Rational Inattention with Multiple Attributes

David Walker-Jones[*]

University of Surrey

January 30, 2023

Abstract

This paper studies a new measure for the cost of learning that allows the different attributes of the options faced by an agent to differ in their associated learning costs. The new measure maintains the tractability of Shannon's classic measure but produces richer choice predictions and identifies a new form of informational bias significant for welfare and counterfactual analysis that is conducted with the multinomial logit model. Necessary and sufficient conditions are provided for optimal agent behavior under the new measure for the cost of learning. Sufficient conditions are also provided for a dataset that identifies the new measure for the cost of learning.

1 Introduction

In many choice environments it is costly for agents to learn about the options that they face because it takes time and effort to acquire and process information. Understanding how agents learn in such environments is crucial for quality economic analysis because the cost of information may result in agents not acquiring all of the relevant information before making a decision. Partially informed agents do not always pick the best available option, which makes welfare analysis more challenging. Further, if what information an agent acquires changes with parameters such as price then counterfactual analysis is also made more difficult.

The standard technique for quantifying the cost of learning in models of rational inattention (RI) is Shannon Entropy (Shannon, 1948; Sims, 2003; Mackowiak, Matejka, & Wiederholt, 2021). Shannon Entropy has an axiomatic foundation, is grounded in the optimal coding of information,

[*] Special thanks to Rahul Deb for all of the support. I would also like to thank Andrew Caplin, Mark Dean, Yoram Halevy, Carolyn Pitchik, and Colin Stewart, for their advice, the government of Ontario for funding, and the anonymous associate editor and referees that provided helpful feedback. An earlier version of this paper was previously circulated under the title "Rational Inattention and Perceptual Distance."

and provides a tractable and flexible framework with which to study agent behavior (Shannon, 1948; Matějka & McKay, 2015; Caplin, Dean, & Leahy, 2018).

While Shannon Entropy has proven to be a valuable tool, it does have limitations in economic environments as they are not what it is designed for. It is, for instance, natural to think that some attributes of the choice environment might be more difficult to learn about than others. Shannon Entropy, however, does not allow for attributes of the choice environment to differ in their associated learning costs because it is a one parameter model for the cost of information, and thus there can only be one level of difficulty when learning. Without a mechanism to allow for what is referred to in the literature as "perceptual distance,"[1] the choice behavior predicted by Shannon Entropy can differ from observed behavior, as is discussed in Example 1 in Section 2.1, which can limit the effectiveness of Shannon Entropy in empirical settings (Dean & Neligh, 2022).

This paper studies a new measure for the cost of learning, Multi-Attribute Shannon Entropy (MASE), that allows for attributes of the choice environment to differ in their associated learning costs. MASE maintains much of the desired tractability of Shannon's classic measure when incorporated into a model of RI because this paper provides the MASE analogues of the famous necessary conditions provided by Matějka and McKay (2015) and necessary and sufficient conditions provided by Caplin et al. (2018) for optimal agent behavior in RI models that use Shannon Entropy.

MASE provides a natural multi-parameter generalization of Shannon Entropy and predicts behavioral patterns that have been identified as problematic for Shannon Entropy. MASE is flexible enough to, for instance, be the foundation of a model of obfuscation in which different firms choose how difficult it is for consumers to learn about the different attributes of their products, while Shannon Entropy is not even flexible enough to allow for different options to differ in their associated learning costs.

MASE also predicts a new informational bias in the multinomial logit random utility (RU) model that should be considered a natural consequence of different learning costs in the same choice environment. Matějka and McKay (2015) show that, in settings where the cost of learning is measured with Shannon Entropy, the value of options that seem appealing to the agent *a priori* are overvalued by multinomial logit in each state of the world and that this bias can be identified with the agent's average choice probabilities[2] as the options that are overvalued are the ones that

[1] If two outcomes are more difficult to differentiate between it is said that they have less perceptual distance between them.

[2] The average choice probability of an option is the weighted average of the probabilities of it being selected in the

have higher average choice probabilities. When the cost of learning is measured with MASE, in contrast, the value of an option according to a multinomial logit regression can be biased upwards in some states of the world and downwards in others, and the presence of a bias cannot necessarily be identified with the agent's average choice probabilities, as is demonstrated by Example 2 in Section 2.2. This is because, with MASE, cheaper to learn about attributes have an overestimated difference between the value of their positive and negative realizations because agent behavior is more sensitive to the realization of cheaper to learn about attributes.

Conditions are also provided in this paper that describe when a dataset is sufficient for the unique identification of the MASE cost function for learning. Such a dataset features observed behavior from simple choice problems, choice problems where two options are available and only a few states of the world occur with a positive probability, and identifies the MASE cost function, both a set of attributes and their associated learning costs, that fully determines the cost of differentiating between outcomes when any subset of the potential states of the world occur with a positive probability.

1.1 Organization of Paper

The remainder of the paper is organized as follows: Section 2 introduces Shannon Entropy, discusses models of RI, and provides motivating examples. In Section 3 MASE, a flexible cost of acquiring information that allows for attributes of the choice environment to differ in their associated learning costs, is introduced and is embedded into a model of RI. Section 3 also discusses the agent behavior predicted by MASE, showing that much of the coveted tractability of Shannon Entropy is maintained by this paper's generalization by establishing necessary and sufficient conditions for optimal behavior. Section 4 discusses the relationship between RU models and the agent behavior found in Section 3, and revisits the motivating examples from Section 2.1 and Section 2.2. Section 5 provides conditions for a dataset that are sufficient for the identification of the MASE cost function for information. Section 6 provides a literature review, and Section 7 concludes. An axiomatic foundation for MASE that weakens Shannon's (1948) original axioms is provided in Appendix 2.

different potential states of the world. Later in the paper this is referred to as the unconditional probability of the option being selected, as is standard in the literature, since the probability does not condition on the state of the world.

2 Rational Inattention and Shannon Entropy

Suppose that the uncertainty faced by the agent is described by a measurable space $(\Omega,\, \mathcal{F})$, where Ω is a finite set of possible **states of the world** (the state space), and $\mathcal{F}$ is the set of **events** generated by Ω (the power set of Ω). The probability measure $\mu : \mathcal{F} \to [0,\, 1]$, which assigns probabilities to events, is referred to as the **prior** belief of the agent. To ease exposition, for the rest of the paper it is assumed that $\mu(\omega) > 0$ for all $\omega \in \Omega$ unless stated otherwise.

Suppose that the agent must make a selection from a set of **options**, denoted $\mathcal{N} = \{1,\, \ldots,\, N\}$. Each option $n \in \mathcal{N}$ in each state of the world $\omega \in \Omega$ has a (finite) **value** to the agent $\mathbf{v}_n(\omega) \in \mathbb{R}$.

In the rational inattention (RI) literature learning by the agent is typically modelled as the choice of a signal structure, which means the agent chooses the probability of receiving different signals in the different states of the world. Receiving a signal updates the agent's belief about the state of the world, giving them a more informed posterior belief. More informative signal structures are more costly for the agent, but allow them to make a more informed decision about which option to select.

The agent's problem is thus to maximize the expected value of their selected option less the cost of learning. They do this by choosing an **information strategy** $F \in \Delta(\mathbb{R} \times \Omega)$, which is a joint distribution between s, the observed **signal**, and the states of the world.[3] The only restriction on the information strategy is that the marginal, $F(\omega) : \mathcal{F} \to \mathbb{R}_+$, must equal the prior μ.

After a signal s is realized, the agent simply picks an action with the highest expected value, denote it by $a(s|F) \in \mathcal{N}$, which thus solves: $\max_{n\in\mathcal{N}} \mathbb{E}_{F(\omega|s)}[\mathbf{v}_n(\omega)]$. Ignoring the cost of learning momentarily, the value to the agent of receiving a signal s, which induces posterior $F(\omega|s)$, is then:

$$V(s|F) = \max_{n\in\mathcal{N}} \mathbb{E}_{F(\omega|s)}[\mathbf{v}_n(\omega)].$$

The agent's problem is to maximize the expected value of the option they select less the cost of learning by choosing an optimal information strategy, and subsequently selecting an option based on the signal produced by their information strategy. Let the expected cost of a particular information strategy, given the agent's prior, be denoted $\mathbf{C}(F,\, \mu)$, and note that the particular form of the cost function studied in this paper is defined by equation (5) in Section 3. The agent's

[3] The decision to allow s to be any real number is rather arbitrary. This is a richer signal space than is required in practice. It is shown that an optimal strategy only results in one of at most N different signals being observed.

problem can thus be written:

$$\max_{F\in\Delta(\mathbb{R}\times\Omega)} \sum_{\omega\in\Omega} \int_s V(s|F)F(ds|\omega)\mu(\omega) - \mathbf{C}(F,\, \mu), \tag{1}$$

$$\text{such that } \forall\omega\in\Omega : \int_s F(ds,\, \omega) = \mu(\omega). \tag{2}$$

The choice behavior the agent exhibits depends on the cost function for information. Shannon Entropy is a measure of total uncertainty that is frequently used to assign costs to information (Matějka & McKay, 2015; Mackowiak et al., 2021). Given a partition (defined formally in Section 3) of the possible states of the world $\mathcal{P} = \{A_1,\, \ldots,\, A_m\}$, and a probability measure μ over these events, the uncertainty about which event has occurred, as measured by **Shannon Entropy**, is defined:[4]

$$\mathcal{H}(\mathcal{P},\, \mu) = -\sum_{i=1}^{m} \mu(A_i)\log(\mu(A_i)). \tag{3}$$

The convention used is this paper is to set $0\log(0) = 0$.

