Pseudo-Bayesian Updating

Chen Zhao∗
Princeton University
December 12, 2016

Abstract

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 conservatism. 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. Conservatism requires the decision maker to adjust her beliefs no more than is necessary to accommodate the new 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.

∗Chen Zhao, chenzhao@princeton.edu; I am deeply indebted to Faruk Gul and Wolfgang Pesendorfer for their invaluable advice, guidance and encouragement. I am also very grateful to Roland Benabou, Dilip Abreu, Stephen Morris, Marek Pycia, Shaowei Ke for their insightful comments. All errors are mine.
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, quantitative information of the form “event $A$ has probability $r$” could also be translated into a collection of qualitative statements and be processed accordingly.

To motivate the model, consider the following example: Doctor 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 Doctor X’s beliefs, reject the hypothesis that procedure $A_1$ results in less morbidity than $A_2$. Doctor X is not a researcher but she does fully understand that the study provides contrary evidence to her long-held position.

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: Doctor 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, Doctor X should disregard morbidity when choosing between the procedures.

The primitive of our model is a nonatomic subjective probability $\mu$ 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 updating rule associates a posterior $\mu^\alpha$, with every $\mu$ and each qualitative statement $\alpha$. 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 updating 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 updating rule to characterize the following formula:

\[ \mu^\alpha = \arg \min_{\nu \ll \mu} d(\mu||\nu) \]

s.t. \( \nu(A_1) \geq \nu(A_2) \)

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

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

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

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

\[ \mu^\alpha = \mu. \] (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,

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

and for an arbitrary measurable event \( C \),

\[ \mu^\alpha(C) = \mu(C \cap R_\alpha) + \frac{\mu(C \cap G_\alpha)}{\mu(G_\alpha)} \mu^\alpha(G_\alpha) + \frac{\mu(C \cap B_\alpha)}{\mu(B_\alpha)} \mu^\alpha(B_\alpha). \] (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

\[ \mu^\alpha(C) = \frac{\mu(C \cap R_\alpha)}{\mu(R_\alpha)} \] (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 $G_\alpha$ is more likely than the event $B_\alpha$ 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 $R_\alpha$, as if she is conditioning on the event $R_\alpha$ according to Bayes’ rule. We will call the preceding equations (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 $\mu$ is over $S \times I$ where $I$ is the set of signals. Then, learning that signal $i \in I$ occurred amounts to learning $(\emptyset, S \times (I \setminus \{i\}))$; 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 $\alpha$ and $\beta$ are orthogonal. We define and discuss orthogonality in section 2. Second, *conservatism*, which requires the decision maker to adjust her beliefs no more than is necessary to accommodate the new information; that is, $\alpha = (A_1, A_2)$ and $\mu(A_2) \geq \mu(A_1)$ implies $\mu^\alpha(A_1) = \mu^\alpha(A_2)$.

In section 2, we prove that the pseudo-Bayes’ rule characterized by equations (1)-(4) is equivalent to the axioms and is the unique solution to the relative entropy minimization problem above. In section 3, we discuss the related literature. In section 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 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.

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

Since $\mu$ 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.

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 $\alpha, \beta$ and $\gamma$ to respectively denote news $(A_1, A_2)$, $(B_1, B_2)$ and $(C_1, C_2)$. Given prior $\mu$, 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 em-
brace a non-credible \( \alpha = (A_1, A_2) \): doing so would 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 \((S, \Sigma)\). The DM’s information set
before updating is an element of \( \mathbb{I}(\Delta(S, \Sigma)) = \{(\mu, \alpha) | \alpha \) is credible given \( \mu \} \).

**Definition.** A function \( r : \mathbb{I}(\Delta(S, \Sigma)) \to \Delta(S, \Sigma) \) is a **one-step updating 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 updating stages. A decision maker equipped with a one-
step updating rule can process any finite string of statements, \( \alpha_1, \alpha_2, \ldots, \alpha_n \) sequentially:
let \( \mu_0 = \mu \) and \( \mu_k = \mu^\alpha_{k-1} \) 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 \( \mu_n \). That is, each time the decision maker learns a new
qualitative statement, she applies it to her current beliefs according to the one-step updating
rule.

In section 5, we show that with our one-step updating rule, the DM could also update on
quantitative statements of the form “\( \Pr(A) = r \)” where \( r \) is a rational number within \([0, 1]\). In
particular, we show that such a quantitative statement could be interpreted as a sequence of
qualitative statements to which the one-step updating rule is applicable.

2.1 Orthogonality

Our main axiom, exchangeability, asserts that if two qualitative statements are orthogonal given prior $\mu$; that is, if they neither reinforce 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, $\alpha$ conveys exactly the same information to the DM as $(A_1 \setminus A_2, A_2 \setminus A_1)$. Thus, each $\alpha$ partitions the event space into three sets: the good news region $G_\alpha = A_1 \setminus A_2$, the bad news region $B_\alpha = A_2 \setminus A_1$ and the remainder $R_\alpha = S \setminus (G_\alpha \cup B_\alpha)$.

If $\mu(G_\alpha) > 0$, we call $\Pi_\alpha = \{G_\alpha, B_\alpha, R_\alpha\}$ the effective partition generated by $\alpha$. Also, we say that $D_\alpha = G_\alpha \cup B_\alpha$ is the domain of $\alpha$ since $D_\alpha$ contains all of the states that are influenced by $\alpha$. If $\mu(G_\alpha) = 0$, the DM interprets $G_\alpha$ as a synonym for impossibility and therefore views $\alpha$ as equivalent to $(\emptyset, B_\alpha)$. Therefore, to accommodate $\alpha$, the DM must lower the probability of $B_\alpha$ to zero and distribute the probability $\mu(B_\alpha)$ among the states in $S \setminus B_\alpha$. Hence, when $\mu(G_\alpha) = 0$, we call $\Pi_\alpha = \{G_\alpha \cup R_\alpha, B_\alpha\}$ the effective partition generated by $\alpha$. Since every state in $S$ is influenced by $(\emptyset, B_\alpha)$, the domain of $\alpha$ 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(a). Qualitative statement $\beta$ demands that the probability of $B_2$ be decreased and hence be brought closer to that of $B_1$; $\alpha$ however requires the probability of $A_1$ and therefore $B_2$ to be increased and thus increasing the difference between the probability of $B_2$ and that of $B_1$. Therefore, Figure 1(a) depicts a situation in which $\alpha$ and $\beta$ are in conflict and hence are not orthogonal.

