Dynamic Decisions with Short-term Memories

LI, HAO[1]
Dept. of Economics, University of Toronto

SUMON MAJUMDAR
Dept. of Economics, Queen's University

This version: August 2005.

[1]We thank Tilman Börgers and James Dow for comments.

Abstract

ABSTRACT: To model decisions and learning under short-term memories, a two armed bandit problem is studied where the decision maker can only recall the most recent outcome of his past decisions. Unlike the full memory case, optimal learning strategies are shown to involve random and periodic experimentation (in choosing the risky arm). We find that any optimal strategy is necessarily time inconsistent, unless it calls for experimentation with probability one or zero regardless of history. We show through an example that the decision maker with short-term memories can benefit from memory manipulation, sometimes by choosing to forget his past experience.

1 Introduction

People and society frequently learn from past experiences, either their own or of past generations. Often however the complete record of past outcomes maybe lost and only the recorded experiences of recent generations maybe available to guide the actions of the present generation. In choosing this action, it may differ from the advice of previous generations as each generation maybe skeptical of the experiences and stories of previous generations as told to them by their parents. Further, in passing down their experience to future generations, one may choose what memory to pass on, if at all. In this paper we study issues of optimal decision-making, time inconsistency and memory management in generational learning under the constraint of a short-term memory.

The formal model we use is that of a single decision maker with short term memories facing a two armed bandit. The two armed bandit problem, with one safe arm yielding a known, constant per-period payoff and one risky arm with a stochastic per-period payoff and an unknown mean, is a canonical model of dynamic learning and has been well-studied for a decision-maker with full memory of all past experiences. In each period the decision maker chooses, based on his entire history of past outcomes, whether or not to experiment, i.e. play the risky arm. Instead we study here the problem when the decision maker can only recall the outcome of his previous period's decision and must therefore make his decision based on this one-period experience only. We use this particular form of imperfect recall to capture an essential feature of generational learning, namely that memory, both personal and societal, is often short-lived. Relatedly, without imperfect recall, the issues of time inconsistency in decision-making and deliberate memory manipulation would not arise.

We show that optimal learning strategies generally involve random and periodic experimentation. Unlike in the case of full memory, here the optimal probability of experimentation after getting an unfavorable payoff from

the risky arm can be strictly between zero and one. Without such randomization, the decision maker with short term memories would be forced to make a tough choice between stopping experimentation right after the first unfavorable payoff form or continuing experimentation even after repeated negative information about the risky arm. Instead, the optimal strategy carefully calibrates the probability of experimentation to balance the need to engage in some experimentation and the need to respond to negative information. In periodic experimentation, the decision maker adopts a positive probability of resuming experimentation after having drawn the safe arm in the previous period. Optimal strategies require the right combination of periodic experimentation with random experimentation as a response to the constraint of short term memories.

In deriving optimal learning strategies, we assume that at the start of time the decision maker can commit not to modify his plan along the entire learning process. Such commitment is extreme in the context of generational learning. We show that optimal strategies are generally time inconsistent if the decision maker is introspective in spite of the constraint of a short term memory. That is, if he updates his belief based on the experience that he recalls *and* the knowledge that he has acted according to the optimal strategy in the past, then there exists points at which the decision maker would want to deviate from this strategy. Only when the optimal strategy calls for experimentation with probability one or zero regardless of information, would it be time consistent. This happens only when the prior about the risky arm is either extremely optimistic or extremely pessimistic, so that the decision maker optimally disregards any information. Thus, the urge to respond to new information and the incentive to deviate from ex ante optimal learning go hand-in-hand in generational learning.

Random and periodic experimentation as part of an optimal learning strategy reflects the need for the decision maker to retain the flexibility in how to make use of information. This raises the possibility that the

decision maker may benefit from managing or manipulating his "memory" in the sense of not necessarily recording his experience truthfully (or even not at all). Of course we assume that the decision maker knows his own manipulation strategy, so that a simple relabeling of memory states has no effect on his ex ante welfare. We demonstrate through an example how memory manipulation can work without assuming that the decision maker engages in any form of direct self deception. In this example, instead of recording the outcome resulting from his most recent experimentation, the decision maker chooses to sometimes retain the "clean slate" of null history . This form of memory manipulation via "endogenous forgetfulness" allows the decision maker to enrich the state space of his strategy, and helps improve his ex ante welfare by responding better to new information.

The two armed bandit problem without the short term memory constraint is a simple example of a class of problems studied by Gittins (1989). The short term memory constraint considered here is a type of complexity constraints that focus on limited memory (see Lipman (1995) for a survey of the literature). A standard way of modeling limited memory is a finite automaton, which consists of a finite set of memory states, an action rule that maps the set of states to a finite set of choices, and a transition rule that maps the set of states and (finite set of) outcomes to the set of states (see Rubinstein (1986) for an application of finite automata to repeated games and for references to the literature on finite automata). A feasible strategy for our decision maker with short term memories can be thought of as a finite automaton with the set of per-period payoffs as the set of states and the fixed transition rule that gives the state of the next period as the payoff resulting from the current choice. To our knowledge, there is no work on finite automata playing bandits; the closest is a recent paper by Börgers and Morales (2004), who study a bandit model but with perfectly revealing outcomes (about the two arms) and limited scope for learning. The present paper is motivated by issues of time consistency and memory manipulation

in generational learning, and we find the assumption of short term memories more natural than generic finite automata. In particular, the issue of memory manipulation cannot be addressed with a finite automaton approach, as the meaning of each memory state is optimally chosen and is therefore endogenous with such an approach.[1] Our assumption of short term memories is also a form of imperfect recall. The need for randomization and the problem of time inconsistency under imperfect recall have been pointed out by Piccione and Rubinstein (1997) using a particular example.[2] We also add to this literature by characterizing the solution to a well-studied dynamic learning model under an intuitive constraint on memory capacity and highlighting similar randomization and time consistency properties.

The rest of the paper is organized as follows. In the next section, we describe the basic problem of a two armed bandit with short term memories. Section 3 characterizes optimal learning strategies and shows that this can involve random and periodic experimentation. In section 4 we show that any optimal strategy is necessarily time-inconsistent, unless it calls for experimentation with probability one or zero regardless of history. In section 5 we consider the case of memory manipulation and show through an example that memory management can improve ex ante welfare. Section 6 lists some topics for further research. Detailed proofs of the propositions can be found in the appendix.

[1] The same is true for decision models with limited memory and one time decisions, such as Wilson (2004). Her characterization of optimal recording of outcomes in coarse learning shares similarities with what we call memory manipulation here. The issue of time inconsistency and experimentation however does not arise there as it involves once and for all decisions. The same is true in Meyer (1991) where the memory states have fixed, exogenous meanings.

[2] Studies of randomization under imperfect recall go back to Kuhn (1953). Kalai and Salon (2003) define "non interactive" Markov decision problems, and show that under imperfect recall, optimal strategies generally require randomization, but not in the action rule. Our two armed bandit problem is interactive because the decision maker controls the set of possible outcomes through his choice in each period.