If an agent has prior μ about the state of the world, and their beliefs are updated to the posterior $\mu(\cdot|s)$ after they receive a signal s, then there is a change in the uncertainty as measured by Shannon Entropy. Typically, when Shannon Entropy is used in RI models, the cost of an information strategy F is measured as the expected reduction in total uncertainty as measured by Shannon Entropy:

$$\mathbb{E}\Big[\mathcal{H}(\mathcal{P},\, \mu) - \mathcal{H}(\mathcal{P},\, \mu(\cdot|s))\Big],$$

where $\mathcal{P} = \{\{\omega_1\},\, \{\omega_2\},\, \ldots\}$ is the finest partition of the state space. Henceforth, such a model that uses the expected reduction in Shannon Entropy to measure the cost of an information strategy is referred to as the **Shannon RI model**.

Problems can occur, however, when the Shannon RI model is applied in settings with attributes that differ in their associated learning costs, as is discussed in Example 1, or in settings with options that have different associated learning costs, as is discussed in Example 2.

[4] This measure is only unique up to a positive multiplier.

Table 1: Example 1

State:	ω_1	ω_2	ω_3	ω_4
Balls in State:	60 Blue & 40 Red	51 Blue & 49 Red	49 Blue & 51 Red	40 Blue & 60 Red
Probability of State:	1/4	1/4	1/4	1/4
$\mathbf{v}_1(\omega)$:	y	y	-y	-y
$\mathbf{v}_2(\omega)$:	0	0	0	0

2.1 Example 1: Multiple Attributes and Problems with Predictions

Caplin, Dean, and Leahy (2022, p. 26) show that the Shannon RI model results in choice behavior that satisfies "invariance under compression." That is, when Shannon Entropy is used to measure information, if there are two states of the world, ω_1 and ω_2, across which payoffs are identical for each option ($\mathbf{v}_n(\omega_1) = \mathbf{v}_n(\omega_2)\ \forall n \in \mathcal{N}$), then the probability of each option being selected is the same in ω_1 and ω_2. The invariance under compression that is predicted by Shannon Entropy is, unfortunately, not found in many settings, as is shown by the work of Dean and Neligh (2022). This subsection describes an environment akin to the experiments in Dean and Neligh (2022).

Consider the environment described in Table 1 where an agent is faced with a screen that shows 100 balls, each of which is either red or blue. The agent is offered a prize that they may either accept (option 1), or reject to get a payoff of zero (option 2). The agent is told that if the majority of the balls on the screen are blue then the prize is $y \in \mathbb{R}_{++}$, and if the majority of the balls on the screen are red then the prize is $-y$. Suppose further that the agent is also told that there is a 1/4 chance of each of four different states of the world in which there are either 40, 49, 51, or 60 red balls.

The Shannon RI model, which imposes invariance under compression, predicts that the agent has the same probability of selecting option 1 when there are 40 red balls as when there are 49 red balls, and that the agent has the same probability of selecting option 1 when there are 60 red balls as when there are 51 red balls. This predicted behavior is not intuitive because it should be easier for the agent to differentiate between the states that are more different (40 versus 60 red balls) than the states that are more similar (49 versus 51 red balls). One should instead expect that the probability of option 1 being selected is decreasing in the number of red balls, as is demonstrated by the experiments of Dean and Neligh (2022), because it should be easier to determine which color of ball constitutes the majority the more of that color ball there are.

Why does Shannon Entropy impose this type of behavior? In short, Shannon Entropy results in invariance under compression because of Shannon's third axiom (Shannon, 1948). In the context of Example 1, let $\mathcal{P} = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}\}$, and $\tilde{\mathcal{P}} = \{\{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}\}$, be two partitions of the state space. Shannon's third axiom requires that the total uncertainty about the state of the world is equal to the uncertainty about which event in $\tilde{\mathcal{P}}$ has occurred plus the expected amount of uncertainty that remains about which event in $\mathcal{P}$ has occurred after which event in $\tilde{\mathcal{P}}$ occurred has been learned. This equality means that the reduction in uncertainty caused by a signal, which is the cost of the signal, is equal to the reduction in uncertainty about which event in $\tilde{\mathcal{P}}$ has occurred, plus the expected reduction in uncertainty about which event in $\mathcal{P}$ has occurred given which event in $\tilde{\mathcal{P}}$ has occurred. The agent, however, is only concerned with which event in $\tilde{\mathcal{P}}$ has occurred, as this fully determines payoffs. If agent behavior is different in ω_1 compared to ω_2, or ω_3 compared to ω_4, so that their behavior does not satisfy invariance under compression, then the agent is, to an extent, differentiating between these states, and paying for information that does not benefit them, and their information strategy is thus not optimal.

While other information cost functions do not require that choice behavior satisfies invariance under compression (Caplin et al., 2022; Morris & Yang, 2022), they lack the tractability and flexibility of Shannon Entropy,[5] which limits the potential for their application. This has led to the following open question: "what workable alternative models allow for the complex behavioral patterns identified in practice?" (Caplin, Dean, & Leahy, 2017, p. 2), a question that this paper attempts to answer. MASE solves the problem with predictions outlined in this example by allowing option 1 to have multiple attributes that differ in their learning cost, as is explained in Section 4.2.

2.2 Example 2: Options that Differ in Learning Costs and Biases in Fitting

If attributes vary in their learning costs then RU models are susceptible to a form of informational bias that has not previously been identified, as demonstrated by the following example. This is significant for those who wish to conduct welfare or counterfactual analysis because there are many economically significant examples where, for instance, one option is easier to learn about, as in Example 2.

Consider a choice environment where an agent has two options: option 1 and option 2, which can each be of high value H, or low value $L < H$, as is described in Table 2. Assume, contrary to what is possible with Shannon Entropy, that learning the value of option 1 is less costly than learning the value of option 2. For example, perhaps the agent is interested in investing in one of

[5] Shannon Entropy has a number of mathematical properties that make it easy to use for predicting behavior in a wide range of environments.

Table 2: Example 2

State:	ω_1	ω_2	ω_3	ω_4
Probability of State:	1/4	1/4	1/4	1/4
Value of option 1 in state ($\mathbf{v}_1(\omega)$):	H	H	L	L
Value of option 2 in state ($\mathbf{v}_2(\omega)$):	H	L	H	L

two businesses that are *a priori* identical except for the fact that one is local and easier to learn about, while the other is foreign and harder to learn about. It is not difficult to come up with more examples along these lines.

Because payoffs are symmetric, any knowledge about the value of option 1 has the same value to the agent as the same knowledge about option 2. Further, the cost of said information about option 1 is lower. As such, while the marginal benefit of information about option 1 or option 2 is the same, the marginal cost of information about option 1 is lower. One should thus expect research of a rational agent to be more attentive to option 1, and they should more cognisant of its value as a result.

If both option 1 and option 2 have realized their high value H, one should thus expect that the agent is more likely to select option 1 since our intuition is that the agent should be more cognisant of option 1's high value. Similarly, if option 1 and option 2 have both realized their low value L, then one should expect that the agent is more likely to select option 2.[6]

Because of this, if an econometrician tried to deduce the two values of option 1, H_1 and L_1, and the two values of option 2, H_2 and L_2, using a multinomial logit regression, they would decide that H_1 is more than the true value H, and that L_1 is less than the true value L (as is shown rigorously in Section 4). Fitting thus falls prey to an informational bias, undermining the value of any counterfactual or welfare analysis.

This type of bias has not previously been identified in the literature on RI: Matějka and McKay (2015) show that fitting of multinomial logit results in the value of any option n being biased by the (weighted) average probability of it being selected over states ω. The bias found by Matějka and McKay (2015) can thus be identified by examining the average probabilities of the agent selecting each option because the driving mechanism is that the cost of learning causes the agent to be biased towards options that they have a higher probability of selecting *a priori*. The bias previously found by Matějka and McKay (2015) is fundamentally different than the bias demonstrated in this example because their bias does not allow for an option to be over valued in

[6] Our intuition is that the agent should be more cognisant of option 1's low value.

some states and under valued in others, which is in contrast with our setting where option 1 is over valued when it is of high value, and is undervalued when it is of low value.

An econometrician who observes equal average choice probabilities in this setting, as is predicted by MASE (this is shown rigorously in Section 4.3), might be tempted conclude, based on the previous literature, that their analysis is not susceptible to informational biases since each option has the same unconditional probability of being selected *a priori*, and thus any counterfactual or welfare analysis that they conduct is valid. This conclusion may not be correct given the results in this paper.

