Essays on Probabilistic Belief Revision and Updating

Chen Zhao

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Economics
Adviser: Faruk Gul

June 2017

Abstract

This dissertation studies probabilistic belief revision and updating. In the first chapter, we propose an axiomatic framework for belief revision when new information is of the form "event A is more likely than event B." Our decision maker need not have beliefs about the joint distribution of the signal she will receive and the payoff-relevant states. With the *pseudo-Bayesian* updating rule that we propose, the decision maker behaves as if she selects her posterior by minimizing Kullback-Leibler divergence (or, maximizing relative entropy) subject to the constraint that A is more likely than B. The two axioms that yield the representation are *exchangeability* and *symmetry*. Exchangeability is the requirement that the order in which the information arrives does not matter whenever the different pieces of information neither reinforce nor contradict each other. Symmetry requires that the decision maker be neutral when receiving two directly opposite signals. We show that pseudo-Bayesian agents are susceptible to recency bias and honest persuasion. We also show that the beliefs of pseudo-Bayesian agents communicating within a network will converge but that they may disagree in the limit even if the network is strongly connected.

In the second chapter, we focus on belief updating. We provide a framework for analyzing a range of well-documented non-Bayesian updating behaviors including base rate neglect, conjunction fallacy and disjunction fallacy. Our model links the concept of similarity in theoretical psychology with belief updating. We follow Kahneman and Tversky (1974) and assume that when attempting to respond to the question "How likely is A given B?", people mistakenly respond to the question "How representative is A of B (i.e. how similar are A and B)?" With a similarity-based updating rule the conditional probability of $A \cup C$ given B might be less than the conditional probability of A given B if $B \cap C = \emptyset$, simply because the pair of events $A \cup C$ and B differ more from each other. Our axioms yield a Cobb-Douglas weighted geometric mean of $P(A|B)$ and $P(B|A)$ as the behavioral conditional probability of A given B, where

P is the correct subjective probability and $P(\cdot|\cdot)$ is the Bayesian conditional of P. That is, we provide a model of behavioral decision makers who confuse these two conditional probabilities but have correct unconditional beliefs. This combination of correct priors and incorrect updating occurs often since in many experiments subjects are explicitly given the relevant prior probabilities.

In the third chapter we present the tools that we developed through the course of writing the second chapter. In particular, we extend the Anscombe-Aumman theorem of subjective probability to allow for general mixture operations. Applying our theorem, we characterize quasi-linear means with a simple condition that resembles the classic independence axiom. We show that within the framework introduced in the second chapter, in addition to our Cobb-Douglas similarity index, the condition also enables us to recover Tversky's similarity index, which is a weighted harmonic mean of $P(A|B)$ and $P(B|A)$.

Acknowledgements

Albeit a decision theorist, I have inevitably made many dumb decisions in my academic life. The only decision that I will never regret, however, is to have Faruk Gul and Wolfgang Pesendorfer as my advisors. Throughout the years, their wisdom have profoundly shaped my development as an economist. I am deeply indebted to them for their invaluable guidance and support.

I would like to thank Roland Benabou. The second chapter of this dissertation stems from a term paper under his supervision. I would not be able to complete this project without his sharp insights and continuous encouragement.

I also benefited a lot from discussions with Dilip Abreu and Stephen Morris. The theory group at Princeton is of no doubt the most knowledgeable and approachable one among all departments.

And finally, I thank my parents Xinlin Zhao, Qing Cheng and my wife Ning Yan for their endless and unconditional care along this long but fruitful journey.

To my dearest parents, Xinlin Zhao and Qing Cheng

And to my beloved wife, Ning Yan

Contents

List of Figures

Chapter 1

Pseudo-Bayesian Belief Revision

1.1 Introduction

People often receive unexpected information. Even when they anticipate receiving information, they typically do not form priors over what information they will receive. Seldom, if ever, do decision makers bother to construct a joint prior on the set of signals and payoff-relevant states: waiting for the signal to arrive before making assessments enables the decision maker to eliminate many contingencies and hence reduces her workload. In short, due to the informational and computational demands, most people cannot reason within the framework of Bayes' rule and even those who can, often find it more economical to avoid Bayesian thinking. This gap between the Bayesian model and the typical experience of most decision makers suggests that a model that resembles the latter more closely may afford some insights into many observed biases in decision making under uncertainty.

In this paper, we propose a framework for belief revision that is closer to the typical experience in two ways. First, the signal that the decision maker receives is explicitly about probabilities of payoff-relevant states. This feature enables our model to be prior-free. That is, the decision maker need not have beliefs about the joint distribution of the signal she will receive and the payoff-relevant states; she need not know the full set of possible realizations of the signal and most importantly, whether the information was anticipated or a surprise does not matter. Second, the information is qualitative; that is, the decision maker is told that some event A_1 is more likely than some other event A_2. This qualitative setting nests the usual form of information that specifies occurrences of events: that event A occurs is simply equivalent to the information that $\emptyset$ is more likely than the complement of A. In addition, with our representation, *quantitatve* information of the form "event A is $q-$times as likely as B" could also be translated into a collection of qualitative statements and be processed accordingly.

To motivate the model, consider the following example: Physician X regularly chooses between two surgical procedures for her patients. She believes that, while both procedures are effective treatments, A_1 results in less morbidity than A_2 so A_1 is better. This belief is based not only on her personal experience but also on her reading of the relevant literature analyzing and comparing the two procedures. A recent paper in a prestigious journal provides some new information on this issue. The paper tracks the experiences of hundreds of patients for one year post surgery. After controlling for various factors, the authors of the new study, contrary to Physician X's beliefs, *reject* the hypothesis that procedure A_1 results in less morbidity than A_2. Physician X is not a researcher but she does fully understand that the study provides contrary evidence to her long-held position.

What is a plausible heuristic for learning from such a message without having a joint prior over all possible results, methodologies of the study? How should she revise her assessment of the relative morbidity of the two procedures? Should she revise her ranking of other events and if so, how? Below, we present an axiomatic model of belief revision relevant to the kind of situation depicted in the example above. Our model offers the following answers to these questions: Physician X should revise her beliefs by moving probability mass from the states in which A_2 will result in morbidity but A_1 will not, to the states in which A_1 will result in morbidity but A_2 will not; she should move probability mass in proportion to her prior beliefs, just as much as necessary to equate the morbidity of A_1 and A_2. Hence, due to the conflict in the literature, Physician X should disregard morbidity when choosing between the procedures.

The primitive of our model is a nonatomic subjective probability μ on a set of payoff-relevant states S. The decision maker learns some qualitative statement $\alpha = (A_1, A_2)$ which means "event A_1 is more likely than event A_2." The statement may or may not be consistent with the decision maker's prior ranking of A_1 and A_2. A *one-step revision rule* associates a posterior μ^α, with every μ and each qualitative statement α. We assume that the decision maker does not change her beliefs when she hears something that she already believes; that is, $\mu^\alpha = \mu$ whenever $\mu(A_1) \geq \mu(A_2)$.

A decision maker equipped with a one-step revision rule can process any finite string of qualitative statements sequentially: each time the decision maker learns a new qualitative statement, she applies it to her current beliefs according to the one-step updating rule.

We impose two axioms on the revision rule to characterize the following formula:

$$\begin{aligned} \mu^\alpha = \quad & \arg\min_{\nu \ll \mu} d(\mu||\nu) \\ s.t. \quad & \nu(A_1) \geq \nu(A_2) \end{aligned}$$

where $\nu \ll \mu$ means ν is absolutely continuous with respect to μ and $d(\mu||\nu)$ is the Kullback-Leibler divergence (i.e., relative entropy) of ν with respect to μ; that is,

$$d(\mu||\nu) \equiv -\int_S \ln\left(\frac{d\nu}{d\mu}\right) d\mu.$$

Hence, μ^α is the unique probability measure that minimizes $d(\mu||\cdot)$ among all μ-absolutely continuous probability measures consistent with the qualitative statement (A_1, A_2).

To see how μ^α is derived, note that each statement $\alpha = (A_1, A_2)$ partitions the state space into three sets: the *good news region* $G_\alpha = A_1 \backslash A_2$; the *bad news region* $B_\alpha = A_2 \backslash A_1$ and the *remainder* $R_\alpha = S \backslash (G_\alpha \cup B_\alpha)$. Then, if $\mu(A_1) \geq \mu(A_2)$,

$$\mu^\alpha = \mu. \tag{1.1}$$

That is, the decision maker does not update when she hears what she already knows. If instead $\mu(A_1) < \mu(A_2)$ and $\mu(G_\alpha) > 0$, then,

$$\mu^\alpha(G_\alpha) = \mu^\alpha(B_\alpha) = \frac{\mu(G_\alpha \cup B_\alpha)}{2} \tag{1.2}$$

and for an arbitrary measurable event C,

$$\mu^\alpha(C) = \mu(C \cap R_\alpha) + \frac{\mu(C \cap G_\alpha)}{\mu(G_\alpha)}\mu^\alpha(G_\alpha) + \frac{\mu(C \cap B_\alpha)}{\mu(B_\alpha)}\mu^\alpha(B_\alpha). \tag{1.3}$$

Hence, the probability of states in the remainder stays the same while states in the good news and bad news regions share $\mu(G_\alpha \cup B_\alpha)$ in proportion to their prior probabilities.

Finally, if $\mu(A_2) > \mu(A_1)$ and $\mu(G_\alpha) = 0$, then

$$\mu^\alpha(C) = \frac{\mu(C \cap R_\alpha)}{\mu(R_\alpha)} \tag{1.4}$$

for all $C \in \Sigma$. Our decision maker, like her Bayesian counterpart cannot assign a positive posterior probability to an event that has a zero prior probability. Thus, she deals with the information that the zero-probability event G_α is more likely than the event B_α by setting the posterior probability of the latter event to zero and redistributing the prior probability $\mu(B_\alpha)$ evenly among the states in the remainder R_α, as if she is conditioning on the event R_α according to Bayes' rule. We will call the preceding equations (1.1)-(1.4) together the *pseudo-Bayes' rule*.

The last case above enables us to incorporate Bayesian updating as a special case of the pseudo-Bayes' rule: Suppose the prior μ is over $S \times I$ where I is the set of signals. Then, learning that signal $i \in I$ occurred amounts to learning $(\emptyset, S\times(I\backslash\{i\}))$; that is, the statement "the empty set is more likely than any signal other than i." Then, the preceding display equation yields Bayes' rule:

$$\mu^\alpha(C) = \frac{\mu(C \times \{i\})}{\mu(S \times \{i\})}$$

for $C \in \Sigma$.

The two axioms that yield the pseudo-Bayes' rule above are the following: first, *exchangeability*, which is the requirement that the order in which the information arrives does not matter whenever the different pieces of information neither reinforce nor contradict each other. Hence, $(\mu^\alpha)^\beta = (\mu^\beta)^\alpha$ whenever α and β are *orthogonal*. We define and discuss orthogonality in section 2. Second, *symmetry*, which requires the decision maker to be neutral when receiving two directly opposite pieces of information; that is, $\alpha = (A_1, A_2)$ and $\bar{\alpha} = (A_2, A_1)$ implies $(\mu^\alpha)^{\bar{\alpha}}(A_1) = (\mu^\alpha)^{\bar{\alpha}}(A_2)$.

In section 1.2, we prove that the pseudo-Bayes' rule characterized by equations (1.1)-(1.4) is equivalent to the axioms and is the unique solution to the relative entropy minimization problem above. In section 1.3, we discuss the related literature. In section 1.4, we analyze communication among pseudo-Bayesian agents and extend our model to allow for quantitative information. We show that pseudo-Bayesian agents are susceptible to recency bias and honest persuasion. We also show that the beliefs of pseudo-Bayesian agents communicating within a network will converge but that they may disagree in the limit even if the network is strongly connected. In section 1.5, we illustrate the relationship between pseudo-Bayesian updating and Bayesian updating. In particular, we prove a result that shows how the latter can be interpreted as a special case of the former when the state space is rich enough; that is, when each state identifies both a payoff relevant outcome and a signal realization. Then section 6 concludes.

1.2 Model

In the section we first describe the primitives of our model. Then, we introduce the orthogonality concept which plays a key role in our main axiom, *exchangeability*. We show that *exchangeability* and our other axiom *conservatism* together are equivalent to the pseudo-Bayes' rule and also to the aforementioned constrained optimization characterization.

Let Σ be a σ-algebra defined on state space S. We will use capital letters $A, B, C, \ldots$ to denote generic elements of Σ. The decision maker (DM) has a *nonatomic* prior μ defined on (S, Σ); that is, $\mu(A) > 0$ implies there is $B \subset A$ such that $\mu(A) > \mu(B) > 0$.[1]

The DM encounters a qualitative statement (A_1, A_2): "A_1 is more likely than A_2". She interprets this information as $\Pr(A_1) \geq \Pr(A_2)$. We will use greek letters α, β and γ to repectively denote news (A_1, A_2), (B_1, B_2) and (C_1, C_2). Given prior μ, we call $\alpha = (A_1, A_2)$ *credible* if $\mu(A_1) \neq 0$ or $\mu(A_2) \neq 1$ and assume that the DM simply ignores non-credible statements. Our DM, like her Bayesian counterpart cannot assign positive posterior probability to an event that has a zero prior probability; there is simply no coherent way to distribute probability within the previously-null event. Therefore, no posterior could embrace a non-credible $\alpha = (A_1, A_2)$: doing so would

[1] Since μ is nonatomic and countably additive, it is also *convex-ranged*; that is, for any $a \in (0, 1)$ and A such that $\mu(A) > 0$, there is $B \subset A$ such that $\mu(B) = a\mu(A)$. We will exploit this property throughout the paper to identify suitable events.

require either increasing the probability A_1 or decreasing the probability of A_2 both of which necessitates increasing the probability of some previously-null event.

We focus on weak information; that is, we only consider qualitative statements of the form "A_1 is more likely than A_2" and, for simplicity assume that the DM translates a statement like "A_1 is strictly more likely than A_2" to one of our qualitative statements "A_1' is more likely than A_2" for some $A_1' \subset A_1$. We assume that the DM's procedure of for translating all statements into credible qualitative statements is exogenous and objective. That is, we ignore factors that might influence how the DM evaluates information; factors such as effectiveness of the wording and trustworthiness of the source.

Let $\Delta(S, \Sigma)$ be the set of nonatomic probabilities on (S, Σ). The DM's information set before revision is an element of $\mathbb{I}(\Delta(S, \Sigma)) = \{(\mu, \alpha) | \alpha \text{ is credible given } \mu\}$.

Definition. *A function $r : \mathbb{I}(\Delta(S, \Sigma)) \to \Delta(S, \Sigma)$ is a* ***one-step revision rule*** *if $r(\mu, \alpha) \equiv \mu^\alpha = \mu$ whenever $\mu(A_1) \geq \mu(A_2)$.*

Hence, if the qualitative statement does not contradict the DM's prior, she will keep her beliefs unchanged. The DM has no prior belief over the qualitative statements that she might receive. Thus, she interprets a statement that conforms to her prior simply as a confirmation of her prior beliefs and leaves them unchanged.

Our model permits multiple stages of learning. A decision maker equipped with a one-step revision rule can process any finite string of statements, $\alpha_1, \alpha_2, \ldots, \alpha_n$ sequentially: let $\mu_0 = \mu$ and $\mu_k = \mu_{k-1}^{\alpha_k}$ for $k = 1, 2, \ldots, n$. Hence, after learning $\alpha_1, \alpha_2, \ldots, \alpha_n$, the DM's beliefs change from $\mu = \mu_0$ to μ_n. That is, each time the decision maker learns a new qualitative statement, she applies it to her current beliefs according to the one-step revision rule.

In section 1.4, we show that with our one-step revision rule, the DM could also process quantitive statements of the form "$\Pr(A) = q \Pr(B)$" where q is rational.

In particular, we show that such a quantitative statement could be interpreted as a sequence of qualitative statements to which the one-step revision rule is applicable.

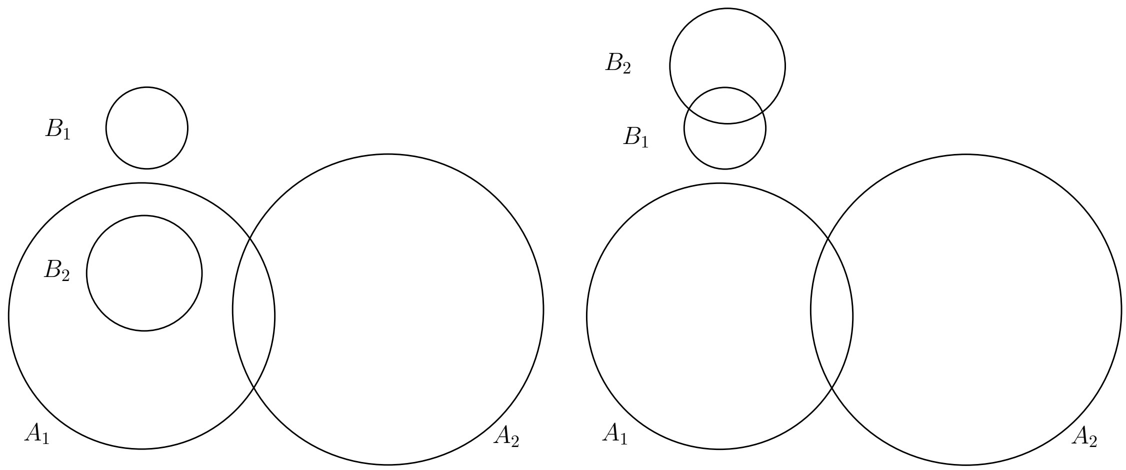
1.2.1 Orthogonality

Our main axiom, exchangeability, asserts that if two qualitative statements are *orthogonal* given prior μ; that is, if they neither reinforece nor contradict each other, then the order in which the DM receives these qualitative statements does not affect her posterior. In this subsection we provide a formal definition and discussion of this notion of orthogonality.

Since probabilities are additive, α conveys exactly the same information to the DM as $(A_1 \backslash A_2, A_2 \backslash A_1)$. Thus, each α partitions the event space into three sets: the good news region $G_\alpha = A_1 \backslash A_2$, the bad news region $B_\alpha = A_2 \backslash A_1$ and the remainder $R_\alpha = S \backslash (G_\alpha \cup B_\alpha)$.

If $\mu(G_\alpha) > 0$, we call $\Pi_\alpha = \{G_\alpha, B_\alpha, R_\alpha\}$ the *effective* partition generated by α. Also, we say that $D_\alpha = G_\alpha \cup B_\alpha$ is the *domain* of α since D_α contains all of the states that are influenced by α. If $\mu(G_\alpha) = 0$, the DM interprets G_α as a synonym for impossibility and therefore views α as equivalent to $(\emptyset, B_\alpha)$. Therefore, to accommodate α, the DM must lower the probability of B_α to zero and distribute the probability $\mu(B_\alpha)$ among the states in $S \backslash B_\alpha$. Hence, when $\mu(G_\alpha) = 0$, we call $\Pi_\alpha = \{G_\alpha \cup R_\alpha, B_\alpha\}$ the *effective* partition generated by α. Since every state in S is influenced by $(\emptyset, B_\alpha)$, the *domain* of α is S.

Our orthogonality concept identifies qualitative statements pairs that are neither conflicting with nor reinforcing each other. To understand what this means, consider Figure 1.1(a). Qualitative statement β demands that the probability of B_2 be decreased and hence be brought closer to that of B_1; α however requires the probability of A_1 and therefore B_2 to be increased and thus increasing the difference between



(a) α and β are in conflict (b) α and β are orthogonal

Figure 1.1: Orthogonality

the probability of B_2 and that of B_1. Therefore, Figure 1.1(a) depicts a situation in which α and β are in conflict and hence are not orthogonal.

In contrast, consider Figure 1.1(b): now β compares two subsets of R_α. Therefore, α does not affect the relative likelihood of any state in B_1 versus any state in B_2. In fact, $D_\alpha \cap D_\beta = \emptyset$; the domains of α and β do not overlap, and thus, α and β are orthogonal.

Similarly, if $\mu(G_\beta) > 0$ and β compares two subsets of B_α or two subsets of G_α; that is, if $D_\beta \subset G_\alpha$ or $D_\beta \subset B_\alpha$, α and β are again orthogonal since α affects both B_1 and B_2 equally.

The preceding observations motivate the following definition of orthogonality for two pieces of information:

Definition. *Let α and β be credible given μ. We say that α and β are **orthogonal** (or, $\alpha \perp \beta$) given μ if $D_\alpha \subset C \in \Pi_\beta$ or $D_\beta \subset C \in \Pi_\alpha$ for some C.*

1.2.2 Pseudo-Bayesian Revision

In the statement of the axioms $\mu^{\alpha\beta}$ represents the posterior after DM updates on α first and then on β, applying *the* one-step revision rule that we are axiomatizing.

Axiom 1.1. *(Exchangeability)* $\mu^{\alpha\beta} = \mu^{\beta\alpha}$ *if* $\alpha \perp \beta$ *given* μ.

Axiom 1 states that if qualitative statements α and β are orthogonal, then the sequence in which the DM updates does not matter. This concept of exchangeability closely resembles the standard exchangeability notion of statistics. In contrast to the standard notion, the one here only considers situations where the qualitative statements are orthogonal. When statements are not orthogonal we allow (but do not require) the revision rule to be non-exchangeable.

To state our next axiom, let $\bar{\alpha}$ denote (A_2, A_1), the opposite statement to α.

Axiom 1.2. *(Symmetry)* $\mu^{\alpha\bar{\alpha}}(A_1) = \mu^{\alpha\bar{\alpha}}(A_2)$.

Axiom 2 says that the DM, upon receiving two directly opposite statements regarding how A_1 compares to A_2, resolves the conflict by not taking either side; that is, when told $\Pr(A_1) \geq \Pr(A_2)$ and $\Pr(A_2) \geq \Pr(A_1)$, she concludes that $\Pr(A_1) = \Pr(A_2)$. This axiom does not impose exchangeability on opposite statements. Rather, it is only a restriction on the relative likelihood of A_1 versus A_2 and is silent on the DM's beliefs of any other event.

In a Bayesian framework, the joint prior over signals and states determines the content of a signal. Absent such a joint prior, our DM is ignorant about how the signals are drawn and therefore will not be able to distinguish between different signal sources.[2] In our leading example, there is only one way for Physician X to alleviate the tension between the new study and the previous literature without disregarding either: she can interpret that the morbidity difference between procedures A_1 and A_2 is insignificant, and therefore assume that the morbidity of the two procedures are comparable. When evaluating A_1 and A_2 in the future, Physician X will simply disregard morbidity due to this conflict in the literature.

[2] In the appendix, we present an intermediate case in which the DM has a slightly richer prior and is able to discount the content according to a subjective reliability assessment of the signal source.

Next, we state our main theorems. If $\mu(A_1) \geq \mu(A_2)$, the definition of a one-step revision rule requires $\mu^\alpha = \mu$; the theorem below characterizes the DM's learning behavior when $\mu(A_2) > \mu(A_1)$.

Theorem 1.1. *A one-step revision rule satisfies Axiom 1.1 and 1.2 if and only if for all $C \in \Sigma$,*

$$\mu^\alpha(C) = \begin{cases} \mu(C \cap R_\alpha) + \left(\frac{\mu(C \cap G_\alpha)}{\mu(G_\alpha)} + \frac{\mu(C \cap B_\alpha)}{\mu(B_\alpha)}\right) \cdot \frac{\mu(G_\alpha \cup B_\alpha)}{2}, & \text{if } \mu(G_\alpha) > 0, \\ \frac{\mu(C \cap R_\alpha)}{\mu(R_\alpha)}, & \text{otherwise.} \end{cases}$$

for any credible α given $\mu \in \Delta(\Sigma, S)$ such that $\mu(A_2) > \mu(A_1)$.

We call the formula in Theorem 1.1 together with $\mu^\alpha = \mu$ if $\mu(A_1) \geq \mu(A_2)$ the *pseudo-Bayes' rule*. Suppose $\mu(G_\alpha) > 0$, then

$$\mu^\alpha(G_\alpha) = \mu^\alpha(B_\alpha) = \frac{\mu(G_\alpha \cup B_\alpha)}{2}$$

and for $C \subset G_\alpha$, we have that

$$\mu^\alpha(C) = \frac{\mu(C \cap G_\alpha)}{\mu(G_\alpha)} \cdot \frac{\mu(G_\alpha \cup B_\alpha)}{2} = \mu(C|G_\alpha) \cdot \mu^\alpha(G_\alpha)$$

where the conditional probability is defined as in Bayes' rule. The $C \subset B_\alpha$ case is symmetric. Hence, the probability of states in the remainder stays the same while states in the good news and bad news regions share $\mu(G_\alpha \cup B_\alpha)$ in proportion to their prior probabilities.

If $\mu(G_\alpha) = 0$, then the DM sets the posterior probability of B_α to zero and redistributes the prior probability $\mu(B_\alpha)$ evenly among the states in the remainder R_α, as if she is conditioning on the event R_α according to Bayes' rule. This case enables us to incorporate Bayesian revision as a special case of the pseudo-Bayesian rule: suppose the prior μ is over $S \times I$ where I is a set of signals. Then, learning that signal i occurred amounts to learning $(\emptyset, S \times (I \backslash \{i\}))$; that is, the statement "any signal other than i is impossible." We formally present this result and discuss how our pseudo-Bayes' rule relates to Bayesianism in section 1.5.

Like her Bayesian counterpart, our DM cannot assign a positive posterior probability to an event that has a zero prior probability; that is, $\mu(A) = 0$ implies $\mu^\alpha(A) = 0$, or equivalently, μ^α is absolutely continuous with respect to μ, denoted as $\mu^\alpha \ll \mu$. To see why the pseudo-Bayes' rule implies absolute continuity, assume for now $\mu(G_\alpha) > 0$. For any event $C \in \Sigma$ such that $\mu(C) = 0$, $C \cap R_\alpha$ is clearly null according to the posterior since we have kept the probability distribution over R_α unchanged. The set $C \cap G_\alpha$ is also null in the posterior by the previous display equation and similarly so is $C \cap B_\alpha$. The case where $\mu(G_\alpha) = 0$ is trivial since it resembles conditioning on R_α according to Bayes' rule.

The following two examples illustrate how pseudo-Bayes' rule works.

Dice Example.[3] Suppose that the DM initially believes that the dice is fair and then encounters the qualitative statement $\alpha = (\{1, 2\}, \{2, 3, 4\})$. First, she eliminates $\{2\}$ from both A_1 and A_2 and hence identifies α with $(\{1\}, \{3, 4\})$. Keeping her beliefs on $\{2, 5, 6\}$ unchanged, she then moves probability from $\{3, 4\}$ proportionately to $\{1\}$ just enough to render the two sets equiprobable. Hence, her posterior probabilities of the states $(1, 2, \ldots, 6)$ are $(\frac{1}{4}, \frac{1}{6}, \frac{1}{8}, \frac{1}{8}, \frac{1}{6}, \frac{1}{6})$. If the DM had instead received the qualitative statement $(\emptyset, \{2\})$, her posterior would have been $(\frac{1}{5}, 0, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$. If the DM hears $(\{7\}, \{2\})$, she interprets it as $(\emptyset, \{2\})$.

Uniform Example. Suppose the DM has a uniform prior on $[0, 1]$ and encounters the qualitative statement $(\{0 < x < 0.2\}, \{0.6 < x < 1\})$. Then, the density of her posterior will be the step function depicted in Figure 1.2. That is, the density at states in $(0.2, 0.6)$ will remain unchanged and mass will shift proportionally from the

[3]The discrete probability space here should be viewed as a partition of the state space S with a nonatomic prior; that is, each state should be viewed as an event in S.

