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
1. Introduction

The idea that agents evaluate outcomes relative to reference points has been influential in the economics literature since the pioneering work of Kahneman and Tversky (1979). It also has support in the literature on 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 on neuroscience much interest has focused on modeling learning by reinforcement learning. Simple procedures can lead an 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 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.

In this paper, agents 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 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 to learn optimal actions in a static environment. The models here attempt to learn optimal behavior in a dynamic environment and draw inspiration from the models of reinforcement learning in the tradition Sutton and Barto (1998), 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 than conventional models. They use the model of Mihatsch and Neuneier (2002) from the the machine 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 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 learning 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 background on learning from Neuroscience. Section 3 outlines the basic framework and Section 4 presents a special case of static learning to improve intuition. These initial sections also
discuss related literature. Section 5 presents the main results. Local convergence to recursive preference is shown in the case of general gain-loss functions. Global convergence is shown when losses and gains only depend on the difference between outcomes and reference points. Section 6 considers extensions, in particular to the case when preferences need not be intertemporally separable. Section 7 discusses the implications for choice of actions and also for learning the optimal policies. Results for local and global convergence are 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

\[
W_{t+1} = W_t + \alpha_t (R_t - W_t)
\]  

(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. \( \alpha_t \) 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)).\(^1\) One 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:

\[
E\left(\sum_{t=0}^{\infty} \beta^t R_t\right)
\]

\(^1\)Much of its interest derives from its explanation of phenomena involving multiple stimuli, which it assumes affect rewards additively.
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

$$V(s) = R(s) + \beta V(s')$$

(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 $s_{t+1}$.

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

(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.

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

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

(4)

$V$ is left unchanged for states other than $s_t$. $\alpha_t$ 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 $\alpha_t$ 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

$$Q(a, s) = R(a, s) + \beta EV(s')$$

(5)
If the policy is optimal then \( V(s) = \max_a Q(a, s) \) for all \( s \) so this is equivalent to

\[
Q(a, s) = R(a, s) + \beta E \max_{a'} Q(a', s')
\]  (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

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

Values of \( Q \) for other state-action pairs are left unchanged. \( \alpha_t(s, a) \) 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 \( \epsilon \)-greedy rule: with probability \( 1 - \epsilon \) the action with the highest value of \( Q_t(a_t, s_t) \), with probability \( \epsilon \) all actions 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 \( \delta_t \) tends to zero at an appropriate rate.

Other rules of 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:

\[
V_{t+1}(s_t) = V(s_t) + \alpha_t \phi(\delta_t)
\]  (8)

where

\[
\phi(x) = \begin{cases} 
ax & x < 0 \\
bx & x > 0
\end{cases}
\]  (9)
with $a > b > 0$. 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. This also 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 $p_{ss'}^a$.
- The agent has discount factor $\beta$.

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

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

or equivalently

$$V^\pi(s) = R(\pi(s), s) + \beta E(V^\pi(s')|s)$$

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

$$V(s) = \max_a R(a, s) + \beta E(V(s')|s, a)$$
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 $\pi$, 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.

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

$$V^\pi_{t+1}(s) = \begin{cases} V^\pi_t(s_{t+1}) + \alpha_t (R(a_t, s_t) + \beta V^\pi_t(s_{t+1}) - V^\pi_t(s_t)) & s = s_t \\ V^\pi_t(s) & \text{otherwise} \end{cases}$$

(13)

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

If $V^\pi_t = V^\pi$, then the expected change in $V^\pi_t$ is zero from (12). If $\alpha_t$ tends to zero at an appropriate rate it can be shown that $V^\pi_t$ converges to $V^\pi$, 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 6.

The term $R(a_t, s_t) + \beta V^\pi_t(s_{t+1}) - V^\pi_t(s_t)$ 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 $s_t$, $V^\pi_t(s_t)$ 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

$$V^\pi_{t+1}(s) = \begin{cases} V^\pi_t(s_{t+1}) + \alpha_t \phi(R(a_t, s_t) + \beta V^\pi_t(s_{t+1}) - V^\pi_t(s_t)) & s = s_t \\ V^\pi_t(s) & \text{otherwise} \end{cases}$$

(14)

where $\phi$ is a function measuring losses and gains or more generally
\[ V_{t+1}^\pi(s) = \begin{cases} 
 V_t^\pi(s_{t+1}) + \alpha_t \psi(\beta V_t^\pi(s_{t+1}), V_t^\pi(s_t) - R(a_t, s_t)) & \text{if } s = s_t \\
 V_t^\pi(s) & \text{otherwise} 
\end{cases} \]  

(15)

where \( \psi \) is again a gain-loss function.

