.00	N	40
Prelim-A	40	20	40	5	1.00	N	40
Prelim-P	40	10	40	10,0	0.50	N	40
Con1	20	10	40,0	10	0.50	N	40
Con4	15	10	60,0	10	0.25	N	40
Con5	40	10	60,20	10	0.50	N	40
Amt1	40	5	40	20,0	0.25	N	40
Amt2	40	5	40	30,0	0.17	N	40
2001-1	40	5	40	30,0	0.17	Y	24,48
2001-2	40	5	40	30,0	0.33	N	40
2001-3	15	10	60,0	10	0.50	N	40
2001-4A	40	5	40	30,0	0.17	N	72
2001-4B						Y	24,48

different nectar rewards with different distributions. The nectar distributions were associated with two different scents, and the bumblebees learned to associate the scents with their respective distributions. All of the experiments in Shapiro (2000) and all but two from Shapiro et al. (2001) utilized a block structure with a single choice trial followed by a single forced trial. The option that was not chosen on the choice trial was presented on the forced trial, ensuring that the bee received one sample from each food source in each block. Two of these experiments (2001-1, 2001-4B) reversed the reward distributions midway through the experiment: the scent associated with the constant reward distribution in the first portion of these experiments was associated with the variable distribution in the second portion. The last experiments from Shapiro et al. (2001) did not utilize forced trials, instead presenting an unbroken sequence of consecutive choice trials.

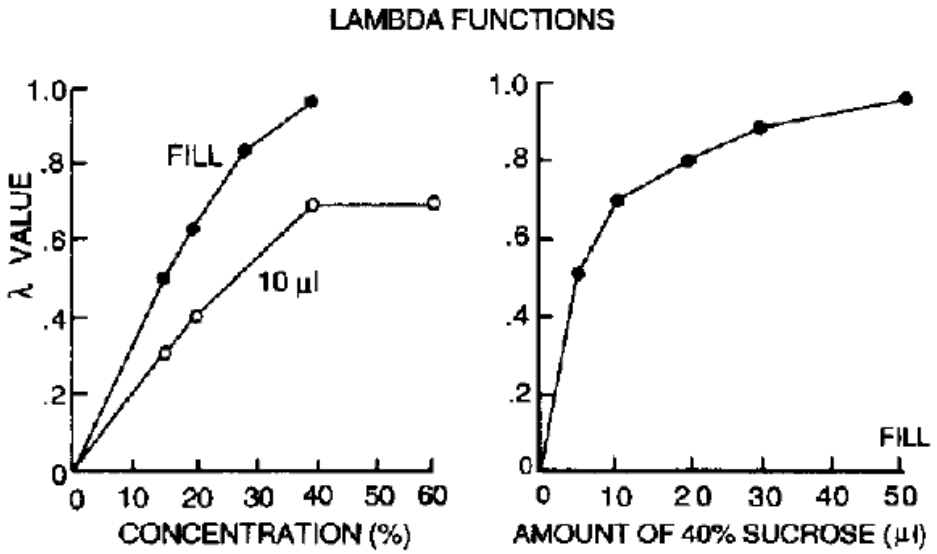


Figure 6.1- The utility values derived by Shapiro (2000). The left hand graph shows the currency values as a function of the sucrose concentration at two different volumes (Fill = 50 μl). The right hand graph shows the currency values as a function of nectar volume. Note that all functions are concave. (Reprinted with permission from Shapiro 2000).

Nectar rewards were characterized by a sucrose concentration and a nectar volume. Bees prefer both higher sucrose concentrations and larger volumes, but it is not immediately clear how the bees should integrate these relative reward dimensions. Shapiro (2000) introduced utility functions to translate the physical rewards into subjective rewards (although Shapiro did not phrase it as such). Rather than try and fit a parameterized function, Shapiro introduced individual parameters for each concentration/volume combination in the experiments. Shapiro thus introduced 9 additional parameters for a total of 13 (9 + two choice function parameters + two averaging function parameters). Shapiro searched the parameter space factorially, and obtained plausible values for the utility function (Figure 6.1). These utility values were also used by Shapiro et al. (2001) in their analysis.

Note that the utility values derived by Shapiro are concave with respect to both sucrose concentration and nectar volume. Thus we might expect some risk aversion based on the non-linear utility function alone. Indeed, in all of the experiments but one a preference for the constant option is predicted on the basis of the average utility provided alone (Table 6.2). In the terminology of Chapter 4, this is apparent risk sensitivity. The

Table 6.2 The utility values for each experiment as presented in Shapiro (2000). For the variable option, the mean value of the rewards provided, computed using the utility values, is also shown. If the mean is less than the utility value of the constant option, preference for the constant option is expected (these values are in bold). Also presented is the observed proportion of choices for the constant option over the last 4 choice trials in each experiment (these values were estimated from the graphs in Shapiro, 2000 and Shapiro et al., 2001). The asterisks denote experiments with reward reversals, and thus the value reported is actually the proportion of choices for the option with the currently variable distribution.

Experiment	Constant	Variable				Observed θ
	r_c	r_+	r_-	p_+	$\overline{r}_v$	
Prelim-C	0.7	0.4	0.4	1.00	**0.40**	**0.75**
Prelim-A	0.8	0.5	0.5	1.00	**0.50**	**0.6**
Prelim-P	0.7	0.7	0.0	0.50	**0.35**	**0.7**
Con1	0.4	0.7	0.0	0.50	**0.35**	0.49
Con4	0.3	0.7	0.0	0.25	**0.18**	**0.7**
Con5	0.7	0.7	0.4	0.50	**0.55**	**0.6**
Amt1	0.5	0.8	0.0	0.25	**0.20**	**0.6**
Amt2	0.5	0.88	0.0	0.17	**0.15**	**0.75**
2001-1	0.5	0.88	0.0	0.17	**0.15**	**0.1***
2001-2	0.5	0.88	0.0	0.33	**0.29**	**0.7**
2001-3	0.3	0.7	0.0	0.50	0.35	0.4
2001-4A*	0.5	0.88	0.0	0.17	**0.15**	**0.75**
2001-4B*						**0.35***

CB model can also generate genuinely risk sensitivity due to the dynamics of the averaging function.

The CB model

The most unique feature of the CB model is its averaging function. The model utilizes a version of the standard EWMA averaging function but introduces two memory coefficients, m_+ and m_-. When the reward received from an action is larger than the current estimate for that action, the memory coefficient m_+ is used, and when the reward received is smaller than the current estimate, the coefficient m_- is used:

$$\hat{r}_i(n+1) = \begin{cases} (1-m_+)r(n) + m_+\hat{r}_i(n) & \text{if } a(n) = a_i \text{ and } r(n) > \hat{r}_i(n) \\ (1-m_-)r(n) + m_-\hat{r}_i(n) & \text{if } a(n) = a_i \text{ and } r(n) < \hat{r}_i(n) \\ \hat{r}_i(n) & \text{if } a(n) \neq a_i \end{cases}. \qquad (5.1)$$

In the special case where $m_+ = m_-$, Equation (5.1) is equivalent to the EWMA averaging function (i.e. Equation (2.6)).

The CB choice function (presented in Shapiro, 2000) is given by

$$\mathcal{C}(x,y) = \begin{cases} .5 + s\left|\frac{x-y}{x+y}\right|^k & \text{if } x > y \\ .5 - s\left|\frac{x-y}{x+y}\right|^k & \text{if } x < y \end{cases}. \tag{5.2}$$

This choice function is ratio-based, but it is not proper with the parameter values used in the published papers ($s = .625, k = .75$). With these parameters, the choice function is not necessarily confined to $[0,1]$. As long as there are only two possible actions, the function can be confined to this interval by simply taking 0 or 1 when $\mathcal{C}(x,y)$ is less than zero or greater than one respectively. However, this choice function can also be approximated fairly well by a modified Boltzmann choice function with $\gamma = 4.586$ (obtained using logistic regression).

On the BRSE the updating rule for the CB averaging function can be simplified asymptotically. Consider the variable option which provides either a big reward r_+ or a small reward r_-. Whatever the starting value for the associated estimator, $\hat{r}_v(0)$, the value of $\hat{r}_v(n)$ will eventually be confined to $[r_-, r_+]$ after enough samples from the variable option. As a result, it will eventually hold that $r_- \le \hat{r}_v(n) \le r_+$ for large enough n. Thus asymptotically the CB averaging function simplifies to

$$\hat{r}_v(n+1) = \begin{cases} (1-m_+)r_+ + m_+\hat{r}_v(n) & \text{if } a(n) = a_v \text{ and } r(n) = r_+ \\ (1-m_-)r_- + m_-\hat{r}_v(n) & \text{if } a(n) = a_v \text{ and } r(n) = r_- \\ \hat{r}_v(n) & \text{if } a(n) \neq a_v \end{cases}. \tag{5.3}$$

Unlike in Equation (5.1), here each memory coefficient is always associated with the same reward independent of the current value of the estimator.

In order to analyze the risk sensitive behavior of this model using the results from Chapter 5, we need to be able to compute the asymptotic distribution of the estimator, $\hat{r}_v(n)$. To do so, I will need some formalism and results from iterated function systems.

Iterated function systems

An iterated function system (IFS) is a dynamical system defined by a set of N functions that map a metric space to itself and N associated probabilities that define the probability of applying each map (Barnsley, 1987). The evolution of the system is determined by random application of the different maps, and proceeds in discrete time steps. Let X be a compact metric space and let $f_i : X \to X$ $(i = 1,...,N)$ be functions acting on the metric space. Associated with each function is a probability, p_i $(i = 1,...,N)$; in general these probabilities can depend on the current state of the system, in which case they are functions $p_i : X \to [0,1]$. If the probabilities depend on the current state, the IFS has *place dependent* probabilities, otherwise the probabilities are *place independent*. The IFS is defined by the set $\{X, f_i, p_i, i = 1,...,N\}$.

If the mapping functions, f_i, are contractive mappings, the IFS is called *hyperbolic*. The following discussion will deal exclusively with contractive affine maps acting on the positive real numbers $\mathbb{R}_+$. Thus we will only be considering functions $f_i : \mathbb{R}_+ \to \mathbb{R}_+$ of the form

$$f_i(x) = b_i + m_i x, \tag{5.4}$$

where $b_i \geq 0$ and $0 \leq m_i < 1$ are constants associated with each map. Contractive affine functions have attractive fixed points, $\tilde{x}_i$, where $\tilde{x}_i = f_i(\tilde{x}_i) = \dfrac{b_i}{1-m_i}$. For any given set of affine transformations, f_i $(i = 1,...,N)$, there will exist two transformations $f_{\min}, f_{\max}$ with associated fixed points $\tilde{x}_{\min}, \tilde{x}_{\max}$ such that $\tilde{x}_{\min} \leq \tilde{x}_i \leq \tilde{x}_{\max}$ for all i. Given any initial starting position $x(0) \in \mathbb{R}_+$ this IFS will eventually end up confined to the region $[x_{\min}, x_{\max}]$ and will never leave.