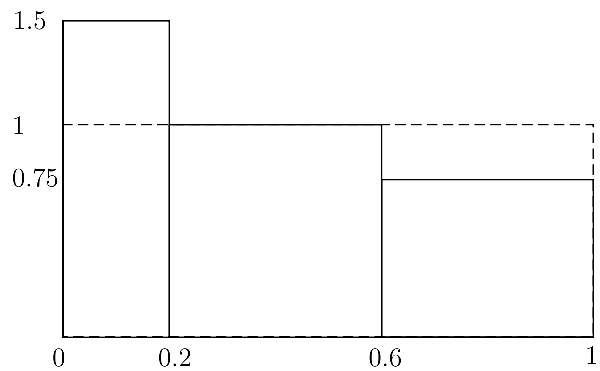


Figure 1.2: Uniform prior and $\alpha = (\{0 < x < 0.2\}, \{0.6 < x < 1\})$.

interval $(0.6, 1)$ to the interval $(0, 0.2)$ just enough to make the probabilities of $(0, 0.2)$ and $(0.6, 1)$ equal.

Alternatively, we could express the formula in Theorem 1.1 as the unique solution to a constrained optimization problem. This way of describing the pseude-Bayesian revision rule reveals that the DM's posterior is the closest probability distribution from the prior that is consistent with the new information. The notion of closeness here is Kullback-Leibler divergence, defined below. For $\mu, \nu \in \Delta(S, \Sigma)$ such that $\nu \ll \mu$, the Radon-Nikodym derivative of ν with respect to μ, denoted $\frac{d\nu}{d\mu}$, is defined as the measurable function $f : S \to [0, \infty)$ such that $\nu(A) = \int_A f d\mu$ for all $A \in \Sigma$. By the Radon-Nikodym Theorem, such an f exists and is unique up to a zero μ-measure set.

Definition. *For $\mu, \nu \in \Delta(S, \Sigma)$ such that $\nu \ll \mu$, the Kullback-Leibler divergence (henceforth KL-divergence) from μ to ν is given by*

$$d(\mu||\nu) \equiv -\int_S \ln\left(\frac{d\nu}{d\mu}\right) d\mu.$$

It is well-known that for $\nu \ll \mu$, KL-divergence $d(\mu||\nu)$ always exists and is strictly convex in ν. Moreover $d(\mu||\nu) \geq 0$; which holds with equality if and only if $\mu = \nu$. See appendix for a proof of these properties. The function $-d(\mu||\nu)$ (sometimes $d(\mu||\nu)$) is also called the relative entropy of μ with respect to ν.

Theorem 1.2. *A one-step revision rule satisfies Axiom 1.1-1.2 if and only if*

$$\begin{aligned} \mu^\alpha = \quad & \arg\min_{\nu \ll \mu} d(\mu || \nu) \\ s.t. \quad & \nu(A_1) \geq \nu(A_2) \end{aligned} \tag{P}$$

for any credible α given $\mu \in \Delta(S, \Sigma)$.

All we have to prove here is that the unique solution to the constrained optimization (P) is the pseudo-Bayesian posterior μ^α. The concavity of the logarithm function renders moving probability mass in proportion to the prior the most economic way to revise beliefs. In other words, the optimal Radon-Nikodym derivative must be constant within each element of the effective partition. This observation allows us to reduce P to a readily analyzable finite-dimensional convex optimization problem similar to the one below:

$$\begin{aligned} & \min_{q_i \geq 0} - \sum_{i=1}^{3} p_i \ln \frac{q_i}{p_i} \\ s.t. \quad & q_1 \geq q_2 \\ & q_i = 0 \text{ if } p_i = 0 \\ & \sum_{i=1}^{3} q_i = 1 \end{aligned}$$

where p's are the prior probabilities of G_α, B_α and R_α and q's are their posterior counterparts. Then, the Kuhn-Tucker conditions ensure that our pseudo-Bayes' rule is the unique solution.

Note that we are minimizing $d(\mu||\nu)$ given prior μ, while in information theory the method of maximum relative entropy calls for the objective to be $d(\nu||\mu)$ given prior μ. We postpone the discussion of how the two procedures are related to section 3, where we provide a survey of the related literature. Next, we provide a sketch of the proof of Theorem 1.1.

1.2.3 Proof of Theorem 1.1

In this subsection, we describe the key steps of the proof of Theorem 1.1. A formal proof is provided in the appendix. Theorem 1.1 can be broken into three parts: the first says that if the DM receives an α that contradicts her prior, she sets the likelihoods of A_1 and A_2 equal; the second says that upon receiving α the DM moves probability mass in proportion to the prior between elements of Π_α; the third says that learning α does not affect the probability of D_α.

The first part is implied by the symmetry axiom. Since α contradicts μ it must be the case that $\bar{\alpha}$ conforms to μ. By definition of a revision rule, $\mu = \mu^{\bar{\alpha}}$; that is, the DM behaves as if she receives an $\bar{\alpha}$ before α. Therefore, by the symmetry axiom the DM will set the likelihoods of A_1 and A_2 equal.

To see why the second part is true, first note that if α and β are orthogonal given μ then $\mu(B_1) \geq \mu(B_2)$ implies $\mu^\alpha(B_1) \geq \mu^\alpha(B_2)$. To see this, note that if $\mu^\alpha(B_2) > \mu^\alpha(B_1)$ then $\mu^{\alpha\beta} \neq \mu^\alpha$. But note that $\mu^\beta = \mu$ and, therefore, $\mu^{\beta\alpha} = \mu^\alpha$. Exchangeability requires that $\mu^{\alpha\beta} = \mu^{\beta\alpha}$, delivering the desired contradiction. Thus, we conclude that for all B_1, B_2 that are nonnull sub-events of the same element of Π_α, the likelihood ranking cannot be affected by the arrival of α. Since μ is nonatomic, this, in turn, implies that the probability of sub-events in the same element of Π_α must be updated in proportion to the prior.

The preceding argument identifies μ^α for the case where A_1 is empty (that is, $\alpha = (A_1, A_2)$ says that A_2 has probability zero): to accommodate this information, the decision maker has to distribute the mass of A_2 proportionally on $G_\alpha \cup R_\alpha$; that is, she must behave as if she is conditioning of $S \backslash A_2$ according to Bayes rule.

Now, let $\beta = (\emptyset, B_2)$ and assume that A_1 and A_2 are both nonnull events contained in $S \backslash B_2$. Since α and β are orthogonal given μ it follows that the order of revision does not matter. This yields the following equation:

$$\frac{\mu^{\alpha}(D_{\alpha})}{1-\mu^{\alpha}(B_2)} = \mu^{\beta\alpha}(D_{\alpha})$$

We have established that the probability of all events within R_{α} must change in proportion to their prior, that is,

$$\mu^{\alpha}(B_2) = \frac{1-\mu^{\alpha}(D_{\alpha})}{1-\mu(D_{\alpha})}\mu(B_2).$$

Combining the equations above we have

$$\frac{\mu^{\alpha}(D_{\alpha})}{1-\frac{1-\mu^{\alpha}(D_{\alpha})}{1-\mu(D_{\alpha})}\mu(B_2)} = \mu^{\beta\alpha}(D_{\alpha}). \tag{1.5}$$

Next, we establish a connection between $\mu^{\alpha}(D_{\alpha})$ and $\mu^{\beta\alpha}(D_{\alpha})$ in order to turn (1.5) into a functional equation. Consider two priors μ_1 and μ_2 which differ only within B_2. Immediately we have $\mu_1^{\beta} = \mu_2^{\beta}$ and therefore $\mu_1^{\beta\alpha} = \mu_2^{\beta\alpha}$. On the other hand, since α and β are orthogonal given both priors, we expect that $\mu_1^{\alpha\beta} = \mu_2^{\alpha\beta}$ as well. Again due to the proportionate movement of probability mass, it must be the case that $\mu_1^{\alpha}(D_{\alpha}) = \mu_2^{\alpha}(D_{\alpha})$. Using an analogous argument, we show that $\mu^{\alpha}(D_{\alpha})$ can never depend on the prior distribution within R_{α}. Hence, $\mu^{\alpha}(D_{\alpha}) = g(\mu(\cdot|D_{\alpha}), \mu(D_{\alpha}))$; that is, $\mu^{\alpha}(D_{\alpha})$ is a function of $\mu(D_{\alpha})$ and how μ behaves within D_{α}. It follows that

$$\mu^{\beta\alpha}(D_{\alpha}) = g(\mu^{\beta}(\cdot|D_{\alpha}), \mu^{\beta}(D_{\alpha})) = g\left(\mu(\cdot|D_{\alpha}), \frac{\mu(D_{\alpha})}{1-\mu(B_2)}\right).$$

The observation above turns (1.5) into a functional equation. Let $\mu(B_2) = b$ and $\mu(D_{\alpha}) = x$. Holding $\mu(\cdot|D_{\alpha})$ constant and abusing the notations a little bit we arrive at

$$\frac{g(x)}{1-\frac{1-g(x)}{1-x}b} = g\left(\frac{x}{1-b}\right) \tag{1.6}$$

which holds for $0 < x < 1$ and $0 < b < 1 - x$ since we could vary B_2 within R_α continuously thanks to the convex-rangedness of μ. The solution to (1.6) is simply

$$g(x) = \frac{a}{2a - 1 + \frac{1-a}{x}}$$

where $a \in [0, 1]$. If $a > 1/2$, $g(x) > x$ which means that the probability of D_α increases if the DM learns α; if $a < 1/2$, $g(x) < x$ so D_α shrinks; if $a = 1/2$, the probability of D_α is unchanged.

The interaction between Axiom 1.1 and 1.2 then dictates that $a = 1/2$. To see that, suppose $a > 1/2$. Pick mutually exclusive C_1, C_2 such that $\mu(C_2) \geq \mu(C_1) > 0$ and $D_\alpha \subset C_2$. First note that $\mu^\alpha(C_2) > \mu^\alpha(C_1)$ since when D_α expands, R_α shrinks proportionately. Then by the conservatism axiom, $\mu^{\alpha\gamma}(C_1) = \mu^{\alpha\gamma}(C_2)$. On the other hand, also by conservatism $\mu^\gamma(C_1) = \mu^\gamma(C_2)$. Then if learning α increases D_α and therefore C_1, the DM will believe that $\mu^{\gamma\alpha}(C_1) > \mu^{\gamma\alpha}(C_2)$, which contradicts exchangeability since $\alpha \perp \gamma$ given μ. A similar argument holds if $a < 1/2$. Hence $a = 1/2$ and we have sketched the "only if" part of Theorem 1.1. We provide the full proof in the appendix.

1.3 Related Literature

Existing models of non-Bayesian revision in the economics literature consider a setting where Bayes rule is applicable but the DM deviates from it due to a bias, due to bounded rationality or due to temptation. By contrast, we assume that the agent uses Bayes rule when it is applicable but extends revision to situations in which Bayes rule does not apply. In particular, given a state space (S, Σ), Bayes' rule only applies to information of the form "$A \in \Sigma$ has occurred." While nesting such information as

$(\emptyset, S\backslash A)$, our qualitative setting also permits statements that are not events in the state space.[4]

For behavioral models of non-Bayesian revision, see, for example, Barberis, Shleifer, and Vishny (1998); Rabin and Schrag (1999); Mullainathan (2002a); Rabin (2002); Mullainathan, Schwartzstein, and Shleifer (2008); Gennaioli and Shleifer (2010). In the decision theory literature, Zhao (2016) formally links the concept of similarity with belief revision to generate a wide class of non-Bayesian fallacies. Ortoleva (2012) proposes a hypothesis testing model in which agents reject their prior when a rare event happens. In that case, the DM looks at a prior over priors and chooses the prior to which the rationally updated second-order prior assigns the highest likelihood. Epstein (2006) and Epstein, Noor, and Sandroni (2008) build on Gul and Pesendorfer (2001)'s temptation theory and show that the DM might be tempted to use a posterior that is different from the Bayesian update. All of the contributions above focus on non-Bayesian deviations when Bayes' rule is applicable while our DM sticks to Bayes' rule whenever possible. Our model weighs in when the DM encounters generic qualitative statements.

Our work is also related to the literature in information theory on maximum entropy methods. This literature aims to find universally applicable algorithms for belief revision when new information imposes constraints on the probability distribution. Papers in this literature typically posit a well-parameterized class of constrained optimization models. In particular, the statistician is *assumed* to be able to use standard Lagrangian arguments to optimize his posterior subject to the constraints. In contrast, we consider a choice-theoretic state space and start from a general mapping that assigns a posterior to each piece of information given a prior.

Within this literature, Caticha (2004) achieves the same constrained optimization representation as in our Theorem 1.2; that is, the statistician minimizes $d(\mu||\cdot)$

[4] Enlarging the state space will not help, since with a more general state space comes a much larger set of qualitative information. See section 5 for a detailed argument.

subject to newly-received constraints given prior μ. As is common in the literature, Caticha (2004) assumes that the statistician chooses his posterior by minimizing a smooth function defined on the space of probabilities. He then regulates this objective function with axioms that require, for example, invariance to coordinates changes. Because of the parametrization, it is difficult to translate his axioms into revealed preference statements.

In a similar vein, Shore and Johnson (1980), Skilling (1988), Caticha (2004) and Caticha and Giffin (2006) propose the method of maximum relative entropy[5] which minimizes $d(\cdot||\mu)$ instead of $d(\mu||\cdot)$. Karbelkar (1986) and Uffink (1995) relax Shore and Johnson (1980)'s axioms and show that all $\eta-$entropies[6] survive scrutiny as objective functions.

1.4 Recency Bias, Persuasion and Communication

In this section, we first show that pseudo-Bayesian agents are susceptible to recency bias, and that this bias could be mitigated by repeated learning. Our results imply that if the DM is presented repeatedly with true and representative qualitative statements, she will learn the correct distribution in the limit. We then extend our model to allow for quantitative information by showing that any quantitative statement can be translated into a representative collection of qualitative statements. After that, we consider a situation where an information sender knows the correct distribution of a random variable and intends to persuade a pseudo-Bayesian information receiver

[5]The concept of relative entropy originates from Kullback-Leibler divergence. When first introduced in Kullback and Leibler (1951), the KL-divergence between probability measures μ and ν is defined as $d(\mu||\nu) + d(\nu||\mu)$. Such a symmetric divergence is designed to measure how difficult it is for a statistician to discriminate between distributions with the best test.

[6]The form of η-entropy is given by

$$d_\eta(\nu||\mu) \equiv \frac{1}{\eta(\eta+1)} \left(\int_S \left(\frac{d\nu}{d\mu} \right)^\eta d\nu - 1 \right).$$

In the limit $d_0(\nu||\mu) = d(\nu||\mu)$ and $d_{-1}(\nu||\mu) = d(\mu||\nu)$. In fact, among all η-entropies, only $d(\mu||\cdot)$ ($\eta = -1$) is consistent with our axioms.

with true qualitative statements. We show that honest persuasion is almost always possible. Finally, we consider a network of pseudo-Bayesian agents who would like to reach consensus on a pair of qualitative statements. We show that agents' beliefs will always converge as they interact with each other but consensus might not be reached in the limit, even if the network is strongly connected.

1.4.1 Recency Bias and the Qualitative Law of Large Numbers

Our DM employs a step-by-step learning procedure and views each piece of new information as a constraint. For example, if she is given two possibly conflicting pieces of information, α and then β, her ultimate beliefs must be consistent with β but need not be consistent with α. More specifically, suppose there are three states $\{s_1, s_2, s_3\}$ and the prior is $(0.2, 0.3, 0.5)$.[7] Suppose the decision maker receives (s_2, s_3) (i.e., s_2 is more likely than s_3) then (s_1, s_2). After the first step, her posterior is $(0.2, 0.4, 0.4)$. Then the second statement induces the belief $(0.3, 0.3, 0.4)$. Notice that s_2 now has a lower probability than s_3; that is, the first statement has been "forgotten". However, the DM has not completely forgotten (s_2, s_3) since her ultimate posterior differs from what she would believe had she received only (s_1, s_2).

When the arrival of a second statement prompts the DM to "forget" the first-learned statement, we say that recency bias has been induced. A qualitative statement α is said to be *degenerate* given μ if $\mu(G_\alpha) = 0$. Due to the absolute continuity of our pseudo-Bayes' rule, once a degenerate statement is learned, it remains true in the DM's beliefs ever after. Therefore, we focus on statements that are nondegenerate given μ for the remainder of this subsection. Since our revision formula creates extra

[7] Again, the reader should view the three states as a partition of state space S with a nonatomic prior.

null events only if a degenerate statement is received, we will drop the qualification "given μ" for simplicity of exposition.

Definition. *We say that an ordered pair of nondegenerate statements* $[\alpha, \beta]$ *induces* ***recency bias*** *on* μ *if* $\mu^{\alpha\beta}(A_2) > \mu^{\alpha\beta}(A_1)$.

If $[\alpha, \beta]$ does not induce recency bias on μ, then clearly the DM has no problem incorporating α and β at the same time: all she needs to do is to learn α then β. The more interesting case is when recency bias occurs. In that case, the next proposition shows that the DM could never accommodate both α and β with finite stages of learning.

Proposition 1.1. *If* $[\alpha, \beta]$ *induces recency bias on* μ*, then* $[\beta, \alpha]$ *also induces recency bias on* μ^{α}.

In other words, if $\mu^{\alpha\beta}(A_2) > \mu^{\alpha\beta}(A_1)$ then $\mu^{\alpha\beta\alpha}(B_2) > \mu^{\alpha\beta\alpha}(B_1)$; that is if the DM learns β for a second time, she will forget about α, so on and so forth. As she learns the statements repeatedly, however, the DM will be able to accommodate both α and β in the limit.

Proposition 1.2. *If* $[\alpha, \beta]$ *induces recency bias on* μ*, then* $\mu^{(\alpha\beta)^n}$ *and* $\mu^{(\alpha\beta)^n\alpha}$ *converge in total variation to* μ^* *such that* $\mu^*(A_1) = \mu^*(A_2)$ *and* $\mu^*(B_1) = \mu^*(B_2)$.

Proposition 1.2 implies that repetition plays a role in learning. This is so since at each step the DM only partially forgets the statement that is biased against; like her Bayesian counterpart, her beliefs at any step are a function of the whole history of statements learned. In the Bayesian setting, however, hearing an old piece of information again does nothing to the decison maker's beliefs.

DeMarzo, Vayanos, and Zwiebel (2003) consider a setting where agents treat any information they receive as new and independent information. Since the agents do not adjust properly for repetitions, repeated exposure to an opinion has a cumulative effect on their beliefs. In our model, however, repetition plays a role only in the presence of contradictory information and, in particular, when recency bias occurs. If the DM learns a single statement repeatedly, she will simply stop revising her beliefs after the first time.

The next theorem generalizes Proposition 1.2 to any finite set $\{\alpha_1, \ldots, \alpha_n\}$ of nondegenerate statements. For $\mu, \nu \in \Delta(S, \Sigma)$, write $\mu \sim \nu$ if $\mu \ll \nu$ and $\nu \ll \mu$.

Definition. *Let $\alpha_i = (A_{i1}, A_{i2})$. We say that a set of nondegenerate qualitative statements $\{\alpha_i\}_{i=1}^n$ is **compatible** if there exists $\nu \sim \mu$ such that $\nu(A_{i1}) \geq \nu(A_{i2})$ for all i. Such ν is called a **solution** to $\{\alpha_i\}_{i=1}^n$.*

Put differently, $\{\alpha_i\}_{i=1}^n$ is compatible if the statements are sampled from a probability distribution ν which agrees with μ on zero-probability events. To mitigate possible recency bias, the DM will have to learn each statement often enough. Let $\{\pi_n\}$ be the sequence of statements that the DM learns.

Definition. *A sequence of nondegenerate qualitative statements $\{\pi_n\}$ is **comprehensive** for $\{\alpha_i\}_{i=1}^n$ if there is N such that $\{\alpha_i\}_{i=1}^n = \bigcup_{i=1}^N \pi_{kN+i}$ for all $k \geq 0$.*

That is, within each block of N steps, the DM learns each statement in $\{\alpha_i\}_{i=1}^n$ at least once. Notably, the frequency of each α_i in $\{\pi_n\}$ does not have to converge as $n \to \infty$. As long as the learning sequence $\{\pi_n\}$ is comprehensive, the DM's beliefs will converge.

Theorem 1.3. *(Qualitative Law of Large Numbers) Let $\{\alpha_i\}_{i=1}^n$ be compatible. If $\{\pi_n\}$ is comprehensive for $\{\alpha_i\}_{i=1}^n$, then $\mu^{\pi_1\pi_2\cdots\pi_n}$ converges in total variation to a solution to $\{\alpha_i\}_{i=1}^n$.*

The theorem implies that if the DM correctly identifies zero-probability events, she could digest any finite collection of objectively true qualitative statements by repeated learning. Note that if the DM assigns positive probability to some objectively zero-probability event A, she could easily correct her mistake by learning $(\emptyset, A)$. However, if the DM fully neglects certain probable event, she could never correct her mistake, for her posterior has to be absolutely continuous with respect to her prior.

By learning each α_i, the DM projects her beliefs onto the closed and convex set of probabilities that assign a weakly higher likelihood to A_{i1} than A_{i2}. Bregman (1966) proves that if the notion of distance is well-behaved, cyclically conducting projections onto a finite collection of closed and convex sets converges to a point in the intersection. Although our procedure P is not a Bregman-type projection, we are able to adapt Bregman (1966)'s proof to our situation.

The limit in Theorem 1.3 depends on the sequence $\{\pi_n\}$ in general.[8] However, if the collection of qualitative statements uniquely represents a distribution, the limit does not depend on $\{\pi_n\}$. To formally define representativeness, let $\bigvee_{j=1}^m \Pi_j$ denote the coarsest common refinement of partitions $\Pi_1, \Pi_2, \ldots, \Pi_m$ of S.

Definition. *We say that a compatible collection $\{\alpha_i\}_{i=1}^n$ is **representative** if*

$$\left.\begin{array}{l} \text{(i) } \nu, \nu' \text{ are solutions to } \{\alpha_i\}_{i=1}^n \\ \text{(ii) } \nu(\bigcup_{i=1}^n D_{\alpha_i}) = \nu'(\bigcup_{i=1}^n D_{\alpha_i}) \end{array}\right\} \Longrightarrow \nu(C) = \nu'(C) \text{ for all } C \in \bigvee_{i=1}^n \Pi_{\alpha_i}.$$

That is, a collection of statements $\{\alpha_i\}_{i=1}^n$ is representative if it uniquely identifies the likelihood ratios among all elements of $\bigvee_{i=1}^n \Pi_{\alpha_i}$ that is a subset of $\bigcup_{i=1}^n D_{\alpha_i}$. For an example let A, B, C, D be mutually exclusive and nonnull according to μ. Consider the collection

$$\{(A \cup B, C \cup D), (D, C), (C, B), (B, A)\}.$$

Let $\nu(A \cup B \cup C \cup D) = p > 0$. Then if each statement in the collection is true under ν it has to be the case that $\nu(A) = \nu(B) = \nu(C) = \nu(D) = p/4$. Hence the collection above is representative.

[8]For example, let $S = \{s_1, s_2, s_3\}$ and the DM's prior be $(0.1, 0.3, 0.6)$. Suppose $\alpha_1 = (s_1, s_2)$ and $\alpha_2 = (s_1, s_3)$. Then $\mu^{\alpha_1\alpha_2\cdots} = (0.4, 0.2, 0.4)$ but $\mu^{\alpha_2\alpha_1\cdots} = (0.35, 0.3, 0.35)$.

By Theorem 1.3, if $\{\alpha_i\}_{i=1}^n$ is representative, then the sequence in which the DM learns does not matter as long as it is comprehensive.

Corollary. *Let* $\{\alpha_i\}_{i=1}^n$ *be representative. If* $\{\pi_n\}$ *is comprehensive for* $\{\alpha_i\}_{i=1}^n$*, then* $\mu^{\pi_1\pi_2\cdots\pi_n}$ *converges in total variation to* μ^**, a solution to* $\{\alpha_i\}_{i=1}^n$*.*

In particular, if we present the DM with a representative collection of *correct* qualitative statements covering the wholes state space, i.e. $\bigcup_i D_{\alpha_i} = S$, the DM would eventually have correct beliefs on any event $C \in \bigvee_{i=1}^n \Pi_{\alpha_i}$. This corollary and the previous theorem establishes the legitimacy of our pseudo-Bayes' rule as a model of learning. In a word, truth could eventually be learned under very relaxed conditions.

With the preceding corollary, for any rational numbers q, the DM is in fact able to incorporate quantitative statements in the form of $\Pr(A) = q\Pr(B)$ where $A \cap B = \emptyset$. In particular, any quantitative statement of such kind could be interpreted as a representative collection of qualitative statements. For example, $\Pr(A) = 2\Pr(B)$ could be interpreted as "half of A is as likely as B" which boils down to the following collection of qualitative information:

$$\{(A, A\backslash A'), (A\backslash A', B), (B, A')\}$$

where $A' \subset A$ with $\mu(A') = \mu(A)/2$. If $\mu(A)\cdot\mu(B) > 0$, the collection is representative since any probability measure which embraces the collection has $\Pr(A') = \Pr(A\backslash A') = \Pr(B) = \Pr(A \cup B)/3$. By previous results the DM must hold the same believe in the limit if she learns this collection repeatedly in any comprehensive sequence. This limit, depends on neither the sequence nor how A' is chosen.

In general, we call $\{\alpha_i\}_{i=1}^{m+n}$ an *interpretation* of quantitative information $\Pr(A) = \frac{m}{n}\Pr(B)$ if it is given by

$$\{(A_1, A_2), (A_2, A_3), \ldots, (A_{m-1}, A_m), (A_m, B_1), (B_1, B_2), \ldots, (B_{n-1}, B_n), (B_n, A_1)\}$$

where $\{A_i\}_{i=1}^{m}$ and $\{B_i\}_{i=1}^{n}$ are respectively equal-probability partitions of A and B under μ. The interpretation $\{\alpha_i\}_{i=1}^{m+n}$ basically conveys the information "1/m of A is as likely as $1/n$ of B."

Our next corollary extends the pseudo-Bayes rule to allow for quantitative information based on the idea that each quantitative statement has an interpretation.

Corollary. *Suppose* $A \cap B = \emptyset$ *and* $\mu(A) \cdot \mu(B) > 0$. *Let* $\{\alpha_i\}_{i=1}^{m+n}$ *be an interpretation of quantitative information* $\Pr(A) = \frac{m}{n}\Pr(B)$ *with* $m, n \in \mathbb{N}_+$. *If* $\{\pi_n\}$ *is comprehensive for* $\{\alpha_i\}_{i=1}^{m+n}$, *then* $\mu^{\pi_1\pi_2\cdots\pi_n}$ *converges in total variation to* μ^* *such that*

$$\begin{aligned} \mu^* = \quad & \arg\min_{\nu \ll \mu} d(\mu || \nu) \\ s.t. \quad & \nu(A) = \frac{m}{n}\nu(B). \end{aligned}$$

As in Theroem 1.1 and 1.2, the DM in fact reallocates probability mass between A, B in proportion to her prior and never touches $S \backslash (A \cup B)$. The posterior of A will be exactly equal to m/n times the posterior of B.