RU models and RI models with Shannon Entropy can both be rejected for RI with MASE in this environment if it is possible to alter the correlation between the values of the two options while holding the marginal distributions over values fixed for each option, as is discussed in Section 4.3 when this example is revisited.[7]

3 Inattentive Learning with MASE

This section introduces and solves a model of RI that uses MASE, a multi-parameter generalization of Shannon Entropy, to measure the cost of acquiring information and establishes that MASE can be incorporated tractably into a model of RI, which is not an obvious result. Apart from the use of MASE instead of Shannon Entropy for the measurement of uncertainty, this section follows the work of Matějka and McKay (2015) closely so as to aid comparison between the two models.

The idea behind MASE is that the state of the world may be determined by the realization of multiple attributes and how costly it is for the agent to learn about the realization of a given attribute may differ across attributes, i.e., some attributes may be more costly to learn about than others. As is introduced formally in the coming paragraphs, the different attributes are modelled as different partitions of the state space. This is a natural way of modelling attributes because learning the realization of one attribute rules out some states of the world, but does not necessarily remove all uncertainty about the state, much like learning the realized event of a partition.

A **partition** $\mathcal{P}$ of a state space Ω is a set of more than one disjoint events in $\mathcal{F}$ whose union is Ω.[8] For each event $A \in \mathcal{F}$, define the **complement** of the event, denoted A^c, to be the set of states that are not in A, so $A^c = \Omega \backslash A$, and thus $\{A, A^c\}$ forms a partition.

[7] This assertion is not difficult to show with Theorem 1 and Lemma 2.

[8] Notice that the definition of a partition excludes trivial partitions that only contain a single event.

If $\omega \in \Omega$ is the state of the world, let the **realized event** of the partition $\mathcal{P} = \{A_1, \ldots, A_m\}$ be denoted by $\mathcal{P}(\omega)$, that is $\mathcal{P}(\omega) = A_i \in \{A_1, \ldots, A_m\}$ iff $\omega \in A_i$. Given a prior μ, if the agent **learns the realized events** $\mathcal{P}_1(\omega), \ldots, \mathcal{P}_n(\omega)$ of a collection of partitions, their updated belief is denoted $\mu(\cdot|\cap_{i=1}^n \mathcal{P}_i(\omega))$, and is defined on states $\tilde{\omega}$ as follows:

$$\mu(\tilde{\omega}|\cap_{i=1}^n \mathcal{P}_i(\omega)) = 0 \text{ if } \tilde{\omega} \notin \cap_{i=1}^n \mathcal{P}_i(\omega), \text{ and otherwise } \mu(\tilde{\omega}|\cap_{i=1}^n \mathcal{P}_i(\omega)) = \frac{\mu(\tilde{\omega})}{\mu(\cap_{i=1}^n \mathcal{P}_i(\omega))}.$$

The **attributes**, denoted $\mathcal{A}_1, \ldots, \mathcal{A}_M$, with $M \geq 1$, are a group of partitions whose realized events together indicate the state of the world: $\cap_{i=1}^M \mathcal{A}_i(\omega) = \omega$ for all $\omega \in \Omega$. Each attribute $\mathcal{A}_i$ has a **multiplier**, a strictly positive (finite) constant, λ_i, associated with it, and to ease exposition assume that the attributes are ordered by their multipliers: $0 < \lambda_1 < \ldots < \lambda_M$. The multipliers reflect the difficulty of learning about a given attribute: attributes with larger multipliers associated with them are more costly to learn about. To ease exposition, it is assumed that no attribute is redundant in the sense that they each provide information that is not provided by attributes with lower multipliers: for each $m \in \{2, \ldots, M\}$ there is a state $\omega \in \Omega$ such that $\cap_{i=1}^m \mathcal{A}_i(\omega) \subset \cap_{i=1}^{m-1} \mathcal{A}_i(\omega)$.

Using these attributes and their associated multipliers, define **Multi-Attribute Shannon Entropy** (MASE), $\mathbb{H} : \Delta(\Omega) \to \mathbb{R}_+$, to be the measure of total uncertainty:

$$\mathbb{H}(\mu) \equiv \lambda_1 \mathcal{H}\Big(\mathcal{A}_1, \mu\Big) + \mathbb{E}\Big[\lambda_2 \mathcal{H}\Big(\mathcal{A}_2, \mu(\cdot|\mathcal{A}_1(\omega))\Big) + \cdots + \lambda_M \mathcal{H}\Big(\mathcal{A}_M, \mu(\cdot|\cap_{i=1}^{M-1} \mathcal{A}_i(\omega))\Big)\Big], \tag{4}$$

where $\mathcal{H}$ is Shannon Entropy, which is defined in equation (3). This paper refers to $\mathbb{H}$ as a measure of total uncertainty because, given any probability measure over states, it describes the minimal expected cost of perfectly observing the state of the world, as is true with Shannon Entropy in the Shannon RI model. The formula for $\mathbb{H}$ can be interpreted as describing the agent as learning the state of the world by successively learning the realizations of the different attributes, learning about the less costly to learn about attributes first so as to minimize the cost of learning any information that the attributes share. A more detailed discussion of the rationale for this functional form can be found in Appendix 2.

Define $\mathbf{C}(F, \mu)$, the expected cost of a particular information strategy, to be the expected reduction in total uncertainty the information strategy causes as measured by $\mathbb{H}$:

$$\mathbf{C}(F, \mu) \equiv \mathbb{E}\Big[\mathbb{H}(\mu) - \mathbb{H}(\mu(\cdot|s))\Big], \tag{5}$$

where $\mathbb{H}(\mu)$ is as defined in equation (4) and $\mu(\cdot|s) : \Omega \to \mathbb{R}_+$ is the distribution over states induced by the signal s, the prior μ, the information strategy F, and Bayes' Rule, which is to say $\mu(\omega|s) = F(\omega|s)$ for each state ω. This definition of the cost of learning is the same as in the standard Shannon model of RI studied by Matějka and McKay (2015) except Shannon Entropy is replaced by MASE. Because $\mathbb{H}(\mu)$ is shown by Lemma 3 to be a strictly concave function of μ, the work of Mensch (2018) indicates that $\mathbf{C}(F,\, \mu)$ is monotone in Blackwell (1953) informativeness.

3.1 Selecting Optimal Choice Probabilities

Finding optimal information strategies, joint distributions between signals and states that are a solution to (1) subject to (2), is a complicated and not particularly tractable problem, so this paper follows Matějka and McKay (2015) and re-writes the agent's problem directly in terms of the choice probabilities of the agent. This process requires the development of some new notation. Define $\mathcal{S}(n|F) = \{s \in \mathbb{R} :\, F(s) > 0,\, a(s|F) = n\}$, to be the set of signals that result in the agent selecting option n. Next, define the probability of option n being selected conditional on any state of the world ω to be:

$$\Pr(n|\omega) = \int\limits_{s\in\mathcal{S}(n|F)} F(ds|\omega), \tag{6}$$

and for each event $A \in \mathcal{F}$, define the probability of n being selected conditional on A being realized to be:

$$\Pr(n|A) = \sum_{\omega\in A} \Pr(n|\omega)\mu(\omega|A). \tag{7}$$

Define the **unconditional probability** of option n being selected to be:

$$\Pr(n) = \sum_{\omega\in\Omega} \Pr(n|\omega)\mu(\omega). \tag{8}$$

Denote the collection of $\Pr(n|\omega)$ for each $n \in \mathcal{N}$ and $\omega \in \Omega$ by $\mathbb{P}$, which is referred to as the agent's observable **behavior**. Using this notation, the agent's problem can be re-written:

Lemma 1. Behavior $\mathbb{P}$ is the outcome of a solution to the agent's problem in (1) subject to (2) iff it solves:

$$\max_{\mathbb{P}} \sum_{n\in\mathcal{N}} \sum_{\omega\in\Omega} \mathbf{v}_n(\omega)\Pr(n|\omega)\mu(\omega) - \mathbf{C}(\mathbb{P},\, \mu), \tag{9}$$

$$\text{such that: } \forall\, n \in \mathcal{N},\ \Pr(n|\omega) \geq 0,\ \forall\, \omega \in \Omega, \tag{10}$$

$$\text{and } \sum_{n \in \mathcal{N}} \Pr(n|\omega) = 1 \ \forall\, \omega \in \Omega, \tag{11}$$

where $\mathbf{C}(\mathbb{P},\, \mu)$ is as defined in Lemma 5. Further, the objective described by equation (9) is concave on the set of $\mathbb{P}$ that satisfy (10) and (11).

Proofs for results in Section 3, Section 4, and Section 5, can be found in Appendix 1.

Th new problem outlined in Lemma 1, where the agent selects their conditional choice behavior $\mathbb{P}$, is substantially easier to solve than the problem where the agent picks their information strategy F. If behavior solves (9) subject to (10) and (11) then it is referred to as **optimal**.