For any hyperbolic IFS such as the one above, there are several well established results (see Barnsley, 1993 and Slomczynski et al. 2000 for reviews). Most importantly,

there exists a unique invariant probability measure μ corresponding to the asymptotic distribution for the system. This distribution is attractive, and the system will converge to this distribution independent of the starting state of the system. In order to compute the asymptotic expected value of a continuous function $g(x)$ on an IFS, the function must be integrated over this invariant measure:

$$E\left(g(x)\right) = \int_X g(x) d\mu(x). \tag{5.5}$$

This integral can be computed in at least two different ways. The random iteration algorithm (Barnsley, 1993) chooses some starting value for the state of the system, $x_0 \in X$, and then generates a random sequence of values from the IFS, $\{z_1 = x_0, z_2, ..., z_n\}$. The integral in Equation (5.5) is then equivalent to

$$\int_X g(x) d\mu(x) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} g(z_i). \tag{5.6}$$

Thus the asymptotic expected value of the function can be computed by simply generating a sequence of values from the IFS and computing the mean of the function over the sequence. Alternatively, the deterministic algorithm (Barnsley, 1993; Hepting et al., 1991; Edalat, 1996) computes the exact distribution explicitly on each step by summing over all possible states for the system. Given an initial state, $x_0 \in X$, after n steps there are N^n possible states for the system (N possible operators on each step). Expected values can be computed by summing over all permutations of these operators and taking the limit:

$$\int_X g(x) d\mu(x) = \lim_{n\to\infty} \sum_{i_1, i_2, ... i_n = 1}^{N} g\left(f_{i_1} f_{i2} ... f_{i_n}(x_0)\right) p_{i_1} p_{i_2} ... p_{i_n}, \tag{5.7}$$

where the sum is taken over all permutations of the operators. Note that Equation (5.7) justifies the finite memory approximation to recursive averaging functions introduced in Chapter 2 and used in the proof of the hot stove effect.

The CB model as an IFS

In general, the CB model defines an IFS with place dependent probabilities, but asymptotically the probabilities can be treated as place-independent. The estimator for the constant side converges asymptotically to the value of the reward delivered, $\hat{r}_c(n) \to r_c$, so we can focus on the dynamics of the estimator for the variable option, $\hat{r}_v(n)$. Different memory coefficients are used depending on the relationship between the current value of this estimator and the value of the reward received. Each memory coefficient and reward combination represents a different mapping function for the associated IFS. These different combinations are shown in Figure 6.2; there are four possible combinations and thus four maps, f_i. The probability of each of the different maps depends on the current state of the system, however, and these probabilities can go to zero.

As discussed above, we know the system will eventually enter the range $[r_-, r_+]$ and never leave. In this range, there are only two possible maps,

$$\begin{aligned} f_-(\hat{r}_v) &= (1-m_-)r_- + m_-\hat{r}_v \\ f_+(\hat{r}_v) &= (1-m_+)r_+ + m_+\hat{r}_v \end{aligned}, \tag{5.8}$$

and the probability of each map is equal to the probability of receiving the associated reward: p_+, p_-. Thus the asymptotic dynamics for this system are given by an IFS with place independent probabilities defined as $\{\{[r_-, r_+]; f_-, f_+; p_-, p_+\}$.

On the BRSE, we are interested in the asymptotic expected proportion of choices for the variable option, θ, in the special case where $\overline{r_v} = r_c$ (assuming a linear utility function). From Chapter 5, we know that, when each block of trials has a single choice and single forced trial, θ is equal to the arithmetic mean, $E\left(\mathcal{C}(\hat{r}_v, r_c)\right)$, but when the experiment provides uninterrupted choice trials, θ is equal to the harmonic mean, $H\left(\mathcal{C}(\hat{r}_v, r_c)\right)$. In the former case, the expected value is computed as

$$E\left(\mathcal{C}(\hat{r}_v, r_c)\right) = \int_X \mathcal{C}(\hat{r}_v, r_c) d\mu(\hat{r}_v). \tag{5.9}$$

and in the latter as

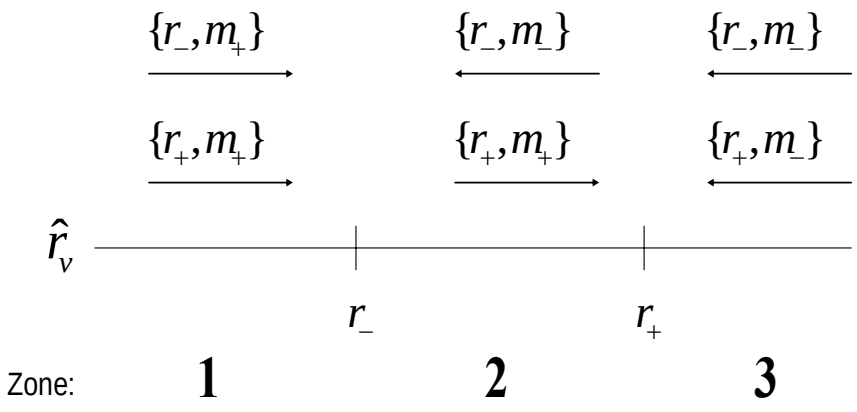


Figure 6.2. Possible combinations of reward and memory coefficient for the CB model. Shown as a function of the current value of the estimator for the variable option, $\hat{r}_v$. The arrows show the direction which each map will move the system. In each zone there are two possible maps. In zone 1 both maps increase the value of $\hat{r}_v$, while in zone 3 both maps decrease the value. In zone 2, the maps move in opposite direction.

$$H\left(\mathcal{C}(\hat{r}_v, r_c)\right) = \frac{1}{E\left(\frac{1}{\mathcal{C}(\hat{r}_v, r_c)}\right)} = \frac{1}{\int_X \frac{1}{\mathcal{C}(\hat{r}_v, r_c)} d\mu(\hat{r}_v)}. \tag{5.10}$$

These integrals can be computed directly using either the deterministic or random iteration algorithm presented in Equations (5.6) and (5.7). However, we could also approximate them naively using $\mathcal{C}(E(\hat{r}_v), r_c)$. Appendix 3, proves that

$$E(\hat{r}_v) = Mr_+ + (1-M)r_- \tag{5.11}$$

where

$$M = \frac{1}{1+\frac{p_-(1-m_-)}{p_+(1-m_+)}}. \tag{5.12}$$

I will refer to M as the memory bias. In the special case of equiprobable rewards, the memory bias simplifies to

$$M = \frac{1}{1+\frac{1-m_-}{1-m_+}}. \tag{5.13}$$

As a direct result of Equation (5.13), whenever $m_+ < m_-$ it follows that $M > \frac{1}{2}$, $E(\hat{r}_v) > \overline{r}_v = r_c$, and hence that $\mathcal{C}(E(\hat{r}_v), r_c) \geq \frac{1}{2}$. Similarly, the converse is true when $m_+ > m_-$, and when $m_+ = m_-$ we have the unbiased EWMA averaging function with $\mathcal{C}(E(\hat{r}_v), r_c) = \frac{1}{2}$. Thus a memory bias greater than one half will increase risk prone behavior, but a memory bias of less than one half will increase risk averse behavior.

Choosing $m_+ \neq m_-$ effectively weights either the big reward or the small reward more strongly, leading to a biased estimator. The CB model can generate risk preferences because it explicitly weights one of the two possible rewards more strongly than the other. Weighting the small reward more strongly will lead to risk aversion while more weight on the big reward will lead to risk prone behavior. In their initial paper, CB found that $m_+ = m_-$ (i.e. the standard EWMA) provided the best fit to their bumblebee data, but the two more recent studies found $m_+ < m_-$. Shapiro (2000) reported $m_+ = .96, m_- = .98$ and Shapiro et al (2001) reported $m_+ = .92, m_- = .94$ as the best fitting parameterization. Either parameterization will produce a memory bias of more than one half and increase the magnitude of risk prone behavior.

On the experiments with one choice and one forced trial per block, there will be no hot stove effect. In this case the CB model generates risk averse behavior due to the concave utility function, but the magnitude of the risk aversion is reduced by the biased averaging function ($M > \frac{1}{2}$). For these experiments, the algorithms presented above can be used to compute $\theta = E\left(\mathcal{C}(\hat{r}_v, \hat{r}_c)\right)$. On the lone experiment with repeated choice trials and no forced trials (Experiment 2001-4), the hot stove effect will hold, and the harmonic mean of the choice function must be computed over the attractive probability measure. Table 6.3 presents the results from each algorithm on the twelve experiments that did not involve a reversal of the reward distributions. The results from each algorithm are similar and provide a reasonable fit to the experimentally observed data. Note that, at least for the parameter values used, the direct and simple computation of $\mathcal{C}(E(\hat{r}_v), r_c)$ with Equation (5.11) is largely equivalent to the two more complex computation methods

Table 6.3 Empirical results and CB model predictions for each honeybee experiment. The second column was computed exactly using Equation (5.11). Column 3 presents the results from the deterministic algorithm run with $\hat{r}_v(0) = E(\hat{r}_v)$ and up to $n = 18$. Column 4 presents the results from the random iteration algorithm over 100,000 iterations. Observed values were estimated from the published papers, and the model parameters were taken from Shapiro et al. (2001): $m_+ = .92, m_- = .94, s = .625, k = .75$. The predictions from the three methods are quite similar, at least with the parameter values used here.

Experiment	Observed	$\mathcal{C}(E(\hat{r}_v), r_c)$	Deterministic	Random Iteration
Prelim-C	**0.75**	0.74	0.74	0.74
Prelim-A	**0.6**	0.71	0.71	0.71
Prelim-P	**0.7**	0.74	0.74	0.74
Con1	0.49	0.5	0.51	0.51
Con4	**0.7**	0.66	0.66	0.66
Con5	**0.6**	0.61	0.61	0.61
Amt1	**0.6**	0.78	0.78	0.78
Amt2	**0.75**	0.85	0.85	0.85
2001-2	**0.7**	0.67	0.67	0.67
2001-3	0.4	0.35	0.37	0.37
2001-4A	**0.75**	0.85	0.84	0.85

in Equations (5.6) and (5.7). More exact approximations can be computed using higher order moments of $\hat{r}_J$. Appendix 3 also presents a formula for the variance of $\hat{r}_J$, and this value could be used in a second order Taylor series to approximate $\theta = E\left(\mathcal{C}(\hat{r}_v, \hat{r}_c)\right)$.