2 A Two Armed Bandit Problem with Short Term Memories

To model a simple situation of experimentation and learning, we consider an infinite horizon two armed bandit problem, with discrete time $t = 1, \ldots, \infty$. One of the arms is safe and gives a certain per-period payoff of 0. The risky arm has either high average payoffs (state h) or low payoffs (state l), with the decision maker's prior probability in period 0 equal to η for the state being h. We assume that the normalized per-period payoff from the risky arm is either $+1$ or -1, with $\Pr[+|h] = \Pr[-|l] = q$ and $\frac{1}{2} < q < 1$. Thus, the risky arm has a symmetric binary signal structure.

In each period a decision maker must choose between the risky arm (experimentation, e) and the safe arm (stop, s). The decision maker maximizes the period 0 discounted sum of his expected utility, with a per-period discount factor $\delta \in (0, 1)$.

Without any memory constraint, in each period the decision maker can recall all of his past experience and can base his action on the entire history of past occurrences. His optimal learning strategy is then given by the solution to a Bellman equation. Let p denote the current belief for state h, and $U(p)$ denote the optimal value of the decision maker's objective function. The Bellman equation for this problem is

$$U(p) = max\{\delta U(p), (2p-1)(2q-1) + \delta(pq + (1-p)(1-q))U(p(+)) + \delta((1-p)q + p(1-q))U(p(-))\},$$

where

$$p(+) = \frac{pq}{pq + (1-p)(1-q)}, \qquad p(-) = \frac{p(1-q)}{p(1-q) + (1-p)q}$$

are Bayesian updates of the belief after getting payoffs of $+1$ and -1 respectively from the risky arm. It is straightforward to establish the following: (i) there is a unique function $U(p)$ that satisfies the Bellman equation; (ii) $U(p)$

is increasing and convex; and (iii) there exists $\hat{p} < \frac{1}{2}$ such that $U(p) = 0$ (and the optimal choice is s) if $p \leq \hat{p}$ and $U(p) > 0$ (and the optimal choice is e) if $p > \hat{p}$.

In the present paper we are interested in exploring the consequences of short-term memory on learning and experimentation strategies in this framework. We thus make the extreme but simple assumption that in any period the decision maker can only remember his experience from the previous period. To model this memory constraint, we assume that there are four memory states: null memory ($\emptyset$), a positive payoff of 1 from the risky arm ($+$), a negative payoff of -1 from the risky arm ($-$), and a payoff of 0 from the safe arm (c). Denote a memory state as $m \in \{\emptyset, +, -, c\}$. In line with our focus on short-term memory, we make the assumption that except for the first period, the decision-maker is unable to distinguish calendar-time; thus, his chosen strategy is required to be time-independent. A pure strategy sends each memory state m to a choice of experiment (e) or stop (s). A behavioral strategy β maps each m to a probability β_m of playing e.[3] The decision maker chooses $\beta = (\beta_\emptyset, \beta_+, \beta_-, \beta_c)$ to maximize his period 0 discounted sum of expected utilities.

3 Optimal Learning Strategies

Fix a strategy β. Suppose that the state is h. From the perspective of period 0, the probability X_t^h of choosing the risky arm in period $t = 1, 2, \ldots$ satisfies

$$X_{t+1}^h = (1 - X_t^h)\beta_c + X_t^h(q\beta_+ + (1-q)\beta_-).$$

Denoting $B^h = q\beta_+ + (1-q)\beta_- - \beta_c$, we thus have

$$X_{t+1}^h = B^h X_t^h + \beta_c.$$

[3] Since this is a decision problem with imperfect recall, Kuhn's (1953) theorem of equivalence of behavioral and mixed strategies does not hold. A mixed strategy in our model is a period 0 randomization of pure strategies. It is easy to see that mixed strategies will not improve over pure strategies given the von-Neumann expected utility formulation here.

Using the above formula recursively and $X_1^h = \beta_\emptyset$, we obtain

$$X_t^h = \beta_\emptyset (B^h)^{t-1} + \frac{\beta_c(1-(B^h)^{t-1})}{1-B^h}.$$

Symmetrically, in state l, from the perspective of period 0 the probability X_t^l of experimenting in period t is given by

$$X_t^l = \beta_\emptyset (B^l)^{t-1} + \frac{\beta_c(1-(B^l)^{t-1})}{1-B^l},$$

where $B^l = (1-q)\beta_+ + q\beta_- - \beta_c$.

From a period 0 perspective, the expected payoff to experimentation in any period t is $2q-1$ in state h, and $-(2q-1)$ in state l. Thus, the decision maker's period 0 discounted sum of expected utilities from the strategy β is given by

$$V(\beta) = (2q-1)(\eta V^h(\beta) - (1-\eta)V^l(\beta)),$$

where

$$V^h(\beta) = \sum_{t=1}^{\infty} \delta^t X_t^h; \qquad V^l(\beta) = \sum_{t=1}^{\infty} \delta^t X_t^l$$

Completing the geometric sums, we have

$$V(\beta) = \delta(2q-1)\left(\beta_\emptyset + \frac{\delta\beta_c}{1-\delta}\right)\left(\frac{\eta}{1-\delta B^h} - \frac{1-\eta}{1-\delta B^l}\right). \tag{1}$$

An optimal strategy β maximizes $V(\beta)$ subject to $\beta_m \in [0,1]$ for each $m \in \{\emptyset, +, -, c\}$.

Intuitively, when the prior η on the state being h is very high, it would be optimal to experiment for all memory states, while if it is very low, then playing the safe arm would be optimal. The interesting decisions on whether to experiment or not (and for which memory states) occur for intermediate ranges of η. To characterize optimal strategies, we need the following three threshold values for the prior η. Define

$$\eta_0 = \frac{1-\delta q}{2-\delta} \qquad \text{and} \qquad \eta_1 = q.$$

Note that η_0 and η_1 satisfy

$$\frac{1-\eta_0}{\eta_0} = \frac{1-\delta(1-q)}{1-\delta q};$$
$$\frac{1-\eta_1}{\eta_1} = \frac{1-q}{q}.$$

Since $q > \frac{1}{2}$, we have $\eta_0 < \frac{1}{2} < \eta_1$. Also define η_* such that

$$\frac{1-\eta_*}{\eta_*} = \left(\frac{1-\eta_1}{\eta_1}\right)\left(\frac{1-\eta_0}{\eta_0}\right)^2.$$

It is straightforward to verify that $\eta_* \in (\eta_0, \eta_1)$ because $q > \frac{1}{2}$ and $\delta < 1$. Next, for each $\eta \in [\eta_*, \eta_1]$, define $K(\eta)$ such that

$$\left(\frac{1-q}{q}\right)\left(\frac{1+\delta q K(\eta)}{1+\delta(1-q)K(\eta)}\right)^2 = \frac{1-\eta}{\eta}. \tag{2}$$

Note that K is a strictly decreasing function in η, with $K(\eta_*) = 1/(1-\delta)$ and $K(\eta_1) = 0$.