1.4.2 Persuasion

In this subsection we apply our model to a persuasion problem. Let $X : S \to \mathbb{R}$ be a random variable. Suppose μ^* is the true probability measure on (S, Σ) while μ is the DM's beliefs. We assume that μ and μ^* are both *nonatomic*. The information sender, knowing the true distribution μ^* and also μ, intends to increase the DM's expectation on X but is bound to provide only correct qualitative statements.

Definition. *We say that the DM can be* ***persuaded*** *of X if there is α such that $\mu^*(A_1) \geq \mu^*(A_2)$ and $\mathbb{E}_{\mu^\alpha}(X) > \mathbb{E}_\mu(X)$. Such an α is said to be* ***persuasive*** *of X.*

The next theorem states that it is always possible to move the information receiver's expectation towards the correct direction, if the receiver believes that X has positive variance and bounded support.

Proposition 1.3. *Suppose $var_\mu(X) > 0$ and there exists M such that $\mu(\{|X| \leq M\}) = 1$. Then if $E_\mu(X) < E_{\mu^*}(X)$, the DM can be persuaded of X.*

To find a persuasive statement, the sender can partition $\mathbb{R}$ into equiprobable intervals according to the true distribution of X, then ask how likely that the receiver thinks each interval is. If the receiver assigns lower probability to interval I_1 than interval I_2 but I_1 is to the right of I_2 along the real axis, then $(\{X \in I_1\}, \{X \in I_2\})$ is persuasive. As the partition becomes finer, if such persuasive statements never exist, it must be that $E_\mu(X) \geq E_{\mu^*}(X)$, since the receiver always assigns higher probability to intervals with larger values.

If $E_\mu(X) \geq E_{\mu^*}(X)$, scope for persuasion might not exist. Consider the following situation: Let $X(s) = \mathbf{1}\{s \in A\}$ and $\mu(A) \geq \mu^*(A) = 1/2$. Also suppose that μ and μ^* have the same conditional probabilities on A and on $S \backslash A$. Picking a persuasive α amounts to picking $B, C \subset A$ and $B', C' \subset S \backslash A$ such that the following inequalities hold:

$$\begin{aligned}
\mu(B) + \mu(B') &< \mu(C) + \mu(C'), \\
\frac{\mu(B)}{\mu(A)} + \frac{\mu(B')}{1-\mu(A)} &\geq \frac{\mu(C)}{\mu(A)} + \frac{\mu(C')}{1-\mu(A)}, \\
\frac{\mu(B)}{\mu(B)+\mu(B')} &> \frac{\mu(C)}{\mu(C)+\mu(C')}.
\end{aligned}$$

The first two inequalities state that $(B \cup B', C \cup C')$ is a surprise so the statement prompts the DM to move probability from $C \cup C'$ to $B \cup B'$. The third inequality ensures that such revision enhances the receiver's expectation of X. The first two inequalities together imply that $\mu(B') > \mu(C')$ and $\mu(C) > \mu(B)$, which contradicts the third. Therefore, no qualitative statements could increase the receiver's expectation of X.

In Kamenica and Gentzkow (2011), the signal sender is able to design the informational structure and therefore the joint prior of the signal receiver. In contrast, we allow the sender and receiver to disagree on how signals are generated (μ and μ^* could be different joint priors) but permit the sender to send qualitative statements, not just signals. In our setting, even if conditioning on A favors X under μ^* it might not be the case under μ. In an extreme situation, the signal receiver could believe that any state outside A is impossible, so if she receives A, she simply will not change her beliefs.

1.4.3 Reaching Consensus in a Network

Let $N = \{1, 2, \ldots, n\}$ be a set of pseudo-Bayesian decision makers. The agents communicate with each other according to a social network. We describe the network as a non-directed graph where the presence of a link between i, j indicates that agent i communicates with agent j. Let $S(i)$ be the set of neighbors of i plus i herself.

Agents in the network would like to reach a consensus on qualitative beliefs over the following pairs of events: A_{11} versus A_{12} and A_{21} versus A_{22}. For each j, let $\alpha_j = (A_{j1}, A_{j2})$ and $N_j \subset N$ be the set of α_j-experts, who exogenously learns α_j at the start of every period.

In each period, each decision maker communicates to her neighbors her qualitative beliefs about A_{j1} versus A_{j2} for each j. If the decision maker believes that A_{j1} and A_{j2} are equally likely, then by default α_j is communicated.[9] If $N_j \cap S(i) \neq \emptyset$; that is if agent i is an α_j-expert or has an α_j-expert neighbor, she will simply ignore any

[9]Same results hold if α_j and $\bar{\alpha}_j = (A_{j2}, A_{j1})$ are both communicated.

$\bar{\alpha}_j = (A_{j2}, A_{j1})$. Then, each agent updates on the rest of qualitative statements in any finite sequence, repetitions permitted. Let μ_i denote agent i's prior belief and μ_i^k denote her belief at the end of period k. We assume that $\mu_i(R_{\alpha_1} \cap R_{\alpha_2}) > 0$ for all i, so that $\alpha_1, \bar{\alpha}_1, \alpha_2$ and $\bar{\alpha}_2$ are all credible along any agent's belief path.[10]

We say that a network is *strongly connected* if for any $i, j \in N$ there exists $k_1, k_2, \ldots, k_r$ such that $i \in S(k_1), k_1 \in S(k_2), \ldots, k_{r-1} \in S(k_r), k_r \in S(j)$. Note that if a network is strongly connected and there exists an α_j-expert, then the consensus ranking between A_{j1} and A_{j2}, if reached, must be the correct one: $A_{j1} \geq A_{j2}$.

Proposition 1.4. *μ_i^k converges in total variation to some μ_i^* for each i. Moreover, $N_j \cap S(i) \neq \emptyset$ implies $\mu_i^*(A_{j1}) \geq \mu_i^*(A_{j2})$.*

Hence, the beliefs of pseudo-Bayesian agents communicating within a network will converge. The results in the previous subsection imply that the limit μ_i^* depends on agent i's learning protocol. The proposition also states that experts' opinions regarding their specific expertise will always reach to their neighbors. However, although neighbors of experts could learn the truth in the limit, it is not guarenteed that they can spread the truth to their own neighbors.

In fact, agents can disagree with each other in the limit even if the network is strongly connected. Suppose three decision makers form a line as in Figure 1.3. Agent 1 is an α-expert and agent 2 is a β-expert. Agent 3, however, has expertise in neither α nor β. Suppose $[\beta, \alpha]$ induces recency bias on μ_2. Then, although in the limit agent 2 learns both α and β, each time she communicates with agent 3, she will say β and $\bar{\alpha} = (A_2, A_1)$. Therefore agent 3 can never learn α.

Hence, recency bias may prohibit consensus. DeMarzo, Vayanos, and Zwiebel (2003) adopt a similar network setting where agents communicate repeatedly with their neighbors but each time treat the information received as if it were new and

[10] Otherwise, if, for example, α_1 and α_2 are degenerate given μ_i, then after agent i learns α_1, statement α_2 becomes non-credible.

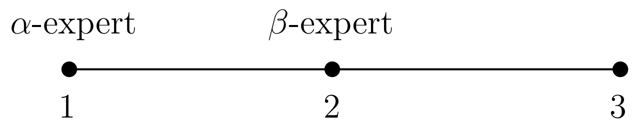


Figure 1.3: A network where consensus might not be reached.

independent. In their model, consensus is always reached on multiple issues partly because information on different issues are updated simultaneously so that recency bias is precluded. The next proposition shows that if recency bias is ruled out, in particular when $\alpha_1 \perp \alpha_2$ given any μ_i, consensus can be reached within a rich enough network.

Proposition 1.5. *Suppose the network is strongly connected and* $\#N_j > 0$ *for each* j*. If* $\alpha_1 \perp \alpha_2$ *given* μ_i *for all* i*, then* $\mu_i^*(A_{j1}) \geq \mu_i^*(A_{j2})$ *for all* i, j*. In particular, there is* K *such that* $\mu^k = \mu^*$ *for any* $k \geq K$*.*

That is, given that every agent believes that α_1 are α_2 are orthogonal, if there exists an expert on each pair of events and a path between any pair of agents, experts' ideas will spread to everyone within a finite time. If the orthogonality condition is violated, then Proposition 1.4 implies that consensus is also guarenteed when $N_j \cap S(i) \neq \emptyset$ for all i, j.

1.5 Pseudo-Bayesian versus Bayesian revision

It is well-known that Bayes' rule is a special case of the method of maximum relative entropy.[11] Not surprisingly, our pseudo-Bayes' rule also includes Bayesianism as a special case. In this section, let the state space be $S \times I$ and Σ, Ω be respectively σ-algebras on S, I. The DM's prior μ is therefore a nonatomic probabililty measure defined on the product measure space $(S \times I, \Sigma \times \Omega)$.

[11] See Caticha and Giffin (2006).

Suppos the DM learns signal $I' \in \Omega$. If $\mu(S \times I') > 0$, then her rational Bayesian posterior should be $\mu(\cdot|I')$. Alternatively, if the DM translates signal I' into qualitative information $\alpha = (\emptyset, S \times (I \backslash I'))$ which means "any signal other than I' cannot occur", the pseudo-Bayes' rule also assigns $\mu(\cdot|I')$ as the posterior. In particular, our pseudo-Bayes' rule requires that

$$\mu^{\alpha}(A) = \frac{\mu(A \cap (S \times I'))}{\mu(S \times I')} = \mu(A|S \times I') = \mu(A|I')$$

for any $A \in \Sigma \times \Omega$, which proves the following theorem.

Theorem 1.4. *If a one-step revision rule satisfies Axiom 1 and 2, then $\mu^{\alpha} = \mu(\cdot|I')$ for any $I' \in \Omega$ such that $\mu(S \times I') > 0$, where $\alpha = (\emptyset, S \times (I \backslash I'))$.*

Hence, Bayes' rule is a special case of our pseudo-Bayes' rule. However, given any fixed state space, the latter allows the decision maker to process a richer set of statements. For example, with a prior on $S \times I$, a Bayesian decision maker cannot process statements such as $(S \times \{i\}, S \times \{i'\})$ where $i, i' \in I$; that is, the statement "signal i is more likely than signal i'." Such information is not measurable with respect to the space if it contradicts μ. To process this qualitative statement, the decision maker needs a prior μ' on the extended state space $S \times I^2$. Nevertheless, with a larger and more flexible state space comes a larger collection of qualitative statements. Even with μ', the decision maker still could not update on qualitative statements such as $(S \times (i, j), S \times (i, j'))$; that is, " signal i is more likely to beat j than j' in terms of likelihood." Hence, given the state space, no matter how flexible and high-dimensional it is, the set of qualitative statements is always strictly larger than what Bayesians are able to process.

The greater flexibility that the pseudo-Bayes' rule affords is important. In our leading example, Physician X receives unexpected but payoff-relevant qualitative information from recent developments in scientific research. That a piece of information is unexpected means that a Bayesian decision maker lacks the ability to incorporate this information in a consistent manner. With the pseudo-Bayes' rule, decision makers like Physician X can incorporate such payoff-relevant information into her beliefs in order to make more informed decisions.

1.6 Conclusion

In this paper, we considered a situation in which the decision maker receives unexpected qualitative information. Two simple axioms delivered a closed-form formula, the pseudo-Bayes rule, which assigns a posterior to each qualitative statement given a prior. The pseudo-Bayesian posterior turns out to be the closest probability measure, in the Kullback-Leibler sense, to the decision maker's prior consistent with the newly-received information.

We showed that our DM is susceptible to recency bias and that repetition enables her to overcome it. This last observation implies that through repetition the decision maker eventually learns the truth no matter how complicated it is.

We then describe how pseudo-Bayesian agents interact. We first consider a situation where an information sender knows the true distribution of a random variable and intends to persuade the receiver with honest qualitative statements. We show that if the receiver's belief has bounded support, it is possible to move her towards the truth. Second, we consider a network of pseudo-Bayesian agents who seek to reach consensus on qualitative rankings of events. We show that the beliefs of these agents communicating within a network will converge but that they may disagree in the limit even if the network is strongly connected.

This network analysis provides a different interpretation of our model: it describes two agents who seek to reach consensus on qualitative rankings over multiple pairs of events according to some agenda. In this context, the symmetry axiom implies that if they disagree with each other on A versus B, both will concede and the concensus will be that A and B are equally likely. Exchangeability now states that the agenda does not matter when the pairs of events are orthogonal to each other.

1.7 Appendix

1.7.1 Properties of the Kullback-Leibler Divergence

Let μ, ν be probability measures defined on measure space (S, Σ) such that $\nu \ll \mu$. The Kullback-Leibler divergence from μ to ν, $d(\mu||\nu) = -\int_S \ln(\frac{d\nu}{d\mu})d\mu$.

Fact 1. *The integral $\int_S \ln(\frac{d\nu}{d\mu})d\mu$ exists.*

Proof. It suffices to show that $\int_S \max\{\ln \frac{d\nu}{d\mu}, 0\}d\mu < \infty$. By definition of the Radon-Nikodym derivative $\frac{d\nu}{d\mu}$ is a measurable function. Therefore In the second equality $\frac{d\mu}{d\nu} = (\frac{d\nu}{d\mu})^{-1}$ is well-defined within the range $\frac{d\nu}{d\mu} \geq 1$. □

Fact 2. *$d(\mu||\cdot)$ is strictly convex in $\{\nu \in \Delta(S, \Sigma) | \nu \ll \mu\}$.*

Proof. Let $\nu = a\nu_1 + (1-a)\nu_2$ for $\nu_1, \nu_2 \in \{\nu \in \Delta(S, \Sigma) | \nu \ll \mu\}$ and $a \in (0, 1)$. It is clear that $\nu \in \Delta(S, \Sigma)$ and $\nu \ll \mu$. Also

$$\frac{d\nu}{d\mu} = a\frac{d\nu_1}{d\mu} + (1-a)\frac{d\nu_2}{d\mu} \qquad \mu\text{-almost everywhere.}$$

Hence with equality if and only if

$$\frac{d\nu_1}{d\mu} = \frac{d\nu_2}{d\mu} \qquad \mu\text{-almost everywhere.}$$

By definition of Radon-Nikodym derivatives, if $\frac{d\nu_1}{d\mu} = \frac{d\nu_2}{d\mu}$ μ-almost everywhere, for any $A \in \Sigma$ we have that

$$\nu_1(A) = \int_A \frac{d\nu_1}{d\mu}d\mu = \int_A \frac{d\nu_2}{d\mu}d\mu = \nu_2(A).$$

therefore $\nu_1 = \nu_2$. □

Fact 3. *$d(\mu||\nu) \geq 0$ and equality is attained if and only if $\mu = \nu$.*

Proof. By Jensen's inequality

$$-\int_S \ln(\frac{d\nu}{d\mu})d\mu \geq -\ln(\int_S \frac{d\nu}{d\mu}d\mu) = -\ln 1 = 0$$

with equality attained if and only if $\frac{d\nu}{d\mu} = C$ μ-almost everywhere. By definition of the Radon-Nikodym derivative it is clear that $C = 1$. Hence for any $A \in \Sigma$,

$$\nu(A) = \int_A \frac{d\nu}{d\mu}d\mu = \int_A d\mu = \mu(A).$$

□

1.7.2 Proof of Theorem 1.2

Proof. It suffices to prove that optimization (P) in Theorem 1.2 has the pseudo-Bayes' rule as the unique solution. Recall optimization (P):

$$\begin{aligned} \min_{\nu \ll \mu} \quad & -\int_S \ln(\frac{d\nu}{d\mu})d\mu \\ s.t. \quad & \nu(A_1) \geq \nu(A_2). \end{aligned} \tag{P}$$

First of all note that since we are picking $\nu \ll \mu$ the Radon-Nikodym derivative is well-defined. The constraint optimization problem (P) is equivalent to the following (P*).

$$\begin{aligned} \min_{f:S\to[0,\infty) \text{ measurable}} \quad & -\int_S \ln f \ d\mu \\ s.t. \quad & \int_{A_1 \backslash A_2} f d\mu \geq \int_{A_2 \backslash A_1} f d\mu \\ & \int_S f d\mu = 1. \end{aligned} \tag{P*}$$

This is an infinite-dimensional optimization problem. The next lemma reduces (P^*) to a finite dimensional problem.

Lemma 1.1. *Let $\mu(C) > 0$ and $p \in [0,1]$, the following optimization problem has solution $f = p/\mu(C)$ and the solution is unque up to μ-almost everywhere equality.*

$$\min_{f:C\to[0,\infty)\ measurable} -\int_C \ln f\ d\mu$$
$$s.t.\quad \int_C f d\mu = p.$$

Proof. Let $\nu = \mu/\mu(C)$. It is clear that

$$-\int_C \ln f\ d\mu = -\mu(C)\mathbb{E}_\nu[\ln f] \geq -\mu(C)\ln \mathbb{E}_\nu[f] = -\mu(C)\ln\frac{p}{\mu(C)}$$

by Jensen's inequality. Equality is attained if and only if f is a constant μ-almost everywhere. Then the constraint demands that $f = p/\mu(C)$ μ-almost everywhere. □

With Lemma 1.1, (P^*) reduces to the following finite-dimensional optimization problem, for Lemma 1.1 basically demands that within G_α, B_α or R_α, f is a constant. Let $\mu(G_\alpha) = p_1, \mu(B_\alpha) = p_2$ and $p_3 = 1 - p_1 - p_2$.

$$\min_{q_1,q_2,q_3\geq 0} -\sum_{i=1}^{3} p_i \ln\frac{q_i}{p_i} \tag{P**}$$
$$s.t.\quad q_1 \geq q_2$$
$$q_i = 0 \text{ if } p_i = 0$$
$$\sum_{i=1}^{3} q_i = 1.$$

Let us first analyze the salient case when $p_1 > 0$. We will first ignore the absolute continuity constraint. Since our objective function is strictly convex, Kuhn-Tucker conditions are necessary and sufficient. Let

$$\mathcal{L}(\mathbf{q}, \lambda, \eta) = -\sum_{i=1}^{3} p_i \ln \frac{q_i}{p_i} + \lambda(q_1 - q_2) + \eta(1 - \sum_{i=1}^{3} q_i).$$

It is easy to verify that the Kuhn-Tucker conditions imply $q_3 = p_3$. In addition if $p_2 > p_1$ then $q_1 = q_2 = (p_1 + p_2)/2$; if $p_1 \geq p_2$ then $q_i = p_i$ for all i. Clearly in all circumstances the absolute continuity constraint is not violated. On the other hand, if $p_1 = 0$ we need $q_1 = 0$ and then the constraints demand $q_2 = 0$. Therefore the only feasible solution is $(0, 0, 1)$.

Let the unique solution of (P**) be (q_1, q_2, q_3). Let $h : [0, 1] \to \mathbb{R}$ be such that $h(0) = 0$ and $h(x) = 1/x$ if $x > 0$. By Lemma 1.1 and the above analysis, (P*) has a solution f which is unique up to μ-everywhere equality, given by

$$f(s) = \begin{cases} q_1 h(\mu(G_\alpha)), & \text{if } s \in G_\alpha, \\ q_2 h(\mu(B_\alpha)), & \text{if } s \in B_\alpha, \\ q_3 h(\mu(R_\alpha)), & \text{if } s \in R_\alpha. \end{cases}$$

Since Kullback-Leibler divergence is strictly convex, the solution to (P) is unique. It is easy to verify that the solution is exactly characterized by our pseudo-Bayes' rule. □

1.7.3 Proof of Theorem 1.1

Proof. We first prove the "only if" part. Consider the following lemma.

Lemma 1.2 (*Conservatism*). *If a one-step revision rule satisfies Axiom 1.2 then* $\mu(A_1) < \mu(A_2) \Longrightarrow \mu^\alpha(A_1) = \mu^\alpha(A_2)$.

Proof. Since $\mu(A_1) < \mu(A_2)$, $\bar{\alpha}$ conforms to μ. Therefore $\mu = \mu^{\bar{\alpha}}$ by definition of a one-step revision rule. Hence $\mu^\alpha = \mu^{\bar{\alpha}\alpha}$ and the rest is implied by symmetry. □

Lemma 1.3. *If a one-step revision rule satisfies Axiom 1.1 and 1.2, then*

$$\mu(A) \geq \mu(B) \Longrightarrow \mu^\alpha(A) \geq \mu^\alpha(B)$$

if $A \cup B \subset C \in \Pi_\alpha$ for some C and α is credible given μ.

Proof. First we prove that if the one-step revision rule satisfies Axiom 1.1 and 1.2, $\alpha \perp \beta$ given μ implies $\mu(B_1) \geq \mu(B_2) \Longrightarrow \mu^\alpha(B_1) \geq \mu^\alpha(B_2)$. Suppose $\mu(B_1) \geq \mu(B_2)$ but $\mu^\alpha(B_2) > \mu^\alpha(B_1)$. By definition of the revision rule we must have $\mu = \mu^\beta$ so $\mu^{\beta\alpha} = \mu^\alpha$ and hence $\mu^{\beta\alpha}(B_2) > \mu^{\beta\alpha}(B_1)$. But by conservatism we have $\mu^{\alpha\beta}(B_1) = \mu^{\alpha\beta}(B_2)$, a contradiction to our exchangeability axiom.

Next suppose $A \cup B \subset C \in \Pi_\alpha$. Wlog we could assume $A \cap B = \emptyset$ since μ and μ^α are well-defined probabilities. If $\mu(A) > 0$, by definition $\alpha \perp (A, B)$ with respect to μ and we are done. When $\mu(A) = 0$, if $\mu(B) > 0$ we have $\mu(B) > \mu(A)$ so the left hand side of the implication will not be met. The only case left is when $\mu(A) = \mu(B) = 0$. If $\mu(C) > 0$, then there exist mutually exclusive $C', C'' \subset C$ with $(C' \cup C'') \cap (A \cup B) = \emptyset$ such that $\mu(C') = \mu(C'') > 0$ by the convex-rangedness of μ. Clearly we have $\alpha \perp (A \cup C', B \cup C'')$ and $\alpha \perp (C', C'')$. Hence we have $\mu^\alpha(A \cup C') = \mu^\alpha(B \cup C'')$ and $\mu^\alpha(C') = \mu^\alpha(C'')$. Therefore $\mu^\alpha(A) = \mu^\alpha(B)$ and we are done. If $\mu(C) = 0$ and $C = B_\alpha = A_2 \backslash A_1$, then we already have $\mu(A_1) \geq \mu(A_2)$, by definition of a one-step revision rule, $\mu^\alpha = \mu$ and we are done. If $\mu(C) = 0$ and $C = R_\alpha$, then clearly $D_\alpha \subset R_{(A,B)}$ and hence $\alpha \perp (A, B)$. Finally, $\mu(C) = 0$ and $C = G_\alpha \cup R_\alpha$ implies that α is non-credible. □

Then, we take advantage of the convex-rangedness of μ and further characterize the revision rule by the following lemma.

Lemma 1.4. *If an one-step revision rule satisfies Axiom 1.1 and 1.2, then $\mu^\alpha \ll \mu$, and for $C \in \Pi_\alpha, \mu^\alpha(\cdot|C) = \mu(\cdot|C)$ if $\mu(C) \cdot \mu^\alpha(C) > 0$.*

Proof. We first prove that for $C \in \Pi_\alpha$, $\mu^\alpha(\cdot|C) = \mu(\cdot|C)$ if $\mu(C) \cdot \mu^\alpha(C) > 0$. By Lemma 1.2, for any $A, B \subset C$, $\mu(A) = \mu(B)$ implies $\mu^\alpha(A) = \mu^\alpha(B)$. Therefore for any partition $\{C_i\}_{i=1}^n$ of C such that $\mu(C_i) = \mu(C_1)$ for all i, we also have $\mu^\alpha(C_i) = \mu^\alpha(C_1)$ for all i. Therefore $\mu(C_i|C) = \mu^\alpha(C_i|C)$ for any i. Note that since μ is nonatomic, such a partition exists for any n. Moreover for any $A \subset C$ such that $\mu(A) = \frac{\mu(C)}{n}$, A belongs to some partition $\{C_i\}_{i=1}^n$. Therefore for any event $B \subset C$ such that $\mu(B|C) = r$ where r is rational, $\mu(B|C) = \mu^\alpha(B|C)$. Then countable additivity of probabilities finishes the proof.

To prove that $\mu^\alpha \ll \mu$, first note that for any $C \in \Pi_\alpha$, by Lemma 1.2 if $\mu(C) = \mu(\emptyset) = 0$, $\mu^\alpha(C) = \mu^\alpha(\emptyset) = 0$ as well. Together with the result above, we know that for any $A \in \Sigma$ and $C \in \Pi_\alpha$, $\mu(A \cap C) = 0$ implies $\mu^\alpha(A \cap C) = 0$ and the rest is trivially implied by addivity. □

Next, we show that $\mu^\alpha(D_\alpha) = \mu(D_\alpha)$. When $\mu(A_1) \geq \mu(A_2)$ there is nothing to prove. Thus we assume that $\mu(A_2) > \mu(A_1)$ and hence $\mu(B_\alpha) > 0$. Since $\mu^\alpha \ll \mu$ the case when $\mu(D_\alpha) = 1$ is trivial. Hence the salient case is when $\mu(G_\alpha)$, $\mu(B_\alpha)$ and $\mu(R_\alpha)$ are all positive. Consider β where $B_1 = \emptyset$ and $B_2 \subset R_\alpha$ such that $\mu(B_2) > 0$. It is easy to see that $\alpha \perp \beta$ given μ. Axiom 1.1 then requires that $\mu^{\alpha\beta} = \mu^{\beta\alpha}$. Therefore

$$\mu^{\alpha\beta}(D_\alpha) = \mu^{\beta\alpha}(D_\alpha).$$

Note that Lemma 1.4 implies that revision on β is equivalent to conditioning on $S \backslash B_2$ according to Bayes' rule. Therefore the above display equation is equivalent to

$$\frac{\mu^\alpha(D_\alpha)}{1 - \mu^\alpha(B_2)} = (\mu(\cdot|S \backslash B_2))^\alpha(D_\alpha) \tag{1.7}$$

where $\mu(\cdot|A)$ denotes the probability measure conditioning on $A \in \Sigma$ given μ. Moreover, we know by the previous lemma that

$$\mu^\alpha(B_2) = \frac{1-\mu^\alpha(D_\alpha)}{1-\mu(D_\alpha)}\mu(B_2).$$

Substituting into (1.7), we get

$$\frac{\mu^\alpha(D_\alpha)}{1-\frac{1-\mu^\alpha(D_\alpha)}{1-\mu(D_\alpha)}\mu(B_2)} = (\mu(\cdot|S\backslash B_2))^\alpha(D_\alpha) \tag{1.8}$$

The next lemma turns equation (8) above into a functional equation. It states that the sufficient statistic for $\mu^\alpha(D_\alpha)$ when revision on α from μ is $\mu(\cdot|D_\alpha)$ and $\mu(D_\alpha)$.

Lemma 1.5. *If Axiom 1.1 and 1.2 are satisfied, for each α, $\mu^\alpha(D_\alpha) = g(\mu(\cdot|D_\alpha), \mu(D_\alpha))$.*

Proof. Let $A \subset R_\alpha$ such that $0 < \mu(A) < \mu(R_\alpha)$ and γ be $(\emptyset, A)$. Also let $B = R_\alpha \backslash A$.

Step 1: If $\mu(\cdot|S\backslash A) = \mu_1(\cdot|S\backslash A)$ and $\mu(A) = \mu_1(A)$, then $\mu^\alpha(D_\alpha) = \mu_1^\alpha(D_\alpha)$.