In contrast, consider Figure 1(b): now $\beta$ compares two subsets of $R_\alpha$. Therefore, $\alpha$ 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 $\alpha$ and $\beta$ do not overlap, and thus, $\alpha$ and $\beta$ are orthogonal.
(a) $\alpha$ and $\beta$ are in conflict

(b) $\alpha$ and $\beta$ are orthogonal

Figure 1: Orthogonality

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

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

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

### 2.2 Pseudo-Bayesian Updating

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

**Axiom 1.** *(Exchangeability)* $\mu^{\alpha \beta} = \mu^{\beta \alpha}$ if $\alpha \perp \beta$ given $\mu$.

Axiom 1 states that if qualitative statements $\alpha$ and $\beta$ 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 updating rule to be non-exchangeable.
Axiom 2. (*Conservatism*) $\mu(A_2) \geq \mu(A_1) \implies \mu^\alpha(A_1) = \mu^\alpha(A_2)$.

Axiom 2 says that the DM, upon receiving information that contradicts her prior, makes minimal adjustments to her beliefs. The prior $\mu$ encapsulates all the credible information that the DM received in the past and, therefore, stands on equal footing with new information. Hence, the DM resolves the conflict between the two by not taking either side.

In our leading example, there is only one way for Doctor X to alleviate the tension between the previous literature and the recent paper 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, Doctor X will simply disregard morbidity due to the conflict in the literature.

Such behavior need not be inconsistent with Bayesianism. Consider the following situation: suppose the state space is $\{0, 1\}$. A Bayesian decision maker believes that $\Pr(0)$ is uniformly distributed within $[0, 1/2]$ then receives qualitative information ($\{0\}, \{1\}$). Bayes’ rule commands that she set $\Pr(0) = 1/2 = \Pr(1)$. In this example, ($\{0\}, \{1\}$) is unexpected in the sense that it is a null event given the decision maker’s prior: the event “$\{0\}$ and $\{1\}$ are equally likely” is in the support of her prior but has zero probability. In similar situations when unexpected information is received, Bayesian decision makers would always resort to equality to resolve the conflict.

One would be tempted to add an extra “strength” dimension to the signal; that is, if signal $\alpha$ is “strong” enough the DM could possibly lean more towards $A_1$. Incorporating such a notion of “strength” would require constructing a full joint prior where the strength of signals is encoded in the differences between conditional probabilities. Hence, Axiom 2 is the key simplifying assumption that releases the DM from having to contruct a joint distribution on pay-off relevant states and possible qualitative signals. Even with Axiom 2, information does implicitly differ in “strength”. Compared to $\alpha = (A_1, A_2)$, a “stronger piece” of information could be $(A_1', A_2)$ with $A_1' \subset A_1$, which could lead DM to update to $\mu^\alpha(A_1) > \mu^\alpha(A_2)$.

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

**Theorem 1.** A one-step updating rule satisfies Axiom 1 and 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 $\alpha$ given $\mu \in \Delta(\Sigma, S)$ such that $\mu(A_2) > \mu(A_1)$.

We call the formula in Theorem 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 $B_\alpha$ to zero and redistributes the prior probability $\mu(B_\alpha)$ evenly among the states in the remainder $R_\alpha$, as if she is conditioning on the event $R_\alpha$ according to Bayes’ rule. This case enables us to incorporate Bayesian updating as a special case of the pseudo-Bayesian rule: suppose the prior $\mu$ is over $S \times I$ where $I$ is a set of signals. Then, learning that signal $i$ occurred amounts to learning $(\emptyset, S \times (I \setminus \{i\}))$; 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 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, $\mu^\alpha$ is absolutely continuous with respect to $\mu$, 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 \( R_\alpha \) 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 \( R_\alpha \) according to Bayes’ rule.

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

**Dice Example.** 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 \( \alpha \) 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 \((1,2,...,6)\) are \((\frac{4}{11},\frac{2}{11},\frac{1}{11},\frac{1}{11},\frac{1}{11},\frac{1}{11})\). 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 2. That is, the density at states in \((0.2,0.6)\) will remain unchanged and mass will shift proportionally from the 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 as the unique solution to a constrained optimization problem. This way of describing the pseudo-Bayesian updating 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 \( \nu \) with respect to \( \mu \), 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

---

1The 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 \( S \).
unique up to a zero $\mu$-measure set.

**Definition.** For $\mu, \nu \in \Delta(S, \Sigma)$ such that $\nu \ll \mu$, the Kullback-Leibler divergence (henceforth KL-divergence) from $\mu$ to $\nu$ 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 $\nu$. 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 $\mu$ with respect to $\nu$.

**Theorem 2.** A one-step updating rule satisfies Axiom 1-2 if and only if

$$\mu^\alpha = \arg\min_{\nu \ll \mu} d(\mu||\nu)$$

\[(P)\]

$$s.t. \quad \nu(A_1) \geq \nu(A_2)$$

for any credible $\alpha$ 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 update $\mu^\alpha$. 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-Nichodym 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:

$$\min_{q_i \geq 0} - \sum_{i=1}^{3} p_i \ln \frac{q_i}{p_i}$$

s.t.  
$$q_1 \geq q_2$$

$$q_i = 0 \text{ if } p_i = 0$$

$$\sum_{i=1}^{3} q_i = 1$$

where $p$'s are the prior probabilities of $G_\alpha$, $B_\alpha$ and $R_\alpha$ 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 $\mu$, while in information theory the method of maximum relative entropy calls for the objective to be $d(\nu || \mu)$ given prior $\mu$. 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.

