ISSN 1471-0498

DEPARTMENT OF ECONOMICS

DISCUSSION PAPER SERIES

Reference Points and Learning

Alan Beggs

Number 767
November 2015

Manor Road Building, Oxford OX1 3UQ

Reference Points and Learning

Alan Beggs
Wadham College
Oxford
OX1 3PN
UK
alan.beggs@economics.ox.ac.uk

Current Version: November 2015

Abstract

This paper studies learning when agents evaluate outcomes in comparison to a reference point. It shows that certain models of reinforcement learning lead to classes of recursive preferences.

Keywords: reference points, reinforcement learning, recursive preferences

JEL Nos: D830, D870

I am grateful to participants in the Transatlantic Theory Workshop in Paris in September 2014 for helpful comments.

1. Introduction

The idea that agents evaluate outcomes relative to reference points has been been influential in the economics literature since the pioneering work of Kahneman and Tversky (1979). It also has support in the literature in neuroscience, as outlined in the stimulating book by Glimcher (2011). How reference points are determined is an open question. Köszegi and Rabin (2006) suggest that reference points are determined by expectations. Glimcher (2011) argues that this theory is consistent with models of learning in neuroscience. This paper explores this link from a theoretical perspective.

In the Köszegi and Rabin (2006) approach expectations, and so reference points, are determined by rational expectations. In many environments the assumption of rational expectations seems too strong and it seems more reasonable to assume that expectations and reference points are determined adaptively.

In the literature in neuroscience much interest has focused on modeling learning by reinforcement learning. An agent is assumed to learn in a dynamic but stationary environment of the kind studied in dynamic programming. Simple procedures can lead the agent to learn values of policies, and indeed optimal actions, even if she is unaware of the true stochastic process governing the environment and the relationship between actions and payoffs. In these papers agents form expectations and adjust them linearly according as outcomes are above or below their expectations or reference points. This paper studies generalizations of these models where the relationship between gains and losses need not be symmetric or even linear, as suggested by Kahneman and Tversky (1979) in the context of prospect theory.

Agents are therefore assumed in the current paper to learn according to models of reinforcement learning inspired by those in this neuroscience literature. It is, however, assumed that agents evaluate losses and gains using gain-loss functions which need not be linear. It shows that if an agent learns in a stationary environment then her preferences over policies will converge to recursive preferences of the kind introduced by Epstein and Zin (1989). In the general case convergence is local but global convergence can be shown in some cases. Such preferences have been widely studied in macroeconomics (see for example

Backus et al. (2005)). These are sometimes regarded as rather exotic but the current paper shows they emerge as a result of simple learning procedures and so may provide some motivation for their use.

The paper also shows that under such preferences action choice can be represented as agents seeking to maximize gains or minimize losses relative to reference points as in the behavioral literature. The reference points are, however, the result of long-run learning and so agents respond rationally to shocks given their induced preferences.

The models of reinforcement learning studied here are somewhat different to those familiar in the economics literature from the work of Erev and Roth (1998). In that literature the focus is on simple rules for a single player attempting to learn in a static environment or in a static game interacting with other players. Formal analyses of their convergence properties of these models can be found in Beggs (2005) and Hopkins and Posch (2005). The models here instead analyze a single player learning in a dynamic, but stationary, environment and draw inspiration from the models of reinforcement learning in the of tradition Sutton and Barto (1998) in machine learning, which in turn have heavily influenced neuroscience.

In the neuroscience literature Niv et al. (2012) find that a model with a piecewise-linear gain-loss function may fit neural data better than conventional models. They use the model of Mihatsch and Neuneier (2002) from the machine-learning literature literature on risk-sensitive reinforcement learning. The current paper extends this work to general non-linear loss-gain functions and gives it an economic interpretation. In recent independent work in the machine learning literature Shen et al. (2014) have also considered non-linear loss-gain functions but they allow for a less general class than those considered here. In addition, they give a different interpretation and do not make the link with Epstein and Zin (1989) preferences.

In the economics literature the paper closest to the current one is probably Sarver (2012). He considers a model where consumers may gain utility from anticipation if they choose a high reference point but must balance these against losses from realized outcomes below the reference point. He gives an axiomatic characterization of the resulting preferences. The model here is one of learn-

ing rather optimal anticipation and the focus is on the convergence of learning schemes rather than axiomatic characterizations. Related literature is discussed in more detail in the body of the paper.

The paper proceeds as follows. Section 2 outlines the background on learning from neuroscience to motivate the models studied. Section 3 outlines the basic environment, which is one of stationary dynamic programming. Section 4 examines the special case of learning in a static environment to improve intuition. It shows that an agent using reinforcement learning with a non-linear loss-gain function can be thought of as estimating a generalized certainty-equivalent of the payoffs of actions. These generalized certainty equivalents usually do not coincide with the standard certainty equivalent but belong to the class introduced by Chew (1989). Examples of the resulting preferences are given. These include disappointment aversion (Gul (1991)) but also less familiar ones.

Section 5 presents the main results. It studies reinforcement learning in a dynamic environment. It shows that if agents use reinforcement learning with non-linear loss-gain functions then their preferences over policies converge to recursive preferences of the kind introduced by Epstein and Zin (1989). Local convergence to recursive preferences is shown in the case of general gain-loss functions under some mild conditions. Global convergence is shown when losses and gains only depend on the difference between outcomes and reference points.

Section 6 considers extensions. In Section 5 it is assumed that preferences are intertemporally separable. Section 6 shows that this assumption can be relaxed and the local convergence results extended to general Epstein-Zin preferences. It also discusses robustness of the results to other assumptions, including the form of the reference point and the timing of shocks.

Section 7 discusses the implications for choice of actions and also for learning the optimal policies. Results for local and global convergence are again given. It is shown that optimal actions can be interpreted as maximizing gains relative to reference points. Section 8 concludes.

2. Background

Consider a subject who receives a signal s and a random reward R, which may depend on s. In classical conditioning, interest centers on the extent to

which the subject learns to predict the reward. If the prediction by the subject at time t on receiving signal s is W_t and R_t is the reward received at time t, then a natural learning model is

$$W_{t+1} = W_t + \alpha_t(R_t - W_t) \tag{1}$$

This is essentially the Rescorla-Wagner model in psychology (see for example Dayan and Abbott (2001)). That is the subject raises his prediction if the reward is greater than the prediction and lowers it otherwise. α_t is a parameter which determines the rate of adjustment.

The Rescorla-Wagner model gives a reasonable explanation of some features of conditioning (see for example Dayan and Abbott (2001)).[1] A situation it does not fit so well is one where rewards may occur at different points in time. A signal may predict that rewards will arise in future and so its occurrence may affect the agent's expectation of reward even if there is no immediate payoff.

To model this situation suppose that signals or states follow some stationary Markov chain and that the agent is interested in his total expected discounted reward from the present onwards:

$$E(\sum_{t=0}^{\infty} \beta^t R_t)$$

Assume that the reward depends on the current state but not otherwise on time. If the current state is s and $V(s)$ is his expected discounted reward then a standard argument shows that

$$V(s) = R(s) + \beta E V(s') \tag{2}$$

where s' is the random state tomorrow.

Suppose that at time t the state is s_t and at time $t+1$ the state is s_{t+1}.

$$\delta_t = R(s_t) + \beta V(s_{t+1}) - V(s_t) \tag{3}$$

can be thought of as an estimate of the extent to which realized payoffs differ from expectations or more poetically as a measure of disappointment or elation.

[1] Much of its interest derives from its explanation of phenomena involving multiple stimuli, which it assumes affect rewards additively.

Equation (3) suggest a learning rule. Let $V_t(s)$ be the current estimate of payoffs in state s. Then a natural learning rule is

$$V_{t+1}(s_t) = V_t(s_t) + \alpha_t \delta_t \tag{4}$$

V is left unchanged for states other than s_t. α_t is a parameter governing the extent of adjustment each period. This is somewhat like value-function iteration except that realized values rather than expectations are used and values are adjusted only gradually.

This model, and variants, of it have attracted much attention since the work of Schultz (1998) suggesting that patterns of dopamine activation in the brain, which are thought to represent reaction to rewards, follow a pattern similar to that suggested by (4). This suggests that the brain is forming expectations of reward in the way suggested by this equation.

The model in (4) is known as temporal difference learning in the literature on machine learning (see for example Sutton and Barto (1998)). If α_t is chosen appropriately then the values of V_t converge to those satisfying (2). In the learning literature this result is of considerable interest because it implies that the true values can be learned without the probabilities governing the evolution of state being known (and without any attempt to estimate them).