In this formulation the gain (or loss) experienced if the outcome is \( x \) and the anticipated payoff is \( \mu \) is \( \psi(x, \mu) \).

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 \( \mu \),  
(iii) \( \psi \) is Lipschitz in \( x \), and  
(iv) there exist \( k \) and \( K \), \( K \geq k > 0 \), such that for all \( \mu, \mu' \neq \mu \), and for all \( x \)

\[ k \leq \left| \frac{\psi(x, \mu) - \psi(x, \mu')}{\mu - \mu'} \right| \leq K \]  

(16)

In the case when \( \psi(x, \mu) = \phi(x - \mu) \), this is implied by

**Assumption 2**  
\( \phi \) 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 \( \phi \) 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 results can be extended to them. They do not study general gain-loss functions, as is done here. They offer an interpretation in terms of risk-sensitive programming with coherent 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. Static Learning

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 \( p_{as}^{a'} \) is independent of \( s \) for all \( a \) and \( s' \). With some abuse of notation the probability of observing state \( s' \) will be written as \( p_{s'}^s \).2

Since the states in all periods are independent, the decision maker’s problem reduces to a myopic one of attempting to maximize her expected return each period. Assume that if she takes action \( a \) now and the realized state next period is \( s' \) she obtains reward \( R(a, s') \). To save notation the state will be denoted by \( s \) rather than \( s' \) but note that in contrast to other sections \( s \) refers to the state in the next rather than the current period. This contrast will be discussed further in Section 6.

Since states are independent across periods, agents’ reference points will be independent of the current state and the learning rule (15) can be simplified to

\[
V_{t+1}^a = V_t^a + \alpha_t \psi (R(a, s), V_t^a))
\]

Note that as the problem is now effectively a single period one the discount factor, \( \beta \), is irrelevant.

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, \( \alpha_t \), 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.

2The assumption that the number of states is finite is not essential for the results of this section.
Standard results in stochastic approximation show that the long-run behavior of the learning process can be examined by looking at the ordinary differential equation:

\[ \dot{V}^a = E_s \psi(R(a, s), V^a) \]  

(18)

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

\[ E_s \psi(R(a, s), V^a) = 0 \]  

(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 \( U - 1(E_s U(R(a, s))) \). More generally, however, \( V^a \) would be belong to the class of generalized certainty equivalents defined by Chew (1989).

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 \( R(a, s) \).

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

Let \( GC \) denoted the generalized certainty equivalent defined by (19). Some immediate properties following from Assumption 1 are

**Lemma 1** (i) (constancy): if \( R(a, s) = k \) for all \( s \), then \( GC(R(a, s)) = k \).

(ii) (monotonicity) if \( R'(s, a) \geq R(a, s) \) for all \( s \) then \( GC((R'(a, s)) \geq GC((R(a, s))) \).

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

$$E_s \phi(R(a, s) - V^a) = 0 \quad (20)$$

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

$$C^a = U^{-1}(E_s U(R, s)) \quad (21)$$

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 (19) is the compensating variation.\(^3\) 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.\(^4\)

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

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

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

---

\(^3\)See for example Pope and Chavas (1985) or Kimball (1990) for further discussion.

\(^4\)As noted for example by LaValle (1968) and Raiffa (1968).
Example 2  Piecewise-Linear Kahneman-Tversky Form

Suppose $\phi$ has the piecewise-linear form

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

(22)

$V^a$ is then the solution to

$$
E_s I(R(a, s) < V^a)(R - V^a) + \lambda E_s I(R(a, s) > V^a)(R - V^a) = 0
$$

(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:

$$
\min_{V^a} E_s I(R(a, s) < V^a)(R - V^a)^2 + \lambda E_s I(R(a, s) > V^a)(R - V^a)^2
$$

(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

$$
L(x, d) = \begin{cases} 
U(x) - U(d) & x \leq d \\
a(U(x) - U(d)) & x > d 
\end{cases} 
$$

(25)

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

$$
E_s I(R(a, s) < V^a)(U(R) - U(V^a)) + \lambda E_s I(R(a, s) > V^a)(U(R) - U(V^a)) = 0
$$

(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, $\mu$, 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 notions are distinct. There is no concept of anticipation in the current framework. The decision maker is assumed simply to adjust his reference point to minimize expected losses.