We have the following characterization of optimal learning strategies:[4]

Proposition 1 *An optimal strategy β satisfies: (i) (no experimentation) $\beta_\emptyset = \beta_c = 0$ for $\eta \leq \eta_0$; (ii) (pure experimentation) $\beta_\emptyset = \beta_+ = 1$ and $\beta_- = \beta_c = 0$ for $\eta \in (\eta_0, \eta_*]$; (iii) (random and periodic experimentation) $\beta_\emptyset = \beta_+ = 1$, and β_- and β_c satisfy $(1-\beta_-)/(1-\delta(1-\beta_c)) = K(\eta)$ for $\eta \in (\eta_*, \eta_1]$; and (iv) (always experiment) $\beta_\emptyset = \beta_+ = \beta_- = 1$ for $\eta > \eta_1$.*

Thus, a pure strategy is uniquely optimal in cases (i), (ii) and (iv) above. For a sufficiently pessimistic prior (case (i), $\eta \leq \eta_0$), the optimal strategy calls for no experimentation from the start and no experimentation ever. In the opposite extreme when the prior is sufficiently strong (case (iv), $\eta > \eta_1$), the optimal strategy calls for experimentation from the start and continuing

[4] We do not give the value of β_m for an optimal strategy if m occurs with 0 probability under the strategy. Thus, β_+ and β_- are unrestricted in case (i) below and β_c is unrestricted in case (iv).

experimentation regardless of the experience from the risky arm. For intermediate priors just above the *no experiment* region (case (iii), $\eta \in [\eta_0, \eta_*)$), the optimal strategy calls for initial experimentation, continuing experimentation until the first negative payoff from the risky arm and no experimentation thereafter.

The most interesting region is the case of intermediate priors just below the *always experiment* region. Here there exists a continuum of optimal strategies involving β_- and β_c. From the expression for $V(\beta)$ in (1), we can see that it is always optimal to set $\beta_\emptyset$ to 0 or 1. Further, it is intuitive that as the memory state "+" is the most favorable, β_+ should be set to 1 if there is a positive probability of experimentation in any memory state.[5] However, the memory states "−" and "c" can involve random and periodic experimentation. By "random experimentation," we mean that β_- is strictly between 0 and 1 i.e. in the memory state "−", it is optimal to randomize between experimentation and not. By "periodic experimentation," we mean that β_c is great than 0 i.e. there maybe a continuous series of adopting the safe arm, followed by experimentation (even though no new information has been obtained). Such random and/or periodic experimentation is never observed under full memory. Optimal strategies under short-term memory (if the prior η is in the intermediate range $(\eta_*, \eta_1]$)require the right combination of periodic experimentation with random experimentation, so that[6]

$$\frac{1-\beta_-}{1-\delta(1-\beta_c)} = K(\eta).$$

Since $K(\eta)$ is decreasing in η, a more favorable prior about the risky arm tends to increase both β_- and β_c. However, due to the multiplicity of op-

[5] The proof of Proposition 1 makes this point formal by showing that the derivative of V with respect to β_+ is strictly positive whenever the derivatives of V with respect to β_c or β_- are weakly positive.

[6] Since $\beta_+ = 1$, how frequent a learning strategy plays the risky arm is determined by β_- and β_c. Intuitively, the ratio $\frac{1-\beta_-}{1-\delta(1-\delta_c)}$ measures how frequent the learning strategy plays the safe arm. The constraint on β_- and β_c below shows that β_- and β_c matter only through their effects on this ratio.

timal strategies, the experimentation probabilities in memory states "$-$" and "c" are not necessarily monotone in the prior η. Instead, the two variables β_- and β_c are carefully calibrated to balance the need to engage in some experimentation and at the same time the need to respond to negative information.

Since $K(\eta)$ satisfies

$$0 \leq K(\eta) \leq \frac{1}{1-\delta}$$

for all $\eta \in [\eta_*, \eta_1]$, the constraint on β_c and β_- can always be satisfied by $\beta_c = 0$ and $\beta_- = 1-(1-\delta)K(\eta)$. Thus, there is always an optimal strategy that uses random experimentation alone. Periodic experimentation can be optimal but is not implied by optimality.

On the other hand, for a range of values of η in the *random and periodic experimentation* region, there is an optimal learning strategy that does not use random experimentation. Define η_{**} such that $K(\eta_{**}) = 1$, or

$$\left(\frac{1-q}{q}\right)\left(\frac{1+\delta q}{1+\delta(1-q)}\right)^2 = \frac{1-\eta_{**}}{\eta_{**}}.$$

Since $K(\eta)$ is a decreasing function, we have $\eta_* < \eta_{**} < \eta_1$. Then, for all $\eta \in [\eta_*, \eta_{**}]$, we can find $\beta_c \in [0,1]$ such that

$$\frac{1}{1-\delta(1-\beta_c)} = K(\eta).$$

Thus, in this range random experimentation can be optimal but is not implied by optimality. Alternately, for $\eta \in (\eta_{**}, \eta_1]$, any optimal strategy under short-term memory must involve random experimentation i.e. $\beta_- > 0$.

4 Time (In)consistency

In this section we ask whether any of the optimal strategies characterized in the previous section is time consistent. The decision-maker here chooses his strategy β once-and-for-all at the beginning of period 0; so the question here is whether given a chance, he would like to change his optimal strategy

at any point in time. To answer this question, we need to assume that the decision maker is "introspective" in spite of the short memory constraint. This assumption requires that the decision maker remember the strategy he is carrying out, and be capable of updating his belief about the risky arm based on the current memory state and the knowledge that he has acted (previously) according to the optimal strategy. The issue of time (in)consistency of an optimal strategy then reduces to the question of whether or not there is a memory state m along the path at which the decision maker wants to deviate from the (original) prescribed choice β_m, with his updated belief now taken as the prior i.e. whether $\beta_m = \beta_\emptyset(\eta^u_m)$, where η^u_m is the updated prior when the memory state is m.

The short term memory constraint means that the decision maker can not recall the calendar time except at the very first period, i.e. when the memory state m is $\emptyset$. Thus, we have $\Pr[h|\emptyset] = \eta$, and there remain three updated beliefs to compute, $\Pr[h|m]$ for $m \in \{+, -, c\}$. To define how the belief about the risky arm is updated under any given strategy β, we use the concept of "consistent beliefs" a la Piccione and Rubinstein (1997). The idea is to use "Bayes' rule" to compute the updated beliefs along the path implied by β, even though the constraint of short term memory implies that the numbers assigned to events are not probability numbers because they can exceed 1. Further, due to the infinite horizon in our model, these numbers can be infinity. We resolve this issue by introducing a small probability τ in every period that the decision problem terminates in that period after the choice between e and s is made, and then take τ to zero in the limit.[7] Then, we have

$$\Pr[h|+] = \lim_{\tau \to 0} \frac{\eta \sum_{t=1}^{\infty} \tau(1-\tau)^t q X^h_t}{\eta \sum_{t=1}^{\infty} \tau(1-\tau)^t q X^h_t + (1-\eta) \sum_{t=1}^{\infty} \tau(1-\tau)^t (1-q) X^l_t}.$$

The interpretation is the decision maker assesses the belief about the risky arm conditional on that the decision problem has stopped and the memory

[7] We are inspired by Wilson's (2004) model of limited memory capacity with one time decisions and an exogenous termination probability.

state is $+$. Using the expressions for X_t^h and X_t^l and taking the limit, we have

$$\eta_+^u = \Pr[h|+] = \frac{\eta q(1-B^l)}{\eta q(1-B^l) + (1-\eta)(1-q)(1-B^h)}.$$

Similar calculations lead to

$$\eta_-^u = \Pr[h|-] = \frac{\eta(1-q)(1-B^l)}{\eta(1-q)(1-B^l) + (1-\eta)q(1-B^h)},$$

and

$$\eta_c^u = \Pr[h|c] = \frac{\eta(1-\beta_c - B^h)}{\eta(1-\beta_c - B^h) + (1-\eta)(1-\beta_c - B^l)}.$$

Using these updated beliefs as the priors in the respective memory states, we have the following result regarding time consistency of optimal learning strategies.