The model can be modified to allow for an optimal choice of action to be learned. One popular model is so-called Q-learning. Let $Q(a, s)$ be the payoff to taking action a in state s if the optimal strategy is followed in future. Then

$$Q(a, s) = R(a, s) + \beta E V(s') \tag{5}$$

If the policy is optimal then $V(s) = \max_a Q(a, s)$ for all s so this is equivalent to

$$Q(a, s) = R(a, s) + \beta E \max_{a'} Q(a', s') \tag{6}$$

Q-learning takes this equation and uses an analogous procedure to (4). If action a_t is played in state s_t at time t then

$$Q_{t+1}(a_t, s_t) = Q_t(a_t, s_t) + \alpha_t(a_t, s_t) \left(\beta \max_{a'} Q_t(a', s') + R(a_t, s_t) - Q_t(a_t, s_t) \right) \tag{7}$$

Values of Q for other state-action pairs are left unchanged. $\alpha_t(a,s)$ is an adjustment parameter. This rule can be shown to converge to the values corresponding to the optimal policy provided sufficient experimentation is ensured. One example would be the so-called ϵ-greedy rule: with probability $1-\epsilon$ play the action with the highest value of $Q_t(a_t,s_t)$, with probability ϵ all action are equally likely to be played. Another possibility is to choose randomly with a logit probability function (so-called softmax) : action a_t is played with probability $\exp(\delta_t Q_t(a_t,s_t)) / \sum_a \exp(\delta_t Q_t(a,s_t))$, where δ_t tends to zero at an appropriate rate.[2]

Other methods for learning the optimal action are possible and is unclear whether Q-learning is the best model for describing learning in the brain. Other less sophisticated models of learning may be more appropriate — see for example Niv and Montague (2009) for a discussion. For simplicity the paper will assume this is used. The paper will in any case for the most part concentrate on the implicit preferences described by temporal difference learning rather than action learning.

Niv et al. (2012) find some evidence that a model incorporating a kind of loss aversion or risk-sensitivity may fit data from brain scans better (4). Following Mihatsch and Neuneier (2002) in the reinforcement learning literature they examine a variant of (4) where over-predictions are weighted more heavily than under-predictions:

$$V_{t+1}(s_t) = V(s_t) + \alpha_t \Phi(\delta_t) \tag{8}$$

where

$$\Phi(x) = \begin{cases} bx & x < 0 \\ cx & x > 0 \end{cases} \tag{9}$$

with $b > c > 0$. That is over-predictions cause more disappointment than an under-prediction of the same magnitude causes joy. They find this fits their data better than (4) or a version in which the learning rule remains unchanged but payoffs have an expected utility form ($U(R)$).

Niv et al. (2012) and Mihatsch and Neuneier (2002) do not offer an interpretation of the value function to which this procedure converges. They also

[2] In the economics literature inspired by Erev and Roth (1998) one of the key issues is to show that the reinforcement rules used do guarantee enough experimentation — see for example Beggs (2005).

restrict attention to piecewise linear loss-functions. This paper investigates the interpretation of this learning rule and shows that the limiting value function can be interpreted as one of the Epstein and Zin (1989) class. It shows this interpretation holds for general loss functions.

3. General Model

The framework the agent operates in is the standard one of infinite horizon stationary dynamic programming:

- There is infinite number of discrete time periods, $t = 0, 1, 2, \ldots$.
- Each period she must choose one of a finite number of actions from the set $A = \{1, \ldots, n\}$.
- There is a finite number of states, $S = \{1, \ldots, m\}$.
- The payoff to action a in state s is $R(a, s)$.
- If action a is chosen in state s the state will be s' next period with probability $p^a_{ss'}$
- The agent has discount factor β.

A (deterministic) stationary policy, π is a function $\pi : S \longrightarrow A$. Its expected discounted payoff, or value, in state s, $V^\pi(s)$, satisfies the recursive equations:

$$V^\pi(s) = R(\pi(s), s) + \beta \sum_{s'=1}^{m} p^{\pi(s)}_{ss'} V^\pi(s') \quad i = 1, \ldots, m \tag{10}$$

or equivalently

$$V^\pi(s) = R(\pi(s), s) + \beta E\left(V^\pi(s')|s\right) \tag{11}$$

The value function, V, of the optimal policy satisfies the Bellman equation:

$$V(s) = \max_a R(a, s) + \beta E\left(V(s')|s, a\right) \tag{12}$$

To apply these equations to find the optimal decision rule, the agent needs to know both the transition probabilities, $p_{ss'}$, and the reward function R. The literature on reinforcement learning shows that the optimal rule can be learned without these being known.

It will also be assumed that

- Under each policy π, the set of states forms an ergodic Markov chain.

This ensures that all states are eventually visited, so it is possible to learn about them. The assumption that the number of states is finite can be relaxed by using function approximation (see for example Sutton and Barto (1998)) but will be maintained in this paper.

Temporal-difference learning proceeds by iteratively updating an estimate of a value of an current policy π. Let V_t^π be the current estimate of V^π. Let s_t be the state at time t, a_t the action specified by policy π and s_{t+1} the state at time $t+1$. Then

$$V_{t+1}^\pi(s) = \begin{cases} V_t^\pi(s_{t+1}) + \alpha_t \left(R(a_t, s_t) + \beta V_t^\pi(s_{t+1}) - V_t^\pi(s_t)\right) & s = s_t \\ V_t^\pi(s) & \text{otherwise} \end{cases} \tag{13}$$

where α_t is a parameter. It is assumed that both the realized reward and subsequent state are observed by the agent.

If $V_t^\pi = V^\pi$, then the expected change in V_t^π is zero from (12). If α_t tends to zero at an appropriate rate it can be shown that V_t^π converges to V^π, as will be discussed further in Section 4.

As discussed in the previous section, the optimal policy can be learned if temporal-difference learning is combined with an appropriate learning rule, for example Q-learning. This will be explained and discussed further in Section 7.

The term $R(a_t, s_t) + \beta V_t^\pi(s_{t+1}) - V_t^\pi(s_t)$ can be thought of as estimate of extent to which realised utility exceeds or is less than the current expected of total utility in state s_t, $V_t^\pi(s_t)$ or in other words of losses or gains in comparison with expectations. These expectations are revised until the expected losses and gains are zero. A loss averse agent may weight losses and gains differently so a natural generalization would be to replace (13) by

$$V_{t+1}^\pi(s) = \begin{cases} V_t^\pi(s_{t+1}) + \alpha_t \Phi\big(R\left(a_t, s_t\right) + \beta V_t^\pi\left(s_{t+1}\right) - V_t^\pi(s_t)\big) & s = s_t \\ V_t^\pi(s) & \text{otherwise} \end{cases} \tag{14}$$

where Φ is a function measuring losses and gains or more generally

$$V_{t+1}^\pi(s) = \begin{cases} V_t^\pi(s_{t+1}) + \alpha_t \Psi\big(\beta V_t^\pi\left(s_{t+1}\right), V_t^\pi(s_t) - R\left(a_t, s_t\right)\big) & s = s_t \\ V_t^\pi(s) & \text{otherwise} \end{cases} \tag{15}$$

where Ψ is again a gain-loss function. In this formulation the gain (or loss) experienced if the outcome is x and the anticipated payoff is μ is $\Psi(x, \mu)$. Expectations of future payoffs at time t, $V_t^\pi(s_t) - R(a_t, s_t)$, are compared with a revised estimate at time $t+1$, $\beta V_t^\pi(s_{t+1})$. The interpretation of (15) is discussed in more detail in Section 5.

It will be assumed that

Assumption 1 *(i)* $\Psi(x, x) = 0$ *for all* x, *(ii)* $\Psi(x, \mu)$ *is increasing in* x *and decreasing in* μ, *(iii)* Ψ *is Lipschitz in* x, *and (iv) there exist* k *and* K, $K \geq k > 0$, *such that for all* μ, μ', $\mu' \neq \mu$, *and for all* x

$$k \leq \left| \frac{\Psi(x, \mu) - \Psi(x, \mu')}{\mu - \mu'} \right| \leq K \qquad (16),$$

In the case when $\Psi(x, \mu) = \Phi(x - \mu)$, this is implied by

Assumption 2 Φ *satisfies* $\Phi(0) = 0$ *and there exist* $m > 0$ *and* $M > 0$ *such that for all* $x \neq y$,

$$m \leq \frac{\Phi(y) - \Phi(x)}{y - x} \leq M$$

.

That is Φ is Lipschitz-continuous and has slope bounded away from zero, which is clearly satisfied by the piecewise-linear form.