### 2.3 Proof of Theorem 1

In this subsection, we describe the key steps of the proof of Theorem 1. A formal proof is provided in the appendix. Theorem 1 can be broken into two parts: the first says that upon learning $\alpha$ the DM moves probability mass in proportion to the prior between elements of $\Pi_\alpha$; the second says that learning $\alpha$ does not affect the probability of $D_\alpha$.

To see why the first part is true, first note that if $\alpha$ and $\beta$ are orthogonal given $\mu$ 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 $\Pi_\alpha$, the likelihood ranking cannot be affected by the arrival of $\alpha$. Since $\mu$ is nonatomic, this, in turn, implies that the probability of sub-events in the same element of $\Pi_\alpha$ must be updated in proportion to the prior.

The preceding argument identifies $\mu^\alpha$ 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 $S \setminus A_2$ according to Bayes rule.

Now, let $\beta = (\emptyset, B_2)$ and assume that $A_1$ and $A_2$ are both nonnull events contained in $S \setminus B_2$. Since $\alpha$ and $\beta$ are orthogonal given $\mu$ it follows that the order of updating 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 $R_\alpha$ 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

$$\frac{\mu^\alpha(D_\alpha)}{1 - \frac{1 - \mu^\alpha(D_\alpha)}{1 - \mu(D_\alpha)} \mu(B_2)} = \mu^{\beta\alpha}(D_\alpha). \quad (5)$$

Next, we establish a connection between $\mu^\alpha(D_\alpha)$ and $\mu^{\beta\alpha}(D_\alpha)$ in order to turn (5) into a functional equation. Consider two priors $\mu_1$ and $\mu_2$ 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 $\alpha$ and $\beta$ 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 $R_\alpha$. 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 $\mu$ behaves within $D_\alpha$. 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 (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

$$\frac{g(x)}{1 - \frac{1 - g(x)}{1 - b} b} = g \left( \frac{x}{1 - b} \right). \quad (6)$$
which holds for $0 < x < 1$ and $0 < b < 1 - x$ since we could vary $B_2$ within $R_\alpha$ continuously thanks to the convex-rangedness of $\mu$. The solution to (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 $D_\alpha$ increases if the DM learns $\alpha$; if $a < 1/2$, $g(x) < x$ so $D_\alpha$ shrinks; if $a = 1/2$, the probability of $D_\alpha$ is unchanged.

The interaction between Axiom 1 and 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 $D_\alpha$ expands, $R_\alpha$ 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 $\alpha$ increases $D_\alpha$ and therefore $C_1$, the DM will believe that $\mu^{\alpha\alpha}(C_1) > \mu^{\alpha\alpha}(C_2)$, which contradicts exchangeability since $\alpha \perp \gamma$ given $\mu$. A similar argument holds if $a < 1/2$. Hence $a = 1/2$ and we have established the “only if” part of Theorem 1. We provide the “if” part in the appendix.

3 Related Literature

Existing models of non-Bayesian updating 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 updating to situations in which Bayes rule does not apply. In particular, given a state space $(S, \Sigma)$, Bayes’ rule only applies to information of the form “$A \in \Sigma$ has occurred.” While nesting such information as $(\emptyset, S\setminus A)$, our qualitative setting also permits statements that are not events in the state space.\footnote{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.}

For behavioral models of non-Bayesian updating, see, for example, Barberis, Shleifer, and
Vishny (1998); Rabin and Schrag (1999); Mullainathan (2002); 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 updating 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 updating 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 2; that is, the statistician minimizes $d(\mu|\cdot)$ subject to newly-received constraints given prior $\mu$. 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\(^3\) 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\(^4\) survive scrutiny as objective functions.

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 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.

\(^3\)The concept of relative entropy originates from Kullback-Leibler divergence. When first introduced in Kullback and Leibler (1951), the KL-divergence between probability measures \(\mu\) and \(\nu\) 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.

\(^4\)The form of \(\eta\)-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 \(\eta\)-entropies, only \(\eta = -1\) is consistent with our axioms.
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, $\alpha$ and then $\beta$, her ultimate beliefs must be consistent with $\beta$ but need not be consistent with $\alpha$. More specifically, suppose there are three states $\{s_1, s_2, s_3\}$ and the prior is $(0.2, 0.3, 0.5)$. 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 $\alpha$ is said to be degenerate given $\mu$ 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 $\mu$ for the remainder of this subsection. Since our updating formula creates extra null events only if a degenerate statement is received, we will drop the qualification “given $\mu$” for simplicity of exposition.

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

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

**Proposition 1.** If $[\alpha, \beta]$ induces recency bias on $\mu$, then $[\beta, \alpha]$ also induces recency bias on $\mu^\alpha$.

---

5 Again, the reader should view the three states as a partition of state space $S$ with a nonatomic prior.
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 $\beta$ for a second time, she will forget about $\alpha$, so on and so forth. As she learns the statements repeatedly, however, the DM will be able to accommodate both $\alpha$ and $\beta$ in the limit.

**Proposition 2.** If $[\alpha, \beta]$ induces recency bias on $\mu$, then $\mu^{\alpha\beta\alpha}_n$ and $\mu^{\alpha\beta\alpha\alpha}_n$ converge in total variation to $\mu^*$ such that $\mu^*(A_1) = \mu^*(A_2)$ and $\mu^*(B_1) = \mu^*(B_2)$.

Proposition 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 decision 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 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 $\nu$ 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 $\nu$ which agrees with $\mu$ 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 \( \alpha_i \) 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 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 \( \alpha_i \), 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 3 depends on the sequence \( \{\pi_n\} \) in general.\(^6\) 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 \( S \).