Matějka and McKay (2015) show that in the Shannon RI model, which is the special case of MASE where one attribute is used to measure the cost of learning with associated multiplier $\lambda_1 = \lambda$, optimal agent behavior is such that for each option $n \in \mathcal{N}$ the probability of it being selected in state $\omega \in \Omega$ satisfies:

$$\Pr(n|\omega) = \frac{\Pr(n) e^{\frac{\mathbf{v}_n(\omega)}{\lambda}}}{\sum\limits_{\nu \in \mathcal{N}} \Pr(\nu) e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda}}}. \tag{12}$$

In the Shannon RI model the probability of the agent selecting an option in any given state thus depends on both the unconditional probabilities of the options being selected and the realized values of the options in said state.

3.2 Behavior of a Rationally Inattentive Agent with MASE

Using Lemma 1 and MASE instead of Shannon Entropy, a necessary condition for the optimal behavior of the agent in the more general context is established by Theorem 1 below. Said necessary condition simplifies the maximization problem undertaken by the agent, as is demonstrated by Lemma 2.

Theorem 1.

If $\mathbb{P}$ is the solution to (9) subject to (10) and (11) then $\forall\, n \in \mathcal{N}$ if option n is selected with a positive probability, $\Pr(n) > 0$, then $\forall\, \omega \in \Omega$ the probability of it being selected in said state is positive, $\Pr(n|\omega) > 0$, and satisfies:

$$\Pr(n|\omega) = \frac{\Pr(n)^{\frac{\lambda_1}{\lambda_M}} \Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots \Pr(n| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}}{\sum\limits_{\nu \in \mathcal{N}} \Pr(\nu)^{\frac{\lambda_1}{\lambda_M}} \Pr(\nu|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots \Pr(\nu| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_M}}}. \tag{13}$$

Those familiar with the work of Matějka and McKay (2015) will recognize the above formula as the MASE analogue of the necessary condition for optimal behavior stated in Matějka and McKay (2015)'s Theorem 1 and depicted in equation (12) for the Shannon RI model. When there is only one attribute to learn about and $\lambda_1 = \lambda_2 = \dots = \lambda_M = \lambda$, the above formula collapses to the one from Matějka and McKay (2015)'s Theorem 1.

For $m \in \{1, \dots, M-1\}$, $\Pr(n| \cap_{i=1}^{m} \mathcal{A}_i(\omega))$ is the average probability of n being selected given the realizations of the m easiest to learn about attributes. With MASE, as the above formula indicates, the chance that the agent selects an option n in a particular state of the world ω depends on the unconditional probabilities of the options being selected and the realized values of the options, $\Pr(\nu)$ and $\mathbf{v}_\nu(\omega)$ for each $\nu \in \mathcal{N}$, as is the case with Shannon Entropy, but also depends on the realizations of the attributes that are easier to learn about. It makes sense that when easier to observe pieces of information indicate that an option n is likely of above average value, that the agent should select option n with a higher probability, even if the above average value has not been realized. A more complete discussion of the intuitive properties of the choice behavior described in Theorem 1 can be found after its proof in Appendix 1.

Behavior that is consistent with Theorem 1 is not necessarily optimal because in many settings it is not optimal for the agent to choose all available options with a positive probability, and though such a corner solution may be optimal, there are many corners that are consistent with Theorem 1 but are not optimal. For instance, for any $n \in \mathcal{N}$, if the agent selects n with probability one in all states of the world, then their behavior is consistent with Theorem 1, but it is easy to come up with examples where this would not be optimal for any n, as is demonstrated when Example 1 and Example 2 are revisited in Section 4.

Lemma 2.

If behavior $\mathbb{P}$ is such that $\Pr(n|\omega)$ is described by (13) and $\Pr(n|\omega) > 0$ for all $n \in \mathcal{N}$ and $\omega \in \Omega$, then it is a solution to (9) subject to (10) and (11). $\mathbb{P}$ is a solution to (9) subject to (10) and (11) iff it is defined using equation (13) and a solution to:

$$\max_{\mathbb{P}} \sum_{\omega \in \Omega} \lambda_M \log \left(\sum_{n \in \mathcal{N}} \Pr(n)^{\frac{\lambda_1}{\lambda_M}} \Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2 - \lambda_1}{\lambda_M}} \ldots \Pr(n| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega))^{\frac{\lambda_M - \lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}} \right) \mu(\omega),$$

such that:

$$\forall\, A \in \mathcal{F} : \; \Pr(n|A) \geq 0 \; \forall\, n, \quad \text{and} \quad \sum_{n \in \mathcal{N}} \Pr(n|A) = 1.$$

Further, the objective of this new maximization problem is concave on vectors of non-negative numbers of appropriate dimension.

Lemma 2 is helpful for two main reasons. First, Lemma 2 indicates that if behavior is such that all options are selected with a positive probability and Theorem 1 is satisfied, then it is optimal. Second, Lemma 2 reduces the number of choice variables faced by the agent, which means it is easier for the researcher to find optimal agent behavior.

Theorem 4, which can be be found in Appendix 1, provides necessary and sufficient conditions for optimal behavior in settings where MASE is used to measure the cost of information. Theorem 4 thus establishes the MASE analogue of Caplin et al. (2018)'s Proposition 1, their central proposition.

As is true with standard Shannon Entropy, optimal choice behavior may not be unique. If two options are known *a priori* to take the same value in each state of the world, for instance, then the agent can shift probability from one of these two options to the other whenever the former has a strictly positive probability of being selected in an optimal solution. While these sorts of environments are possible, optimal behavior is unique generically. This feature of optimal behavior should be evident since payoffs are linear and costs are convex. The exact sufficient conditions for the uniqueness of a solution are withheld, but for the solution not to be unique, similar to the case with Shannon Entropy studied by Matějka and McKay (2015), a very rigid form of co-movement is required between payoffs and states.

4 Comparisons with Standard Models

This section discusses the relationship between RU models and RI with MASE before revisiting the two motivating examples, Example 1 and Example 2.

4.1 Comparison with Random Utility Model

RU models are frequently used to fit behavior in discrete choice settings. In such a model, the agent picks the option $n \in \mathcal{N}$ with the largest sum $u_n = v_n + \epsilon_n$. Generally, u_n represents the value of the option to the agent, v_n represents the average value of the option across agents, and ϵ_n represents an idiosyncratic value to the agent. The role ϵ_n plays is up to interpretation, however, and is determined by the researchers specification (Train, 2009). In a setting where agents are thought to be rationally inattentive, the above terms are interpreted in a different way because the agent's noisy behavior is generated by perceptual error instead of idiosyncratic differences in taste. In such settings, u_n represents the perceived value to the agent, v_n represents the true value to the agent, and ϵ_n is interpreted as an unobservable perceptual error that results from the noisy information strategy selected by the agent. Woodford (2014) argues that this latter interpretation is necessary in many contexts due to the stochastic responses observed in perceptual discrimination tasks such as those administered by Dean and Neligh (2022), which are akin to Example 1 in Section 2.1. While the interpretation of ϵ_n is relevant for welfare analysis, it is inconsequential for the description of choice behavior. How then can MASE be interpreted in terms of an RU framework, and what insights may be provided about the fitting of RU models?

Matějka and McKay (2015) point out that choice probabilities predicted by RI with Shannon Entropy correspond to multinomial logit choice probabilities where it is as if option values have been shifted due to the agent's prior about potential values. An option that seems more desirable *a priori* is more likely to be selected by the agent in every state of the world, and thus is overvalued by a multinomial logit regression.

Rational inattention with MASE takes this one step further, as is shown by Theorem 2, allowing the shift in perceived value to also depend on easier to observe attributes (attributes that have an associated multiplier that is less than λ_M). This flexibility seems natural in many real world environments. Consider an agent that is trying to select a restaurant to go to. One may expect that the probability of the agent selecting a given option to increase not only with the quality of the restaurant, and their prior impression of it, but also with easy to observe positive pieces of information, such as high on-line ratings the restaurant has received.

Theorem 2:

The choice behavior described by $\mathbb{P}$, a solution to (9) subject to (10) and (11), is identical to the behavior produced by an RU model where each option $n \in \mathcal{N}$ has perceived value:

$$u_n = \tilde{v}_n + \alpha_n + \epsilon_n,$$

where $\tilde{v}_n = \dfrac{\mathbf{v}_n(\omega)}{\lambda_M}$, ϵ_n has an iid Gumbel distribution, and:

$$\alpha_n = \tfrac{\lambda_1}{\lambda_M} \log(N\Pr(n)) + \tfrac{\lambda_2-\lambda_1}{\lambda_M} \log(N\Pr(n|\mathcal{A}_1(\omega))) + \cdots + \tfrac{\lambda_M-\lambda_{M-1}}{\lambda_M} \log(N\Pr(n| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega))).$$

Theorem 2 is meant to provide insight into the outcome of attempting to fit a RU model in an environment where agents are rationally inattentive with a cost function for information described by MASE. Theorem 2 does not say that a model of RI with MASE is equivalent to a RU model. Even if choice data from a given choice problem cannot be used to reject one for the other, across choice problems MASE produces behavior that can reject the hypothesis of a RU model. With MASE, for instance, as with standard Shannon Entropy, adding an option can increase the probability of an existing option being selected, which is not possible with a RU model.