Recently Shen et al. (2014) have independently studied the case of gain-loss functions of the form $\Phi(x - \mu)$ and noted, as here, that Mihatsch and Neuneier (2002)'s convergence results can be extended to them. They do not, however, study general gain-loss functions, as is done here. They offer an interpretation in terms of risk-sensitive programming with risk-measures (see for example Shapiro et al. (2014) Chapter 6 and Rusczyński (2010)) rather than the economic one given here. In particular, they do not make the link with Epstein and Zin (1989) preferences.

4. Learning in a Static Environment

This section takes a diversion and considers a simpler model in order to understand the properties of the learning rules studied. The model of the previous section is complicated in that the agent's reference point or expectations

vary with the current state, as it helps to predict the future. In this section it is assumed that the future is independent of the present, so the reference point for the agent is simply a scalar. This simpler case will help with the interpretation of the general model. In particular it shows that the agent can be thought of estimating a kind of certainty equivalent but not necessarily of the standard sort.

More formally, it is assumed that the transition probabilities have the property that[3]

- $p^a_{ss'}$ is independent of s for all a and s'.

It will also be assumed, for simplicity, that the action does not affect payoffs directly but only through their effect on transition probabilities, that is

- $R(a, s) = R(s)$ for all a.

These assumptions imply that the problem is essentially static, as optimal actions are not affected by the state. Since future payoffs are independent of the state, expectations or reference points will also be independent of the current state. The learning rule (15) can be simplified to

$$V^a_{t+1} = V^a_t + \alpha_t \Psi \left(R(s), V^a_t \right) \tag{17}$$

Note that as the problem is now effectively a single period one the discount factor, β, is irrelevant. Equivalently one can consider the agent as caring about long-run average payoffs.

Assume that

Assumption 3 *$\alpha_t \geq 0$, $\sum_t \alpha_t = \infty$ and $\sum_t \alpha_t^2 < \infty$.*

That is, the adjustment parameter, α_t, becomes small fast enough that the influence of random shocks dies out but not so quickly that initial conditions dominate. The assumption is satisfied by $\alpha_t = 1/t$, for example.

Standard results in stochastic approximation show that the long-run behavior of the learning process can be determined by looking at the ordinary differential equation:

$$\dot{V}^a = E_s \Psi(R(s), V^a) \tag{18}$$

[3]The assumption that the number of states is finite is not essential for the results of this section.

Expectations also depend on a as the action taken affects the probabilities of the states but this will suppressed from the notation. Under the assumptions made, V^a will converge to the unique stationary point of this system. That is the long-run estimate of V^a is the unique solution to

$$E_s \Psi(R(s), V^a) = 0 \tag{19}$$

Theorem 1 *Under Assumption 1 and Assumption 3, V_t^a converges to the unique solution of (19) with probability 1.*

V^a can be regarded as a certainty-equivalent of action a. If $\Psi(x, \mu) = U(x) - U(\mu)$, where U is a utility function, then V^a is simply the usual certainty-equivalent $U^{-1}(E_s U(R(s)))$. More generally, however, V^a belongs to the class of generalized certainty equivalents defined by Chew (1989).[4]

If actions are ranked by their long-run certainty equivalent, then (19) implies that the class of preferences encompassed by this formulation are exactly those introduced by Chew (1989). V^a is the generalized certainty equivalent of $R(s)$.

These preferences can also be understood from the point of view of robust estimation, in particular of M-estimation (see Huber (1981)). If $M(x, \mu)$ represents the loss experienced by the statistician if his estimate is μ and the observed value is x then, then if $\Psi = -\partial M / \partial \mu$, (19) are the first-order conditions for minimizing the expected loss.

Let GC denoted the generalized certainty equivalent defined by (19). For future reference some of its properties are recorded below. It follows immediately from Assumption 1 that

Lemma 1 *(i) (constancy): if $R(s) = k$ for all s, then $GC(R(s)) = k$. (ii) (monotonicity) if $R'(s) \geq R(s)$ for all s then $GC(R'(s)) \geq GC(R(s))$.*

The certainty equivalent will be said to be *translation-subinvariant* if $GC(R(s) + k) \leq GC(R(s)) + k$. It is well known that (see for example Appendix C of Marinacci and Montruchio (2010)) that

[4] See also Dekel (1986).

Lemma 2 *If the certainty-equivalent is derived from expected utility then it is translation subinvariant for all risks if and only if the utility function displays increasing-absolute risk aversion.*

Some examples of certainty equivalents not derived from expected utility are:

Example 1 *Kahnemann-Tversky Form:* $\Psi(x, \mu) = \Phi(x - \mu)$.

In this case V^a satisfies

$$E_s \Phi(R(s) - V^a) = 0 \tag{20}$$

Given that $\Phi(0) = 0$, V^a represents the amount of money to be subtracted from ex post payoffs to make the decision-maker indifferent between taking and not taking the gamble. In the literature on insurance this is sometimes referred to as the 'zero-utility principle' (see for example Buhlmann (1970)). If the agent has utility function U, the usual certainty equivalent is defined as

$$C^a = U^{-1}\left(E_s U\left(R\left(s\right)\right)\right) \tag{21}$$

That is the usual certainty equivalent is the amount of money which gives the agent as much utility as taking the lottery, so is in a sense the equivalent variation, while (20) is the compensating variation.[5] Pratt (1964) refers to the certainty equivalent as the selling price of a lottery and (19) as the buying price. In general the two notions differ unless the utility function is exponential, that is displays constant absolute risk aversion.[6]

As a piece of shorthand, let CCE (compensating certainty equivalent) denote the certainty equivalent defined by (20). One can write $V^a = CCE(R(s))$

For future reference, some immediate properties of CCE are recorded in the following lemma:

Lemma 3 *(i) (constancy): if $R(s) = k$ for all s, then $CCE(R(s)) = k$. (ii) (monotonicity) if $R'(s) \geq R(s)$ for all s then $CCE\left((R'(s)\right) \geq CCE((R(s))$. (iii) (translation invariance) If $R'(s) = R(s) + k$ for all s then $CCE(R'(s)) = CCE\left((R(s)\right) + k$.*

[5] See for example Pope and Chavas (1985) or Kimball (1990) for further discussion.
[6] As noted for example by LaValle (1968) and Raiffa (1968).

Example 2 *Piecewise-Linear Kahneman-Tversky Form*

Suppose Φ has the following piecewise-linear form with $0 < \lambda < 1$:

$$\Phi(x) = \begin{cases} \lambda x & x > 0 \\ x & x \leq 0 \end{cases} \tag{22}$$

V^a is then the solution to

$$E_s I(R(s) < V^a)(R - V^a) + \lambda E_s I(R(s) > V^a)(R - V^a) = 0 \tag{23}$$

where $I(A)$ denotes the indicator function for the event A. That is V^a is an expectile. An expectile can be thought of as a generalization of the idea of a quantile but applied to expectations rather than probability. They appear in the literature on estimation with asymmetric least-squares (see Newey and Powell (1987)). In particular V^a minimizes the asymmetric least-squares error, where errors are weighted differently according as outcomes are less or greater than the prediction:

$$\min_{V_a} E_s I(R(s) < V^a)(R - V^a)^2 + \lambda E_s I(R(s) - V^a)(R - V^a)^2 \tag{24}$$

This is consistent with the idea of loss-aversion: the decision-maker is more concerned with losses when the outcome is less than the predicted value than with gains when the outcome exceeds the prediction.

Example 3 *Disappointment Aversion*

Suppose

$$\Psi(x, \mu) = \begin{cases} U(x) - U(\mu) & x \leq \mu \\ \lambda(U(x) - U(\mu)) & x > \mu \end{cases} \tag{25}$$

where U is concave and increasing and $\lambda < 1$. Then V^a is the solution to

$$E_s I(R(s) < V^a)(U(R) - U(V^a)) + \lambda E_s I(R(s) > V^a)(U(R) - U(V^a)) = 0 \tag{26}$$

That is V^a is the generalized-certainty equivalent in Gul (1991)'s theory of disappointment-aversion. If U is linear, $U(x) = x$, this coincides with the linear Kahneman-Tversky form. Equivalently, with a linear KT gain-loss function the two notions will coincide if outcomes, x, are measured in utils. With non-linear gain-loss functions, the two will differ.

Sarver (2012) proposes a theory of reference points resulting from optimal anticipation: consumers derive utility from the anticipation caused by a high expectation or reference point but must balance this against the loss caused by disappointment of the outcome being below the reference point. If losses are given by the KT-form then this implies choosing the reference point, μ, to maximize $\mu + E\Phi(x - \mu)$. A similar notion appears in the literature on risk measures under the name of the 'optimized certainty equivalent' (see for example Ben-Tal and Teboulle (2007)). In the case of a piecewise linear gain-loss function, Sarver (2012) shows that the resulting reference point is quantile, which differs from the expectile found here. More generally, the two approaches are different. Sarver (2012) gives an axiomatic characterization of preferences. The approach here is not axiomatic but asks what preferences may emerge if the decision-maker adjusts reference points to minimize expected losses.