Proposition 2 *An optimal strategy for prior η is time consistent if and only if $\eta \in [0, \eta_0] \cup [\eta_1, 1]$.*

One can easily verify that

$$\left(\frac{1-q}{q}\right)\left(\frac{1-B^h}{1-B^l}\right) \leq 1$$

for any β, with equality if and only if $\beta_- = 1$ and $\beta_c = 0$. Therefore,

$$\eta_+^u = \Pr[h|+] \geq \eta$$

i.e. the decision maker always becomes more optimistic about the risky arm after a positive payoff regardless the strategy he is using (not just the optimal strategies). Note that the optimal strategies given by Proposition 1 have the property that β_+ is either 0 or 1, and whenever $\beta_+ = 1$ for some η then $\beta_\emptyset = 1$ for all higher priors. Since a positive payoff never depresses the decision maker's belief, if an optimal strategy calls for experimentation after a positive payoff i.e. if $\beta_+ = 1$, he would not want to change the decision if he takes the updated belief as his prior i.e. $\beta_\emptyset(\eta_+^u) = 1$. Therefore,

the issue of time inconsistency does not arise after a positive payoff from experimentation.

Time consistency does not necessarily arise after a negative payoff from the risky arm. When the decision maker starts with a very optimistic belief (in the *always experiment* region i.e. for $\eta > \eta_1$), it turns out that his updated belief after a negative payoff, η^u_-, remains sufficiently upbeat so that he will not deviate from the prescribed choice of $\beta_- = 1$.

However, time consistency becomes a problem for all intermediate values of the prior; this is however for different reasons, depending on whether the prior is in the *pure experimentation* region or in the *random and periodic experimentation* regions. In the former case, the decision maker is supposed to stop at the first instance of a negative payoff, but the updated belief η^u_- would suggest experimentation is optimal. In fact, in this case the updated belief in the memory state "$-$" is equal to the prior η — according to the optimal strategy in this region, the first negative payoff could be either after a series of positive payoffs from the risky arm, which would lead to a rather favorable belief, or actually the first payoff, which would result in an unfavorable belief. The situation in the *random and periodic experimentation* region is more complicated. Essentially, since the probability of experimentation at the beginning of the decision process (i.e. for the null history $\emptyset$) is either 0 or 1 in any optimal strategy (except for the prior $\eta = \eta_0$), random and periodic experimental decisions (i.e. β_- and/or β_c in the interior) can not be time consistent because they would require the updated beliefs, η^u_- and/or η^u_c, to be precisely η_0.

Thus, an optimal strategy is time consistent only in the *never experiment* and *always experiment* regions. These two regions are precisely where the decision maker does not respond to new information, and there is no learning going on. In our model of dynamic decisions with short term memory, optimal learning and time consistency are necessarily linked to each other. Since η_0 decreases with q and η_1 increases with q, the incidence of

time inconsistency in optimal learning increases with the quality of signal. Further, since η_0 decreases with δ, time inconsistency in optimal learning is more likely to arise with a more patient decision maker.

5 Memory Manipulation

If one casts the behavioral strategies of the decision maker with a short term memory as finite automata, then we have considered only varying the action rule while exogenously fixing the transition rule from one memory state to another, namely the memory in any period is the previous period's experience. However, the characterization of optimal learning strategies in Proposition 1, and in particular, random and periodic experimentations being optimal, strongly suggests that the decision maker may want to vary the transition rule as well. In other words, given that has a short-term memory constraint, he may choose to manipulate what he remembers from a particular period's experience.

In general, different forms of memory manipulation may be considered. For example, the decision maker may record a negative payoff as being positive. Since we assume that the decision maker can recall his own strategy, including possible manipulations of memory states, pure relabeling of memory states will not add value. Hence we will directly assume that while the decision-maker cannot deliberately misrepresent his current experience, he may choose to simply not record it, instead retaining his memory state at the beginning of the period. This may be thought of as "endogenous forgetfulness." The questions we are interested in exploring are when is such forgetfulness optimal, and relatedly, what type of experience will it be optimal to forget?

In this section we focus on the possible manipulation of only the initial memory. Recall that the decision maker begins period 0 with the memory state $\emptyset$. Here we investigate whether the decision maker can improve his period 0 discounted sum of expected utilities by retaining the "clean slate"

of null history (i.e., the memory state $\emptyset$) instead of recording the payoff from the most recent experimentation. The interpretation in generational learning would be that the generation that has made their choice does not always admit its experience to the next generation of decision makers.

Formally, if the memory state at the beginning of a period is $\emptyset$, then for each current period outcome $m \in \{+, -, c\}$, let γ_m be the probability of replacing the memory state $\emptyset$ with m. While in the previous section $\gamma_m = 1$, now γ_m is a choice variable for the decision-maker. Memory manipulation with respect to the null history state $\emptyset$ occurs when $\gamma_m < 1$ for some $m \in \{+, -, c\}$. To simplify the analysis and build insight, memory manipulation in memory states other than γ_m are ruled out by assumption, so that when the beginning of period memory state is any $m \neq \emptyset$, then with probability 1 the decision maker replaces m with his current period experience.

Denote $\gamma = (\gamma_+, \gamma_-, \gamma_c)$. Along with β, the decision maker now chooses γ as well to maximize $W(\beta; \gamma)$, the period 0 discounted sum of expected utilities.

Fix a strategy β and γ. Suppose that the underlying true state is h. Let P_t^h, N_t^h, Z_t^h and F_t^h be the ex ante probability (i.e., from period 0 perspective) of the memory state being $+$, $-$, c and ϕ respectively, at the beginning of period t, $t = 1, 2, \ldots$. The decision maker then makes her experimental decision according to β, and her memory decision for the following period according to γ. Together, the evolution of $(P_t^h, N_t^h, Z_t^h, F_t^h)$ is determined by the following transition matrix:

$$\begin{bmatrix} \beta_+ q & \beta_- q & \beta_c q & \beta_\emptyset q \gamma_+ \\ \beta_+(1-q) & \beta_-(1-q) & \beta_c(1-q) & \beta_\emptyset(1-q)\gamma_- \\ 1-\beta_+ & 1-\beta_- & 1-\beta_c & (1-\beta_\emptyset)\gamma_c \\ 0 & 0 & 0 & \Lambda^h \end{bmatrix}$$

where

$$\Lambda^h = (1-\beta_\emptyset)(1-\gamma_c) + \beta_\emptyset(q(1-\gamma_+) + (1-q)(1-\gamma_-)).$$

Note that $\Lambda^h = 0$ if there is no memory manipulation, and we are back in the formulation of the previous section. The initial values are given by $P_1^h = N_1^h = Z_1^h = 0$ and $F_1^h = 1$. It follows from the transition matrix that

$$F_t^h = (\Lambda^h)^{t-1}$$

for each t.