Also, it is worth mentioning that since optimal behavior may result in some options being selected with probability zero, Theorem 2 implicitly defines each α_n on the extended reals so that $\alpha_n = -\infty$ if $\Pr(n) = 0$.[9]

4.2 Example 1 Revisited

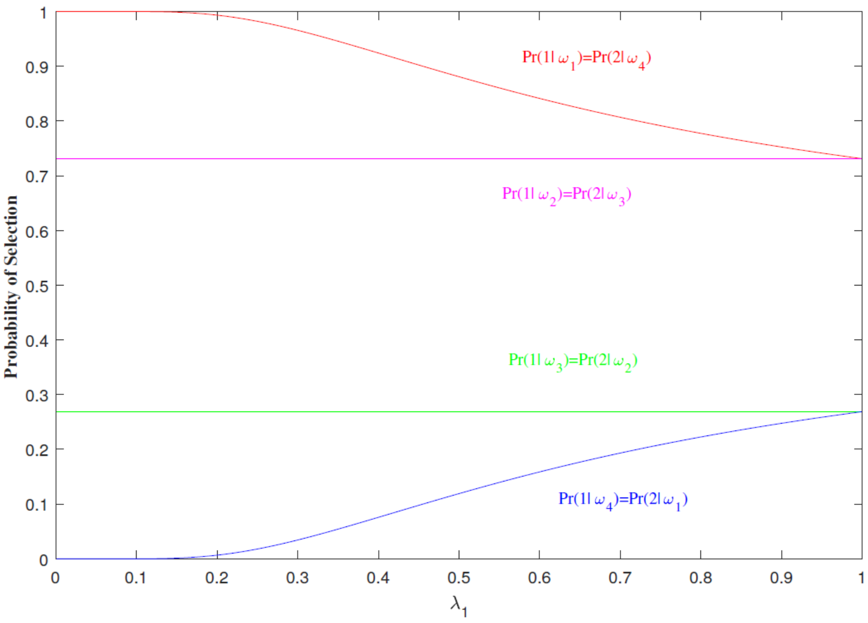
This subsection revisits Example 1 from Section 2.1, which is described in Table 1. It seems natural that it should be easier for the agent to answer the question ‘Are 60 or more of the balls blue?’, than it is for them to answer ‘Are 51 or more of the balls blue?’, as is demonstrated by the experiments of Dean and Neligh (2022) because it should be easier to determine the color of ball that constitutes the majority the more of that color ball there are. Similarly, it seems natural that it should be easier for the agent to answer the question ‘Are 60 or more of the balls red?’, than it is for them to answer ‘Are 51 or more of the balls red?’. Symmetry also implies that the questions ‘Are 60 or more of the balls blue?’ and ‘Are 60 or more of the balls red?’ should have the same expected cost, and the questions ‘Are 51 or more of the balls blue?’ and ‘Are 51 or more of the balls red?’ should have the same expected cost. Thus, assume that $\mathcal{A}_1 = \{A_1, A_2, A_3\} = \{\{\omega_1\}, \{\omega_2, \omega_3\}, \{\omega_4\}\}$ and $\mathcal{A}_2 = \{\{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}\}$.

Solutions to Lemma 2 combined with Theorem 1 mean that the probability of the agent

[9] Theorem 1 shows that if optimal behavior results in $\Pr(n) > 0$, then $\Pr(n|\omega) > 0 \ \forall \omega \in \Omega$.

selecting option 1 is increasing in the number of blue balls, as can be seen in Figure 1, which depicts optimal behavior in each state of the world for a range of λ_1. When λ_1 is small relative to λ_2 the agent chooses option 1 in state ω_1 with a high probability, and choose option 2 in state ω_4 with a high probability. The agent is thus better able to discern the state of the world when there are 40 of one color ball and 60 of the other than when there are 49 of one color and 51 of the other. This is supported by the experimental work of Dean and Neligh (2022), and is in contrast with the behavior predicted by the Shannon RI model.

Figure 1: Optimal Behavior in Example 1 for a Range of λ_1 if $y = 1$ and $\lambda_2 = 1$:



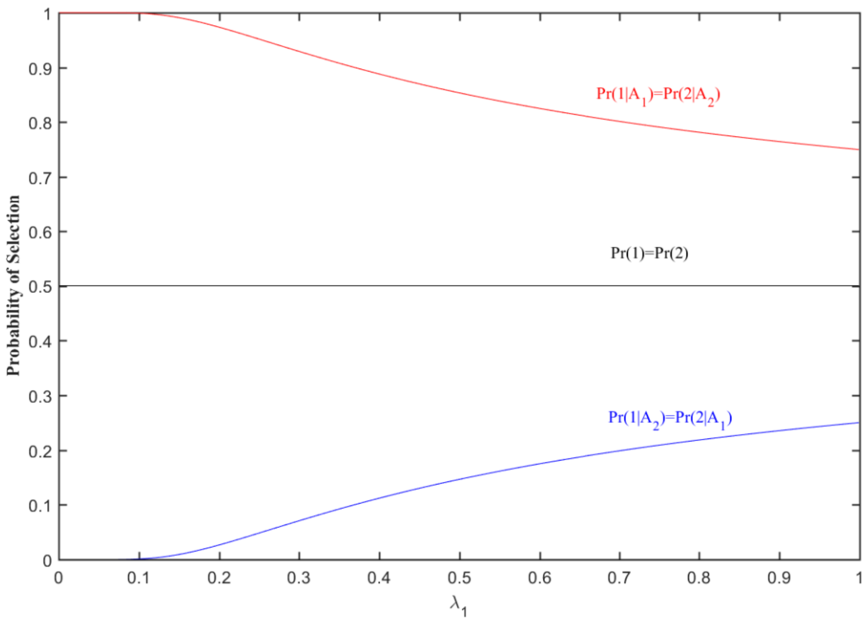
Morris and Yang (2022) identify a related issue with Shannon Entropy's lack of perceptual distance, and warn against its use in some continuous settings because it predicts discontinuous changes in behavior at places where payoffs change discontinuously. In the limit, as the number of different attributes is allowed to grow, MASE can be used to produce the kind of continuous behavior that Morris and Yang (2022) desire.

4.3 Example 2 Revisited

This subsection revisits Example 2 from Section 2.2, which is described in Table 2. It is assumed that learning the value of option 1 is less costly than learning the value of option 2. That is to say, there are two attributes of the choice environment, one determines the value of option 1, the other determines the value of option 2, and the attribute that determines the value of option 1 is less costly to learn about. Thus, assume that: $\mathcal{A}_1 = \{A_1, A_2\} = \{\{\omega_1, \omega_2\}, \{\omega_3, \omega_4\}\}$ and $\mathcal{A}_2 = \{\{\omega_1, \omega_3\}, \{\omega_2, \omega_4\}\}$.

Solutions to Lemma 2 in this environment for a range of λ_1 can be found in Figure 2, which shows that when λ_1 is small compared to λ_2, the agent selects option 1 with a high probability when it is of value H, and selects option 2 with a high probability when option 1 is of value L. As λ_1 increases relative to λ_2, the probability of option 1 being selected when it is of value H decreases. Similarly, as λ_1 increases relative to λ_2, the chance of option 1 being selected when it is of value L increases. Note that the solutions to Lemma 2 mean that the agent is more likely to select option 1 when state ω_1 has been realized since $\Pr(1|A_1) > \Pr(2|A_1)$, and more likely to select option 2 when state ω_4 has been realized since $\Pr(1|A_2) < \Pr(2|A_2)$, as can be observed with Theorem 1.

Figure 2: Solutions to Lemma 2 in Example 2 for a Range of of λ_1 if $H = 10$, $L = 0$, and $\lambda_2 = 1$:



Solutions to Lemma 2 combined with Theorem 2 mean that if an econometrician tries to fit this environment with a multinomial logit model that their estimate of H_1, the high value of option 1, is biased upwards by $\frac{\lambda_2-\lambda_1}{\lambda_2}\log(2\Pr(1|A_1))$, which is greater than zero since $\Pr(1|A_1) > 1/2$, and their estimate of L_1, the low value of option 1, is biased downwards by $\frac{\lambda_2-\lambda_1}{\lambda_2}\log(2\Pr(1|A_2))$, which is less than zero since $\Pr(1|A_2) < 1/2$. These biases are despite the fact that the unconditional probability of either option being selected is the same: $\Pr(1) = \Pr(2) = 1/2$. As such, the econometrician may have believed their analysis was not susceptible to informational biases if they had used Shannon Entropy to model the environment.

Further, as is mentioned in Section 2, RU models and the Shannon RI model can both be rejected for RI with MASE if it is possible to alter the correlation between the values of the two options while holding the marginal distributions over values fixed for each option by changing the distribution over states in this example so it is no longer uniform.[10] If a RU model describes the agent, then changing the correlation between the values of the two options would not change the choice behavior of the agent in any state. If the behavior of the agent is instead described by MASE, then changing the correlation between the values of the two options would change the choice behavior of the agent in individual states because the total information that can be acquired from learning the value of option 1 (the option that is easier to learn about) changes with the correlation of the options' values. Further, if the above MASE specification is correct, the unconditional choice probabilities of the agent would remain constant when correlation is changed due to the symmetry of the environment, as long as the agent is doing some learning.[11] Finally, if the behavior of the agent is instead described by Shannon Entropy, in contrast, then the choice behavior in the individual states could only change if the unconditional choice probabilities changed.