These results suggest some simpler methods for fitting model parameters: rather than simulate the model or use the (eventually) exact computation algorithms, quickly compute predictions using moments of the asymptotic distribution. The results of this faster search can then be used to narrow the parameter space for more intensive exploration using simulation or exact computation methods. Of course, this approach will only work if the experiments provide asymptotic data (i.e. the experiments are long enough and the choice proportions reach an asymptote). In such cases, however, the approach should be quite useful since computing the exact distribution over the course of the learning process is more computationally involved

Short term model predictions

Each of the published papers focused the acquisition of preferences over the course of an experiment and not on the asymptotic preferences. In fitting the model, the authors simulated the entire experimental run and compared the predicted and observed choice proportions at fixed intervals throughout the experimental period. In order to predict the proportion of choices throughout an experiment (or deal with reward distribution reversals), the dynamics of the estimator for the constant option must also be considered. To derive predictions for the short term dynamics, the system can be treated as an IFS with place dependent probabilities defined on a bivariate metric space, $X = \mathbb{R}_+ \times \mathbb{R}_+$, corresponding to the possible values for the estimators $\hat{r}_v(n)$ and $\hat{r}_c(n)$. The experimental structure of one choice trial followed by one forced trial ensures that each estimator is updated exactly once in each block. There are two possible maps for $\hat{r}_c(n)$ corresponding to $\hat{r}_c(n) < r_c$ and $\hat{r}_c(n) > r_c$. In combination with the four possible maps for $\hat{r}_v(n)$ (as depicted in Figure 2), there are 8 possible maps, $(\hat{r}_c, \hat{r}_v) = f_i(\hat{r}_c, \hat{r}_v)$, in all.

Given some initial state $\left(\hat{r}_c(0), \hat{r}_v(0)\right)$, the deterministic algorithm can be used to compute the exact distribution for the system on each trial. Similarly, the deterministic algorithm can be used to compute the exact distribution given any initial probability kernel on $(\hat{r}_c, \hat{r}_v)$. The number of computations increases quickly as a function of the number of trials ($\geq O(2^n)$), and exact computation quickly becomes infeasible. However, these computations can be simplified by imposing a discrete grid on the state space and approximating the distribution. Given a computed distribution on each trial, more precise model selection and inferential methods become feasible.

Discussion

The Couvillon-Bitterman model is one of the few published models that can generate genuinely risk prone behavior. This ability is due to its novel averaging

function which is neither unbiased nor commutative. Depending on the parameter values used, the CB averaging function can generate either risk-prone or risk-averse behavior. In the former case the averaging function places more weight on the larger reward (memory bias > 0.5) and in the latter it weights the smaller reward more strongly (memory bias < 0.5), thus biasing the estimates. This model is able to fit the data from the honeybee experiments quite well.

Although the honeybees on the whole display risk averse behavior, the averaging function parameter values that fit the honeybee data best actually decrease the magnitude of the risk aversion (memory bias for the parameters used is > 0.5). The risk aversion is due to the non-linear utility function; with a linear utility function the CB model would actually be risk prone using the published parameter values.

While not explored here, the possibility remains that a more traditional averaging function could fit the honeybee foraging data equally if different utility and choice functions were used. The next chapter, presents a more formal model selection using a larger set of data from some foraging experiments with starlings.

Chapter 7 Starling experiments

Bateson and Kacelnik, together and with other authors, have conducted an extensive series of bird in a box experiments with starlings. Many of these experiments used similar methodologies but differed in experimental details. Several experiments (BK, 1995a; BK, 1996; BK, 1997) utilized the titration procedure from Chapter 3, and presented the birds with a transient environment. Other experiments did not titrate the reward distributions (BK, 1995b; Schuck-Paim & Kacelnik, 2002; Bateson, 2002), and thus presented stationary environments. This chapter evaluates the performance of some simple DELA models, using the published data from these two sets of experiments.

Unfortunately, I do not have access to the full data sets from these experiments and must make do with the published summary statistics. This fact limits the scope of any possible model selection. Thus my goal here is not to find a 'best' model but instead to explore the behavior of some simple 2 and 3 parameter models in order to diagnose discrepancies between model predictions and the observed results. The main emergent pattern is that it is difficult for the same choice function to fit the data from the transient and stationary experiments simultaneously. This result, which was also discussed in Chapter 3, suggests a major possible shortcoming of DELA models.

The experiments

All of the experiments utilized the same basic bird in a box procedure presenting a choice between two foraging options providing food rewards. Rewards were characterized by both a waiting time and a reward size, and there were two types of experiments. The first type of experiment offered a choice between a high variance option and a constant or low variance option, with variability in either the waiting times or the reward sizes but not both. The response variable from these experiments was a proportion (proportion of choices for the more variable option), so I will refer to them as P-experiments (Table 7.1). The other type of experiments, the T-experiments (Table

7.2), offered a choice between a standard option and an adjusting option and used a titration procedure to modify the reward provided by the adjusting option. Either the reward size or the waiting time could be titrated, and the response variable was the average value of the titrating dimension. The P-experiments presented a stationary environment, and the T-experiments presented a transient one.

Both types of experiments were organized into blocks of forced and choice trials, but the relative number of forced and choice trials varied between experiments. In all experiments, each option was presented on exactly half of the forced trial in each block, ensuring that the birds received equal exposure to each option, but the order of presentation was chosen randomly. Similarly, all of the experiments structured the forced trials so that the birds were exposed to the 'true' reward distributions in each block: each possible reward was presented with a frequency proportional to its probability. For example, in experiment T-Time2 the standard option provided two equiprobable rewards, there were 8 forced trials per block, and on 4 of these forced trials the variable option was presented (Figure 7.1). On exactly two of those 4 forced trials, the bird received a big reward and on the other two the bird received a small reward, with the order of presentation chosen randomly. This method ensured that the birds experienced the 'true' distributions in each block but also reduced the effective variance in the rewards from the variable option, eliminating the possibility of a block presenting all/none big rewards.

In lieu of access to the original experimental data, I use the reported summary statistics as response variables. In the P-experiments (Table 7.3) the response variable was either the mean or median proportion of choices for the variable option computed across all the birds in the experiment. Similarly, in the T-experiments (Table 7.4), the response variable was the mean or median value of the adjusting option across all birds. Note that the response variables were not always reported numerically in the text, in which cases I estimated the values from published graphs.

Table 7.1 Characteristics of the12 P-experiments. Side V is the high variance option and provides a big reward, r_+', with probability, p_+, and a small reward, r_-' with probability, $1-p_+$; w is the waiting time and s is the reward size. Side C is the low variance or constant option. For the block structure, an F denotes a forced trial and a C denotes a choice trial. In the Bateson 2002 experiments, each food item was a .045 g starling pellet, while in all others each reward was ~.012g of turkey starter crumbs. The Bateson 2002 experiments had 8 birds per experiment, while all others used 6 birds.

Source	Experiment	Side V					Side C					Block Structure
		r_-'		r_+'		p_+	r_-'		r_+'		p_+	
		w	s	w	s		w	s	w	s		
B & K 1995B	P_Amt1	20	3	20	7	0.50			20	5	1	8 F- 1 C
	P_Amt1_B											1C
	P_Time1	60.5	5	2.5	5	0.50			20	5	1	8 F- 1 C
	P_Time1_B											1C
SP & K 2002	P_Time2	12	5	28	5	0.50			20	5	1	4 F- 1C
	P_Time3	4	5	36	5	0.50			20	5	1	4 F- 1C
	P_Time4	4	5	36	5	0.50	28	5	12	5	0.50	4 F- 1C
B 2002	P_Amt2	5	2	5	8	0.33			5	4	1	18 F- 18 C
	P_Amt3	5	1	5	10	0.33			5	4	1	18 F- 18 C
	P_Amt4	5	1	5	10	0.33	5	2	5	8	0.33	18 F- 18 C
B & K 1997	P_RF_1	18	1	3	1	0.50			18	1	1	2 F- 1 C
	P_RY_1	18	1	3	1	0.50			18	1	1	2 F- 1 C

Table 7.2 Characteristics of the 12 T experiments. Side S is the standard side and side J is the adjusting side. The dimension with the * is the titrating dimension, and the value shown is the value at the start of the titrations. Note that in the experiments with **, the waiting time started with the ending value from the previous experiment. All of these experiments used 6 birds, and each food item represented ~.012g of turkey crumb.

Source	Experiment	Side S					Side J		Block Structure
		r_-'		r_+'		p_+	w	s	
		w	s	w	s				
B & K 1995A	T_Amt1			5	3	1	5	9*	8 F- 2 C
	T_Amt2			5	9	1	5	3*	8 F- 2 C
B & K 1996	T_Amt3			20	5	1	20	15*	8 F- 2 C
	T_Time1			20	5	1	5*	5	8 F- 2 C
	T_Time2	60.5	5	2.5	5	0.50	20*	5	8 F- 2 C
	T_Time3	20	2	5	2	0.50	20	2*	8 F- 2 C
B & K 1997	T_RF_2	18	1	3	1	0.50	18*	1	2 F- 1 C
	T_RF_3	18	1	3	1	0.50	**	1	2 F- 1 C
	T_RF_4	18	2	3	2	0.50	**	2	2 F- 1 C
	T_RY_2	18	1	3	1	0.50	18*	1	2 F- 1 C
	T_RY_3	18	1	3	1	0.50	**	1	2 F- 1 C
	T_RY_4	18	2	3	2	0.50	**	2	2 F- 1 C

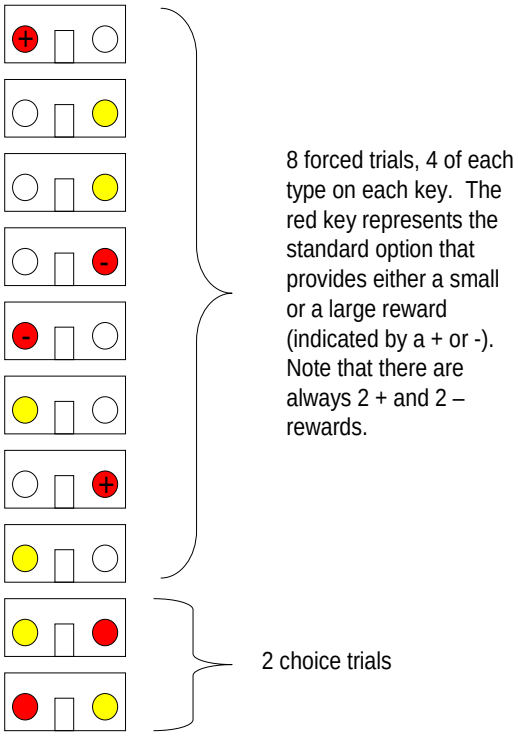


Figure 7.1. A possible block of trials. The experimental protocol from experiment T_Time2 is shown, but all of the experiments used similar structures. (After Bateson and Kacelnik, 1996).

Each experiment was divided into three parts. A pre-training period exposed the birds to the experimental apparatus and some representative reward values. After pre-training, the birds were exposed to the actual reward schedules as summarized in Table 7.1 and Table 7.2. However, since the authors were interested in the asymptotic preferences, the data from the first few blocks was discarded. Thus each experiment had a brief training period followed by a testing period. The birds were exposed to the actual reward schedules in both periods, but only the data from the testing phase was used to derive the summary statistics presented in Table 7.3 and Table 7.4. The number of data points used to compute the summary statistics, N, differed between experiments. In the P- experiments, N was equivalent to the number of choice trials in the testing phase of the experiment. In the T-experiments N denoted the number of possible titrations during the testing phase.

Table 7.3 Results from the P-experiments. In all cases, the response variable is the proportion of choices for the more variable option. N denotes the number of choice trials in the testing phase of the experiment. In the comments, RP stands for risk prone and RA for risk averse.