Köszegi and Rabin (2006) suggest that reference points are determined by expectations. They, examine, however, a more complex notion where agents consider the entire distribution of returns rather than a single reference point. Gains and losses are evaluated by comparing the realized outcome to every outcome considered possible. Their theory bears some resemblance to earlier models of regret and disappointment aversion and Fishburn's SSB theory (see for example Fishburn (1988)). A detailed discussion of the relationship of their model to other theories can be found in Masatlioglu and Raymond (2014)

5. Learning in a Dynamic Environment

The paper now returns to the case of dynamic learning. It shows that under some conditions the learning rules adopted imply that preferences converge to those belonging to the Epstein-Zin class of recursive preferences. In the general case the convergence is local in the sense that if preferences start close enough to their equilibrium values then the system converges to them with arbitrarily high probability. In the case of gain-loss functions of the Kahneman-Tversky form then convergence is with probability one for any starting values.

As laid out in section 3, the agent adopts the following learning rule:

$$V_{t+1}^{\pi}(s) = \begin{cases} V_t^{\pi}(s_{t+1}) + \alpha_t \Psi\big(\beta V_t^{\pi}(s_{t+1}), V_t^{\pi}(s_t) - R(a_t, s_t)\big) & s = s_t \\ V_t^{\pi}(s) & \text{otherwise} \end{cases} \tag{27}$$

or equivalently

$$V_{t+1}^{\pi}(s) = V_t^{\pi}(s) + \alpha_t^s \Psi\big(\beta V_t^{\pi}(s_{t+1}), V_t^{\pi}(s_t) - R(a_t, s_t)\big) \tag{28}$$

where

$$\alpha_t^s = \begin{cases} \alpha_t & s = s_t \\ 0 & \text{otherwise} \end{cases}$$

This reflects the fact that values for only one state are updated each period.

(28) may be interpreted as saying that in state s_t, $V_t^{\pi}(s_t) - R(a_t, s_t)$ are the player's expectations of future payoffs which she compares to $\beta V_t^{\pi}(s_{t+1})$, her updated estimate of future payoffs, to determine her realized gain or loss at time $t+1$. In the special case when $\Psi(x, \mu) = \Phi(x - \mu)$, (28) can be rewritten as

$$V_{t+1}^{\pi}(s) = V_t^{\pi}(s) + \alpha_t^s \Phi\big(R(a_t, s_t) + \beta V_t^{\pi}(s_{t+1}) - V_t^{\pi}(s_t)\big) \tag{29}$$

In this case one can also interpret the player as comparing current expectations, $V^{\pi}(s_t)$ with an estimate of overall payoffs, $R(a_t, s_t) + \beta V(s_{t+1})$.

Attention will focus on the case of a fixed policy so the superscript, π, will be dropped. $a(s)$ will denote the action to be taken in state s. As in the previous section, standard results in stochastic approximation[7] show that the long run behavior of the system is governed by the differential equation

$$\dot{V}(s) = H_s^1(V) = \delta_s E_{s'} \Psi\left(\beta V(s'), V(s) - R(a(s), s)\right) \tag{30}$$

where δ_s reflects the amount of time spent in state s, since states are only updated one at a time.

The stationary points of this equation are given by

$$E_{s'} \Psi\left(\beta V(s'), V(s) - R(a, s)\right) = 0 \tag{31}$$

That is $V(s) - R(a, s)$ is the generalized certainty equivalent of $\beta V(s)$, so that in the notation of the previous section

$$V(s) = R(a(s), s) + GC(\beta V(s')) \tag{32}$$

[7] See for example Borkar (2008) Section 7.4.

This is essentially a Bellman equation but with the certainty equivalent operator of the previous section replacing the expectation operator. The β occurs inside the certainty equivalent operator but replacing βV by $\tilde{V}$ and βR by $\tilde{R}$, one can check can see that (32) is equivalent to

$$\tilde{V}(s) = \tilde{R}(a, s) + \beta GC(\tilde{V}(s')) \tag{33}$$

This is similar to the kind of equation familiar in the models of risk-sensitive control studied in the engineering literature (see for example Whittle (1996)) and introduced into economics by Hansen and Sargent (see for example Hansen and Sargent (2008)). The original formulation was non-recursive but Hansen and Sargent (1995) suggested a formulation of risk-sensitive control with

$$\tilde{V}(s) = \tilde{R}(a(s), s) + \beta\, C\left(\tilde{V}(s')\right) \tag{34}$$

with C denoting the usual certainty equivalent. In the applications studied in engineering and by Hansen and Sargent, C is usually derived from exponential utility, so

$$C\left(V(s')\right) = -\frac{1}{\gamma} \ln E\left(-\gamma \exp V(s')\right) \tag{35}$$

The above formulation allows for a general certainty equivalent and is an instance of Epstein and Zin (1989) preferences with intertemporally additive utilities.

A standard sufficient condition (see for example Marinacci and Montrucchio (2010)), for (33) to have a unique solution is that the generalized certainty-equivalent be translation sub-invariant (see Section 4), which as noted in Lemma 2 corresponds in the expected-utility framework to increasing absolute risk-aversion:

Lemma 4 *Under Assumption 1, (32), which is equivalent to (31), has a unique solution if GC is translation sub-invariant.*

Each potential value function is a function defined on S. Since S is finite one can simply regard each value function as an element Euclidean space, with dimension the number of states in S, and measure the distance between value functions by Euclidean distance. One then has

Theorem 2 *Under Assumption 1 and Assumption 3 if Ψ is C^1 and the corresponding certainty-equivalent translation-subinvariant, then for any $\epsilon > 0$, there is a neighborhood of the unique solution to (31), $\{V^*(s)\}$, such that if the initial values of V lie in this neighborhood then the values of V_t generated by (27) converge to V^* with probability at least $1 - \epsilon$.*

Translation sub-invariance implies that the Jacobian of H^1 is a dominant-diagonal matrix, that is own effects offset cross-effects, with a negative diagonal at $\{V^*(s)\}$. This in turn implies that the differential equation (30) is locally asymptotically stable there. Standard results in stochastic approximation now yield the result. The assumption of translation sub-invariance is only sufficient for the convergence and numerical results suggest it may not be necessary. This is discussed further in Section 6.

The C^1 assumption is made to rule out complications due to non-smoothness of the right-hand side of (30). The proof proceeds by looking at local linear stability near the equilibrium point but the assumption can probably relaxed, as it is in the next result. It rules out kinks in the loss function but such a loss function can be well approximated be a smooth one which is very curved near the kink points.

It is not clear if the result can be extended to global convergence in general. It can be when $\Psi(x, \mu) = \Phi(x - \mu)$. In this case (31) specializes to

$$E_{s'} \Phi \left(R(a(s), s) + \beta V(s') - V(s) \right) = 0 \tag{36}$$

and (32) becomes

$$V(s) = R(a(s), s) + CCE(\beta V(s')) \tag{37}$$

where CCE denotes the compensating certainty equivalent introduced in the last section. As noted there, CCE is translation-invariant, so Lemma 4 implies that (37) has a unique solution.

Theorem 3 *Under Assumption 2 and Assumption 3, the values of V_t generated by (14) converge with probability 1 to the unique solution of (37).*

In essence this is because if one considers the corresponding differential equation

$$\dot{V}(s) = \delta_s E_{s'} \Phi \left(R(a(s), s) + \beta V(s') - V(s) \right) \tag{38}$$

then the Jacobian of the right-hand side is globally a dominant-diagonal matrix, which implies global convergence as with suitable rescaling the right-hand side can be rewritten as $\delta_s \left(T(V)(s) - V(s)\right)$, where T is a contraction. The formal proof is in the Appendix and is an easy extension of the one used by Mihatsch and Neuneier (2002) in the piecewise-linear case. They interpret T as a generalization of the usual dynamic programming operator. It is not clear, however, what it represents in economic terms. Armed with the results of the last section, one can, however, interpret the model easily.

The key feature of the models in this section is that the limiting preferences can be written as satisfying equations which are linear in the probabilities (see for example (31)). These equations are therefore amenable to stochastic approximation, which essentially estimates expectations by calculating sample averages recursively, even if the probabilities are unknown. This linearity is a well known useful feature of the certainty-equivalent equations from the Chew-Dekel class. The direct recursive formulation in (37) is in general non-linear in the probabilities and so is not suitable for stochastic approximation.