5 Identification of the Cost of Learning

What data is required to uniquely identify $\mathbb{H} : \Delta(\Omega) \to \mathbb{R}_+$? Theorem 3 demonstrates that if the payoff functions $\mathbf{v}_n : \Omega \to \mathbb{R}$ for the options $n \in \mathcal{N}$ satisfy certain properties, then variation in the belief of the agent and the set of options that they choose from is sufficient for uniquely identifying $\mathbb{H}$.

Let $\mathcal{M} \subseteq \mathcal{N}$ denote a non-empty subset of the options available to the agent, and let $\mathbb{P}^*(\mathcal{M}, \mu)$ denote an **optimal behavior** of the agent when their set of options is $\mathcal{M}$ and their prior belief is μ, that is, a set of $\Pr(m|\omega)$ for each $m \in \mathcal{M}$ and $\omega \in \Omega$ that solve (9) subject to (10) and (11) when the prior over states is μ and the agent is further constrained so $\Pr(n) = 0$ if $n \in \mathcal{N} \backslash \mathcal{M}$. Further, for each pair of states ω_i and ω_j in Ω such that $\omega_i \neq \omega_j$, let $\lambda(\omega_i, \omega_j)$ denote the multiplier associated with the cheapest attribute that allows for differentiating between the two states, that is, $\lambda(\omega_i, \omega_j)$ is the unique constant such that if $\mu(\omega_i) = \mu(\omega_j) = \frac{1}{2}$, then $\mathbb{H}(\mu) = \lambda(\omega_i, \omega_j)(-\log(\frac{1}{2}))$.

[10] This assertion and the assertions that follow in this paragraph are not difficult to show with Theorem 1 and Lemma 2.

[11] The agent is doing some learning if their choice probabilities differ at all in states of the world that are realized with positive probability.

Notice that the attributes $\mathcal{A}_1, \ldots, \mathcal{A}_M$, with $M \geq 1$, whose realized events together indicate the state of the world: $\cap_{i=1}^{M} \mathcal{A}_i(\omega) = \omega$ for all $\omega \in \Omega$, and their associated multipliers $\lambda_i > 0$ for each attribute $\mathcal{A}_i$ with $\lambda_M > \ldots > \lambda_1 > 0$, define $\mathbb{H} : \Delta(\Omega) \to \mathbb{R}$, and as a result determine the cost of any behavior, denoted $\mathbf{C}(\mathbb{P}, \mu)$, as defined in Lemma 5.

Theorem 3: Assume $\mathbb{P}^*(\mathcal{M}, \mu)$ is known for each $\mathcal{M} \subseteq \mathcal{N}$ with exactly two options and each $\mu \in \Delta(\Omega)$ that assigns a strictly positive probability to four or less states. If for each pair of states ω_i and ω_j in Ω with $\omega_i \neq \omega_j$ there are options n and m in $\mathcal{N}$ such that at least one of the following conditions **(i)**-**(v)** are satisfied, then $\mathbb{H}$ is uniquely identified. Further, for each such pair of states, whether or not at least one of the following conditions **(i)**-**(v)** are satisfied is observable.

Condition **(i)**: One of the options is better in ω_i while the other is better in ω_j:

$$\mathbf{v}_n(\omega_i) - \mathbf{v}_m(\omega_i) > 0 \text{ and } \mathbf{v}_m(\omega_j) - \mathbf{v}_n(\omega_j) > 0.$$

Condition **(ii)**: One of the options is better in both ω_i and ω_j, but is better by different amounts in these two states, and there is a third state ω_k where the other option is better:

$$\mathbf{v}_n(\omega_i) - \mathbf{v}_m(\omega_i) > 0,\, \mathbf{v}_n(\omega_i) - \mathbf{v}_m(\omega_i) \neq \mathbf{v}_n(\omega_j) - \mathbf{v}_m(\omega_j) > 0, \text{ and } \mathbf{v}_m(\omega_k) - \mathbf{v}_n(\omega_k) > 0.$$

Condition **(iii)**: One of the options is better in one of the states, assuming without loss that this state is ω_i, neither option is better in the other state ω_j, and there is a third state ω_k such that the option that is not better in ω_i is better in ω_k and the cost of differentiating between ω_i and ω_j differs from the cost of differentiating between ω_j and ω_k:

$$\mathbf{v}_n(\omega_i) - \mathbf{v}_m(\omega_i) > \mathbf{v}_n(\omega_j) - \mathbf{v}_m(\omega_j) = 0 < \mathbf{v}_m(\omega_k) - \mathbf{v}_n(\omega_k) \text{ and } \lambda(\omega_i, \omega_j) \neq \lambda(\omega_j, \omega_k).$$

Condition **(iv)**: One of the options is better in both ω_i and ω_j by the same amount, and there is a third state ω_k such that the other option is better in ω_k and the cost of differentiating between ω_i and ω_k differs from the cost of differentiating between ω_j and ω_k:

$$\mathbf{v}_n(\omega_i) - \mathbf{v}_m(\omega_i) = \mathbf{v}_n(\omega_j) - \mathbf{v}_m(\omega_j) > 0 < \mathbf{v}_m(\omega_k) - \mathbf{v}_n(\omega_k) \text{ and } \lambda(\omega_i, \omega_k) \neq \lambda(\omega_j, \omega_k).$$

Condition **(v)**: Neither option is better in either ω_i or ω_j and there are two more states ω_k and ω_r such that one of the options is better in ω_k while the other is better in ω_r, the cost of differentiating between ω_i and ω_k differs from the cost of differentiating between ω_i and ω_r, the cost of differentiating between ω_j and ω_k differs from the cost of differentiating between ω_j and ω_r, and, in addition, either the cost of differentiating between ω_i and ω_k differs from the cost of differentiating between ω_j and ω_k or the cost of differentiating between ω_i and ω_r differs from the cost of differentiating between ω_j and ω_r:

$$\mathbf{v}_n(\omega_i) - \mathbf{v}_m(\omega_i) = 0 = \mathbf{v}_n(\omega_j) - \mathbf{v}_m(\omega_j),\, \mathbf{v}_n(\omega_k) - \mathbf{v}_m(\omega_k) > 0 < \mathbf{v}_m(\omega_r) - \mathbf{v}_n(\omega_r),$$

$$\lambda(\omega_i,\, \omega_k) \neq \lambda(\omega_i,\, \omega_r),\ \lambda(\omega_j,\, \omega_k) \neq \lambda(\omega_j,\, \omega_r),\ \text{and}\ \lambda(\omega_i,\, \omega_k) \neq \lambda(\omega_j,\, \omega_k)\ \text{or}\ \lambda(\omega_i,\, \omega_r) \neq \lambda(\omega_j,\, \omega_r).$$

The proof of Theorem 3 demonstrates that if for each pair of states one of the conditions **(i)-(v)** are satisfied, then there is a finite number of $\mathbb{P}^*(\mathcal{M},\, \mu)$ that demonstrate this and uniquely identify $\mathbb{H}$. Theorem 3 does not say that behavior uniquely identifies the attributes, as there can be different sets of attributes that produce the same $\mathbb{H}$, as is discussed in Appendix 2.

The intuition behind the proof of Theorem 3 is as follows: $\mathbb{H}$ can be identified as long as for each pair of states ω_i and ω_j the multiplier associated with the cheapest attribute that allows for differentiating between them can be identified. Such multipliers can be identified as long as optimal behavior is observed in a choice environment with limited options and possible states and the agent has choice probabilities for the options that optimally differ across the two states in said choice environment. If one option is better in one state while another option is better in the other state, then identifying the multiplier associated with the cheapest attribute that allows for differentiating between the pair of states is simple as Theorem 4 indicates that there is a distribution over these two states that results in the agent optimally selecting choice probabilities that differ in the two states when the two options are the only ones available, and Theorem 1 then indicates that this difference across states identifies the desired multiplier. If two states feature the same ranking of the values of all options, or produce the same value for each option, then the task is made more difficult, but not impossible if other states exist that can be introduced into the choice environment that result in the agent optimally selecting differing choice probabilities in the two states of interest. The proof of Theorem 3 is constructive in the sense that, if a pair of states satisfies one of the five conditions, the proof of Theorem 3 demonstrates how to achieve a closed-form solution for the lowest cost of differentiating between the two states, and how to use these lowest costs for each pair of states to construct $\mathbb{H}$.

With the assumptions that are outlined in Section 4.2 and Section 4.3, each pair of states from both Example 1 and Example 2 satisfy one of the five conditions from the statement of Theorem 3. As each of these examples only features two options for the agent, variation in the distribution over states and observation of the resultant optimal agent behavior would thus be sufficient for identifying the cost of learning in both instances.