Köszegi and Rabin (2006) suggest that reference points are determined by expectations. They, however, consider 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. Dynamic Learning

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 additive form then convergence is with probability one for any starting values.

As laid out in section 3, the agent adopts the following learning rule:
\[ V_{t+1}^\pi(s) = \begin{cases} V_t^\pi(s_{t+1}) + \alpha_t \psi(\beta V_t^\pi(s_{t+1}), V_t^\pi(s_t) - R(a_t, s_t)) & s = s_t \\ V_t^\pi(s) & \text{otherwise} \end{cases} \]  

or equivalently

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

where

\[ \alpha_t = \begin{cases} \alpha_t & s = s_t \\ 0 & \text{otherwise} \end{cases} \]

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. In the special case when \( \psi(x, \mu) = \phi(x - \mu) \), (28) can be rewritten as

\[ V_{t+1}^\pi(s) = V_t^\pi(s) + \alpha_t \phi(R(a_t, s_t) + \beta V_t^\pi(s_{t+1}) - V_t^\pi(s_t)) \]  \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, \( R(a_t, s_t) + \beta V(s_{t+1}) \).

Note that the reward, \( R \), depends on the current state not the state next period in contrast to the previous section. The action, \( a_t \), does, however, affect the evolution of future states. The case when \( R \) may depend on the state next period is discussed in the next section.

Attention will focus on the case of a fixed policy so the superscript, \( \pi \), 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\(^5\) show that the long run behavior of the system is governed by the differential equation

\[ \dot{V}(s) = H_s^1(V) = \delta_s E_s \psi(\beta V(s'), V(s) - R(a(s), s)) \]  \tag{30}

where \( \delta_s \) 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

\[ E_s \psi (\beta V(s'), V(s) - R(a, s)) = 0 \] (31)

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

\[ V(s) = R(a(s), s) + GC(\beta V(s')) \] (32)

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

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

This is similar to the kind of equation familiar in 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

\[ \tilde{V}(s) = R(a(s), s) + \beta C(\tilde{V}(s')) \] (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

\[ C(V(s')) = -\frac{1}{\gamma} \ln E (-\gamma \exp V(s')) \] (35)

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

As is well-known a sufficient condition (see for example Marinacci and Montrucchio (2010)), (33) to have a unique solution is that it 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, (33) has a unique solution if $GC$ is translation sub-invariant.

One then has

**Theorem 2** Under Assumption 1 and Assumption 3 if $\psi$ 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. In any case it is a fairly standard assumption in the literature on recursive preferences (see for example Marinacci and Montruchio (2010)). 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

$$E_{\phi'} (R(a(s), s) + \beta V(s') - V(s)) = 0$$

and (32) becomes

$$V(s) = R(a(s), s) + CCE(\beta V(s'))$$

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

$$\dot{V}(s) = \delta_s E_s' \phi (R(a(s), s) + \beta V(s') - V(s))$$  \hspace{1cm} (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 $\delta_s (T(V)(s) - V(s))$, where $T$ is a contraction. The formal proof is in the Appendix and an easy extension of the proof 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 if the probabilities are known. 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 third other qualifications.
6.1 Non-additive Intertemporal Preferences

The full family of Epstein-Zin preferences allows for non-additive preferences over time. 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

\[ V(s) = W(R(a(s), s), GC(V(s'))) \] (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

\[ W(x, y) = (x^\rho + by^\rho)^{1/\rho} \] (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 \( \beta < 1 \). 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

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

or equivalently

\[ E_{s'}\psi(V(s'), \tilde{W}(R(a(s), s), V(s))) = 0 \] (42)

and one can apply the learning scheme

\[ V^\pi_{t+1}(s) = \begin{cases} V^\pi_t(s_{t+1}) + \alpha_t \psi \left( \beta V^\pi_t(s_{t+1}), \tilde{W}(R(a_t, s_t), V^\pi_t(s_t)) \right) & s = s_t \\ V^\pi_t(s) & \text{otherwise} \end{cases} \] (43)
As in the previous case, the asymptotic behavior of the system is related to that a differential equation, in this case

$$\dot{V}(s) = \delta_s E_s \psi(V(s'), \tilde{W}(R(a(s), s), V(s)))$$ (44)

where again $\delta_s$ 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. $\psi$ and $\tilde{W}$ are $C^1$, and $\tilde{W}_2 \geq \beta > 1$ for all $x, y$, some $\beta$.

