INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

Learning Through Stories and Other Essays

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Ricky Pak Ki Lam

Dissertation Director: David Pearce

December 2000

UMI Number: 9991175

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Acknowledgments

I would like to express my gratitude to Professors Dirk Bergemann, Stephen Morris, David Pearce and Ben Polak for teaching me economic theory, and for invaluable discussions throughout the progress of this dissertation. The work presented here also benefited enormously from conversations with Ettore Damiano, Jason Draho, Yianis Sarafidis and Mario Simon.

I am grateful to David Pearce and to Peter Phillips for all their advice during my time at Yale, and to Bob Shiller for introducing me to behavioral economics. Ben Polak's counsel and encouragement during the job market is very much appreciated. Finally, I would like to give special thanks to Eugene Choo, Ettore Damiano, Mario Simon and Lori Snyder, and to my parents, Fung Ping and Wai Ying, and to my brothers, Patrick, Stephen and Alan. Without their support, I would never have stayed in graduate school long enough to finish this dissertation.

Contents

Introduction

Each chapter presented in this dissertation is motivated in some way by an interest in *learning*. The first two chapters are concerned with how people learn. An understanding of how people learn is of course crucial for understanding how they behave in economic situations where information is received over time. For many economists, "learning" is synonymous with Bayesian updating. Uncertainty is modeled as a set of possible states of the world over which the agent assigns a subjective probability measure. Learning is probabilistic conditioning on subsets of this state space. Despite its widespread use, this model is at odds with introspection and with a large body of experimental evidence.

Chapter 1, *Learning Through Stories*, is motivated by an observation from cognitive psychology: when faced with certain complex situations, people deal with uncertainty by forming "stories". A juror may form a story to explain the evidence presented. A central banker may have a story about the current state of the economy. Other examples where the idea of a story seems closer to the way we speak and think about learning include a doctor diagnosing a patient, an investor deciding whether to purchase stocks in a company, and a teacher assessing her student.

Of course, a juror's story about a case is very different in content to a doctor's story about her patient, or to a teacher's story about her student's ability. Nevertheless,

there are common features to stories. Within the theoretical model presented in this chapter, a story is one possible scenario of "what happened", or one possible state of the world. However, it is not just any scenario. A story must be coherent (at least from the storyteller's perspective); that is, it must be consistent, both internally—the different parts of a story have to "fit"—and externally with the evidence.

The chapter presents a stylized procedure of how people construct stories. The model is rich enough to capture the confusion that can be associated with difficult choices. The main result establishes a set of conditions under which this will not occur: that is, conditions under which the agent will be able to form a coherent story. Loosely, the key assumption is that the agent recognizes the contrapostive: she realizes that alternative ways of stating the same inference are in fact equivalent.

Among the other results, there can be multiple coherent interpretations of the same evidence. People who construct stories typically do not undo inferences made from evidence, even after the evidence is discredited. More generally, the order in which evidence is presented can affect the stories that they form. In addition, apparently weak evidence can trigger large changes in people's stories.

Though only hinted at in this chapter, these features of the model have important implications for behavior. The dependence on the order of signals has troubling consequences for a doctor's assessment of a patient, or a central banker's assessment of the economy. The irreversibility of signals encourages smear campaigns in politics, and may lead to lasting effects from the release of economic statistics, even when they are subsequently revised. The discontinuity in the response to signals may account for

under and overreactions in financial markets. And finally, the model has implications for the tactics that advocates should use.

Chapter 1 argues that, in many real-world situations, the Bayesian model is limited in its representation of human learning. Nevertheless, the Bayesian approach is justified by axioms on preferences which have a great deal of normative appeal. An alternative approach toward more realistic models of learning is to relax or change these axioms to be more consistent with the experimental evidence on choice under uncertainty. Within this literature, a number of theories attempt to explain the so-called paradoxes of Ellsberg.[1] Ellsberg's research indicates that most people exhibit "uncertainty-aversion". That is, they have a preference for situations where probabilities are known.

Recent work in decision theory has sought to represent the subjective beliefs of such individuals in the form of a convex non-additive measure over the state space. Chapter 2, *Revising Non-Additive Priors*, considers how these non-Bayesian beliefs can be updated. It is joint work with Yianis Sarafidis. Consider an employer who has a subjective prior over the quality of a worker and who knows the distribution for output conditioned on each level of quality. How does she learn (update her beliefs) about quality upon observing some output level? If her prior is additive, this problem is trivial: Bayes's rule suffices. First, a distribution over all possible pairs of quality and output is constructed. She can then condition on the appropriate subset

---

[1] Ellsberg, D. (1961), "Risk, Ambiguity and the Savage Axioms", *Quarterly Journal of Economics*, 75, 643-669.

of this product space to calculate the posterior on quality.

When the employer's beliefs are non-additive, calculating a measure over the product space of pairs of quality and output is no longer so simple. We propose two rules: the first uses the idea of Choquet integration over identity functions and produces a non-additive measure over the product space; the second converts the initial non-additive measure to a set of additive priors, and then applies Bayes's rule to each element in this set. We argue that the non-equivalence of these two rules highlights a limitation of non-additive measures. While this limitation does not matter for the representation of uncertainty-averse preferences, it results in a loss of information when beliefs have to be revised.

The final chapter is coauthored with Ettore Damiano. It represents the start of a research project motivated by the question of how learning restricts outcomes in matching markets. In this chapter, we focus on a special class of matching markets, termed marriage markets. Theses are trading arrangements where participants belong to two disjoints sets, and where trades require one-to-one matches. An obvious restriction on outcomes is that they be "stable". Here we have in mind cooperative concepts such as the core. A large literature has considered stability in the case of a static market with perfect information. In many real-world situations, however, the value of a proposed match is not known until after trade has taken place; participants have to learn about these parameters of the game as it is played over time. In *Self-Sustaining Stability in Dynamic Matching Markets*, we undertake the first step toward a more realistic theory of marriage markets by incorporating dynamics into a notion of stability. For most of the chapter, however, we maintain the assumption

of complete information.

We label our definition "self-sustaining stability". This concept can be viewed as the core with two additional requirements. Loosely, these requirements are the cooperative counterparts of subgame perfection and coalition proofness. We provide a justification for the concept, sufficient conditions for its existence, and an algorithm for computing it. At the end of the chapter, we extend the definition to a dynamic, incomplete information setting with learning. Characterizing this new definition is the subject of ongoing work.

Chapter 1

Learning Through Stories

Many thanks to Stephen Morris, David Pearce and Ben Polak for their invaluable advice and support. Discussions with Dirk Bergemann, Ettore Damiano, Jason Draho, John Geanakoplos, Brian Lonergan, Giuseppe Moscarini, Yianis Sarafidis, Robert Shiller, Mario Simon were very helpful. Financial support from the Cowles Foundation is gratefully acknowledged.

[Humans are...] "primates who tell stories".

Stephen Gould[1]

1.1 Introduction

An understanding of how agents learn is crucial for understanding how they behave in economic situations where information is received over time. For many economists, "learning" is synonymous with Bayesian updating.[2] Uncertainty is modeled as a set of possible states of the world over which the agent assigns a subjective probability measure. Learning is probabilistic conditioning on subsets of this state space. Despite its widespread use, this model is at odds with introspection and with evidence from psychology about how we reason.[3]

In economic theory, the Bayesian approach is justified by placing normative axioms on preferences and then *deriving* Bayes's rule as part of a utility-representation result. While elegant, this focus on preferences, to the neglect of cognition, implies that the model sometimes fails to capture essential aspects of people's internal deliberations. More damagingly (perhaps precisely because it does not explicitly model

---

[1]Citation from Dawes (1999) who modifies this definition to "primates whose cognitive capacity shuts down in the absence of a story". Thanks to Robert Forsythe for bringing this reference to my attention.

[2]This is not true in game theory. This paper, however, undertakes the simpler task of studying learning in non-strategic situations.

[3]Kahneman, Slovic and Tversky (1982), Rabin (1996), as well as Shiller (1997), summarize a large body of evidence in support of the claim that people systematically violate the laws of probability.

the procedures people use to incorporate information), the implications of the theory are frequently at odds with observed behavior.

This paper develops a stylized, more "structural", model of how people reason when faced with uncertainty. The model is meant to be descriptive rather than normative. Examples of real-world situations to which it is meant to be applied include: a juror deciding on a verdict: a manager considering whether to undertake a merger with another firm: a doctor diagnosing a patient; a teacher assessing a student; and a central banker trying to form a view of the economy's current performance.

Common to these examples, objective probabilities are usually unavailable and there can be a large amount of implication-rich evidence to contemplate. In addition, we often speak of constructing "stories" to facilitate learning in such situations. A doctor may form a "story" of the cause of her patient's symptoms. A central banker may have a "story" about the potential sources of inflationary pressure over some time horizon. Before the model is presented, I discuss the psychological research which motivates it.

In a series of experiments, Pennington and Hastie (1986, 1988, 1990, 1992 and 1993) find that jurors in a trial do not form complicated probability distributions over states of the world—descriptions of the defendant's guilt or innocence, and all the evidence that could be presented. Nor do they incorporate information by calculating posteriors. Instead, they construct "cognitive representations of the evidence in the form of stories"—scenarios describing "what happened" during the events in question. Stories facilitate evidence comprehension and are constructed by jurors

even though evidence is often presented out of temporal and causal order. Moreover, Pennington and Hastie find that jurors base their verdicts on the story representation of information, rather than on the raw evidence.

Of course, a juror's story about a case is very different in content to a doctor's story about her patient, or to a teacher's story about her student's ability. Nevertheless, there are common features to stories. The theoretical definition below will capture two ideas. First, a story is one possible scenario, or state of the world. In the juror example, it is one possible theory of what happened. Second, it is not just any scenario, but must be consistent (at least from the storyteller's perspective), both internally, and with the evidence. The process of reasoning will place restrictions on which scenarios are viewed as coherent.[4]

Outline and Summary

In the next section, I present a theory of how people construct stories. The foundation of the model is a set of statements that an agent has in her mind. In the trial, statements may include: "the defendant is guilty". In order to facilitate reasoning, the agent assigns *attitudes* toward statements: each statement is either accepted or not accepted.

A signal provides, or alters, the attitudes toward some statements. For example,

---

[4]The premise that people like to hold consistent views of the world is supported by research in psychology. A very influential approach to explaining this human tendency was developed by Festinger (1957), who proposed that any perceived inconsistency among various aspects of beliefs, emotions, memory, and behavior, causes an unpleasant state that he termed *cognitive dissonance*, which people try to reduce whenever possible.

a convincing account from a witness may lead our juror to accept that "the defendant has an alibi".

Statements are related to one another through *reasons* which I model as links in two directed networks. The idea is that the attitude toward a statement can provide some justification for accepting, or not accepting, another statement. One network provides reasons from accepted statements: the other from non-accepted statements.

Because of reasons, the initial set of attitudes will typically be modified by a sequence of inferences. Continuing with the example, the juror's acceptance that "the defendant has an alibi" may lead her to accept that "the accused did not commit the crime himself". This may in turn give her some reason to revise her attitude toward the sentence "there is an accomplice to the crime", and toward the statement "the defendant is innocent", and so on.

To capture inferences, a function which maps from patterns of attitudes (over the set of statements) to patterns of attitudes is defined. Iterations of this function are interpreted as *reasoning* on the part of the agent. In this framework, a *story* is a coherent pattern of attitudes, with the attitude toward each statement being supported and built on the attitudes toward other statements. We therefore define stories to be fixed points in the reasoning function.

In section 3, some theoretical properties of the model are presented. The model is rich enough to capture the confusion that can be associated with difficult choices. This occurs when the agent's inferences lead her to cycle endlessly between patterns of attitudes, unable to come to a coherent view of the evidence. Surprisingly, if the

agent satisfies a very weak form of rationality—she recognizes that different ways to state the same reasoning are equivalent—then her inferences will always converge to a story. An upper bound is placed on how long this can take.

Section 4 compares the processing of information in the Story model to the Bayesian calculus. I argue that there are elements of both models in how people learn.

In sections 5, I discuss how an agent who already possesses a story incorporates new evidence. Formally, I show how multiple signals are accommodated. With this extension, the theory is able to explain phenomena observed in the psychology and behavioral economics literatures. The order of signals can affect the story that is formed, and a discredited piece of evidence can leave lingering effects—signals are in general irreversible. Incredibly, "fresh thinkers"—agents who have received less information—can have an advantage in learning the truth. These properties have important consequences for many economic applications.

In the model, it is also possible for an apparently "weak" signal to lead to large changes in the agent's story.[5] This occurs when the signal triggers a long sequence of reasoning; informally, stories can "collapse". A prediction of the model, therefore, is that advocates should use evidence of this type to argue against the stories presented by their opponents.[6] In addition, this possibility of a sudden and dramatic change in an agent's story captures some aspects of the phenomena of epiphany and

---

[5]We will be more precise about what "weakness" means.

[6]Recall the "if the glove doesn't fit, you must acquit" defense of O.J. Simpson.

overreaction. The latter, in particular, may prove to be very relevant for economics.

Sections 6 discuss a number of possible extensions to the basic model. The paper concludes in section 7.

1.2 The Model

1.2.1 Statements and Attitudes

Assume that there exists a set of propositions or *statements*, $\mathcal{P}$.[7] This set is subjective and contains all the sentences which the agent views as relevant for learning about the situation at hand. It is assumed to be finite.[8] During the course of learning and reasoning, the agent assigns an *attitude* toward each statement, which I group into a

---

[7]The mathematics of the model will turn out to be a modification of the Hopfield (1982) network—an example of an artificial neural network—which has been used to solve combinatorial optimization and pattern recognition problems. However, the "neurons" and "synaptic connections" in our model take on very different interpretations compared with these applications. Moreover, the Story model requires two "networks"; restrictions between these two networks will be important for many of the results.

[8]Sentences or statements are also used in models of epistemic states and in propositional logic. To be formal, they are elements in some object language. It is usually assumed that the language contains expressions for the standard sentential connectives, such as negation, conjunction and implication. An infinite number of statements can then be generated from a set of primitive statements using these connectives. Rather than thinking of $\mathcal{P}$ as the set of all possible sentences, it is best thought of as a subjective set of statements to which the agent is paying attention. The finiteness of this set is important for what follows.

vector $\mathbf{a}$ in $\{0,1\}^{\#\mathcal{P}}$. Elements in this vector are indexed by $\mathcal{P}$. $\mathbf{a}(x) = 1$ signifies that statement x is *accepted*; $\mathbf{a}(x) = 0$ refers to a statement which is *not accepted*.[9] To simplify notation, I will also use $\mathbf{a}(x)$ to denote the acceptance of x, and $\neg\mathbf{a}(x)$ to denote non-acceptance.

If a certain statement, x, is not accepted, this does not entail that its negation, $\neg x$, be accepted. If x and $\neg x$ are both in $\mathcal{P}$, then $\mathbf{a}(x) = 0$ and $\mathbf{a}(\neg x) = 0$ can be interpreted as the agent feeling *indetermined* toward statement x. Because indeterminacy can only be represented using both a statement and its negation, we may want to assume that $\mathcal{P}$ is closed under negations. However, this is unimportant for the results which follow.

1.2.2 Modeling Reasons

In this model, the attitude toward one statement can be a *reason* or a justification for the acceptance, or non-acceptance, of another. Reasons can come from knowledge of the physical world, knowledge of human motivations, from logic, as well as from common sense.[10]

Consider two statements, x and y, in isolation; I will discuss the interaction and

---

[9] If the model is viewed as an artificial neural network, then 1 denotes the firing or the activation of the neuron and 0 represents inactivity. The binary assumption reflects the "all-or-none" law in neural biology.

[10] Shafir, Simonson and Tversky (1993) present very convincing experimental evidence which confirm that people seek reasons to justify their decisions, to themselves, and to others. The premise here is that reasons matter, not only to choices, but to attitudes.

aggregation of reasons subsequently. One can reason from statement y to x in four ways. The acceptance of y can be a reason to accept x; or it may inhibit the acceptance of x. Similarly, the non-acceptance of y can provide the agent with a justification to accept, or not to accept, statement x. I model the first two types of inferences from accepted statements with one directed and valued network. Inferences from non-accepted statements are modeled with another. The set of statements, $\mathcal{P}$, are the nodes in these networks. The direction of a link corresponds to the direction of the inference. The magnitude of the value attached to a link represents the strength of the reason. Its sign denotes whether the attitude toward the first statement reinforces or inhibits acceptance of the second.

These networks are represented by two $\#\mathcal{P} \times \#\mathcal{P}$ matrices, $\mathbf{R}$ and $\mathbf{Q}$. Assume, just for now, that all elements in these two matrices are drawn from the binary set $\{-1, 1\}$. I index cells in the matrices by statements in $\mathcal{P}$. For example, $\mathbf{R}(x, y)$ refers to the cell in row x and column y.

Let us begin with a discussion of reasons from accepted statements. In the absence of interaction with other statements, $\mathbf{R}(x, y) = 1$ means that the agent views the acceptance of statement y as a sufficient reason for the acceptance of x. That is, $\mathbf{a}(y) \Rightarrow \mathbf{a}(x)$.[11] Notice that reasoning flows from the column to the row statement. Reasons can also inhibit acceptance. $\mathbf{R}(x, y) = -1$ denotes the inference $\mathbf{a}(y) \Rightarrow \neg\mathbf{a}(x)$.

Implications in the opposite direction do not hold automatically. The fact that

[11] Recall that $\mathbf{a}(x)$ represents $\mathbf{a}(x) = 1$ and $\neg\mathbf{a}(x)$ represents $\mathbf{a}(x) = 0$.

accepting y provides reason for accepting x does not mean that accepting x should provide a reason to accept y. To put this another way, $\mathbf{R}$ need not be symmetric.

Now, consider the implication $\neg\mathbf{a}(y) \Rightarrow \mathbf{a}(x)$. Ignoring effects from other statements, the agent infers from the non-acceptance of y, that x should be accepted. Notice that this is not equivalent to $\mathbf{a}(\neg y) \Rightarrow \mathbf{a}(x)$ because the non-acceptance of y is not the same as the acceptance of its negation. If both $\neg y$ and x are in $\mathcal{P}$, the inference $\mathbf{a}(\neg y) \Rightarrow \mathbf{a}(x)$ can be modeled as $\mathbf{R}(x, \neg y) = 1$, but to model the inference $\neg\mathbf{a}(y) \Rightarrow \mathbf{a}(x)$, the $\mathbf{Q}$ matrix needs to defined.

In the absence of interaction with other statements, $\mathbf{Q}(x, y) = 1$ represents $\neg\mathbf{a}(y) \Rightarrow \mathbf{a}(x)$. A value $\mathbf{Q}(x, y) = -1$ denotes the inference $\neg\mathbf{a}(y) \Rightarrow \neg\mathbf{a}(x)$.[12] One particularly interesting implication which can be captured by the $\mathbf{Q}$ matrix is: $\neg\mathbf{a}(\neg x) \Rightarrow \mathbf{a}(x)$, which corresponds to $\mathbf{Q}(x, \neg x) = 1$. Unless the negation of x is accepted, the agent accepts x. This is essentially saying that x is an *presumption*. If x corresponds to the statement: "defendant is innocent", then a juror with the reason $\mathbf{Q}(x, \neg x) = 1$ is one who presumes that the defendant is innocent unless proven guilty.

In order to allow for *strength* in reasoning, I generalize the above discussion. Elements in $\mathbf{R}$ and $\mathbf{Q}$ now lie in the interval $[-1, 1]$. Agents are permitted to have greater confidence in some inferences than others. For example, blood on the hands of the defendant is a fairly strong reason for believing that he is guilty of murder. The lack of an alibi will also contribute to the belief, but probably to a lesser degree. The

---

[12] It may be obvious at this point that this inference can also be written as $\mathbf{a}(x) \Rightarrow \mathbf{a}(y)$ or $\mathbf{R}(y, x) = 1$. Consistency requirements between $\mathbf{R}$ and $\mathbf{Q}$ will be discussed shortly.

bound of 1 on the absolute value merely represents a normalization. What matters is the magnitude, or strength, of a reason relative to the magnitude of others.[13]

One important validation for this generalization is that it allows a chain of reasoning to have a "weak" link. In arguments that obey the rules of classical logic, the idea of weakness does not make sense. As we are about to see, valued links also allow the agent to trade the magnitude of a reason for a multitude of weak reasons. The idea of a weak link and the idea of magnitude from multitude are present in everyday reasoning.

Aggregation of Reasons

Of course, the strength of reasons only matters if the attitude toward any given statement x is allowed to depend on more than one attitude. I assume that reasons from different statements are additive.[14] Let $\mathbf{R}(x,.)$ and $\mathbf{Q}(x,.)$ denote the xth rows of the reason matrices. The *cumulative reason* for accepting a statement x in $\mathcal{P}$, given the current vector of attitudes $\mathbf{a}$, is:

$$\mathbf{R}(x,.)\mathbf{a} + \mathbf{Q}(x,.)\,(\mathbf{1} - \mathbf{a}) \tag{1.1}$$

where $\mathbf{1}$ denotes a $(\#\mathcal{P} \times 1)$ vector of ones. The assumption of additivity is meant to capture deliberation in the mind of the agent. She contemplates and weighs the

---

[13] Although the precise bound does not matter, for some of the theoretical results which follow, the assumption that elements in these matrices are bounded is important.

[14] If the model were viewed as an artificial neural network, this assumption is consistent with the observation in physical neural networks that signals from different neurons satisfy spacial summation.

reasons for, and against, the acceptance of the statement; the cumulative reason is the result of looking for where the balance lies.[15]

An implicit assumption in the way I have formalized reasons is that they are "independent". Whether the acceptance of y is a good reason to accept x—that is, the value of $\mathbf{R}(x, y)$—does not depend on the attitudes toward other statements. Consider the following, somewhat gruesome, example. The testimony of a witness who claims that "the defendant killed the victim with a gun" (y) is a good reason for thinking that "the defendant is guilty of murder" (x). Similarly, if another witness testifies that "the defendant stabbed the victim to death with a knife" (z), then we have good reason for thinking that the defendant is guilty. However, if both testimonies are accepted, it is likely to cause doubt in the juror's mind. Essentially, we would like our reasons to be able to capture the exclusive-or Boolean function.

This example seems to suggest that additivity is a severe limitation. In fact, because no restrictions are placed on the content of sentences, the assumption is without loss of generality and does allow for interaction between reasons. In the above example, the expression "y and z" is a valid statement for the agent to have in her mind, and its acceptance can be an overriding reason not to accept x.

---

[15]Notice that when we allow for aggregation, values of $\mathbf{R}(x, y)$ and $\mathbf{Q}(x, y)$ in $\{-1, 1\}$ do not have the simple interpretation of logical deductions. For example, a value of $\mathbf{R}(x, y) = 1$ may be "canceled" or offset by a value of $\mathbf{R}(x, z) = -1$. To capture a deductive inference which holds by strict necessity, we need to ensure that the reason outweighs all others.

Contrapositive Requirement and Irreflexivity

Rationality imposes at least two conditions on the agent's reasons: they should contain no contradictions and incorporate all deductive consequences. It is clear that these conditions are not necessarily compatible with human ability. In particular, this is the case for the latter requirement: we often do not see all the consequences of what we accept. Moreover, in order to impose these two conditions, the content of statements must be explicitly specified.

A much weaker rationality criterion—one which does not involve the content of sentences—will play an important role in the results below. Because an inference can always be stated in the contrapositive, certain restrictions are necessary for reason matrices to be sensible. Ignoring effects from other statements,[16] $\mathbf{R}(x, y) = 1$ denotes the implication $\mathbf{a}(y) \Rightarrow \mathbf{a}(x)$. This can be rewritten as $\neg\mathbf{a}(x) \Rightarrow \neg\mathbf{a}(y)$, or $\mathbf{Q}(y, x) = -1$. A much stronger requirement is that this relationship holds even when we do not have implications that hold with strict necessity: $\mathbf{R}(x, y) = r \gtrless 0$ if and only if $\mathbf{Q}(y, x) = -r$. If the acceptance of the sentence "the defendant had a motive" is a reason for accepting that "the defendant is guilty", then not accepting that "the defendant is guilty" should be *as good* a reason for not accepting that "the defendant had a motive". This condition is listed under (C3) below.