6 Literature Review

Shannon Entropy has been used in several contexts to demonstrate informational biases in RU models. Matějka and McKay (2015) use the Shannon RI model to demonstrate the potential for informational biases in multinomial logit, while Steiner, Stewart, and Matějka (2017) use Shannon Entropy in a model of RI to demonstrate the potential for a similar bias in dynamic Logit. These results are significant for those who wish to fit RU models because, while observational data may coincide with the assumptions of a fitted RU model, informational biases can potentially invalidate counterfactual and welfare analysis, two common goals of such a fitting.

The Shannon RI model has also led to a number of predictive successes. Acharya and Wee (2020) show that using Shannon Entropy to model firms as rationally inattentive results in a better fitting of labor market dynamics after the great depression. Dasgupta and Mondria (2018) show that using Shannon Entropy to model importers as rationally inattentive results in novel predictions that are supported by trade data. Ambuehl, Ockenfels, and Stewart (2022) experimentally verify predictions of Shannon Entropy in environments where agents are rationally inattentive to the consequences of participating in different transactions.

Perhaps as a response to the success Shannon Entropy has enjoyed, several recent papers have noted that Shannon Entropy may be a poor measure of the cost of acquiring information in some environments (Caplin et al., 2022; Morris & Yang, 2022) because it lacks what is called "perceptual distance" (Caplin et al., 2022, p. 31). As was alluded to previously, these papers argue that (i) more similar outcomes (outcomes that have less perceptual distance between them) should be more difficult to differentiate between, and (ii) when this property is missing, predicted behavior can differ signficantly from the type of behavior that it would seem natural to expect (Morris & Yang, 2022; Dean & Neligh, 2022).

To better understand the relationship between the cost of learning and agent behavior, a number of papers have studied axiomatic models of rational inattention. Different papers, however, choose to focus their axioms on different aspects of the choice environment. Caplin et al. (2022), for instance, develop axioms that focus on the choice behavior of an agent after they expend effort to learn about the state of the world. In contrast, de Oliveira (2014) and de Oliveira, Denti, Mihm, and Ozbek (2017) develop axioms that focus on an agent's preferences over choice menus before they expend effort to learn about the state of the world. Broadly, these papers aim to understand what implications rational agent behavior has for the form of information cost functions.

Ellis (2018) features axioms that focus on choice behavior and studies the implications for information cost functions, but further assumes that the agent learns by picking a partition of the state space. While MASE uses the cost of learning the realized event of partitions as a primitive, the model studied in this paper does not constrain agents so that they must learn using partitions of the state space, and it can be shown that in a model of RI with MASE it is never optimal for the agent to choose an information strategy that is equivalent to a partition of the state space.[12]

Closer in nature to the work done in this paper, Pomatto, Strack, and Tamuz (2022) develop axioms that focus directly on the costs of information. Axioms that focus on costs for information are interesting because intuitive properties for costs of information can lead to unintuitive agent behavior that is compelling given real-world observations (Gigerenzer & Todd, 1999), but is often mistaken for irrational when axioms that appear rational are imposed on behavior. MASE, for instance, predicts 'non-compensatory' behavior, whereby changing an option so that it is more valuable to the agent can result in a lower chance of it being selected. This type of behavior raises important questions for welfare and counterfactual analysis, making effective policy design more challenging.

Unlike the work of Pomatto et al. (2022), which features axioms that are concerned with probabilistic experiments that can result in different outcomes in the same state of the world, this paper's cost of information is based on axioms, which can be found in Appendix 2, that are concerned with deterministic experiments (questions) that always result in the same outcome in a given state of the world, and contradicts the form of constant marginal cost assumed in their paper.

The cost functions defined with MASE are in the class of posterior-separable cost functions and are, in particular, uniformly posterior separable (Caplin et al., 2022; Denti, 2022). There is a recent literature that has provided foundations for posterior-separable cost functions using models of optimal sequential information sampling (Morris & Strack, 2019; Bloedel & Zhong, 2021).

The cost functions explored in this paper that measure the reduction in MASE are a strict

[12] This is true whenever the agent does some costly learning.

subset of the neighborhood-based cost functions described by Hébert and Woodford (2021). While symmetry imposes a unique set of partitions in Example 1 when MASE is used, there are numerous representations that can be used when a neighborhood-based cost function is assumed. Hébert and Woodford (2021) suggest a way of modelling the neighborhoods in such a setting, which is fitted by Dean and Neligh (2022), that is not equivalent to the unique partitions suggested by MASE.

Huettner, Boyacı, and Akçay (2019), in turn, create an ad hoc group of cost functions that are also a generalization of Shannon Entropy, but are a subset of the cost functions studied in this paper that measure reduction in MASE. The cost functions studied by Huettner et al. (2019) allow different options to have different learning costs associated with them, but are not capable of predicting the behavior that is argued to be intuitive in Example 1 without additional states of the world being introduced. Further, the sufficient conditions for optimal behavior provided by Huettner et al. (2019) contradict the sufficient conditions provided by this paper's Theorem 4.

7 Conclusion

Models of rational inattention that use Shannon Entropy to measure the cost of learning can help to better fit observed data in a range of contexts and also demonstrate that informational biases in random utility models can be significant for welfare and counterfactual analysis. While Shannon Entropy is a flexible and tractable tool, it does not allow for the attributes of the options an agent is choosing between to differ in their associated learning costs, which limits its application in economic environments.

This paper contributes to the literature by proposing and axiomatizing a new measure of uncertainty, Multi-Attribute Shannon Entropy (MASE) that allows for the different attributes of the options faced by an agent to differ in their associated learning costs. MASE is shown to be a natural multi-parameter generalization of Shannon Entropy that maintains much of the tractability of Shannon's standard measure. Theorem 1 establishes the MASE analogue of Matějka and McKay (2015)'s necessary conditions for optimal behavior in the context of Shannon Entropy, and Theorem 4 establishes the MASE analogue of Caplin et al. (2018)'s necessary and sufficient conditions for optimal behavior in the context of Shannon Entropy.

Sufficient conditions for uniquely identify MASE from choice data are also provided. The identification of MASE relies on variation of the set of options the agent can select from and variation in the distribution over states, though binary choice data when relatively few states occur with a positive probability is sufficient for identification in many settings.

MASE also identifies a new form of informational bias demonstrated in Theorem 2. The new form of bias can be present even when the agent has the same probability of selecting each option, which may seem to indicate an unbiased environment based on the previous literature. The biases that have previously been identified in the literature are independent of the realized state of the world, depending only on the agent's prior about the environment. The informational biases that MASE identify are caused by attributes varying in their associated learning costs and can result in the same option being overvalued by a multinomial logit model for some realizations of its attributes and undervalued for other realizations of its attributes.

References

Acharya, S., & Wee, S. L. (2020). Rational inattention in hiring decisions. *American Economic Journal: Macroeconomics*, *12*(1), 1–40.

Ambuehl, S., Ockenfels, A., & Stewart, C. (2022). Who opts in? composition effects and disappointment from participation payments. *The Review of Economics and Statistics*, 1–45.

Blackwell, D. (1953). Equivalent comparisons of experiments. *The annals of mathematical statistics*, 265–272.

Bloedel, A. W., & Zhong, W. (2021). The cost of optimally acquired information. *Unpublished Manuscript, June*.

Caplin, A., Dean, M., & Leahy, J. (2017). *Rationally inattentive behavior: Characterizing and generalizing shannon entropy* (Tech. Rep.). National Bureau of Economic Research.

Caplin, A., Dean, M., & Leahy, J. (2018). Rational inattention, optimal consideration sets, and stochastic choice. *The Review of Economic Studies*, *86*(3), 1061–1094.

Caplin, A., Dean, M., & Leahy, J. (2022). Rationally inattentive behavior: Characterizing and generalizing shannon entropy. *Journal of Political Economy*, *130*(6), 1676–1715.

Dasgupta, K., & Mondria, J. (2018). Inattentive importers. *Journal of International Economics*, *112*, 150–165.

Dean, M., & Neligh, N. L. (2022). Experimental tests of rational inattention.

Denti, T. (2022). Posterior separable cost of information. *American Economic Review*, *112*(10), 3215–59.

de Oliveira, H. (2014). *Axiomatic foundations for entropic costs of attention* (Tech. Rep.). Mimeo.

de Oliveira, H., Denti, T., Mihm, M., & Ozbek, K. (2017). Rationally inattentive preferences and hidden information costs. *Theoretical Economics*, *12*(2), 621–654.

Ellis, A. (2018). Foundations for optimal inattention. *Journal of Economic Theory*, *173*, 56–94.

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In *Simple heuristics that make us smart* (pp. 3–34). Oxford University Press.

Hébert, B., & Woodford, M. (2021). Neighborhood-based information costs. *American Economic Review*, *111*(10), 3225–55.

Huettner, F., Boyacı, T., & Akçay, Y. (2019). Consumer choice under limited attention when alternatives have different information costs. *Operations Research*.

Lange, K. (2013). *Optimization* (Second Edition ed.; G. Casella, I. Olkin, & S. Fienberg, Eds.).

Springer.

Mackowiak, B., Matejka, F., & Wiederholt, M. (2021). Rational inattention: A review.

Matějka, F., & McKay, A. (2015). Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review*, *105*(1), 272–98.