Define

$$X_t^h = P_t^h \beta_+ + N_t^h \beta_- + Z_t^h \beta_c + F_t^h \beta_\emptyset$$

as the aggregate probability of experimentation in period t from a period 0 perspective. We claim that

$$X_{t+1}^h = B^h X_t^h + \beta_c + G^h F_t^h$$

for each $t \geq 1$, where B^h is as defined in section 3 and

$$G^h = (\beta_\emptyset - \beta_+)\beta_\emptyset q(1-\gamma_+) + (\beta_\emptyset - \beta_-)\beta_\emptyset(1-q)(1-\gamma_-) + (\beta_\emptyset - \beta_c)(1-\beta_\emptyset)(1-\gamma_c).$$

This can be verified by using

$$P_t^h + N_t^h + Z_t^h + F_t^h = 1$$

for each $t \geq 1$ and the transition matrix to establish it as an identity in P_t^h, N_t^h, Z_t^h and F_t^h. The explicit solution to the above difference equation is then

$$X_t^h = \beta_\emptyset (B^h)^{t-1} + \frac{\beta_c(1-(B^h)^{t-1})}{1-B^h} + \frac{G^h((\Lambda^h)^{t-1} - (B^h)^{t-1})}{\Lambda^h - B^h}.$$

As in section 2, define

$$\begin{aligned} W^h(\beta;\gamma) &= \sum_{t=1}^{\infty} \delta^t X_t^h \\ &= \frac{\delta}{1-\delta B^h}\left(\beta_\emptyset + \frac{\delta\beta_c}{1-\delta} + \frac{\delta G^h}{1-\delta\Lambda^h}\right). \end{aligned}$$

This reduces to $V^h(\beta)$ of section 2 when there is no memory manipulation.

Symmetrically, defining

$$W^l(\beta;\gamma) = \sum_{t=1}^{\infty} \delta^t X_t^l$$

and deriving X_t^l in the same way as for X_t^h, we have

$$W^l(\beta;\gamma) = \frac{\delta}{1-\delta B^l}\left(\beta_\emptyset + \frac{\delta\beta_c}{1-\delta} + \frac{\delta G^l}{1-\delta\Lambda^l}\right),$$

where

$$\Lambda^l = (1-\beta_\emptyset)(1-\gamma_c) + \beta_\emptyset((1-q)(1-\gamma_+) + q(1-\gamma_-)),$$

and

$$G^l = (\beta_\emptyset - \beta_+)\beta_\emptyset(1-q)(1-\gamma_+) + (\beta_\emptyset - \beta_-)\beta_\emptyset q(1-\gamma_-) + (\beta_\emptyset - \beta_c)(1-\beta_\emptyset)(1-\gamma_c).$$

Finally, we can write

$$W(\beta;\gamma) = (2q-1)(\eta W^h(\beta;\gamma) - (1-\eta)W^l(\beta;\gamma))$$

which is equal to $V(\beta)$ when $\gamma = 1$. Comparing the case of manipulation versus no-manipulation of memory, we have the following result:

Proposition 3 *For all $\eta \in (\eta_*, \eta_1)$, $\max_\beta V(\beta) < W(\beta';\gamma)$ for some β' and $\gamma \neq 1$.*

By definition, $\max_\beta W(\beta;1)$ is the optimal value of the period 0 discounted sum of utilities when there is no memory manipulation. From the characterization of Proposition 1 over the range $\eta \in (\eta_*, \eta_1)$, this optimal value can be attained by using random and periodic experimentation, with $\beta_\emptyset = \beta_+ = 1$ and β_- and β_c such that $(1-\beta_-)/(1-\delta(1-\beta_c)) = K(\eta)$. The claim of Proposition 3 is established by showing at any such optimal β with no manipulation, there exists $\gamma \neq 1$ such that $W(\beta;\gamma) > W(\beta;1)$.

The rough intuition behind Proposition 3 may be understood as follows. Without memory manipulation, there are effectively only three memory states, $+$, $-$ and c, because the initial memory state of $\emptyset$ exists only for

the first period.[8] Unlike those for $+$, $-$ and c, the experimental decision corresponding to $\emptyset$ is one time only. By the characterization of the optimal learning strategy in Proposition 1, this decision $\beta_\emptyset$ is equal to 1 if the value of objective function under an optimal learning strategy is positive, and 0 otherwise. In contrast, memory manipulation allows the decision maker to make the memory state $\emptyset$ a recurring state. This can help improve the decision maker's ex ante welfare because an additional memory state can be used to enrich the state space and allow the strategy to better respond to new information. Of somewhat greater interest is the precise way that this increase in value is achieved through memory manipulation i.e. which is/are the memory states that the decision chooses to forget?

This can be made more precise by following the steps of the proof of Proposition 3. We first observe that with $\beta_\emptyset = \beta_+ = 1$, the decision maker attains the same ex ante payoff by setting $\gamma_+ = 0$ and $\gamma_- = 1$ as by no memory manipulation (i.e. by setting $\gamma_+ = \gamma_- = 1$). The path of decisions is identical in these two scenarios: if the payoff from the risky arm in the first period is negative then because $\gamma_- = 1$, one exits from the memory state $\emptyset$, and follows the same path as no memory manipulation; while if the first period payoff is positive, then too the same sequence of decisions are made even though $\gamma_+ = 0$, as $\beta_\emptyset = \beta_+$.

We ask if the decision maker can improve his ex ante payoff by reducing γ_- from 1 while maintaining $\gamma_+ = 0$. The key is to note that under $\gamma_+ = 0$ and $\gamma_- = 1$ the memory state of $\emptyset$ carries information distinct from the memory state of $+$: the state of $\emptyset$ occurs only after a string of positive payoffs from the risky arm, whereas the state of $+$ occurs only after getting at least one negative payoff in the past. The former suggests a more favorable belief about the risky arm and thus should lead to a greater probability of experimentation than the latter, but such a distinction cannot be made when there is no memory manipulation by the decision maker. With memory

[8] By assumption, the decision maker does not recall calendar time but is able to distinguish the first period from the rest of decision nodes.

manipulation, this can be exploited by the decision maker by reducing γ_- to just below 1. Then, the decision maker has a positive probability of ignoring a negative payoff when the current memory state is $\emptyset$. For small reductions in γ_-, the benefit of increasing experimentation when the state is likely to be h outweighs the potential cost of repeatedly ignoring the unfavorable information of negative payoffs.