Source	Experiment	Statistic	Observed value	N	Comments
B & K 1995B	P_Amt1	mean	0.54	216	slightly RP
	P_Amt1_B	mean	0.475	216	slightly RA, Hot stove Effect
	P_Time1	mean	0.98	216	RP
	P_Time1_B	mean	0.97	216	RP
SP & K 2002	P_Time2	mean	0.66	432	RP
	P_Time3	mean	0.89	432	RP
	P_Time4	mean	0.85	432	RP
B 2002	P_Amt2	median	0.53	720	slightly RP
	P_Amt3	median	0.48	720	slightly RA
	P_Amt4	median	0.4	720	RA
B & K 1997	P_RF_1	median	0.98	180	RP
	P_RY_1	median	0.96	180	RP

Table 7.4 Results from the T-experiments. The response variable is the average value of the titrating options. N denotes the number of blocks in the testing phase. In the comments, TB stands for titration bias. Note that in the BK 1997 experiments, the titrating value is larger in the risk-free treatments than in the risky treatments, suggesting that the birds preferred the risky option.

Source	Experiment	Statistic	Observed value	N	Comments
B & K 1995A	T_Amt1	mean	3.97	2247	TB
	T_Amt2	mean	12.1	1613	TB
B & K 1996	T_Amt3	mean	5.22	600	small TB
	T_Time1	mean	21.77	600	small TB
	T_Time2	mean	5.6	600	
	T_Time3	mean	5.25**	600	
B & K 1997	T_RF_2	median	7.56	1620	Titrating waiting time is Larger in the RF treatment
	T_RF_3	median	7.5	60	
	T_RF_4	median	7	60	
	T_RY_2	median	6.34	1620	
	T_RY_3	median	5.2	60	
	T_RY_4	median	4.9	60	

There are some consistent patterns in the results across experiments. The P-experiments show clear risk preference with respect to variability in waiting time and a weaker tendency towards risk aversion with respect to variability in amount. The results from the T-experiments are slightly more complicated, but there is evidence of a titration bias in several of the experiments. Many of the experiments had unique features, so I now briefly summarize the procedures from each publication. See the individual publications for further details.

Bateson and Kacelnik 1995A

These are the titration experiments discussed in Chapter 3. Only the titrating (stage 2) results are considered here. Experiment T_Amt1 used the smaller 3 pellet standard, while T_Amt2 used the larger standard. In both experiments the mean value of the adjusting option displayed a titration bias and was larger than the value of the standard (Table 7.4). The CV of the adjusting option was similar in both experiments but slightly larger in T_Amt2.

Bateson and Kacelnik 1995B

This paper presented four different versions of the BRSE. Two of the experiments, Amt1 and Amt1_B, introduced variability into the reward sizes, while the other two, Time1 and Time1_B, introduced variability into the waiting times. For Amt1 and Time1, each block consisted of 4 forced trials followed by one choice trial. On the other hand, experiments Amt1_B and Time1_B eliminated the forced trials and presented uninterrupted choice trials instead. These later two experiments should thus be subject to the hot stove effect.

However, there was little observable difference in the results (Table 7.3). The proportion of choices for the variable option was only marginally lower in Amount1_B than in Amount1, and there was even less difference in the choice probabilities between Time1 and Time1_B. Although consistent with the hot stove effect, these results suggest that it had little impact on the Starling's choices in this experiment.

Bateson and Kacelnik 1996

These experiments employed the same titration procedure as in BK 1995, but in some of the experiments the standard option provided rewards stochastically. Experiment T_Amt3 was functionally equivalent to the experiments in BK 1995 but used different reward sizes and waiting times. Experiment T_Time1 was also quite similar, but here the waiting time was adjusted and the reward amount held constant. In Experiments T_Time2 and T_Time3, the standard option provided rewards stochastically with the same amount but either a small or a large waiting time. In T_Time2, the waiting time on the adjusting option was titrated, while in T_Time3 the reward amount was titrated.

The first two experiments from this paper showed a titration bias but the magnitude was quite small. The presence of a titration bias in the other two experiments depends on the utility function used, so it is not obvious whether these data display a titration bias.

Bateson and Kacelnik 1997

The experiments in this paper were unique in that they attempted to gauge the source of risk sensitivity. Specifically, these experiments were designed to test whether variability or uncertainty was the key factor. In each treatment a titrating option provided rewards with no variability, but the waiting time to the reward was titrated throughout the experiment. The alternative option provided rewards with variability, but with different amounts of uncertainty. The risk free (RF) treatment delivered either a large or a small reward but always delivered them in the same repetitive sequence (i.e. $+-+-+-$). The risky (RY) treatment delivered the same two rewards but the sequences were quasi-random (for example $+--++-$ or $-+--+-$).

Each trial in this experiment was composed of a sequence of 6 sub-trials. Thus upon making a choice on a choice trial, the organism initiated a sequence of 6 sub-trials and received six rewards in a row. On the forced trials, the source of the rewards was determined by the experimenter, and the birds had one forced trial (6 sub trials) from each option in each block. There were 3 trials per block, two forced followed by two

choice, so the bird received 18 samples in each block (18 sub-trials): 6 from each of the options on the forced trials, and 6 from an option of its own choosing.

The experiment took place over 4 stages. In stage one (Experiments P_RF_1, and P_RY_2), the values of the adjusting waiting time was held constant at a large value until the starlings developed a preference for the standard option. The response variable in this stage was thus a probability or preference, so the results from this stage are included with the P-experiments. The remaining stages are classified as T-experiments. In stage 2 the titrations began, and the value of the adjusting option was changed after each block. If the adjusting option was chosen on the choice trial, the adjusting waiting time was increased by one second, while if the standard option was chosen, the adjusting waiting time was decreased by one second. Stage 3 began using the final titration value from stage 2, but now the adjusting option was only titrated at the end of each session (30 blocks). If, in a given block, the adjusting option was chosen significantly more often than the standard, as indicated by a two-tailed binomial test ($p<0.05$), the value of the adjusting time was increased by one unit, otherwise it was decreased. Finally, Stage 4 used the same titration rule as in stage 3 but the reward amount was increased from one pellet to two.

In Stage 1, the birds developed strong preferences under both treatments (Table 7.3). The results from the various titration stages were quite similar within treatments (Table 7.4), but the titration value in stage 2 was slightly larger than in stages 3 or 4. However, there was a difference between treatments, as the mean value of the titrating time was significantly lower in the risky than in the risk free group. Bateson and Kacelnik interpreted this result as suggesting that the birds preferred the risky option to the risk free option, since a longer waiting time was needed to make the birds indifferent between the risk free standard option and the titrating option. Surprisingly, the birds seemed to prefer an unpredictable reward to a predictable one.

Bateson and Kacelnik (2002)

These P-experiments introduced variability into the reward size, and held the waiting time constant for all options. There were three possible reward distributions: a

constant reward distribution always providing the same reward, a high variance distribution, and a low variance distribution. The three experiments from this paper explore all three pair-wise combinations of these distributions (Const., Low; Const., High; Low; High); note that the last experiment (P_Amt4) provides a choice between two variable options.

Surprisingly, the birds seem to prefer the intermediate level of variability the most. The birds preferred the low to the constant option (P_Amt2), the constant to the high option (P_Amt3), and the low to the high option (P_Amt4), as shown in Table 7.3. While the preferences in the former two experiments were slight, the preference was significant in the third. This intermediate level of preference was borne out in an additional experiment (not discussed here) that presented all three options simultaneously.

Note that this experiment differed from the others in several important respects. First of all, more birds (8) were used in this experiment and the reward units were different: this experiment offered .045 g starling pellets as rewards, while the others provided .012g of turkey starter crumbs. Second, these experiments utilized many consecutive choice trials per block (18 forced trials followed by 18 choice trials), while most of the other experiments had at most 2 consecutive choice trials. As a result, the hot stove effect should have more impact in these experiments.

Schuck-Paim and Kacelnik (2002)

As with the Bateson (2002) experiments, these experiments utilized three possible reward distributions (a constant, a low variance, and a high variance) and tested all 3 pair-wise combinations, but here variability was introduced into the waiting times. The birds were strongly risk prone in all three experiments, but showed no evidence of the intermediate risk preference displayed in Bateson (2002).

Simulating the experiments

Although the previous chapters present analytic results that could be used to derive model predictions analytically on several of the experiments, I instead chose to simulate all the experiments. By using simulations on all experiments, the methodology is standardized and results can still be presented for models or experiments that are not amenable to an analytic treatment. The simulations attempt to replicate the experimental structure in all aspects relevant to the models under study. Thus in each experiment I simulated the published number of pre-training, training, and testing trials exactly. Several aspects of the experimental procedures, such as the inter-trial interval or the number of blocks per day are ignored completely, since they do not impact the dynamics of the tested models.

A simple Monte-Carlo method was used to explore the parameter space of each model. Uniform distributions were defined for each parameter, and parameterizations were generated by sampling independently from these distributions. On each experiment, the predictions for a given parameterization were derived by simulating an experimental run for the appropriate number of replicate birds. Each bird was thus assumed to have exactly the same parameter values across all of the experiments. This potentially represents a major simplification, as individual birds might express different behavior. However, in order to add some variability to the parameterizations across birds, prior distributions on the parameter values, requiring additional parameters to define, would be needed. Thus in the interest of simplicity, the simulations assume that all the birds are identical.

Evaluating model fit

A metric is needed to evaluate model performance, and the metric used can have substantial ramifications for the resulting analysis. In the current context, the choice of metric is complicated by the diverse nature of the response variables. The response variables in the P-experiments are probabilities bounded in $[0,1]$, while in the T-

experiments they range from approximately 4-20, and so are an order of magnitude larger. Some traditional metrics, such as the sum of squares, will be biased in the face of this divergence, since the maximum possible squared error in a P-experiment is 1, while in a T-experiment it is functionally unbounded. Thus a sum of squares metric will be more sensitive to model prediction errors in the T-experiments than in the P-experiments.

Consider instead a metric based on the negative log-likelihood. Rather than compute the empirical likelihood for a model through simulation, I will assume simple distributions for the response variables. The P-experiments naturally imply a binomial distribution with success probability equal to the observed proportion of choices for the variable option, θ_i, based on a sample size of N_i, the number of choice trials in the testing phase of experiment i. With this assumption, and letting $\hat{\theta}_i$ be the choice proportion predicted by a given model, the negative log likelihood for P-experiment i is given by:

$$LL_i^P = N_i\left(\theta_i \ln\left(\hat{\theta}_i\right) + \left(1-\theta_i\right)\ln\left(1-\hat{\theta}_i\right)\right). \tag{6.1}$$

For the titration experiments, a normal distribution with parameters (μ_i, σ_i^2) and a mean equal to the average value of the adjusting option, $\mu_i = \overline{r}_{Ji}$, is an obvious choice. Without access to the data, exact measures of variance can't be computed, but based on the evidence discussed in Chapter 3, I'll assume a constant coefficient of variation, with the variance proportional to the square of the mean, $\sigma_i^2 \propto \mu_i^2$. The negative log-likelihood on T-experiment i is then proportional to

$$LL_i^T = N_i \frac{\left(\hat{\mu}_i - \mu_i\right)^2}{\mu_i^2}, \tag{6.2}$$

where $\hat{\mu}_i$ is the predicted mean value for the adjusting option. Effectively Equation (6.2) is the weighted squared error of the model's prediction on experiment i. Note that, for both sets of experiments these assumptions are somewhat arbitrary; a more detailed analysis of the data could help determine whether these distributions are appropriate.