Similarly $\mathbf{a}(y) \Rightarrow \neg\mathbf{a}(x)$ is equivalent to $\mathbf{a}(x) \Rightarrow \neg\mathbf{a}(y)$, and $\neg\mathbf{a}(y) \Rightarrow \mathbf{a}(x)$ is equivalent to $\neg\mathbf{a}(x) \Rightarrow \mathbf{a}(y)$. The stronger versions of these requirements correspond

---

[16] What I really mean is: "assuming that $\mathbf{R}(x, y) = 1$ overrides all other reasons". See footnote 15.

to conditions (C1) and (C2) respectively. They entail that $\mathbf{R}$ be symmetric with respect to its negative elements and $\mathbf{Q}$ be symmetric with respect to its positive elements.

Definition (Contrapositive Condition). *An agent's reasons, $\mathbf{R}$ and $\mathbf{Q}$, are said to satisfy the* contrapositive condition *if the following hold for all $x, y \in \mathcal{P}$,*

$$\mathbf{R}(x, y) < 0 \text{ implies } \mathbf{R}(x, y) = \mathbf{R}(y, x) \tag{C1}$$

$$\mathbf{Q}(x, y) > 0 \text{ implies } \mathbf{Q}(x, y) = \mathbf{Q}(y, x) \tag{C2}$$

$$\mathbf{R}(x, y) > 0 \text{ or } \mathbf{Q}(y, x) < 0 \text{ implies } \mathbf{R}(x, y) = -\mathbf{Q}(y, x) \tag{C3}$$

In the real world, there are many situations where this definition is not met because the agent does not realize the equivalence between alternative ways of stating the same implication. The contrapositive requirement should be viewed as a condition that we may not always want to impose.

Another restriction on reasons, namely *irreflexivity*, will also appear in the results that follow.

Definition (Irreflexivity). *An agent's reasons, $\mathbf{R}$ and $\mathbf{Q}$, are said to be* irreflexive *if the following hold for all $x \in \mathcal{P}$,*

$$\mathbf{R}(x, x) = 0 \tag{C4}$$

$$\mathbf{Q}(x, x) = 0 \tag{C5}$$

Equations (C4) and (C5), ensure that there are no self-loops in the two networks. This is so that no baseless and circular arguments can occur. The acceptance of a sentence x cannot in itself be a reason for accepting x. Nor can the non-acceptance of x be a reason to accept x.

1.2.3 States, Signals and Reasoning

Having described the agent's reasons, I now describe how they are used in her internal deliberations. The vector **a** assigns to each statement a subjective attitude. There exists a vector $\boldsymbol{\omega}$ in $\{0,1\}^{\#\mathcal{P}}$ that determines whether each statement is objectively true (1) or false (0). The task of learning involves deducing this pattern of "0"s and "1"s over the set $\mathcal{P}$. Think of $\boldsymbol{\omega}$ as the binary representation of the *state of the world.*[17]

Given the true state $\boldsymbol{\omega}$, a *signal*—denoted by $\mathcal{S}$—is a subset of the statements in $\mathcal{P}$. The interpretation is that the truth values of sentences in $\mathcal{S}$ are revealed to the agent. Typically, each signal, or each piece of evidence, will be informative with respect to only one statement. In the criminal trial, the testimony of a forensic detective may, for example, reveal the truth of the statement: "the defendant's prints

[17]Because $\boldsymbol{\omega}$ refers to objective truth, it should not contain any contradictions. For example, this would require the following condition on statements and their negation: for any statement x in $\mathcal{P}$, $\boldsymbol{\omega}(x) = 1$ if and only if $\boldsymbol{\omega}(\neg x) = 0$. Further conditions on $\boldsymbol{\omega}$ would entail knowing more about the content of the statements. Although intuitive, assumptions on $\boldsymbol{\omega}$ do not have implications for what follows.

were found on the murder weapon".

Also, there will usually be statements whose truth values are never made known to the agent. Since the juror was not present when the crime took place, she will never observe a direct signal regarding: "the defendant is guilty". She must infer her attitude toward this statement from her other attitudes.

The remainder of this section describes how the agent processes *one* signal, $\mathcal{S}$. I will consider the processing of multiple signals in section 5. A time subscript t is introduced in order to model explicitly the inferential process which occurs in the mind of the agent. For example, $\mathbf{a}_t$ are the attitudes after the signal has been processed for t *periods*. I will interpret the sequence of vectors $\{\mathbf{a}_t\}_t$ as the outcome of reasoning.

One more piece of notation is needed. For any subset of statements $\mathcal{E}$ in $\mathcal{P}$, and any vector $\mathbf{u}$ in $\mathbb{R}^{\#\mathcal{P}}$, let $\mathbf{u}^{\mathcal{E}}$ in $\mathbb{R}^{\#\mathcal{E}}$ denote the projection of $\mathbf{u}$ onto the coordinates representing statements in $\mathcal{E}$. For example, $\mathbf{a}^{\mathcal{S}}$ is a vector in $\{0,1\}^{\#\mathcal{S}}$ that contains attitudes toward sentences in the set $\mathcal{S}$.

Upon observing the signal $\mathcal{S}$, the agent provisionally assigns attitudes. The only requirement is that she assigns the correct attitudes toward statements whose truth values have been revealed.

$$\mathbf{a}_0^{\mathcal{S}} = \boldsymbol{\omega}^{\mathcal{S}} \tag{1.2}$$

No requirement on the remaining set of attitudes, $\mathbf{a}_0^{\mathcal{P}-\mathcal{S}}$, is necessary for the results below.

The vector $\mathbf{a}_0$ is not what the agent concludes from the signal; it is merely an

initial view of the world from which she makes inferences. Each period t, the attitudes toward some subset of the sentences in $(\mathcal{P} - \mathcal{S})$ are revised. Naturally, the attitudes toward statements whose truth values have been revealed—that is, attitudes toward statements in $\mathcal{S}$—are left unchanged. Consider a statement x from the subset of the sentences whose attitudes will be revised. The attitude assigned to x depends on the cumulative reason toward x, which in turn depends on the current vector of attitudes. If a preponderance of reasons suggests that the statement is true, then it remains accepted, or is revised to be accepted:

$$\mathbf{a}_{t+1}^{\mathcal{P}-\mathcal{S}}(x) = \begin{cases} 1 & \text{if } \mathbf{R}(x..)\,\mathbf{a}_t + \mathbf{Q}(x..)\,(\mathbf{1} - \mathbf{a}_t) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1.3}$$

The same procedure is applied to all elements in the subset of sentences whose attitudes are to be revised. The resulting vector of attitudes is used to deduce yet another vector, and so on.

A couple of observations. First, the assumption that attitudes are binary—and the associated assumption of a threshold criterion on the cumulative reason (equation 1.3)—may appear to be extreme. Acceptance, however, does not correspond to knowledge, or even to belief with probability one. An attitude is merely a "guess" about whether the statement is true or false. Because the strength of the cumulative justification can differ across statements with the same attitude, the agent can feel more confident, in her (non-)acceptance of one statement than in her (non-)acceptance of another. In short, degrees of belief are possible even though attitudes are binary.

Obviously the precise subset of statements selected to be updated at each period will affect the time-path of attitudes. I distinguish between updating attitudes in

sequential mode and in *parallel mode.*[18]

Definition (Modes of Reasoning). *If in every period the attitude toward one, and only one, statement in the set* $(\mathcal{P}-\mathcal{S})$ *is updated, then the agent is said to be undertaking* sequential reasoning. *If in every period the attitudes toward all sentences in* $(\mathcal{P}-\mathcal{S})$ *are updated, then the agent is said to be undertaking* parallel reasoning.

When reasoning is in parallel, we can rewrite equation (1.3) in matrix notation:

$$\mathbf{a}_{t+1}^{\mathcal{P}-\mathcal{S}} = \left(\text{sgn}\left[\mathbf{R}\mathbf{a}_t + \mathbf{Q}\left(\mathbf{1}-\mathbf{a}_t\right)\right]\right)^{\mathcal{P}-\mathcal{S}} \tag{1.4}$$

where for a scalar u, $\text{sgn}(u) = 1$ if $u > 0$, and $\text{sgn}(u) = 0$ if $u \leqslant 0$; for vectors, sgn operates element-by-element. Whereas there is only one way of making inferences in parallel, there are many different ways of carrying out sequential reasoning depending on the precise order in which sentences are considered. The *cyclic mode* of serial reasoning plays a role in one of the results that follow.

Definition (Cyclic Mode of Inference). *If the agent is reasoning sequentially, and if each of the sentences in* $(\mathcal{P}-\mathcal{S})$ *is revised in every* $\#(\mathcal{P}-\mathcal{S})$ *periods, then the agent is said to be undertaking* cyclic reasoning. *The order of statements is:* $x_{11}, x_{12}, \ldots, x_{1,\#(\mathcal{P}-\mathcal{S})}, x_{21}, x_{22}, \ldots, x_{2,\#(\mathcal{P}-\mathcal{S})}, \ldots$, *where* $\left\{x_{i1}, x_{i2}, \ldots, x_{i,\#(\mathcal{P}-\mathcal{S})}\right\} = (\mathcal{P}-\mathcal{S})$ *for every cycle* i.

---

[18] These definitions are borrowed from the artificial neural networks literature.

Notice that the order through statements in $(\mathcal{P} - \mathcal{S})$ may differ with each cycle; all the definition requires is that each attitude be updated once in each cycle. Before the dynamics are studied, I define a story.

1.2.4 Stories

In everyday language, when we say that a story "makes sense", we mean that it is consistent with the evidence and that its components are compatible with each other. The following definition provides a way of capturing this in our formal model.

Definition (Story). *Given a collection of reasons* $\mathbf{R}$ *and* $\mathbf{Q}$, *a state of the world* $\boldsymbol{\omega}$, *and a signal* $\mathcal{S}$, *a vector of attitudes* $\mathbf{a}$ *constitutes a* story *if the following conditions hold:*

$$\mathbf{a}^{\mathcal{S}} = \boldsymbol{\omega}^{\mathcal{S}} \tag{1.5}$$

$$\mathbf{a}^{\mathcal{P}-\mathcal{S}} = \left(\operatorname{sgn}\left[\mathbf{Ra} + \mathbf{Q}\left(\mathbf{1} - \mathbf{a}\right)\right]\right)^{\mathcal{P}-\mathcal{S}} \tag{1.6}$$

How does this definition correspond to the real-world notion of a story? First, notice that a story is one possible scenario. It is not a subjective prior over the state space. An agent who thinks according to the Story model recognizes that only one state can correspond to reality and constructs a pattern of attitudes which is her best "guess" at what this state may be.[19]

---

[19] This is not to say that the agent does not have an opinion about the likelihood of her story. We discuss the idea of confidence in section 4.

Second, a story is coherent. It is consistent with evidence because equation (1.5) provides the attitudes in $\mathbf{a}^{\mathcal{S}}$. It is internally consistent because the vector $\mathbf{a}^{\mathcal{P}-\mathcal{S}}$ satisfies the fixed point condition of (1.6). At a story, if the agent was asked why she held a particular attitude toward some statement, she would be able to justify it with the attitudes toward some other statements. These other attitudes can be justified by yet other attitudes, and so on. At no point would she feel the need to revise her pattern of acceptances. The sequence of justifications ends when the attitude toward a statement is given by the external signal. From the agent's perspective, a story is a *justifiable state of the world.*[20]

---

[20]To further convince the reader that this definition captures some aspect of human reasoning, consider an observation made by Marvin Minsky (1985): we often speak of beliefs in terms of structural or architectural expressions, as if they were buildings:

"Your beliefs have no *foundation.*"
"You must *support* that with more evidence."
"That argument is *shaky.* It will *collapse.*"
"Your story cannot *stand up* to scrutiny."

The definition of a story as a pattern of attitudes, where the attitude assigned to each statement is supported and built on the attitudes toward other statements, is consistent with the observation. Also, as we will investigate further in section 6, like buildings, stories can "fall apart" if the attitude assigned to one of the statements is unsound. Of course, there are other plausible explanations for Minsky's observation, and we also speak of beliefs in other metaphors.

1.3 Properties of the Model

1.3.1 Confusion

The first observation to make is that confusion is possible. Recall that we are considering the processing of *one* signal. This signal will typically trigger a chain of inferences. Convergence to a story is the reaching of a conclusion. Non-convergence is a situation of conflict. The agent is unable to reconcile the evidence with her knowledge of reasons.

From a Bayesian perspective, it is difficult to explain why a hypothetical juror may ever have trouble reaching a verdict. At the end of the trial, she merely has to select the partition of the state space—"guilty" or "not guilty"—with the higher probability. Value comparisons are easy.[21] However, experience tells us that confusion often occurs in learning. One can imagine the juror throwing up her hands and exclaiming that she cannot make sense of the evidence.

What conditions on the agent's reasons, $\mathbf{R}$ and $\mathbf{Q}$, guarantee convergence? Notice from the definition of a story that attitudes will not alter for any mode of inference once we are at a story. Whether convergence occurs, however, does depend on the mode of inference. It turns out that, if reasoning is sequential, sufficient conditions for convergence are the contrapositive requirement (equations C1, C2 and C3) and

---

[21] To incorporate the burden of proof required for a guilty verdict, the threshold probability required for the juror to vote for a guilty verdict may be increased to above $\frac{1}{2}$. But the point remains. Whether the posterior assigned to the "guilty" event is above this threshold is easy to answer for a probabilistic thinker.

irreflexivity (equations C4 and C5). I state this claim as theorem 1 below.[22]

Although it is certainly true that people are often confused by alternative ways of stating the same inference, the contrapositive condition is a weak form of rationality compared with that implied by formal logic. It is an assumption made on the architecture of the reason networks without any reference to the content of sentences. In contrast, the rules of logic dictate the relationship between truth values assigned to sentences derived using propositional connectives. For example, the truth or falsity of a statement $z =$ "x or y" is determined by the truth values for x and y.

Theorem 1 (Convergence with Serial Reasoning). *Assume that the reason matrices* **R** *and* **Q** *satisfy the contrapositive requirement and irreflexivity. Then, when reasoning is sequential, and the attitude of each statement is updated sufficiently often, the agent is able to construct a story for any signal* $\mathcal{S}$.

Proof. See the appendix. ■

In the theorem, there is a requirement that the attitude toward each statement be updated sufficiently often. This is necessary to rule out certain sequences of revisions. For example, if the agent considers the same statement in each period, then the set

[22]There are other sufficient conditions for convergence when inferences are made sequentially. For example, the contrapositive requirement can be replaced with a requirement that all reasons be bidirectional: **R** and **Q** are symmetric. Clearly this is a restrictive assumption with no normative appeal. The fact that it implies convergence is of little interest.

of attitudes will not change after the first iteration, but the attitudes will not in general constitute a story. Cyclic reasoning satisfies the requirement. Note that the assumptions in the theorem are not necessary for convergence.

Of course, conditions which establish convergence to a story also establish the existence of a story.

Remark (Existence of a Story). *If the reason matrices* $\mathbf{R}$ *and* $\mathbf{Q}$ *satisfy the contrapositive requirement and irreflexivity, then a story exists for any signal* $\mathcal{S}$.

Looking for a coherent pattern of attitudes among all possible patterns is a difficult task. Even for a situation involving relatively few sentences, the number of combinations that need to be considered $\left(2^{\#(\mathcal{P}-\mathcal{S})}\right)$ is large. Further complication arises if the agent has a complex network of reasons. Making inferences sequentially is an easy procedure to implement, although convergence is not obvious. The eventual effects of a change in the attitude of one statement on distant parts of the network are difficult to foresee. Somewhat surprisingly, theorem 1 shows that an agent who carries out sequential reasoning can only be confused if her reasons contain some inconsistency.

From introspection, it would seem that conscious thought is sequential. Attentional limitations prevent us from considering multiple inferences at once. Therefore, the rest of the paper will focus on sequential inferences. Nevertheless, for the sake of completeness, I present the following theorem:

Theorem 2 (Convergence with Parallel Reasoning). *If the reason matrices* $\mathbf{R}$ *and* $\mathbf{Q}$ *satisfy the contrapositive condition as well as irreflexivity, and if reasoning is in parallel, then the set of attitudes always converges to a cycle of length at most two for any signal* $\mathcal{S}$.

Proof. See the appendix. ■

This theorem implies that parallel inferences may never lead to a story. Interestingly, however, this conflict in the mind of the agent must be of a relatively simple form: she can only cycle between two sets of attitudes.

1.3.2 Multiplicity

At this point, it is useful to consider a simple example which illustrates the concepts that have been introduced. The example will also introduce an important feature of the model: the possibility of multiple stories.

Example

Assume that an agent considers two statements, $\mathcal{P} = \{x, y\}$, and possesses the following reasons:

$$\mathbf{R} = \begin{array}{c} \\ x \\ y \end{array} \begin{array}{c} \begin{array}{cc} x & y \end{array} \\ \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \end{array} \qquad \mathbf{Q} = \begin{array}{c} \\ x \\ y \end{array} \begin{array}{c} \begin{array}{cc} x & y \end{array} \\ \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} \end{array} \tag{1.7}$$

Ignore the signal and assume that the initial set of attitudes is:

$$\mathbf{a}_0 = \begin{matrix} x \\ y \end{matrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{1.8}$$

Notice that the reason matrices satisfy equations (C1) to (C5).[23] Convergence to a story is assured with sequential revision. Indeed, in this example, a story is reached after just one inference. Say that the agent reasons by first updating the attitude toward statement x.

In period 1, the attitude toward x is given by a preponderance of reasons:

$$\begin{aligned} \mathbf{a}_1(x) &= \text{sgn}\left[\mathbf{R}(x,.)\,\mathbf{a}_0 + \mathbf{Q}(x,.)\,(\mathbf{1} - \mathbf{a}_0)\right] \\ &= \text{sgn}\left(\begin{bmatrix} 0 & 1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 & -1 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) \\ &= 1 \end{aligned} \tag{1.9}$$

Thus, $\mathbf{a}_1 = [\ 1 \quad 1\ ]'$. This is a story; subsequent reasoning results in the same pair of attitudes. Had statement y been the first statement updated, the opposite conclusion, $\mathbf{a}_1 = [\ 0 \quad 0\ ]'$, would have been reached. This is hardly surprising. In this example, the acceptance of x and of y are mutually supporting. In the absence of evidence, it is clear that there are two stories, both equally coherent and compelling. Which one is reached depends on which statement is first updated to be consistent with the attitude toward the other statement.

---

[23] In fact, $\mathbf{R}$ and $\mathbf{Q}$ are symmetric. This is of course not necessary for the contrapositive condition.

Multiplicity and the Order of Inferences

This example is trivial but one can imagine more complicated networks where an agent reasons to a particular story without being aware of the existence of other stories. The possibility of multiple, coherent interpretations of the same evidence is important to many features of the Story model. In contrast, Bayes's rule ensures that there is a unique way for a probabilistic agent to incorporate a signal into her beliefs.

When there are multiple stories, and the agent reasons sequentially, the order in which attitudes are revised can affect which story is reached. Agents who receive the same signal, and who have the same reasons, can still form different stories if they undertake different chains of inferences.

Novelists and scriptwriters frequently take advantage of multiple stories to generate surprise. For example, throughout the movie, the scriptwriter may steer the viewer toward one particular story. Only in the final act is it revealed that an alternative scenario is in fact the truth. This is only possible because the evidence presented prior to the final act is consistent with multiple coherent interpretations. Interestingly, when the truth is presented, we are often not totally convinced and think back to earlier parts of the movie to verify that the new story is in fact consistent with what we have seen. This highlights the importance of coherence in our thinking.

As we will see when the model is extended to consider multiple signals, people are susceptible to being led toward a particular story partly because they sometimes

interpret new evidence to "fit" with their current story.

Back to the Example

Before leaving the example, I use it to illustrate a couple more points. Theorem 2 states that making inferences in parallel leads to a cycle of length at most two periods. We can verify this for the initial set of attitudes given above:

$$\begin{aligned} \mathbf{a}_1 &= \text{sgn}\left(\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right) \\ &= \begin{bmatrix} 1 \\ 0 \end{bmatrix} \end{aligned} \tag{1.10}$$

Reasoning from this new vector of attitudes leads to $\mathbf{a}_2 = [\ 0\ \ 1\ ]' = \mathbf{a}_0$, and so on. The agent cycles between the two scenarios: $[\ 1\ \ 0\ ]'$ and $[\ 0\ \ 1\ ]'$.

Finally, I alter the example to illustrate the consequences of violating the contrapositive condition. This will clarify its role in the convergence theorems. Replace the above reason matrices by:

$$\mathbf{R} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \qquad \mathbf{Q} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \tag{1.11}$$

These justifications violate equations (C1) to (C3). It can be shown that no story can be constructed from any initial set of attitudes, in either sequential or parallel mode. In fact, no state of the world is coherent. This is obvious if we rewrite the reasons as the following contradictory chain of inferences:

$$\mathbf{a}(y) \Rightarrow \mathbf{a}(x) \Rightarrow \neg\mathbf{a}(y) \Rightarrow \neg\mathbf{a}(x) \Rightarrow \mathbf{a}(y) \Rightarrow \ldots \tag{1.12}$$

The agent cycles endlessly without coming to a conclusion.

The next obvious issue to consider is how quickly the agent is able to form a story after the presentation of a signal.

1.3.3 Time Required for Convergence

Definition (Time to Convergence). *Upon the receipt of a signal $\mathcal{S}$, the* time to convergence, *T, is defined as:*

$$T = \min \{t \mid \mathbf{a}_{i+1} = \mathbf{a}_i = \mathbf{a} \ \textit{for all} \ \ i \geqslant t\} \tag{1.13}$$

where **a** *is a story with respect to the signal $\mathcal{S}$.*

Time to convergence[24] can be viewed as a measure of the agent's perception of the difficulty of the situation. It is the number of inferences that she has to make before being able to construct a story. In general, this will depend on: the mode of reasoning; the agent's reasons; and the signal.

Consider a worst-case, upper bound for this time. No upper bound exists in parallel mode because convergence may not occur. Even with sequential reasoning, no upper bound exists without further restrictions on the precise sequence of inferences. The agent might, for example, contemplate the same sentence for some arbitrarily large number of periods. We therefore derive the upper bound for cyclic mode.

---

[24] In discrete neural computation, the number of iterations required before the network reaches an equilibrium is called the *transient period.*

Proposition 3 (Time to convergence). *Assume that the contrapositive condition and irreflexivity are met, and that reasoning is in cyclic mode. Then the agent converges to a story with a time to convergence bounded above by:*

$$T \leqslant \left[\frac{\#(\mathcal{P}-\mathcal{S})}{\delta}\right]\left(\sum_{x\in\mathcal{P}-\mathcal{S}}\left|\sum_{y\in\mathcal{S}}\left[\mathbf{R}(x,y)-\mathbf{Q}(x,y)\right]\mathbf{a}(y)+\mathbf{Q}(x,.)\ \mathbf{1}\right|+\frac{1}{2}\sum_{x\in\mathcal{P}-\mathcal{S}}\sum_{y\in\mathcal{P}-\mathcal{S}}\left|\mathbf{R}(x,y)-\mathbf{Q}(x,y)\right|\right) \tag{1.14}$$

where

$$\delta = \min_{(x,\mathbf{a}^{\mathcal{P}-\mathcal{S}})\in(\mathcal{P}-\mathcal{S})\times\{0,1\}^{\#(\mathcal{P}-\mathcal{S})}}\left|\mathbf{R}(x,.)\ \mathbf{a}+\mathbf{Q}(x,.)\left(\mathbf{1}-\mathbf{a}\right)\right| \tag{1.15}$$

subject to

$$\left|\mathbf{R}(x,.)\ \mathbf{a}+\mathbf{Q}(x,.)\left(\mathbf{1}-\mathbf{a}\right)\right| > 0 \tag{1.16}$$

$$\mathbf{a}^{\mathcal{S}} = \boldsymbol{\omega}^{\mathcal{S}} \tag{1.17}$$

Proof. See the appendix. ■

Notice that the bound is increasing in the number of statements which the agent has to consider, $\#(\mathcal{P}-\mathcal{S})$. A learning situation that requires a greater number of statements, or in which fewer truth values are provided by the signal, is likely to be more computationally intensive. The upper bound is negatively related to δ, which can be interpreted as the weakest possible inference. A small value of δ means that the story constructed may be very fragile.