---

\(^6\)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 \( \mu^{\alpha_2\alpha_1\cdots} = (0.35, 0.3, 0.35) \).
Definition. We say that a compatible collection \( \{\alpha_i\}_{i=1}^n \) is **representative** if

\[
\begin{align*}
(i) & \quad \nu, \nu' \text{ are solutions to } \{\alpha_i\}_{i=1}^n \\
(ii) & \quad \nu\left(\bigcup_{i=1}^n D_{\alpha_i}\right) = \nu'(\bigcup_{i=1}^n D_{\alpha_i}) \Rightarrow \nu(C) = \nu'(C) \text{ for all } C \in \bigvee_i \alpha_i.
\end{align*}
\]

That is, a collection of statements \( \{\alpha_i\}_{i=1}^n \) is representative if it uniquely identifies a probability measure defined on the discrete state space \( \bigvee_{i=1}^n \alpha_i \). For an example let \( A, B, C, D \) be mutually exclusive and nonnull according to \( \mu \). 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 \( \nu \) it has to be the case that \( \nu(A) = \nu(B) = \nu(C) = \nu(D) = p/4 \). Hence the collection above is representative.

By Theorem 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 \( \mu^* \), 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 whole state space, i.e. \( \bigcup_i D_{\alpha_i} = S \), the DM would eventually have correct beliefs on any event \( C \in \bigvee_{i=1}^n \alpha_i \). 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, the DM is in fact able to incorporate quantitative statements in the form of \( \Pr(A) = r \) where \( r \) is a rational number within \([0, 1]\). In particular, any quantitative statement of such kind could be interpreted as a representative collection of qualitative statements. For example, \( \Pr(A) = 2/3 \) could be interpreted as “half of \( A \) is as likely as its complement” which boils down to the following collection of qualitative
information:
$$\{(A', A \backslash A'), (A \backslash A', S \backslash A), (S \backslash A, A')\}$$

where $$A' \subset A$$ with $$\mu(A') = \mu(A) / 2$$. If $$0 < \mu(A) < 1$$, the collection is representative since any probability measure which embraces the collection has $$\Pr(A') = \Pr(A \backslash A') = \Pr(S \backslash A) = 1 / 3$$. 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 $$\Pr(A) = \frac{m}{m+n}$$ 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 $$S \backslash A$$ under $$\mu$$. The interpretation $$\{\alpha_i\}_{i=1}^{m+n}$$ basically conveys the information “$$1/m$$ of $$A$$ is as likely as $$1/n$$ of $$S \backslash A$$.”

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.** Let $$0 < \mu(A) < 1$$ and $$\{\alpha_i\}_{i=1}^{m+n}$$ be an interpretation of quantitative information $$\Pr(A) = \frac{m}{m+n}$$ 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 $$\mu^*$$ such that
$$\mu^* = \arg \min_{\nu \ll \mu} d(\mu \| \nu)$$
$$s.t. \quad \nu(A) = \frac{m}{m+n}.$$

By Theorem 1 and 2, $$\mu^*$$ in the corollary above is also given by
$$\mu^*(C) = \frac{m}{m+n} \cdot \frac{\mu(C \cap A)}{\mu(A)} + \frac{n}{m+n} \cdot \frac{\mu(C \cap (S \backslash A))}{\mu(S \backslash A)}$$

for any $$C \in \Sigma$$.

Hence, the DM can process frequency data without making assumptions on the data-
generating process. Imagine the following situation: an experiment is repeated for a large number of times but the DM could only observe whether or not event $A$ happens in each experiment. The Law of Large Numbers requires her to set the probability of $A$ to be the proportion of experiments in which $A$ happens but the law is silent on any subevent of $A$ and $S \backslash A$. Without assumptions regarding the data-generating process, she simply has no idea how to revise her beliefs on these sub-events. In this situation, the above corollary weighs in and proposes that the DM should keep the likelihood ratios among the sub-events of $A$ and the likelihood ratios among the sub-events of $S \backslash A$.

### 4.2 Persuasion

In this subsection we apply our model to a persuasion problem. Let $X : S \rightarrow \mathbb{R}$ be a random variable. Suppose $\mu^*$ is the true probability measure on $(S, \Sigma)$ while $\mu$ is the DM’s beliefs. We assume that $\mu$ and $\mu^*$ are both nonatomic. The information sender, knowing the true distribution $\mu^*$ and also $\mu$, 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 $\alpha$ such that $\mu^*(A_1) \geq \mu^*(A_2)$ and $E_{\mu^*}(X) > E_\mu(X)$. Such an $\alpha$ 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 3.** Suppose $\text{var}_\mu(X) > 0$ 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 $X(s) = 1\{s \in A\}$ and $\mu(A) \geq \mu^*(A) = 1/2$. Also suppose that $\mu$ and $\mu^*$ have the same conditional probabilities on $A$ and on $S\setminus A$. Picking a persuasive $\alpha$ amounts to picking $B, C \subset A$ and $B', C' \subset S\setminus A$ such that the following inequalities hold:

$$
\frac{\mu(B) + \mu(B')}{\mu(A)} < \frac{\mu(C) + \mu(C')}{1 - \mu(A)},
$$

$$
\frac{\mu(B)}{\mu(B) + \mu(B')} > \frac{\mu(C)}{\mu(C) + \mu(C')}.
$$

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 ($\mu$ and $\mu^*$ 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 $\mu^*$ it might not be the case under $\mu$. 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.

### 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 \( \alpha_j \)-experts, who exogenously learns \( \alpha_j \) 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 \( \alpha_j \) is communicated.\(^7\) If \( N_j \cap S(i) \neq \emptyset \); that is if agent \( i \) is an \( \alpha_j \)-expert or has an \( \alpha_j \)-expert neighbor, she will simply ignore any \( \bar{\alpha}_j = (A_{j2}, A_{j1}) \). Then, each agent updates on the rest of qualitative statements in any finite sequence, repetitions permitted. Let \( \mu_i \) denote agent \( i \)'s prior belief and \( \mu^k_i \) 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.\(^8\)

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 \( \alpha_j \)-expert, then the consensus ranking between \( A_{j1} \) and \( A_{j2} \), if reached, must be the correct one: \( A_{j1} \geq A_{j2} \).

**Proposition 4.** \( \mu^k_i \) converges in total variation to some \( \mu^*_i \) 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 \( \mu^*_i \) 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 guaranteed that they can spread the truth to their own neighbors.