The claim here basically states that $\mu^\alpha(D_\alpha)$ does not depend on how μ behaves within A. Since $\nu^\gamma = \nu(\cdot|S\backslash A)$ for any ν we have that $\mu^\gamma = \mu_1^\gamma$. Therefore we must have $\mu^{\gamma\alpha} = \mu_1^{\gamma\alpha}$. Note that $\alpha \perp \gamma$ given both μ and μ_1. Therefore $\mu^{\alpha\gamma} = \mu_1^{\alpha\gamma}$. Hence by the same logic in the derivation of equation (1.8),

$$\frac{\mu^\alpha(D_\alpha)}{1-\frac{1-\mu^\alpha(D_\alpha)}{1-\mu(D_\alpha)}\mu(A)} = \frac{\mu_1^\alpha(D_\alpha)}{1-\frac{1-\mu_1^\alpha(D_\alpha)}{1-\mu_1(D_\alpha)}\mu_1(A)}$$

which reads

$$(1-\mu(D_\alpha)-\mu(A))\mu^\alpha(D_\alpha) = (1-\mu_1(D_\alpha)-\mu_1(A))\mu_1^\alpha(D_\alpha).$$

It is clear that $\mu(D_\alpha) = \mu_1(D_\alpha)$ and $1-\mu(D_\alpha)-\mu(A) > 0$ so we have established the claim.

Step 2: If $\mu(\cdot|D_\alpha) = \mu_1(\cdot|D_\alpha)$, $\mu(D_\alpha) = \mu_1(D_\alpha)$ and $\mu(A) = \mu_1(A)$, then $\mu^\alpha(D_\alpha) = \mu_1^\alpha(D_\alpha)$.

Let $\mu_2 = \mu(S\backslash A)\mu(\cdot|S\backslash A) + \mu(A)\mu_1(\cdot|A)$. So μ_2 agrees with μ within $S\backslash A$ and agrees with μ_1 within A. It is clear that by Step 1 $\mu^\alpha(D_\alpha) = \mu_2^\alpha(D_\alpha)$. Moreover we also have that $\mu_2 = \mu_1(S\backslash B)\mu_1(\cdot|S\backslash B) + \mu_1(B)\mu(\cdot|B)$; that is μ_1 and μ_2 only differ within B. Hence by Step 1, $\mu_1^\alpha(D_\alpha) = \mu_2^\alpha(D_\alpha)$ and we are done.

Step 3: If $\mu(\cdot|D_\alpha) = \mu_1(\cdot|D_\alpha)$ and $\mu(D_\alpha) = \mu_1(D_\alpha)$ then $\mu^\alpha(D_\alpha) = \mu_1^\alpha(D_\alpha)$.

If $\mu_1(A) = 0$ then $\mu_1(B) > 0$. Let $\mu_2 = \mu(S\backslash B)\mu(\cdot|S\backslash B) + \mu(B)\mu_1(\cdot|B)$, clearly $\mu^\alpha(D_\alpha) = \mu_2^\alpha(D_\alpha)$ by Step 1 since μ and μ_2 only differ within B. Pick $A' \subset A$ such that $\mu(A') = \mu(A)/2$ and $B' \subset B$ such that $\mu_1(B') = \mu_1(B)/2$. It is clear that $\mu_1(A' \cup B') = \mu_2(A' \cup B') = \mu(S\backslash D_\alpha)/2$ since $\mu_1(A') = \mu_1(A) = 0$, then Step 2 implies that $\mu_1^\alpha(D_\alpha) = \mu_2^\alpha(D_\alpha)$.

If $0 < \mu_1(A) < \mu_1(R_\alpha)$, let $\mu_3 = \mu(D_\alpha)\mu(\cdot|D_\alpha) + \mu(A)\mu_1(\cdot|A) + \mu(B)\mu_1(\cdot|B)$. By Step 2 it is clear that $\mu^\alpha(D_\alpha) = \mu_3^\alpha(D_\alpha)$. Pick $A' \subset A$ such that $\mu_1(A') = \mu_1(A)/2$ and $B' \subset B$ such that $\mu_1(B') = \mu_1(B)/2$. It is clear that $\mu_1(A' \cup B') = \mu_3(A' \cup B') = \mu(R_\alpha)/2$, then Step 2 finishes the proof.

The case when $\mu_1(A) = \mu_1(R_\alpha)$ is symmetric to the case when $\mu_1(A) = 0$, all we need is to switch the identities between A and B. □

By Lemma 1.5, equation (1.8) is equivalent to

$$\frac{g(\mu(\cdot|D_\alpha), \mu(D_\alpha))}{1 - \frac{1-g(\mu(\cdot|D_\alpha),\mu(D_\alpha))}{1-\mu(D_\alpha)}\mu(B_2)} = g\left(\mu(\cdot|D_\alpha), \frac{\mu(D_\alpha)}{1-\mu(B_2)}\right)$$

Let $\mu(B_2) = b, \mu(D_\alpha) = x$. Holding $\mu(\cdot|D_\alpha)$ constant and abusing the notations a little bit we arrive at

$$\frac{g(x)}{1 - \frac{1-g(x)}{1-x}b} = g(\frac{x}{1-b}) \tag{1.9}$$

which holds for $0 < x < 1$ and $0 < b < 1 - x$, since we could always vary B_2 within R_α continuously thanks to the convex-rangedness of μ.

Lemma 1.6. *The solutions to functional equation (19) which maps from* $(0, 1)$ *to* $[0, 1]$ *are*

$$g(x) = \frac{a}{2a - 1 + \frac{1-a}{x}} \tag{1.10}$$

where $a \in [0, 1]$.

Proof. Let $g(\frac{1}{2}) = a$. For any $0 < x < \frac{1}{2}$ let $b = 1 - 2x$, therefore

$$a = g(\frac{1}{2}) = \frac{g(x)}{1 - \frac{1-g(x)}{1-x}(1 - 2x)}$$

which implies equation (7). For any $\frac{1}{2} < x < 1$ let $b = 1 - \frac{1}{2x}$. We have

$$g(x) = \frac{g(\frac{1}{2})}{1 - \frac{1-g(\frac{1}{2})}{1-1/2}(1 - \frac{1}{2x})} = \frac{a}{1 - (1 - a)(2 - \frac{1}{x})} = \frac{a}{2a - 1 + \frac{1-a}{x}}.$$

Clearly there are no other solutions to the functional equation since $g(\frac{1}{2})$ uniquely defines $g(x)$ on $(0, \frac{1}{2})$ and $(\frac{1}{2}, 1)$. □

The functional form in (1.10) has the property that if $a = \frac{1}{2}, g(x) = x$; if $a > \frac{1}{2}$, $g(x) > x$ and if $a < \frac{1}{2}$, $g(x) < x$. Thus the last step of the proof is to show that $a = \frac{1}{2}$.

Suppose $a > \frac{1}{2}$ therefore $\mu^\alpha(D_\alpha) = g(x) > x = \mu(D_\alpha)$. Pick mutually exclusive C_1, C_2 such that $C_1 \cup C_2 = S, \mu(C_2) \geq \mu(C_1) > 0$ and $D_\alpha \subset C_2$. First note that $\mu^\alpha(C_2) > \mu^\alpha(C_1)$. To see this note that $g(x) > x$ implies $1 - g(x) < 1 - x$, and by Lemma 3, $\mu^\alpha(\cdot|R_\alpha) = \mu(\cdot|R_\alpha)$. Hence

$$\mu^\alpha(C_2) = g(x) + (\mu(C_2) - x)\frac{1 - g(x)}{1 - x} > \mu(C_2)$$

and

$$\mu^{\alpha}(C_1) = \mu(C_1)\frac{1-g(x)}{1-x} < \mu(C_1).$$

If $a = 1$ now γ becomes non-credible given μ^{α}, a contradiction to Axiom 1.1. For $a < 1$ by conservatism, $\mu^{\alpha\gamma}(C_1) = \mu^{\alpha\gamma}(C_2)$. On the other hand, also by conservatism $\mu^{\gamma}(C_1) = \mu^{\gamma}(C_2) = 1/2$. Lemma 3 ensures that $\mu^{\gamma}(\cdot|D_{\alpha}) = \mu(\cdot|D_{\alpha})$ since $D_{\alpha} \subset C_2$. Therefore we are able to use the same a for revision on α next. By the same logic as above if $a > \frac{1}{2}$ we have $\mu^{\gamma\alpha}(C_2) > \mu^{\gamma\alpha}(C_1)$, which is a contradiction since $\alpha \perp \gamma$ with respect to μ.

Suppose $a < \frac{1}{2}$ therefore $g(x) < x$. For x small pick mutually exclusive C_1, C_2 such that $C_1 \cup C_2 = S, \mu(C_2) \geq \mu(C_1) > 0$ and $D_{\alpha} \subset C_1$. By similar logic we have that $\mu^{\alpha\gamma}(C_1) = \mu^{\alpha\gamma}(C_2)$ but $\mu^{\gamma\alpha}(C_2) > \mu^{\gamma\alpha}(C_1)$, a contradiction to our exchangeability axiom.

Therefore we must have $a = \frac{1}{2}$ and hence $\mu^{\alpha}(D_{\alpha}) = \mu(D_{\alpha})$, which together with Lemma 1.4, implies the pseudo-Bayes rule. To see why, first of all, if $\mu(A_1) \geq \mu(A_2)$ the definition of a one-step revision rule requires $\mu^{\alpha} = \mu$. Secondly, if $\mu(A_1) < \mu(A_2)$ and $\mu(G_{\alpha}) > 0$, the DM moves probability from B_{α} to G_{α} in proportion to the prior until these two sets have the same probability $\mu(D_{\alpha})/2$, i.e.

$$\mu^{\alpha}(C) = \mu(C \cap R_{\alpha}) + \left(\frac{\mu(C \cap G_{\alpha})}{\mu(G_{\alpha})} + \frac{\mu(C \cap B_{\alpha})}{\mu(B_{\alpha})}\right) \cdot \frac{\mu(D_{\alpha})}{2}$$

for all $C \in \Sigma$. Finally, if $\mu(G_{\alpha}) = 0$, by absolutely continuity $\mu^{\alpha}(G_{\alpha}) = 0$, therefore the DM has no choice but redistribute all the probability of B_{α} to R_{α}, and hence $\mu^{\alpha} = \mu(\cdot|R_{\alpha})$. Hence we establish the "only if" part of Theorem 1.1.

Next we show the "if" part.

Axiom 1.2 is immediately implied by the pseudo-Bayes' rule. Hence it suffices to prove that if $\alpha \perp \beta$ given μ, the procedure implies $\mu^{\alpha\beta} = \mu^{\beta\alpha}$. First we prove that β is credible given μ^{α}. Clearly when $\mu(G_{\alpha}) > 0$ it is true since in this case $\mu \ll \mu^{\alpha}$.

When $\mu(G_\alpha) = \mu(G_\beta) = 0$, α and β cannot be orthogonal unless at least one of B_α and B_β is $\emptyset$. In thes cases β is always credible given μ^α if it is credible given μ. Finally assume that $\mu(G_\beta) > \mu(G_\alpha) = 0$. If $\alpha \perp \beta$ it has to be either $D_\beta \subset B_\alpha$ or $D_\beta \subset G_\alpha \cup R_\alpha$. In the former case $\mu^\alpha(B_\beta) = 0$ and in the latter case $\mu^\alpha(G_\beta) > 0$ hence β is always credible given μ^α.

Next we prove that $\mu^{\alpha\beta} = \mu^{\beta\alpha}$. Suppose $\mu(G_\alpha) > 0$ and $\mu(G_\beta) > 0$. The pseudo-Bayes' rule requires proportional change of local probabilities and fixes the masses of the domains, in our case D_α and D_β. Here we will prove that $\mu^{\alpha\beta} = \mu^{\beta\alpha}$ if $D_\alpha \subset G_\beta$ and leave the remaining cases to the reader since they are very similar. We know that $\mu^\alpha(D_\beta) = \mu(D_\beta) \equiv b$ since $\mu^\alpha(D_\alpha) = \mu(D_\alpha) \equiv a$. We also know the formula requires $\mu^{\alpha\beta}(D_\beta) = \mu^\alpha(D_\beta) = b$. Hence

$$\mu^{\alpha\beta}(D_\alpha) = \frac{b/2}{\mu(G_\beta)} a.$$

On the other hand,

$$\mu^{\beta\alpha}(D_\alpha) = \mu^\beta(D_\alpha) = \frac{\mu^\beta(D_\beta)/2}{\mu(G_\beta)} a = \frac{b/2}{\mu(G_\beta)} a.$$

The first equality also implies that $\mu^{\beta\alpha}(D_\beta) = \mu^\beta(D_\beta) = b$. Therefore we have $\mu^{\beta\alpha}(D_\alpha) = \mu^{\alpha\beta}(D_\alpha)$ and $\mu^{\beta\alpha}(D_\beta) = \mu^{\alpha\beta}(D_\beta)$, which is enough to ensure that $\mu^{\alpha\beta} = \mu^{\beta\alpha}$. Suppose $\mu(G_\alpha) = 0$ and $\mu(G_\beta) > 0$, $\alpha \perp \beta$ given μ implies $D_\beta \subset B_\alpha$ or $D_\beta \subset G_\alpha \cup R_\alpha$. In the former case $\mu^{\alpha\beta}(B_\alpha) = \mu^{\beta\alpha}(B_\alpha) = 0$ so our claim is trivial. In the latter case we have $\mu^{\alpha\beta}(G_\alpha \cup R_\alpha) = \mu^{\beta\alpha}(G_\alpha \cup R_\alpha) = 1$. It is easy to verify that $\mu^{\alpha\beta}(D_\beta) = \mu^{\beta\alpha}(D_\beta)$ and we are done. When $\mu(G_\alpha) = \mu(G_\beta) = 0$, α and β cannot be orthogonal unless at least one of B_α and B_β is $\emptyset$. If so, then at least one of α and β is redundant when learning. $\square$

1.7.4 Pseudo-Bayesian Updating with a Richer Prior

Suppose the state space of the DM is now $\{T, F\} \times S$, where T means that the signal source is reliable and F means otherwise; that is, the DM now has a subjective assessment of how reliable the source is. When she receives α from the signal source, the DM naturally translates the signal into $(T, A_1) \succsim (T, A_2)$, meaning "if the source is reliable, then A_1 is more likely than A_2". In case that the signal contradicts her prior, our pseudo-Bayesian decision maker will set (T, A_1) and (T, A_2) to be equally as likely but will not touch (F, A_1) and (F, A_2). Hence, when she evaluate the marginals of A_1 and A_2, the extent to which she increases the probability of A_1 and decreases the probability of A_2 will depend on her reliability assessment; that is, the DM is now capable of discounting the information content of the signal.

1.7.5 Proofs Related to Recency Bias

Lemma 1.7. *Let $[\alpha, \beta]$ be an ordered pair of nondegenerate qualitative statements. If $\mu(A_1) \geq \mu(A_2)$ and $\mu^\beta(A_2) > \mu^\beta(A_1)$, then $\mu^{\beta\alpha}(B_2) > \mu^{\beta\alpha}(B_1)$.*

Proof. WLOG assume $A_1 \cap A_2 = B_1 \cap B_2 = \emptyset$. Let $a = \mu(A_1 \cap B_1), c = \mu(A_1 \cap B_2), e = \mu(B_2 \cap A_2), g = \mu(A_2 \cap B_1), b = \mu(A_1 \backslash (B_1 \cup B_2)), d = \mu(B_2 \backslash (A_1 \cup A_2)), f = \mu(A_2 \backslash (B_1 \cup B_2)), h = \mu(B_1 \backslash (A_1 \cup A_2))$. Note that $\mu(A_1) = a + b + c, \mu(B_2) = c + d + e, \mu(A_2) = e + f + g$ and $\mu(B_1) = g + h + a$. Also let $a', b', c', d', e', f', g', h'$ be the corresponding probabilities for μ^β.

Since $A_1 \succsim A_2$ we have $a + b + c \geq e + f + g$. Moreover for $\mu^\beta(A_2) > \mu^\beta(A_1)$ to be true it has to be the case that $\mu(B_2) > \mu(B_1)$, so $a + h + g < c + d + e$. Since α, β are nondegenerate, $a + h + g > 0$. The pseudo-Bayes' rule requires $b' = b$, $f' = f$ and

$$a' = \frac{a+h+g+c+d+e}{2}\frac{a}{a+h+g}, \quad c' = \frac{a+h+g+c+d+e}{2}\frac{c}{c+d+e}$$
$$e' = \frac{a+h+g+c+d+e}{2}\frac{e}{c+d+e}, \quad g' = \frac{a+h+g+c+d+e}{2}\frac{g}{a+h+g}$$

By $\mu^\beta(A_2) > \mu^\beta(A_1)$, $a'+b'+c' < g'+f'+e'$, which implies $a'-g'+c'-e'+g-a+e-c < 0$ since $b'-f' = b-f \geq g-a+e-c$. It follows that

$$\frac{c+d+e}{a+h+g}(a-g) + \frac{a+h+g}{c+d+e}(c-e) + g - a + e - c < 0.$$

which reads

$$\frac{a-g}{a+h+g} < \frac{c-e}{c+d+e}. \tag{1.11}$$

Intuitively, if the probability decrease of B_2 is Δp_1 then such decrease reduces the difference in probability between A_1 and A_2 by $\frac{c-e}{c+d+e}\Delta p_1$. Let the probability increase of B_1 be Δp_2 then such increase raises the difference between A_1 and A_2 by $\frac{a-g}{a+h+g}\Delta p_2$. Our learning process requires $\Delta p_1 = \Delta p_2$, therefore to make $\mu^\beta(A_2) > \mu^\beta(A_1)$ inequality (1.11) has to be true. When $\mu(A_1) = \mu(A_2)$, (1.11) becomes also sufficient since $b - f = g - a + e - c$.

Now we prove that $\mu^{\beta\alpha}(B_2) > \mu^{\beta\alpha}(B_1)$. Since we know $\mu^\beta(B_1) = \mu^\beta(B_2)$, by symmetry the following condition is necessary and sufficient.

$$\frac{a'-c'}{a'+b'+c'} < \frac{g'-e'}{g'+f'+e'} \tag{1.12}$$

Inequality (1.11) is equivalent to $a'-c' < g'-e'$. If $g'-e' \leq 0$, since $0 < a'+b'+c' < g'+f'+c'$, inequality (1.12) is obviously true. Next we assume $g'-e' > 0$. Inequality (1.12) reads

$$(a'-c')(g'+e') + (a'-c')f - (g'-e')(a'+c') - (g'-e')b < 0. \tag{1.13}$$

By inequality (1.11), $a' - c' < g' - e'$. Hence we have

$$\begin{aligned}
\text{LHS of (1.13)} &\leq (a'-c')(g'+e') + (g'-e')f - (g'-e')(a'+c') - (g'-e')b \\
&\leq (a'-c')(g'+e') - (g'-e')(a'+c') - (g'-e')(g-a+e-c) \\
&= 2(a'e' - c'g') - (g'-e')(g-a+e-c) \\
&= \frac{a+h+g+c+d+e}{2}(g+e)\left(\frac{a-g}{a+h+g} - \frac{c-e}{c+d+e}\right) \leq 0
\end{aligned}$$

by inequality (1.11). If $a+b+c > e+f+g$ the second inequality above is strict. Since (α, β) is nondenerate $a+b+c > 0$. If $a+b+c = e+f+g > 0$ then either $f > 0$ or $g+e > 0$, both of which imply that LHS of (1.13) < 0 and we have established our claim. □

Proposition 1.1 *If $[\alpha, \beta]$ induces recency bias on μ, then $[\beta, \alpha]$ induces recency bias on μ^α.*

Proof. By $[\alpha, \beta]$ induces recency bias for μ, we know that $\mu^{\alpha\beta}(A_2) > \mu^{\alpha\beta}(A_1)$. It is clear that $\mu^\alpha(A_1) \geq \mu^\alpha(A_2)$. Hence by the previous lemma, $\mu^{\alpha\beta\alpha}(B_2) > \mu^{\alpha\beta\alpha}(B_1)$. □

Proposition 1.2 *If $[\alpha, \beta]$ induces recency bias on μ, then $\mu^{(\alpha\beta)^n}$ and $\mu^{(\alpha\beta)^{n-1}\alpha}$ converges in total variation to μ^* such that $\mu^*(A_1) = \mu^*(A_2)$ and $\mu^*(B_1) = \mu^*(B_2)$.*

Proof. Wlog assume that $A_1 \cap A_2 = B_1 \cap B_2 = \emptyset$. Let $\mu_0 = \mu$, $\mu_{2n-1} = \mu^{(\alpha\beta)^{n-1}\alpha}$ and $\mu_{2n} = \mu^{(\alpha\beta)^n}$ for all n.

It is clear that $\mu_n(A_1 \cap B_1)$ and $\mu_n(A_2 \cap B_2)$ are monotone therefore must be convergent. When revision on α, the probabilities of $B_1 \backslash (A_1 \cup A_2)$ and $B_2 \backslash (A_1 \cup A_2)$ will not change but they will respectively increase and decrease when revision on β. Hence $\mu_n(B_1 \backslash (A_1 \cup A_2))$ and $\mu_n(B_2 \backslash (A_1 \cup A_2))$ are also convergent. Similarly $\mu_n(A_1 \backslash (B_1 \cup B_2))$ and $\mu_n(A_2 \backslash (A_1 \cup A_2))$ converge as well. We are left with $A_1 \cap B_2$ and $A_2 \cap B_1$. Since $\mu_n((A_1 \cup A_2) \cup (B_1 \cup B_2))$ is constant, $\mu_n((A_1 \cap B_2) \cup (A_2 \cap B_1))$

converges. Therefore, to prove that μ_n converges strongly, by the second part of Theorem 1, it suffices to prove that $\mu_n(A_1 \cap B_2)$ converges. We know that

$$\mu_{2n+1}(A_1 \cap B_1) = \frac{\mu_{2n}(A_1 \cup A_2)}{2} \frac{\mu_{2n}(A_1 \cap B_1)}{\mu_{2n}(A_1)}$$

which implies $\mu_{2n}(A_1 \cup A_2)/(2\mu_{2n}(A_1)) \to 1$ and

$$\mu_{2n}(A_1 \cap B_2) = \frac{\mu_{2n}(A_1 \cup A_2)}{2} \frac{\mu_{2n}(A_1 \cap B_1)}{\mu_{2n+1}(A_1 \cap B_1)} - \mu_{2n}(A_1 \cap B_1) - \mu_{2n}(A_1 \backslash (B_1 \cup B_2)).$$

Note that we already know $\mu_n(A_1 \cup A_2) = \mu_n(A_1 \backslash B_2) + \mu_n(A_2 \backslash B_1) + \mu_n((A_1 \cap B_2) \cup (A_2 \cap B_1))$ converges. Hence $\mu_{2n}(A_1 \cap B_2)$ converges. Similarly $\mu_{2n+1}(A_1 \cap B_2)$ converges. Moreover,

$$\mu_{2n+1}(A_1 \cap B_2) = \frac{\mu_{2n}(A_1 \cup A_2)}{2} \frac{\mu_{2n}(A_1 \cap B_2)}{\mu_{2n}(A_1)}.$$

Since $\mu_{2n}(A_1 \cup A_2)/(2\mu_{2n}(A_1)) \to 1$, we have that $\mu_n(A_1 \cap B_2)$ is convergent. Similarly to Theorem 3, the convergence is in total variation.

The fact that $\mu^*(A_1) = \mu^*(A_2)$ and $\mu^*(B_1) = \mu^*(B_2)$ is obvious. □

1.7.6 Proof of Theorem 1.3

Proof. Let the sequence of beliefs be $\{\mu_k\}$ with $\mu_k = \mu^{\pi_1 \pi_2 \cdots \pi_k}$. Let ν be a solution to $\{\alpha_i\}_{i=1}^n$ such that $\nu(\cdot|C) = \mu(\cdot|C)$ if $C \in \bigvee_{i=1}^n \Pi_{\alpha_i}$ and $\mu(C) > 0$.

First we prove that

$$d(\mu_{k+1}||\mu_k) \cdot \min_i \nu(D_{\alpha_i}) \leq d(\nu||\mu_k) - d(\nu||\mu_{k+1}).$$

Since $\mu(G_{\alpha_i}) > 0$ for all i, it is clear that $\mu_k \sim \mu \sim \nu$ for all k. Therefore the KL-divergences in the above equation are all finite. Wlog let $\pi_{k+1} = \alpha$ and $\mu_k(G_\alpha) =$

$p_1, \mu_k(B_\alpha) = p_2$, also $\nu(G_\alpha) = q_1, \nu(B_\alpha) = q_2$. If $p_1 \geq p_2$ we are done since $\mu_k = \mu_{k+1}$. So we assume otherwise. We have picked ν in a way such that the above inequality reduces to

$$\left(\frac{p_1+p_2}{2}\ln\frac{p_1+p_2}{2p_1} + \frac{p_1+p_2}{2}\ln\frac{p_1+p_2}{2p_2}\right)\cdot\min_i \nu(D_{\alpha_i}) \leq q_1\ln\frac{p_1+p_2}{2p_1} + q_2\ln\frac{p_1+p_2}{2p_2}.$$

Since $\min_i \nu(D_{\alpha_i}) \leq q_1 + q_2$; $p_1 + p_2 \leq 1$ and $d(\mu_{k+1}||\mu_k) \geq 0$, it suffices to prove that

$$\left(\frac{q_1+q_2}{2}\ln\frac{p_1+p_2}{2p_1} + \frac{q_1+q_2}{2}\ln\frac{p_1+p_2}{2p_2}\right) \leq q_1\ln\frac{p_1+p_2}{2p_1} + q_2\ln\frac{p_1+p_2}{2p_2}$$

which reads

$$0 \leq \frac{p_2-p_1}{2}\ln\frac{2q_1}{p_1+p_2} + \frac{p_1-p_2}{2}\ln\frac{2q_2}{p_1+p_2}$$

which is obvious since $p_1 < \frac{p_1+p_2}{2} < p_2$ and $q_1 \geq q_2$. It follows that $d(\nu||\mu_k) \geq d(\nu||\mu_{k+1}) \geq 0$ and therefore $\lim_{k\to\infty} d(\nu||\mu_k)$ exists and $d(\mu_{k+1}||\mu_k) \to 0$. By Pinsker's inequality

$$d(\mu_{k+1}||\mu_k) \geq 2||\mu_{k+1} - \mu_k||_{TV}^2$$

where $||\cdot||_{TV}$ is the total variation norm. Hence $||\mu_{k+1} - \mu_k||_{TV}$ converges to 0. Therefore, $\max_{1\leq i,j\leq N} ||\mu_{kN+i} - \mu_{kN+j}||_{TV}$ converges to 0 as $k \to \infty$. For any j, statement $\alpha_j \in \bigcup_{i=1}^N \pi_{kN+i}$ for any k. It follows that any limit point μ^* of $\{\mu_k\}$ must satisfy $\mu^*(A_{j1}) \geq \mu^*(A_{j2})$ for all j. Moreover, since $d(\nu||\mu_k)$ is convergent, by continuity $d(\nu||\mu^*)$ is also bounded, which implies that $\nu \ll \mu^*$. Since $\mu^k \ll \mu \sim \nu$ for any k, it must be the case that $\mu^* \ll \nu$ as well. Therefore μ^* must be a solution to $\{\alpha_i\}_{i=1}^n$. Clearly μ^* satisfies $\mu^*(\cdot|C) = \mu(\cdot|C)$ if $C \in \bigvee_{i=1}^n \Pi_{\alpha_i}$ and $\mu(C) > 0$. Hence by continuity $d(\mu^*||\mu_k)$ converges to 0 and then Pinsker's inequality finishes our proof. □

1.7.7 Proof of Proposition 1.3

Proof. Suppose the proposition is not true. We first prove that for any n, there exists pairwise disjoint $P_1^n, P_2^n, \dots, P_n^n$ such that $\mu^*(P_k^n) = 1/n$ and $s \in P_k^n, s' \in P_{k+1}^n$ implies $X(s) \leq X(s')$. Our proof is constructive. For $k = 1, 2, ..., n-1$, let $a_k^n = \sup\{a | \mu^*(\{X \leq a\}) \leq k/n\}$. By nonatomicity pick $A_k^n \subset \{X = a_k^n\}$ such that $\mu^*(A_k^n) = k/n - \mu^*(\{X < a_k^n\})$. Inductively, let $P_1^n = \{X < a_1^n\} \cup A_1^n$, $P_k^n = \{X < a_k^n\} \cup A_k^n \backslash (\bigcup_{j<k} P_j^n)$. It is easy to show that the P_k^n's satisfy the conditions.