If $|\mathbf{R}(x,y) - \mathbf{Q}(x,y)|$ is small, then the bound on the transient period is tighter. Consider what it means if $\mathbf{R}(x,y) = \mathbf{Q}(x,y)$. Abstract from other statements and say this common value is 1. Then, the agent's reasons include: $\mathbf{a}(y) \Rightarrow \mathbf{a}(x)$ and $\neg\mathbf{a}(y) \Rightarrow \mathbf{a}(x)$. But this is simply saying that the sentence x should be accepted whatever the attitude toward y. If inferences are all of this spurious form, then it cannot take very long for the agent to reach a coherent set of attitudes.

Because the order in which attitudes are revised can differ across agents, even agents with identical reasons, who receive the same signal, and who reach the same story, can have different convergence times. That is, they can have dissimilar opinions about the difficulty of a problem. We have all been in situations where we took a roundabout way to reach a conclusion when a more direct route was available.

1.3.4 Local Uniqueness of Stories

It turns out that states of the world which are local to a story—that is, which differ from a story only in the attitude toward one statement—are not stories.

Proposition 4 (Local Uniqueness). *Assume that the reason matrices* $\mathbf{R}$ *and* $\mathbf{Q}$ *satisfy irreflexivity. Then for any signal* $\mathcal{S}$*, two vectors of attitudes which are both stories must differ in the attitude toward at least two statements.*

Proof. See the appendix. ■

This result is not surprising from the perspective of a reason-based model. The

interpretation is as follows. Take two agents with the same set of reasons who have seen the same evidence, and who have each constructed a story. If these stories are not identical, they must differ in the attitude they assign to at least two statements. If this were not the case, then the set of undisputed attitudes would be supporting both possible attitudes toward the sentence in dispute. One of the agents must not be justified in her attitude toward this sentence. Put simply, if a trial attorney wants to change a juror's attitude toward, say, "the defendant is guilty", he must provide the juror with a reason to do so.

When irreflexivity is not satisfied, the attitude toward a statement can be a reason for itself. If a difference in opinion can justify itself, then obviously two stories can differ in only one attitude. This is ruled out by the conditions for the proposition.

As an aside, a similar local-uniqueness result holds for the set of Nash equilibria in game theory (which are also fixed points of an appropriately defined mapping). For a generic normal form game, if we begin at a Nash strategy profile and change the strategy of one of the players, then the new profile cannot be an equilibrium unless some other player also changes her strategy.[25]

---

[25]Consider a function g which maps from a product space $X = \prod_{i=1}^{I} X_i$ to itself. The property which ensures that the set of fixed points is locally unique is the following: for all i, g_i must only be a function of $\prod_{j \neq i} X_j$, where g_i is the ith component of g. Notice that best response mappings in game theory satisfy this. Equations (C4) and (C5) are needed precisely so that this condition is satisfied. This observation was pointed out to me by Stephen Morris.

1.4 The Bayesian Model

In this section, I contrast the model presented above with the Bayesian model. For simplicity, consider an agent who contemplates only two statements, $\mathcal{P} = \{x, y\}$. The true state of the world lies in the following product space:

		y	
		$\omega(y) = 0$	$\omega(y) = 1$
x	$\omega(x) = 0$	(0, 0)	(0, 1)
	$\omega(x) = 1$	(1, 0)	(1, 1)

It is clear that a truth value assigned to a statement--for example, $\omega(x) = 1$—can be viewed as an event, a subset of the state space.

The Bayesian model assumes that all information that is relevant for decision making can be captured by probability measures. In particular, five measures may be involved:

$\pi_x : 2^{\{0,1\}} \to [0, 1]$	Marginal measure over the truth value of x
$\pi_y : 2^{\{0,1\}} \to [0, 1]$	Marginal measure over the truth value of y
$\pi : 2^{\{0,1\}^2} \to [0, 1]$	Joint measure over the state space $\{0, 1\}^2$
$\pi_{y\|\omega(x)} : 2^{\{0,1\}} \to [0, 1]$	Measure over y, conditioned on a truth value for x
$\pi_{x\|\omega(y)} : 2^{\{0,1\}} \to [0, 1]$	Measure over x, conditioned on a truth value for y

Bayesian beliefs can be represented in three equivalent ways: (i) by the joint measure π; (ii) by the marginal π_x and the conditional measures $\{\pi_{y|\omega(x)}\}_{\omega(x)\in\{0,1\}}$; and (iii)

by the distribution π_y and the conditional measures $\left\{\pi_{x|\omega(y)}\right\}_{\omega(y)\in\{0,1\}}$. Bayesian beliefs are "coherent" in the sense that Bayes's rule allows us to move among these representations.

Another way of seeing this point is via the following system of equations. It should be obvious that these two equations have to hold given Bayes's rule. The first equation states that the probability of x being true is equal to the probability that x is true conditioned on y being true, multiplied by the probability that y is true, plus the probability that x is true given that y is false, multiplied by the probability that y is false.

$$\begin{aligned}\begin{bmatrix} \pi_x(1) \\ \pi_y(1) \end{bmatrix} &= \begin{bmatrix} 0 & \pi_{x|\omega(y)=1}(1) \\ \pi_{y|\omega(x)=1}(1) & 0 \end{bmatrix} \begin{bmatrix} \pi_x(1) \\ \pi_y(1) \end{bmatrix} \\ &+ \begin{bmatrix} 0 & \pi_{x|\omega(y)=0}(1) \\ \pi_{y|\omega(x)=0}(1) & 0 \end{bmatrix} \begin{bmatrix} 1-\pi_x(1) \\ 1-\pi_y(1) \end{bmatrix}\end{aligned} \tag{1.18}$$

or

$$\mathbf{p} = \mathbf{M}\mathbf{p} + \mathbf{N}(\mathbf{1} - \mathbf{p}) \tag{1.19}$$

where $\mathbf{p} = [\ \pi_x(1) \quad \pi_y(1)]'$ and the matrices $\mathbf{M}$ and $\mathbf{N}$ have the obvious definitions. For given matrices of conditional probabilities, $\mathbf{M}$ and $\mathbf{N}$, equation (1.19) can be thought of as a fixed-point condition on the marginal measures π_x and π_y. Now, recall that in the absence of a signal, $\mathcal{S} = \emptyset$, the definition of a story, $\mathbf{a}$, is:

$$\mathbf{a} = \operatorname{sgn}\left[\mathbf{R}\mathbf{a} + \mathbf{Q}\left(\mathbf{1} - \mathbf{a}\right)\right] \tag{1.20}$$

The similarity with equation (1.19) is obvious. Moreover, notice that the matrices of conditional probabilities contain "0"s on the diagonals. The is similar to the irreflexivity requirement, (C4) and (C5), for reason matrices.

There are at least three important differences between the two models. First, probabilities allow for degrees of belief; attitudes do not. An agent who learns through stories is concerned with whether states of the world are coherent; a Bayesian cares about whether probabilities assigned to subsets of the state space are coherent. Second, for given generic conditional probabilities, the fixed point in equation (1.19) is unique.[26] In contrast, we have seen that multiple stories are possible. The third difference is that Bayesian learning is not explicitly dynamic, whereas the Story model is. A Bayesian does not have to "find" coherent beliefs. Probabilistic beliefs are coherent—and remain so—as long as changes satisfy Bayes's rule. On the other hand, the Story model is precisely about how people come to coherent views of the world.

Reconciling the Two Models

In many real-world circumstances, there are aspects of both models in how people reason under uncertainty. A juror cares about whether a story constructed by an attorney is consistent, but the story's likelihood is also important. It may appear from the above discussion that the two models are incongruous. One way toward a reconciliation is to view them as operating at different stages of the processing of information.

In a typical trial, there are numerous possible scenarios. Most of these either contain internal inconsistencies, or are inconsistent with the evidence. The procedure described in this paper allows a juror to limit attention to the small subset of states

---

[26]This is because, the matrix $(\mathbf{I} - \mathbf{M} - \mathbf{N})$ is invertible for generic conditional probabilities.

that are coherent. This subset can contain more than one story if the juror undertakes more than one sequence of reasoning, or if she is presented with different coherent theories by the different trial lawyers.

When aware of multiple stories, the juror is required to assess which is most plausible. These confidence assessments may of course take the form of probabilities satisfying Bayes's rule. One can then view the Story model as providing the support for an agent's subjective probability measure.

There is one important qualification. A state that is not a story—and so not in the support of the agent's probabilistic beliefs—may become coherent after a subsequent signal.[27] Bayes's rule, however, does not allow one to update from zero to some strictly positive probability. This is particularly problematic when the set of stories after some signal have no elements in common with the set of stories prior to the signal. A procedural theory of how people make subjective confidence assessments becomes crucial. Ideas embedded in the Story model could provide the basis for such a theory. The model suggests a number of dimensions to confidence. (This is in contrast to the Bayesian model where confidence has only one aspect: probabilistic likelihood.) Loosely, an agent assigns high confidence to a story if it is: unique;[28]

---

[27]The next section describes how multiple signals can be accommodated.

[28]The claim that uniqueness increases confidence finds support in Baltser and Pennington's (1995) experimental studies. Of course, what is important is the perception of uniqueness. This leads to the interesting idea of overconfidence. Overconfidence occurs when the agent does not realize that alternative coherent interpretations exist; she will then attribute too much confidence to her story.

supported by strong reasons;[29] and "simple"[30].

1.5 Incorporating New Evidence

I now consider how stories adjust—or fail to adjust—when additional evidence is introduced. Many of the economically-relevant implications of the model arise when we allow for multiple signals. I will refer to the processing of each signal as a *stage* and index it by a subscript n. $\mathcal{S}_n$ is the nth signal: $\mathbf{a}_{nt}$ is the set of attitudes after the nth signal has been processed for t periods: and the time to convergence for stage n is denoted by T_n.

At the end of stage $(n-1)$, the agent possesses a story $\mathbf{a}_{n-1,T_{n-1}}$. Consider the nth stage. In the initial period of this stage, the agent alters her attitudes to be consistent with the truth values revealed by the new signal. Attitudes toward the other sentences are unchanged. The resulting vector of attitudes is denoted by $\mathbf{a}_{n0}$. We have:

$$\mathbf{a}_{n0}^{\mathcal{S}_n} = \boldsymbol{\omega}^{\mathcal{S}_n} \tag{1.21}$$

$$\mathbf{a}_{n0}^{\mathcal{P}-\mathcal{S}_n} = \mathbf{a}_{n-1,T_{n-1}}^{\mathcal{P}-\mathcal{S}_n} \tag{1.22}$$

A story for stage n is a vector of attitudes that satisfies equations (1.5) and (1.6), with the signal $\mathcal{S}$ replaced by the union of signals $\bigcup_{i=1}^{n}\mathcal{S}_i$. The vector $\mathbf{a}_{n0}$ satisfies

---

[29]Both in an average sense, and also in the sense that the strength of the weakest link in a story is large. The latter makes a story less "fragile". More on this point in section 5.

[30]One aspect of simplicity is the number of sentences that a story incorporates. To be precise about this point, one needs a theory of how agents choose the set of statements.

(1.5). However, it will in general violate the fixed-point condition of (1.6). A new sequence of reasoning ensues and comes to an end when internal consistency is restored. During this time, only statements in $\mathcal{P} - \bigcup_{i=1}^{n} \mathcal{S}_i$ are revised.

Now that the incorporation of new evidence has been described, a number of results suggest themselves. Although this paper does not undertake the important next step of developing detailed applications of these results, I will suggest some applications that appear promising.

The Possibility of "Collapse"

A signal in the Story model is "weak" if its statements are the causes of weak reasons. Even weak signals can have large effects on the agent's story. This occurs when the truth values revealed by the signal trigger a long sequence of inferences to a very different fixed point. Informally, stories can "collapse". This feature of the model has implications for how advocates should try to convince agents who think with stories.

More ambitiously, this feature may provide a starting point for thinking about both underreactions and overreactions in markets. Without being precise, assume that a trader possesses a story about the prospects for the market in question, and that a signal is received. Some attitudes are changed to make them compatible with the signal. If the reasons emanating from these statements are weak relative to other reasons, then the news may have no effect on other parts of the agent's story: the agent appears to underreact. However, a sequence of signals, each being the cause of the same effect, will finally cause one of the other attitudes to change, possibly

triggering a "jump" to a very different story: the trader appears to overreact to the last signal.[31]

Finally, this discontinuity in the response to information captures one aspect of the phenomenon of epiphany.

The Order of Signals Matters

In this model, the order of *signals* can matter.[32,33] The existence of multiple stories is central to this: order matters because it affects which of the fixed points is reached. This occurs because the current story is a starting point from which a new signal is interpreted and processed. (See equation 1.22.) Contrast this with the fact that Bayesian posteriors do not depend on the arrangement of signals over time.

Many applications suggest themselves. The model predicts that a physician's conclusion about her patient's health depends on the order in which the results of medical tests are presented. A trial lawyer's scheduling of testimonies can be crucial to the trial's outcome. A central banker's story about the current state of the economy will depend on the order in which economic statistics are released.

---

[31]This is suggestive of empirical observations about underreactions and overreactions of stock prices. Barberis, Shleifer and Vishny (1998) summarize this literature.

[32]In section 3, we were speaking about the order of *inferences* in sequential reasoning, not the order of signals.

[33]General conditions that determine when early signals matter more than later ones, and when later signals matter more, remain to be found. Rabin and Schrag (1999) consider some of the economic implications of the *confirmatory bias*: the premise that first impressions matter more.

The Advantage of Fresh Thinkers

Closely related to the observation that the order of signals matters is the idea that "fresh thinkers"—agents who have received less information—may have an advantage in learning the truth. In this model, even a weak signal can lead to a story in which a subset of attitudes is "entrenched"—that is, mutually supported by strong reasons. If the truth is in fact different, then the entrenched attitudes put the agent at a disadvantage. It would now require a very strong signal to change the agent's mind.

This is one argument frequently cited for why external consultants to corporations can be very valuable: they bring a fresh perspective to old problems.

Irreversibility of Signals

This result is also related to, but is more specific than, the idea that order matters. Say that the agent receives a signal, $\mathcal{S}_1 = \{x\}$, revealing that $\omega(x) = 1$. This can provide some reason to accept or not accept other statements, which in turn can cause other attitudes to change, and so on. Now imagine that the agent receives a new signal $\mathcal{S}_2 = \{x\}$ with $\omega(x) = 0$; that is, it is revealed that the original signal is incorrect. In this model, this sequence of two signals will typically alter the agent's story. This is because inferences which take place after the first signal may not be undone when the agent subsequently finds out that the signal is wrong.

One interesting consequence of this result is that smear campaigns in politics can be effective, even after the facts of the campaign have been discredited.[34] The

---

[34]Mullainathan (1997) produces a similar result using a model based on imperfect memory. In his model, a discredited signal still "influences beliefs because the memories evoked by the signal

irreversibility of signals may also have economic implications through revisions of macroeconomic statistics, or of company earnings releases.

1.6 Extensions

This section mentions a number of particularly interesting extensions to the basic model.

Learning Reasons

Thus far, I have taken the agent's reasons, **R** and **Q**, to be fixed. Fixing reasons is only valid if they capture persistent knowledge. A more detailed model should endogenize reasons. One idea would be to allow agents to alter their reasons based on the stories that they hold at the end of each stage.

Attention and the Order of Inferences

An assumption of the model is that the agent is aware of all statements which are relevant for reasoning about the problem at hand. What factors determine which statements are in the agent's focus of attention? It would seem natural that a signal regarding a statement draws attention to the statement. Beyond this, if reasoning occurs sequentially, then the precise sequence of inferences should affect which statements are brought into the agent's focus. One obvious chain of inferences is along the path of "strongest reasons": each period, the agent updates the statement

---

continue to be memorable".

for which the cumulative reason has the greatest magnitude.[35]

Concepts from Social Networks

Graph theory has been used extensively in the study of social networks. In that literature, various notions exist for social concepts such as *cliques* and agent *centrality*. Do these concepts have natural interpretations when applied to the reason graphs, **R** and **Q**? Central statements in these networks may correspond to the "crux" of a story. Cliques may correspond to collections of statements that form "episodes", or "substories".

Game Theoretic Refinements

Many refinements of equilibria in extensive-form games are based on placing restrictions on players' beliefs off the equilibrium path. For players with Bayesian beliefs, this is a difficult thing to do because, in equilibrium, "off the equilibrium path" is an event with zero probability: an event which leads to "confusion" for a Bayesian. More importantly, the Bayesian model does not specify how players can return to coherent probabilistic beliefs when such an event occurs. The Story model on the other hand explicitly models how an agent reaches a coherent scenario from a state of confusion. *If* one could model players who construct stories, interesting equilibrium refinements may result.

---

[35]Bringing a new statement into the agent's focus can dramatically change her story, creating a "flash of insight", or epiphany.

1.7 Conclusion

A cognitive model of how agents reason and learn was introduced. The model is most applicable to situations where a large body of implication-rich evidence must be evaluated, and where objective probabilities are unavailable. Many important decisions have to be made in such environments.

The central premise of the paper is that people like to hold coherent scenarios in their minds: they like to form "stories". From this idea, the formal model produces rich predictions. Among these, agents who think with stories can suffer from confusion and can also come to different interpretations of the same evidence. They typically do not undo inferences made from evidence, even after the evidence is discredited. More generally, the order in which evidence is presented can affect their conclusions. Apparently weak evidence can trigger large changes in their stories.

Though only hinted at in this paper, these features of the model have important implications for behavior. The dependence on the order of signals has troubling consequences for a doctor's assessment of a patient, or a central banker's assessment of the economy. The irreversibility of signals encourages smear campaigns in politics, and may lead to lasting effects from the release of economic statistics, even when they are subsequently revised. The discontinuity in the response to signals may account for under and overreactions in financial markets. And finally, though not really discussed in this paper, the model has implications for the tactics that advocates should use.

Compared with other models of learning, the Story model may be less tractable

and more difficult to apply. Despite this, it does have attractive features. Foremost among these, a focus on reasons and stories seems closer to the way we think and talk about learning.

On a broader level, it is hoped that this paper hints at the possibility that incorporating cognitive elements into theories of learning and choice may yield more realistic, interesting, and testable, implications which complement those from more traditional theories. Viewed in this light, the Story model is an example of a growing literature in economics which focuses explicitly on the procedures by which decisions of economic units are made.[36]

---

[36]See Simon (1982) and Rubinstein (1998).

1.8 Appendix

1.8.1 Proof of Theorem 1

I begin by showing that conditions (C1), (C2), and (C3), imply that $\mathbf{R}-\mathbf{Q}$ is symmetric. Define $\mathbf{R}^+(x,y) = \max\{\mathbf{R}(x,y), 0\}$ and $\mathbf{R}^-(x,y) = \min\{\mathbf{R}(x,y), 0\}$; similarly for $\mathbf{Q}^+$ and $\mathbf{Q}^-$. We have:

$$\begin{aligned}
\mathbf{R}(x,y) - \mathbf{Q}(x,y) &= \mathbf{R}^+(x,y) + \mathbf{R}^-(x,y) - \mathbf{Q}^+(x,y) - \mathbf{Q}^-(x,y) \\
&= \mathbf{R}^+(x,y) + \mathbf{R}^-(y,x) - \mathbf{Q}^+(y,x) - \mathbf{Q}^-(x,y) \\
&= -\mathbf{Q}^-(y,x) + \mathbf{R}^-(y,x) - \mathbf{Q}^+(y,x) + \mathbf{R}^+(y,x) \\
&= \mathbf{R}(y,x) - \mathbf{Q}(y,x) \qquad (1.23)
\end{aligned}$$

The second equality follows from (C1) and (C2); the third from (C3).

For a set of statements $\mathcal{E}$ in $\mathcal{P}$, let $\mathbf{A}^{\mathcal{E}}$ denote the submatrix of the $\#\mathcal{P}\times\#\mathcal{P}$ matrix $\mathbf{A}$ spanned by the elements in $\mathcal{E}$. For example, $\mathbf{R}^{\mathcal{P}-\mathcal{S}}$ is a $\#(\mathcal{P}-\mathcal{S}) \times \#(\mathcal{P}-\mathcal{S})$ matrix formed by removing columns and rows in $\mathbf{R}$ corresponding to statements in $\mathcal{S}$.

By construction, equation (1.5) is satisfied. Rewrite equation (1.6) as:

$$\mathbf{b} = \operatorname{sgn}[\mathbf{W}\mathbf{b} + \mathbf{k}] \qquad (1.24)$$

where:

$$\mathbf{b} = \mathbf{a}^{\mathcal{P}-\mathcal{S}} \qquad (1.25)$$

$$\mathbf{W} = \mathbf{R}^{\mathcal{P}-\mathcal{S}} - \mathbf{Q}^{\mathcal{P}-\mathcal{S}} \qquad (1.26)$$

$$\mathbf{k} = \left[\sum_{y\in\mathcal{S}} (\mathbf{R}(.,y)\,\mathbf{a}(y) + \mathbf{Q}(.,y)\,[1-\mathbf{a}(y)])\right]^{\mathcal{P}-\mathcal{S}} + \mathbf{Q}^{\mathcal{P}-\mathcal{S}}\mathbf{1}^{\mathcal{P}-\mathcal{S}} \qquad (1.27)$$

Define for all t, the following function:[37]

$$G(t) = -2\mathbf{b}_t'\mathbf{k} - \mathbf{b}_t'\mathbf{W}\mathbf{b}_t \tag{1.28}$$

I now show that the function $G(t)$ is decreasing. Because reasoning is in sequential mode, only one statement in $\mathcal{P} - \mathcal{S}$ will be updated in any given period, say x. It is easy to show that:

$$\begin{aligned}
\Delta G(t+1) &= G(t+1) - G(t) \\
&= -2\Delta\mathbf{b}_{t+1}(x)\mathbf{k}(x) - \mathbf{W}(x,x)\left[\Delta\mathbf{b}_{t+1}(x)\right]^2 \\
&\quad -\Delta\mathbf{b}_{t+1}(x)\left[\mathbf{b}_t'\mathbf{W}(.,x) + \mathbf{W}(x,.)\mathbf{b}_t\right]
\end{aligned} \tag{1.29}$$

From (1.23) above, we know that $\mathbf{W}(.,x) = \mathbf{W}(x,.)'$. Together with consistency requirements (C4) and (C5), we have:

$$\Delta G(t+1) = -2\Delta\mathbf{b}_{t+1}(x)\left[\mathbf{W}(x,.)\ \mathbf{b}_t + \mathbf{k}(x)\right] \tag{1.30}$$

To determine the sign of this expression, notice from equation (1.3) that:

$$\Delta\mathbf{b}_{t+1}(x) = \begin{cases} 1 & \text{if } \mathbf{b}_t(x) = 0 \text{ and } \operatorname{sgn}\left[\mathbf{W}(x,.)\ \mathbf{b}_t + \mathbf{k}(x)\right] = 1 \\ 0 & \text{if } \mathbf{b}_t(x) = \operatorname{sgn}\left[\mathbf{W}(x,.)\ \mathbf{b}_t + \mathbf{k}(x)\right] \\ -1 & \text{if } \mathbf{b}_t(x) = 1 \text{ and } \operatorname{sgn}\left[\mathbf{W}(x,.)\ \mathbf{b}_t + \mathbf{k}(x)\right] = 0 \end{cases} \tag{1.31}$$

Thus $\Delta\mathbf{b}_{t+1}(x)\left[\mathbf{W}(x,.)\ \mathbf{b}_t + \mathbf{k}(x)\right] \geqslant 0$, which implies that $\Delta G(t+1) \leqslant 0$ for all t.

Next I note that the function G is bounded from below. In particular,

$$\min G(t) \geqslant -2\sum_{x\in\mathcal{P}-\mathcal{S}} \mathbf{k}^+(x) - \sum_{x\in\mathcal{P}-\mathcal{S}}\sum_{y\in\mathcal{P}-\mathcal{S}} \mathbf{W}^+(x,y) > -\infty \tag{1.32}$$

---

[37]This is known as an *energy function* in the artificial neural networks literature. The use of energy functions to prove convergence in artificial neural networks is described in Hopfield (1982) and Goles, Fogelman and Pellegrin (1985).

where $\mathbf{k}^+(x) = \max\{0, \mathbf{k}(x)\}$ and $\mathbf{W}^+(x, y) = \max\{0, \mathbf{W}(x, y)\}$.