In fact, agents could disagree with each other in the limit even if the network is strongly connected.

\(^7\)Same results hold if \( \alpha_j \) and \( \bar{\alpha}_j = (A_{j2}, A_{j1}) \) are both communicated.

\(^8\)Otherwise, if, for example, \( \alpha_1 \) and \( \alpha_2 \) are degenerate given \( \mu_i \), then after agent \( i \) learns \( \alpha_1 \), statement \( \alpha_2 \) becomes non-credible.
connected. Suppose three decision makers form a line as in Figure 3. Agent 1 is an $\alpha$-expert and agent 2 is a $\beta$-expert. Agent 3, however, has expertise in neither $\alpha$ nor $\beta$. Suppose $[\beta, \alpha]$ induces recency bias on $\mu_2$. Then, although in the limit agent 2 learns both $\alpha$ and $\beta$, each time she communicates with agent 3, she will say $\beta$ and $\bar{\alpha} = (A_2, A_1)$. Therefore agent 3 can never learn $\alpha$.

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 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 $\mu_i$, consensus can be reached within a rich enough network.

**Proposition 5.** Suppose the network is strongly connected and $\#N_j > 0$ for each $j$. If $\alpha_1 \perp \alpha_2$ given $\mu_i$ 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 $\alpha_1$ are $\alpha_2$ 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 4 implies that consensus is also guaranteed when $N_j \cap S(i) \neq \emptyset$ for all $i, j$. 

![Figure 3: A network where consensus might not be reached.](image)
5 Pseudo-Bayesian versus Bayesian Updating

It is well-known that Bayes’ rule is a special case of the method of maximum relative entropy.\textsuperscript{9} 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 $\Sigma, \Omega$ be respectively $\sigma$-algebras on $S, I$. The DM’s prior $\mu$ is therefore a nonatomic probability measure defined on the product measure space $(S \times I, \Sigma \times \Omega)$.

Suppose 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 $\alpha = (\emptyset, S \times (I \setminus I'))$ 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

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

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

**Theorem 4.** If a one-step updating 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 $\alpha = (\emptyset, S \times (I \setminus I'))$.

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 $\mu$.

To process this qualitative statement, the decision maker needs a prior $\mu'$ 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 $\mu'$, 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

\textsuperscript{9}See Caticha and Giffin (2006).
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, Doctor 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 Doctor X can incorporate such payoff-relevant information into her beliefs in order to make more informed decisions.

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 conservatism axiom states that if they disagree with each other on $A$ versus $B$, both will concede and the consensus 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.
References

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.

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.

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.

DeMarzo, P., D. Vayanos, and J. Zwiebel (2003): “Persuasion Bias, Social Influence, 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.

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.

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

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

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.

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. (2002): “A Memory-Based Model of Bounded Rationality,” *Quarterly Journal of Economics*, 117, 735–774.

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.

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.

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.

Zhao, C. (2016): “Representativeness and Similarity,” Working Paper, Princeton University.
A Properties of the Kullback-Leibler Divergence

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

**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

$$\int S \max\{\ln(\frac{d\nu}{d\mu}), 0\} d\mu = \int S \{\frac{d\nu}{d\mu} \geq 1\} \ln(\frac{d\nu}{d\mu}) d\mu = \int S \{\frac{d\nu}{d\mu} \geq 1\} \ln(\frac{d\nu}{d\mu}) d\nu \leq \int S \frac{1}{e} d\nu = \frac{1}{e}.$$ 

In the second equality $\frac{d\nu}{d\mu} = (\frac{d\mu}{d\nu})^{-1}$ 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

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

Hence

$$-\int S \ln(\frac{d\nu}{d\mu}) d\mu = -\int S \ln(a\frac{d\nu_1}{d\mu} + (1-a)\frac{d\nu_2}{d\mu}) d\mu \leq -a\int S \ln(\frac{d\nu_1}{d\mu}) d\mu - (1-a) \int S \ln(\frac{d\nu_2}{d\mu}) d\mu$$

with equality if and only if

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

By definition of Radon-Nikodym derivatives, if $\frac{d\nu_1}{d\mu} = \frac{d\nu_2}{d\mu}$ $\mu$-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 $\mu = \nu$. 

33
Proof. By Jensen’s inequality
\[-\int_S \ln\left(\frac{d\nu}{d\mu}\right) d\mu \geq -\ln\left(\int_S \frac{d\nu}{d\mu} d\mu\right) = -\ln 1 = 0\]
with equality attained if and only if \(\frac{d\nu}{d\mu} = C\) \(\mu\)-almost everywhere. By definition of the Radon-Nikodym derivative it is clear that \(C = 1\). Hence for any \(A \in \Sigma\),
\[\nu(A) = \int_A \frac{d\nu}{d\mu} d\mu = \int_A d\mu = \mu(A).\]

\[\Box\]

## B Proof of Theorem 2

Proof. It suffices to prove that optimization \(P\) in Theorem 2 has the pseudo-Bayes’ rule as the unique solution. Recall optimization \(P\): First of all note that we are picking \(\nu \ll \mu\) so the Radon-Nikodym derivative is well-defined. The constraint optimization problem \(P\) is equivalent to the following \(P^*\).

\[
\begin{align*}
\min_{f : S \rightarrow [0, \infty) \text{ measurable}} & -\int_S \ln f \, d\mu \\
\text{s.t.} & \int_{A_1 \setminus A_2} f \, d\mu \geq \int_{A_2 \setminus A_1} f \, d\mu \\
& \int_S f \, d\mu = 1.
\end{align*}
\]

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

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

\[
\begin{align*}
\min_{f : C \rightarrow [0, \infty) \text{ measurable}} & -\int_C \ln f \, d\mu \\
\text{s.t.} & \int_C f \, d\mu = p.
\end{align*}
\]
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 $\mu$-almost everywhere. Then the constraint demands that $f = p/\mu(C)$ $\mu$-almost everywhere.