If $\mu(P_i^n) = 0$, then let $j = \min\{k | \mu(P_k^n) > 0\}$. Set $\alpha = (P_i^n, P_j^n)$. If α is credible, that is if $\mu(P_j^n) < 1$, α increases the receiver's expectation on X. If $\mu(P_j^n) = 1$, since $var_\mu(X) > 0$ there exists $B \subset P_j^n$ such that $E_\mu(X|B) < E_\mu(X|P_j^n \backslash B)$, then $\alpha = (P_i^n, B)$ would do the job. Therefore, if the scope of persusaion does not exist, we must have $\mu(P_k^n) > 0$ for any n, k. It follows that we also have $\mu^*(\{|X| \leq M\}) = 1$. Let $a_0^n = -M$ and $a_n^n = M$.

Let X_n be such that $s \in P_k^n$ implies $X_n(s) = E_\mu(X|P_k^n)$. It is easy to see $E_\mu(X) = E_\mu(X_n)$. Next we show that $E_\mu(X_n) \geq E_{\mu^*}(X_n)$. By construction we know $E_\mu(X|P_k^n)$ is weakly increasing in k. Notice that $E_\mu(X|P_{k+1}^n) > E_\mu(X|P_k^n)$ implies $\mu(P_{k+1}^n) \geq \mu(P_k^n)$, since if $\mu(P_{k+1}^n) < \mu(P_k^n)$, $\alpha = (P_{k+1}^n, P_k^n)$ would be persuasive. It then follows that

$$\sum_{k=1}^{n} \mu(P_k^n) E_\mu(X|P_k^n) \geq \sum_{i=1}^{n} \frac{1}{n} E_\mu(X|P_k^n).$$

and hence $E_\mu(X_n) \geq E_{\mu^*}(X_n)$.

Next, we show that X_n converges to X in distribution in probability space (S, Σ, μ^*). For $z \in \mathbb{R}$, we want to prove that $\mu^*(\{X_n \leq z\})$ converges to $\mu^*(\{X \leq z\})$. By construction, $a_{k-1}^n \leq E_\mu(X|P_k^n) \leq a_k^n$. Let $k_n = \max\{k | E_\mu(X|P_k^n) \leq z\}$. It is clear that $\mu^*(\{X_n \leq z\}) = k_n/n$. Suppose $k_n < n$, then since $z < E_\mu(X|P_{k_n+1}^n)$, it must be the case that $a_{k_n-1}^n \leq z < a_{k_n+1}^n$. If $k_n = n$, then $a_{n-1}^n \leq z$. In both cases

$(k_n - 1)/n \leq \mu^*(\{X \leq z\}) \leq (k_n + 1)/n$. It is then clear that $\mu^*(\{X_n \leq z\})$ converges to $\mu^*(\{X \leq z\})$.

Since X_n is uniformly bounded by M it is uniformly integrable. It follows that $E_{\mu^*}(X_n)$ converges to $E_{\mu^*}(X)$. Since $E_\mu(X) \geq E_{\mu^*}(X_n)$ for all n we have a contradiction. □

1.7.8 Proof of Proposition 1.4

Proof. For agent i, let $\{\pi_n\} \subset \{\alpha_1, \bar{\alpha}_1, \alpha_2, \bar{\alpha}_2\}$ be the sequence of qualitative statements that she internalizes. Let $\{\hat{\mu}_n\}$ be the subsequence (or subset if finite) of $\{\mu_i^{\pi_1\pi_2\cdots\pi_n}\}$ such that $\hat{\mu}_1 = \mu_i$, and $\mu_i^{\pi_1\pi_2\cdots\pi_n} \neq \mu_i^{\pi_1\pi_2\cdots\pi_{n+1}}$ implies $\mu_i^{\pi_1\pi_2\cdots\pi_{n+1}} \subset \{\hat{\mu}_n\}$. That is, $\{\hat{\mu}_n\}$ is the resulting subsequence (or subset if finite) if consecutive identical elements are eliminated. If $\{\hat{\mu}_n\}$ is finite, then it is clear that $\{\mu_i^{\pi_1\pi_2\cdots\pi_n}\}$ converges. It is then clear that if $N_j \cap S(i) \neq \emptyset$ she must believe in α_j since she receives α_j every period. Suppose $\{\hat{\mu}_n\}$ is a subsequence. Let $\{\hat{\pi}_n\}$ be the corresponding effective statements. It is clear that if $\hat{\pi}_n \in \{\alpha_1, \bar{\alpha}_1\}$ then $\hat{\pi}_{n+1} \in \{\alpha_2, \bar{\alpha}_2\}$ and vice versa. Moreover, we know that none of $\{\hat{\pi}_n\}$ is degenerate since if so the collection cannot have infinite number of elements: one more step after the degenerate step, α_j will hold as equality for both j. Wlog assume that $\hat{\pi}_1 = \alpha_1$. Then there are two situations:

First, $\hat{\mu}_1(A_{21}) > \hat{\mu}_1(A_{22})$. In this case the next effective statement must be $\hat{\alpha}_2$. Then if $\hat{\mu}_2(A_{11}) > \hat{\mu}_2(A_{12})$, by Lemma 1.7 it must be the case that $[\hat{\alpha}_2, \hat{\alpha}_1]$ induces recency bias on $\hat{\mu}_1$. It follows that the rest of the sequence can only be alternating between $\hat{\alpha}_1$ and $\hat{\alpha}_2$. Then by Proposition 1.2 we are done. If $\hat{\mu}_2(A_{12}) > \hat{\mu}_2(A_{11})$, it must be the case that $[\hat{\alpha}_2, \alpha_1]$ induces recency bias on $\hat{\mu}_1$. In this case Proposition 2 is also applicable.

Second, if $\hat{\mu}_1(A_{22}) > \hat{\mu}_1(A_{21})$, then the second effective statement must be α_2. Then depending on how $\hat{\mu}_2(A_{11})$ compares with $\hat{\mu}_2(A_{12})$ we also know that the remainder of the sequence must be alternating between two statements which induces recency bias on $\hat{\mu}_1$. Proposition 1.2 again finishes our proof. □

1.7.9 Proof of Proposition 1.5

Proof. From the proof of the previous proposition, the only case that any agent's beliefs are not converging in finite steps is when $\alpha \in \{\alpha_1, \bar{\alpha}_1\}$ and $\beta \in \{\alpha_2, \bar{\alpha}_2\}$ are nondegenerate and induces recency bias on the agent's belief along the revision path. If $\mu_i(G_{\alpha_j}) \cdot \mu_i(B_{\alpha_j}) = 0$ for some j, then α_j and $\bar{\alpha}_j$ are either degenerate or redundant along any possible revision path of agent i. In all cases agent i's beliefs will converge in finite steps. Suppose $\mu_i(G_{\alpha_j}) \cdot \mu_i(B_{\alpha_j}) > 0$ for both j. Then any such α and β will be orthogonal along any possible path. Therefore agent i's beliefs will also converge in finite steps. It is clear that any neighbors of an α_j-expert will believe in α_j after finite steps. Inductively, since the network is strongly connected and there exists both α_1- and α_2-experts, after a finite number of periods all decision makers must have reached consensus that α_1 and α_2 are both true. □

Chapter 2

Similarity-Based Belief Updating

2.1 Introduction

In this chapter, we focus on belief updating rather than revision; that is, we consider the usual Bayesian learning setting where the decision maker, who has a prior over payoff-relevant states, observes that an event occurs. Below, we present a model of behavioral decision makers who exhibits a range of well-documented deviations from Bayesianism. Among these deviations, the salient ones include base rate neglect, conjunction fallacy and disjunction fallacy.

Agents subject to *Base rate neglect* undervalue the base rate and update too much on the signal. In Kahneman and Tversky (1982), subjects are presented with the following problem:

> A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data: (i) 85% of the cabs in the city are Green and 15% are Blue, (ii) A witness identified the cab as a Blue cab. The court tested his ability to identify cabs under the appropriate visibility conditions. When presented with a sample of cabs (half of which were Blue and half of which were

> Green) the witness made correct identifications in 80% of the cases and erred in 20% of the cases. Question: What is the probability that the cab involved in the accident was Blue rather than Green?

While the Bayesian conditional being 0.41, the median and mode response of the subjects is 0.8—the accuracy of the witness' identification. The vast majority conclude that the culprit is more likely to drive a blue cab. These responses show that subjects tend to place too small a weight on the base rate 15/85 and thereby rely too much on the signal.[1]

Kahneman and Tversky (1983) document subjects who commit the *conjunction fallacy* by attributing a higher probability to event $A \cap B$ than to A. In their experiments, subjects are given a description (D) of a stereotypical leftist, called Linda. They are asked to choose the more probable option over:

> (T) Linda is a bank teller;
>
> (T∧F) Linda is a bank teller and is active in feminist movements.

T∧F, albeit a sub-event of T, is selected by 85% of the subjects as the more probable one.

In a similar vein, Bar-Hillel and Neter (1993) define the *disjunction fallacy* as assigning higher probability to A than to $A \cup B$.[2] In the experiments, subjects are given a description of Danielle, who is sensitive and writes poetry secretly. The majority of subjects are willing to bet on her being a literature major rather than a humanities major even though all subjects seem to know that the department of literature is a part of the faculty of humanities.

---

[1] There are, also, many extreme examples of base rate neglect in the legal process. In these examples, the very low probability of a particular set of evidence (B) given innocence (A) is cited as proof that given the evidence, the individual is very unlikely be innocent (A given B). For a detailed discussion see Fienberg (1989); DeGroot, Fienberg, and Kadane (1994) and Tribe (1971).

[2] See, also Fischhoff, Slovic, and Lichtenstein (1978).

Many models of updating heuristics have been proposed in order to explain these non-Bayesian experimental findings. These models assume that the decision maker has imperfect memory or recall (Mullainathan (2002c); Wilson (2014); Gennaioli and Shleifer (2010); Bordalo, Coffman, Gennaioli, and Shleifer (2016)), or thinks categorically (Mullainathan (2002b); Mullainathan, Schwartzstein, and Shleifer (2008)), or has an incorrect model in mind and interprets data accordingly (Barberis, Shleifer, and Vishny (1998); Rabin and Schrag (1999); Rabin (2002); and Schwartzstein (2014)). These heuristics vividly describe various thinking processes of behavioral agents, but none of them seems to be consistent with the taxicab experiment. The distinctive feature of the taxicab experiment (and many others) is that either (a) subjects are given the relevant prior probabilities, or (b) most subjects do not err in ranking unconditional events. In contrast, previous literature tends to assume that the decision maker does not have the full correct prior to reason with. Clearly this behavioral assumption has tension with experiments that feature well-defined priors.

In this paper, we provide a micro-foundation for behavioral subjects who tend to commit the fallacies above but have correct unconditional beliefs. Our decision maker assigns subjective probabilities, rationally and reasonably, to all events but she mistakes the question "How likely is A given B?" for the question "How representative is A of B (i.e., how similar are A and B)?". We base this behavioral twist on Kahneman and Tversky (1974), who propose that,

> *"In answering such [conditional probability] questions, people typically rely on the representativeness heuristic, in which probabilities are evaluated by the degree to which A is representative of B, that is, by the degree to which A resembles B."*

Hence, we *assume* that our decision maker assesses the degree to which A is similar to B when asked to evaluate $\Pr(A|B)$.

Our primitive is a binary relation between *two pairs* of events in the state space Ω: $(A,B) \succsim (C,D)$ means that A is more similar to B than C is to D. We impose axioms on $\succsim$ to identify a similarity index, $S(A,B)$, such that $(A,B) \succsim (C,D)$ if and only if $S(A,B) \geq S(C,D)$. We assume that our decision maker assesses $S(A,B)$ when asked to evaluate $\Pr(A|B)$. We follow the psychology literature and allow $S(A,B)$ to be asymmetric; that is, $S(A,B)$ need not equal $S(B,A)$.[3]

Our characterization of $S(A,B)$ are twofold. First, our axioms yield a subjective probability measure μ and a monotone aggregator f such that

$$S(A,B) = f\left(\mu(A|B), \mu(B|A)\right).$$

where $\mu(\cdot|\cdot)$ is given by Bayes' rule. That is, the decision maker confuses the conditional probabilities $\mu(A|B)$ and $\mu(B|A)$. Second, by imposing an additional robustness condition on $\succsim$ we obtain the following updating rule:

$$S(A,B) = \mu(A|B)^{\alpha} \mu(B|A)^{1-\alpha}$$

where $\alpha \in (0,1]$ measures the decision maker's deviation from Bayesianism. In particular, when $\alpha = 1$, the decision maker is Bayesian.

Our similarity representation, $S(\cdot,\cdot)$, provides a theory consistent with the experimental evidence above. Consider evaluating whether A or B is more likely given C. Given the Cobb-Douglas index, we have

$$\frac{S(A,C)}{S(B,C)} = \frac{\mu(A|C)^{\alpha}\mu(C|A)^{1-\alpha}}{\mu(B|C)^{\alpha}\mu(C|B)^{1-\alpha}} = \frac{\mu(C|A)}{\mu(C|B)} \cdot \left(\frac{\mu(A)}{\mu(B)}\right)^{\alpha}.$$

Since $\alpha \in (0,1]$, $(\mu(A)/\mu(B))^{\alpha}$ is closer to 1 than the correct prior odds ratio $\mu(A)/\mu(B)$. That is, α specifies the extent to which the decision maker neglects

[3]For example, we say that "an ellipse is similar to a circle" but not "a circle is similar to an ellipse." See Tversky (1977) for a detailed analysis.

the prior odds ratio.[4] In the taxicab problem, the likelihood ratio of the witness' signal (correct over incorrect) is $80/20 = 4$ where as the base rate (blue over green) is $15/85 = 0.176$. A decision maker with an $\alpha < 0.8$ would conclude, like the experimental subjects, that the hit and run cab was more likely to be blue than green.

In the Linda problem, although $\Pr(\mathrm{T}\wedge\mathrm{F}|\mathrm{D}) < \Pr(\mathrm{T}|\mathrm{D})$, arguably we have $\Pr(\mathrm{D}|\mathrm{T}\wedge\mathrm{F}) > \Pr(\mathrm{D}|\mathrm{T})$; that is, Linda is more likely to match the stereotypical description D if she is a feminist bank teller than if she is just a bank teller. Then, if α is low, an α-mixture of $\Pr(\mathrm{T}\wedge\mathrm{F}|\mathrm{D})$ and $\Pr(\mathrm{D}|\mathrm{T}\wedge\mathrm{F})$ may well be higher than an α-mixture of $\Pr(\mathrm{T}|\mathrm{D})$ and $\Pr(\mathrm{D}|\mathrm{T})$. In that case a decision maker would choose the conjunctive description T∧F as the more "probable" option. A similar logic applies to the results in Bar-Hillel and Neter (1993).

In our axiomatic system, *monotonicity* is the only postulate which permits deviation from Bayes' rule: it allows the decision maker to specify $\Pr(A\cup C|B) < \Pr(A|B)$ if $B \cap C = \emptyset$, simply because the pair of events $A \cup C$ and B differ more from each other. This axiom reflects exactly Kahneman and Tversky (1974)'s insight that people tend to confuse conditional probability with representativeness or similarity.

As we have mentioned, the probability measure μ in our representation is subjective; that is, it is extracted from the decision maker's similarity-based conditional probability assessments. In particular, μ is revealed from $\succsim$ on the collection $\{(A,B)|A \subset B\}$. Within the restricted domain, $B \cap C = \emptyset$ implies $(A \cup C, B) = (A, B)$, so the aforementioned departure from Bayesianism cannot occur. Therefore, μ is indeed the *correct* rational subjective probability.

Given that a decision maker fails to apply Bayes' rule, it is unclear whether her posterior can still be described as a function of any probability, not to mention the correct ones. In response to that, our results show that the correct subjective proba-

[4] Note that our similarity index S is ordinal in nature; that is, the decision maker compares the similarity ratio $S(A,C)/S(B,C)$ with 1. If $S(A,C)/S(B,C)$ is larger than 1, she concludes that A is more "likely"; if the similarity ratio is less than 1, then she concludes that B is more "likely".

bility can be used as a descriptive statistic of how agents update. In fact, μ is ordinally equivalent to $S(\cdot,\Omega)$; that is, μ describes the decision maker's relative rankings over unconditional events. Hence, if the decision maker does not err in comparing unconditional events, or is given directly the prior probabilities, μ is indeed the correct *objective* prior.

Our approach in the paper incorporates results from Savage (1954) and Villegas (1964) to obtain a nonatomic subjective probability. Then, we exploit the nonatomicity and apply a Debreu-type ordinal representation argument to prove the existence of a similarity index. Finally, we embed the similarity relation into a two-state Anscombe-Aumann framework with a novel mixture operation and prove that such an index must be a weighted geometric average of the conditional probabilities.

In section 2.2, we discuss previous axiomatization results for similarity indices. In section 2.3, we present our systems of axioms and show that they are equivalent to our (f,μ) representation. We also show that with one more robustness condition, f must be of the Cobb-Douglas form. Section 2.4 then concludes.

2.2 Related Literature

Psychologists have proposed many similarity indices in the context of pattern recognition. Within that literature, this paper is related to Tversky (1977) and Krantz, Luce, Suppes, and Tversky (1999).

Tversky (1977) models the similarity relation as an additive conjoint structure and proves that there exists a representation

$$S\left(A,B\right)=\theta f\left(A\cap B\right)-\alpha f\left(A\backslash B\right)-\beta f\left(B\backslash A\right)$$

for some $\theta, \alpha, \beta \geq 0$ and an interval scale f. In the paper, Tversky also proposes, albeit without axiomatization, the well-known Tversky index:

$$S_T(A, B) = \frac{f(A \cap B)}{f(A \cap B) + \alpha f(A \backslash B) + \beta f(B \backslash A)}$$

which encompasses a large class of similarity indices proposed by previous literature.[5]

In a similar vein, Krantz, Luce, Suppes, and Tversky (1999) axiomatize the following similarity index:

$$S(A, B) = \phi_1(A \cap B) - \phi_2(A \backslash B) - \phi_3(B \backslash A)$$

where ϕ_1, ϕ_2, ϕ_3 are functions that map into $[0, 1]$. They also provide conditions under which ϕ_i's are additive.

In our context, similarity is a substitute for conditional probability. Therefore, it is natural that $(A, A) \succsim (C, D)$ for any events A, C, D; that is, the decision maker assigns maximum similarity to any event with itself. Such a property precludes the linear representations in both Tversky (1977) and Krantz, Luce, Suppes, and Tversky (1999): both representations above imply that $S(A, A)$ is larger than $S(B, B)$ if A is more likely than B. In fact, the difficulty of the proofs in this paper arises from abandoning this linear structure. Without linearity, we are forced to use a Debreu-type ordinal representation argument for identification of a similarity index.

2.3 Model

In this section we first describe the primitives of our model. Then we introduce our axioms and show that they are equivalent to the (f, μ) representation; that is, the decision maker has a subjective prior μ and assesses $f(\mu(A|B), \mu(B|A))$ when asked

[5]In chapter 3, we provide an axiomatic foundation for S_T when $\alpha + \beta = 1$.

to evaluate $\mu(A|B)$. After that, we identify a robustness condition which ensures that f is of the Cobb-Douglas form.

Let $\mathcal{E}$ be a σ-algebra of events defined on state space Ω. We will use capital letters to denote generic events. Let $\mathcal{N} \subset \mathcal{E}$ be a σ-ideal, representing the collection of all null events. We consider a standard learning situation: our decision maker (DM) observes that an event occurs and then updates her beliefs over $(\Omega, \mathcal{E})$.

Following Kahneman and Tversky (1974), we assume that the DM mistakes the question "How likely is A given B?" for the question "How representative is A of B (i.e., how similar are A and B)?". In other words, the DM assesses the degree to which event A is similar to event B when asked to evaluate the probability of A conditioning on B.

Our DM's similarity assessment is summarized by $\succsim$, a binary relation defined on $\mathcal{E}^2\backslash\mathcal{N}^2$; that is, we will not consider similarity between null events. We write $(A,B) \succsim (C,D)$ if "A is more similar to B than C is to D". We follow the psychology literature and allow $\succsim$ to be asymmetric; that is, it is *not* necessarily true that $(A,B) \sim (B,A)$.[6] For each (A,B), we will call A the *stimulus* and B the *standard*.

Our decision maker mistakenly ranks the conditional likelihood of event pairs according to $\succsim$; that is, if $(A,B) \succsim (C,D)$, she concludes that A given B is more likely than C given D.

Definition. $(\Omega, \mathcal{E}, \mathcal{N}, \succsim)$ *is a similarity structure if*

(i) $\succsim$ *is complete and transitive;*

(ii) $(A,A) \succsim (B,C)$ *for all* A, B, C*;*

(iii) $C \cap D \in \mathcal{N} \iff (A,B) \succsim (C,D)$ *for all* A, B*.*

Hence, any nonnull event and itself are maximally similar. In addition, any pair of events with a null intersection is assigned the minimum degree of similarity.

[6]See Tversky (1977) for a detailed analysis.

Each pair (A, B) partitions $A \cup B$ into three subsets: the *stimulus distinction* $A \backslash B$, the *standard distinction* $B \backslash A$ and the *intersection* $A \cap B$. We will call the collection of these three events the *contrast partition* generated by (A, B). For expositional purpose, we explicitly specify each element of the contrast partition and

$$\text{write} \quad A \backslash B \ A \cap B \ B \backslash A \quad \text{in lieu of} \quad (A, B).$$

Hence, we write

$$ACB \succsim A'C'B' \quad \text{if} \quad (A \cup C, B \cup C) \succsim (A' \cup C', B' \cup C')$$

provided that the collections $\{A, B, C\}$ and $\{A', B', C'\}$ are both pairwise disjoint.

We consider only nondegenerate similarity structures; that is, Ω cannot belong to $\mathcal{N}$ and there exists at least one pair of events which have an intermediate degree of similarity between the maximum and minimum possible value.

Definition. *A similarity structure* $(\Omega, \mathcal{E}, \mathcal{N}, \succsim)$ *is nondegenerate if there exist* $\widetilde{A}, \widetilde{B}, \widetilde{C}$ *such that* $\emptyset\Omega\emptyset \succ \widetilde{A}\widetilde{C}\widetilde{B} \succ \widetilde{A}\emptyset\widetilde{B}$.

2.3.1 Similarity Index

In this subsection we present the axioms that we impose on $\succsim$ and show that the system is equivalent to the (f, μ) representation; that is, the decision maker has a subjective prior μ and assesses $f(\mu(A|B), \mu(B|A))$ when asked to evaluate $\mu(A|B)$.

Since Bayes' rule will be a special case of our representation (when $f(x, y) = x$ for all x, y), all axioms below are satisified by the Bayesian model. Among them, Axiom 2.1-2.3 are natural adaptations of respective axioms in Savage (1954) and Villegas (1964) that ensure the existence and uniqueness of a subjective prior.

Axiom 2.1. *(Additivity) For* $(A \cup B \cup C \cup D) \cap E = \emptyset$*, then*

$$ACB \succsim ADB \iff A(C \cup E)B \succsim A(D \cup E)B.$$

The additivity axiom states that if two pairs of events only differ in their intersections, then adding a disjoint subset to the intersection of both pairs does not change their relative similarity ranking. Take a dice as an example, if the event $\{1,2,3\}$ is more similar to $\{2,3,5\}$ than $\{1,2\}$ is to $\{2,5\}$, then $\{1,2,3,4\}$ should be more similar to $\{2,3,4,5\}$ than $\{1,2,4\}$ is to $\{2,4,5\}$ and vice versa; that is, adding $\{4\}$ to or subtraction $\{4\} from$the intersection of both pairs should not change their relative similarity ranking.

The following two axioms are for technical purposes. The nonatomicity axiom ensures that similarity assessments fill up a continuum on the real line. In particular, given two pairs with different degrees of similarity, one could always find an intermediate pair by shrinking the intersection of the more similar pair or shrinking the distinctions of the less similar pair. The countable additivity axiom ensures that $\succsim$ is Archimedean.

Axiom 2.2. *(Nonatomicity) Let* $\emptyset\Omega\emptyset \succ A'CB' \succ ADB$*. Then there is* $\widehat{C} \subset C$ *such that* $A'CB' \succ A'\widehat{C}B' \succ ADB$*. Moreover,* $\emptyset DB \succ ADB$ *(resp.* $AD\emptyset \succ ADB$*) implies there is* $\widehat{A} \subset A$ *(resp.* $\widehat{B} \subset B$*) such that* $A'CB' \succ \widehat{A}DB$ *(resp.* $AD\widehat{B}) \succ ADB$.

Axiom 2.3. *(Countable Additivity)* $AC_nB \succsim A'C'B' \succsim A_nCB_n$ *and* $A_{n+1} \subset A_n, B_{n+1} \subset B_n, C_{n+1} \subset C_n$ *implies* $A\left(\bigcap_n C_n\right)B \succsim A'C'B' \succsim (\bigcap_n A_n)C(\bigcap_n B_n)$.

Axioms 2.4-2.6 have no counterpart in Savage's theorem but again, they are satisfied by rational Bayesian updating.

Axiom 2.4. *(Separability) Let* $\emptyset\Omega\emptyset \succ A'CB'$*. Then* $ACB \succ ADB$ *implies* $A'CB' \succ A'DB'$.

The separability axiom states that when comparing the similarity of ACB with ADB, the agent cancels A and B from both pairs. Therefore in this case the relative similarity ranking between these two pairs depends only on the difference in intersections. The qualification $\emptyset\Omega\emptyset \succ A'CB'$ rules out the case in which the agent simply ignores the states in the distinctions A' and B'. In that case, the agent would find the difference between $A \cup D$ and $B' \cup D$ negligible and therefore assign maximum similarity to $A'DB'$.

Axiom 2.5. *(Scale Invariance) Let A, A', B, B', C, C' be pairwise disjoint. If $AC\emptyset \sim A'C'\emptyset$ and $\emptyset CB \sim \emptyset C'B'$, then $(A \cup A')(C \cup C')(B \cup B') \sim ACB$.*

Our idea behind this axiom is that only the relative salience among elements of the contrast partition should matter. For example, when ranking $AC\emptyset$ and $A'C'\emptyset$, the DM essentially compares the relative salience between distinction A and intersection C with the relative salience between A' and C'. Scale invariance implements such a notion by assuming that if the relative salience among A', C', B' matches the relative salience among A, C, B, then combining the standards and the stimuli respectively will give rise to an equivalent pair in the similarity hierarchy.