Define the following minimum value:

$$\delta = \min_{(x,\mathbf{b}) \in (\mathcal{P}-\mathcal{S}) \times \{0,1\}^{\#(\mathcal{P}-\mathcal{S})}} |\mathbf{W}(x, .)\ \mathbf{b} + \mathbf{k}(x)| \tag{1.33}$$

$$\text{subject to } |\mathbf{W}(x, .)\ \mathbf{b} + \mathbf{k}(x)| > 0 \tag{1.34}$$

From equations (1.30) and (1.31), we see that $\Delta G(t+1) < 0$ implies that $\Delta G(t+1) \leqslant -2\delta < 0$. Because $G(t)$ is bounded from below, after a finite number of inferences, $G(t)$ must converge to some value.

It remains to be shown that $\Delta G(t) = 0$ for all $t > T$ corresponds to a coherent set of attitudes. Recall that each attitude is revised infinitely often. Assume that $\Delta G(t) = 0$ for all $t > T$ but that we are not at a story. From equation (1.30), whenever $\Delta \mathbf{b}_{t+1}(x) \neq 0$, we must have $\mathbf{W}(x, .)\ \mathbf{b}_t + \mathbf{k}(x) = 0$, which in turn implies that $\Delta \mathbf{b}_{t+1}(x) = -1$. Because the number of statements in $(\mathcal{P} - \mathcal{S})$ is finite, we cannot have an infinite sequence of such revisions. A final remark. Notice that the consistency assumption of equations (C4) and (C5) are stronger than necessary for this proof. $\mathbf{R}(x, x) \geqslant 0$ and $\mathbf{Q}(x, x) \leqslant 0$ would have sufficed. ■

1.8.2 Proof of Theorem 2

To show that consistency implies convergence in parallel mode, I use theorem 1 together with a general result in Bruck and Goodman (1988) which enables transformation of a neural network in parallel mode to a equivalent network in sequential mode. What follows is a rephrasing of part of their argument for the transformation

of such networks. Construct from the original reason matrices the following mapping involving statements in a set $\widetilde{\mathcal{P}}$. This set contains $2\#(\mathcal{P}-\mathcal{S})$ statements. For all x in $\widetilde{\mathcal{P}}$.

$$\tilde{\mathbf{b}}_{t+1}(x) = \operatorname{sgn}\left[\widetilde{\mathbf{W}}(x,.)\ \tilde{\mathbf{b}}_t + \tilde{\mathbf{k}}(x)\right] \tag{1.35}$$

where

$$\widetilde{\mathbf{W}} = \begin{bmatrix} \mathbf{0} & \mathbf{W} \\ \mathbf{W} & \mathbf{0} \end{bmatrix} \qquad \tilde{\mathbf{k}} = \begin{bmatrix} \mathbf{k} \\ \mathbf{k} \end{bmatrix} \tag{1.36}$$

Observe that the statements in $\widetilde{\mathcal{P}}$ can be partitioned into two sets:

$$\begin{aligned} \widetilde{\mathcal{P}}_1 &= \left\{x_1, x_2, ..., x_{\#(\mathcal{P}-\mathcal{S})}\right\} \\ \widetilde{\mathcal{P}}_2 &= \left\{x_{\#(\mathcal{P}-\mathcal{S})+1}, x_{\#(\mathcal{P}-\mathcal{S})+2}, ..., x_{2\#(\mathcal{P}-\mathcal{S})}\right\} \end{aligned} \tag{1.37}$$

where the weight between any two statements in $\widetilde{\mathcal{P}}_1$ $\left(\text{respectively } \widetilde{\mathcal{P}}_2\right)$ is zero.

Let the sequence of attitudes resulting from our original reasoning mapping be $\{\mathbf{b}_t\}_t$, and set the initial attitudes in the constructed mapping to $\tilde{\mathbf{b}}_0 = [\ \mathbf{b}_0\ \ \mathbf{b}_0\ ]'$. Suppose that reasoning with the constructed mapping is in cyclic mode with the following order: $x_1, x_2, ..., x_{2\#(\mathcal{P}-\mathcal{S})}, x_1, ...$

Since the attitudes toward statements in $\widetilde{\mathcal{P}}_1$ do not affect each other, after $\#(\mathcal{P}-\mathcal{S})$ sequential inferences, the attitudes toward statements in $\widetilde{\mathcal{P}}_1$ will be the same as the attitudes resulting from one parallel iteration in the original reasoning mapping.

Applying this argument inductively:

$$\mathbf{b}_{2t} = \left[\tilde{\mathbf{b}}_{2\#[\mathcal{P}-\mathcal{S}]t}\left(x_{\#[\mathcal{P}-\mathcal{S}]+1}\right), ..., \tilde{\mathbf{b}}_{2\#[\mathcal{P}-\mathcal{S}]t}\left(x_{2\#[\mathcal{P}-\mathcal{S}]}\right)\right]' \tag{1.38}$$

$$\mathbf{b}_{2t+1} = \left[\tilde{\mathbf{b}}_{\#[\mathcal{P}-\mathcal{S}][2t+1]}\left(x_1\right), ..., \tilde{\mathbf{b}}_{\#[\mathcal{P}-\mathcal{S}][2t+1]}\left(x_{\#[\mathcal{P}-\mathcal{S}]}\right)\right]' \tag{1.39}$$

Under the consistency assumptions, $\widetilde{\mathbf{W}}$ is symmetric and has a zero diagonal; thus the proof for theorem 1 holds. ■

1.8.3 Proof of Proposition 3

At the end of a cycle of inferences, either the set of attitudes remains unchanged, or some attitudes have changed. In the first case, we have constructed a story. In the second, the G function from the proof of theorem 1 must have decreased at some point during the course of the cycle. From the definition of the stopping time, T, in equation (1.13), and because $\Delta G(t) \leqslant 0$ has been established, we have the following inequality:

$$\left[\frac{T}{\#(\mathcal{P}-\mathcal{S})}\right] \cdot \min_{\Delta G(t)<0} |\Delta G(t)| \leqslant G(0) - G(T) \tag{1.40}$$

Each strict decrease in G can be no smaller than $\min|\Delta G(t)|$ and the distance by which the G function has to fall to reach a story is $G(0) - G(T)$. T is divided by $\#(\mathcal{P}-\mathcal{S})$ because, even prior to convergence, the G function does not necessarily decrease at every iteration: it only necessarily does so during every cycle.

Now, from the proof of convergence, we know the following:

$$\min_{\Delta G(t)<0} |\Delta G(t)| \geqslant 2\delta \tag{1.41}$$

where δ is given by (1.33). It is also obvious that:

$$G(0) \leqslant -2 \sum_{x\in\mathcal{P}-\mathcal{S}} \mathbf{k}^{-}(x) - \sum_{x\in\mathcal{P}-\mathcal{S}} \sum_{y\in\mathcal{P}-\mathcal{S}} \mathbf{W}^{-}(x,y) \tag{1.42}$$

$$G(T) \geqslant -2 \sum_{x\in\mathcal{P}-\mathcal{S}} \mathbf{k}^{+}(x) - \sum_{x\in\mathcal{P}-\mathcal{S}} \sum_{y\in\mathcal{P}-\mathcal{S}} \mathbf{W}^{+}(x,y) \tag{1.43}$$

where $u^+ = \max\{0, u\}$ and $u^- = \min\{0, u\}$. It follows that:

$$G(0) - G(T) \leqslant 2 \sum_{x \in \mathcal{P}-\mathcal{S}} |\mathbf{k}(x)| + \sum_{x \in \mathcal{P}-\mathcal{S}} \sum_{y \in \mathcal{P}-\mathcal{S}} |\mathbf{W}(x, y)| \tag{1.44}$$

Substituting (1.41) and (1.44) into the inequality in (1.40), we obtain the required result. ■

1.8.4 Proof of Proposition 4

Let $\mathbf{a} \in \{0, 1\}^{\#\mathcal{P}}$ and $\mathbf{b} \in \{0, 1\}^{\#\mathcal{P}}$ be two stories for a given set of reasons $\mathbf{R}$ and $\mathbf{Q}$, and a given signal $\mathcal{S}$. Obviously, $\mathbf{a}^{\mathcal{S}} = \mathbf{b}^{\mathcal{S}}$. Suppose, contra-hypothesis, that they only differ over one statement in $\mathcal{P} - \mathcal{S}$, say x. From the definition of equilibrium,

$$\mathbf{a}(x) = \operatorname{sgn}\left(\sum_{y \in \mathcal{P}-\mathcal{S}} \mathbf{W}(x, y)\ \mathbf{a}^{\mathcal{P}-\mathcal{S}}(y) + \mathbf{k}(x) \right) \tag{1.45}$$

$$\mathbf{b}(x) = \operatorname{sgn}\left(\sum_{y \in \mathcal{P}} \mathbf{W}(x, y)\ \mathbf{b}^{\mathcal{P}-\mathcal{S}}(y) + \mathbf{k}(x) \right) \tag{1.46}$$

By assumption, $\mathbf{W}(y, y) = 0$ for all $y \in \mathcal{P}$, and $\mathbf{a}(y) = \mathbf{b}(y)$ for all $y \in \mathcal{P} \backslash \{x\}$. So I can rewrite (1.45) as:

$$\mathbf{a}(x) = \operatorname{sgn}\left(\sum_{y \in \mathcal{P}-\mathcal{S}} \mathbf{W}(x, y)\ \mathbf{b}^{\mathcal{P}-\mathcal{S}}(y) + \mathbf{k}(x) \right) = \mathbf{b}(x) \tag{1.47}$$

which contradicts the assumption that $\mathbf{a}(x) \neq \mathbf{b}(x)$. ■

1.9 References

Baltzer, A. and N. Pennington (1983), "Reasoning About Conjunctions and Disjunctions of Events: An Explanation-Based Account, mimeo., Psychology Department, University of Colorado. Boulder.

Barberis, N., A. Schleifer and R. Vishny (1998), "A Model of Investor Sentiment", *Journal of Financial Economics*, 49.

Bruck, J. and J. Goodman (1988), "A Generalized Convergence Theorem for Neural Networks", *IEEE Transactions on Information Theory*, 34(5).

Bruner, J. and M. Potter (1964), "Inference in Visual Recognition", *Science*, 144.

Dawes, R. (1999), "A Message From Psychologists to Economists: Mere Predictability Doesn't Matter Like it Should (Without a Good Story Appended to It.)", *Journal of Economic Behavior & Organization*, 39(1).

Festinger, L. (1957), *A Theory of Cognitive Dissonance*, Row Peterson, Evanston, Illinois.

Goles, E., F. Fogelman and D. Pellegrin (1985), "Decreasing Energy Functions as a Tool for Studying Threshold Networks", *Discrete Applied Mathematics*, 12(3).

Hebb, D. (1949), *The Organization of Behavior: A Neuropsychological Theory*, Wiley, New York, UK.

Hopfield, J. (1982), "Neural Networks and Physical Systems with Emergent Collective Computational Abilities", *Proceedings of the National Academy of Sciences*, April, 79.

Kahneman, D., P. Slovic and A. Tversky (Eds.) (1982), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, UK.

Minsky, M. (1985), *The Society of Mind*, Simon & Schuster, New York, New York.

Mullainathan, S. (1997), "A Memory-Based Model of Bounded Rationality", mimeo., Department of Economics, Harvard University.

Pennington, N. and R. Hastie (1986), "Evidence Evaluation in Complex Decision Making", *Journal of Personality and Social Psychology*, 51(2).

Pennington, N. and R. Hastie (1988), "Explanation-Based Decision Making: Effects of Memory Structure on Judgment", *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14(3).

Pennington, N. and R. Hastie (1990), "Practical Implications of Psychological Research on Juror and Jury Decision Making", *Personality and Social Psychology Bulletin*, 16(1).

Pennington, N. and R. Hastie (1992), "Explaining the Evidence: Tests of the Story Model for Juror Decision Making", *Journal of Personality and Social Psychology*, 62(2).

Pennington, N. and R. Hastie (1993), "Reasoning in Explanation-Based Decision Making", *Cognition*, 49.

Rabin, M. (1998), "Psychology and Economics", *Journal of Economic Literature*, 36(1).

Rabin, M. and J. Schrag (1999), "First Impressions Matter: A Model of the Confirmatory Bias", *Quarterly Journal of Economics*.

Rubinstein, A. (1998), *Modeling Bounded Rationality*, MIT Press, Cambridge, Massachusetts.

Shafir, E., I. Simonson, A. Tversky (1993), "Reason-Based Choice", *Cognition*, 49.

Shiller, R. (1997), "Human Behavior and the Efficiency of the Financial System", mimeo., Cowles Foundation for Research in Economics, Yale University.

Simon, H. (1982), *Models of Bounded Rationality*, Vol.2, The MIT Press, Cambridge, Massachusetts.

Chapter 2

Revising Non-Additive Priors

With Yianis Sarafidis

Many thanks to Ben Polak for teaching us decision theory, and for many fruitful discussions. Ettore Damiano, David Pearce and Mario Simon made many helpful comments during the course of this project. Lam gratefully acknowledges financial support from the Cowles Foundation in the form of an Anderson fellowship.

2.1 Introduction

In a wide range of dynamic economic situations with incomplete information, agents are required to update their initial beliefs upon the receipt of some informative signal or message. For example, an employer may update her prior on the quality of a worker after observing the worker's output. Or the manager of a potential entrant in an industry may update his prior of being fought, after observing the actions of the incumbent firm to previous entrants.

In the belief revision process, five different probability measures may be involved. In order to clarify this, and to illustrate the questions addressed by this paper, we consider the following concrete example. An employer has just hired an employee. The worker can either be one who exerts high effort (θ_H) or one who exerts low effort (θ_L). It is assumed that the employer does not observe the worker's type but she does observe the worker's output, which again can be either high (y_H) or low (y_L). Define the set of possible types and output levels to be $\Theta = \{\theta_H, \theta_L\}$ and $Y = \{y_H, y_L\}$, respectively. The state space is all possible combinations of the worker's type and the output produced: we denote this by: $S = \Theta \times Y = \{(\theta_H, y_H), (\theta_H, y_L), (\theta_L, y_H), (\theta_L, y_L)\}$. The five probability measures are:

$\nu : 2^{\Theta} \to [0,1]$	The unconditional prior over types Θ
$\mu : 2^{Y} \to [0,1]$	The unconditional measure over signals Y
$\sigma : 2^{S} \to [0,1]$	The joint measure over the product space $S = \Theta \times Y$
$\mu(. \mid \theta) : 2^{Y} \to [0,1]$	Conditional likelihood over signals Y, given a type θ in Θ
$\nu(. \mid y) : 2^{\Theta} \to [0,1]$	Posterior over types Θ, given a signal y in Y

If these measures are all additive, the information contained in them can be summarized in three equivalent ways: (a) by σ, the joint probability over the product space: (b) by the prior for types ν, together with the set of likelihoods $\{\mu(. \mid \theta)\}_{\theta \in \Theta}$; and (c) by the unconditional measure over signals μ, together with the set of posteriors over types $\{\nu(. \mid y)\}_{y \in Y}$. Bayes's theorem allows us to move among these representations.

From the point of view of economic applications, agents typically possess information in the form of (b). It is natural to assume that the employer has initial beliefs over the quality of the worker, and that her knowledge of the production process implies knowledge about the distribution over output, conditioned on each of the worker's types.

Now, imagine that the employer's knowledge is indeed in the form of (b), and that a low output (y_L) is realized. How does she update her beliefs on the employee's type? The updating problem is trivial. First, the employer transforms the representation in (b) to that in (a) by a simple rearrangement of Bayes's rule. For all $\theta \times y$ in $\Theta \times Y$,

$$\sigma[(\theta, y)] = \nu(\theta) \cdot \mu(y \mid \theta) \tag{2.1}$$

Having obtained the joint beliefs over the product space, another application of Bayes's rule produces the posterior probability that the employee is type θ:

$$\nu(\theta \mid y_L) = \sigma[(\theta, y_L) \mid \{(\theta_H, y_L), (\theta_L, y_L)\}] = \frac{\sigma[(\theta, y_L)]}{\sigma[(\theta_L, y_L)] + \sigma[(\theta_H, y_L)]} \tag{2.2}$$

This is essentially moving from representing the information using (a) to representing it by (c).

In this updating framework, the prior measure ν is subjective while the likelihood distributions $\{\mu(. \mid \theta)\}_{\theta \in \Theta}$ are often objective.[1] Recent work in decision theory has sought to represent the subjective beliefs of uncertainty-averse agents in the form of a non-additive measure. Schmeidler (1989) and Gilboa (1987) show that if the decision maker's preferences satisfy certain axioms that are consistent with uncertainty aversion, then they choose as if they are maximizing Choquet expected utility. That is, preferences can be represented by a utility function which requires an expectation with respect to a non-additive measure.

If our hypothetical employer possesses such a prior, the three ways of representing information discussed above are no longer equivalent: revising her beliefs over types in the light of signals is no longer so obvious. This non-equivalence arises because Bayes's theorem does not hold for non-additive measures. One may think that the Dempster-Shafer rule for calculating conditional capacities—which we will describe subsequently—can be used in place of Bayes's rule. This is partly justified by the work of Gilboa and Schmeidler (1993), who show that a particular form of "pessimism" in preferences leads to the rule as an updating device for non-additive measures.[2]

---

[1] In the traditional view among economists, the subjective probability measure should be thought of as arising from some representation of the decision maker's preferences. In particular, Savage (1954) and Anscombe and Aumann (1963) outline the axioms for an expected-utility representation. Bayes's rule can also be justified through preferences. (See Myerson 1991.)

[2] The assumption is that when conditioning preferences on a particular event, the agent assumes that the best possible outcome obtains in the impossible states. See Gilboa and Schmeidler for details.

By analogy to the additive case, one may attempt to use the Dempster-Shafer rule to construct joint beliefs σ and then condition on the relevant partition of the product space to obtain posterior beliefs over types $\nu(. \mid y)$. Unfortunately, the first stage of this procedure fails. In general, unique beliefs over the state space S cannot be obtained from the Dempster-Shafer rule alone. This has important implications because in many economic applications, such as the example here, beliefs over the state space S are not given in the specification of the problem. Although the Dempster-Shafer rule can be used to calculate posteriors once the joint measure is known, our maintained assumption is that information is presented to the economist in the form of (b).

We propose two rules for defining a measure over the space S. Under the first proposal, the value of a set in 2^S is given by the iterated expectation of the corresponding indicator function. Expectation is first taken with respect to $\{\mu(. \mid \theta)\}_{\theta \in \Theta}$, and then with respect to ν. We refer to this procedure as the *Choquet-indicator rule*. With additive probability measures, this is of course the correct thing to do because the expectation of an indicator function over a set is the probability of that set. When beliefs are non-additive, we show that this rule still has desirable properties. In the second approach, we recognize that the perception of uncertainty embodied in ν can be equivalently represented by a set of additive measures, which we denote by P. Each of these distributions over the type space Θ can be taken in turn and used to construct a probability over the state space S. We refer to this rule as the *multiple-priors rule*. It produces a set of distributions, denoted by Q.

The two rules are closely related, but not equivalent. This non-equivalence

arises because non-additive measures are unable to capture certain restrictions on the relative likelihood of events. While this does not matter for the representation of uncertainty-averse beliefs, it results in a loss of information when beliefs have to be revised.

The updating problem considered in this paper is in fact closely related to a theoretical question which has received some attention in the literature. When an individual has non-additive beliefs, whether the objects of choice are Anscombe-Aumann "*horse-lotteries*" (functions from states to *lotteries* over consequences) or Savage *acts* (functions from states to consequences) affects her preference for randomization. (Eichberger and Kelsey 1996) This in turn has implications for the desirability of mixed strategies in games with uncertainty-averse players.

To see the relationship between this literature and our paper, note that the signal processing example has two stages of randomness. The first relates to *uncertainty* about which element of Θ corresponds to reality (no objective probabilities are available), and the second relates to *risk* about which signal from Y will be received (objective probabilities). This problem can thus be placed within the Anscombe-Aumann model, where the objects of choice are precisely such two-stage lotteries. Within this framework, the non-additive measure over types ν should not be viewed as primitive, but rather arising from the representation of some preference ordering $\succsim^{AA}$.[3]

Obtaining beliefs over the state space can now be rephrased in terms of preferences.

---

[3]The superscript refers to the Anscombe-Aumann setting.

We will show how the binary relation $\succsim^{AA}$ over two-stage horse-lotteries *induces* an ordering over one-stage acts in the Savage framework. Denote this induced relation by $\succsim^{SV}$. Finding a measure over the product space S is then equivalent to finding a Choquet expected utility representation for the Savage preferences $\succsim^{SV}$.

Based on the non-equivalence of the Choquet-indicator rule and the multiple-priors rule, we argue that the difference between Anscombe-Aumann decision making and the Savage framework arises, not from inherent differences between one and two-stage lotteries, but from the inability of non-additive priors to model uncertainty as precisely as multiple priors.

The rest of the paper is organized as follows. Section 2 provides the notation and outlines some existing results. Section 3 presents the theoretical framework for our updating problem. In section 4, we introduce the two rules for constructing beliefs and discuss the relationship between them. Section 5 makes the case that, at least within dynamic updating problems, a multiple-priors representation of uncertainty is more appropriate. A summary, together with some conclusions, are to be found in section 6.

2.2 Notation and Preliminaries

Let Θ be a finite set of *types* and Y denote the set of *signals*. From the specification of the problem, we have a convex *capacity* ν over Θ.

Definition (Capacity). *A* capacity *or non-additive measure over* Θ *is a function* $\nu : 2^{\Theta} \to [0, 1]$ *satisfying the following:*

(i) $\nu(\emptyset) = 0$, $\nu(\Theta) = 1$

(ii) *For* $A_1, A_2 \subseteq \Theta$, $A_1 \subseteq A_2 \Rightarrow \nu(A_1) \leq \nu(A_2)$

If (ii) holds, ν *is* monotone.

We say that ν *is* convex, *or supermodular, if in addition, the following holds:*

(iii) $\nu(A_1 \cup A_2) \geqslant \nu(A_1) + \nu(A_2) - \nu(A_1 \cap A_2)$, *for all* $A_1, A_2 \in 2^{\Theta}$

It is superadditive, *if (iii) holds for disjoint* A_1 *and* A_2.

For each type $\theta \in \Theta$, there is an additive probability distribution over the set of signals Y which may be received. These lotteries represent objective risk and we denote them by $\{\mu(\cdot \mid \theta)\}_{\theta \in \Theta}$.

The most popular scheme for updating a convex capacity is the Dempster-Shafer rule. For additive measures, this rule corresponds to Bayes's rule.

Definition (Dempster-Shafer). *The* Dempster-Shafer update *of a convex capacity* ν *conditioned on event is* $A \subseteq \Theta$ *is defined by the following expression. For all* $A_1 \subseteq A$,

$$\nu(A_1 \mid A) = \frac{\nu(A_1 \cup A^c) - \nu(A^c)}{1 - \nu(A^c)} \tag{2.3}$$

To see how our problem relates to the literature on decision theory, we need to introduce preferences. Let $\succsim^{AA}$ represent a preference ordering over *horse-lotteries*. A horse-lottery in our notation is simply a mapping from Θ onto the set of probability

distributions over consequences, $h : \Theta \to \Delta C$, where C is the set of consequences. Denote the set of horse-lotteries by H.

Throughout, we assume that preferences $\succsim^{AA}$ are primitive and that they satisfy the Schmeidler (1989) axioms for representation as a Choquet expected utility function so that for all h and h' :

$$h \succsim^{AA} h' \text{ if and only if } \int U \circ h \, d\nu \geqslant \int U \circ h' \, d\nu \tag{2.4}$$

where U is a von-Neumann-Morgenstern linear utility function with a Bernoulli utility function, $u : C \to \mathbb{R}^+$, over consequences. The capacity over types ν is obtained from this representation. Calculating utility involves a two-stage expectation. In the von-Neumann-Morgenstern utility, the expectation is with respect to the lotteries over consequences. Expectation is then carried out using the Choquet integral which is defined as follows:

Definition (Choquet Integral). *Let $g : \Theta \to \mathbb{R}$ be a random variable. The* Choquet integral *of g with respect to the capacity ν is defined as:*

$$\int g \, d\nu = g_1 \nu(A_1) + \sum_{i=2}^{n} g_i \left[\nu \left(\cup_{j=1}^{i} A_j \right) - \nu \left(\cup_{j=1}^{i-1} A_j \right) \right] \tag{2.5}$$

where g_i is the i^{th} highest consequence under g and $A_i \in 2^\Theta$ is the event in which the consequence g_i occurs.