Proposition 3 is proved by changing γ while maintaining the same optimal β as under no manipulation. This raises the question of whether the decision maker not only wants to change γ but also wishes to deviate from the optimal β with no manipulation. The answer is yes. To see this, for any β and γ such that $\beta_\emptyset = \beta_+ = 1$, we can write $W(\beta;\gamma)$ as

$$\frac{\delta(2q-1)\eta}{1+\delta(1-q)K}\left(\frac{1}{1-\delta}+\frac{\delta(1-q)(1-\gamma_-)K}{1-\delta\Lambda^h}\right) - \frac{\delta(2q-1)(1-\eta)}{1+\delta qK}\left(\frac{1}{1-\delta}+\frac{\delta q(1-\gamma_-)K}{1-\delta\Lambda^l}\right),$$

where

$$K=\frac{1-\beta_-}{1-\delta(1-\beta_c)},$$

and

$$\Lambda^h = q(1-\gamma_+)+(1-q)(1-\gamma_-); \qquad \Lambda^l = (1-q)(1-\gamma_+)+q(1-\gamma_-).$$

Thus, as in section 2, β_- and β_c matter only through K. The derivative of $W(\beta;\gamma)$ with respect to γ_+ has the same sign as

$$-\frac{\eta(1-q)}{1+\delta(1-q)K}\frac{q}{(1-\delta\Lambda^h)^2}+\frac{(1-\eta)q}{1+\delta qK}\frac{(1-q)}{(1-\delta\Lambda^l)^2}.$$

It is straightforward to verify that the second derivative of $W(\beta;\gamma)$ with respect to γ_+ is strictly positive when the first derivative is zero. Similarly, the derivative of $W(\beta;\gamma)$ with respect to γ_- has the same sign as

$$-\frac{\eta(1-q)}{1+\delta(1-q)K}\frac{1-\delta q(1-\gamma_+)}{(1-\delta\Lambda^h)^2}+\frac{(1-\eta)q}{1+\delta qK}\frac{1-\delta(1-q)(1-\gamma_+)}{(1-\delta\Lambda^l)^2},$$

with a strictly negative sign for the second derivative when the first derivative is zero. Further, one can easily check that $\partial W/\partial\gamma_+ \geq 0$ implies that $\partial W/\partial\gamma_- > 0$. It follows that the optimal value for γ_+ is either 0 or 1, and $\gamma_- \geq \gamma_+$ at an optimum.

The derivative of $W(\beta;\gamma)$ with respect to K has the same sign as

$$-\frac{\eta(1-q)}{(1+\delta(1-q)K)^2}A^h + \frac{(1-\eta)q}{(1+\delta qK)^2}A^l,$$

where

$$\begin{aligned} A^h &= 1 - \frac{(1-\delta)(1-\gamma_-)}{1-\delta+\delta q\gamma_+ + \delta(1-q)\gamma_-}; \\ A^l &= 1 - \frac{(1-\delta)(1-\gamma_-)}{1-\delta+\delta(1-q)\gamma_+ + \delta q\gamma_-}. \end{aligned}$$

If $\gamma_+ < \gamma_-$, then $A^h < A^l$ and therefore $\partial W(\beta;\gamma)/\partial K > 0$ at $K = K(\eta)$. We already know from Proposition 3 above that for any $\eta \in (\eta_*, \eta_1)$ (i.e. in the *random and periodic experimentation* region), the decision maker can improve ex ante welfare by memory manipulation without changing the optimal learning strategy β under no manipulation. At any such manipulation we must have $\gamma_+ < \gamma_-$, which then implies that the decision maker could further increase his ex ante payoff with changes in the learning strategy β by increasing K.[9]

This issue of doing better through memory manipulation by at the same time deviating from the no-manipulation β acquires more importance in the pure experimentation range i.e. $\eta \in (\eta_0, \eta_*]$ where the optimal no manipulation β consists of $\beta_\emptyset = \beta_+ = 1$ and $\beta_- = \beta_c = 0$. Denote this β as β^{nm}. Now for a given γ, we have:

$$W^h(\beta^{nm};\gamma) - W^h(\beta^{nm};1) = \frac{\delta^2}{1-\delta q}\frac{(1-q)(1-\gamma_-)}{1-\delta[q(1-\gamma_+)+(1-q)(1-\gamma_-)]}$$

[9] Given the interpretation of K as a measure of the frequency of playing the safe arm, an increase in K compensates the increase in the probability of experimentation that comes with memory manipulation (i.e. a decrease in γ_- to below 1).

Similarly computing $W^l(\beta^{nm};\gamma) - W^l(\beta^{nm};1)$ and combining the two, we obtain the difference between memory-manipulation and no-manipulation as $W(\beta^{nm};\gamma) - W(\beta^{nm};1) =$

$$\delta^2(2q-1)\{\frac{\eta}{1-\delta q}\frac{(1-q)(1-\gamma_-)}{1-\delta[q(1-\gamma_+)+(1-q)(1-\gamma_-)]} - \frac{1-\eta}{1-\delta(1-q)}\frac{q(1-\gamma_-)}{1-\delta[(1-q)(1-\gamma_+)+q(1-\gamma_-)]}\}$$

This expression has the same sign as:

$$E = \frac{\eta}{1-\eta} - \frac{q}{1-q}\frac{1-\delta q}{1-\delta(1-q)}\frac{1-\delta[q(1-\gamma_+)+(1-q)(1-\gamma_-)]}{1-\delta[(1-q)(1-\gamma_+)+q(1-\gamma_-)]}$$

It is easy to check that the second term here is increasing in γ_+ and decreasing in γ_-, and thus achieves its lowest value when $\gamma_+ = 0$ and $\gamma_- = 1$. Therefore we have

$$E \leq \frac{\eta}{1-\eta} - \frac{q}{1-q}\left(\frac{1-\delta q}{1-\delta(1-q)}\right)^2 = \frac{\eta}{1-\eta} - \frac{\eta_*}{1-\eta_*}$$

which is non-positive for $\eta \in (\eta_0, \eta_*]$. Thus, in the pure experimentation region, we have $W(\beta^{nm};\gamma) - W(\beta^{nm};1)$ with equality only if and only if $\eta = \eta_*$. This implies that in this range if there is to be any gain from memory-manipulation, one has to necessarily deviate from the no-manipulation optimal strategy of β^{nm}.

6 Open Questions

This paper analyzes a simple model of dynamic decisions with short term memories. We have looked at a two armed bandit problem as a framework suggestive of the time inconsistency and memory manipulation issues we want to study in generational learning. The short term memory constraint takes a particularly simple form in our model. It will be worthwhile to pursue more general forms of such a constraint, for example by allowing the decision maker to recall the past experience of more than a single period.

In particular, we have shown that optimal learning strategies are necessarily time inconsistent if they are responsive to new information. Whether this is true with more general dynamic decision problems and more general short term memory constraints remain to be seen. Further, the memory manipulation considered in this paper is one of many ways available to the decision maker. Whether, and how, other kinds of manipulation can improve the ex ante welfare of the decision maker are interesting topics that we plan to pursue in future research. Finally, we have treated the issues of time inconsistency and memory manipulation separately. Is there a link between these two issues? In particular, does manipulation make the optimal policy more likely to be time inconsistent?