As noted in the previous section, if outcomes are measured in utility then the piecewise-linear loss-gain function delivers the certainty-equivalent corresponding to Gul (1991)'s version disappointment aversion. The result in Theorem 3 implies that version of Epstein and Zin (1989) preferences with additive utility and disappointment aversion are relatively easy to learn.

6. Extensions

This section discusses some extensions and limitations of the results. The first sub-section shows that the results can be extended to the general case of Epstein-Zin preferences. The second discusses the case of random payoffs and the issues of time consistency encountered if these are allowed. The third considers the form of the reference point and the fourth robustness to other assumptions.

6.1 Non-additive Intertemporal Preferences

The full family of Epstein-Zin preferences allows for non-additive intertemporal preferences. It is shown in this section that these can also be interpreted as the outcome of a learning rule.

Epstein and Zin (1989) assume that recursive preferences have the form

$$V(s) = W(R(a(s), s), GC(V(s')) \tag{39}$$

where $W(x, y)$ is an inter-temporal aggregator and GC a generalized certainty equivalent. A popular choice for the aggregator W is the CES form

$$W(x, y) = (x^\rho + by^\rho)^{1/\rho} \tag{40}$$

It will be assumed that

Assumption 4 *W is increasing in x and y and for some $\beta < 1$, $|W(x, y) - W(x, y')| \leq \beta|y - y'|$ for all x,y,y'. GC is translation-subinvariant.*

This is a standard assumption in the literature on recursive preferences (see for example Marinacci and Montruchio (2010)) and the assumption on W is satisfied by the CES form if $\rho > 1$ and $b < 1$. It guarantees that there is a unique solution to (39).

Provided W is invertible in the second argument one can apply stochastic approximation to this form of preferences. If $z = W(x, y)$ let $y = \tilde{W}(x, z)$ be the inverse function. $\tilde{W}$ is well-defined in the CES case.

(39) can be re-written as

$$\tilde{W}(R(a(s), s), V(s)) = GC(V(s')) \tag{41}$$

or equivalently

$$E_{s'} \Psi \left(V(s'), \tilde{W} \left(R \left(a \left(s \right), s \right), V \left(s \right) \right) \right) = 0 \tag{42}$$

and one can apply the learning scheme

$$V_{t+1}^{\pi}(s) = \begin{cases} V_t^{\pi}(s_{t+1}) + \alpha_t \Psi \left(\beta V_t^{\pi} \left(s_{t+1} \right), \tilde{W} \left(R \left(a_t, s_t \right), V_t^{\pi}(s_t) \right) \right) & s = s_t \\ V_t^{\pi}(s) & \text{otherwise} \end{cases} \tag{43}$$

As in the previous case, the asymptotic behavior of the system is related to that a differential equation, in this case

$$\dot{V}(s) = \delta_s E_{s'} \Psi(V(s'), \tilde{W}(R(a(s), s), V(s)) \tag{44}$$

where again δ_s denotes the long-run proportion of time spent in state s.

Let subscripts denote partial derivatives:

Assumption 5 *In addition to Assumption 4, $\tilde{W}$ is well-defined. Ψ and $\tilde{W}$ are C^1, and $\tilde{W}_2 \geq \tilde{\beta} > 1$ for all x,y, some $\tilde{\beta}$.*

These are standard assumptions in the literature on recursive preferences (see for example Marinacci and Montruchio (2010), who refer to this as the Blackwell case). W satisfies this in the CES case if $\rho > 1$ and $b < 1$.

One then has

Theorem 4 *Under Assumption 3, and Assumption 5, then for any $\epsilon > 0$, there is a neighborhood of the unique solution to (42), $\{V^*(s)\}$, such that if the initial values of V lie in this neighborhood then the values of V_t generated by (43) converge to V^* with probability at least $1 - \epsilon$.*

Thus general Epstein-Zin preferences can be regarded as the outcome of a learning procedure. The process is here is, however, arguably rather complex.

6.2 Random Payoffs and Time-Consistency

It has been assumed that rewards are a deterministic function of the current state. In dynamic programming it is common to assume that rewards may be random. For example, one could assume as in Stokey et al. (1989), write the reward as a function of the current and future state, $R(a, s, s')$. One can still apply the learning algorithm to this case, as do Mihatsch and Neuneier (2002), but the resulting preferences will be time-inconsistent in general.

In more detail

Assumption 6 *The assumptions are unchanged except current rewards are a function of the current and future state: $R(a, s, s')$.*

For simplicity consider the case of preferences of the Kahneman-Tversky form ($\Phi(x - \mu)$). The issues are the same in the general case.

Theorem 5 *The values of V_t generated by (14) converge to the unique solution of*

$$E\Phi\left(R(a, s, s') + \beta V(s') - V(s)\right) = 0 \tag{45}$$

(45) is equivalent to

$$V(s) = CCE\left(R(a, s, s') + \beta V(s')\right) \tag{46}$$

If the latter could be written as

$$V(s) = ER(a, s, s') + CCE\left(\beta V(s')\right) \tag{47}$$

then preferences would be time-consistent.

R cannot, however, be extracted from the CCE even in expectation in general. Unlike (37), therefore, (46) is in general not weakly separable between current and future states, so preferences are not recursive. Decisions will therefore be time-inconsistent in general (see Johansen and Donaldson (1985) for example). Preferences will be time-consistent provided the state can be defined in such a way that any randomness regarding payoffs is resolved in the next period rather than the current one. As is well understood in the Epstein and Zin (1989) and Kreps and Porteus (1978) framework, when uncertainty is resolved is crucial once one goes beyond additively separable expected utility preferences.

6.3 Form of Reference Point in the General Case

In the case of preferences of the Kahneman-Tversky form the loss or gain experienced by the agent if the state is s' is $\Phi(R(a, s) + \beta V(s') - V(s))$. One could interpret this in two ways. One could regard the reference point, $V(s)$, as measuring total expected payoffs in state s and so it is compared with current payoff plus a revised estimate of future payoffs, $R(a, s) + \beta V(s')$. Another interpretation would be that as payoffs in state s are known the reference point $V(s) - R(a, s)$ measures expected future payoffs and this is compared with a revised estimate one state s' is known. Either interpretation yields the same limiting values for V.

Consider the general case, with time-separable preferences for simplicity. Here the paper has implicitly taken the second interpretation: the gain is $\Psi(\beta V(s'), V(s) - R(a, s))$. If one were instead to take the first interpretation the relevant gain would be $\Psi(R(a, s) + \beta V(s'), V(s))$. In this case expected gain is zero when

$$V(s) = GC\left(R(a, s) + \beta V(s')\right) \tag{48}$$

In this case preferences are no longer of Epstein-Zin form. As in the previous subsection preferences are in general no longer weakly separable, so will

be time-inconsistent. In the Kahneman-Tversky case the certainty-equivalent is translation-invariant so

$$CCE\left(R(a,s)+\beta V(s')\right)=R(a,s)+CCE(\beta V(s')) \tag{49}$$

and the two formulations are equivalent.

Which is more behaviorally appealing in general is open to debate. The view taken here is that once state s is known the agent revises expectations to take into account information on payoffs accrued, so the second is utilized.

6.4 Robustness to Other Assumptions

6.4.1 Assumptions on Preferences and Loss Functions

The assumption of translation-subinvariance of the certainty-equivalent is fairly common in the literature on recursive preferences (see for example Marinacci and Montruchio (2010)) but it would desirable to relax it. One easy relaxation is to note that because of the discounting term (β in (30)), the diagonal dominance assumption will hold provided β is not too close to 1 and translation sub-invariance is not violated too strongly. The translation-invariant will said to translation d-subinvariant if $GC(X+k)\leq G(X)+dk$ for all X and k. The following is an easy corollary of the proof of Theorem 2:

Corollary 1 *Theorem 2 holds if the certainty-equivalent is d-subinvariant with* $d<1/\beta$.

Similarly Theorem 4 can be extended. For the local convergence it is enough that the certainty-equivalent be d-subinvariant at the solution to the dynamic programming equations, $\{V^*(s)\}$, considered but without it holding globally the dynamic programming equations might not have a unique solution.

The assumption of smoothness of Ψ has already been discussed. As noted, although it rules out kinked loss-gain functions, such functions can be arbitrarily well approximated by smooth ones. Similarly in the global convergence results, the requirement that the slope of Φ be bounded above or low rules out power loss-gains functions which have infinite slopes at the origin, which are popular in the literature on reference points, but again these can again be arbitrarily well approximated by ones with large bounded slopes. The assumption also

rules out exponential cost functions and bounded loss functions. It is enough, however, that the assumption holds within the domain of interest. Since payoffs are bounded the optimal values will lie with in a bounded region so it is enough that the assumptions hold in this region. If random shocks, as in Section 6.2, mean that the system may leave this region one could modify the algorithm by truncating it to ensure that it always remains in it, as is common in reinforcement learning — see for example Kushner and Yin (1997) Chapter 12.8 for details.