Because a preference ordering which admits a Choquet expected utility representation can always be represented as a maxmin expected utility, we know from Gilboa

and Schmeidler (1989) that there exists a closed convex set P of additive probability measures on Θ, such that for all h and h' :

$$h \succsim^{AA} h' \text{ if and only if } \min_{p \in P} \int U \circ h \, dp \geqslant \min_{p \in P} \int U \circ h' \, dp \tag{2.6}$$

Moreover, the set of multiple priors P is the *core* of ν. We will abuse notation by referring to p as both a measure and a vector.

Definition (Core). *The* core *of a non-additive measure ν, denoted by $core(\nu)$ is defined, as in the cooperative theory for transferable-utility games, by:*

$$core(\nu) = \left\{ p = (p_1, \ldots, p_{|\Theta|}) \in \Delta^{|\Theta|-1} \mid \sum_{i \in A} p_i \geqslant \nu(A), \text{ for all } A \subseteq \Theta \right\} \tag{2.7}$$

For the purpose of the signaling problem, the set of probability distributions, $\{\mu(. \mid \theta)\}_{\theta \in \Theta}$, are objective and fixed. Therefore, to place our problem within the Anscombe-Aumann decision setting, we have to restrict the set of horse-lotteries to those in which the second-stage risk is given by some element of $\{\mu(. \mid \theta)\}_{\theta \in \Theta}$. The consequences attached to these probabilities can differ between horse-lotteries. We denote this set of restricted horse-lotteries by $H_\mu \subseteq H$. To illustrate in the context of our motivating example, consider the following capacity and likelihoods:

$$\begin{gathered} \nu(\theta_H) = \tfrac{1}{4}, \ \nu(\theta_L) = \tfrac{1}{4}, \text{ and } \nu(\{\theta_H, \theta_L\}) = 1 \\ \mu(y_H \mid \theta_H) = \tfrac{4}{5}, \ \mu(y_L \mid \theta_H) = \tfrac{1}{5} \\ \mu(y_H \mid \theta_L) = 0, \ \mu(y_L \mid \theta_L) = 1 \end{gathered} \tag{2.8}$$

The employer's prior over the worker's type is characterized by ambiguity and results in a non-additive measure. The production process yields a high output (y_H) with

probability $\frac{4}{5}$ when the worker is a high-effort type (θ_H). It yields low output (y_L) with probability 1 if the worker is of the low-effort type (θ_L). With these numbers, elements of H_μ take the form of the following pair of lotteries: $h(\theta_H) = \langle \frac{4}{5}, c_1;\ \frac{1}{5}, c_2 \rangle$: $h(\theta_L) = \langle 0, c_3;\ 1, c_4 \rangle$ where $c_i \in C$, for all $i \in \{1, 2, 3, 4\}$.

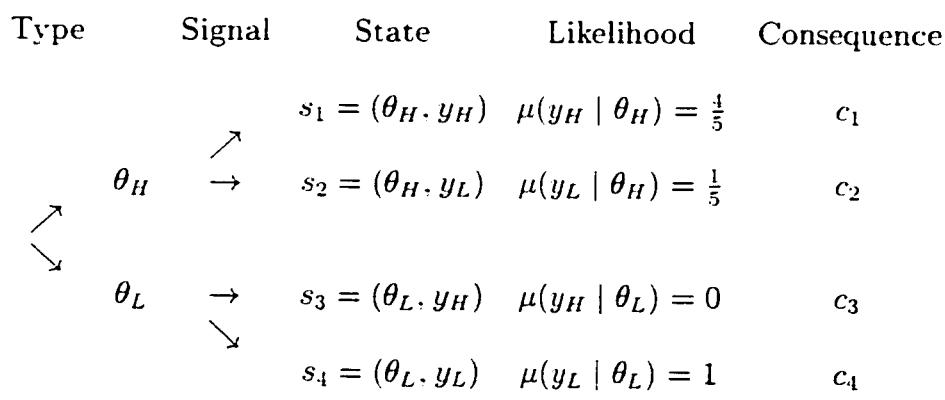
In the next section, it is necessary to compare preference orderings under $\succsim^{AA}$ with those under Savage preferences $\succsim^{SV}$. Call the product space $S = \Theta \times Y$ the set of *states*. Savage preferences are defined over *acts*, which are mappings from states to consequences, $f : S \to C$. Denote the set of acts by F. To facilitate comparison between $\succsim^{AA}$ and $\succsim^{SV}$, we need this additional definition:

Definition (Induced Act). *Write* $Y = \{y_1, y_2, \ldots, y_n\}$ *and consider lotteries over consequences with the dimension of the support equal to the cardinality of* Y*; that is, for all* θ *in* Θ, $h(\theta) = \langle \mu(y_1 \mid \theta), c_{\theta,y_1}; \ldots; \mu(y_n \mid \theta), c_{\theta,y_n} \rangle$. *The* act over states induced by the horse-lottery h in H_μ *is a mapping,* $f^h : S \to C$, *defined as:*

$$f^h(\theta \times y) = c_{\theta,y} \tag{2.9}$$

Notice that induced acts do not depend on the probabilities which are part of the specification of horse-lotteries. In the Savage setting, the risk contained in lotteries over consequences is modeled explicitly as part of the description of the state. Continuing with the example above, the horse-lottery—$h(\theta_H) = \langle \frac{4}{5}, c_1;\ \frac{1}{5}, c_2 \rangle$, $h(\theta_L) = \langle 0, c_3;\ 1, c_4 \rangle$—in the Anscombe-Aumann framework induces the following Savage act: $f^h = (c_1, c_2, c_3, c_4)$. Figure 1 illustrates this example.

Figure 1



Axiomatizations for both Choquet and maxmin expected utility exist in the Savage setting. For Choquet expected utility, see Gilboa (1987) and Sarin and Wakker (1992). Casadesus-Masanell *et al.* (1998) axiomatize the maxmin expected utility representation. One final piece of notation. We denote the capacity and the set of multiple priors over the state space, $S = \Theta \times Y$, by σ and Q, respectively.

2.3 Theoretical Framework

As we pointed out in the introduction, updating the capacity over types upon the receipt of a signal requires the construction of beliefs on the product space $S = \Theta \times Y$. How should this—in general, non-additive—measure be constructed? What desiderata should σ possess?

By placing our problem within the framework of preferences, we obtain a very natural property that σ should satisfy. Assume that the capacity ν over Θ is the result of a representation of the primitive ordering $\succsim^{AA}$ over horse lotteries in H_μ.

Based on this preference ordering, we can *define* a relation $\succsim^{SV}$, over acts, according to the following. For all $h, h' \in H_\mu$,

$$h \succsim^{AA} h' \Leftrightarrow f^h \succsim^{SV} f^{h'} \tag{2.10}$$

Having done so, constructing beliefs on the state space S amounts to finding a measure σ that represents $\succsim^{SV}$. That is, we want σ to satisfy the following utility representation:

$$f^h \succsim^{SV} f^{h'} \Leftrightarrow \int_S u\left[f^h(s)\right]\, d\sigma(s) \geq \int_S u\left[f^{h'}(s)\right]\, d\sigma(s) \tag{2.11}$$

We can obtain another perspective by re-stating the requirement in (2.10) as that of finding a σ such that the expected utility representation of the two preference orderings are equivalent. For all h in H_μ we want,

$$\int_\Theta U\left[h(\theta)\right]\, d\nu(\theta) = \int_S u\left[f^h(s)\right]\, d\sigma(s) \tag{2.12}$$

The left-hand-side contains the utility function which represents $\succsim^{AA}$; the right-hand-side contains the representation of $\succsim^{SV}$. This equation can be written more explicitly as:

$$\int_\Theta \int_Y u\left[f^h(\theta \times y)\right]\, d\mu(y \mid \theta)\, d\nu(\theta) = \int_{\Theta \times Y} u\left[f^h(\theta \times y)\right]\, d\sigma(\theta \times y) \tag{2.13}$$

Equation (2.13) allows us to restate the problem. The aim is to find a measure, σ, on the product space, $S = \Theta \times Y$, for which part of Fubini's theorem holds. Fubini's theorem states that the order of the iterated integrals with respect to two marginal measures do not matter and that both are equal to integration with respect to the product measure. Condition (2.13) requires only that integration over Y, *then* Θ, be equivalent to integration with respect to the product measure.

Sarin and Wakker (1992) were the first to observe that, in general, it is not possible to find a capacity σ on S which satisfies (2.13). This can be illustrated in the context of our example. Consider the following three horse-lotteries, h_1, h_2, h_3:

$$\begin{aligned} h_1(\theta_H) &= \left\langle \tfrac{4}{5}, 1;\ \tfrac{1}{5}, 0 \right\rangle, \quad h_1(\theta_L) = \langle 0.0;\ 1.0 \rangle \\ h_2(\theta_H) &= \left\langle \tfrac{4}{5}, 1;\ \tfrac{1}{5}, 0 \right\rangle, \quad h_2(\theta_L) = \langle 0.0;\ 1.1 \rangle \\ h_3(\theta_H) &= \left\langle \tfrac{4}{5}, 2;\ \tfrac{1}{5}, 0 \right\rangle; \quad h_3(\theta_L) = \langle 0.0;\ 1.1 \rangle \end{aligned} \tag{2.14}$$

Recall that each lottery is of the form $h(\theta) = \langle \mu(y_1 \mid \theta), c_1; \mu(y_2 \mid \theta), c_2 \rangle$. These are all elements of H_μ because the probabilities are identical and given by $\{\mu(\cdot \mid \theta)\}_{\theta \in \Theta}$. Respectively, they induce the following acts in the Savage formulation:

$$\begin{aligned} f^{h_1} &= (1.0.0.0) \\ f^{h_2} &= (1.0.0.1) \\ f^{h_3} &= (2.0.0.1) \end{aligned} \tag{2.15}$$

Without loss of generality, we assume that consequences are in utils, or that the Bernoulli utility function is given by $u(c) = c$. To satisfy equation (2.13) for f^{h_1} and f^{h_2}, it can then be verified that we need:

$$\begin{gathered} \sigma\left[(\theta_H, y_H)\right] = \tfrac{4}{20} \\ \sigma\left[\{(\theta_H, y_H), (\theta_L, y_L)\}\right] = \tfrac{17}{20} \end{gathered} \tag{2.16}$$

However, with these values, the Choquet expected utility of the horse lottery h_3 in the Anscombe-Aumann framework is given by $\int_X \int_Y u\left(f^{h_3}\right) d\mu d\nu = \frac{23}{20}$ while the Choquet expected utility of the corresponding induced act f^{h_3} in the Savage framework is given by $\int_S u\left(f^{h_3}\right) d\sigma = \frac{21}{20}$.

This difference between the two frameworks does have important implications. For example, Eichberger and Kelsey (1996) show that in the Anscombe-Aumann framework (represented by the left-hand-side of equation 2.13), uncertainty-averse agents exhibit a preference for randomization, but they do not necessarily do so when the objects of choice are Savage acts (the representation of the right-hand-side of 2.13). We have shown that the difference also matters when agents are revising non-additive priors upon the receipt of some signal.

2.4 Obtaining Beliefs Over the State Space

Having established that it is impossible to obtain a capacity σ which satisfies the desideratum of (2.13), we now consider some weaker desirable properties which we may want a capacity over the state space to satisfy.

One obvious feature which we would like our rule to possess is that it should correspond to Bayes's rule in the special case of additive distributions. In order to ensure this, rectangular sets formed by partitioning S according to some element of Θ that is, sets of the form $\theta \times B$, where $\theta \in \Theta$ and $B \in 2^Y$—must have measure given by:

$$\sigma(\theta \times B) = \nu(\theta) \cdot \mu(B \mid \theta) \tag{2.17}$$

This is of course just Bayes's rule when ν is additive. We will refer to (2.17) as the *multiplicative property*. However, this still leaves the measure of many subsets in S, including many rectangles, unspecified.

From the Dempster-Schafer rule, we have the following:

$$\mu(B \mid \theta) = \frac{\sigma\left(\theta^C \cup B\right) - \sigma\left(\theta^C\right)}{1 - \sigma\left(\theta^C\right)} \tag{2.18}$$

Rearranging.

$$\begin{aligned} \sigma\left(\theta^C \cup B\right) &= \mu(B \mid \theta)\left[1 - \sigma\left(\theta^C\right)\right] + \sigma\left(\theta^C\right) \\ &= \mu(B \mid \theta)\left[1 - \nu\left(\theta^C\right)\right] + \nu\left(\theta^C\right) \end{aligned} \tag{2.19}$$

The right-hand-side of (2.19) is given by the specification of the problem. Thus using the Dempster-Shafer rule we can obtain the value of the joint capacity on sets of the form $\theta^C \cup B$, where $B \in 2^Y$ and $\theta \in \Theta$. We say that a capacity σ satisfies the *Dempster-Shafer property* if it obeys (2.19).

Even if we impose the multiplicative property of (2.17), as well as the Dempster-Shafer property of (2.19), we do not obtain a unique capacity. In a similar spirit to Hendon *et al.* (1991), we can characterize the *set* of capacities that we do obtain by some limits if we require that σ be monotone. Take for example the set $E = \{(\theta_H, y_L), (\theta_L, y_L)\}$ from figure 1. Although equations (2.17) and (2.19) do not provide a unique value for its measure, we can derive the following bound using a simple set inclusion argument:

$$\max\{\sigma(\theta_H, y_L), \sigma(\theta_L, y_L)\} \leqslant \sigma(E) \leqslant \min\{\sigma(\theta_H \cup Y), \sigma(\theta_L \cup Y)\} \tag{2.20}$$

These bounds can be calculated using the multiplicative property and the Dempster-Shafer property.

2.4.1 The Choquet-Indicator Rule

We now propose a rule for obtaining a *unique* capacity over the state space which satisfies the multiplicative and Dempster-Shafer properties. The definition is as follows.

Definition (Choquet-Indicator Rule). *A capacity σ on S with marginal ν over Θ and a set of likelihood distributions $\{\mu(\cdot \mid \theta)\}_{\theta \in \Theta}$ over Y, is said to be generated by the* Choquet-indicator rule *if for every $E \in 2^{\Theta \times Y}$:*

$$\sigma(E) = \int_{\Theta} \int_{Y} 1_E \, d\mu(y \mid \theta) \, d\nu(\theta) \tag{2.21}$$

where 1_E is an indicator function over E.

Result. *Let σ be a capacity on $\Theta \times Y$ calculated using the Choquet-indicator rule. Then σ satisfies the multiplicative and Dempster-Shafer properties in (2.17) and (2.19), respectively.*

Remark. *By construction, over the set of acts, $\{f \in F \mid u[f(s)] \in \{0, 1\}$ for all s in $S\}$, the Choquet-indicator rule satisfies (2.13), our original desideratum.*

The result is easy to verify. The remark says that, if one restricts attention to acts which take on only two consequences, the Choquet-indicator (CI) rule maintains the equivalence between the Anscombe-Aumann and the Savage frameworks. It is constructed to do so. A comparison of equations (2.13) and (2.21) makes this obvious.

To illustrate how the CI rule works, consider the set $\{(\theta_H, y_H), (\theta_L, y_L)\}$ from our example above. Naively, one could make the following calculation:

$$\begin{aligned} \sigma[\{(\theta_H, y_H), (\theta_L, y_L)\}] &= \nu(\theta_H) \cdot \mu(y_H \mid \theta_H) + \nu(\theta_L) \cdot \mu(y_L \mid \theta_L) \\ &= \frac{1}{4} \times \frac{4}{5} + \frac{1}{4} \times 1 = \frac{9}{20} \end{aligned} \tag{2.22}$$

This calculation assigns a probability of $\frac{1}{4}$ to each of the two types, θ_H and θ_L. It ignores the fact that, with the residual probability of $\frac{1}{2}$, either θ_H or θ_L will necessarily occur. The CI rule corrects for this in the most "pessimistic" way. It assigns the residual probability to that outcome which would produce the lowest Choquet expectation, in this case (θ_H, y_H), as $\mu(y_H \mid \theta_H) = \frac{4}{5} < \mu(y_L \mid \theta_L) = 1$.

Despite being intuitive, and despite satisfying the multiplicative property and the Dempster-Shafer property—both of which seem desirable—the CI rule does not imply an equivalence between the Anscombe-Aumann and Savage frameworks. As we pointed out in the previous section, no rule which generates a capacity can. If we are willing to leave the non-additive framework, and allow uncertainty-averse beliefs to be represented by a set of multiple *additive* priors, can we do better? In the next subsection, we present a rule for calculating multiple priors over states which yields an equivalence result between the one and two-stage frameworks for decision making.

Before this is done, we discuss the relationship between the Choquet-indicator rule and the work of Ghirardato (1997). Ghirardato considered a situation where two non-additive marginal measures are known. He asked what conditions are necessary to obtain a capacity σ on the product space which satisfies the Fubini theorem. He showed that the theorem will hold if one restricts the set of acts to those which

are *slice comonotonic* and imposes on σ a strengthening of independence, which he termed the *Fubini property*[4].

Definition (Fubini Property). *A function $g : \Theta \times Y \to \mathbb{R}$ is slice comonotonic if for every $\theta, \theta' \in \Theta$, $g(\theta, \cdot)$ and $g(\theta', \cdot)$ are comonotonic, and if for every $y, y' \in Y$, $g(\cdot, y)$ and $g(\cdot, y')$ are comonotonic. Define a* comonotonic set *as one in which the indicator function over that set is* slice-comonotonic. *Now, a capacity σ is said to satisfy the* Fubini property *with respect to marginals, ν over Θ and μ over Y, if the following equation holds for every comonotonic set $E \in 2^{\Theta \times Y}$:*

$$\sigma(E) = \int_{\Theta} \int_{Y} 1_E \, d\mu(y) \, d\nu(\theta) \tag{2.23}$$

where 1_E is the indicator function over E.

Our Choquet-indicator (CI) rule can be viewed as a strengthening of the Fubini property; it imposes that (2.23) hold for *all* sets $E \in 2^S$. (The CI rule also differs from equation (2.23) in that one of the measures in the integral is a conditional one.)

Ghirardato's definition does not require that equation (2.23) hold for all sets because, when applied to all elements in 2^S, the capacity generated by (2.23) does not satisfy the "iterated integration" part of the Fubini theorem. It is not clear whether one should define $\sigma(E)$ as $\int_{\Theta} \int_{Y} 1_E d\mu d\nu$ or $\int_{Y} \int_{\Theta} 1_E d\nu d\mu$. For comonotonic sets, these two integrals are equivalent.

---

[4]This is not to be confused with the Fubini theorem. The Fubini property is a characteristic of capacities.

In our updating framework, the order of integration is clear so this part of Fubini's theorem is not a desirable restriction. By strengthening the Fubini property, we are able to obtain a *unique* capacity over the product space: there are multiple capacities over the product space S which satisfy the Fubini property. Ghirardato requires the additional assumption of convexity to obtain uniqueness. In general, our CI rule does not produce a convex capacity. However, we argue that this is not a weakness since convexity restricts the kind of uncertainty which one can model. This point will become clearer when we define the alternative way to obtain beliefs in the next subsection.

2.4.2 The Multiple-Priors Rule

For an agent with preferences $\succsim^{AA}$ over lotteries $h \in H$ who satisfy the axioms for a Choquet expected utility representation with a convex capacity ν over Θ, Gilboa and Schmeidler (1989) showed that the agent is behaviorally equivalent to one with a maxmin expected utility: that is, one who maximizes $\min_{p \in P} \int_\Theta U \circ h \, dp$, where $P = core(\nu)$. In the other direction, assume that an agent possesses a maxmin expected utility representation with a set of multiple priors P. The agent is identical to one who maximizes Choquet expected utility with a capacity ν, defined by $\nu(A) = \min_{p \in P} p(A)$, if and only if ν is convex and $core(\nu) = P$.

In light of these results, we can convert the capacity ν, which is convex by assumption, to the corresponding set of multiple priors $P = core(\nu)$. Is it then possible to obtain a set of *additive* beliefs Q over the product space $S = \Theta \times Y$ so as to obtain equivalence between the one and two-stage formulations? More formally, we require

Q to satisfy:

$$\min_{p \in P} \int_{\Theta} U\left[h(\theta)\right] \, dp(\theta) = \min_{q \in Q} \int_{S} u\left[f^{h}(s)\right] \, dq(s) \tag{2.24}$$

where again U is a von-Neumann-Morgenstern utility function with Bernoulli utility u. This is simply the multiple priors analogue to equation (2.12). The left-hand-side is the maxmin expected utility representation of $\succsim^{AA}$; the right-hand-side is the representation of $\succsim^{S}$. Consider the following rule for calculating a set of distributions Q over the state space using the priors over types, $P = core(\nu)$, and the likelihoods, $\{\mu(. \mid \theta)\}_{\theta \in \Theta}$.

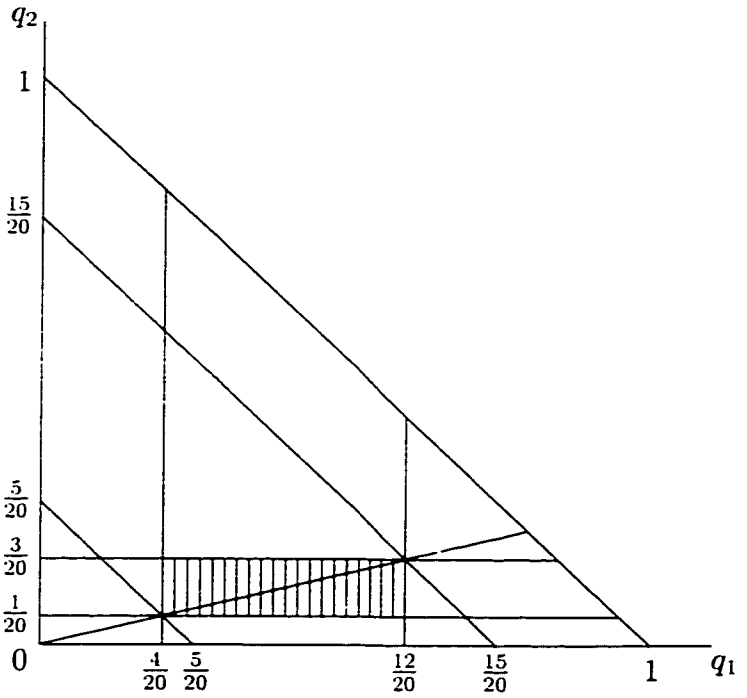
$$Q = \left\{q = (q_1, ..., q_{|S|}) \in \Delta^{|S|-1} \mid q_s = p(\theta) \times \mu(y \mid \theta) \text{ for all } p \in P\right\} \tag{2.25}$$

where $s = \theta \times y$. We refer to this as the multiple-priors (MP) rule. It simply takes each prior in P in turn and applies a rearrangement of Bayes's rule. In the case of our motivating example, we obtain:

$$Q = \left\{q \in \Delta^{3} \mid q_1 \in \left[\tfrac{4}{20}, \tfrac{12}{20}\right], \; q_2 \in \left[\tfrac{1}{20}, \tfrac{3}{20}\right], \; q_3 = 0, \; q_4 \in \left[\tfrac{5}{20}, \tfrac{15}{20}\right], \; q_1 = 4q_2\right\} \tag{2.26}$$

Figure 2 illustrates the projection of this set onto the (q_1, q_2) space. We can completely represent the set in two dimensions because $q_3 = 0$ and q is on the simplex. These two restrictions reduce the degrees of freedom to two.

Figure 2



Proposition (Multiple-Priors Rule). *The multiple-priors rule satisfies the Anscombe-Aumann and Savage equivalence. That is, it satisfies equation (2.24) for all* $h \in H_\mu$.

We omit the proof, since the result is a direct consequence of the fact that the Fubini theorem holds with additive priors. This result, though simple, is somewhat surprising given that it cannot be obtained in terms of a capacity on S. Intuition will be provided in the next subsection.