7 Appendix: Proofs

7.1 Proof of Proposition 1

Since $V(\beta)$ is linear in $\beta_\emptyset$, we have $\beta_\emptyset = 1$ if $V(\beta) > 0$ at any optimal β, and $\beta_\emptyset = 0$ otherwise. The 0 payoff can be implemented by setting $\beta_\emptyset = \beta_c = 0$, regardless of β_+ and β_-. We have:

$$V(\beta_\emptyset = \beta_c = 0) = 0.$$

For the remainder of the proof, we assume that $\beta_\emptyset = 1$.

The derivatives of $V(\beta)$ with respect to β_-, β_c and β_+ are given by:

$$\frac{\partial V}{\partial \beta_-} = \frac{\delta^2}{1-\delta}(2q-1)\left(1-\delta+\delta\beta_c\right)\left(\frac{\eta(1-q)}{(1-\delta B^h)^2} - \frac{(1-\eta)q}{(1-\delta B^l)^2}\right),$$

$$\frac{\partial V}{\partial \beta_c} = \frac{\delta^2}{1-\delta}(2q-1)\left(\frac{\eta(1-(q\beta_+ + (1-q)\beta_-))}{(1-\delta B^h)^2} - \frac{(1-\eta)(1-((1-q)\beta_+ + q\beta_-)}{(1-\delta B^l)^2}\right),$$

$$\frac{\partial V}{\partial \beta_+} = \frac{\delta^2}{1-\delta}(2q-1)\left(1-\delta+\delta\beta_c\right)\left(\frac{\eta q}{(1-\delta B^h)^2} - \frac{(1-\eta)(1-q)}{(1-\delta B^l)^2}\right).$$

It is straightforward to verify that

$$\frac{1-q}{q} \leq \frac{1-(q\beta_+ + (1-q)\beta_-)}{1-((1-q)\beta_+ + q\beta_-)} \leq \frac{q}{1-q},$$

with the first as an equality if and only if $\beta_+ = 1$, and the second as an equality if and only if $\beta_- = 1$.

It follows that the signs of the derivatives of $V(\beta)$ with respect to β_-, β_c and β_+ are ordered: $\partial V/\partial\beta_- \geq 0$ implies that $\partial V/\partial\beta_c > 0$ if $\beta_+ < 1$, and the two have the same sign if $\beta_+ = 1$; while $\partial V/\partial\beta_c \geq 0$ implies that $\partial V/\partial\beta_+ > 0$ if $\beta_- < 1$, and the two have the same sign if $\beta_- = 1$. We distinguish the following three cases.

(1) If $\beta_+ = 0$, then $\partial V/\partial\beta_+ \leq 0$ at the optimum, implying that $\partial V/\partial\beta_c < 0$ and $\partial V/\partial\beta_- < 0$, and therefore $\beta_c = \beta_- = 0$. In this case

$$V(\beta_\emptyset = 1, \beta_+ = \beta_- = \beta_c = 0) = \delta(2q-1)(2\eta-1).$$

(2) If β_+ is in the interior, then $\beta_c = \beta_- = 0$ as in case (1). We have:

$$\frac{\partial V}{\partial \beta_+} = \frac{\delta^2(2q-1)}{1-\delta}\left(\frac{\eta q}{(1-\delta q\beta_+)^2} - \frac{(1-\eta)(1-q)}{(1-\delta(1-q)\beta_+)^2}\right).$$

Thus, there can be at most one critical point at which $\partial V/\partial\beta_+ = 0$. Evaluating the second derivative at this point, we find that $\partial^2 V/\partial\beta_+^2$ has the same sign as

$$\frac{q}{1-\delta q\beta_+} - \frac{1-q}{1-\delta(1-q)\beta_+},$$

which is positive because $q > \frac{1}{2}$. It follows that an interior β_+ can not be optimal.

(3) If $\beta_+ = 1$, then $\partial V/\partial\beta_+ \geq 0$ at the optimum. This case allows for interior solutions in β_c and β_-. Since $\beta_+ = 1$, the signs of $\partial V/\partial\beta_c$ and $\partial V/\partial\beta_-$ are the same, and so both β_c and β_- can be interior at the same time. Indeed, with $\beta_\emptyset = \beta_+ = 1$, we can rewrite V as:

$$V = \frac{\delta(2q-1)}{1-\delta}\left(\frac{\eta}{1+\delta(1-q)K} - \frac{1-\eta}{1+\delta qK}\right),$$

where

$$K = \frac{1-\beta_-}{1-\delta(1-\beta_c)}.$$

By definition, we have $0 \leq K \leq 1/(1-\delta)$. Since V depends on β only through K, we can take derivatives with respect to K and get the following first order condition:

$$-\frac{\eta(1-q)}{(1+\delta(1-q)K)^2} + \frac{(1-\eta)q}{(1+\delta qK)^2} = 0.$$

Define $K(\eta)$ as the point that satisfies the above first order condition. This is done in (2). It is straightforward to verify that the second order condition is satisfied at $K = K(\eta)$. Thus, if $K(\eta) \in [0, 1/(1-\delta)]$, the maximal payoff with $\beta_\emptyset = \beta_+ = 1$ is reached when β_c and β_- satisfy

$$\frac{1-\beta_-}{1-\delta(1-\beta_c)} = K(\eta).$$

Using the definitions of $K(\eta)$, η_1 and η_*, we have $K(\eta) \geq 0$ if and only if $\eta \leq \eta_1$, while $K(\eta) \leq 1/(1-\delta)$ if and only if $\eta \geq \eta_*$. The maximal payoff with $\beta_\emptyset = \beta_+ = 1$ for $\eta \in [\eta_*, \eta_1]$ is thus given by

$$\begin{aligned} V(\beta_\emptyset &= \beta_+ = 1, \beta_- \text{ and } \beta_c \text{ s.t. } \frac{1-\beta_-}{1-\delta(1-\beta_c)} = K(\eta)) \\ &= \frac{\delta(2q-1)}{1-\delta}\left(\frac{\eta}{1+\delta(1-q)K(\eta)} - \frac{1-\eta}{1+\delta qK(\eta)}\right). \end{aligned}$$

For all $\eta > \eta_1$, one can verify that $\partial V/\partial K < 0$ for all $K > 0$, implying that $K = 0$ at the optimum and thus $\beta_- = 1$ (β_c is unrestricted). The maximal payoff with $\beta_\emptyset = \beta_+ = 1$ for $\eta \geq \eta_1$ is then given by

$$V(\beta_\emptyset = \beta_+ = \beta_- = 1) = \frac{\delta(2q-1)(2\eta-1)}{1-\delta}.$$

For all $\eta < \eta_*$, we have $\partial V/\partial K > 0$, implying that $K = 1/(1-\delta)$ at the optimum and thus $\beta_- = \beta_c = 0$. The maximal payoff with $\beta_\emptyset = \beta_+ = 1$ for $\eta \leq \eta_*$ is then given by

$$V(\beta_\emptyset = \beta_+ = 1, \beta_c = \beta_- = 0) = \delta(2q-1)\left(\frac{\eta}{1-\delta q} - \frac{1-\eta}{1-\delta(1-q)}\right).$$

From the definition for η_0, this is positive if and only if $\eta > \eta_0$.