The preceding equations present formulas for computing the negative log likelihood on individual experiments. To evaluate model performance on the entire set of

experiments these individual measures must be combined. I used three statistics to summarize model fit. The first is the total negative log likelihood on the probability experiments:

$$LL^P = \sum_{\substack{\text{probability} \\ \text{experiments}}} LL_i^P \ . \tag{6.3}$$

Similarly, the total negative log likelihood on the titration experiments is

$$LL^T = \sum_{\substack{\text{titration} \\ \text{experiments}}} LL_i^T \ , \tag{6.4}$$

and the overall negative log likelihood is the sum of these two values

$$LL = LL^T + LL^P \ . \tag{6.5}$$

The statistic LL^T summarizes model performance on the T-experiments, LL^P summarizes model performance on the P-experiments, and LL summarizes model performance on the entire set of experiments. For each statistic, smaller is better, and we want to minimize each of these summary statistics. Note that these statistics are relative measures of model performance and mean little in isolation; instead they are useful for comparing the performance of different models.

Models and results

There are many possible combinations of utility functions, choice functions, and averaging functions. However, many of these combinations are functionally equivalent, and the lack of data obscures the model selection: deciding between these similar stochastic models becomes somewhat arbitrary. So here I present results for some simple models in an attempt to diagnose model deficiencies. Table 7.5 summarizes all of the different functions that are explored in the following simulations. A preliminary analysis indicated that there was little difference between results obtained using a TWMA and those obtained using a EWMA, so no models with TWMAs are presented here. Model characteristics are summarized by a three part code denoting the different function used in the model. For example, the model N-EW-MB uses the null utility function, a EWMA averaging function, and a modified Boltzmann choice function.

Table 7.5 The different functions used in the model selection.

	Name	Code	Function	Num Par
Currency			$\mathcal{U}(s,w) =$	
	Null	N	$\frac{s}{1+w}$	0
Choice			$\mathcal{C}(x,y) =$	
	Matching	M	$\frac{x}{x+y}$	0
	Boltzmann	B	$\frac{1}{1+e^{\frac{1}{\gamma}(y-x)}}$	1
	Modified-Boltz	MB	$\frac{1}{1+e^{\frac{1}{\gamma}\left(\frac{y-x}{y+x}\right)}}$	1
Averaging			$\hat{r}(n+1) =$	
	Exponential-Weighted	EW	$(1-m)r(n)+m\hat{r}(n)$	1

Perhaps the simplest possible model, in terms of the number of parameters, combines the null utility function with a EWMA and a matching choice function (N-EW-M). The null utility function, shown in Table 7.5, avoids division by zero but has no free parameters, and the matching function also has no free parameters. With only one parameter, the memory coefficient m from the EWMA, it is easy to explore the parameter space for this model. I conducted 2000 Monte Carlo simulations with m ranging from 0.2 to 0.99. Table 7.6 presents the best results for each metric.

More informative perhaps is a graph of LL^P against LL^T with color indicating the value of m (Figure 7.2). Happily, the best fits in both sets of experiments are obtained with similar parameter values. Moreover, model fit appears to improve as m increases, with the best fits for $m > .5$. Note that, due to the stochastic nature of the models, there is some inherent variability in the log likelihood values, and duplicate runs with identical parameter values can produce different summary statistics.

Table 7.6 Numerical results. The best parameterization, as measured by each of the three statistics, is shown. Note how γ descends for each model as performance improves on the P-experiments.

	LL_P	LL_T	LL	m	γ	γ_T
N-EW-M						
Best LL_T	228.22	144.47	372.69	0.91		
Best LL	228.22	144.47	372.69	0.91		
Best LL_P	190.36	326.69	517.05	0.86		
N-EW-B						
Best LL_T	300.53	235.72	536.26	0.84	0.41	
Best LL	147.87	257.62	405.49	0.98	0.16	
Best LL_P	116.42	346.32	462.74	0.98	0.13	
N-EW-MB						
Best LL_T	344.97	111.96	456.93	0.99	0.61	
Best LL	28.31	247.67	275.98	0.95	0.20	
Best LL_P	12.81	307.92	320.74	0.98	0.16	
N-EW-G						
Best LL_T	428.56	118.44	547.01	0.67	1.59	
Best LL	36.89	239.57	276.46	0.98	0.43	
Best LL_P	11.23	327.04	338.27	0.99	0.36	
N-EW-MBS						
Best LL_T	283.51	42.72	326.23	0.98	0.47	1.61
Best LL	35.79	57.84	93.63	0.99	0.15	1.58
Best LL_P	7.55	201.97	209.52	0.99	0.16	1.26

More complex choice functions can substantially improve model fit. Table 7.6 also shows results for three other models using different choice functions. The model using the Boltzmann choice function (N-EW-B) was clearly the worst, out performed even by the matching choice function (N-EW-M) under two of the three metrics. However, the models using the Gaussian (N-EW-G) and the modified Boltzmann (N-EW-MB) functions fared better, especially on the P-experiments. Table 7.6 also suggests a pattern in the relationship between model fit and the choosiness of the associated choice function. Choosier choice functions (small values of γ) perform better on the P-

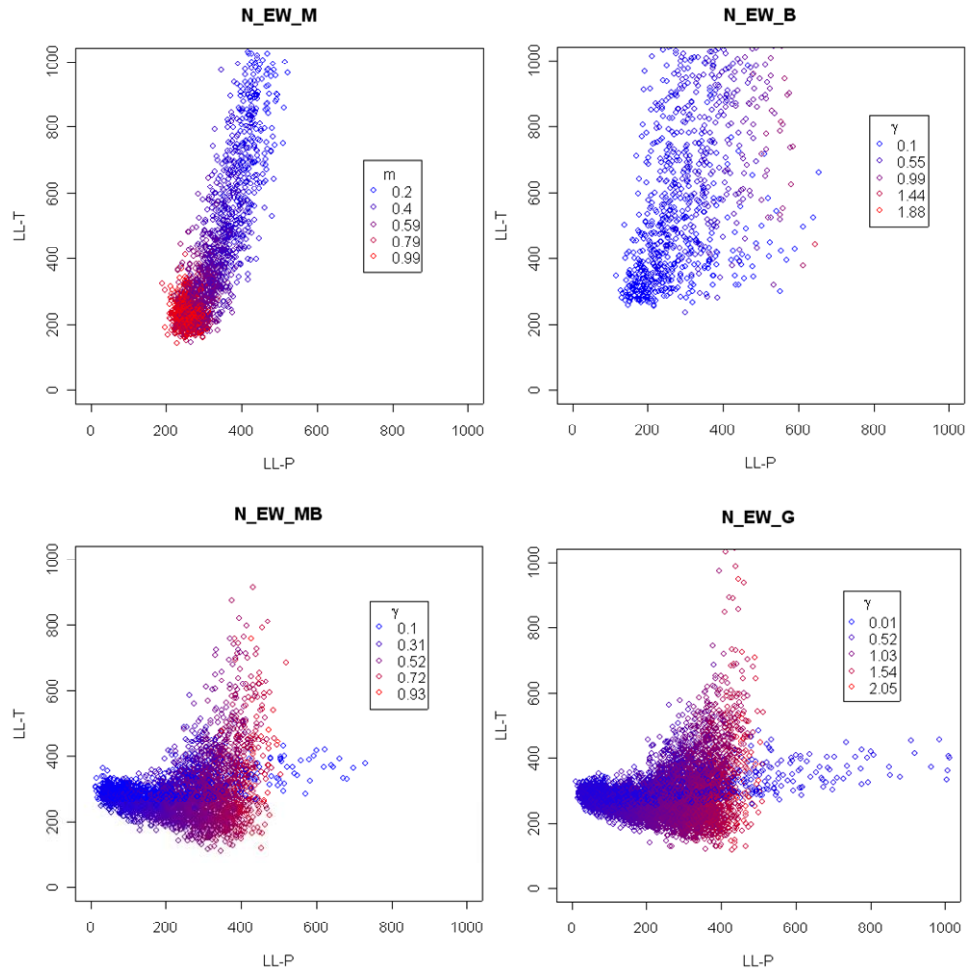


Figure 7.2 Log-likelihood on theT-experiments vs. log-likelihood on the P-experiments. On each graph, the color denotes the value of one of the free parameters. Note that on the bottom two graphs, smaller values of γ do well on the P-experiments, while larger values do well on the T-experiments.

experiments, but less choosy choice functions (large values of γ) are better on the T-experiments.

Graphing LL^T against LL^P further substantiates these observations. Figure 7.2 presents graphs of the negative log-likelihoods for these three more complex models, with the color on the graph indicating the value of γ. Again, the model with the Boltzmann choice function performs much worse than all others, but the Gaussian and modified Boltzmann choice functions produce similar results, suggesting that these choice functions might be largely equivalent. Moreover, the change in model performance as a function of γ is quite obvious in both the Gaussian and modified

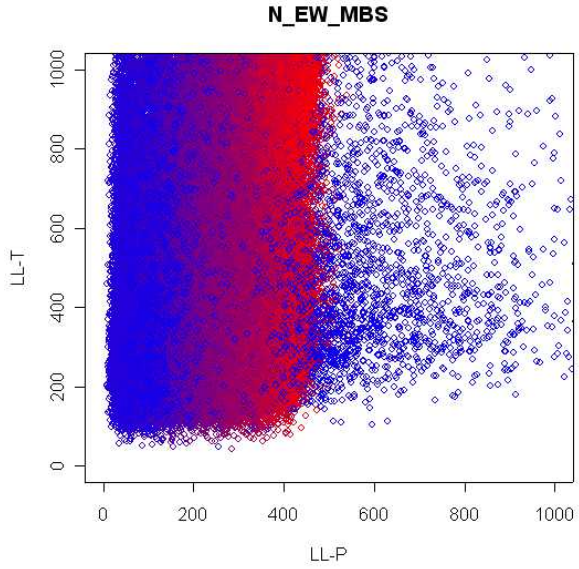


Figure 7.3 Results with two different γ values. The model with a modified Boltzmann choice function and two parameters γ_P (shown here as the color) and γ_T (not depicted).

Boltzmann graphs. Clearly there is tension between the two groups of experiments, and choosier choice functions do better on the P-experiments.

This tension is made clearer by introducing different values for γ when the environment is transient and when it is stationary. The model N-EW-MBS uses the modified Boltzmann choice function but introduces an additional parameter, γ_T, for use in the transient T-experiments. In the portions of the experiments where reward distributions are stationary (the pre-training periods and all stages of the P-experiments), the parameter γ is used, while in the transient parts of the experiments (training and testing stages in the T-experiments) γ_T is used. Over 200,000 simulations, shown in Figure 7.3, simultaneous performance on both the P and T experiments is improved substantially, as indicated by the LL score in Table 7.6.