Remark. *The set of additive measures* Q *generated from the multiple-priors rule cannot in general be expressed as the core of any capacity.*

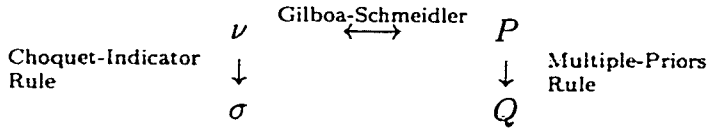
The figure above is an example of this remark. From the definition of equation (2.7), we can see that a set can only be expressed as the core of some capacity if it can be defined by a system of linear inequalities of the form $\sum_{i \in A} p_i \geqslant \nu(A)$. Geometrically, the set must have sides which are parallel to the sides of the simplex. In figure 2, the set of distributions, Q, is represented by the line segment connecting the points $\left(\frac{4}{20}, \frac{1}{20}, 0, \frac{15}{20}\right)$ and $\left(\frac{12}{20}, \frac{3}{20}, 0, \frac{5}{20}\right)$. Since this line is not parallel to any of the sides of the triangle, it cannot be expressed as the core of any capacity.

2.4.3 The Relationship Between the Two Rules

The two procedures for updating beliefs over types can be summarized as follows. We begin with a convex, non-additive prior ν over the set of types Θ. One can think of these beliefs as deriving from some Choquet expected utility representation of the agent's preferences in an Anscombe-Aumann setting. To obtain posterior beliefs after the receipt of some signal from Y, we need to first define beliefs over the product space, $S = \Theta \times Y$. One way to do this is to use a rule based on the Choquet integration of indicator acts. This yields a measure σ which is in general non-additive, reflecting the transfer of uncertainty and uncertainty aversion over the types to uncertainty and uncertainty aversion over states. The measure σ satisfies the multiplicative property and the Dempster-Shafer property. It does not, however, represent an ordering over Savage acts which is equivalent to the Anscombe-Aumann preference ordering.

An alternative approach is to convert the non-additive measure ν to an equivalent set of multiple priors, P. Multiple probability distributions over states can then be obtained using Bayes's rule. The resulting set is labeled Q.

Figure 3



The main advantage of the multiple-priors approach is that it ensures an equivalence between the one and two-stage frameworks.

If one wanted to remain within the non-additive framework, then an obvious question is: which capacity comes "closest" to representing the uncertainty over states embodied in the set Q? It turns out that the capacity which does so is indeed the one calculated from the CI rule. This idea can best be described graphically using our example. Figure 2 shows that, among all sets of distributions which can be expressed as the core of some capacity, the shaded rectangle is the *smallest* one which contains Q. One can verify that, if we define σ using the CI rule, then this rectangle is precisely the set $core(\sigma)$. The theorem and corollary below formalize this.

Theorem (CI and MP Rules). *Assume that ν on Θ is convex. Let σ be the capacity on S defined by the Choquet-indicator rule and let Q be the set of multiple additive measures on S derived from the multiple-priors rule. Then,*

$$\sigma(E) = \min_{q \in Q} q(E) \tag{2.27}$$

for all $E \in S$.

Proof. From Schmeidler (1989), proposition (x), we have:

$$\int_{\Theta} U[h(\theta)]\, d\nu(\theta) = \min_{p \in P} \int_{\Theta} U[h(\theta)]\, dp(\theta) \tag{2.28}$$

for any act h where ν is convex. From our rules, $\sigma(E) = \int_{\Theta} 1_E d\nu(\theta)$ and $\min_{q \in Q} q(E) = \min_{p \in P} \int_{\Theta} 1_E dp(\theta)$. The result follows, by using the act 1_E for h in (2.28). ■

Corollary. *Let σ be the capacity on S defined under the Choquet-indicator rule and let Q be the set on multiple additive measures on S derived from the multiple-priors rule. Then,*

$$Q \subseteq core(\sigma) \tag{2.29}$$

Proof. Let $q \in Q$. Assume, contra-hypothesis, that $q \notin core(\sigma)$. Then, there exists $E \in S$, such that $q(E) < \sigma(E)$. But this contradicts the theorem above. ■

As an aside, the CI rule produces a capacity σ which is not necessarily convex. This is easy to verify in the example above. Despite this, we have the following remark.

Remark. *Because Q is always non-empty, the corollary implies that σ has a core which is non-empty.*[5]

The theorem of this section, together with its corollary, provides a formal justification for the Choquet-indicator rule. Given that the multiple-priors rule satisfies the requirement of (2.24), we argue that the agent's beliefs over the product space should in fact be given by this rule. The Choquet-indicator rule can then be justified on the grounds that it produces that capacity which comes the closest to the "correct" beliefs.

Before concluding this section, we return to the original motivation for this paper and construct, for our example, the posteriors on types, Θ, after the observation of a low output, y_L. Applying the Dempster-Schafer rule to the capacity σ, generated by the CI rule, we obtain the following posterior:

$$\begin{aligned} \nu(\theta_H \mid y_L) &= \tfrac{1}{16} \\ \nu(\theta_L \mid y_L) &= \tfrac{13}{16} \\ \nu(\{\theta_H, \theta_L\} \mid y_L) &= 1 \end{aligned} \tag{2.30}$$

---

[5] In a recent working paper, Ghirardato and Marinacci (1998) argue that, within the Savage framework, ambiguity aversion corresponds to nonemptiness of the core, a property strictly weaker than convexity. In light of this result, the CI rule maintains the initial uncertainty aversion even though the capacity σ on S is not convex.

This measure has the following core:

$$core\,[\nu(.\mid y_L)] = \left\{p = (p_1, p_2) \in \Delta^1 \mid p_1 \in \left[\tfrac{1}{16}, \tfrac{3}{16}\right]\right\} \tag{2.31}$$

We now wish to construct posterior beliefs over types using the set of additive distributions on the product space calculated from the multiple-priors rule. Gilboa and Schmeidler (1993) show that, for decision makers who can be represented both by Choquet expected utility and by maxmin expected utility, the Dempster-Shafer rule on capacities coincides with the combination of maximum likelihood and Bayes's rule applied to the set of multiple priors. Therefore, to enable comparison between the Choquet indicator rule and the multiple-priors rule, we apply maximum likelihood to the set Q and then use Bayes's rule, element-by-element, to obtain posterior beliefs over types. This gives the following unique additive posterior:

$$\left\{p = (p_1, p_2) \in \Delta^1 \mid p_1 = \tfrac{1}{16}\right\} \tag{2.32}$$

Comparing equations (2.31) and (2.32), it is clear that the CI rule yields posterior beliefs which contain greater uncertainty than those obtained from the multiple-priors rule.

2.5 The Argument for Multiple Priors

From the work of Gilboa and Schmeidler (1989), we know that the multiple-priors framework is more general than that of convex capacities. Any convex capacity can be represented as a set of multiple priors, whereas the converse is not true. What is surprising about the updating example is that, even though we begin with beliefs on

types which can be represented equivalently by a capacity ν or by a set of multiple priors P, as soon as we introduce signals and attempt to construct beliefs on the product space, the two frameworks diverge.

We believe that the updating problem considered in this paper highlights the importance of the additional generality of multiple priors. A comparison of the core of σ from the CI rule, with Q from the MP rule, makes this point. In the example:

$$core(\sigma) = \left\{q \in \Delta^3 \mid q_1 \in \left[\tfrac{4}{20}, \tfrac{12}{20}\right],\ q_2 \in \left[\tfrac{1}{20}, \tfrac{3}{20}\right],\ q_3 = 0,\ q_4 \in \left[\tfrac{5}{20}, \tfrac{15}{20}\right]\right\} \qquad (2.33)$$

$$Q = \left\{q \in \Delta^3 \mid q_1 \in \left[\tfrac{4}{20}, \tfrac{12}{20}\right],\ q_2 \in \left[\tfrac{1}{20}, \tfrac{3}{20}\right],\ q_3 = 0,\ q_4 \in \left[\tfrac{5}{20}, \tfrac{15}{20}\right],\ q_1 = 4q_2\right\} \qquad (2.34)$$

These two sets are identical except for the equation $q_1 = 4q_2$ in (2.34). *Capacities are unable to capture restrictions on the relative likelihood of some events.* The capacity σ is unable to restrict the probability of the first state, (θ_H, y_H), to be four times that of the second, (θ_H, y_L). Clearly, given that the risk associated with the signal is objective, no matter what the agent's original beliefs over types, the likelihood ratio between these two states should remain 4 to 1. By trying to use capacities to capture beliefs on the product space, the agent loses some of the information contained in the signal and attributes to the problem greater uncertainty than is in fact present. In turn, this leads to greater uncertainty in the posterior, as the sets in (2.31) and (2.32) demonstrate.

Moreover, the inability of capacities to capture relative likelihoods is the reason for the non-equivalence between the Anscombe-Aumann and the Savage frameworks.

(Recall that in the multiple-priors setting, the one and two-stage frameworks are equivalent.) As a result, Eichberger and Kelsey's (1996) claim that the Savage framework is more appropriate for modeling uncertainty aversion is not justified. There is really no inherent difference between one and two-stage lotteries as objects of choice. Differences arise from a limitation of capacities.

2.6 Conclusion

In many dynamic economic situations, beliefs over the relevant state space are not given by the specification of the problem. With additive measures, Bayes's rule usually suffices to define a unique distribution over states. However, with ambiguous beliefs represented by a non-additive measure, unique beliefs over the state space cannot be obtained from the Dempster-Shafer rule alone.

We argued that obtaining beliefs over the state space is closely related to the issue of whether one can move from the Anscombe-Aumann to the Savage setting while maintaining the "same" preference ordering. This is impossible when capacities are used to model uncertainty-aversion. However, using multiple additive distributions, this equivalence between the two frameworks is possible. We then set out to find the capacity which comes closest to the beliefs obtained using multiple-priors. Such a capacity can be constructed by taking Choquet expectations of appropriate indicator functions.

Finally, we showed that the updating problems studied in our paper highlight the advantage of multiple priors relative to non-additive measures. Capacities are unable

to place restrictions on the relative likelihood of events. This is a severe limitation in dynamic problems where such ratios arise naturally from the updating of beliefs.

2.7 References

Anscombe, F. and R. Aumann (1963), "A Definition of Subjective Probability", *Annals of Mathematical Statistics*. 34, 199-205.

Casadesus-Masanell, R., P. Klibanoff, and E. Ozdenoren (1998), "Maxmin Expected Utility Over Savage Acts With a Set of Priors", mimeo., Northwestern University.

Eichberger, J. and D. Kelsey (1996), "Uncertainty Aversion and Preference for Randomization", *Journal of Economic Theory* 71, 31-43.

Ghirardato, P. (1997), "On Independence for Non-Additive Measures With a Fubini Theorem", *Journal of Economic Theory* 73, 261-291.

Ghirardato, P. and M. Marinacci (1998), "Ambiguity Made Precise: A Comparative Foundation", mimeo..

Gilboa, I. (1987), "Expected Utility Theory With Purely Subjective Non-Additive Probabilities", *Journal of Mathematical Economics* 16, 65-88.

Gilboa, I. and D. Schmeidler (1989), "Maxmin Expected Utility With a Non-Unique Prior", *Journal of Mathematical Economics* 18, 141-153.

Gilboa, I. and D. Schmeidler (1993), "Updating Ambiguous Beliefs", *Journal of Economic Theory* 59, 33-49.

Hendon, E., H. Jacobsen, B. Sloth and T. Tranæs (1991), "The Product of Capacities and Lower Probabilities", mimeo., University of Copenhagen.

Mukerji, S. (1997), "Understanding the Non-Additive Probability Decision Model", *Economic Theory* 9, 23-46.

Myerson, R. (1991), *Game Theory: Analysis of Conflict*, Harvard University Press, Cambridge, Massachusetts.

Sarin, R. and P. Wakker (1992), "A Simple Axiomatization of Non-Additive Expected Utility", *Econometrica* 60, 1255-1272.

Savage, L. (1954), *The Foundations of Statistics*, Wiley, New York.

Schmeidler, D. (1989), "Subjective Probability and Expected Utility Without Additivity", *Econometrica* 57, 571-587.

Chapter 3

Self-Sustaining Stability in Dynamic Matching Markets

With Ettore Damiano

Many thanks to Dirk Bergemann for introducing us to matching models and for his invaluable advice and constant encouragement. David Pearce, Herbert Scarf and Abhijit Sengupta provided helpful comments. Financial support from Nicola Damiano and Fung Ping Lam is most gratefully acknowledged.

3.1 Introduction

Many trading arrangements in the real world do not satisfy the assumptions of a Walrasian model of exchange. A special class of such arrangements is *two-sided matching markets*. These markets are characterized by two important features. First, agents belong to two disjoint sets: they cannot switch from one side of the market to the other no matter what the market condition. A second feature is the bilateral nature of exchange: the contrast is with centralized goods markets where the identity of one's trading partner is a matter of indifference. Examples of two-sided markets include many labor markets, as well as auction markets.

In this paper, we are concerned with a subclass of two-sided matching markets, namely those in which matches are *one-to-one*: each agent may be matched with at most one partner from the opposite set. Historically, the two sides of the market have been labeled *males* and *females*, and the model termed a *marriage market*. In many applications, a many-to-one relationship is more realistic but the issues we are concerned with can be discussed in the simpler class of markets.

Any testable theory of a matching market must place some restrictions on the kind of outcomes that one expects to observe. An obvious restriction is that outcomes be "stable". In thinking about stability, we have in mind cooperative concepts similar to the *core*. An outcome which is not in the core is, by definition, susceptible to blocking by rational agents. In general, there will of course be many other restrictions imposed by the incentives and rules associated with a particular trading institution. We consider the requirement of stability for two main reasons. First, in markets

where participation is voluntary, it represents a minimal constraint. Second, the cooperative notion of stability requires only a very general description of the game[1] and so is applicable to many markets, whereas issues of non-cooperative, strategic behavior depend crucially on the particular trading arrangement and information structure of the market under consideration.[2]

A large and very successful literature has considered stability in the special case of a static market with perfect information. (Roth and Sotomayor 1990 provide an excellent summary.) Existence of the core has been established and many of its interesting characteristics noted. In many markets, however, agents trade repeatedly. In fact, the value from a proposed match is often not known until after trade has taken place, so that agents have to learn about the parameters in the game as it is played over time. This paper takes a first step toward studying stability in such a setting by considering matching markets where trade occurs repeatedly but where there is complete information on the value of matches.

In a repeated market, a *matching plan* specifies a partner for each agent, at each

---

[1] Cooperative game theory considers the *characteristic form* of the game, which specifies the set of achievable payoffs for all possible coalitions.

[2] The consideration of stability in matching markets can be of practical importance. In a famous example in the United States, the market for matching physicians in their first position following medical school with hospitals was very disorderly until the Association of American Medical Colleges adapted a centralized mechanism in the 1951-2 market. The procedure is a version of the Gale and Shapley (1962) algorithm and was successful largely because it implemented matches which were in the core. See Roth and Sotomayor (1990) for details.

point in time. An obvious candidate notion for stability in these markets is the core over the set of feasible matching plans. However, the core has a particularly unsatisfactory property in a dynamic game: it can admit matching plans which are not time-consistent. A plan is in the core as long as it is stable at the beginning of the game; its continuation need not be stable at any other point in time. If agents cannot make binding agreements, elements in the core may be blocked at some later point in time. For our definition of stability, we impose a requirement which can be viewed as the cooperative analogue of *subgame-perfection*. Becker and Chakrabarti (1995), impose a similar condition in their definition of the *recursive core*.

Although time-consistent, there remains an incongruity associated with the recursive core. In judging the stability of the grand coalition's matching plan—that is, the plan for all agents—the recursive core requires that the plan be immune to blocking by coalitions at every point in time. However, no deviating coalition is subject to the same requirement. That is, a set of players may be able to block the grand coalition's original plan using an "incredible" or unstable plan. This motivates our second requirement that blocking coalitions be self-enforcing. They must choose matching plans in which no subset of the coalition can reach an agreement to deviate from the deviation. These sub-coalitions have to satisfy the same requirement, and so on. This is the cooperative analogue to Bernheim, Peleg and Whinston's (1987) non-cooperative notion of *coalition proofness*. This condition is independent of the dynamics of the game and can be applied to static matching markets.

Our definition of stability imposes both of these requirements and we call it *self-sustaining stability*. The paper proceeds as follows. We begin by briefly summarizing

the literature on marriage models in a static, complete-information setting. We then introduce dynamics by allowing agents to match repeatedly over time. We adapt the standard definition of the core, as well as Becker and Chakrabarti's (1995) definition of the recursive core, to this repeated game and argue that they both have limitations.

In section 3, we formally define and illustrate the concept of self-sustaining stability. In sections 4 and 5, we consider its existence and computation. Unfortunately, existence is only guaranteed with very strong conditions. When it does exists, the self-sustaining stable set can be calculated using backward recursion for finitely-repeated markets. In infinite-horizon games, we propose an algorithm, which employs the idea of dynamic programming, for computing the set. Section 6 compares self-sustaining stability to alternative concepts such as the core and the recursive core.

Finally, in section 7, we discuss the limitations of our definition and some initial thoughts on how our concept can be generalized to a market with incomplete information on the value of matches.

3.2 Stability in Marriage Markets

3.2.1 Static, Complete-Information Marriage Market

In this subsection, we summarize some of the results in the literature regarding stability in single-period marriage markets with complete information on the values of all matches. This will also serve to introduce notation and concepts. We denote the two disjoint sides of the market, the males and the females, by $M = \{m_1, m_2, ..., m_n\}$

and $F = \{f_1, f_2, ..., f_p\}$, respectively. We will also refer to the set of players, $M \cup F$, as the *grand coalition.*

Each individual has preferences over the other side of the market. Although none of the results in this section depend on the cardinality of preferences, we nevertheless assign actual values to matches; this will be convenient when we consider a dynamic market and have to aggregate payoffs over time. Throughout the paper, we assume that the outside option associated with being single is normalized to zero for all agents. Payoffs can be summarized by a matrix. For example, with three agents on each side of the market, we may have the following 3×3 matrix:

	f_1	f_2	f_3
m_1	2, 3	3, 2	1, 3
m_2	3, 1	1, 1	2, 1
m_3	3, 2	2, 3	1, 2

(3.1)

The $(i, j)^{th}$ cell contains two numbers, being the payoffs to male m_i and female f_j, respectively, from a match with each other. In this example, all elements are strictly positive implying that it is never rational to remain unmatched. An outcome in the single-period market is referred to as a *matching.*

Definition (Matching). *In a static marriage model, where (M, F) are the two disjoint sets of players, a* matching *is a one-to-one function μ satisfying the following:*

$$\mu \quad : \quad (M \cup F) \rightarrow (M \cup F) \tag{3.2}$$

$$\text{if } \mu(m) \neq m \in M \text{ then } (m) \in F \tag{3.3}$$

$$\text{if } \mu(f) \neq f \in F \text{ then } \mu(w) \in M \tag{3.4}$$

$\mu(i) = i$ implies that individual i is unmatched; we say that agent i is self-matched or *single.*

For a group of players, S, we denote the set of all possible matchings by $\mathcal{M}_S$. Also, let $\pi_S(\mu_S)$ be a vector of payoffs for each agent in the set S from matching with the partner specified under $\mu_S \in \mathcal{M}_S$. If any element of $\pi_S(\mu_S)$ is strictly negative, we say that μ_S is not *individually rational.* If S is the grand coalition, $M \cup F$, we drop the subscript on π, μ and $\mathcal{M}$ for ease of notation. We are now in a position to define stability in this market.

Definition (Core Matching). *A matching $\mu \in \mathcal{M}$ is in the* core *if there does not exist a coalition of players $S \subseteq (M \cup F)$ with a matching $\mu_S \in \mathcal{M}_S$ such that:*

$$\pi_S(\mu_S) > \pi(\mu)^S \tag{3.5}$$

where x^S is the projection of a vector $x \in \mathbb{R}^{|M \cup F|}$, onto the subspace $\mathbb{R}^{|S|}$, $|M \cup F| \geqslant |S|$.

If a coalition which satisfies equation (3.5) does exist, it is referred to as the *blocking coalition.*

For the example of equation (3.1), the matching $\{\mu(m_1) = f_1, \mu(m_2) = f_2, \mu(m_3) = f_3\}$[3] is not in the core. It yields a payoff of $\pi(\mu) = (\ 2 \quad 1 \quad 1 \quad 3 \quad 1 \quad 2\ )$

[3]We only need to specify the partners of one side of the market because of the two-sidedness of the game.

where payoffs are ordered: $m_1, m_2, m_3, f_1, f_2, f_3$. This matching is blocked by a coalition of $\{m_1, f_2\}$ which can achieve for its two members $\pi_{\{m_1,f_2\}}(\mu_{\{m_1,f_2\}}) = (\ 3 \quad 2\ ) > (\ 2 \quad 1\ ) = \pi(\mu)^{\{m_1,f_2\}}$. However, $\{\mu(m_1) = f_2,\ \mu(m_2) = f_3,\ \mu(m_3) = f_1\}$ is a core matching.

An interesting property of the core in this market is that it is equivalent to the set of matchings which are not blocked by any one or two-player coalitions. That is, no matching can be blocked by a larger coalition if it is not blocked by a coalition with either one player or a pair of players. This result holds because of the "independence" of preferences: each player's preference over alternative matchings correspond exactly to his/her preference over his/her partners at these matchings. Roth and Sotomayor (1990) give a proof. Note however that, with a repeated matching game, coalitions of more than two players provide the possibility of altering partners and so matter when considering stability.[4]

There are a number of other well-known results relating to the static marriage market. Two important ones are:

Theorem 1 (Gale and Shapley 1962). *The core is non-empty for every marriage game.*

In light of this result, and to avoid confusion when we introduce dynamics, we refer

---

[4]This point is even more forceful with incomplete information. Agents may learn about parameters from the matches of others and so will in general no longer be indifferent between the alternative ways in which others are matched.

to the core in the static market as the *Gale-Shapley set.*

Theorem 2 (Knuth 1976). *When all agents have strict preferences, the common preferences of the two sides of the market are opposed on the Gale-Shapley set. That is, if μ and μ' are in the Gale-Shapley set, then all males in M like μ at least as well as μ' if and only if all females in F like μ' at least as well as μ.*

Theorem 1 is self-explanatory. One possible proof employs Scarf's (1967) theorem on the existence of a core. Gale and Shapley (1962) prove the claim by constructing an algorithm which always leads to a matching in the Gale-Shapley set. Theorem 2 is less intuitive. It says that over matchings in the Gale-Shapley set, agents on the same side of the market have a coincidence of preferences, while agents on different sides of the market have a conflict of interests. We will see that the analogue of this theorem does not hold in the dynamic market.

3.2.2 Stability in Dynamic Markets

We now consider two possible definitions of stability when a marriage game, such as that in equation (3.1), is played repeatedly over time, with players receiving payoffs each period.

Definition (Matching Plan). *In a marriage market repeated for T periods, a* matching plan *for a group of players, S, is a function $\boldsymbol{\mu}_S : \mathbb{N}_T \rightarrow \mathcal{M}_S$, where we define $\mathbb{N}_T$ to be the set of natural numbers up to and including T: $\{1, 2, ..., T\}$.*

Thus $\boldsymbol{\mu}_S$ specifies a matching at each point in time. The definition applies to both finite and infinite T.

An obvious notion of stability in this game is the core over the set of matching plans. Let $\beta \in [0, 1]$ be the discount factor and define the payoff function $\boldsymbol{\pi}_S$[5] over matching plans as the sum of discounted period payoffs:

$$\boldsymbol{\pi}_S(\boldsymbol{\mu}_S) = \sum_{t=1}^{T} \beta^{t-1} \pi(\boldsymbol{\mu}_S(t)) \tag{3.6}$$

As before, the absence of a subscript implies that the grand coalition, $M \cup F$, is being referred to.

Definition (Core Matching Plan). *A matching plan $\boldsymbol{\mu} : \mathbb{N}_T \to \mathcal{M}$ for a group of agents, $M \cup F$, is in the* core *if there does not exist a coalition $S \subseteq M \cup F$ such that:*

$$\boldsymbol{\pi}_S(\boldsymbol{\mu}_S) > \boldsymbol{\pi}(\boldsymbol{\mu})^S \tag{3.7}$$

for some matching plan $\boldsymbol{\mu}_S : \mathbb{N}_T \to \mathcal{M}_S$.

Consider this definition applied to the following 2×2 marriage game, played repeatedly for two periods with no discounting ($\beta = 1$):

	f_1	f_2
m_1	$5, -1$	$-1, 5$
m_2	$-1, 5$	$5, -1$

(3.8)

[5] In our notation, μ is a particular element of $\mathcal{M}$; while a bold $\boldsymbol{\mu}$ is a mapping from time onto $\mathcal{M}$. Similarly, π is a function over the set of matchings μ; while $\boldsymbol{\pi}$ is function over the set of matching plans $\boldsymbol{\mu}$.