For a rational Bayesian, for pairwise disjoint A, A', B, B', C, C',

$$
\begin{aligned}
& \Pr(C|C \cup B) = \Pr(C'|C' \cup B') \\
\Longrightarrow \quad & \Pr(C \cup C'|(C \cup C') \cup (B' \cup B'')) = \Pr(C|C \cup B) \\
\Longrightarrow \quad & \Pr((A \cup A') \cup (C \cup C')|(C \cup C') \cup (B' \cup B'')) = \Pr(A \cup C|C \cup B).
\end{aligned}
$$

In other words, the requirement $AC\emptyset \sim A'C'\emptyset$ is redundant for a Bayesian. This is because as long as the standard distinction is empty, a Bayesian will always assign 1 to the posterior; that is, $\Pr(A \cup C|C) = \Pr(A' \cup C'|C') = 1$ always hold. This extra requirement reflects our notion that agents with a similarity-based updating procedure can value the stimulus distinction when evaluating the posterior.

We now introduce our main axiom, *monotonicity*. We know from Axiom 2.4 that when comparing ACB with ADB, the decision maker is essentially evaluating the relative salience of intersections C and D. Monotoncity states that if C is more salient than D, then C must also have a higher similarity cost when it appears as a distinction. Conversely, as in Bayes' rule, if C has a higher similarity cost than D as a standard distinction, it must be the case C is more salient than D as an intersection. To summarize, monotonicity states that states of the intersection and states of the distinctions work in opposite ways in determining similarity, which almost sounds tautological by definition of the word "similarity".

Axiom 2.6. *(Monotonicity) $ACB \succsim ADB$ implies $ABD \succsim ABC$ and $DAB \succsim CAB$. Moreover, if $B \notin \mathcal{N}$, $ABD \succsim ABC$ implies $ACB \succsim ADB$.*

Monotonicity is the key axiom that generates possible deviation from Bayes' rule. To see that, consider evaluating $\Pr(C \cup A|A \cup B)$ where A, B, C are pairwise disjoint. Bayesian subjects should only care about $\Pr(A)$ and $\Pr(B)$ since $\Pr(C \cup A|A \cup B) = \Pr(A)/(\Pr(A) + \Pr(B))$. In other words, she would always set $\emptyset AB \sim CAB$. Our monotonicity axiom, however, merely requires $\emptyset AB \succsim CAB$ because $ACB \succsim A\emptyset B$.[7] Together with the additivity axiom, monotonicity allows the decision maker to specify $C'AB \succ CAB$ if $C' \subset C$, as the pair of events $(C \cup A, A \cup B)$ differ more from each other than $(C' \cup A, A \cup B)$.

Given any probability measure μ, write $\mu(A|B) = \mu(A \cup B)/\mu(B)$ if $\mu(B) > 0$. If $\mu(B) = 0$, let $\mu(A|B) = 0$ for expositional purpose. Next, we introduce our first theorem.

Theorem 2.1. *A nondegenerate similarity structure $(\Omega, \mathcal{E}, \mathcal{N}, \succsim)$ satisfies Axiom 1-6 if and only if there exist a nonatomic probability μ and a nondecreasing function*

[7] In fact, if we impose $\emptyset AB \sim CAB$ in addition to Axiom 2.1-2.6, the only representation that survives is consistent with what Bayesian theory predicts.

$f : [0,1]^2 \longrightarrow [0,1]$ which is continuous on $(0,1]^2$ such that

$$(A,B) \succsim (C,D) \iff f\left(\mu\left(A|B\right), \mu\left(B|A\right)\right) \geq f\left(\mu\left(C|D\right), \mu\left(D|C\right)\right)$$

where (i) $\mu(A) = 0$ if and only if $A \in \mathcal{N}$ and (ii) $f(\cdot, y)$ is strictly increasing if $y > 0$.

The theorem states that Axioms 2.1-2.6 is equivalent to the (f, μ) representation; that is, the decision maker has a subjective prior μ over $(\Omega, \mathcal{E})$ and assigns a similarity index $f(\mu(A|B), \mu(B|A))$ to each pair of events (A, B). Such a similarity index $f(\mu(A|B), \mu(B|A))$ is mistakenly used by our behavioral decision maker as the answer when asked to evaluate the Bayesian conditional probability $\mu(A|B)$. In other words, our DM confuses $\mu(A|B)$ with $\mu(B|A)$.

We call an (f, μ) that satisfies the conditions in Theorem 2.1 a similarity representation of $(\Omega, \mathcal{E}, \mathcal{N}, \succsim)$. In the proof, we first incorporate results from Savage (1954) and Villegas (1964) to extract a nonatomic subjective probability from $\succsim$. Then, we exploit the nonatomicity and apply a Debreu-type ordinal representation argument to prove the existence of an aggregator and therefore a similarity index.

Note that the μ in Theorem 2.1 is the *correct* subjective unconditional prior. In particular, μ is revealed from $\succsim$ on the collection $\{(A, B)|A \subset B\}$. Within this restricted domain, $B \cap C = \emptyset$ implies $(A \cup C, B) = (A, B)$, so the aforementioned departure from Bayesianism cannot occur. Therefore, μ is indeed the *correct* subjective probability. That is, our model allows the decision maker to have the correct prior but commit non-Bayesian fallacies in updating.

Corollary. *Suppose (f, μ) is a similarity representation of a nondegenerate similarity structure $(\Omega, \mathcal{E}, \mathcal{N}, \succsim)$, then*

$$\mu(A) \geq \mu(B) \iff (A, \Omega) \succsim (B, \Omega).$$

Proof. Follows from (ii) in Theorem 2.1. □

The corollary states that μ is ordinally equivalent to $S(\cdot, \Omega)$; that is, μ describes the decision maker's relative rankings over unconditional events. Hence, if the decision maker does not err in comparing unconditional events, or is given directly the prior probabilities, μ is indeed the correct *objective* prior.

Given that a decision maker fails to apply Bayes' rule, it is unclear whether her posterior can still be described as a function of any probability, not to mention the correct ones. In response to that, the corollary above shows that the correct probability can be viewed as a descriptive statistic of how similarity-based agents update.

Our next theorem show that although the aggregator f is only identified ordinally, the subjective prior μ is unique.

Theorem 2.2. *Let (f, μ) be a similarity representation of a nondegenerate similarity structure $(\Omega, \mathcal{E}, \mathcal{N}, \succsim)$. Then (f', μ') is a similarity representation of $(\Omega, \mathcal{E}, \mathcal{N}, \succsim)$ if and only if $\mu = \mu'$ and there is a strictly increasing function g, continuous on $f((0,1]^2)$, such that $f' = g \circ f$.*

Proof. Suppose (f', μ') is a similarity representation. By the corollary we know that $\mu'(A) \geq \mu'(B)$ if and only if $(A, \Omega) \succsim (B, \Omega)$. Hence $\mu(A) \geq \mu(B)$ if and only if $\mu'(A) \geq \mu'(B)$. By the nonatomicity of μ, it is clear that $\mu = \mu'$. The second part is standard and therefore its proof is omitted. □

Hence, Theorem 2.1 provides a micro-foundation for behavioral subjects who know the correct prior but confuse $\mu(A|B)$ with $\mu(B|A)$. For example, subjects who answer 0.8 to the taxicab problem confuse $\mu(\text{Blue}|\text{Testimonial})$ with $\mu(\text{Testimonal}|\text{Blue}) = 0.8$. In the next subsection, we characterize a special class of aggregators and then discuss the experimental evidence in detail.

2.3.2 The Cobb-Douglas Index

In this subsection we take the similarity representation (f, μ) as given and impose a robustness condition on $\succsim$ which ensures that f has the Cobb-Douglas form. In particular, we assume that the relative similarity ranking between two pairs of events does not change when we intersect both standards or both stimuli with an independent event. In the context of the taxicab experient, if one believes that the culprit is more likely to drive a blue cab than a green one given witness' testimonal, she will also believe that the cab is more likely to be a blue ford than a green ford. Moreover, had the witness said additionally that the cab was a ford, the DM will still conclude that the cab was more likely to be blue.[8] Formally, we define independence and then robustness as follows.

Definition. *$A, B \in \mathcal{E}$ are **independent**, denoted $A \perp B$, if $\mu(A \cap B) = \mu(A)\mu(B)$.*

For any collection of subsets $\mathcal{A} \subset \mathcal{E}$, we say that $A \perp \mathcal{A}$ if $A \perp B$ for any $B \in \mathcal{A}$. Also, let $\sigma(\mathcal{A})$ be the smallest σ-algebra that contains all of the elements in $\mathcal{A}$.

Definition. *$\succsim$ is said to be **robust** if $(A, B) \succsim (A', B)$ implies $(A \cap C, B) \succsim (A' \cap C, B)$ and $(A, B \cap C) \succsim (A', B \cap C)$ for any $C \perp \sigma(A, A', B)$.*

With robustness, we can embed our similarity representation in a two-state generalized Anscombe-Aumann framework with a nonstandard mixture operation. We summarize our methodology in Chapter 3. The two states of the world can be interpreted as the state in which B caused A and the state in which A caused B. Conditional probabilities $\mu(A|B)$ and $\mu(B|A)$ are viewed as lotteries in the corresponding state. Then, for each pair of events (A, B), $(\mu(A|B), \mu(B|A))$ is an Anscombe-Aumann act that maps subjective states into lotteries. Our aggregator f, in turn, implies a complete and transitive ranking over all these Anscombe-Aumann acts. Ap-

[8] Of course, this is assuming that colors and brands are independent.

plying a generalized version of Anscombe and Aumann (1963)'s theorem, we arrive at the following representation.

Theorem 2.3. *Suppose (f, μ) is a similarity representation of nondegenerate similarity structure $(\Omega, \mathcal{E}, \mathcal{N}, \succsim)$. Then $\succsim$ is robust if and only if there is a unique $\alpha \in (0, 1]$ such that*

$$(A, B) \succsim (C, D) \iff \mu(A|B)^{\alpha}\, \mu(B|A)^{1-\alpha} \geq \mu(C|D)^{\alpha}\, \mu(D|C)^{1-\alpha}.$$

Consider evaluating whether A or B is more likely conditioning on C. In this case, our behavioral decision maker is comparing $S(A, C)$ with $S(B, C)$. With the Cobb-Douglas index, we have

$$\frac{S(A,C)}{S(B,C)} = \frac{\mu(A|C)^{\alpha}\mu(C|A)^{1-\alpha}}{\mu(B|C)^{\alpha}\mu(C|B)^{1-\alpha}} = \frac{\mu(C|A)}{\mu(C|B)} \cdot \left(\frac{\mu(A)}{\mu(B)}\right)^{\alpha}.$$

Since $\alpha \in (0, 1]$, $(\mu(A)/\mu(B))^{\alpha}$ is closer to 1 than the correct prior odds ratio $\mu(A)/\mu(B)$. That is, α specifies the extent to which the decision maker neglects the prior odds ratio; when $\alpha = 1$, our DM is Bayesian. Note that our similarity index is ordinal in nature; that is, the decision maker compares the similarity ratio $S(A, C)/S(B, C)$ with 1. If $S(A, C)/S(B, C)$ is larger than 1, she concludes that A is more "likely"; if the similarity ratio is less than 1, then she concludes that B is more "likely".

In the taxicab problem, the likelihood ratio of the witness' signal (correct over incorrect) is $80/20 = 4$ where as the base rate (blue over green) is $15/85 = 0.176$. A decision maker with an $\alpha < 0.8$ would conclude, like the experimental subjects, that the hit and run cab was more likely to be blue than green.

In the Linda problem, although $\Pr(\mathrm{T} \wedge \mathrm{F}|\mathrm{D}) < \Pr(\mathrm{T}|\mathrm{D})$, arguably we have $\Pr(\mathrm{D}|\mathrm{T} \wedge \mathrm{F}) > \Pr(\mathrm{D}|\mathrm{T})$; that is, Linda is more likely to match the stereotypical description D if she is a feminist bank teller than if she is just a bank teller. Then, if α is low, an α-mixture of Pr(T∧F|D) and Pr(D|T∧F) may well be higher than an α-mixture of Pr(T|D) and Pr(D|T). In that case a decision maker would choose the conjunctive description T∧F as the more "probable" option. A similar logic applies to the results in Bar-Hillel and Neter (1993).

2.4 Conclusion

In this paper, we provided a framework for analyzing a range of well-documented non-Bayesian behaviors including base rate neglect, conjunction fallacy and disjunction fallacy. We assumed that our decision maker mistakenly assesses the similarity of A to B when evaluating the probability of A given B.

Our similarity-based updating procedure allows $A \cup C$ given B to be less likely than A given B if $B \cap C = \emptyset$, simply because the pair of events $A \cup C$ and B differ more from each other. By allowing for this type of similarity-based departure from Bayesian updating, the axioms that we proposed yield an updating formula that is a Cobb-Douglas weighted geometric mean of $\mu(A|B)$ and $\mu(B|A)$, where μ is the correct rational subjective probability and $\mu(\cdot|\cdot)$ is given by Bayes' rule. That is, we have provided a micro-foundation for behavioral subjects who confuse these two conditional probabilities but have correct unconditional beliefs.

In the proof, we applied a generalized version of the Anscombe-Aumann Theorem to show that the aggregator of the conditional probabilities must be of the Cobb-Douglas form. We will show in the next chapter that this method could be applied to micro-found a large class of means.

2.5 Appendix

2.5.1 Proof of Theorem 2.1

In this section, we present our proof of Theorem 2.1. To show the "only if" part, we first incorporates results from Savage (1954) and Villegas (1964) to obtain a nonatomic subjective probability. Then, we exploit the nonatomicity and apply a Debreu-type ordinal representation argument to prove the existence of an aggregator and therefore a similarity index.

First of all, we introduce Savage (1954) and Villegas (1964)'s result for subjective probabilities representations, which we will cite from time to time in the proof. Let $\succsim^*$ be a binary relation defined on a σ-algebra $\mathcal{A}$ of subsets of Ω.

Definition. $\succsim^*$ *is a* ***Savagian qualitative probability*** *on* $\mathcal{A}$ *if it is a preference relation satisfying the following axioms,*

(i) $\Omega \succ^* \emptyset$,

(ii) $A \succsim^* \emptyset$,

(iii) $A \succsim^* B$ *iff.* $A \cup C \succsim^* B \cup C$ *for* $C \cap (A \cup B) = \emptyset$,

(iv) $A_n \succsim^* B$ *and* $A_{n+1} \subset A_n$ *implies that* $\bigcap_n A_n \succsim^* B$,

(v) $A \succ^* B$ *implies there is* $A' \subset A$ *such that* $A \succ^* A' \succ^* B$.

Savage (1954) shows that if a qualitative probability (i.e. satisifes (i),(ii) and (iii)) is fine[9] and tight[10], then there exists a unique finitely additive probability representation. Villegas (1964) then identifies (iv) as a necessary and sufficient condition for such probability representation to be countably additive. Moreover, he note that

[9] $\succsim$ is fine if given any $A \succ \emptyset$ there is a finite partition of Ω, $\{\Omega_n\}$, such that $A \succ \Omega_n$ for all n.

[10] $\succsim$ is tight if $A \sim B$ for all A, B such that $A \cup A' \succsim B$ and $B \cup B' \succsim A$ for all nonnull A', B' with $A \cap A' = B \cap B' = \emptyset$.

with (iv), (v) is enough to ensure fineness and tightness, hence a unique probability representation.[11]

Lemma A. *(Savage (1954);Villegas (1964))* $\succsim^*$ *is a Savagian qualitative probability if and only if there is a unique nonatomic probability* μ *such that*

$$A \succsim^* B \Longleftrightarrow \mu(A) \geq \mu(B).$$

Since μ is nonatomic and countably additive, it is also *convex-ranged*; that is, for any $a \in (0,1)$ and A such that $\mu(A) > 0$, there is $B \subset A$ such that $\mu(B) = a\mu(A)$. We will exploit this property throughout the paper to identify suitable events.

Throughout this subsection, it is assumed that $(\Omega, \mathcal{E}, \mathcal{N}, \succsim)$ is a nondegenerate similarity structure that satisfies Axiom 2.1-2.6. Since $\mathcal{N}$ is a σ-ideal, $ACB \in \mathcal{E}^2 \backslash \mathcal{N}^2$ if not all of A, B, C are null. Write $\overline{C} \equiv \Omega \backslash C$. Since $\Omega \notin \mathcal{N}$, if $C \in \mathcal{N}$, then $\overline{C} \notin \mathcal{N}$. Also note that by (iii) in the definition of a similarity structure, if $C \in \mathcal{N}$ then $ACB \sim A\emptyset B$. Conversely, if $ACB \sim A\emptyset B$ for some A, B, by nondegeneracy and separability, $C \in \mathcal{N}$.

We first consider the cases when maximum similarity is assigned.

Lemma 2.1. $A, B \in \mathcal{N}$ *implies* $ACB \sim \emptyset C \emptyset$.

Proof. If $A, B \in \mathcal{N}$, then $CAB \sim C\emptyset B$ and $\emptyset BC \sim \emptyset\emptyset C$. Therefore by monotonicity we have $ACB \sim \emptyset CB$ and $\emptyset CB \sim \emptyset C \emptyset$ and we are done. □

Lemma 2.2. *If* $C' \subset C$*, then* $ACB \succsim AC'B, C'AB \succsim CAB$ *and* $ABC' \succsim ABC$*.*

Proof. By Lemma 2.1 the case when $A, B \in \mathcal{N}$ is trivial. Now suppose $A\emptyset B$ is in our domain. By definition $A(C \backslash C')B \succsim A\emptyset B$. Then Axiom 1 implies that $ACB \succsim AC'B$. Then monotonicty proves the result. □

[11] In particular, with (i)-(v) we have enough structure to partition any event into two equally probable ones so that we are able to assign dyadic probabilities to such partitions. Then a continuity argument derived from (iv) completes the subjective probability assignment.

Lemma 2.3. *$ACB \sim \emptyset C \emptyset$ implies $B \in \mathcal{N}$.*

Proof. Suppose $A \cup B$ are nonnull but $\emptyset C \emptyset \sim ACB$. Lemma 2.2 implies that $\emptyset C \emptyset \succsim AC\emptyset \succsim ACB \sim \emptyset C \emptyset$. Clearly $C \notin \mathcal{N}$. Then by monotonocity, $A\emptyset C \sim ABC$. Suppose there is A', C' such that $A'BC' \succ A'\emptyset C'$, then since $\emptyset \Omega \emptyset \succ ABC$, by separability, $ABC \succ A\emptyset C$ delivering the desired contradiction. Hence B is null. □

For A such that both A and $\overline{A}$ are nonnull, define $C \succsim_A D$ if $\emptyset CA \succsim \emptyset DA$. Note that $\succsim_A$ is not complete in the space $\mathcal{E}$. However, it is complete and transitive in $\mathcal{E} \cap 2^{\overline{A}}$; that is, measurable subsets of $\overline{A}$. Then by Axiom 2.1,2.2 and 2.3, $\succsim_A$ is a Savagian qualitative probability. Then by Lemma A there exists a unique nonatomic probability μ_A such that $\mu_A(C) \geq \mu_A(D)$ if and only if $C \succsim_A D$. Note that $\mu_A(B) = 0$ if only if $B \in \mathcal{N}$ and $B \subset A$.

Now, define $\succsim^*$ as follows. On $\mathcal{R} \equiv \{A \in \mathcal{E} | \overline{A} \notin \mathcal{N}\}$, define $C \succsim^* D$ if there exist nonnull events A, A', B with $A \sim_B A'$ such that $\emptyset CA \succsim \emptyset DA'$. Then, add $\mathcal{E} \backslash \mathcal{R}$ as the single highest equivalence class.

Next, we prove that $\succsim^*$ is a Savagian qualitative probability. We proceed by first introducing the following two lemmas.

Lemma 2.4. *If $ACB \sim ADB$ and B is nonnull, then $A'CB' \sim A'DB'$, $CA'B' \sim DA'B'$ and $A'B'C \sim A'B'D$.*

Proof. We prove that $A'CB' \sim A'DB'$ and the rest is given by monotonicity. Since B is nonnull we know that $\emptyset \Omega \emptyset \succ ACB$. Therefore if $A'CB' \succ A'DB'$, then by separability we have $ACB \succ ADB$, a contradiction. Hence $A'DB' \succsim A'CB'$. Similarly, $A'CB' \succsim A'DB'$. The rest is by monotonicity. □

Lemma 2.5. *For mutually exclusive $A, B, C \in \mathcal{R} \backslash \mathcal{N}$, let $A_n \subset A, B_n \subset B, C_n \subset C$ and $\mu_{\overline{A}}(A_n) = \mu_{\overline{B}}(B_n) = \mu_{\overline{C}}(C_n) = 2^{-n}$, then $ACB \sim A_nC_nB_n$, $\emptyset CB \sim \emptyset C_nB_n$ and $AC\emptyset \sim A_nC_n\emptyset$.*

Proof. Consider any partitions $\{A_n^m\},\{B_n^m\},\{C_n^m\}$ of respectively A, B, C such that for all m, $\mu_{\overline{A}}(A_n^m) = \mu_{\overline{B}}(B_n^m) = \mu_{\overline{C}}(C_n^m) = 2^{-n}$. By nonatomicity such partitions always exist. Clearly by the previous lemma,

$$\emptyset C_n^m B_n^m \sim \emptyset C_n^{m'} B_n^m \sim \emptyset C_n^{m'} B_n^{m'}.$$

Similarly $A_n^m C_n^m \emptyset \sim A_n^{m'} C_n^{m'} \emptyset$. Inductively applying scale invariance gives the result.

□

Lemma 2.5 states that we could always shrink the events proportionately without changing the similarity assessment. This is essential since if C and D are so large that $\overline{C \cup D}$ is null, it is not possible to find A, B such that ACB and ADB are both in our domain. In that case we cannot extract the qualitative probability preference between C and D. Lemma 5, however, gives us the luxury to shrink C and D so that $\overline{C \cup D}$ is no longer null.

The next lemma shows that $\succsim^*$ is well-defined.

Lemma 2.6. *Let $C, D \in \mathcal{R}$. If $C \succsim^* D$, then for **any** nonnull events $A \subset \overline{C}, A' \subset \overline{D}, B$ with $A \sim_B A'$, $\emptyset CA \succsim \emptyset DA'$.*

Proof. By definition there are nonnull events $\widehat{A}, \widehat{A}', \widehat{B}$ with $\widehat{A} \sim_{\widehat{B}} \widehat{A}'$ such that $\emptyset C\widehat{A} \succsim \emptyset D\widehat{A}'$. If D is null the result is trivial. If D is nonnull so is C. Let $\{C_n\}$ and $\{D_n\}$ be sequences of sets such that $C_0 = C, D_0 = D$, $C_{n+1} \subset C_n, D_{n+1} \subset D_n$ and

$$\frac{\mu_{\overline{C}}(C_n)}{\mu_{\overline{C}}(C_{n+1})} = \frac{\mu_{\overline{D}}(D_n)}{\mu_{\overline{C}}(D_{n+1})} = 2.$$

If $\widehat{A} \cap D$ is null, in other words $\mu_{\widehat{B}}(\widehat{A}) = \mu_{\widehat{B}}(\widehat{A} \backslash D)$, then

$$\emptyset C\widehat{A} \sim \emptyset C(\widehat{A} \backslash D) \succsim \emptyset D\widehat{A}' \sim \emptyset D(\widehat{A} \backslash D).$$

Let $n = 0$ and $\widehat{A}^* = \widehat{A}\backslash D$. If $\widehat{A} \cap D$ is nonnull, pick n large enough such that there is $\widehat{A}^* \subset (\widehat{A} \cap D)\backslash D_n, \widehat{A}'^* \subset \widehat{A}'$ such that

$$\mu_{\overline{\widehat{A}}}(\widehat{A}^*) = \mu_{\overline{\widehat{A}'}}(\widehat{A}'^*) = \frac{1}{2^n}.$$

This is possible by the convexed-rangeness of the probabilities extracted above. Hence we have $\emptyset C_n \widehat{A}^* \succsim \emptyset D_n \widehat{A}'^* \sim \emptyset D_n \widehat{A}^*$. Using the same procedure we obtain $\emptyset C_m A^* \sim \emptyset CA$ and $\emptyset D_m A^* \sim \emptyset DA'$. Assume that $m = n$. This is wlog since we could always shrink the larger further with the preference unchanged. It is clear by separability that $\emptyset C_m A^* \succsim \emptyset D_m A^*$ and we complete the proof. $\square$

With the above lemma, $C \succ^* D$ if there exist A, A', B nonnull with $A \sim_B A'$ such that $\emptyset CA \succ \emptyset DA'$. Moreover, given the procedure in the proof, if $C \succsim^* D$ we could wlog assume that there is nonnull A such that $\emptyset CA \succsim \emptyset DA$. Then, we show that $\succsim^*$ is complete and transitive.

Lemma 2.7. $\succsim^*$ *is complete and transitive.*

Proof. It suffices to show that $\succsim^*$ is complete and transitive in $\mathcal{R}$.

(i) Completeness. It suffices to show that for any C, D in $\mathcal{R}$ there exists $A \subset \overline{C}, A' \subset \overline{D}$ and $B \subset \overline{A \cup A'}$ such that $A \sim_B A'$.

Case 1: $C \cup D$ is null. It is clear that C, D are both null. Then $\emptyset CA \sim \emptyset DA'$ for any A, A'. Consdier $\widetilde{A}, \widetilde{B}, \widetilde{C}$ such that $\emptyset \Omega \emptyset \succ \widetilde{A}\widetilde{C}\widetilde{B} \succ \widetilde{A}\emptyset\widetilde{B}$. Let $E = \widetilde{C} \cap \overline{C \cup D}$. It is easy to see that E and $\overline{E}$ are both nonnull. Therefore $E \sim_{\overline{E}} E$ and we are done.

Case 2: $C \cup D$ and $\overline{C \cup D}$ are both nonnull. Then $\overline{C \cup D} \sim_{C \cup D} \overline{C \cup D}$ and we are done.

Case 3: $C \cup D$ is nonnull but $\overline{C \cup D}$ is null. Clearly $C\backslash D$ and $D\backslash C$ are both nonnull since $C, D \in \mathcal{R}$. Pick $E \subset C\backslash D, E' \subset D\backslash C$ such that $\mu_{\overline{C\backslash D}}(E) = \mu_{\overline{D\backslash C}}(E') = \frac{1}{2}$. Clearly $C\backslash(D \cup E)$ is nonnull since $\mu_{\overline{C\backslash D}}(C\backslash(D \cup E)) = \frac{1}{2}$. If $E \sim_{C\backslash(D\cup E)} E'$ we are done. If $E \succ_{C\setminus(D\cup E)} E'$ by the convex-rangedness of $\mu_{C\setminus(D\cup E)}$ there is $E'' \subset E$ such that $E'' \sim_{C\setminus(D\cup E)} E'$.

(ii) Transitivity. Let $C \succsim^* D$ and $D \succsim^* E$. The case when at least one of them is null is trivial. So assume that C, D, E are all nonnull. By definition there are $A \sim_B A', \widehat{A} \sim_{\widehat{B}} \widehat{A}'$ such that $\emptyset CA \succsim \emptyset DA'$ and $\emptyset D\widehat{A} \succsim \emptyset E\widehat{A}'$. By the procedure in the previous lemma we could wlog assume that $A = A'$ and $\widehat{A} = \widehat{A}'$. Also assume that $\mu_{\overline{D}}(A) > \mu_{\overline{D}}(\widehat{A})$. Pick $A^* \subset A$ such that $\mu_{\overline{D}}(A^*) = \mu_{\overline{D}}(\widehat{A})$. Then by Lemma 4 and 6,

$$\emptyset CA^* \succsim \emptyset DA^* \sim \emptyset D\widehat{A} \succsim \emptyset E\widehat{A}$$

and we are done. $\square$

The next lemma proves that $\succsim^*$ is a qualitative probability defined on $\mathcal{E}$.