6.4.2 Step Sizes

The rule for step-sizes in (14) can be generalized. One can for example allow the updated rate for each state to depend on the number of times it has been visited rather than being linked to the total number of periods that have elapsed. One could for example set

$$\alpha^s(t) = \begin{cases} 1/(n_s(t)+1) & \text{if } s \text{ is visited at } t \\ 0 & \text{otherwise} \end{cases} \tag{50}$$

where $n_s(t)$ is the number of visits to state s by time t. This is referred to in the literature as asynchronous updating. Assumption 3 needs to be replaced by

Assumption 7 *With probability 1,* $\sum_t \alpha^s(t) = \infty$ *and* $\sum_t (\alpha^s(t))^2 \leq C$ *for some constant* C *for all* s*.*

This requires that each state is visited infinitely often, which holds here since the chain is ergodic. It is satisfied by (50).

Theorem 6 *Theorem 3 holds if Assumption 3 is replaced by Assumption 7.*

7. Action Choice

This section outlines the implications of the model for action choices and also briefly notes how the learning model can be extended to this case. Recursive preferences have been widely used in finance and macroeconomics — see for example Backus et al. (2005) for a survey. Disappointment aversion in particular has been widely used. One can interpret the current model as providing support for the use of this model. The first subsection outlines what further light the model sheds on behavior. The second subsection briefly outlines how the optimal policy may be learned in this context.

7.1 Implications for Behavior

In the case of preferences of the Kahneman-Tversky form optimal decisions can be represented in a particularly appealing form. The Bellman equation for optimal action choice in this case is

$$V(s) = \max_a R(a,s) + CCE\left(\beta V(s')\right) \tag{51}$$

This is equivalent to V solving the set of equations

$$\max_a \quad E_{s'}\ \Phi\left(R(a,s) + \beta V(s') - V(s)\right) \tag{52}$$

since for any number k and random variable X

$$k \geq CCE(X) \iff E\Phi(X-k) \leq 0 \tag{53}$$

and (51) is equivalent, as CCE is translation-invariant, to

$$V(s) = \max_a\ CCE\left(R(a,s) + \beta V(s')\right) \tag{54}$$

Hence,

Lemma 5 *(51) and (52) have the same, unique solutions.*

In particular, this implies that the optimal choice of action solves

$$a(s) \quad \max_a\ E_{s'}\ \Phi\left(R(a,s) + \beta V(s') - V(s)\right) \tag{55}$$

This gives an intuitive representation in terms of loss-aversion. The optimal action maximizes the expected gain relative to the reference point $V(s)$. Note, however, that the agent's behavior is forward-looking in that she takes into account future losses and gains not directly whether the current period is good or bad. This is made clear in the formulation of (51). To reconcile with (55) note that $V(s)$ already incorporates the fact that the agent may be in a disappointing state.

If the agent did not adjust her reference point in response to the state then (55) would become

$$a(s) \quad \max_a\ E_{s'}\ \Phi\left(R(a,s) + \beta V(s') - \bar{V}\right) \tag{56}$$

or even

$$a(s) \quad \max_a \; E_{s'} \; \Phi \left(R(a,s) + \beta \bar{V} - \bar{V} \right) \tag{57}$$

The agent might then, for example, choose to work less hard if the state were relatively good even though the marginal payoff to effort is higher if Φ is sufficiently concave — compare the literature on income targets in labor supply originated by the work of Camerer et al. (1997). In the forward-looking case this is less clear as the effect will depend on how curvature of Φ affects the marginal impact of effort, in a given state, on the certainty equivalent in future — see (51).

A similar, if less transparent representation can be given in the general case. The standard Bellman equation is

$$V(s) = \max_a \; W \left(R(a,s), GC(V(s')) \right) \tag{58}$$

This can be written equivalently as

$$a(s) \quad \max_a \; E_{s'} \Psi(V(s'), \tilde{W}(R(a,s), V(s))) \tag{59}$$

$\tilde{W}$ is the inverse of temporal aggregator W introduced in the previous section. One has (proof in Appendix):

Lemma 6 *(58) and (59) have the same solutions.*

The form here is more convoluted but (59) can still be regarded as representing choice as maximizing gains relative to a reference point.

7.2 Action Learning

Q-learning can be adapted to this case. Recall from Section 2 that optimal values of $Q(a,s)$ described the payoff to playing action a today but behaving optimally thereafter. So with Kahneman-Tversky loss-gain functions

$$Q(a,s) = R(a,s) + CCE \left(\beta V(s') \right) \tag{60}$$

In equilibrium $V(s') = \max_a Q(a,s')$, so this is equivalent to

$$Q(a,s) = R(a,s) + CCE \left(\beta \max_a Q(a,s') \right) \tag{61}$$

or

$$E_{s'}\Phi\left(R(a,s)+\beta V(s')-Q(a,s)\right) \tag{62}$$

A version of Q-learning for this context would be

$$\begin{array}{l} Q_{t+1}(a,s)=Q_t(a,s)+ \\ \quad \begin{cases} \alpha_t(a,s)\Phi\left(\beta\max_{a'}Q_t(a',s')+R(a,s)-Q_t(a,s)\right) & s=s_t, a=a_t \\ 0 & \text{otherwise} \end{cases} \end{array} \tag{63}$$

Assumption 8 *$\sum_t \alpha_t(a,s)=\infty$ and $\sum_t \alpha_t^2(a,s)<\infty$ with probability 1 for all s, a.*

As noted in Section 2, Assumption 8 requires that there be sufficient experimentation, in particular each state-action pair must be occur infinitely often. If the chain is ergodic under all policies this will follow under, for example, the ϵ-greedy learning or softmax policies described in Section 2.

Theorem 7 *Under Assumption 2 and Assumption 8, Q_t converges to the unique solution of (61) or equivalently of (62).*

The result can be extended to local convergence in the case of general preferences. The Q values again give the payoff to taking action a in state s and playing optimally in future so satisfy

$$Q(a,s)=W\left(R(a,s),GC\left(\left(\max_{a'}Q(a',s')\right)\right)\right) \tag{64}$$

So that

$$\tilde{W}(R(a,s),Q(a,s))=GC\left(\left(\max_{a'}Q(a',s')\right)\right) \tag{65}$$

or equivalently

$$E_{s'}\Psi\left(\max_{a'}Q(a',s'),\tilde{W}\left(R\left(a,s\right),Q(a,s)\right)\right)=0 \tag{66}$$

The algorithm becomes

$$\begin{array}{l} Q_{t+1}(a,s)=Q_t(a,s) \\ \quad +\begin{cases} \alpha_t(a,s)\Psi\left(\max_{a'}Q_t(a',s'),\tilde{W}\left(R\left(a_t,s_t\right),Q_t(a_t,s_t)\right)\right) & s=s_t, a=a_t \\ 0 & \text{otherwise} \end{cases} \end{array} \tag{67}$$

Asymptotically the evolution of the system is related to that of the differential equation

$$\dot{Q}(a,s) = \delta_{a,s} E_{s'} \Psi \left(\max_{a'} Q(a',s'), \tilde{W}\left(R\left(a,s\right), Q(a,s)\right)\right) \tag{68}$$

One can apply a similar argument to that in Section 5 to prove local convergence. To avoid complication caused by possible non-smoothness of the right-hand side it is assumed that optimal policy specifies a unique action in each state. This is probably not necessary.

Theorem 8 *Under Assumption 3, and Assumption 5, if the optimal policy specifies a unique action in each state then for any $\epsilon > 0$, there is a neighborhood of the corresponding Q-values, $\{Q^*(a,s)\}$, such that if the system starts in that neighborhood then $\{Q_t(a,s)\}$ converges to $\{Q^*(a,s)\}$ with probability at least $1-\epsilon$.*

Whether Q-learning is the best model of learning is debated in the neuroscience literature. If it is then the results here show that the optimal policy can be learned for a wide range of recursive preferences.

8. Conclusion

This paper has studied models of learning where agents compare outcomes to reference points and adjust their reference points in light of outcomes. It has shown that such a process can lead agents to have recursive preferences and so strengthened the case for their use.