Despite the relatively good performance of the N-EW-MBS model, it cannot match some important characteristics of the data. Figure 7.4 shows the T-experiment results from 500 simulations of the N-EW-MBS parameterization with the best LL value (taken from Table 7.6). By no means does this model match all of the significant results

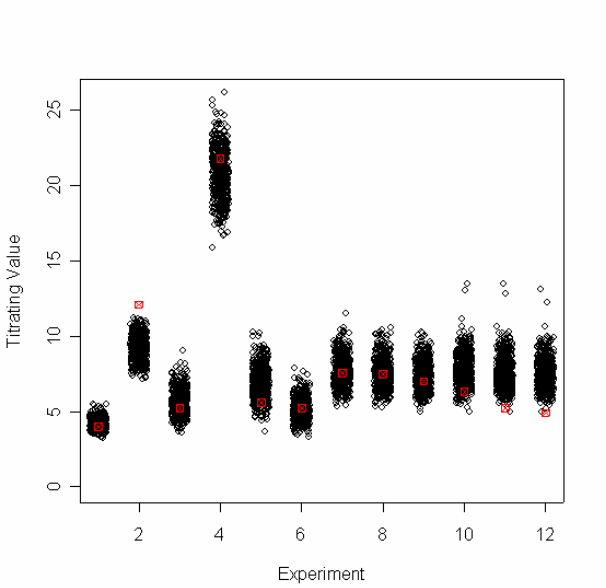


Figure 7.4 Predictions from the best N-EW-MBS model parameterization on the T-experiments. Results from 500 simulations are shown.

from the data; it never produces the correct value on the second experiment (T_Amt2) and only rarely matches the results from the last two experiments (T_RY_3, T_RY_4).

Model performance could perhaps be improved by using other averaging or utility functions. However, a cursory analysis suggested that differentiating between these models was difficult with the data on hand, and no such analysis will be attempted here. Suffice it to say that the N-EW-MBS model is superior to all the other models presented in Table 7.6. The discrepancy in γ values needed to fit the T and P experiments suggests a potential failure of the simple DELA models.

Discussion

The poor performance of the Boltzmann choice function relative to the others can be taken as more evidence against difference-based choice functions. The difference-based Boltzmann choice function, with one free parameter, on the whole provides a worse fit to the data than do any of the ratio-based choice functions, even the matching choice function with no free parameters. These results suggest strongly that the starlings are making choices by taking ratios, not differences. Although none of the single parameter choice functions performed well simultaneously on both sets of experiments, the modified Boltzmann and Gaussian functions emerged as the best of the group. The

results from the modified Boltzmann were marginally better than those from the Gaussian, but on the whole the functions behaved quite similarly.

The modified Boltzmann choice function with two parameters, one for each type of experiment, produced the best results. Of course this is an inelegant and unrealistic solution. If the bird is to use different parameters in different types of environments, it must have a mechanism for detecting the difference between the environments and a mechanism for changing its choice parameter. Yet it seems quite natural that a living bird would be less choosy in a transient environment, if only because its knowledge is more uncertain: if the reward distributions are changing, the bird should have less certainty in its estimates and be more likely to explore by choosing options with a currently smaller estimate. In transient environments, estimates are less certain, exploration is more valuable, and choosiness should decrease.

The basic DELA framework has no way to represent certainty or confidence in the value of the estimators. Similarly, there is no mechanism for evaluating whether an environment is transient or detecting environmental changes and trends. In the conclusion, I will argue that this is a fundamental flaw in the DELA framework, and suggest some possible mechanisms for addressing these considerations.

Conclusion

This thesis had three main goals. First, to establish a framework and formalism for thinking about learning models that would be useful for organizing the published models from psychology and animal behavior. Second, to establish some formal analytic results for DELA models in the context of some simple discrete trials experiments. Finally, to use a set of data from actual experiments to evaluate some simple models and diagnose model deficiencies.

To the first end, DELA models were introduced using the terminology and formalism from learning automata. The utility-estimator-choice structure was then presented to help structure the analysis of DELA models and their application to living organisms. This structure proved useful for deriving some analytic results about DELA models, and these results can be applied to other learning models that meet the given mathematical criteria. Moreover these results demonstrated that the different model elements (utility-estimator-choice) need to be considered together when analyzing model dynamics. The interaction of these elements can lead to some surprising results, such as the hot stove effect and the titration bias.

There were three main mathematical results. Chapter 3 showed how simple DELA models with R-concave choice functions would always produce a titration bias in the face of the titration experiment conducted by Bateson and Kacelnik (1995a). Chapter 5 proved that the hot stove effect will lead to genuine risk aversion for all DELA models with unbiased and commutative averaging functions and R-concave choice functions confronted by a BRSE with equiprobable rewards. Finally, Chapter 6 established some results about the Couvillon Bitterman learning model, and presented some mathematical techniques for evaluating its behavior.

One consequence of the hot stove effect is a relationship between the number of consecutive forced trials in an experiment and the magnitude of risk aversion. Recall, from Equation (4.21), that the proportion of choices for the variable option will be bounded:

$$E_e\left(\mathcal{C}\left(\mathcal{A}(\Psi(n)), r_c\right)\right) \geq \theta \geq H_e\left(\mathcal{C}\left(\mathcal{A}(\Psi(n)), r_c\right)\right).$$

As the number of consecutive choice trials increases, the proportion of choices for the variable option will approach the lower bound. There is a simple experimental test of this result: vary the number of choice trials and examine the impact on risk sensitive preferences. If the proportion of choices for the variable option does not decrease as the number of consecutive choice trials increases, the decision mechanisms of the experimental subjects must be able to avoid the hot stove effect in some way.

Comparing model predictions with the results from the large group of Starling experiments (Chapter 7) was informative, suggesting a clear model deficiency: the inability to simultaneously fit the results from both the transient T-experiments and stationary P experiments with a single set of parameters. As a quick hack solution, I proposed using two different choice function parameters, different levels of choosiness, in different types of environments. While there may be other ways to improve DELA model performance on both sets of experiments, such as by introducing different utility or averaging functions, introducing different choice parameters was an easy, if unsatisfying, solution to the problem. Moreover, there is some intuitive appeal to the idea that organisms would be less choosy, and more explorative, in transient environments.

Gallistel (1990) criticized LA-type models because their decision mechanism has no place for estimates of certainty. Not only the value of an estimate matters, the certainty, or confidence, of that estimate should also impact the decision making process. Many factors can affect the certainty of an estimate for an option's expected value: the total number of samples taken from the option, the amount of time since the last sample was taken, whether the environment has been stable or fluctuating in the past, etc. Organisms should incorporate these factors into their decision making process will be more efficient.

A learning model incorporating uncertainty into the modeling framework might be able to resolve the conflict between the T and P experiments and differentiate between stationary and transient environments. If the environment has been transient there is more uncertainty in the estimates and choosiness should decrease. Modifying a DELA model in this way might also change the risk sensitive preferences expressed in stationary

environments: estimating environmental transience is complicated by environmental stochasticity and runs of bad luck can seem like transitions in the reward distributions.

More generally, I would argue that we should think about organisms as responding to transient time series and engaging in forecasting. Based on the sequence of samples obtained in the past, the organism needs a forecast about the current rewards expected from the environment. The estimators in a DELA model should perhaps instead be thought of as forecasters: forecasts from the past about the present state of the environment are needed. The averaging functions tested in Chapter 7 make very poor forecasters, tending to lag behind the current state of a transient time series. By embellishing the averaging function, for example by introducing separate short term (small m) and a long term (large m) averages and integrating them together into a single forecast, better forecasts can be obtained, perhaps increasing the organism's ability to respond to and exploit environmental trends or transitions.

In an essential way, forecasts about the expected value of engaging in different actions are the basis for decision making, but the certainty of these forecasts should also play a large role in the decision making process. Some forecasts deserve more confidence than others, as many television meteorologists can attest. A more complete DELA model would generate choices by simultaneously incorporating better forecasting methods with estimates of the uncertainty of the forecasts. As discussed above, this could improve the model performance on the body of starling experiments and explain the divergence between the two groups of experiments.

Of course there are many possible ways to integrate forecasting and uncertainty into the DELA framework, and many might be equivalent to Ptolemaic epicycles. My hope is that by exploring how simple models fail, new experiments can be motivated that will then help direct the embellishment of the simple models. In the process, formal mathematical analysis of model dynamics and detailed numerical comparisons of model predictions with experimental data will be essential.

References

Barnsley, M. F. (1993). *Fractals everywhere*. Boston, MA: Academic.

Bateson, M. and Kacelnik, A. (1995a). Accuracy of memory for amount in the foraging starling *(Sturnus vulgaris). Animal Behaviour*. 50: 431-443.

Bateson, M. and Kacelnik, A. (1995b). Preferences for fixed and variable food sources: variability in amount and delay. *Journal of the Experimental Analysis of Behavior* 63: 313-329.

Bateson, M. and Kacelnik, A. (1996). Rate currencies and the foraging starling: the fallacy of the averages revisited. *Behavioral Ecology*, 7: 341–352.

Bateson, M. and Kacelnik, A. (1997). Starlings' preferences for predictable and unpredictable delays to food. *Animal Behaviour*, 53**:** 1129–1142.

Bateson, M. (2002). Context-dependent foraging choices in risk-sensitive starlings. *Animal Behaviour*. 64: 251-260.

Berry, D. A. and Fristedt, B. (1985). *Bandit Problems: Sequential Allocation of Experiments*. London: Chapman and Hall.

Bullen, P. S. (2003). *Handbook of Means and Their Inequalities*. Boston, MA: Kluwer Academic.

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

Bush, R. R. and Estes, W. K.-editors. (1959). *Studies in Mathematical Learning Theory*. Stanford, CA: Stanford University.

Caraco, T., Blackenhorn, W. U., Gregory, G. M., Newman, J. A., Recer, G. M., and Zwicker, S. M. (1990). Risk-sensitivity: Ambient temperature affects foraging choice. *Animal Behaviour*. 44: 441-447.

Caraco, T., Kacelnik, A., Mesnik, N., and Smulewitz, M. (1992). Short term rate maximization when rewards and delays covary. *Animal Behaviour* 44: 441–447.

Cardinal, R. N., Daw, N., Robbins, T. W., and Everitt, B. J. (2002). Local analysis of behaviour in the adjusting-delay task for assessing choice of delayed reinforcement. 15: 617-634.

Couvillon, P. A. and Bitterman, M. E. (1991). How honeybees make choices. In: Goodman, J. L. and Fischer, R. C. (editors), *The Behaviour and Physiology of Bees*, Wallingford, UK: CAB International. 116–130.

Dayan, P. and Abbott, L. F. (2001). *Theoretical neuroscience: Computational and mathematical modeling of neural systems*. Cambridge, MA: Massachusetts Institute of Technology Press.