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

$$E \phi(R(a, s, s') + \beta V(s') - V(s)) = 0$$

(45)

(45) is equivalent to

$$V(s) = CCE(R(a, s, s') + \beta V(s'))$$

(46)

If the latter could be written as

$$V(s) = ER(a, s, s') + CCE(\beta V(s'))$$

(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 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 revised 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 resulting values for $V$.

Consider general time-separable case for simplicity. In this case the paper has implicitly taken the second interpretation: the gain is $\psi(\beta V(s'), V(s) -$
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

\[
V(s) = GC \left( R(a,s) + \beta V(s') \right)
\] (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

\[
CCE \left( R(a,s) + \beta V(s') \right) = R(a,s) + CCE(\beta V(s'))
\] (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 expectation to take into account information on payoffs accrued, so the second is utilized.

### 6.4 Robustness to Other Assumptions

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

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

The requirement that the slopes of \( \phi \) be bounded above and below is somewhat restrictive. It implies that \( \phi \) cannot be too non-linear and plays an important role in the uniqueness and contraction arguments. It is somewhat restrictive.
in that it rules out power loss-gain functions which have infinite slopes at the origin, which are sometimes popular in prospect theory. On the other hand these can readily approximated by loss-gain functions with large but bounded slopes at the origin, so this does not seem too restrictive.

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 and one can replace the loss-gain functions outside this domain by a linear extension. Alternatively one could modify the algorithm by truncating it: pick a very large number \( B > 0 \) and modify (14) to

\[
V^\pi_{t+1}(s) = \begin{cases} 
\Pi_{[-B,B]} \left( V^\pi_t(s_{t+1}) + \alpha_t \phi \left( R(a_t, s_t) + \beta V^\pi_t(s_{t+1}) - V^\pi_t(s_t) \right) \right) & s = s_t \\
V^\pi_t(s) & \text{otherwise}
\end{cases}
\]

(51)

where \( \Pi_{[-B,B]}(x) \) maps \( x \) to the nearest point lying in \([-B,B]\). This ensures the process remains bounded, which is part of the reason for Assumption 2, and leaves the algorithm unchanged except for very large states. Assumption 2 need only hold for values of \( x \) and \( y \) in this bounded region, which allows for exponential loss-gain functions. If \( B \) is large enough the solution is unaffected and for large \( t \) the projection is irrelevant.\(^6\) This may seem artificial but it is arguable that neural states can only hold bounded values.

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.

\(^6\)If one allows for random payoffs then the algorithm can leave the region but for large \( t \) the probability will be very small; see for example Kushner and Yin (1997) Chapter 12.8 for a discussion in a related example.
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

\[ V(s) = \max_a R(a, s) + CCE(\beta V(s')) \]  

(52)

This is equivalent to \( V \) solving the set of equations

\[ \max_a E_{s'} \phi (R(a, s) + \beta V(s') - V(s)) \]  

(53)

since for any number \( k \) and random variable \( X \)

\[ k \geq CCE(X) \iff E\phi(X - k) \leq 0 \]  

(54)

and (52) is equivalent to

\[ V(s) = \max_a CCE(R(a, s) + \beta V(s')) \]  

(55)

Lemma 5  (52) and (53) have the same, unique solutions.

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

\[ a(s) = \max_a E_{s'} \phi (R(a, s) + \beta V(s') - V(s)) \]  

(56)

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 (52). To reconcile with (56) 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 (56) would become

\[ a(s) = \max_a E_{s'} \phi (R(a, s) + \beta V(s') - \bar{V}) \]  

(57)
or even

$$a(s) \max_a E_{s'} \phi \left( R(a, s) + \beta \tilde{V} - \tilde{V} \right)$$

(58)

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 $\phi$ 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 $\phi$ affects marginal effect of effort, given the state, on the certainty equivalent in future — see (52).