Re that all agents receive a payoff of zero when they remain single. In the single-period game, the above matrix implies that males prefer the matching $\{\mu(m_1) = f_1, \mu(m_2) = f_2\}$ where they both receive a payoff of 5. Under this matching, females receive -1; they prefer the matching $\{\mu(m_1) = f_2, \mu(m_2) = f_1\}$. Neither of these two matchings are individually rational in the one-shot game: they are both blocked by some singleton coalition. The unique element of the Gale-Shapley set specifies that players remain self-matched: $\{\mu^{\text{single}}(i) = i, i \in (M \cup F)\}$, which yields $\pi(\mu^{\text{single}}) =$ (0 0 0 0).

In the two period version of this game, it is easy to see that all players can do better by agreeing to implement the male-preferred matching $\{\mu(m_1) = f_1, \mu(m_2) = f_2\}$ in one period and the female-preferred matching $\{\mu(m_1) = f_2, \mu(m_2) = f_1\}$ in the other. At the beginning of the game, such a plan cannot be blocked by any coalition. There are of course two matching plans which do this and they make up the core for this two-period game. Both plans yield a payoff vector: $\boldsymbol{\pi}(\boldsymbol{\mu}^{\text{core}}) = (\ 4\ \ 4\ \ 4\ \ 4\ )$, where $\boldsymbol{\mu}^{\text{core}}$ is one of the core matching plans.

Unfortunately, one would not expect to observe such an outcome in any play of the game if matching plans are not binding. In the second period, one side of the market always has an incentive to withdraw participation and to renege on the plan agreed upon in period 1. For example, assume that the plan requires that the "female-preferred" stage-matching be played in period 1, followed by the "male-preferred" stage-matching in the second period. Come period 2, the females would withdraw from the market rather than receive a payoff of -1. Knowing this, the males will

never agree to the female-preferred matching in period 1; the core matching plan unravels. In a dynamic setting, the core is unsatisfactory because it does not impose stability at each point in time.[6]

One candidate notion, which does impose stability at every point in time, is the *recursive core* of Becker and Chakrabarti (1995). This concept is closely related to the *sequential core* of Gale (1978). Both of these two concepts are motivated by the lack of trust in a general equilibrium model. The Arrow-Debreu treatment of time and uncertainty implicitly assumes that a promise to deliver a commodity is as good as the commodity itself. When this is not the case, Gale argues that only allocations in the sequential core are "trustworthy" and that the institution of money could act as a substitute for trust. Becker and Chakrabarti show that real capital goods can also provide a trust mechanism.

Let $\boldsymbol{\mu}_S \mid_t : \mathbb{N}_T \backslash \{1, 2, \ldots, t-1\} \rightarrow \mathcal{M}_S$, where $t \leqslant T$, be the matching function induced by $\boldsymbol{\mu}_S$ on the continuation game from time t onward. The idea of Becker and Chakrabarti (1995) can be adapted for our matching model according to the following definition.

---

[6]It may be thought that randomization could provide the solution. Players could agree to implement matchings based on a publicly observable flip of a coin each period. This will give each player an expected payoff equal to the payoff from $\boldsymbol{\mu}^{\text{core}}$. However, this requires that the outcome specified by the coin flip be enforceable, which is counter to the spirit of the paper. We want to place some restrictions on the matching plans which may be observed when enforceability is assumed not to be possible.

Definition (Recursive Core). *A matching plan* $\boldsymbol{\mu} : \mathbb{N}_T \to \mathcal{M}$, *for a group of player,* $M \cup F$, *is in the* recursive core *if* $\boldsymbol{\mu}\,|_t$: $\mathbb{N}_T \backslash \{1, 2, ..., t-1\} \to \mathcal{M}$ *is in the core of the continuation game from* t *onward, for all* $t \in \mathbb{N}_T$.

It is clear from this definition that the recursive core is a refinement, or a subset, of the core.

In terms of the two-period example in equation (3.8), this implies that the only candidate elements in the recursive core are the two core matching plans which specify either the male or the female-preferred matching in one period and an exchange of partners in the second. Neither of these two plans are in the recursive core. No matter which matching is carried out in the first period, in the second, the only stable matching for the remaining (one-shot) game is for all players to remain single. This is inconsistent with both matching plans, so the recursive core is empty in this example.

Once again this result seems unsatisfactory. Intuitively, the matching plan which specifies that agents remain unmatched in both periods appears to be robust to "blocking by rational agents". Yet, this matching plan is not in the recursive core. Denote this matching plan by $\boldsymbol{\mu}^{\text{single}}$; $\boldsymbol{\pi}(\boldsymbol{\mu}^{\text{single}}) = (\,0 \quad 0 \quad 0 \quad 0\,)$. In period 2, the continuation of $\boldsymbol{\mu}^{\text{single}}$ is consistent with the recursive core since it specifies a Gale-Shapley matching in the final stage. However, in period 1, $\boldsymbol{\mu}^{\text{single}}$ is blocked by the grand coalition agreeing to play one of the core matching plans. This highlights an inconsistency associated with the recursive core: coalitions are allowed too much freedom in choosing the deviating matching plan. In judging the original matching

plan, the recursive core requires that the plan be immune to blocking by coalitions. However, no deviating group of players (including a deviation of the grand coalition) is subject to the same requirement. In the example above, the grand coalition which blocks μ^{single} does so by agreeing to an alternative matching plan (namely μ^{core}) which we've already argued does not satisfy the requirement of time-consistency. This observation motivates our definition of self-sustaining stability.

3.3 Self-Sustaining Stability

In addition to time-consistency, the idea behind self-sustaining stability is to limit the set of possible plans played by blocking coalitions. It does so by requiring that these plans be self-enforcing. Coalitions must choose matching plans in which, at all points in time, no subset of the coalition can reach an agreement to deviate from the deviation. The sub-coalitions have to satisfy the same requirement, and so on. This implies that we have to define self-sustaining stability inductively beginning with the smallest coalition and ending with the grand coalition.

Definition (Self-sustaining Stability).

(1) *For a coalition of 1 player,* $\{i\}$, *the matching plan* $\boldsymbol{\mu}_{\{i\}} : \mathbb{N}_T \rightarrow \mathcal{M}_{\{i\}}$, *which specifies that the player remains single forever, satisfies* self-sustaining stability.

(2) *Assume that self-sustaining stability has been defined for all proper sub-coalitions,* C, *of some set of players,* S. *A matching plan* $\boldsymbol{\mu}_S : \mathbb{N}_T \rightarrow \mathcal{M}_S$ *satisfies* self-sustaining stability *if:*

(a) *There does not exist a coalition* $C \subset S$ *with a continuation matching plan from some* $t \in \mathbb{N}_T$, $\boldsymbol{\mu}_C \mid_t : \mathbb{N}_T \backslash \{1, 2, ..., t-1\} \rightarrow \mathcal{M}_C$, *which satisfies self-sustaining stability for coalition* C *from time* t *onward, and which satisfies the inequality:*

$$\boldsymbol{\pi}_C(\boldsymbol{\mu}_C \mid_t) > \boldsymbol{\pi}_S(\boldsymbol{\mu}_S \mid_t)^C \tag{3.9}$$

(b) *There does not exist another matching plan* $\boldsymbol{\mu}'_S : \mathbb{Z}_T \rightarrow \mathcal{M}_S$, *satisfying condition (a), such that:*

$$\boldsymbol{\pi}_S(\boldsymbol{\mu}'_S \mid_t) > \boldsymbol{\pi}_S(\boldsymbol{\mu}_S \mid_t) \tag{3.10}$$

at some $t \in \mathbb{N}_T$.

The set of matching plans which satisfy self-sustaining stability (S^3) is labeled the S^3 set.

Condition (1) in this definition serves to initialize the recursion. (2) is the inductive step. (a) ensures that at no point in time can member of a proper coalition do better by leaving the larger market and trading amongst themselves thereafter[7].

[7]Note that in the definition, coalitions which leave the game do not have the possibility of

assuming that they agree to implement self-sustaining stable plans in this smaller market. The final condition (b) simply requires that the S^4 contain no elements which can be dominated by a coalition consisting of all players in S.

Bernheim, Peleg and Whinston's (1987) non-cooperative concept of coalition proofness is motivated by the same considerations. They argue that in games with non-binding preplay communication, the Nash best-response property is a necessary, but not sufficient, condition for self-enforceability. An earlier concept which addresses this issue is *strong Nash equilibrium*. While Nash equilibrium is robust to unilateral deviations, a strong Nash equilibrium allows for deviations by every conceivable coalition of players. Bernheim, Peleg and Whinston argue that this is inconsistent because deviating coalitions are not subject to the same requirement as the grand coalition. Coalition proofness addresses this. In extensive form games, they further require that equilibria be dynamically consistent and define perfect coalition-proof equilibrium. This concept is the precise analogue to ours.

As an aside, if the S^4 is the counterpart of perfect coalition proof equilibria, we can usefully think of the core as the counterpart of strong Nash equilibria. Similarly, the recursive core can be thought of as the analogue to Rubinstein's (1980) concept of *strong perfect equilibrium*.

Although we will continue to discuss S^3 in the context of a two-sided matching market, it can of course be applied to any dynamic cooperative game, and we intend to

---

rejoining it in a subsequent period. It can be argued that this is an undesirable assumption. We will illustrate the weakness of this assumption in the example of (3.11).

analyze its properties in the context of these other models—such as a simple exchange economy—in subsequent research. At this point, it is worth noting a characteristic of S^3 which arises in the special case of marriage game:

Proposition 3 (S^3 in Static Marriage Markets). *In a static marriage market, the self-sustaining stable set is equivalent to the Gale-Shapley set.*

Proof. As we have already mentioned, in a static game only deviations by one or two-player coalitions matter for the Gale-Shapley set. If the matching μ is not in the Gale-Shapley set and gives one agent a negative payoff, then it is dominated by a singleton coalition, which by definition satisfies S^3. If μ yields positive payoffs but is not in the Gale-Shapley set because of some coalition $\{m, f\}$, then the coalition $\{m, f\}$ can achieve a payoff greater than $\left[\pi_{\{m\}}(\mu), \pi_{\{f\}}(\mu)\right]$. This vector is in turn greater than $(0, 0)$, so that the blocking coalition satisfies the self-sustaining criterion. It follows that μ does not satisfy S^3. In the other direction, if μ is not in the S^3 set, then μ is dominated by some matching μ' so it cannot be in the Gale-Shapley set. ■

Put another way, the proposition states that neither time-consistency, nor self-enforceability, have any impact on the Gale-Shapley set. The restriction of time-consistency is trivially satisfied in a static game. However, the ineffectiveness of self-enforceability in modifying the Gale-Shapley set is not true in a general static game. It arises only because of the two-sidedness of marriage markets.

3.4 S^3 in Finitely-Repeated Markets

3.4.1 Computation

In a finite-horizon repeated game, we can construct a S^3 matching plan via backward induction. The recursion is through both time and coalition size. In a marriage market, this task is made slightly easier by the above proposition, which implies that we can begin the recursion at the final stage game with a Gale-Shapley matching. Next, consider the single-agent coalitions in the subgame consisting of the final two time periods. For these singletons, remaining unmatched for two periods satisfies self-sustaining stability. For coalitions of two agents, a matching plan consistent with S^3 must satisfy two requirements. First, it must specify a Gale-Shapley matching in the final period. Second, it must not be blocked by any single agent leaving the game and remaining unmatched for two periods.

Having derived S^3 matching plans for all two-player coalitions in this two-period game, we proceed by considering larger and larger coalitions, until we reach the grand coalition. Along the way we always require that matching plans induce continuations that satisfy S^3 for the continuation game, and that they be undominated by proper coalitions specifying plans from their corresponding S^3 set. The only complication arises when multiple matching plans satisfy these two conditions. When this occurs, all matching plans on the Pareto frontier are in the S^3 set.[8] Having reached the grand

[8] If there are multiple elements in the S^3 set, we consider these in turn, much like one does when finding multiple subgame-perfect equilibria by backward recursion.

coalition in the two-period game, we then solve for S^3 in the continuation subgame beginning at the third-last period, and so on.

Applying this concept to the two-period example of equation (3.8) is relatively simple. Any element of the S^3 set must specify that agents remain single in the last period since this is the unique Gale-Shapley matching. At the beginning of the game, remaining unmatched for both periods is the only matching plan that satisfies S^3 for two-player coalitions. If a match were to occur in the first period, the requirement that players remain single in the second period would cause one agent to receive a negative payoff from the plan. Because of strict preferences in this example, coalitions of three agents cannot block any plan not blocked by coalitions of two or less players.[9]

It follows that for the grand coalition, we only need to ensure that matching plans: satisfy individual rationality; have continuations which are stable; and are not Pareto dominated by other plans which satisfy these two criteria. The only matching plan which meets these requirements is μ^{single}. This plan survives because the deviating plan which blocks μ^{single} from the recursive core is not admitted under S^3.

---

[9]With three players, the agent on the side of the market in short supply determines the matching plan.

3.4.2 Existence

Theorem 4 (Existence in a Finitely-Repeated Market). *There exists a matching plan which satisfies S^3 in a finitely-repeated marriage market if one of the following conditions hold:*

(a) *The discount rate, β, is sufficiently close to zero.*

(b) *There are less than, or equal to, two agents on each side of the market and each player has strict preferences.*

The proof of this theorem is deferred until the infinite horizon model where we have a similar theorem.

Proposition 5 (Non-existence in a Finitely-Repeated Market). *More generally, there may be no S^3 matching plan in a finitely-repeated marriage market.*

The following example illustrates the possibility of non-existence. Consider the following stage-game repeated twice, with no discounting:

	f_1	f_2	f_3	f_4
m_1	1, 1	1, 1	3, 2	2, 1
m_2	1, 1	1, 1	2, 1	3, 2
m_3	2, 3	1, 1	5, 1	1, 5
m_4	1, 2	2, 3	1, 5	5, 1

(3.11)

This stage game has a unique Gale-Shapley matching: $\{\mu^{GS}(m_1) = f_3, \mu^{GS}(m_2) = f_4, \mu^{GS}(m_3) = f_1, \mu^{GS}(m_4) = f_2\}$.[10] Any candidate for inclusion in the S^4 must spec-

[10]This can be found by applying the Gale and Shapley (1962) algorithm.

ify this matching in the final stage. Consider, for example, a matching plan which specifies the Gale-Shapley matching in both periods. We denote this by μ^{GS}. This plan is blocked by the coalition of $\{m_3, m_4, f_3, f_4\}$, which has a self-sustaining, stable matching plan, namely matching in both periods with a switch of partners in the second. This yields its members: $\pi_{\{m_3,m_4,f_3,f_4\}}(\cdot) = (\ 6\ \ 6\ \ 6\ \ 6\ ) > (\ 4\ \ 4\ \ 4\ \ 4\ ) = \pi(\mu^{GS})^{\{m_3,m_4,f_3,f_4\}}$.[11] Other candidates for S^3 can be similarly eliminated.

In addition to demonstrating non-existence, this example is important because it highlights a weakness in the S^3 concept. The candidate plan μ^{GS} is blocked by a coalition agreeing to a plan which is "credible" only because self-sustaining stability does not allow deviating coalitions to rejoin the original set. If it did, the coalition $\{m_3, m_4, f_3, f_4\}$ would fall apart in the second period since the Gale-Shapley matching does not involve these players matching amongst themselves.

3.5 S^3 in Infinitely-Repeated Markets

3.5.1 Characterization

Solving for S^3 in an infinite horizon game is more difficult. In order to solve for S^3 at any point in time, we have to have solved for S^3 in the continuation game, which is of course impossible with an infinite horizon. Fortunately the idea of dynamic programming suggests a solution. The work of Abreu, Pearce and Stacchetti (1986,

[11] In contrast to the example in (2.8), this plan, which involves switching partners, satisfies S^3 because all values in the matrix are strictly positive.

1990), henceforth APS, in the context of looking for sequential equilibria in infinitely-repeated games, employs a similar idea. What is different about the problem here is that the usual maximization operator has to be replaced by a "non-blocking" condition.

We now give an alternative definition of dynamic stability for an infinitely-repeated matching market. This definition makes the recursion through time more explicit than in the earlier definition and is written in terms of conditions on the value set, rather than on matching plans. It provides some additional insights.

Definition (Self-Sustaining Stable Value Set, S^3VS). *In an infinite-horizon game with a group of agents S, the* self-sustaining stable value set, $V_S \subseteq \mathbb{R}^{|S|}$, *is the set of payoff vectors associated with the S^3 set for coalition S :*

$$V_S \equiv \left\{ \boldsymbol{\pi}_S(\boldsymbol{\mu}) \mid \boldsymbol{\mu} : \mathbb{N}^+ \to \mathcal{M}_S \text{ satisfies } S^3 \right\} \tag{3.12}$$

The following definition is adapted from APS (1986).

Definition (Admissibility). *For a group of agents S, a pair (μ_S, w_S) is* admissible *with respect to $W_S \subseteq \mathbb{R}^{|S|}$ if*

(a) $\mu_S \in \mathcal{M}_S$ *and* $w_S \in W_S$

(b) *There does not exist a proper subset $C \subset S$ such that:*

$$v_C > \pi_S(\mu_S)^C + \beta w_S^C \tag{3.13}$$

for some $v_C \in V_C$.

(c) *There does not exist a (μ'_S, v'_S) satisfying (a) and (b) such that:*

$$\pi(\mu'_S) + \beta v'_S > \pi(\mu_S) + \beta v_S \tag{3.14}$$

Definition (S^3 Mapping). *For each $W_S \subseteq \mathbb{R}^{|S|}$, let the* S^3 mapping *be:*

$$\Gamma_S(W_S) = \{\pi_S(\mu_S) + \beta v_S \mid (\mu_S, v_s) \text{ is admissible with respect to } W_S\} \tag{3.15}$$

Proposition 6 (Bellman Equation). *In an infinitely-repeated game, the S^3VS for a groups of players, $M \cup F$, $V^* \subseteq \mathbb{R}^{|M \cup F|}$, is the largest compact set in $\mathbb{R}^{|M \cup F|}$ which satisfies the following fixed-point condition:*

$$V^* = \Gamma(V^*) \tag{3.16}$$

Proof. By definition. ■

It is worth restating the intuition behind S^3 in this new context. Notice first that the above formulation essentially converts an infinite-horizon problem into an infinite

sequence of static games. The payoffs in these static games are not given by π alone but rather by π augmented with some discounted element of V.

Recall that S^3 imposes time consistency and self-enforceability in addition to the core requirement. Notice from equation (3.16) that V is the argument in Γ, which implies that continuation payoffs have to come from the S^3VS (time consistency), and from condition (b) in definition 5.2 that proper coalitions have to take values from their S^3VS, $v_C \in V_C$ (self-enforceability).

The recursive core for an infinitely-repeated game can also be written as conditions on the value set. The definitions would be identical except that in part (b), we dispense with the requirement that $v_C \in V_C$ and replace it by:

$$v_C \in \{\boldsymbol{\pi}_C(\boldsymbol{\mu}_C) \mid \boldsymbol{\mu}_C : \mathbb{Z}_\infty \to \mathcal{M}_C\} \tag{3.17}$$

With this reformulation, it is clear that the recursive core has to be a subset of the S^3 set because blocking coalitions are allowed more freedom in forming their plans in the recursive core.

In the appendix, we outline how this dynamic programming approach can be used to compute the S^3VS.

3.5.2 Existence

Theorem 7 (Existence in Infinitely-Repeated Market). *There exists a matching plan which satisfies S^3 exits in an infinitely-repeated marriage market if one of the following conditions hold:*

(a) *The discount rate, $\beta \in [0, 1)$, is sufficiently close to zero.*

(b) *The limit-of-means criterion is used to aggregate payoffs over time.*[12]

(c) *There are less than, or equal to, two agents on each side of the market and each player has strict preferences.*

Proof. **(a)** We prove the stronger result that, for sufficiently small discount factor, the recursive core is non-empty. With a sufficiently small discount factor, agents view the repeated game as a sequence of static games and we know from theorem 2.4 that the Gale-Shapley set is non-empty. More formally, consider a matching plan $\boldsymbol{\mu}$ such that $\boldsymbol{\mu}(t)$ is an element of the Gale-Shapley set for all $t = 1, \ldots, \infty$. That is, for all t, there does not exist $S \subseteq M \cup F$ with $\mu_S \in \mathcal{M}_S$, such that $\pi_S(\mu_S) > \pi(\boldsymbol{\mu}(t))^S$. For sufficiently small $\beta \in [0, 1)$, the following inequality must also not hold for any coalition S and for any $\boldsymbol{\mu}_S \mid_{t+1}$: $\pi_S(\mu_S) > \pi(\boldsymbol{\mu}(t))^S + \beta\left[\boldsymbol{\pi}(\boldsymbol{\mu} \mid_{t+1})^S - \boldsymbol{\pi}_S(\boldsymbol{\mu}_S \mid_{t+1})\right]$. This implies that $\boldsymbol{\pi}_S(\boldsymbol{\mu}_S) > \boldsymbol{\pi}(\boldsymbol{\mu})^S$ cannot hold at any t. That is, $\boldsymbol{\mu}$ is in the recursive core and the proof is completed by recalling that the recursive core is a subset of the S^3 set.

(b) Let $\boldsymbol{\mu}^{\text{core}}$ be a core-matching plan when the payoffs are evaluated according to

---

[12] The limit-of-means criterion cannot be replaced with "$\beta \in [0, 1)$ is sufficiently large".

the limit-of-means criterion[13]. We will show that such a plan exists in theorem 12. Consider the payoff to a continuation plan for some $t \in \{1, 2, ...\}$:

$$\begin{aligned} \pi(\mu^{\text{core}} \mid_t) &= \lim_{s\to\infty} \frac{1}{s-t+1} \sum_{\tau=t}^{s} \pi(\mu^{\text{core}}(\tau)) \\ &= \pi(\mu^{\text{core}}) - \lim_{s\to\infty} \frac{1}{s-t+1} \sum_{\tau=1}^{t-1} \pi(\mu^{\text{core}}(\tau)) \qquad (3.18) \\ &= \pi(\mu^{\text{core}}) \end{aligned}$$

Since μ^{core} is, by assumption, not blocked by any coalition, $\mu^{\text{core}} \mid_t$ is also in the core of the continuation game, for all $t \in \{1, 2, ...\}$. This implies that μ^{core} is in the recursive core and hence satisfies S^3.

(c) This proposition is obvious when we have less than two agents on either side of the market. The only agent in its side of the market matches with its preferred partner every period if individual rationality is satisfied. Otherwise, everyone remains single. This plan is consistent with S^3. Consider the 2×2 market with $M \cup F = \{m_1, m_2, f_1, f_2\}$. We have just said that the S^3 set is non-empty for all proper subcoalitions $S \subset M \cup F$ and in fact V_S is a singleton for all S. Let $\mathcal{P} = \{..., C, ...\}$ be a partition of $M \cup F$ so that the sets in $\mathcal{P}$ are disjoint, exhaustive and non-empty.

Now, define $v(\mathcal{P})$ to be a stacking of the self-sustaining stable payoff vectors for each $C \in \mathcal{P}$. That is, $v(\mathcal{P})^C \in V_C$, for all $C \in \mathcal{P}$.[14] By definition, $v(\mathcal{P}) \geqslant 0$ because individual rationality must be satisfied for elements in V_C. Suppose contra-hypothesis that $V = \emptyset$, then it must be the case that for any $v(\mathcal{P})$, there is another partition $\mathcal{P}'$

---

[13] $\pi(\mu) = \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=1}^{t} \pi(\mu(\tau))$

[14] This point is unique because V_C is a singleton for each $C \subset \mathcal{P}$.

such that $v(\mathcal{P}')^{C'} > v(\mathcal{P})^{C'}$ for some $C' \in \mathcal{P}'$. Otherwise $v(\mathcal{P}) \in V$, contradicting the emptiness of V. It turns out that this implication cannot be satisfied when there are only two players on each side of the market.