Denrell, J. (2005). Uncertainty aversion in experiential learning. Working paper, Stanford graduate school of business.

Denrell, J. and March, J. G. (2001). Adaptation as information restriction: the hot stove effect. *Organization Science*. 12(5): 523-538.

Edalat, A. (1996). Power domains and iterated function systems. *Information and Computation*. 124: 182-197.

Fantino, E. and Goldshmidt, J. N. (2000). Differences, not ratios, control choice in an experimental analogue to foraging. *Psychological Science*, 11, 229-233.

Gallager, R. G. (1996). *Discrete stochastic processes*. Boston, MA: Kluwer Academic.

Gallistel, C. R. (1990). *The organization of learning*. Cambridge, MA: MIT Press.

Gibbon, J., Church, R. M., and Meck, W. H. (1984). Scalar timing in memory. In: Gibbon, J. and Allan, L. (editors), *Timing and time perception*. New York, NY: Annals of the New York Academy of Sciences. 423: 52-77.

Gibbon, J. and Fairhust, S. (1994). Ratio versus difference comparators in choice. *Journal of the Experimental Analysis of Behavior*. 62(3): 409-434.

Harder, L., and Real, L. 1987. Why are bumble bees risk averse? *Ecology*. 68: 1104-1108.

Hepting, D., Prusinkiewicz, P., and Saupe, D. (1991). Rendering methods for iterated function systems, in *Proceedings IFIP Fractals 90*.

Herrnstein, R. J. (1961). Relative and absolute strength of response as a function of frequency of reinforcement. *Journal of the Experimental Analysis of Behavior*, 4: 267-272.

Howard, R. A. (1971). *Dynamic probabilistic systems volume II: Semi-Markov and decision processes*. New York, NY: Wiley and Sons.

Kacelnik, A. and Bateson, M. (1996). Risky theories – the effects of variance on foraging decisions. *American Zoologist* 36, 402–434.

Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4: 237-285.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*. 47(2): 263-291.

Keaser, T., Rashkovich, E., Shmida, A. (2002). Bees in two-armed bandit situations: Foraging choices and possible decision mechanisms. *Behavioral Ecology*. 13(6): 757-763.

Killeen, P. R. (1984). Incentive Theory III: Adaptive clocks. In: Gibbon, J. and Allan, L. (editors), *Timing and time perception*. New York, NY: Annals of the New York Academy of Sciences. 423: 269-277.

Lea, S. E . G. (1976). Titration of schedule parameters by the pigeon. *Journal of the experimental analysis of behavior*. 25: 43-54.

Lea, S. E. G., and Dow, S. M. (1984). The integration of reinforcements over time. In Gibbon, J. and Allan, L. (editors), *Timing and time perception*. New York, NY: Annals of the New York Academy of Sciences. 423: 269-277.

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. New York, NY: Wiley and Sons.

March, J. G. (1996). Learning to be risk averse. *Psychological Review*. 103(2): 309-319.

Marsh, B., Schuck-Paim, C., and Kacelnik, A. (2004). Energetic state during learning affects foraging choices in starlings. *Behavioral Ecology*. 15(3): 396-399.

Maynard Smith, J. (1984). Game theory and the evolution of behaviour. *Behavioral and Brain Sciences***.** 7: 95-125.

Mazur, J.E. (1984). Tests of an equivalence rule for fixed and variable reinforced delays. *Journal of Experimental Psychology: Animal Behaviour Processes*. 10: 426-436.

Mazur, J. E. (1985). Probability and delay of reinforcement as factors in discrete-trial choice. *Journal of the Experimental Analysis of Behavior*. 43(3): 341-351.

Mazur, J. E. (1986a). Choice between single and multiple delayed reinforcers. *Journal of the Experimental Analysis of Behavior*. 46(1): 67-77.

Mazur, J. E. (1986b). Fixed and variable ratios and delays: Further tests of an equivalence rule. *Journal of Experimental Psychology: Animal Behavior Processes*. 12: 116-124.

Mazur, J. E. (2000). Tradeoffs among delay, rate, and amount of reinforcement. *Behavioural Processes*. 49: 1-10.

Mazur, J.E. (2001). Hyperbolic value addition and general models of animal choice. *Psychological Review*. 108(1): 96-112.

Mazur, J. E. (2002). Evidence against a constant-difference effect in concurrent-chains schedules. *Journal of the Experimental Analysis of Behavior*. 77(2): 147-155.

Mazur, J. E. (2005). Effects of reinforcer probability, delay, and response requirements on the choices of rats and pigeons: Possible species differences. *Journal of the Experimental Analysis of Behavior*. 83 (3): 263-279.

Narendra, K. S., and Thathachar, M. A. L. (1974). Learning automata: a survey. *IEEE Transactions on Systems, Man, and Cybernetics*, 4 (4): 323-334.

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

Niculescu, C. P. (2003). Convexity according to means. *Mathematical Inequalities and Applications*. 6(4): 571-579.

Niv, Y., Joel, D., Meilijson, I., Ruppin, E. (2002). Evolution of reinforcement learning in uncertain environments: a simple explanation for complex foraging behaviors. *Adaptive Behavior*. 10(1): 5-24.

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

Philippou A. N., Georghiou C., and Philippou G. N. (1983). A generalized geometric distribution and some of its properties. *Statistics and Probability Letters*. 1(4): 171-175.

Pompilio, L. and Kacelnik, A. (2005). State-dependent learning and suboptimal choice: when starlings prefer long over short delays to food. *Animal Behaviour*. 70(3): 571-578.

Reboreda, J.C. and Kacelnik, A. (1991). Risk Sensitivity in starlings (*Sturnus vulgaris L.*): the effect of variability in food amount and in delay to food. *Behavioral Ecology*, 2: 301–308.

Savastano, H. I. and Fantino, E. (1996). Differences in delay, not ratios, control choice in concurrent chains. *Journal of the Experimental Analysis of Behavior. 66(1)*: 97-116.

Schuck-Paim, C. and Kacelnik, A. Rationality in risk-sensitive foraging choices by starlings. Animal Behaviour. 64(6): 869-879.

Shapiro, M. S. (2000). Quantitative analysis of risk sensitivity in honeybees (*Apis Mellifera*) with variability in concentration and amount of reward. *Journal of Experimental Psychology: Animal Behavior Processes*. 26(2): 196-205.

Shapiro, M. S., Couvillon, P. A., and Bitterman, M. E. (2001). Quantitative tests of an associative theory of risk-sensitivity in honeybees. *Journal of Experimental Biology*, 204(3): 565-573.

Slomczynski W., Kwapien J., and Zyczkowski K. (2000). Entropy computing via integration over fractal measures. Chaos 10 (1): 180-188.

Stephens D. W. and Krebs J. R. 1986. *Foraging theory*. Princeton, NJ: Princeton University Press.

Sutton, R. S. and Barto, A. G. 1998. *Reinforcement learning: An introduction.* Cambridge, MA: MIT Press.

Vasilakos, A. V. and Papadimitiou, G. I. (1995). A new approach to the design of reinforcement schemes for learning automata: Stochastic estimator learning algorithm. *Neurocomputing.* 7: 275-297.

Von Neumann, J. and Morgenstern, O. (1947). Theory of games and economic behavior. 2nd edition. Princeton, NJ: Princeton University Press.

Weber E. U., Shafir S., and Blais A. R. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review*. 111 (2): 430-445

Yechiam, E. and Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience-based decision making. *Psychonomic Bulletin and Review*. 12(3): 387-402.

Zikun, W. and Xiangqun, Y. (1992). *Birth and death processes and Markov chains.* New York, NY: Springer-Verlag.

Appendix 1- Titration bias proof

Proposition: All strictly R-concave choice functions, $\mathcal{C}(.)$, will produce a titration bias, $\Delta_J = \overline{r}_J - r_S > 0$.

Proof: Define

$$\beta_i = \frac{\alpha_i}{\sum_{i=0}^{\infty} \alpha_i}, \tag{A1.1}$$

and note that $\beta_i \geq 0$ and $\sum_{i=0}^{\infty} \beta_i = 1$. Using Equation (2.42), we can then write the titration bias as

$$\Delta_J = \sum_{i=0}^{\infty} (i - r_S)\beta_i \, . \tag{A1.2}$$

Rearranging this equation gives

$$\Delta_J = \sum_{i=1}^{r_S} i\left(\beta_{r_S+i} - \beta_{r_S-i}\right) + \sum_{i=2r_S+1}^{\infty} (i - r_S)\beta_i \, . \tag{A1.3}$$

Since the second summation in this equation is clearly greater than or equal to zero, the proof is complete if we can show that the first summation is also greater than or equal to zero. Thus if we can show that $\beta_{r_S+i} - \beta_{r_S-i} \geq 0$ for $i = 1,...r_S$, we are done.

The proof will proceed by induction. First, note from Equation (2.44) that

$$\beta_{r_S+1} = \beta_{r_S}\left(\frac{1-\phi_{r_S}}{\phi_{r_S+1}}\right)^2 \tag{A1.4}$$

and similarly

$$\beta_{r_S-1} = \beta_{r_S}\left(\frac{\phi_{r_S}}{1-\phi_{r_S-1}}\right)^2 . \tag{A1.5}$$

Subtracting these, we obtain

$$\beta_{r_S+1} - \beta_{r_S-1} = \beta_{r_S}\left(\frac{\left(1-\phi_{r_S-1}\right)^2\left(1-\phi_{r_S}\right)^2 - \left(\phi_{r_S}\phi_{r_S+1}\right)^2}{\left(\phi_{r_S+1}(1-\phi_{r_S-1})\right)^2}\right). \tag{A1.6}$$

By condition C.2, we know that $\phi_{r_S} = \mathcal{C}(r_S, r_S) = 1 - \mathcal{C}(r_S, r_S) = 1 - \phi_{r_S}$, so (A1.6) simplifies to

$$\beta_{r_S+1} - \beta_{r_S-1} = \beta_{r_S}\phi_{r_S}^2\left(\frac{\left(1-\phi_{r_S-1}\right)^2 - \phi_{r_S+1}^2}{\left(\phi_{r_S+1}(1-\phi_{r_S-1})\right)^2}\right) \tag{A1.7}$$

Since the choice function is assumed to be strictly R-concave, we know from Equation (2.50) that

$$1-\phi_{r_S-1} = \mathcal{C}(r_S, r_S-1) > \mathcal{C}(r_S+1, r_S) = \phi_{r_S+1} \tag{A1.8}$$

and thus that $\beta_{r_S+1} - \beta_{r_S-1} > 0$.