Lemma 2.8. *(i)* $\Omega \succ^* \emptyset$*; (ii)* $C \succsim^* \emptyset$*; (iii)* $(C \cup D) \cap E = \emptyset$ *implies* $[C \succsim^* D \iff C \cup E \succsim^* D \cup E]$.

Proof. (i) and (ii) are by construction. (iii) If $C \cup E, D \cup E \in \mathcal{R}$, we appeal to the procedure in the proof of Lemma 6 and the additivity axiom. If $C \notin \mathcal{R}$ or $D \notin \mathcal{R}$ both directions are trivial. Now assume that $C \in \mathcal{R}$ and $D \in \mathcal{R}$. Suppose $C \cup E \sim^* \Omega$. The "$\Longrightarrow$" direction is trivial. For the other direction, since $D\setminus C$ is null, we have $C \succsim^* C \cap D \sim^* D$. Suppose $C \cup E \in \mathcal{R}$. The "$\Longleftarrow$" direction is clear since $D \cup E \in \mathcal{R}$ and it reduces to the first case that we have considered. For the other direction, we show that if $C \succsim^* D$ and $C \cup E \in \mathcal{R}$ it must also be that $D \cup E \in \mathcal{R}$. Suppose $D \cup E \sim^* \Omega$. We know that $C\setminus(D \cup E) = C\setminus D$ is null. Also, since $C \cup E \in \mathcal{R}$, $D\setminus(C \cup E) = D\setminus C$ is nonnull. Then we must have $D \succ^* C$ delivering the desired contradiction. $\square$

Then, we proceed by showing that $\succsim^*$ is indeed a Savagian qualitative probability.

Lemma 2.9. *(i)* $C \succ^* D$ *implies there exist* $C' \subset C$ *such that* $C \succ^* C' \succ^* D$.

(ii) $C_n \succsim^* D$ *and* $C_{n+1} \subset C_n$ *for all* n *implies that* $\bigcap_n C_n \succsim^* D$.

Proof. (i) If $C \in \mathcal{R}$, it is implied by the nonatomicity axiom. If $C \notin \mathcal{R}$, then by construction $D \in \mathcal{R}$. If D is null, set $C' = C \cap \widetilde{C}$. Since $\widetilde{C}$ is nonnull and $\overline{C}$ is null, C' must be nonnull. Moreover since $\widetilde{C} \in \mathcal{R}$, $\overline{C'}$ cannot be null. Hence $C \succ^* C' \succ^* D$. Now suppose D is nonnull, then $C \cap D$ is nonnull. Since $D \backslash C$ and $\overline{D} \backslash C$ are null we have

$$\emptyset D\overline{D} \sim \emptyset(C \cap D)\overline{D} \sim \emptyset(C \cap D)\left(C \cap \overline{D}\right).$$

Pick $E \subset C \cap \overline{D}$ such that E and $C \cap \overline{D} \backslash E$ are both nonnull. Let $C' = (C \cap D) \cup E$ and we are done.

(ii) It suffices to prove that if $C_n \in \mathcal{E} \backslash \mathcal{R}$ and $C_{n+1} \subset C_n$, then $\bigcap_n C_n \in \mathcal{E} \backslash \mathcal{R}$. It suffices to prove that if D_n are null and $D_n \subset D_{n+1}$, $\bigcup_n D_n$ is null. This is directly given by the fact that $\mathcal{N}$ is a σ-ideal. □

Hence, by Lemma A, there exists a unique nonatomic probability μ defined on $\mathcal{E}$ such that $\mu(C) \geq \mu(D)$ if and only if $C \succsim^* D$. Clearly, $\mu(A) = 0$ if and only if $A \in \mathcal{N}$.

Lemma 2.10. $ACB \sim A'C'B'$ *if* $\mu(A) = \mu(A'), \mu(B) = \mu(B')$ *and* $\mu(C) = \mu(C')$.

Proof. Assume that A, B, C are all nonnull. The cases when any of them is null is similar. By Lemma 5, wlog we could assume that A, B, C, A', B', C' are mutually exclusive. Then we know that

$$ACB \sim A'CB \sim A'C'B \sim A'C'B'$$

by monotonicity and Lemma 2.4. □

Lemma 2.10 implies that if $\succsim$ has a representation, then the representation can only depends on $\mu(A)$, $\mu(B)$ and $\mu(C)$. Since we have abandoned linearity, we are not able to apply the techniques from linear conjoint measurement. Instead, a Debreuian argument is needed for identification of the representation. In particular, we need to show that there exists a countable order dense set. We define order denseness as follows.

Definition. *For any binary relation $\succsim$ on X, we say that the subset Y is $\succsim$-order dense if for any $x, y \in X$ such that $x \succ y$, there exists $z \in Y$ such that $x \succsim z \succsim y$.*

Then we introduce Debreu (1954)'s result as a lemma. Then we show that a representation for our similarity rankings exists.

Lemma B. *(Debreu (1954)) For any set X and binary relation $\succsim$ on X, there exists a function U that represents $\succsim$ if and only if $\succsim$ is a preference relation and X has a countable $\succsim$-order dense subset.*

Lemma 2.11. *$\mathcal{E}^2 \backslash \mathcal{N}^2$ has a countable $\succsim$-order dense subset.*

Proof. By the convex-rangedness of μ, we construct the order dense subset as follows. Pick A_0, B_0, C_0 mutually exclusive such that $A_0 \cup B_0 \cup C_0 = \Omega$ and $\mu(A) = \mu(B) = \mu(C) = 1/3$. Let $\{A_{mn}\}, \{B_{mn}\}, \{C_{mn}\}$ be subsets of respectively A, B, C, such that

$$\mu(A_{mn}) = \mu(B_{mn}) = \mu(C_{mn}) = \frac{m}{n}$$

for $m = 0, 1, 2, \ldots, \lfloor n/3 \rfloor$ and $n = 1, 2, 3, \ldots$

Next, we show that $\{ACB | A \in \{A_{mn}\}, B \in \{B_{mn}\}, A \in \{C_{mn}\}\}$ is a $\succsim$-order dense subset.

Let $A'C'B' \succ ACB$. Wlog assume that $\mu(A), \mu(B), \mu(C), \mu(A'), \mu(B'), \mu(C') \leq 1/3$. If $\emptyset\Omega\emptyset \sim A'C'B'$ we have $A'C'B' \sim A_{01}C_{13}B_{01} \succ ACB$ and we are done.

Suppose $\emptyset\Omega\emptyset \succ A'C'B'$, then by the nonatomicity axiom, there is $\widehat{C} \subset C'$ such that $A'C'B' \succ A'\widehat{C}B' \succ ACB$. Pick m_1, n_1 such that

$$\mu(C') > \frac{m_1}{n_1} > \mu(\widehat{C}).$$

Pick $C^* \subset C'$ such that $\mu(C^*) = m_1/n_1$. It follows that

$$A'C'B' \succ A'C^*B' \succ A'\widehat{C}B'.$$

If $\emptyset C^*B' \sim A'C^*B'$ then set $A^* = \emptyset$ and $m_2 = 0, n_2 = 1$. Suppose $\emptyset C^*B' \succ A'C^*B'$, then by nonatomicity, there is $\widehat{A} \subset A'$ such that

$$A'C'B' \succ \widehat{A}C^*B' \succ A'C^*B'.$$

Pick m_2, n_2 such that

$$\mu(A') > \frac{m_2}{n_2} > \mu(\widehat{A})$$

and $A^* \subset A'$ such that $\mu(A^*) = m_2/n_2$. Then by monotonicity

$$A'C'B' \succ \widehat{A}C^*B' \succsim A^*C^*B' \succsim A'C^*B' \succ ACB.$$

If $A^*C^*\emptyset \sim A^*C^*B'$ then set $B^* = \emptyset$ and $m_3 = 0, n_3 = 1$. Suppose $A^*C^*\emptyset \succ A^*C^*B'$, again by the nonatomicity axiom there is $\widehat{B} \subset B'$ such that

$$A'C'B' \succ A^*C^*\widehat{B} \succ A^*C^*B' \succ ACB.$$

Then similarly pick $\mu(B^*) = m_3/n_3$ such that

$$A'C'B' \succ A^*C^*\widehat{B} \succsim A^*C^*B^* \succsim A^*C^*B' \succ ACB.$$

Therefore

$$A'C'B' \succ A^*C^*B^* \succ ACB$$

and

$$A^*C^*B^* \sim A_{m_2n_2}C_{m_1n_1}B_{m_3n_3}.$$

Thus, we have proved the result. $\square$

Hence $\succsim$ has a numerical representation $S(A, B)$. By Lemma 2.10 we know that $S(A, B) = h(\mu(A\backslash B), \mu(A\cap B), \mu(B\backslash A))$. Then we characterize how h should behave.

Definition. *A function $f : D \subset \mathbb{R}^n \to \mathbb{R}$ is separately continuous if for any fixed n-tuple $(x_1, x_2, \ldots, x_n) \in D$ and for every $i \in \{1, 2, \ldots, n\}$, the mapping $t \mapsto f(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_n)$ is continuous for all t such that $(x_1, \ldots, x_{i-1}, t, x_{i+1}, \ldots, x_n) \in D$.*

Lemma C. *(Kruse and Deely (1969)) Let $f(x_1, x_2, ..., x_{n-1}, y)$ be a real-valued function defined on an open set G in $\mathbb{R}^n, n \geq 2$. Suppose f is separately continuous and is monotone in each x_i separately, $1 \leq i \leq n-1$. Then f is continuous on G.*

By Lemma C the following characterizations of h is true. Let Δ^n be the n-simplex in $\mathbb{R}^n$.

Lemma 2.12. *There exists a function $h : \Delta^3\backslash\{(0, 0, 0)\} \longrightarrow [0, 1]$ such that*

$$ACB \succsim A'C'B' \Longleftrightarrow h(\mu(A), \mu(C), \mu(B)) \geq h(\mu(A'), \mu(C'), \mu(B')).$$

where

(i) $h(a, 0, b) = 0$, $h(0, c, 0) = 1$;

(ii) $h(a, \cdot, b)$ is nondecreasing;

(iii) $h(\cdot, c, b)$ and $h(a, c, \cdot)$ are nondecreasing;

(iv) h is continuous at (a, c, b) if $0 < c < 1$;

(v) $h(a, c, b) = h(ka, kc, kb)$ for $k \in (0, 1)$ if $0 < c < 1$.

Proof. (i) is simple rescaling. (ii) is given by additivity. (iii) is given by monotonicity.

Next we prove (iv) by construction. By (ii), fixed any a, b, $h(a, \cdot, b)$ is monotone and hence continuous almost everywhere. Let $\{x_n\}$ be the discontinuity points of

$h(a,\cdot,b)$. we prove that there cannot exist $(a',b',c')\in\Delta^3$ such that $h(a,x_n^+,b)\geq h(a',c',b')\geq h(a,x_n^-,b)$ with $h(a',b',c')\neq h(a,x_n,b)$. Consider first if

$$h\left(a,x_n^+,b\right)\geq h(a',c',b')>h(a,x_n,b)\geq h\left(a,x_n^-,b\right).$$

Pick any decreasing sequence $\{x_{n,m}\}\in[0,1-a-b]$ such that $x_{n,m}\to x_n$ as $m\to\infty$. There are A,B,A',B',C' such that $\mu(A)=a,\mu(B)=b,,\mu(A')=a',\mu(B')=b'$ and $\mu(C')=c'$. Also pick $\{C_m\}$ such that $C_{m+1}\subset C_m$ and $\mu(C_m)=x_{n,m}$. Hence

$$\lim_{m\to\infty}\mu(C_m)=\mu\left(\bigcap_m C_m\right)=x_n.$$

Therefore $AC_mB\succsim A'C'B'$ however $A'C'B'\succ A(\bigcap_m C_m)B$, a contradiction to Axiom 6. For the other case, that is when

$$h\left(a,x_n^+,b\right)\geq h(a,x_n,b)>h(a',c',b')\geq h\left(a,x_n^-,b\right),$$

it suffices to prove $A'C'B'\succsim AC_mB$ and $C_m\subset C_{m+1}$ implies $A'C'B'\succsim A(\bigcup_m C_m)B$. Suppose $A(\bigcup_m C_m)B\succ A'C'B'$. There exists $C\subset\bigcup_m C_m$ such that $A(\bigcup_m C_m)B\succ ACB\succ A'C'B'$. Therefore there is m^* such that

$$\mu\left(\bigcup_m C_m\right)>\mu(C_{m^*})>\mu(C).$$

By (ii)

$$A(\bigcup_m C_m)B\succsim AC_{m^*}B\succsim ACB\succ A'C'B'$$

which is a contradiction.

By (iii), $h(\cdot,c,b)$ is monotone therefore is continuous almost everywhere. Let $\{y_n\}$ be the discontinuity points of $h(\cdot,c,b)$. The next step is prove that there cannot exist $(a',b',c')\in\Delta^3$ such that $h(y_n^-,c,b)\geq h(a',c',b')\geq h(y_n^+,c,b)$ with $h(a',c',b')\neq$

$h\left(y_n, c, b\right)$. Suppose

$$h\left(y_n^-, c, b\right) \geq h\left(a', c', b'\right) > h\left(y_n, c, b\right) \geq h\left(y_n^+, c, b\right).$$

By the nonatomicty axiom there is $\widehat{y}_n < y_n$ such that

$$h\left(a', c', b'\right) > h\left(\widehat{y}_n, c, b\right) > h\left(y_n, c, b\right)$$

which is a contradiction. Suppose

$$h\left(y_n^-, c, b\right) \geq h\left(y_n, c, b\right) > h\left(a', c', b'\right) \geq h\left(y_n^+, c, b\right).$$

Pick any increasing sequence $\{y_{n,m}\} \in [0, 1-c-b]$ such that $y_{n,m} \to y_n$ as $m \to \infty$. There are B, C, A', B', C' such that $\mu(B) = b, \mu(C) = c, \mu(A') = a', \mu(B') = b'$ and $\mu(C') = c'$. Also pick $\{A_m\}$ such that $A_{m+1} \subset A_m$ and $\mu(A_m) = y_{n,m}$. It follows that

$$(\bigcap_m A_m)CB \succ A'C'B' \succ A_m CB.$$

which is a contradiction to countable-addivity. Similar results obtain for $h(a, c, \cdot)$.

Consider $h\left(0, \frac{1}{2}, \cdot\right)$ and $h\left(0, \cdot, \frac{1}{2}\right)$. By the arguments above, these two functions capture all the discontinuity of range in any $h(\cdot, c, b)$, $h(a, \cdot, b)$ or $h(a, c, \cdot)$. Let $\{a_n\}$ be the set of discontinuity points of $h\left(0, \frac{1}{2}, \cdot\right)$ and $\{b_n\}$ be that of $h\left(0, \cdot, \frac{1}{2}\right)$. Then construct h^* as follows. Let

$$h^1(a, c, b) = \begin{cases} h\left(0, \frac{1}{2}, a_n^+\right) & \text{if } h(a, c, b) = h\left(0, \frac{1}{2}, a_n\right) \text{ for some } n, \\ h\left(0, b_n^-, \frac{1}{2}\right) & \text{if } h(a, c, b) = h\left(0, b_n, \frac{1}{2}\right) \text{ for some } n, \\ h(a, c, b) & \text{otherwise.} \end{cases}$$

Then let

$$v_1(a,c,b) = \sum_{k:\ h^1(a,c,b) > h^1\left(0,\frac{1}{2},a_k^-\right)} \left[h^1\left(0,\frac{1}{2},a_k^-\right) - h^1\left(0,\frac{1}{2},a_k^+\right)\right].$$

That is, the cumulative discontinuity of $h(0,\frac{1}{2},\cdot)$ within the range $h(0,\frac{1}{2},\frac{1}{2})$ to $h(a,c,b)$. Also let,

$$v_2(a,c,b) = \sum_{k: h^1(a,c,b) > h^1\left(0,b_k^+,\frac{1}{2}\right)} \left[h^1\left(0,b_k^+,\frac{1}{2}\right) - h^1\left(0,b_k^-,\frac{1}{2}\right)\right].$$

That is, the cumulative discontinuity of $h(0,\cdot,\frac{1}{2})$ within the range from $h(0,0,\frac{1}{2})$ to $h(a,c,b)$.

$$h^2(a,c,b) = \begin{cases} h^1(a,c,b) - v_2(a,c,b) & \text{if } h^1(a,c,b) \leq h^1\left(0,\frac{1}{2},\frac{1}{2}\right), \\ h^1(a,c,b) - v_1(a,c,b) - v_2(a,c,b) & \text{if } h^1(a,c,b) > h^1\left(0,\frac{1}{2},\frac{1}{2}\right). \end{cases}$$

Then rescale by

$$h^*(a,b,c) = \frac{h^2(a,b,c) - h^2(1,0,0)}{h^2(0,1,0) - h^2(1,0,0)}.$$

Basically h^* is produced from h by squeezing out the discontinuity points and then rescaling. By the argument above the resulting h^* is separately continuous on $\Delta^3\backslash\{(0,0,0)\}$. Extend h^* by setting

$$h^*(a,c,b) = h^*(\max\{a,0\}, c, \max\{b,0\})$$

to $G \equiv \{(a,c,b) \in \mathbb{R}^3 | 0 < c < 1,\ a+b+c < 1,\ b+c < 1, \text{ and } a+c < 1\}$. It is clear that h^* is monotone and separately continuous on G. Therefore, by Lemma C, h^* is continuous on G. It then suffices to prove that h^* is also continuous on $\{(a,c,b) \in \Delta^3 | a+b+c = 1\}$. This is true since $h^*(a,c,b) = h^*(\frac{a}{2},\frac{b}{2},\frac{c}{2})$ by Lemma 5.

By Lemma 5 and scale invariance we have that for all n and $m = 1, 2, \ldots, 2^n$.

$$h(a, c, b) = h\left(\frac{m}{2^n}a, \frac{m}{2^n}c, \frac{m}{2^n}b\right).$$

By continuity (v) follows. □

Let $f : [0,1]^2 \longrightarrow [0,1]$ be defined as follows. For $(x, y) \in (0,1]^2$ let

$$f(x, y) = h\left(\frac{\frac{1}{y} - 1}{2(\frac{1}{x} + \frac{1}{y} - 1)}, \frac{1}{2(\frac{1}{x} + \frac{1}{y} - 1)}, \frac{\frac{1}{x} - 1}{2(\frac{1}{x} + \frac{1}{y} - 1)}\right).$$

Let $f(0,0) = 0$, $f(x, 0) = \lim_{y \to 0^+} f(x, y)$ and $f(0, y) = \lim_{x \to 0^+} f(x, y)$. Clearly f is continuous on $(0,1]^2$ since h is continuous at (a, c, b) such that $0 < c < 1$.

Lemma 2.13. *For $a + c > 0$ and $b + c > 0$, $h(a, c, b) = f(\frac{c}{b+c}, \frac{c}{a+c})$.*

Proof. If $c = 0$, then $a, b > 0$. It follows that

$$f\left(\frac{c}{b+c}, \frac{c}{a+c}\right) = 0 = h(a, c, b).$$

If $0 < c < 1$ then

$$f\left(\frac{c}{b+c}, \frac{c}{a+c}\right) = h\left(\frac{a}{2(a+c+b)}, \frac{c}{2(a+c+b)}, \frac{b}{2(a+c+b)}\right) = h(a, c, b)$$

by Lemma 2.12. If $c = 1$ and therefore $a = b = 0$, then $f(1,1) = h(0,1,0) = h(0, c, 0)$. □

Let $\mu(A|B)$ be given by Bayes' rule if $\mu(B) > 0$. For expositional purpose, let $\mu(A|B) = 0$ if $\mu(B) = 0$. Therefore, it is easy to see that $S(A, B) = f(\mu(A|B), \mu(B|A))$ represents $\succsim$. The fact that f is nondecreasing is given by the properties of h. It suffices to prove that f is strictly increasing in the first argument, or equivalently, that $h(a,c,\cdot)$ is strictly decreasing for $c>0$. Due to additivity $h(a,\cdot,c)$ is strictly increasing for $c>0$. Then monotonicity finishes the argument.

The "if" part of Theorem 2.1 is standard and therefore omitted.

2.5.2 Proof of Theorem 2.3

We only have to prove that f has the Cobb-Douglas form on $(0,1]^2$ since $f(0,0)=0$ and it is simply impossible to have $\mu(A|B)>0$ and $\mu(B|A)=0$ for any A,B. Therefore, in the proof we will assume that the arguments of f are all within $(0,1]$. Lemma 2.14 is a direct implication of the robustness condition.

Lemma 2.14. *$f(p_1,p_2)\geq f(q_1,q_2)$ implies $f(rp_1,p_2)\geq f(rq_1,q_2)$ and $f(p_1,rp_2)\geq f(q_1,rq_2)$.*

Proof. First we prove that for any $(p_1,p_2),(q_1,q_2)$ there are A,A',B such that

$$\mu(A|B)=p_1,\quad \mu(B|A)=p_2,\quad \mu(A'|B)=q_1 \text{ and } \mu(B|A')=q_2.$$

It is equivalent to picking

$$\frac{c}{b+c}=p_1,\quad \frac{c}{a+c}=p_2,\quad \frac{c'}{b'+c'}=q_1,\quad \frac{c'}{a'+c'}=q_2 \text{ and } b'+c'=b+c.$$

Let

$$\begin{aligned} a&=\left(\frac{1}{p_2}-1\right)p_1x,\quad b=(1-p_1)x,\quad c=p_1x,\\ a'&=\left(\frac{1}{q_2}-1\right)q_1x,\quad b'=(1-q_1)x,\quad c'=q_1x,\end{aligned}$$

and pick a positive x small enough such that both $a+b+c$ and $a'+b'+c'$ are not larger than 1. Now we prove for any such A,A',B there is $C\perp\sigma(A,A',B)$ such that $\mu(C)=r$. Let $\{A_1,...,A_n\}$ be the finest partition of Ω which is contained in

$\sigma(A, A', B)$. Pick $C_k \subset A_k$ such that $\mu(C_k) = r\mu(A_k)$. Let $C = \bigcup_{k=1}^{n} C_k$ and we are done. Then it is clear that robustness implies the lemma. □

With the above lemma, the following claim is true.

Lemma 2.15. *For $a \in (0,1)$, $f(p_1, p_2) > f(q_1, q_2)$ implies $f(p_1^a, p_2^a) > f(q_1^a, q_2^a)$.*

Proof. By Lemma 14, $f(p_1, p_2) > f(q_1, q_2)$ implies that

$$f\left(p_1^2, p_2^2\right) \geq f(p_1q_1, p_2q_2) \geq f\left(q_1^2, q_2^2\right).$$

In fact for any n,

$$f(p_1^n, p_2^n) \geq f\left(p_1^{n-1}q_1, p_2^{n-1}q_2\right) \geq f\left(p_1^{n-2}q_1^2, p_2^{n-2}q_2^2\right) \geq \cdots \geq f\left(p_1q_1^{n-1}, p_2q_2^{n-1}\right) \geq f(q_1^n, q_2^n).$$

For any m, it must be the case that

$$f\left(p_1^{\frac{1}{m}}, p_2^{\frac{1}{m}}\right) > f\left(q_1^{\frac{1}{m}}, q_2^{\frac{1}{m}}\right)$$

since if otherwise inductively we would have that $f(p_1, p_2) \leq f(q_1, q_2)$, which is a contradiction. Since f is monotone and continuous on $(0,1]^2$, there are $r > s$ such that

$$f\left(p_1^{\frac{1}{m}}, p_2^{\frac{1}{m}}\right) > f(r,r) > f(s,s) > f\left(q_1^{\frac{1}{m}}, q_2^{\frac{1}{m}}\right)$$

Combining the two claims, for any m, n with $n > 0$

$$f(p_1, p_2) > f(q_1, q_2) \Longrightarrow f\left(p_1^{\frac{n}{m}}, p_2^{\frac{n}{m}}\right) \geq f(r^n, r^n) > f(s^n, s^n) \geq f\left(q_1^{\frac{n}{m}}, q_2^{\frac{n}{m}}\right).$$

Pick $t > u$ such that

$$f(p_1, p_2) > f(t,t) > f(u,u) > f(q_1, q_2).$$

It follows that

$$f\left(p_1^{\frac{n}{m}}, p_2^{\frac{n}{m}}\right) > f\left(t^{\frac{n}{m}}, t^{\frac{n}{m}}\right) > f\left(u^{\frac{n}{m}}, u^{\frac{n}{m}}\right) > f\left(q_1^{\frac{n}{m}}, q_2^{\frac{n}{m}}\right).$$

Let $\frac{n}{m} \to \alpha \in (0,1)$ we have that

$$f\left(p_1^{\alpha}, p_2^{\alpha}\right) \geq f\left(t^{\alpha}, t^{\alpha}\right) > f\left(u^{\alpha}, u^{\alpha}\right) \geq f\left(q_1^{\alpha}, q_2^{\alpha}\right)$$

where the central inequality is strict due to the fact that f is strictly increasing in its first argument. $\square$

With Theorem 3.2, it suffices to prove the following lemma.

Lemma 2.16. *$f(p_1, p_2) > f(q_1, q_2)$ and $a \in (0,1)$ implies $f\left(p_1^a r_1^{1-a}, p_2^a r_2^{1-a}\right) > f\left(q_1^a r_1^{1-a}, q_2^a r_2^{1-a}\right)$.*

Proof. By the previous lemma

$$f\left(p_1, p_2\right) > f\left(q_1, q_2\right) \Longrightarrow f\left(p_1^a, p_2^a\right) > f\left(q_1^a, q_2^a\right)$$

Pick $r > s$ such that

$$f\left(p_1^a, p_2^a\right) > f(r, r) > f(s, s) > f\left(q_1^a, q_2^a\right).$$

Then by Lemma 14

$$\begin{aligned} & f\left(p_1^a r_1^{1-a}, p_2^a\right) \geq f(r r_1^{1-a}, r) > f(s r_1^{1-a}, s) \geq f\left(q_1^a r_1^{1-a}, q_2^a\right) \\ \Longrightarrow \quad & f\left(p_1^a r_1^{1-a}, p_2^a r_2^{1-a}\right) \geq f(r r_1^{1-a}, r r_2^{1-a}) > f(s r_1^{1-a}, s r_2^{1-a}) \geq f\left(q_1^a r_1^{1-a}, q_2^a r_2^{1-a}\right) \end{aligned}$$

where the central inequality is strict due to the fact that f is strictly increasing in its first argument. $\square$

The rest is established by Theorem 3.2. Clearly $\alpha > 0$ since f is strictly increasing in its first argument.

Chapter 3

Anscombe and Aumann (1963) with General Mixture Spaces

3.1 Model

In this chapter we present the tools that we developed through the course of writing the second chapter. In particular, we extend the Anscombe-Aumman theorem of subjective probability to allow for general mixture spaces rather than lotteries. We then apply our theorem to characterize quasi-linear means with a simple condition that resembles the classic independence axiom. We then show that within the framework introduced in the second chapter, in addition to our Cobb-Douglas similarity index, this condition also enables us to recover Tversky's similarity index—a weighted harmonic mean of $\Pr(A|B)$ and $\Pr(B|A)$.