Appendix

Proof of Theorem 1

The law of evolution of V_t^a can be written as

$$V_{t+1}^a = V_a^t + \alpha_t \left(F(V^a) + u_t\right) \tag{69}$$

where $F(V^a) = E\Psi\left(R(s), V^a\right)$ and $u_t = \Psi(R(s_{t+1}), V^a) - F(V^a)$.

It follows from Assumption 1 that the differential equation

$$\dot{V}^a = F(V^a) \tag{70}$$

has a unique stationary state, $V^{a\star}$, which is globally stable. To prove global convergence to it of the stochastic algorithm a Lyapounov function will be constructed.

Consider the function

$$W(V^a) = \frac{1}{2}\left(V^a - V^{a\star}\right)^2 \tag{71}$$

Obviously,

$$W \text{ is } C^2 \text{ with bounded derivatives and tends to } \infty \text{ as } |V^a| \to \infty. \tag{72}$$

Moreover

$$F(V^a)W'(V^a) < 0 \qquad V^a \neq V^{a\star} \tag{73}$$

It follows from Assumption 1 that there exists $C > 0$ such that

$$|F(V^a)| \leq C(W(V^a) + 1) \qquad \forall\, V^a \tag{74}$$

Finally, if $\mathcal{F}_t$ denotes the σ-field of generated by events up to and including to time t, then it follows from this definition that

$$E(u_t|\mathcal{F}_{t-1}) = 0 \tag{75}$$

and, using Assumption 1, that there exists $D > 0$ such that

$$E(u_t^2|\mathcal{F}_{t-1}) \leq D(1 + W(V_t^a)) \qquad \forall\, V^a \tag{76}$$

(72) to (76) imply that the conditions of Theorem 9.3.1 on p. 331 of Duflo (1997) are satisfied, so that V^a converges to $V^{a\star}$ almost surely.

Proof of Lemma 4

Consider the following map defined on the set of bounded functions on S, $B(S)$, with the supremum metric,

$$\Theta(V)(s) = R(a(s), s) + GC(\beta V(s')) \tag{77}$$

Since S is finite R is bounded (and $B(S)$ is in fact just the set of real-valued functions on S). Also by the constancy and and monotonicity properties of GC (Lemma 1), if V uniformly bounded then so is $GC(V(s'))$. Hence Θ maps $B(S)$ to itself.

It follows from the monotonicity property of GC (Lemma 1) that

$$V(s) \geq W(s)\ \forall s \Longrightarrow \Theta V(s) \geq \Theta W(s)\ \forall s \tag{78}$$

and from the assumption of translation-subinvariance that

$$\Theta(V + k)(s) \leq \Theta(V)(s) + \beta k \qquad \forall s \tag{79}$$

(78) and (79) imply that T satisfies Blackwell's sufficient condition for a contraction on $B(S)$ (see for example Stokey et al. (1989) p. 54, Theorem 3.3). Hence T has a unique fixed point, which is equivalent to uniqueness of the solution to (32). The equivalence of (32) and (31) follows from the remarks in the text.

Proof of Theorem 2

If the ODE is written as $\dot{V}(s) = F(V)$ then if the certainty-equivalent is translation-subinvariant, the Jacobian of F at V^* is a dominant-diagonal matrix with a strictly-negative diagonal. This follows since the ODE is

$$\dot{V}(s) = \delta_s \sum_{s'} p_{ss'} \Psi\left(\beta V(s'), V(s) - R(a, s)\right) \tag{80}$$

By Assumption 1 $\Psi_1 \geq 0$ and $\Psi_2 < 0$ (subscripts denoting derivatives). Moreover if μ is the certainty equivalent of a distribution $x_1, \ldots, x_s$ with probabilities $q_1, \ldots, q_s$ then if Ψ is sub-invariant

$$\sum_s q_s \left(\Psi_1(x_s, \mu) + \Psi_2(x_s, \mu)\right) \leq 0 \tag{81}$$

Applying this to the right-hand side of (80) implies, noting that $\beta < 1$, that F is diagonally-dominant with a negative diagonal.

Now the eigenvalues of a dominant-diagonal matrix with strictly-negative diagonal have negative real parts. V^* is therefore locally asymptotically stable. The result then follows from Benaïm (1999) Proposition 7.9 on convergence with positive probability of a stochastic approximation algorithm to an attractor (here a locally linearly stable equilibrium) of the corresponding differential equation.

Proof of Theorem 3

Consider the mapping defined by

$$T(V)(s) = V(s) + \gamma E_{s'} \Phi \left(R(a(s), s) + \beta V(s') - V(s) \right) \tag{82}$$

where γ is a parameter to be chosen. The existence and uniqueness of a solution to (36) is equivalent to T having a unique fixed point. The latter follows from the following lemma

Lemma A1 *T is contraction mapping on $B(S)$ for small enough γ.*

Proof

Let W and V be two elements of $B(S)$.

$$\begin{aligned} TV(s) - TW(s) &= V(s) - W(s) \\ &+ \gamma E_{s'} \left(\Phi \left(R(a(s), s) + \beta V(s') - V(s) \right) - \Phi \left(R(a(s), s) + \beta W(s') - W(s) \right) \right) \end{aligned} \tag{83}$$

Since Φ is Lipschitz by the Intermediate Value Theorem for Lipschitz functions (see Clarke (1990) Theorem 2.3.7) for any x, y there is $\xi \in [m, M]$, dependent on x, y (see Assumption 2 for the definitions of m and M) such that $\Phi(x) - \Phi(y) = \xi(x - y)$. Applying this to (83), there are $\xi_{ss'}$ (suppressing other dependence on arguments for convenience) independent of γ such that

$$\begin{aligned} T(V)(s) - T(W)(s) =& V(s) - W(s) + \\ & \gamma \sum_{s'} p_{ss'} \xi_{ss'} (\beta V(s') - \beta W(s') - (V(s) - V(s')) \end{aligned} \tag{84}$$

which is equivalent to

$$\begin{aligned} T(V)(s) - T(W)(s) =& (1 - \gamma \sum_{s'} p_{ss'} \xi_{ss'})(V(s) - W(s)) \\ & + \sum_{s'} \beta\gamma \, \xi_{ss'} (V(s') - W(s')) \end{aligned} \tag{85}$$

Since $\xi_{ss'} \geq m$ for all s', it follows that if $\gamma m < 1$ the coefficient of $V(s) - W(s)$ is positive and hence

$$|T(V)(s) - TW(s)| \leq \left(1 - \gamma \sum_{s'} p_{ss'}\xi_{ss'} + \beta\gamma \sum_{s'} p_{ss'}\xi_{ss'}\right) \max_{\sigma} |V(\sigma) - W(\sigma)| \tag{86}$$

Hence for small enough γ, T is a contraction mapping.

The convergence of the algorithm can be proven by verifying the conditions of Tsitsklis (1994). (29) can be written in the form

$$V_{t+1}(s) = \tilde{\alpha}_s(t) \left(T(V_t)(s_t) - V(s_t) + w_s(t)\right)$$

where $\tilde{\alpha}_s(t) = \alpha_t^s/\gamma$, where γ is chosen small enough to make T a contraction and $w_s(t)$ is the random error term.

We have that with probability 1

$$\sum_t \tilde{\alpha}_s(t) = \infty, \qquad \sum_t \tilde{\alpha}_s^2(t) \leq C \qquad \forall s \tag{87}$$

for some C from Assumption 3 (the divergence of the first sum with probability 1 follows since the divergence of the sum is a tail event and so has probability 0 or 1 as the chain is ergodic — see Breiman (1992) Lemma 7.43. Assumption 3 implies that this probability must be 1 for some state and so by ergodicity 1 for all states).

T is a contraction with respect to the supremum norm. If $\mathcal{F}_t$ is the information observed by time t then by construction $E\left(w_s(t)|\mathcal{F}_t\right) = 0$ and Assumption 2 implies that for some constants A and B $E\left(w_s^2(t)|\mathcal{F}_t\right) \leq A + B \max_t \max_s |V_s(t)|^2$. Convergence follows from Tsitsklis (1994) Theorem 3.

Proof of Theorem 4

The uniqueness of the solution argument to (42) follows from a similar argument to that in Lemma 4. The assumed properties of W and GC imply that the operator T defined by

$$\Theta(V)(s) = W\left(R(a(s), s), GC(V(s'))\right) \tag{88}$$

maps bounded functions to bounded functions and satisfies Blackwell's sufficient conditions for a contraction.

The remainder of the proof is similar to that of Theorem 2. Here the relevant ODE is

$$\dot{V}(s) = \sum_{s'} p_{ss'} \Psi(V(s'), \tilde{W}(R(a(s), s), V(s)) \tag{89}$$

As in Theorem 2, the Jacobian matrix of the right-hand side is a dominant diagonal matrix at V^*. This follows from the properties of Ψ established in the proof of Theorem 2 and from the fact that the derivative of $\tilde{W}$ with respect to $V(s)$, $\tilde{W}_2$, is strictly greater than 1 by Assumption 5. The rest of the argument is as in Theorem 2.