To complete the proof, note that

$$\beta_{r_S+i+1} - \beta_{r_S-i-1} = \beta_{r_S+i}\left(\frac{1-\phi_{r_S+i}}{\phi_{r_S+i+1}}\right)^2 - \beta_{r_S-i}\left(\frac{\phi_{r_S-i}}{1-\phi_{r_S-i-1}}\right)^2. \tag{A1.9}$$

Assume that $\beta_{r_S+i} - \beta_{r_S-i} > 0$; then Equation (A1.9) can be rewritten as

$$\beta_{r_S+i+1} - \beta_{r_S-i-1} > \beta_{r_S+i}\left(\left(\frac{1-\phi_{r_S+i}}{\phi_{r_S+i+1}}\right)^2 - \left(\frac{\phi_{r_S-i}}{1-\phi_{r_S-i-1}}\right)^2\right). \tag{A1.10}$$

As above, however, it follows directly from the assumed R-concavity that

$$\left(\left(\frac{1-\phi_{r_S+i}}{\phi_{r_S+i+1}}\right)^2 - \left(\frac{\phi_{r_S-i}}{1-\phi_{r_S-i-1}}\right)^2\right) > 0. \tag{A1.11}$$

Thus by induction

$$\beta_{r_S+i} - \beta_{r_S-i} > 0 \tag{A1.12}$$

for all $i = 1, \ldots r_S$.

Proposition: It is more difficult for an R-affine choice function, to produce a titration bias.

Proof: Note that, for an R-affine function, $1-\phi_{r_S+i}=\phi_{r_S-i}$, and thus, by an argument similar to the one presented above, $\beta_{r_S+i}-\beta_{r_S-i}=0$ for all $i=1,...r_S$. So the first summation in Equation (A1.3) is equal to zero, and any titration bias will only be a result of the second summation.

Appendix 2- Hot stove effect proof

Proposition: An agent using a DELA model with an R-concave choice function and an unbiased and commutative averaging function will always display risk averse behavior on the BRSE when $p_+ = p_- = \frac{1}{2}$.

Proof: We need to show that, for the assumed decision model,

$$\theta < \tfrac{1}{2}. \tag{A2.1}$$

From Equation (4.10), this is equivalent to showing that

$$\frac{1}{\theta} = \sum_{i=0}^{2^W-1} \pi_i^e \frac{1}{\mathcal{C}(\mathcal{A}(i), \overline{r}_v)} > 2. \tag{A2.2}$$

Since the rewards are assumed to be equiprobable, $\pi_i^e = \frac{1}{2^W}$, and the left side of Equation (A2.2) becomes

$$\frac{1}{\theta} = \sum_{i=0}^{2^W-1} \frac{1}{2^W} \frac{1}{\mathcal{C}(\mathcal{A}(i), \overline{r}_v)}. \tag{A2.3}$$

We have assumed that the averaging function is commutative, so we can rewrite this sum as

$$\frac{1}{\theta} = \sum_{i=0}^{2^{W-1}-1} \frac{1}{2^W} \left(\frac{1}{\mathcal{C}(\overline{r}_v(1-K_i), \overline{r}_v)} + \frac{1}{\mathcal{C}(\overline{r}_v(1+K_i), \overline{r}_v)} \right). \tag{A2.4}$$

Since the choice function is R-concave, we know that

$$\mathcal{C}(\overline{r}_v(1+K_i), \overline{r}_v) \le 1 - \mathcal{C}(\overline{r}_v(1-K_i), \overline{r}_v) \tag{A2.5}$$

and substituting into Equation (A2.4) we obtain

$$\frac{1}{\theta} \ge \sum_{i=0}^{2^{W-1}-1} \frac{1}{2^W} \left(\frac{1}{\mathcal{C}(\overline{r}_v(1-K_i), \overline{r}_v)} + \frac{1}{1 - \mathcal{C}(\overline{r}_v(1-K_i), \overline{r}_v)} \right). \tag{A2.6}$$

Simple arithmetic then gives

$$\frac{1}{\theta} \ge \sum_{i=0}^{2^{W-1}-1} \frac{1}{2^W} \left(\frac{1}{\mathcal{C}(\overline{r}_v(1-K_i), \overline{r}_v)\left(1 - \mathcal{C}(\overline{r}_v(1-K_i), \overline{r}_v)\right)} \right). \tag{A2.7}$$

Let $\tilde{C}_i = C(\overline{r}_v(1-K_i), \overline{r}_v)$ and substitute into (A2.7) to obtain

$$\frac{1}{\theta} \geq \sum_{i=0}^{2^{W-1}-1} \frac{1}{2^W} \left(\frac{1}{\tilde{C}_i \left(1-\tilde{C}_i\right)} \right). \tag{A2.8}$$

Since $\tilde{C}_i \in (0,1)$ is a probability, $\tilde{C}_i \left(1-\tilde{C}_i\right) \leq \frac{1}{4}$ with equality only when $\tilde{C}_i = 1-\tilde{C}_i = \frac{1}{2}$.

Thus

$$\frac{1}{\theta} \geq \sum_{i=0}^{2^{W-1}-1} \frac{1}{2^W} 4 = \frac{2^{W-1}}{2^W} 4 = 2 \tag{A2.9}$$

If there is at least one memory state for which $\tilde{C}_i \neq 1-\tilde{C}_i$ then the inequality is strict,

$$\frac{1}{\theta} > 2, \tag{A2.10}$$

and the agent will be risk averse.

Appendix 3- Asymptotic characteristics of the CB model

Let $x(t)$ be a variable that is defined on $[0,\infty)$ and evolves in discrete time steps. The evolution of $x(t)$ is given by probabilistically applying one of N contractive linear transformations to x at each time step. That is

$$x(t) = \left\{ \begin{array}{c} f_1(x(t-1)) \text{ with probability } p_2 \\ f_2(x(t-1)) \text{ with probability } p_2 \\ . \\ . \\ f_N(x(t-1)) \text{ with probability } p_N \end{array} \right\} \tag{A3.1}$$

where p_i is the probability of applying transformation $f_i(.)$, and the $f_i(.)$ are affine functions given by:

$$f_i(x) = b_i + m_i x \tag{A3.2}$$

where $b_i \in (0,\infty)$, $m_i \in (0,1)$, and both are constant across time. We want to compute the asymptotic expected value of $x(t)$:

$$\tilde{x} = \lim_{t\to\infty} E(x(t)). \tag{A3.3}$$

For any given initial value $x(0) = x_0$, only certain values for $x(t)$ are possible. Denote by X^t the set of points, s_j^t, that can be reached from x_0 at time *t*. For example:

$$\begin{aligned} X^0 &= \{x_0\} \\ X^1 &= \{f_1(x_0) \bigcup f_2(x_0) .... \bigcup f_M(x_0)\}. \\ X^t &= \left\{ \bigcup_{i=1}^{N} \left(\bigcup_{s_j^{t-1} \in X^{t-1}} f_i(s_j^{t-1}) \right) \right\} \end{aligned} \tag{A3.4}$$

Associated with each point s_j^t, is a probability, μ_j^t, which denotes the probability that $x(t)$ will have the value s_j^t at time *t* given an initial value of x_0. Let P^t represent the set of probabilities associated with X^t. Then:

$$P^0 = \{1\}$$
$$P^1 = \{p_1 \bigcup p_2 ... \bigcup p_i ... \bigcup p_M\}. \quad \text{(A3.5)}$$
$$P^t = \left\{ \bigcup_{i=1}^{N} \left(\bigcup_{\mu_j^{t-1} \in P^{t-1}} p_i \mu_j^{t-1} \right) \right\}$$

The expected value of $x(t)$ at time t, $\tilde{x}^t$, is then given by:

$$\tilde{x}^t = \sum_{s_j^t \in X^t} \mu_j^t s_j^t \,. \quad \text{(A3.6)}$$

So then

$$\tilde{x}^0 = x_0$$
$$\tilde{x}^1 = p_1 f_1(x_0) + p_2 f_2(x_0) + ... + p_N f_N(x_0) \quad \text{(A3.7)}$$
$$\tilde{x}^t = \sum_{i=1}^{N} \sum_{s_j^{t-1} \in X^{t-1}} p_i \mu_j^{t-1} f_i(s_j^{t-1})$$

and $\tilde{x} = \lim_{t \to \infty} \tilde{x}^t$.

Using Equation (A3.2), we can rewrite Equation (A3.7) as

$$\tilde{x}^t = \sum_{i=1}^{N} \sum_{s_j^{t-1} \in X^{t-1}} p_i \mu_j^{t-1} (b_i + m_i s_j^{t-1}) \quad \text{(A3.8)}$$

Note that $\Pr(x(t) \notin X^t) = 0$ and thus

$$\sum_{s_j^t \in X^t} \mu_j^t = 1 \,. \quad \text{(A3.9)}$$

So Equation (A3.8) reduces to

$$\tilde{x}^t = \sum_{i=1}^{N} p_i \left(b_i + m_i \sum_{s_j^{t-1} \in X^{t-1}} \mu_j^{t-1} s_j^{t-1} \right). \quad \text{(A3.10)}$$

By definition (Equation (A3.6)), $\tilde{x}^{t-1} = \sum_{s_j^t \in X^t} \mu_j^{t-1} s_j^{t-1}$, and Equation (A3.10) becomes

$$\tilde{x}^t = \sum_{i=1}^{N} p_i (b_i + m_i \tilde{x}^{t-1}) \,. \quad \text{(A3.11)}$$

Assuming that $\tilde{x}$ exists, then in the limit as $t \to \infty$ $\tilde{x} = \tilde{x}^t = \tilde{x}^{t-1}$ and thus from Equation (A3.11),

$$\tilde{x} = \sum_{i=1}^{N} p_i (b_i + m_i \tilde{x}) . \tag{A3.12}$$

Rewriting Equation (A3.12) gives

$$\tilde{x} = \frac{\sum_{i=1}^{N} p_i b_i}{1 - \sum_{i=1}^{N} p_i m_i} . \tag{A3.13}$$

For the CB model on the BRSE, we have the special case where $N = 2$,

$$f_1(x) = (1 - m_-) r_- + m_- x , \tag{A3.14}$$

$$f_2(x) = (1 - m_+) r_+ + m_+ x , \tag{A3.15}$$

$p_1 = p_-$, and $p_2 = p_+$. In this case Equation (A3.13) can be rewritten as

$$\tilde{x} = \frac{p_+ r_+ (1 - m_+) + p_- r_- (1 - m_-)}{1 - p_+ m_+ - p_- m_-} . \tag{A3.16}$$

Making use of the fact that $p_- + p_+ = 1$, this simplifies to

$$\tilde{x} = r_+ \frac{p_+ (1 - m_+)}{p_+ (1 - m_+) + p_- (1 - m_-)} + r_- \frac{p_- (1 - m_-)}{p_+ (1 - m_+) + p_- (1 - m_-)} . \tag{A3.17}$$

Define

$$M = \frac{1}{1 + \frac{p_- (1 - m_-)}{p_+ (1 - m_+)}} . \tag{A3.18}$$

Then Equation (A3.17) can be rewritten as:

$$\tilde{x} = r_+ M + r_- (1 - M) . \tag{A3.19}$$

Higher order moments can also be computed in this way. For example, the variance of the asymptotic distribution can be shown to be

$$\sigma^2 = \frac{\sum_{i=1}^{N} p_i \left(b_i + m_i \tilde{x}\right)^2 - \tilde{x}^2}{1 - \sum_{i=1}^{N} p_i m_i^2} . \tag{A3.20}$$