With Lemma 1, $P^*$ has a solution if and only if the following $P^{**}$ has a solution, for Lemma 1 basically demands that within $G_\alpha$, $B_\alpha$ or $R_\alpha$, $f$ is a constant. Let $\mu(G_\alpha) = p_1, \mu(B_\alpha) = p_2$ and $p_3 = 1 - p_1 - p_2$.

\[
\begin{align*}
\min_{q_1, q_2, q_3 \geq 0} & \quad -\sum_{i=1}^{3} p_i \ln \frac{q_i}{p_i} \\
\text{s.t.} & \quad q_1 \geq q_2 \\
& \quad q_i = 0 \text{ if } p_i = 0 \\
& \quad \sum_{i=1}^{3} q_i = 1.
\end{align*}
\]

(P**)

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
\[
\mathcal{L}(q, \lambda, \eta) = -\sum_{i=1}^{3} p_i \ln \frac{q_i}{p_i} + \lambda(q_1 - q_2) + \eta(1 - \sum_{i=1}^{3} q_i).
\]
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 and the above analysis, $P^*$ has a solution $f$ which is
unique up to $\mu$-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. \hfill \Box

C Proof of Theorem 1

Proof. We first prove the following lemma.

Lemma 2. If an one-step updating rule satisfies Axiom 1 and 2, then

\[
\mu(A) \geq \mu(B) \implies \mu^\alpha(A) \geq \mu^\alpha(B)
\]

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

Proof. First we prove that if the one-step updating rule satisfies Axiom 1 and 2, $\alpha \perp \beta$ given $\mu$ implies $\mu(B_1) \geq \mu(B_2) \implies \mu^\alpha(B_1) \geq \mu^\alpha(B_2)$. Suppose $\mu(B_1) \geq \mu(B_2)$ but $\mu^\alpha(B_2) > \mu^\alpha(B_1)$. By definition of the updating 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 the conservatism axiom 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 $\mu$ and $\mu^\alpha$ are well-defined probabilities. If $\mu(A) > 0$, by definition $\alpha \perp (A, B)$ with respect to $\mu$ 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 $\mu$. 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 \( C = B_\alpha = A_2 \setminus A_1 \), then we already have \( \mu(A_1) \geq \mu(A_2) \), by definition of a one-step updating 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 \( \alpha \) is non-credible.

Then, we take advantage of the convex-rangedness of \( \mu \) and further characterize the updating rule by the following lemma.

**Lemma 3.** If an one-step updating rule satisfies Axiom 1 and 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, 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 \( \mu \) 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 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 additivity. □

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 \( \beta \) 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 \( \mu \). Axiom 1 then requires that \( \mu^{\alpha \beta} = \mu^{\beta \alpha} \). Therefore

\[
\mu^{\alpha \beta}(D_\alpha) = \mu^{\beta \alpha}(D_\alpha).
\]
Note that Lemma 3 implies that updating on $\beta$ is equivalent to conditioning on $S\setminus B_2$ according to Bayes’ rule. Therefore the above display equation is equivalent to

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

(7)

where $\mu(\cdot|A)$ denotes the probability measure conditioning on $A \in \Sigma$ given $\mu$. 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 (7), we get

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

(8)

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

**Lemma 4.** If Axiom 1 and 2 are satisfied, for each $\alpha$, $\mu^\alpha(D_\alpha) = g(\mu(\cdot|D_\alpha), \mu(D_\alpha))$.

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

**Step 1:** If $\mu(\cdot|S\setminus A) = \mu_1(\cdot|S\setminus A)$ and $\mu(A) = \mu_1(A)$, then $\mu^\alpha(D_\alpha) = \mu^\gamma_1(D_\alpha)$.

The claim here basically states that $\mu^\alpha(D_\alpha)$ does not depend on how $\mu$ behaves within $A$. Since $\nu^\gamma = \nu(\cdot|S\setminus A)$ for any $\nu$ we have that $\mu^\gamma = \mu^\gamma_1$. Therefore we must have $\mu^\alpha = \mu^\gamma_1$. Note that $\alpha \perp \gamma$ given both $\mu$ and $\mu_1$. Therefore $\mu^\alpha = \mu^\gamma_1$. Hence by the same logic in the derivation of equation (2),

$$\frac{\mu^\alpha(D_\alpha)}{1 - \frac{1 - \mu^\alpha(D_\alpha)}{1 - \mu(D_\alpha) \mu(A)}} = \frac{\mu^\gamma_1(D_\alpha)}{1 - \frac{1 - \mu^\gamma_1(D_\alpha)}{1 - \mu_1(D_\alpha) \mu_1(A)}}$$

which reads

$$(1 - \mu(D_\alpha) - \mu(A))\mu^\alpha(D_\alpha) = (1 - \mu_1(D_\alpha) - \mu_1(A))\mu^\gamma_1(D_\alpha).$$

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 \( \mu_2 = \mu(S\setminus A)\mu(\cdot|S\setminus A) + \mu(A)\mu_1(\cdot|A) \). So \( \mu_2 \) agrees with \( \mu \) within \( S\setminus A \) and agrees with \( \mu_1 \) within \( A \). It is clear that by Step 1 \( \mu^\alpha(D_\alpha) = \mu_2^\alpha(D_\alpha) \). Moreover we also have that \( \mu_2 = \mu_1(S\setminus B)\mu_1(\cdot|S\setminus B) + \mu_1(B)\mu(\cdot|B) \); that is \( \mu_1 \) and \( \mu_2 \) 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 \( \mu_2 = \mu(S\setminus B)\mu(\cdot|S\setminus B) + \mu(B)\mu_1(\cdot|B) \), clearly \( \mu^\alpha(D_\alpha) = \mu_2^\alpha(D_\alpha) \) by Step 1 since \( \mu \) and \( \mu_2 \) 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 \( \mu_1(A' \cup B') = \mu_2(A' \cup B') = \mu(S\setminus D_\alpha)/2 \) 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 \( B \).

\[ \square \]