To see this, first consider a partition $\mathcal{P}$ which contains a coalition of three players. We can always find another partition $\mathcal{P}'$, which contains no three-player coalition, such that $v(\mathcal{P}') \geqslant v(\mathcal{P})$. The same is true for partitions containing singletons, or partitions where no coalition contains members of both sides of the market. We can therefore ignore all partitions except for $\mathcal{P} = \{\{m_1, f_1,\}, \{m_2, f_2\}\}$ and $\mathcal{P}' = \{\{m_1, f_2,\}, \{m_2, f_1\}\}$. Without loss of generality, say that $v(\mathcal{P})$ is blocked by $\{m_1, f_2\}$: that is, $v(\mathcal{P}')^{\{m_1,f_2\}} > v(\mathcal{P})^{\{m_1,f_2\}}$. Now, because V is assumed to be empty, $v(\mathcal{P}')$ must be blocked by either $\{m_1, f_1\}$ or $\{m_2, f_2\}$, say $\{m_1, f_1\}$, so that $v(\mathcal{P})^{\{m_1,f_1\}} > v(\mathcal{P}')^{\{m_1,f_1\}}$. For this to be consistent with the inequality above, we must have $v(\mathcal{P}')^{\{m_1\}} = v(\mathcal{P})^{\{m_1\}}$. This contradicts the assumption of strict preferences stated in the theorem. It is thus not possible to block all partitions.[15] ■

Proposition 8 (Non-existence in Infinitely-Repeated Markets). *There may not exists a matching plan which satisfies S^3 in an infinitely-repeated, one-to-one, two-sided matching market.*

Even though condition (a), (b) or (c) in theorem 7 guarantees existence in special

---

[15]Notice that the proof only considers a subset of the possible values in V; V will in general contain points which are not in the set of $v(\mathcal{P})$ values, but we show that under the assumptions of the theorem, for at least one partition $\mathcal{P}$, $v(\mathcal{P}) \in V$.

classes of infinitely-repeated matching games, it is not possible to extend the proof of existence to a generic game. The following example is one in which no matching plan satisfies S^3.

Consider the matching market in equation (3.8) with the addition of an extra male who has positive value only when matched with player f_1. The matrix of payoffs is:

	f_1	f_2
m_1	$5, -1$	$-1, 5$
m_2	$-1, 5$	$5, -1$
m_3	$\frac{4}{5}, \frac{4}{5}$	$-2, -2$

(3.19)

Assume that this stage market is repeated infinitely with a discount factor of $\beta = \frac{1}{5}$.

To find the S^3VS for the game, we need to first compute the value sets for all proper coalitions S. To allow for comparison, we represent elements of these value sets as vectors in $\mathbb{R}^5$ instead of $\mathbb{R}^{|S|}$. The values for players not in the coalition in question are set to zero. Each vector contains payoffs ordered according to: m_1, m_2, m_3, f_1, f_2.

$$V_{\{m_1,m_2,f_1,f_2\}} = \left\{ \begin{array}{ll} (\ 5 \quad 5 \quad 0 \quad 0 \quad 0\ ), & (\ 0 \quad 0 \quad 0 \quad 5 \quad 5\ ), \\ (\ 5 \quad 0 \quad 0 \quad 0 \quad 1\ ), & (\ 0 \quad 5 \quad 0 \quad 1 \quad 0\ ), \\ (\ 1 \quad 0 \quad 0 \quad 5 \quad 0\ ), & (\ 0 \quad 1 \quad 0 \quad 0 \quad 5\ ), \\ (\ 5\frac{1}{5} \quad 0 \quad 0 \quad 0 \quad 0\ ), & (\ 0 \quad 5\frac{1}{5} \quad 0 \quad 0 \quad 0\ ), \\ (\ 0 \quad 0 \quad 0 \quad 5\frac{1}{5} \quad 0\ ), & (\ 0 \quad 0 \quad 0 \quad 0 \quad 5\frac{1}{5}\ ) \end{array} \right\} \tag{3.20}$$

$$\begin{array}{lll} V_{\{m_3,f_1\}} = V_{\{m_2,m_3,f_1\}} & = V_{\{m_1,m_3,f_1\}} & = V_{\{m_1,m_2,m_3,f_1\}} \\ = V_{\{m_3,f_1,f_2\}} & = V_{\{m_2,m_3,f_1,f_2\}} & = V_{\{m_1,m_3,f_1,f_2\}} \\ & & = \{(\ 0 \quad 0 \quad 1 \quad 1 \quad 0\ )\} \end{array} \tag{3.21}$$

$$V_S = \{(\ 0 \quad 0 \quad 0 \quad 0 \quad 0\ )\} \text{ for all other } S \subset \{m_1, m_2, m_3, f_1, f_2\} \tag{3.22}$$

Consider the first element of $V_{\{m_1,m_2,f_1,f_2\}}$. It is obtained from a matching plan which alternates forever between the male-preferred $\{\mu(m_1) = f_1,\ \mu(m_2) = f_2\}$ and the female-preferred $\{\mu(m_1) = f_2,\ \mu(m_2) = f_1\}$, beginning with the male-preferred matching. The second element of $V_{\{m_1,m_2,f_1,f_2\}}$ is associated with the same matching plan, except beginning with the female-preferred matching.[16] The third element derives from a plan with the following first-period matching: $\{\mu(m_1) = f_1,\ \mu(m_2) = m_2\}$ and thereafter alternating between the male and the female-preferred matchings, beginning with the female-preferred one. The next three vectors come from similar plans. The seventh element comes from the following two matchings: $\{\mu(m_1) = f_1,\ \mu(m_2) = m_2\}$ and $\{\mu(m_1) = m_1,\ \mu(m_2) = f_1\}$, followed by alternating male and female-preferred matchings, beginning with the male preferred. This yields the following sequence of period payoffs:

	1	2	3	4	...
m_1	5	0	5	−1	...
m_2	0	−1	5	−1	...
m_3	0	0	0	0	...
f_1	−1	5	−1	5	...
f_2	0	0	−1	5	...

(3.23)

which has the required present value. The final three vectors in $V_{\{m_1,m_2,f_1,f_2\}}$ are obtained using analogous plans. It can be verified that these plans all satisfy the requirements for S^3. The S^3 matching plans for this four-player coalition is easy to derive because all proper subsets of $\{m_1, m_2, f_1, f_2\}$ have singleton S^3 sets, namely

[16]Switching partners is sustainable here because the game is infinite. A backward recursion argument does not apply.

to remain unmatched. Thus we only have to ensure that the present value of continuation payoffs are positive at every point in time. The S^3VS's for the remaining coalitions in equations (3.21) and (3.22) are obvious.

Now, consider any matching plan $\boldsymbol{\mu}$ for the grand coalition which involves, at some point in time, a match between $i \in \{m_1, m_2\}$ and $j \in \{f_1, f_2\}$. This match will yield one of the two players a period payoff of -1. Take, for example, a match between m_1 and f_2, where m_1 receives a payoff of -1. Given the low discount factor of $\frac{1}{5}$, m_1 has to receive a payoff of 5 in the next period in order for the matching plan to satisfy S^3. So m_1 must be matched with f_1 giving f_1 a stage payoff of -1. It can be shown that starting with any $i \in \{m_1, m_2\}$ and $j \in \{f_1, f_2\}$, a candidate S^3 plan will eventually lead to a payoff of -1 for f_1. Let t be the time at which this occurs. It can further shown that the most favorable sequence of matchings for f_1 which is consistent with S^3 is:

	t	$t+1$	$t+2$	$t+3$	$t+4$	...
m_1	5	0	0	-1	5	...
m_2	0	-1	5	0	0	...
m_3	0	0	$\frac{4}{5}$	$\frac{4}{5}$	0	...
f_1	-1	5	$\frac{4}{5}$	$\frac{4}{5}$	-1	...
f_2	0	0	-1	5	0	...

(3.24)

which implies that $\boldsymbol{\pi}(\boldsymbol{\mu}|_t)^{\{f_1\}} \leqslant \frac{1}{26}$. However, a coalition $\{m_3, f_1\}$, deviating from the game forever can achieve a discounted payoff of $\boldsymbol{\pi}_{\{m_3,f_1\}}(.) = (\ 1 \quad 1\ )$. We therefore conclude that any S^3 plan must involve $\{m_1, m_2, f_2\}$ remaining single forever.

The only remaining possible matching plan is between f_1 and m_3. But again a S^3 plan which matches f_1 and m_3 all the time would be blocked by the coalition

$\{f_1, f_2, m_1, m_2\}$ playing any plan in its S^3 set that gives f_1 a payoff greater than 1. Obviously, a plan in which all agents remain single in all periods is blocked. So no plans are consistent with S^3 for this game.

3.6 Comparing Alternative Notions of Stability

In the process of defining S^3, we have already discussed the relationship between the core, the recursive core, and S^3. To summarize, the recursive core imposes the core condition at each point in time. S^3 imposes, in addition, that deviating coalitions be self-enforcing.

We can define a fourth concept, namely one which requires that coalitions be self-enforcing but which does not require dynamic consistency; call this set of matching plans, the *self-enforcing core.*

Definition (Self-Enforcing Core).

(1) *For a coalition of 1 player* $\{i\}$, *the matching plan* $\boldsymbol{\mu}_{\{i\}} : \mathbb{N}_T \rightarrow \mathcal{M}_{\{i\}}$, *which specifies that the player remain single forever, is in the* self-enforcing core.

(2) *Assume that the self-enforcing core has been defined for all proper sub-coalitions* $C \subset S$. *A matching plan* $\boldsymbol{\mu}_S : \mathbb{Z}_T \rightarrow \mathcal{M}_S$ *is in the* self-enforcing core *if:*

(a) *There does not exist a matching plan* $\boldsymbol{\mu}_C : \mathbb{Z}_T \rightarrow \mathcal{M}_C$ *which is in the self-enforcing core for the set of player in* C, *such that:*

$$\boldsymbol{\pi}_C(\boldsymbol{\mu}_C) > \boldsymbol{\pi}_S(\boldsymbol{\mu}_S)^C \tag{3.25}$$

for some proper subset $C \subset S$.

(b) *There does not exist another matching plan* $\boldsymbol{\mu}'_S : \mathbb{N}_T \rightarrow \mathcal{M}_S$, *satisfying condition (a), such that:*

$$\boldsymbol{\pi}_S(\boldsymbol{\mu}'_S) > \boldsymbol{\pi}_S(\boldsymbol{\mu}_S) \tag{3.26}$$

Now, these four concepts of stability can be summarized by the table below.

	Does Not Impose Time Consistency		*Imposes Time Consistency*
Does Not Impose Self-Enforceability	CORE Non-empty Pareto optimal	$\supseteq$	RECURSIVE CORE May be empty Pareto optimal
	$\subseteq$		$\subseteq$
Imposes Self-Enforceability	SELF-ENFORCING CORE Non-empty May not be Pareto optimal		S^3 Non-empty under conditions May not be Pareto optimal

Some of the following properties have already been mentioned, but for the sake of clarity, we re-express them in the following propositions and remarks.

Proposition 9 (Core and Self-Enforcing Core). *The core is a subset of the self-enforcing core.*

Proposition 10 (Recursive Core and S^3). *The recursive core is a subset of the S^3 set.*

Proposition 11 (Core and Recursive Core). *The recursive core is a subset of the core.*

Remark (Self-Enforcing Core and S^4). *The self-sustaining stable set is not necessarily a subset of the self-enforcing core.*

Remark (Non-Pareto Optimality). *In contrast to the core and the recursive core, the S^3 set can contain matching plans that are not Pareto optimal.*[17]

The first two propositions are not particularly surprising. The requirement of self-enforceability limits the feasible set of coalitional deviations and so enlarges the set of matching plans which are "stable". The third proposition is also obvious. The imposition of dynamic consistency reduces the set of "stable" matching plans. The lack of an inclusion result between S^3 and the self-enforcing core is perhaps more surprising since S^3 imposes dynamic consistency in addition to self-enforceability. The explanation lies in the interaction between the two requirements. Although time consistency tends to reduce admissible matching plans for the grand coalition, it also limits the set of coalitional deviations which are possible. This is true because coalitions are also required to satisfy dynamic consistency. The first effect tends to make the S^3 set smaller relative to the self-enforcing core, while the second tends to make it larger. Of course one could define yet another notion of stability which imposes

---

[17]Note that this remark does not contradict condition (c) in the definition of admissibility. That condition simply imposes that no element that satisfies S^3 be dominated *by another element that satisfies* S^3.

time consistency on the original grand coalition, but not on blocking coalitions; this is not pursued here.[18]

In the second remark, the Pareto optimality of the recursive core stems from the fact that it is a refinement/subset of the core. The non-Pareto optimality of S^3 is illustrated by the example in equation (3.8) where the unique plan that is consistent with S^3, μ^{single}, is Pareto dominated by μ^{core}

In a static game, the recursive core is equivalent to the core, and S^3 to the self-enforcing core, because dynamic consistency is trivially satisfied. In the context of a two-sided, static, matching market, proposition 3 implies that all four sets are equivalent. The ineffectiveness of self-enforceability in modifying the core is not true in a general static game.

Finally, there is no inclusion relation between the core and S^3. Compared with the core, S^3 restricts continuation payoffs. Thus S^3 has a tendency to eliminate elements from the core. On the other hand, S^3 also restricts candidate blocking plans which tends to admit matching plans not in the core.

For completeness, we state the existence results for the other stability concepts.

Theorem 12 (Existence of the Core). *The core is always non-empty in a finitely, as well as infinitely-repeated, marriage market.*

---

[18] These inclusion results apply also to the non-cooperative notions of: strong Nash, perfect strong, coalition proof, and perfect coalition proof equilibrium.

Proof. It is not difficult to show that the matching game is *balanced*, so that we can apply directly Scarf's (1967) proof for the existence of the core. ■

Proposition 13 (Non-existence of the Recursive Core). *The recursive core may be empty in finitely, as well as infinitely-repeated, one-to-one, two-sided matching markets.*

However, in both finitely and infinitely-repeated games, the S^3 set may be non-empty even when the recursive core is empty.

Theorem 14 (Existence of Self-Enforcing Core). *The self-enforcing core is always non-empty in a finitely, as well as an infinitely-repeated, one-to-one, two-sided matching market.*

Proof. This follows immediately because the core is a subset of the self-enforcing core. ■

3.7 Extensions

The definition of S^3 can be extended without too much difficulty to matching markets where the value of some matches are only known after trade has taken place. We assume that information is common knowledge and once revealed is never forgotten. We model information as a *state* θ in Θ and we use the following terminology. If θ' is a *successor* to θ, then it contains weakly more information. If θ' is a successor to

θ, and it is not equal to θ, then θ' contains strictly more information than θ.

A matching plan is now a mapping from both time and a state of information onto the set of feasible matchings. The definition of S^3 is again inductive through time and the size of the coalition, but now we add the dimension of information to this recursion.

Definition (Self-sustaining Stability Under Incomplete Information).

(1) Self-sustaining stability *has been defined for all coalitions at information nodes with complete information.*

(2) *Assume that we are at an information node* $\theta \in \Theta$ *which has successors* $\Theta' \cup \{\theta\}$ *where* $\Theta' \subset \Theta$ *does not contain* θ. *For a coalition of one player,* $\{i\}$, *the matching function* $\boldsymbol{\mu}^{\theta}_{\{i\}} : (\Theta' \cup \{\theta\}) \times \mathbb{N}_T \rightarrow \mathcal{M}_{\{i\}}$, *which specifies that the player remain single at every information node, and at every point in time, satisfies* self-sustaining stability.

(3) *Assume that self-sustaining stability for all coalitions at successor information nodes,* Θ', *has been defined. Further assume that self-sustaining stability at node* θ *has been defined for all proper coalitions,* $C \subset S$. *A matching function* $\boldsymbol{\mu}^{\theta}_{S} : (\Theta' \cup \{\theta\}) \times \mathbb{N}_T \rightarrow \mathcal{M}_S$ *satisfies* self-sustaining stability *if:*

(a) *There does not exist a matching function* $\boldsymbol{\mu}^{\theta}_{C} \mid_{\theta',t}$ *which satisfies self-sustaining stability for coalition* C *from time* $t \in \mathbb{N}_T$ *and information node* $\theta' \in (\Theta' \cup \{\theta\}$ *onward, such that*

$$\boldsymbol{\pi}_C(\boldsymbol{\mu}^{\theta}_{C} \mid_{\theta',t}) > \boldsymbol{\pi}_S(\boldsymbol{\mu}^{\theta}_{S} \mid_{\theta',t})^C \tag{3.27}$$

for some proper subset $C \subset S$ *and for some* $(\theta', t) \in (\Theta' \cup \{\theta\}) \times \mathbb{N}_T$.

(b) *There does not exist another* $\tilde{\boldsymbol{\mu}}_S^{\theta}$, *satisfying the condition above, such that*

$$\boldsymbol{\pi}_S(\tilde{\boldsymbol{\mu}}_S^{\theta} \mid_{\theta,t}) > \boldsymbol{\pi}_S(\mu_S^{\theta} \mid_{\theta,t}) \tag{3.28}$$

at some $(\theta, t) \in (\Theta' \cup \{\theta\}) \times \mathbb{N}_T$.

We have yet to examine the properties of this definition.

Despite its advantages, S^3 still has a number of significant weaknesses. As the example of non-existence in equation (3.11) shows, one may want to allow for deviating subsets to rejoin the coalition under consideration. This would limit the set of feasible deviations and may provide a concept which exists under more general conditions. A related criticism is that in considering sub-coalitions which deviate from a coalitional deviation, we do not allow for the possibility that these sub-coalitions can consist of both members of the coalition and those outside the coalition. Allowing for this possibility is appealing since there seems to be no satisfactory reason to limit sub-coalition plans in the way that we do. However, the major disadvantage is that it eliminates the recursive structure of S^3. Both of these criticisms apply to the concept of coalition-proof Nash equilibria. We are currently thinking about how they can be incorporated into the definition or ameliorated.

3.8 Conclusion

In this paper, we introduced a new concept of equilibrium for cooperative games and illustrated it within the context of repeated matching markets. The concept imposes time-consistency and self-enforceability, together with the standard core idea. We

believe that this self-sustaining stable set is a more appropriate idea of "stability" than both the core and the recursive core in a dynamic environment. Although the set is empty in some games, it does exist in many cases where the recursive core does not.

3.9 Appendix: Computing the S^3VS in Infinitely-Repeated Markets

In applications, it is important to have an algorithm capable of computing the S^3VS. In a finitely-repeated game, we have already shown how backward induction can be used to compute this set. This section provides a method for computing the value set in an infinitely-repeated marriage market with discounting. The algorithm, as in the finite-horizon case, has a recursive structure. We begin by computing the S^3VS for single agents, and then proceed with larger and larger coalitions until the grand coalition is reached.

For each of these coalitions, we need to compute an approximation to the fixed point given by equation (3.16). Essentially, the approach is to use value-iteration of the S^3 mapping in equation (3.15). The algorithm will differ from standard value iteration in two important ways. First, the mapping used is set-to-set, rather than function-to-function. Second, the mapping does not involve a standard maximization or supremum operator but rather the requisite of non-blocking by coalitions. As will be seen, the second fact turns out to be helpful because it implies that a Pareto criterion can be used to reduce the number of points for which we have to apply the mapping.

The algorithm outlined here has many similarities with the procedure described in APS (1990) for finding the value set associated with sequential equilibria in infinitely-repeated games. In fact, our Γ mapping exhibits many of the same properties as the "B" operator in APS. For the approach to be valid we need to show that iterating the

S^3 mapping from a sufficiently large initial value set will lead to the largest fixed-point of the Γ mapping. The following theorem establishes this claim and is analogous to theorem 5 in APS (1990).

Define the set of feasible payoff vectors, for a coalition S, to be:

$$U_S \equiv [1/(1-\beta)]\,\mathrm{co}\left[\begin{array}{cccc} \pi(\mu_1) & \pi(\mu_2) & \ldots & \pi(\mu_{|\mathcal{M}_S|}) \end{array}\right] \tag{3.29}$$

where $\{\mu_1, \mu_2, \ldots, \mu_{|\mathcal{M}_S|}\} = \mathcal{M}_S$. The value set, V_S is obviously a subset of U_S. This set, together with the Euclidean norm, $(U_S, \|.\|)$ represents a complete metric space. Let $\mathcal{H}(U_S)$ be the Hausdorff space: that is the collection of all compact subsets of U_S. It can be shown that the pair $[\mathcal{H}(U_S), h]$, with h being the Hausdorff metric, is also a complete metric space.

$$h(X, Y) = \max\left\{\max_{x\in X}\min_{y\in Y}\|x-y\|, \max_{y\in Y}\min_{x\in X}\|x-y\|\right\} \tag{3.30}$$

Theorem 15 (Theorem 5, APS 1990). *Define* $W_0 = U_S$ *and for* $n = 1, 2, \ldots$, *let* $W_n = \Gamma(W_{n-1})$. *Then* $\{W_n\}$ *is a decreasing sequence and* $V_S = \lim_{n\to\infty} W_n$.

Proof. See APS (1990) lemmata 1 through 3 and theorems 4 and 5. ■

3.9.1 The Algorithm

Because the recursion through the size of the coalition has already been described in the dynamic-programming definition of S^3, the discussion here will focus on the iteration required to calculate an approximation to V_S, assuming that value sets V_C, have been computed for all $C \subset S$.

Step 1. Approximate U_S by a grid of points Ξ.

Step 2. Let $(\Xi_1, \Xi_2, ..., \Xi_p)$ be a partition of Ξ such that, for all $j \in \{1, 2, ..., p-1\}$, there does not exists a $\xi \in \Xi_{j+1} \cup ... \cup \Xi_p$ and a $\xi' \in \Xi_j$, such that $\xi > \xi'$. Put more simply, Ξ_j is the Pareto frontier of $\Xi_j \cup ... \cup \Xi_p$.[19] Now, begin with $j = 1$ and normalize a set $\Xi' = \emptyset$.

Step 3. For each $\xi \in \Xi_j$, compute $\Gamma_S(\{\xi\})$. If $\Gamma_S(\{\xi\}) = \emptyset$, eliminate from the grid all points $\xi' \in \Xi_j \cup ... \cup \Xi_p$, for which $\xi > \xi'$. Otherwise, add $\Gamma_S(\{\xi\})$ to Ξ'.

Step 4. Repeat Step 3 for each $j \in \{2, ..., p\}$.

Step 5. If $\Xi' = \emptyset$, then the value set is empty and the algorithm stops. If $h(\Xi, \Xi')$ is sufficiently small, the approximation to the value set is the Pareto frontier of $\Xi'$and the algorithm stops. Otherwise set $\Xi = \Xi'$ and return to Step 2.

This algorithm finds the largest fixed point of the Γ_S, which is by definition the S^3VS for coalition S. Notice that in step 3, we use the fact that if $\xi \in \Xi$ is not in the S^3VS, then any element $\xi' \in \Xi$, which is dominated by ξ, cannot be in the S^3VS. This elimination substantially reduces computational time.

---

[19]The magnitude of p is obviously inversely related to the gridsize.

3.10 References

Abreu, D., D. Pearce and E. Stacchetti (1986), "Optimal Cartel Equilibria with Imperfect Monitoring", *Journal of Economic Theory* 39, 251-269.

Abreu, D., D. Pearce and E. Stacchetti (1990), "Toward a Theory of Discounted Repeated Games with Imperfect Monitoring", *Econometrica* 58(5), 1041-1063.

Becker, R.A. and S.K. Chakrabarti (1995), "The Recursive Core", *Econometrica* 63(2), 401-423.

Bernheim, B.D., B. Peleg and M.D. Whinston (1987), "Coalition-Proof Nash Equilibria: I: Concepts", *Journal of Economic Theory* 42, 1-12.

Gale, D. (1982). *Money in Equilibrium*. Cambridge University Press, Cambridge, England.

Gale, D. and L. Shapley (1962), "College Admissions and the Stability of Marriage", *American Mathematical Monthly*, 69, 9-15.

Knuth, D.E. (1976), *Marriages Stables*, Les Presses de l'Université de Montreal.

Roth, A.E. and M.A.O. Sotomayor (1990), *Two-sided Matching: A Study in Game-theoretic Modeling and Analysis*, Econometric Society Monographs, Cambridge University Press, Cambridge, England.

Rubinstein, A. (1980), "Strong Perfect Equilibrium in Supergames", *International Journal of Game Theory* 9, 1-12.

Scarf, H.E. (1967), "The Core of an N Person Game", *Econometrica* 35(1), 50-69.