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

$$V(s) = \max_a W \left( R(a, s), GC(V(s')) \right)$$

(59)

This can be written equivalently as

$$a(s) \max_a E_{s'} L(V(s'), \tilde{W}(R(a, s), V(s)))$$

(60)

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

**Lemma 6** (59) and (60) have the same solutions.

The form here is more convoluted but (60) 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

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

(61)

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

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

(62)

or

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

(63)
A version of \(Q\)-learning for this context would be

\[
Q_{t+1}(a_t, s_t) = Q_t(a, s) + \begin{cases} 
\alpha_t(a, s) \phi (\beta \max_{a'} Q_t(a', s') + R(a, s) - Q_t(a, s)) & s = s_t, a = a_t \\
0 & \text{otherwise}
\end{cases}
\]  
(64)

**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, \(\epsilon\)-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 (62) or equivalently of (63).

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

\[
Q(a, s) = W \left( R(a, s), GC \left( \max_{a'} Q(a', s') \right) \right)
\]  
(65)

So that

\[
\bar{W}(R(a, s), Q(a, s)) = GC \left( \max_{a'} Q(a', s') \right)
\]  
(66)

or equivalently

\[
E_{s'} \psi \left( \max_{a'} Q(a', s'), \bar{W}(R(a, s), Q(a, s)) \right) = 0
\]  
(67)

The algorithm becomes

\[
Q_{t+1}(a_t, s_t) = Q_t(a, s) + \begin{cases} 
\alpha_t(a, s) \phi (\beta \max_{a'} Q_t(a', s') + R(a, s) - Q_t(a, s)) & s = s_t, a = a_t \\
0 & \text{otherwise}
\end{cases}
\]  
(68)

Asymptotically the evolution of the system is related to that of

\[
\dot{Q}(a, s) = \delta_{a, s} E_{s'} \psi \left( \max_{a'} Q(a', s'), \bar{W}(R(a, s), Q(a, s)) \right)
\]  
(69)
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.
Appendix

Proof of Theorem 1

The law of evolution of $V^a_t$ can be written as

$$V^a_{t+1} = V^t_a + \alpha_t (F(V^a) + u_t) \quad (70)$$

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

It follows from Assumption 1 that the differential equation

$$\dot{V}^a = F(V^a) \quad (71)$$

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

Consider the function

$$W(V^a) = \frac{1}{2} (V^a - V^{a*})^2 \quad (72)$$

Obviously,

$W$ is $C^2$ with bounded derivatives and tends to $\infty$ as $|V^a| \to \infty$. \quad (73)

Moreover

$$F(V^a)W'(V^a) < 0 \quad V^a \neq V^{a*} \quad (74)$$

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

$$|F(V^a)| \leq C(W(V^a) + 1) \quad \forall V^a \quad (75)$$

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

$$E(u_t|\mathcal{F}_{t-1}) = 0 \quad (76)$$

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

$$E(u_t^2|\mathcal{F}_{t-1}) \leq D(1 + W(V^a_t)) \quad \forall V^a \quad (77)$$

(73) to (77) imply that the conditions of Theorem 9.3.1 on p. 331 of Duflo (1997) are satisfied, so that $V^a$ converges to $V^{a*}$ 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,

$$
\Theta(V)(s) = R(a(s), s) + \beta GC(V(s'))
$$

(78)

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 $GCV(s')$. Hence $\Theta$ maps $B(S)$ to itself.

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

$$
V(s) \geq W(s) \forall s \implies \Theta V(s) \geq \Theta W(s) \forall s
$$

(79)

and from the assumption of translation-subinvariance that

$$
\Theta(V + k)(s) \leq \Theta(V)(s) + \beta k \quad \forall s
$$

(80)

(79) and (80) 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 (33).

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

$$
\dot{V}(s) = \delta_s \sum_{s'} p_{ss'} \psi (\beta V(s'), V(s) - R(a, s))
$$

(81)

By Assumption 1 $\psi_1 \geq 0$ and $\psi_2 < 0$. Moreover if $\mu$ is the certainty equivalent of a distribution $x_1, \ldots, x_s$ with probabilities $q_1, \ldots, q_s$ then if $\psi$ is sub-invariant

$$
\sum_s q_s (\psi_1(x_s, \mu) + \psi_2(x_s, \mu)) \leq 0
$$

Applying this to the right-hand side of (81) implies, noting that $\beta < 1$, implies that $F$ is diagonally-dominant with a negative diagonal.