By Lemma 4, equation (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

\[
\frac{g(x)}{1 - \frac{1 - g(x)}{1 - b}} = g \left( \frac{x}{1 - b} \right)
\]

which holds for \( 0 < x < 1 \) and \( 0 < b < 1 - x \), since we could always vary \( B_2 \) within \( R_\alpha \) continuously thanks to the convex-rangedness of \( \mu \).
Lemma 5. The solutions to functional equation (9) which maps from $(0, 1)$ to $[0, 1]$ are

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

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

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

\[ a = g\left(\frac{1}{2}\right) = \frac{g(x)}{1 - \frac{1-g(x)}{1-x} (1-2x)} \]

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

\[ g(x) = \frac{g\left(\frac{1}{2}\right)}{1 - \frac{1-g\left(\frac{1}{2}\right)}{1-x} (1-\frac{1}{2x})} = \frac{a}{1 - (1-a)(2-\frac{1}{x})} = \frac{a}{2a - 1 + \frac{1-a}{x}}. \]

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

The functional form in (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 $\gamma$ becomes non-credible given $\mu^\alpha$, a contradiction to Axiom 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 updating on $\alpha$ next. By the same logic as above if $a > \frac{1}{2}$ we have $\mu^{\alpha\gamma}(C_2) > \mu^{\alpha\gamma}(C_1)$, which is a contradiction since $\alpha \perp \gamma$ with respect to $\mu$.  

40
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 \( \mu^\alpha(C_1) = \mu^\gamma(C_2) \) but \( \mu^\alpha(C_2) > \mu^\alpha(C_1) \), 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 3, 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 updating rule requires \( \mu^\alpha = \mu \). Secondly, if \( \mu(A_1) < \mu(A_2) \) and \( \mu(G_\alpha) > 0 \), the DM moves probability from \( B_\alpha \) to \( G_\alpha \) in proportion to the prior until these two sets have the same probability \( \mu(D_\alpha)/2 \), i.e.

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

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 \( B_\alpha \) to \( R_\alpha \), and hence \( \mu^\alpha = \mu(\cdot | R_\alpha) \). Hence we establish the “only if” part of Theorem 1.

D The If Part for Theorem 1

Proof. Axiom 2 is immediately implied by the pseudo-Bayes’ rule. Hence it suffices to prove that if \( \alpha \perp \beta \) given \( \mu \), the procedure implies \( \mu^\alpha \beta = \mu^\beta \alpha \). First we prove that \( \beta \) is credible given \( \mu^\alpha \). Clearly when \( \mu(G_\alpha) > 0 \) it is true since in this case \( \mu \ll \mu^\alpha \). When \( \mu(G_\alpha) = \mu(G_\beta) = 0 \), \( \alpha \) and \( \beta \) cannot be orthogonal unless at least one of \( B_\alpha \) and \( B_\beta \) is \( \emptyset \). In this cases \( \beta \) is always credible given \( \mu^\alpha \) if it is credible given \( \mu \). 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 \( \beta \) is always credible given \( \mu^\alpha \).

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 \( D_\alpha \) and \( D_\beta \). 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 \( \mu^{\alpha \beta}(D_\alpha) = \mu^{\alpha \beta}(D_\alpha) \) and \( \mu^{\beta \alpha}(D_\beta) = \mu^{\beta \alpha}(D_\beta) \). With Lemma 3 this 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 \( \mu \) 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 \( \mu^{\beta \alpha}(G_\alpha \cup R_\alpha) = \mu^{\beta \alpha}(G_\alpha \cup R_\alpha) = 1 \). It is easy to verify that \( \mu^{\beta \alpha}(D_\beta) = \mu^{\beta \alpha}(D_\beta) \) and we are done. When \( \mu(G_\alpha) = \mu(G_\beta) = 0 \), \( \alpha \) and \( \beta \) cannot be orthogonal unless at least one of \( B_\alpha \) and \( B_\beta \) is \( \emptyset \). If so, then at least one of \( \alpha \) and \( \beta \) is redundant when learning. \( \square \)

**E Proofs Related to Recency Bias**

**Lemma 6.** 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 \( a = \mu(A_1 \cap B_1), e = \mu(A_1 \cap B_2), e = \mu(B_2 \cap A_2), g = \mu(A_2 \cap B_1), b = \mu(A_1 \backslash (B_1 \cup B_2)), d = \mu(B_2 \backslash (A_1 \cup A_2)), f = \mu(A_2 \backslash (B_1 \cup B_2)), h = \mu(B_1 \backslash (A_1 \cup A_2)) \). 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 \( \mu^{\beta} \).

Since \( A_1 \succ_\mathcal{A} A_2 \) 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 \( \alpha, \beta \) are nondegenerate,
\(a + h + g > 0\). The pseudo-Bayes’ rule requires \(b' = b, f' = f\) and

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

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

\[
c' = \frac{a + h + g + c + d + e}{2} \quad \frac{c}{c + d + e}
\]

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

\[
\frac{a - g}{a + h + g} < \frac{c - e}{c + d + e}.
\]

Intuitively, if the probability decrease of \(B_2\) is \(\Delta p_1\) 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 \(\Delta p_2\) 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 (11) has to be true. When \(\mu(A_1) = \mu(A_2)\), (11) becomes also sufficient since \(b - f = g - a + e - c\).