Comparing case (3) with case (1), note that $K = 0$ yields the same value as $\frac{1}{1-\delta}V(\beta_\emptyset = 1, \beta_+ = \beta_- = \beta_c = 0)$. Therefore whenever case (1) gives a positive value for V, it is dominated by case (3). Thus, case (1) can not occur at the optimum.

The complete characterization of optimal strategies in Proposition 1 then follows immediately from the analysis of case (3).

7.2 Proof of Proposition 2

We check time consistency for each of the four cases in Proposition 1 separately.

Case (i) $\eta \leq \eta_0$: The only memory state that occurs with positive probability after the initial period is c. To calculate $\Pr[h|c]$, we need to make assumptions on the values of β_+ and β_-, which are unrestricted in this case. We choose $\beta_+ = \beta_- = 0$, implying that

$$\Pr[h|c] = \frac{\eta}{\eta + (1-\eta)} = \eta.$$

Since the updated belief stays at η, the optimal strategy ($\beta_\emptyset = \beta_+ = \beta_- = \beta_c = 0$) is time consistent in this case.

Case (ii) $\eta \in (\eta_0, \eta_*]$: Here we have

$$\Pr[h|-] = \frac{\eta(1-q)q}{\eta(1-q)q + (1-\eta)q(1-q)} = \eta,$$

which is greater than η_0 by assumption, while $\beta_- = 0$. Thus, the optimal strategy is time inconsistent in this case.

Case (iii) $\eta \in (\eta_*, \eta_1]$. Note that the optimal $\beta_\emptyset$ is either 0 or 1, except when $\eta = \eta_0$, in which case an interior $\beta_\emptyset$ can be optimal because the decision maker is indifferent e and s. Since at least one of β_c or β_- must be in the interior in any optimal strategy in case (iii), time consistency would require the corresponding updated belief $\Pr[h|c]$ or $\Pr[h|-]$ to equal η_0.

In this case,

$$\Pr[h|-] = \frac{1}{1+\frac{(1-\eta)q[\beta_c+(1-q)(1-\beta_-)]}{\eta(1-q)[\beta_c+q(1-\beta_-)]}}, \qquad \Pr[h|c] = \frac{1}{1+\frac{(1-\eta)q(1-\beta_-)}{\eta(1-q)(1-\beta_-)}}$$

Thus,

$$\Pr[h|c] \leq \Pr[h|-],$$

with equality if and only if $\beta_- = 1$, which can only occur when $\eta = \eta_1$. Since at least one of β_c or β_- must be in the interior (and $\beta_- \neq 1$ for $\eta < \eta_1$) the only possibility for consistency is that $\Pr[h|-] = \eta_0 > \Pr[h|c]$ and $\beta_c = 0$. However this would then imply $\Pr[h|-] = \eta$ which in this case exceeds η_0 as $\eta_* > \eta_0$. Thus in this case, there is no optimal strategy which is time consistent.

Case (iv) $\eta > \eta_1$: We have

$$\Pr[h|-] = \frac{\eta(1-q)}{\eta(1-q)+(1-\eta)q}.$$

Using the definitions of η_0 and η_1, we can verify that $\Pr[h|-] > \eta_0$ for all $\eta > \eta_1$ because $\eta_0 < \frac{1}{2}$. Since $\beta_- = 1$ (and also $\beta_+ = 1$), the optimal strategy is time consistent in this case.

7.3 Proof of Proposition 3

For any $\eta \in (\eta_*, \eta_1)$, let β be such that $\beta_\emptyset = \beta_+ = 1$, $\beta_- = 0$ and β_c satisfies

$$\frac{1}{1-\delta(1-\beta_c)} = K(\eta).$$

Then, we can write the difference between $W(\beta;\gamma)$ for any γ and $W(\beta;\gamma_+ = \gamma_- = 1)$ as

$$\delta^2(2q-1)\left(\frac{(1-q)(1-\gamma_-)\eta K}{(1+\delta(1-q)K)(1-\delta\Lambda^h)} - \frac{q(1-\gamma_-)(1-\eta)K}{(1+\delta qK)(1-\delta\Lambda^l)}\right),$$

where

$$\begin{aligned}\Lambda^h &= q(1-\gamma_+)+(1-q)(1-\gamma_-);\\ \Lambda^l &= (1-q)(1-\gamma_+)+q(1-\gamma_-).\end{aligned}$$

Note that

$$W(\beta;\gamma_+=0,\gamma_-=1)-W(\beta;\gamma_+=\gamma_-=1)=0.$$

The derivative of $W(\beta;\gamma)-W(\beta;\gamma_+=\gamma_-=1)$ with respect to γ_-, evaluated at $\gamma_-=1$, has the same sign as

$$-\frac{(1-q)\eta K}{(1+\delta(1-q)K)(1-\delta q(1-\gamma_+)}+\frac{q(1-\eta)K}{(1+\delta qK)(1-\delta(1-q)(1-\gamma_+))}.$$

Since

$$\frac{(1-q)\eta}{(1+\delta(1-q)K(\eta))^2}=\frac{q(1-\eta)}{(1+\delta qK(\eta))^2},$$

the sign of the derivative evaluated at $K=K(\eta)$ is the same as

$$\frac{\delta(2q-1)K(\eta)}{1-\delta(1-\gamma_+)}\left(K(\eta)-\frac{1-\gamma_+}{1-\delta(1-\gamma_+)}\right).$$

For $\gamma_+=0$, we have $K(\eta)>0$ and $K(\eta)<1/(1-\delta)$ for all $\eta\in(\eta_*,\eta_1)$. Thus, the derivative of $W(\beta;\gamma)-W(\beta;\gamma_+=\gamma_-=1)$ with respect to γ_-, evaluated at $\gamma_-=1$, $\gamma_+=0$ and $K=K(\eta)$ is strictly positive for all $\eta\in(\eta_*,\eta_1)$. The proposition follows immediately.

Reference

Börgers, T. and A. Morales, 2004, "Complexity constraints in two-armed bandit problems: an example," University College London working paper.

Gittins, J.C., 1989, *Multi-armed Bandit Allocation Indices*, New York: John Wiley & Sons.

Kalai, E. and E. Solan, 2003, "Randomization and simplification in dynamic decision-making," *Journal of Economic Theory* 111, pp 251–264.

Kuhn, H.W., 1953, "Extensive games and the problem of information," in *Contributions to the Theory of Games III*, pp 79–96, Princeton, NJ: Princeton University Press.

Meyer, M., 1991, "Learning from coarse information: biased contests and career profiles," *Review of Economic Studies* 58, pp 15–41.

Lipman, B., 1995, "Information processing and bounded rationality: a survey," *Canadian Journal of Economics* 28, pp 42–67.

Piccione, M. and A. Rubinstein, 1997, "On the interpretation of decision problems with imperfect recall," *Games and Economic Behavior* 20, pp 2–35.

Rubinstein, A., 1986, "Finite automata play the repeated prisoners' dilemma," *Journal of Economic Theory* 39, pp 83–96.

Wilson, A., 2004, "Bounded memory and biases in information processing," Princeton University working paper.