Now the eigenvalues of dominant-diagonal matrix with strictly-negative diagonal have negative real parts. $V^*$ is therefore locally asymptotically stable. The result then follows from Benaim (1999) Proposition 7.9.
**Proof of Theorem 3**

Consider the mapping defined by

\[
T(V)(s) = V(s) + \gamma E_{ss'} \phi(R(a(s), s) + \beta V(s') - V(s))
\]  \hspace{1cm} (82)

where \( \gamma \) is a parameter. The existence and uniqueness of a solution to (36) is equivalent to \( T \) having a unique fixed point. This follows from the following lemma

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

**Proof** Let \( W \) and \( V \) be two elements of \( B(S) \).

\[
TV(s) - TW(s) = V(s) - W(s) + \gamma E_{ss'}(\phi(R(a(s), s) + \beta V(s') - V(s)) - \phi(R(a(s), s) + \beta W(s') - W(s)))
\]  \hspace{1cm} (83)

Since \( \phi \) is Lipschitz there is by the Intermediate value theorem for Lipschitz functions 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 \( \gamma \) such that

\[
TV(s) - TW(s) = V(s) - W(s) + \gamma \sum_{s'} p_{ss'} \xi_{ss'}(\beta V(s') - \beta W(s') - (V(s) - V(s'))
\]  \hspace{1cm} (84)

which is equivalent to

\[
TV(s) - TW(s) = (1 - \gamma \sum_{s'} p_{ss'} \xi_{ss'}) (V(s) - W(s)) + \sum_{s'} \beta \gamma \xi_{ss'} (V(s') - W(s'))
\]  \hspace{1cm} (85)

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

\[
|TV(s) - TW(s)| \leq \left(1 - \gamma \sum_{s'} p_{ss'} \xi_{ss'} + \beta \gamma \sum_{s'} p_{ss'} \xi_{ss'}\right) \max_\sigma |V(\sigma) - W(\sigma)|
\]  \hspace{1cm} (86)

Hence for small enough \( \gamma \), \( T \) is a contraction mapping.
The convergence of the algorithm can be proven by verifying the conditions of Tsitskli (1994). (29) can be written in the form

\[ V_{t+1}(s) = \alpha_s(t) (T(V_t)(s_t) - V(s_t) + w_s(t)) \]

where \( \alpha_s(t) = \alpha_s(t)/\gamma \), where \( \gamma \) is chosen small enough to make \( T \) a contraction and \( w_s(t) \) is the random error term.

We have that with probability 1

\[ \sum_t \alpha_s(t) = \infty, \quad \sum_t \alpha_s^2(t) \leq C \quad \forall s \]  

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 \( E\left(w_s^2(t)|\mathcal{F}_t\right) = 0 \) and Assumption 2 implies that for some constants \( A \) and \( B \) \( E\left(w_s^2(t)|\mathcal{F}_t\right) \leq A + B \max_s |V_s(t)|^2 \). Convergence follows from Tsitskli (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

\[ \Theta(V)(s) = W(R(a(s), s), GC(V(s'))) \]  

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

\[ \dot{V}(s) = \sum_{s'} p_{ss'} \psi(V(s'), \tilde{W}(R(a(s), s), V(s))) \]

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 \( \psi \) 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 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 stepsize property is now assumed.

**Proof of Lemma 6**

Note that for any \( k \) and random variable \( X \)

\[
\begin{align*}
    k \geq W(x, GC(X)) & \iff \tilde{W}(x, k) \geq GC(X) \\
    & \iff E\psi(X, \tilde{W}(x, k)) \leq 0
\end{align*}
\]  

(90)

which implies the result.

**Proof of Theorem 7.**

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

\[
T(Q)(a, s) = Q(a, s) + \gamma E_{a'} \phi \left( R(a, s) + \beta \max_{a'} Q(a', s') - Q(a, s) \right)
\]  

(91)

is a contraction on the set of bounded functions on \( A \times S \) for suitable \( \gamma \) 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 (69) can be replaced by \( Q(a^*(s'), s') \) in a small enough neighborhood of \( \{Q^*\} \). The right-hand side is a \( C^1 \) function in this neighborhood. Since from (66) \( \tilde{W}(R(a, s), Q(a, s) \) is a generalized certainty equivalent at \( \{Q^*\} \) for each \( a \) and \( s \), a similar argument to that in Theorem 4 shows that its Jacobian matrix is a dominant-diagonal matrix with a negative diagonal. One concludes as there.
References