Now we prove that \(\mu(\beta(A_2)) > \mu(\beta(A_1))\). Since we know \(\mu(\beta(B_1)) = \mu(\beta(B_2))\), by symmetry the following condition is necessary and sufficient.

\[
\frac{a' - c'}{a' + b' + c'} < \frac{g' - e'}{g' + f' + e'}
\]

Inequality (11) is equivalent to \(a' - c' < g' - e'\). If \(g' - e' \leq 0\), since \(0 < a' + b' + c' < g' + f' + e'\), inequality (12) is obviously true. Next we assume \(g' - e' > 0\). Inequality (12) reads

\[
(a' - c')(g' + e') + (a' - c')f - (g' - e')(a' + c') - (g' - e')b < 0.
\]

(13)
By inequality (11), \(a' - c' < g' - e'\). Hence we have

\[
\text{LHS of (13)} \leq (a' - c')(g' + e') + (g' - e')(a' + c') - (g' - e')b \\
\leq (a' - c')(g' + e') - (g' - e')(a' + c') - (g' - e')(g - a + e - c) \\
= 2(a'c' - c'g') - (g' - e')(g - a + e - c) \\
= \frac{a + h + g + c + d + e}{2}(g + e)(\frac{a - g}{a + h + g} - \frac{c - e}{c + d + e}) \leq 0
\]

by inequality (11). If \(a + b + c > e + f + g\) the second inequality above is strict. Since \((\alpha, \beta)\) is nondegenerate \(a + b + c > 0\). If \(a + b + c = e + f + g > 0\) then either \(f > 0\) or \(g + e > 0\), both of which imply that LHS of (13) < 0 and we have established our claim.

\[\square\]

**Proposition 1** If \([\alpha, \beta]\) induces recency bias on \(\mu\), then \([\beta, \alpha]\) induces recency bias on \(\mu^\alpha\).

**Proof.** By \([\alpha, \beta]\) induces recency bias for \(\mu\), 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, \(\mu^{\alpha\beta\alpha}(B_2) > \mu^{\alpha\beta\alpha}(B_1)\).

\[\square\]

**Proposition 2** If \([\alpha, \beta]\) induces recency bias on \(\mu\), then \(\mu^{(\alpha\beta)^n}\) and \(\mu^{(\alpha\beta)^{n-1}\alpha}\) converges in total variation to \(\mu^*\) 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 updating on \(\alpha\), the probabilities of \(B_1 \setminus (A_1 \cup A_2)\) and \(B_2 \setminus (A_1 \cup A_2)\) will not change but they will respectively increase and decrease when updating on \(\beta\). Hence \(\mu_n(B_1 \setminus (A_1 \cup A_2))\) and \(\mu_n(B_2 \setminus (A_1 \cup A_2))\) are also convergent. Similarly \(\mu_n(A_1 \setminus (B_1 \cup B_2))\) and \(\mu_n(A_2 \setminus (B_1 \cup B_2))\) 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 \(\mu_n\) 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

\[
\mu_{2n+1}(A_1 \cap B_1) = \frac{\mu_{2n}(A_1 \cup A_2) \mu_{2n}(A_1 \cap B_1)}{2 \mu_{2n}(A_1)}
\]
which implies $\mu_{2n}(A_1 \cup A_2)/(2\mu_{2n}(A_1)) \to 1$ and

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

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

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

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.

\[\square\]

\section*{F Proof of Theorem 3}

\textit{Proof.} Let the sequence of beliefs be $\{\mu_k\}$ with $\mu_k = \mu_{i_1 \ldots i_k}$. Let $\nu$ 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 $\nu$ 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

$$\left(\frac{q_1 + q_2}{2} \ln \frac{p_1 + p_2}{2p_1} + \frac{q_1 + q_2}{2} \ln \frac{p_1 + p_2}{2p_2}\right) \leq q_1 \ln \frac{p_1 + p_2}{2p_1} + q_2 \ln \frac{p_1 + p_2}{2p_2}.$$
Our proof is constructive. For $k$ disjoint
Suppose the proposition is not true. We first prove that for any $G$
Proof of Proposition 3
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
\[ d(\mu_{k+1}||\mu_k) \geq 2||\mu_{k+1} - \mu_k||_{TV} \]
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 $\mu^*$ 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 $\mu^*$ must be a solution to $\{\alpha_j\}_{i=1}^{n}$. Clearly $\mu^*$ 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.

\[ \square \]

G Proof of Proposition 3

Proof. Suppose the proposition is not true. We first prove that for any $n$, there exists pairwise
disjoint $P_1^n, P_2^n, \ldots, P_n^n$ 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
$P_1^n = \{X < a_1^n\} \cup A_1^n$, $P_k^n = \{X < a_k^n\} \cup A_k^n \setminus (\bigcup_{j<k} P_j^n)$. It is easy to show that the $P_k^n$’s
satisfy the conditions.

If $\mu(P_k^n) = 0$, then let $j = \min\{k|\mu(P_k^n) > 0\}$. Set $\alpha = (P_1^n, P_j^n)$. If $\alpha$ is credible, that is
if $\mu(P_j^n) < 1$, $\alpha$ 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 $E_{\mu}(X|B) < E_{\mu}(X|P_j^n \setminus B)$, then $\alpha = (P_1^n, B)$ would do the job.
Therefore, if the scope of persuaion 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 $a^n_0 = M$.

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 $E_\mu(X) = E_\mu(X|P_k^n)$. 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 $(S, \Sigma, \mu^*)$. For $z \in \mathbb{R}$, we want to prove that $\mu^*(\{X_n \leq z\})$ converges to $\mu^*(\{X \leq z\})$. By construction, $a_k^{n-1} \leq E_\mu(X|P_k^n) \leq a_k^n$. 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 $a_k^{n-1} \leq z < a_k^n$. If $k_n = n$, then $a_k^{n-1} \leq z$. 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. \qed

\section*{H Proof of Proposition 4}

\textit{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 $\mu_{i}^{\pi_1\pi_2\cdots\pi_{n+1}} \in \{\hat{\mu}_n\}$. 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 $\alpha_j$ since she receives $\alpha_j$ 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, $\alpha_j$ 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 6 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 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 $\alpha_2$. 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 2 again finishes our proof.

\[\square\]

I Proof of Proposition 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 updating path. If $\mu_i(G_{\alpha_j}) \cdot \mu_i(B_{\alpha_j}) = 0$ for some $j$, then $\alpha_j$ and $\bar{\alpha}_j$ are either degenerate or redundant along any possible updating 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 $\alpha$ and $\beta$ 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 $\alpha_j$-expert will believe in $\alpha_j$ after finite steps. Inductively, since the network is strongly connected and there exists both $\alpha_1$- and $\alpha_2$-experts, after a finite number of periods all decision makers must have reached consensus that $\alpha_1$ and $\alpha_2$ are both true. \[\square\]