Let X be an arbitrary set and I be a mixture operation defined on X; that is, I is a function that maps $[0, 1] \times X^2$ to X such that for all $a, b \in [0, 1]$ and $x, y \in X$

(i) $I_1(x, y) = x$,

(ii) $I_a(x, y) = I_{1-a}(y, x)$, and

(iii) $I_a(I_b(x, y), y) = I_{ab}(x, y)$.

Our X and I replace respectively the space of lotteries and the weighted arithmetic average in Anscombe and Aumann (1963). Let $S = \{1, 2, \ldots, n\}$ be a finite set of pay-off relevant states. An act is a function $h : S \longrightarrow X$; that is, it specifies an element in X for each state in S. We will use H to denote the set of all acts. Abusing the notation a little bit, write $I_a(f, g) = (I_a(f_1, g_1), ..., I_a(f_n, g_n))$; that is, we are mixing acts in a state-by-state manner.

In Anscombe and Aumann (1963), an act specifies a simple lottery over prizes in each state. By contrast, our acts can be considered as Savage acts: an act specifies directly a prize for each state. For $x, y \in X$, we interpret $I_a(x, y)$ as the certainty equivalent to objectively randomizing between constant acts x and y with probability a and $1 - a$.

Let $\succsim$ be a binary relation defined on H. Axiom 3.1-3.4 are direct translations of the classic Anscombe-Aumann axioms while Axiom 3.5 is Gilboa and Schmeidler (1989)'s version of monotonicity. Henceforth, we identify constant acts by elements in X.

Axiom 3.1. *$\succsim$ is a preference relation.*

Axiom 3.2. *$f \succ g$ and $a \in (0, 1)$ implies $I_a(f, h) \succ I_a(g, h)$.*

Axiom 3.3. *$f \succ g \succ h$ implies that there exist $a, b \in (0, 1)$ such that $I_a(f, h) \succ g \succ I_b(f, h)$.*

Axiom 3.4. *There exist f, g such that $f \succ g$.*

Axiom 3.5. *$f_j \succsim g_j$ for all $j \in S$ implies $f \succsim g$.*

We say that a function U on H is linear if $U(I_a(f, g)) = aU(f) + (1 - a)U(g)$.

Theorem 3.1. *$\succsim$ satisfies Axiom 3.1-3.5 if and only if there exist a non-constant linear function U on X and a probability measure μ on S such that $W(f) = \sum_i U(f_i)\mu(i)$ represents $\succsim$. This U is unique up to a positive affine transformation and μ is unique.*

To see why the theorem is true, first consider only constant acts. By Herstein and Milnor (1953)'s mixture space theorem there is U such that $U(I_a(x,y)) = aU(x) + (1-a)U(y)$. Let $U(f) = (U(f_1), \ldots, U(f_n))$. Then, consider all acts in H, also by the mixture space theorem, there is W such that $W(I_a(f,g)) = aW(f) + (1-a)W(g)$.

We only have to proof the next lemma. The rest is standard, implied by mixture space uniqueness.

Lemma. *W is linear if and only if there exists a collection of linear functions U_j, for $j \in S$, such that $W(f) = \sum_j U_j(f_j)$.*

Proof. **Step 1:** Let $a_j \in [0,1]$ and $\sum_j a_j = 1$. If $U(f) = \sum_j a_j U(f^j)$ then $W(f) = \sum_j a_j W(f^j)$.

First of all, suppose $\#\{a_j > 0\} = 2$. We know that $U(I_a(f^1, f^2)) = aU(f^1) + (1-a)U(f^2)$. Since $U(f) = U(I_a(f^1, f^2))$, by Axiom 3.5, it must be the case that $W(f) = W(I_a(f^1, f^2))$. By linearity of W we prove the claim for $\#\{a_j > 0\} = 2$. An inductive argument finishes this step.

Let $x_0 \in X$ and let h_x^j be the act that yields x in state j and x_0 in every other state. Let x_0 be the constant act.

Step 2: $U(I_{\frac{1}{n}}(f, x^0)) = \frac{1}{n}U(f) + \frac{n-1}{n}U(x^0) = \sum_j \frac{1}{n}U(h_{f_j}^j)$.

Define $U_j(x) = W(h_x^j) - \frac{n-1}{n}W(x_0)$ for all $x \in X$, we have

$$\begin{aligned}
\sum_j U_j(f_j) &= \sum_j W(h_{f_j}^j) - (n-1)W(x_0) \\
&= n\sum_j \frac{1}{n}W(h_{f_j}^j) - (n-1)W(x_0) \\
&= nW(I_{\frac{1}{n}}(f, x_0)) - (n-1)W(x_0) \\
&= n\left(\frac{1}{n}W(f) + \frac{n-1}{n}W(x_0)\right) - (n-1)W(x_0) \\
&= W(f).
\end{aligned}$$

Then we show that U_j is linear.

$$\begin{aligned} U_j(I_a(x,y)) &= W(h^j_{I_a(x,y)}) - \frac{n-1}{n} W(x_0) \\ &= W(I_a(h^j_x, h^j_y)) - \frac{n-1}{n} W(x_0) \\ &= aW(h^j_x) + (1-a)W(h^j_y) - \frac{n-1}{n} W(x_0) \\ &= aU_j(x) + (1-a)U_j(y) \end{aligned}$$

for all $j \in S$. □

Theorem 3.1 essentially states that a unique subjective probability can be revealed from Savage acts on a finite state space if we allow for pairwise objective randomization. By contrast, Anscombe and Aumann (1963) need all simple lotteries defined on the prize space as primitives. Gul (1992) also extracts subjective probability from Savage acts on a finite state space but his independence axiom requires no objective randomization.

3.2 Method for Axiomatizing Quasi-linear Means

Building on Theorem 3.2, we summarize the method used in Chapter 2 for axiomatizing means in this section. Let X be a convex non-singleton subset of $\mathbb{R}$. Let $G : X^n \longrightarrow X$ be a nondecreasing and continuous function. In addition, we assume that G is diagonally-increasing; that is, $x > y$ implies $G(x, x, \ldots, x) > G(y, y, \ldots, y)$. We would like to find condition(s) which ensures that G is ordinally equivalent to a f-mean, defined as follows.

Definition. *Let $f : X \to \mathbb{R}$ be strictly increasing and continuous. $F : X^n \to X$ is a **f-mean** if*

$$F(x_1, x_2, \ldots, x_n) = f^{-1}\left(\sum_{i=1}^{n} \alpha_i f(x_i)\right).$$

for $\alpha_i \in [0,1]$ *such that* $\sum_{i=1}^{n} \alpha_i = 1$.

This class of means is called quasi-linear means. Hong (1983) traces this notion to Kolmogorov (1930) and de Finetti (1931), and provides an axiomatic foundation for a more general, infinite-state version of our definition of f-mean. In contrast, we focus on quasi-linear means of a finite vector of numbers; we appeal to Anscombe and Aumann (1963) and the mixture space theorem in order to provide a simple condition that ensures the form of a quasi-linear mean.

Definition. *Let* $f : X \to \mathbb{R}$ *be strictly increasing and continuous.* $I : [0,1] \times X^2 \to X$ *is a* ***f-mixture operation*** *if*

$$I_a(x,y) = f^{-1}\left(af(x) + (1-a)f(y)\right).$$

Abusing the notation a little bit, write

$$I_a(\boldsymbol{x},\boldsymbol{y}) = \left(I_a(x_1,y_1), I_a(x_2,y_2), \ldots, I_a(x_n,y_n)\right).$$

Theorem 3.2. *Let* $G : X^n \to X$ *be nondecreasing, diagonally-increasing and continuous. Then*

$$G(\boldsymbol{x}) \geq G(\boldsymbol{y}) \iff F(\boldsymbol{x}) \geq F(\boldsymbol{y})$$

where F *is a* f*-mean, if and only if*

$$G(\boldsymbol{x}) > G(\boldsymbol{y}) \text{ implies } G(I_a(\boldsymbol{x},\boldsymbol{z})) > G(I_a(\boldsymbol{y},\boldsymbol{z})) \text{ for } a \in (0,1).$$

The theorem states that all f-means boil down to a condition which resembles the independence in von Neumann and Morgenstern (1944). We provide the following examples to illustrate this condition.

Arithmetic mean. Let $f(x) = x$. Then the condition reduces to

$$
\begin{aligned}
& G(\boldsymbol{x}) > G(\boldsymbol{y}) \\
\Longrightarrow \quad & G(ax_1 + (1-a)z_1, \ldots, ax_n + (1-a)z_n) > G(ay_1 + (1-a)z_1, \ldots, ay_n + (1-a)z_n)
\end{aligned}
$$

for any $a \in (0,1)$ and $\boldsymbol{z}$.

Geometric mean. Let $f(x) = \ln x$. The condition reduces to

$$G(\boldsymbol{x}) > G(\boldsymbol{y}) \Longrightarrow G(x_1^a z_1^{1-a}, \ldots, x_n^a z_n^{1-a}) > G(y_1^a z_1^{1-a}, \ldots, y_n^a z_n^{1-a})$$

for any $a \in (0,1)$ and $\boldsymbol{z}$. This is the condition that we used in Chapter 2 to show that the aggregator f is a weighted geometric mean.

Harmonic mean. Let $f(x) = -1/x$. The condition reduces to

$$G(\boldsymbol{x}) > G(\boldsymbol{y}) \Longrightarrow G\left(\frac{1}{\frac{a}{x_1} + \frac{1-a}{z_1}}, \ldots, \frac{1}{\frac{a}{x_n} + \frac{1-a}{z_n}}\right) > G\left(\frac{1}{\frac{a}{y_1} + \frac{1-a}{z_1}}, \ldots, \frac{1}{\frac{a}{y_n} + \frac{1-a}{z_n}}\right)$$

for any $a \in (0,1)$ and $\boldsymbol{z}$. In the next section, we apply the condition above to provide a micro-foundation for the well-known Tversky index of similarity.

3.2.1 Proof of Theorem 3.2

We prove it using Theorem 3.1. First of all we need the following claim which says that I is indeed a mixture operation.

Claim 1. *If I is a f-mixture operation, it is a mixture operation. In particular, (i) $I_1(x,y) = x$; (ii) $I_a(x,y) = I_{1-a}(y,x)$; (iii) $I_a(I_b(x,y), y) = I_{ab}(x,y)$.*

Proof. (i) and (ii) are trivial. For (iii)

$$\begin{aligned}
I_a\left(I_b\left(x,y\right),y\right) &= I_a\left(f^{-1}\left(bf\left(x\right)+\left(1-b\right)f\left(y\right)\right),y\right) \\
&= f(af\left(f^{-1}\left(bf\left(x\right)+\left(1-b\right)f\left(y\right)\right)\right)+\left(1-a\right)f\left(y\right)) \\
&= f^{-1}\left(a\left(bf\left(x\right)+\left(1-b\right)f\left(y\right)\right)+\left(1-a\right)f\left(y\right)\right) \\
&= f^{-1}\left(abf\left(x\right)+\left(1-ab\right)f\left(y\right)\right)=I_{ab}\left(x,y\right).
\end{aligned}$$

Therefore I is a mixture operation. □

On the state space X^n, define

$$\boldsymbol{x}\succsim_G\boldsymbol{y}\iff G\left(\boldsymbol{x}\right)\geq G\left(\boldsymbol{y}\right).$$

The next lemma shows that $\succsim$ satisfies Axiom 3.1-3.5.

Claim 2. *Let $\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}\in X^n$. If $G\left(\boldsymbol{x}\right)>G\left(\boldsymbol{y}\right)$ implies $G\left(I_a\left(\boldsymbol{x},\boldsymbol{z}\right)\right)>G\left(I_a\left(y,\boldsymbol{z}\right)\right)$ for $a\in\left(0,1\right)$, then*

(i) $\succsim_G$ is a preference relation;

(ii) $\boldsymbol{x}\succ_G\boldsymbol{y}$ and $a\in\left(0,1\right)$ implies $I_a\left(\boldsymbol{x},\boldsymbol{z}\right)\succ_G I_a\left(\boldsymbol{y},\boldsymbol{z}\right)$;

(iii) $\boldsymbol{x}\succ_G\boldsymbol{y}\succ_G\boldsymbol{z}$ implies there exists $a,b\in\left(0,1\right)$ such that $I_a\left(\boldsymbol{x},\boldsymbol{z}\right)\succ_G\boldsymbol{y}\succ_G I_b\left(\boldsymbol{x},\boldsymbol{z}\right)$;

(iv) there are $\boldsymbol{x},\boldsymbol{y}\in X^n$ such that $\boldsymbol{x}\succ_G\boldsymbol{y}$;

(v) $\left(x_i,x_i,\ldots,x_i\right)\succsim_G\left(y_i,y_i,\ldots,y_i\right)$ for $i=1,2,\ldots,n$ implies $\boldsymbol{x}\succsim_G\boldsymbol{y}$.

Proof. (i),(ii),(iv),(v) are trivial. (iii) is implied by the continuity of f and G. □

By Theorem 3.1, there exists a nonconstant function U and unique $\alpha_i\in\left(0,1\right)$ with $\sum_{i=1}^n\alpha_i=1$ such that the function defined by

$$W\left(\boldsymbol{x}\right)=\sum_{i=1}^n\alpha_iU\left(x_i\right)$$

represents $\succsim_G$. Moreover, the utility index U is linear; that is, $U(I_a(x,y)) = aU(x) + (1-a)U(y)$, and is unique up to a positive affine tranformation.

Claim 3. *$U(x) = cf(x) + d$ for $c > 0$.*

Proof. First we prove that such U is linear.

$$\begin{aligned} U(I_a(x,y)) &= U\left(f^{-1}(af(x) + (1-a)f(y))\right) \\ &= cf\left(f^{-1}(af(x) + (1-a)f(y))\right) + d \\ &= aU(x) + (1-a)U(y). \end{aligned}$$

Since f is strictly increasing, it represents $\succsim_G$ within $\{\boldsymbol{x} \in X^n | x_i = x_j \text{ for all } i,j\}$. Therefore by the uniqueness of U we have $U(x) = cf(x) + d$. □

The last step is to recover the form of a f-mean by performing a monotone transformation on $W(\boldsymbol{x})$. We have $W(\boldsymbol{x}) = c\sum_{i=1}^{n} \alpha_i f(x_i) + d$. Since f^{-1} is strictly increasing

$$F(\boldsymbol{x}) = f^{-1}\left(\frac{W(\boldsymbol{x}) - d}{c}\right) = f^{-1}\left(\sum_{i=1}^{n} \alpha_i f(x_i)\right)$$

also represents $\succsim_G$ and we proved our theorem.

3.2.2 Recovering the Tversky Index

Although the Tversky index is popular in the psychology literature, even Tversky himself has not provided an axiomatic foundation. In this section, we utilize Theorem 3.2 to achieve this goal within the framework introduced in the second chapter; that is, we take as given that our DM's similarity relation $\succsim$ has a (f,μ) representation as in Theorem 2.1. We will show that in addition to the Cobb-Douglas similarity index as in Theorem 2.2, Theorem 3.2 also enables us to recover Tversky's similarity index—a weighted harmonic mean of $P(A|B)$ and $P(B|A)$. To see that, consider an alternative definition of robustness.

Definition. $\succsim$ *is said to be **robust*** * *if for any nonnull* C *and mutually exclusive* $\widehat{A}, \widehat{B}, \widehat{C}$ *such that* $(A \cup B \cup C) \cap (\widehat{A} \cup \widehat{B} \cup \widehat{C}) = \emptyset$,

$$ACB \succ A'CB' \Longrightarrow (A \cup \widehat{A})(C \cup \widehat{C})(B \cup \widehat{B}) \succ (A' \cup \widehat{A})(C \cup \widehat{C})(B' \cup \widehat{B}).$$

The condition says that if two pairs of events share the same intersection, then adding the same disjoint states to the standard or the stimulus of both pairs will not change their relative similarity ranking.

Theorem 3.3. *Suppose* (f, μ) *is a similarity representation of nondegenerate similarity structure* $(\Omega, \mathcal{E}, \mathcal{N}, \succsim)$. *Then* $\succsim$ *is robust* if and only if there is a unique* $\alpha \in (0,1]$ *such that*

$$(A, B) \succsim (C, D) \iff \frac{\mu(A \cap B)}{\alpha\mu(B) + (1-\alpha)\mu(A)} \geq \frac{\mu(C \cap D)}{\alpha\mu(D) + (1-\alpha)\mu(C)}.$$

We only have to consider $(0,1]^2$ since $f(0,0) = 0$ and it is simply impossible to have $\mu(A|B) > 0$ and $\mu(B|A) = 0$ for any A, B. By Theorem 3.2, if suffices to prove the following lemma.

Lemma 3.1. $f(p_1, p_2) > f(q_1, q_2)$ *implies*

$$f\left(\frac{1}{\frac{\beta}{p_1} + \frac{1-\beta}{r_1}}, \frac{1}{\frac{\beta}{p_2} + \frac{1-\beta}{r_2}}\right) > f\left(\frac{1}{\frac{\beta}{q_1} + \frac{1-\beta}{r_1}}, \frac{1}{\frac{\beta}{q_2} + \frac{1-\beta}{r_2}}\right)$$

for all $\beta \in (0,1)$.

Proof. Pick any $\beta \in (0,1)$ and $a, b, c, a', b', \widehat{a}, \widehat{b}, \widehat{c}$ such that

$$p_1 = \frac{c}{a+c},\ p_2 = \frac{c}{b+c},\ q_1 = \frac{c}{a'+c},\ q_2 = \frac{c}{b'+c},\ r_1 = \frac{\widehat{c}}{\widehat{a}+\widehat{c}},\ r_2 = \frac{\widehat{c}}{\widehat{b}+\widehat{c}},\ \beta = \frac{c}{c+\widehat{c}}.$$

and $a+b+c+a'+b'+\widehat{a}+\widehat{b}+\widehat{c}\leq 1$. Then

$$\begin{aligned}\left(\frac{1}{\frac{\beta}{p_1}+\frac{1-\beta}{r_1}},\frac{1}{\frac{\beta}{p_2}+\frac{1-\beta}{r_2}}\right) &= \left(\frac{c+\widehat{c}}{a+\widehat{a}+c+\widehat{c}},\frac{c+\widehat{c}}{b+\widehat{b}+c+\widehat{c}}\right),\\ \left(\frac{1}{\frac{\beta}{q_1}+\frac{1-\beta}{r_1}},\frac{1}{\frac{\beta}{q_2}+\frac{1-\beta}{r_2}}\right) &= \left(\frac{c+\widehat{c}}{a'+\widehat{a}+c+\widehat{c}},\frac{c+\widehat{c}}{b'+\widehat{b}+c+\widehat{c}}\right).\end{aligned}$$

Then pick $A,B,C,A',B',\widehat{A},\widehat{B},\widehat{C}$ such that

$$\mu(A)=a,\ \mu(B)=b,\ \mu(C)=c,\ \mu(A')=a',\ \mu(B')=b',\ \mu(\widehat{A})=\widehat{a},\ \mu(\widehat{B})=\widehat{b},\ \mu(\widehat{C})=\widehat{c}.$$

Then robustness implies the result. □

Clearly in the representation $\alpha>0$ since f is strictly increasing in its first argument.

3.3 Conclusion

In this chapter we presented the tools that we developed through the course of writing the second chapter. In particular, we extended the Anscombe-Aumman theorem of subjective probability to allow for general mixture spaces rather than lotteries. We then applied our theorem to characterize quasi-linear means with a simple condition that resembles the classic independence axiom. We then showed that within the framework introduced in the second chapter, in addition to the Cobb-Douglas similarity index, this condition also enables us to recover Tversky's similarity index—a weighted harmonic mean of $P(A|B)$ and $P(B|A)$.

Bibliography

Aczel, J. (1966): *Lectures on functional equations and their applications*, New York: Academic Press.

Anscombe, F. J. and R. J. Aumann (1963): "A Definition of Subjective Probability," *Ann. Math. Statist.*, 34, 199–205.

Aumann, R. J. (1976): "Agreeing to Disagree," *Ann. Statist.*, 4, 1236–1239.

Bar-Hillel, M. and E. Neter (1993): "How alike is it versus how likely is it: A disjunction fallacy in probability judgments." *Journal of Personality and Social Psychology*, 65, 1119.

Barberis, N., A. Shleifer, and R. Vishny (1998): "A Model of Investor Sentiment," *Journal of Financial Economics*, 49, 307–343, reprinted in Richard Thaler, ed., Advances in Behavioral Finance Vol. II, Princeton University Press and Russell Sage Foundation, 2005.

Blume, L., A. Brandenburger, and E. Dekel (1991): "Lexicographic Probabilities and Choice under Uncertainty," *Econometrica*, 59, 61–79.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer (2016): "Stereotypes," *The Quarterly Journal of Economics*, 131, 1753–1794.

Bregman, L. (1966): "The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming," *USSR Computational Mathematics and Mathematical Physics*, 7, 200–217.

Caticha, A. (2004): "Relative Entropy and Inductive Inference," in *Bayesian Inference and Maximum Entropy Methods In Science and Engineering*, ed. by G. Erickson and Y. Zhai, vol. 707, 75–96.

Caticha, A. and A. Giffin (2006): "Updating Probabilities," in *Bayesian Inference and Maximum Entropy Methods In Science and Engineering*, ed. by A. Mohammad-Djafari, vol. 872 of *American Institute of Physics Conference Series*, 31–42.

de Finetti, B. (1931): "Sul Concetto di Media," *Giornale dell' Instituo Italiano degli Attuari*, 2, 369–396.

Debreu, G. (1954): "Representation of a preference ordering by a numerical function," *Decision processes*, 3, 159–165.

DeGroot, M., S. Fienberg, and J. Kadane (1994): *Statistics and the Law*, New York: Wiley.

DeMarzo, P., D. Vayanos, and J. Zwiebel (2003): "Persuasion Bias, Social Infuence, and Unidimensional Opinions," *Quarterly Journal of Economics*, 118, 909–968.

Epstein, L. G. (2006): "An Axiomatic Model of Non-Bayesian Updating," *The Review of Economic Studies*, 73, 413–436.

Epstein, L. G., J. Noor, and A. Sandroni (2008): "Non-Bayesian updating: A theoretical framework," *Theoretical Economics*, 3, 193–229.

Fienberg, S. (1989): *The Evolving role of statistical assessments as evidence in the courts*, New York: Springer-Verlag.

Fischhoff, B., P. Slovic, and S. Lichtenstein (1978): "Fault trees: Sensitivity of estimated failure probabilities to problem representation." *Journal of Experimental Psychology: Human Perception and Performance*, 4, 330.

Geanakoplos, J. D. and H. M. Polemarchakis (1982): "We can't disagree forever," *Journal of Economic Theory*, 28, 192 – 200.

Gennaioli, N. and A. Shleifer (2010): "What Comes to Mind," *Quarterly Journal of Economics*, 125, 1399–1433.

Gilboa, I. and D. Schmeidler (1989): "Maxmin expected utility with non-unique prior," *Journal of Mathematical Economics*, 18, 141–153.

Gul, F. (1992): "Savage's Theorem with a Finite Number of States," *Journal of Economic Theory*, 57, 369–396.

Gul, F. and W. Pesendorfer (2001): "Temptation and Self-Control," *Econometrica*, 69, 1403–1435.

Herstein, I. N. and J. Milnor (1953): "An Axiomatic Approach to Measurable Utility," *Econometrica*, 21, 291–297.

Hong, C. S. (1983): "A Generalization of the Quasilinear Mean with Applications to the Measurement of Income Inequality and Decision Theory Resolving the Allais Paradox," *Econometrica*, 51, 1065–1092.

Jaynes, E. T. (1957): "Information Theory and Statistical Mechanics," *The Physical Review*, 106, 620–630.

Kahneman, D. and A. Tversky (1972): "Subjective probability: A judgment of representativeness," *Cognitive psychology*, 3, 430–454.

——— (1974): "Judgment under Uncertainty: Heuristics and Biases," *Science*, 185, 1124–1131.

——— (1982): "Evidential impact of base rates," in *Judgment under uncertainty: Heuristics and biases*, ed. by D. Kahneman, A. Tversky, and P. Slovic, Cambridge University Press.

——— (1983): "Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment." *Psychological review*, 90, 293–315.

Kamenica, E. and M. Gentzkow (2011): "Bayesian persuasion," *American Economic Review*, 101, 2590–2615.

Karbelkar, S. N. (1986): "On the axiomatic approach to the maximum entropy principle of inference," *Pramana-Journal of Physics*, 26, 301–310.

Kolmogorov, A. (1930): "Sur la Notion de la Moyenne," *Rendiconti Accademia dei Lincei*, 6, 388–391.

Krantz, D., D. Luce, P. Suppes, and A. Tversky (1999): "Foundations of measurement, Vol. I: Additive and polynomial representations," .

Kruse, R. L. and J. J. Deely (1969): "Joint Continuity of Monotonic Functions," *The American Mathematical Monthly*, 76, 74–76.

Kullback, S. (1959): *Information Theory and Statistics*, New York: Wiley.

Kullback, S. and R. A. Leibler (1951): "On Information and Sufficiency," *The Annals of Mathematical Statistics*, 22, 79–86.

Mullainathan, S. (2002a): "A Memory-Based Model of Bounded Rationality," *Quarterly Journal of Economics*, 117, 735–774.

——— (2002b): "A Memory-Based Model of Bounded Rationality," *The Quarterly Journal of Economics*, 117, 735–774.

——— (2002c): "Thinking through Categories," *Working Paper, MIT.*

Mullainathan, S., J. Schwartzstein, and A. Shleifer (2008): "Coarse Thinking and Persuasion," *The Quarterly Journal of Economics*, 123, 577–619.

Myerson, R. B. (1986): "Axiomatic Foundations of Bayesian Decision Theory," *Northwestern University, Center for Mathematical Studies in Economics and Management Science Discussion Paper*, 671.

Ortoleva, P. (2012): "Modeling the Change of Paradigm: Non-Bayesian Reactions to Unexpected News," *American Economic Review*, 102, 2410–2436.

Rabin, M. (2002): "Inference by Believers in the Law of Small Numbers," *The Quarterly Journal of Economics*, 117, 775–816.

Rabin, M. and J. L. Schrag (1999): "First Impressions Matter: A Model of Confirmatory Bias," *The Quarterly Journal of Economics*, 114, 37–82.

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

Schwartzstein, J. (2014): "Selective Attention and Learning," *Journal of the European Economic Association*, 12, 1423–1452.

Shannon, C. E. (1948): "A Mathematical Theory of Communication," *The Bell System Technical Journal*, 27, 379–423.

Shore, J. E. and R. W. Johnson (1980): "Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy," *IEEE Transactions of Information Theory*, IT-26, 26–27.

Skilling, J. (1988): "The Axioms of Maximum Entropy," *Maximum-Entropy and Bayesian Methods in Science and Engineering*, 1, 173–187.

Tribe, L. H. (1971): "Trial by Mathematics: Precision and Ritual in the Legal Process," *Harvard Law Review*, 84, 1329–1393.

Tversky, A. (1977): "Features of Similarity," *Psychological Review*, 84, 327–352.

Uffink, J. (1995): "Can the Maximum Entropy Principle be Explained as a Consistency Requirement?" *Studies in History and Philosophy of Modern Physics*, 26, 223–261.

Villegas, C. (1964): "On Qualitative Probability σ-Algebras," *Ann. Math. Statist.*, 35, 1787–1796.

von Neumann, J. and O. Morgenstern (1944): *Theory of Games and Economic Behavior*, Princeton: Princeton University Press.

Wilson, A. (2014): "Bounded Memory and Biases in Information Processing," *Econometrica*, 82, 2257–2294.

Zhao, C. (2016): "Representativeness and Similarity," *Working Paper, Princeton University.*