Proof of Corollary 1

The uniqueness of the solution follows from the argument of Lemma 4, replacing β by $\beta d < 1$ in (79). As in the proof of Theorem 2, the Jacobian of the right-hand side of (80) is a dominant-diagonal matrix because d-subinvariance implies that (81) becomes $\sum_s q_s \left(\Psi_1(x_s, \mu) + d\Psi_2(x_s, \mu)\right) \leq 0$. The rest of the argument is as in the proof of Theorem 2.

Proof of Theorem 5.

The proof is omitted as it is almost identical to that of Theorem 3.

Proof of Theorem 6.

The proof is identical to that of Theorem 3 as the required step size property is now assumed.

Proof of Lemma 6

Note that for any k, x and random variable X

$$\begin{aligned} k \geq W(x, GC(X)) &\iff \tilde{W}(x, k) \geq GC(X) \\ &\iff E\Psi(X, \tilde{W}(x, k)) \leq 0 \end{aligned} \tag{90}$$

which implies the result.

Proof of Theorem 7.

The proof is very similar to that of Theorem 3. One shows that the map

$$T(Q)(a, s) = Q(a, s) + \gamma E_{s'} \Phi\left(R(a, s) + \beta \max_{a'} Q(a', s') - Q(a, s)\right) \tag{91}$$

is a contraction on the set of bounded functions on $A \times S$ for suitable γ by a similar argument to Lemma A1 and then proceeds as in the the proof of Theorem 3.

Proof of Theorem 8.

The proof is very similar to that of Theorem 4. Since the optimal policy has a unique optimal action in each state, say $a^*(s)$ in state s, in a small enough neighborhood of $\{Q^*\}$, $Q(a^*(s'), s') > Q(a', s')$ for all $a' \neq a(s')$. It follows that $\max_{a'} Q(a', s')$ in the right-hand side of (68) can be replaced by $Q(a^*(s'), s')$ in a small enough neighborhood of $\{Q^*\}$. The right-hand side is therefore a C^1 function in this neighborhood. Since from (65) $\tilde{W}(R(a, s), Q(a, s))$ is a generalized certainty equivalent at $\{Q^*\}$ for each a and s, a similar argument to that in Theorem 4 shows that the Jacobian matrix of the right-hand side is a dominant-diagonal matrix with a negative diagonal. One concludes as there.

References

Backus, D., Routledge, B., and Zin, S. (2005). Exotic preferences for macroeconomists. In *NBER Macroeconomics Annual 2004*, volume 19, pages 319–390. MIT Press, Cambridge.

Beggs, A. (2005). On the convergence of reinforcement learning. *Journal of Economic Theory*, 122:1–36.

Ben-Tal, A. and Teboulle, M. (2007). An old-new concept of convex risk-measures: the optimized certainty equivalent. *Mathematical Finance*, 17:449–476.

Benaïm, M. (1999). Dynamics of stochastic approximation algorithms. In *Seminaire de Probabilités, XXXIII*, volume 1709 of *Lecture Notes in Mathematics*, pages 1–68. Springer Verlag, Berlin.

Borkar, V. (2008). *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, Cambridge.

Breiman, L. (1992). *Probability*. Society for Industrial and Applied Mathematics, Philadelphia.

Buhlmann, H. (1970). *Mathematical Methods of Risk Theory*. Springer Verlag, Berlin.

Camerer, C., Babcock, L., Loewenstein, G., and Thaler, R. (1997). Labor supply of new city cabdrivers: One day at a time. *Quarterly Journal of Economics*, 112.

Chew, S. (1989). Axiomatic theories of utility with the betweenness property. *Annals of Operations Research*, 19:273–298.

Clarke, F. (1990). *Optimization and Nonsmooth Analysis*. Classics in Applied Mathematics. SIAM, Philadelphia, PA.

Dayan, P. and Abbott, L. (2001). *Theoretical Neuroscience*. MIT Press, Cambridge, MA.

Dekel, E. (1986). An axiomatic characterization of preferences under uncertainty: relaxing the independence axiom. *Jiournal of Economic Theory*, 40:304–318.

Duflo, M. (1997). *Random Iterative Models.* Springer Verlag, Berlin.

Epstein, L. and Zin, S. (1989). Substitution, risk aversion, and the temporal behavior of asset returns: A theoretical framework. *Econometrica*, 57:937–69.

Erev, I. and Roth, A. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, 88:848–881.

Fishburn, P. (1988). *Nonlinear Preference and Utility Theory.* Wheatsheaf, Brighton.

Glimcher, P. (2011). *Foundations of Neuroeconomic Analysis.* Oxford University Press, Oxford.

Gul, F. (1991). A theory of disappointment aversion. *Econometrica*, 59:667–686.

Hansen, L. and Sargent, T. (1995). Discounted linear exponential quadratic gaussian control. *IEEE Transactions on Automatic Control*, 40:968–71.

Hansen, L. and Sargent, T. (2008). *Robustness.* Princeton University Press, Princeton, NJ.

Hopkins, E. and Posch, M. (2005). Attainability of boundary points under reinforcement learning. *Games and Economic Behavior*, 53:110–125.

Huber, P. (1981). *Robust Statistics.* Wiley, Chichester.

Johansen, T. and Donaldson, J. (1985). The structure of intertemporal preferences under uncertainty and time consistent plans. *Econometrica*, 53:1451–1458.

Kahneman, D. and Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47:263–292.

Kimball, M. (1990). Precautionary saving in the small and the large. *Econometrica*, 58:53–73.

Köszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences. *Quarterly Journal of Economics*, 121:1133–1165.

Kreps, D. and Porteus, E. (1978). Temporal resolution of uncertainty and dynamic choice theory. *Econometrica*, 46:185–200.

Kushner, H. and Yin, G. (1997). *Stochastic Approximation Algorithms and Applications*. Springer Verlag, New York.

LaValle, I. (1968). On cash equivalents and information evaluation under uncertainty: Part i: Basic theory. *Journal of the American Statistical Association*, 63:253–276.

Marinacci, M. and Montruchio, L. (2010). Unique solutions for stochastic recursive utility. *Journal of Economic Theory*, 145:1776–1804.

Masatlioglu, Y. and Raymond, C. (2014). A behavioral analysis of stochastic reference dependence. Technical report, University of Michigan.

Mihatsch, O. and Neuneier, R. (2002). Risk-sensitive reinforcement learning. *Machine Learning*, 49:267–90.

Newey, W. and Powell, J. (1987). Asymmetric least squares estimation and testing. *Econometrica*, 55:819–847.

Niv, Y., Edlund, J., Dayan, P., and O'Doherty, J. (2012). Neural prediction errors reveal a risk-sensitive learning process in the human brain. *Journal of Neuroscience*, 32:551–562.

Niv, Y. and Montague, R. (2009). Theoretical and empirical studies of learning. In Glimcher, P., Camerer, C., Fehr, C., and Poldrack, R., editors, *Neuroeconomics: Decision Making and the Brain*, chapter 22, pages 331–351. Elsevier, Amsterdam.

Pope, R. and Chavas, J.-P. (1985). Producer surplus and risk. *Quarterly Journal of Economics*, 100:853–869.

Pratt, J. (1964). Risk aversion in the small and in the large. *Econometrica*, 32:122–136.

Raiffa, H. (1968). *Decision Analysis: introductory lectures on choices under uncertainty*. Addison Wesley, Reading, MA.

Rusczyński, A. (2010). Risk-averse dynamic programming for markov decision processes. *Mathematical Programming Series B*, 125:235–261.

Sarver, T. (2012). Optimal reference points and anticipation. Technical report, Northwestern.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80:1–27.

Shapiro, A., Dentcheva, D., and Rusczyński, A. (2014). *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, second edition.

Shen, Y., Tobia, M., Sommer, T., and Obermayer, K. (2014). Risk-sensitive reinforcement learning. arXiv::1311.2097.

Stokey, N., Lucas, R., and Prescott, E. (1989). *Recursive Methods in Economic Dynamics*. Harvard University Press, Cambridge, MA.

Sutton, R. and Barto, A. (1998). *Reinforcement Learning*. MIT Press, Cambridge, MA.

Tsitsklis, J. (1994). Asynchronous stochastic approximation and q-learning. *Machine Learning*, 16:185–202.

Whittle, P. (1996). *Optimal Control: Basics and Beyond*. Wiley, New York.