Mensch, J. (2018). Cardinal representations of information. *Available at SSRN 3148954*.

Morris, S., & Strack, P. (2019). The wald problem and the relation of sequential sampling and ex-ante information costs.

Morris, S., & Yang, M. (2022). Coordination and continuous stochastic choice. *The Review of Economic Studies*, *89*(5), 2687–2722.

Noguchi, T., & Stewart, N. (2014). In the attraction, compromise, and similarity effects, alternatives are repeatedly compared in pairs on single dimensions. *Cognition*, *132*(1), 44–56.

Noguchi, T., & Stewart, N. (2018). Multialternative decision by sampling: A model of decision making constrained by process data. *Psychological review*, *125*(4), 512.

Pomatto, L., Strack, P., & Tamuz, O. (2022). The cost of information.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Sims, C. A. (2003). Implications of rational inattention. *Journal of monetary Economics*, *50*(3), 665–690.

Steiner, J., Stewart, C., & Matějka, F. (2017). Rational inattention dynamics: Inertia and delay in decision-making. *Econometrica*, *85*(2), 521–553.

Stewart, N., Chater, N., & Brown, G. D. (2006). Decision by sampling. *Cognitive psychology*, *53*(1), 1–26.

Train, K. E. (2009). *Discrete choice methods with simulation.* Cambridge university press.

Woodford, M. (2014). Stochastic choice: An optimizing neuroeconomic model. *American Economic Review*, *104*(5), 495–500.

Appendix 1

Lemma 1 shows that the agent's problem can be re-written in terms of selecting the choice probabilities described in equations (6), (7), and (8). Before proving this, three other lemmas are introduced. Lemma 3 shows that $\mathbb{H}(\mu)$ is a strictly concave function of μ. This is a commonly known property of Shannon Entropy, but needs to be established for MASE. Lemma 4 shows that $\mathbf{C}$ (defined in equation (5)) is a convex function of feasible information strategies F and any selected action is associated with a particular posterior distribution with probability one. This is desirable because it means an optimal information strategy is a 'recommendation strategy,' where the signals are simply recommendations of what option $n \in \mathcal{N}$ should be selected by the agent. Lemma 5 shows that the cost function for information can be re-written in terms of the choice probabilities in equations (6), (7), and (8).

Lemma 3. Given a non-empty set of attributes (partitions) $\mathcal{A}_1, \ldots, \mathcal{A}_M$, with associated multipliers $\lambda_M > \ldots > \lambda_1 > 0$, such that $\cap_{i=1}^{M} \mathcal{A}_i(\omega) = \omega$ for all $\omega \in \Omega$, the resultant $\mathbb{H}(\mu)$ (defined using the attributes and the multipliers as described in equation (4)) is a strictly concave function of μ. Namely, if there are probability measures μ_a and μ_b on Ω such that for some $\alpha \in (0,1)$ and $\forall \omega \in \Omega : \mu(\omega) = \alpha\mu_a(\omega) + (1-\alpha)\mu_b(\omega)$, and $\mu_a \neq \mu_b$, then $\mathbb{H}(\mu) > \alpha\mathbb{H}(\mu_a) + (1-\alpha)\mathbb{H}(\mu_b)$.

Proof. For each such μ_a, μ_b, $\alpha \in (0,1)$, and μ, the strict concavity of Shannon Entropy (Matějka & McKay, 2015; Caplin et al., 2022) implies:

$$\mathcal{H}(\mathcal{A}_1, \mu) \geq \alpha\mathcal{H}(\mathcal{A}_1, \mu_a) + (1-\alpha)\mathcal{H}(\mathcal{A}_1, \mu_b).$$

Define a random variable X that takes value 1 with chance α, and takes value 0 with chance $1-\alpha$, so that a draw from μ is equivalent to a draw of X, and then a draw according to the probability measure $X\mu_a + (1-X)\mu_b$. If $M \geq 2$, for each $i \in \{2, \ldots, M\}$ and probability measure $\nu : \mathcal{A}_i \times \{0, 1\} \to [0, 1]$, define:

$$\mathcal{H}(X, \nu) = -\sum_{X} \nu(x)\log(\nu(x)), \;\; \mathcal{H}(\mathcal{A}_i, X, \nu) = -\sum_{A \in \mathcal{A}_i} \sum_{X} \nu(A, x)\log(\nu(A, x)).$$

Then, for each such μ_a, μ_b, $\alpha \in (0,1)$, and μ such that $\mu = \alpha\mu_a + (1-\alpha)\mu_b$, and $i \in \{2, \ldots, M\}$, if $M \geq 2$ the properties of Shannon Entropy tell us:

$$\mathbb{E}\Big[\mathcal{H}(\mathcal{A}_i, X, \mu(\cdot| \cap_{j=1}^{i-1} \mathcal{A}_j(\omega)))\Big] = \mathbb{E}\Big[\mathcal{H}(\mathcal{A}_i, \mu(\cdot| \cap_{j=1}^{i-1} \mathcal{A}_j(\omega)))\Big] + \mathbb{E}\Big[\mathcal{H}(X, \mu(\cdot| \cap_{j=1}^{i} \mathcal{A}_j(\omega)))\Big],$$

$$\mathbb{E}\Big[\mathcal{H}(\mathcal{A}_i,\, X,\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{A}_j(\omega)))\Big] = \mathbb{E}\Big[\mathcal{H}(X,\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{A}_j(\omega)))\Big] + \mathbb{E}\Big[\mathcal{H}(\mathcal{A}_i,\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{A}_j(\omega),\, X))\Big],$$

$$\implies \mathbb{E}\Big[\mathcal{H}(\mathcal{A}_i,\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{A}_j(\omega)))\Big] = \mathbb{E}\Big[\mathcal{H}(\mathcal{A}_i,\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{A}_j(\omega),\, X))\Big]$$

$$+\mathbb{E}\Big[\mathcal{H}(X,\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{A}_j(\omega)))\Big] - \mathbb{E}\Big[\mathcal{H}(X,\, \mu(\cdot|\cap_{j=1}^{i}\mathcal{A}_j(\omega)))\Big]$$

$$\geq \mathbb{E}\Big[\mathcal{H}(\mathcal{A}_i,\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{A}_j(\omega),\, X))\Big]$$

$$= \mathbb{E}\Big[\alpha\mathcal{H}(\mathcal{A}_i,\, \mu_a(\cdot|\cap_{j=1}^{i-1}\mathcal{A}_j(\omega))) + (1-\alpha)\mathcal{H}(\mathcal{A}_i,\, \mu_b(\cdot|\cap_{j=1}^{i-1}\mathcal{A}_j(\omega)))\Big].$$

The above inequality is strict for at least one $i \in \{2,\, \ldots,\, M\}$ if $M \geq 2$ and the inequality from the previous paragraph is not strict, and the inequality from the previous paragraph is strict if $M = 1$ and $\mu_a \neq \mu_b$ as $\mathcal{H}$ is strictly concave. The desired result thus follows from the definition of $\mathbb{H}$ (in equation (4)) and the definition of the attributes and multipliers. ■

Lemma 4. $\mathbf{C}(F,\, \mu)$ is convex in information strategies F that satisfy equation (2): if F^1 and F^2 are two information strategies that satisfy (2), and an information strategy F is defined by $F(s,\, \omega) = \alpha F^1(s,\, \omega) + (1-\alpha)F^2(s,\, \omega)$ for some $\alpha \in (0,\, 1)$ and each $\omega \in \Omega$ and $s \in \mathbb{R}$, then F satisfies (2) and $\mathbf{C}(F,\, \mu) \leq \alpha\mathbf{C}(F^1,\, \mu) + (1-\alpha)\mathbf{C}(F^2,\, \mu)$. Further, if action $n \in \mathcal{N}$ is selected with positive probability, $\Pr(n) > 0$, as the outcome of information strategy F that is a solution to (1) subject to (2), then there exists a posterior belief B_n such that $F(\cdot|s) = B_n$ with probability one whenever n is selected.

Proof. It is evident that F satisfies (2) as for each $\omega \in \Omega$:

$$\mu(\omega) = \alpha \int_s F^1(ds,\, \omega) + (1-\alpha)\int_s F^2(ds,\, \omega) = \int_s F(ds,\, \omega).$$

Next, notice that for each s:

$$F(s) = \sum_{\omega\in\Omega} F(s,\, \omega) = \sum_{\omega\in\Omega}\Big(\alpha F^1(s,\, \omega) + (1-\alpha)F^2(s,\, \omega)\Big) = \alpha F^1(s) + (1-\alpha)F^2(s),$$

and for each s with $F(s) > 0$ and ω:

$$F(\omega|s) = \frac{F(s,\, \omega)}{F(s)} = \frac{\alpha F^1(s,\, \omega)}{F(s)} + \frac{(1-\alpha)F^2(s,\, \omega)}{F(s)} = \frac{\alpha F^1(s)}{F(s)}F^1(\omega|s) + \frac{(1-\alpha)F^2(s)}{F(s)}F^2(\omega|s).$$

As a result, Lemma 3 thus implies that for each s with $F(s) > 0$:

$$\mathbb{H}(F(\cdot|s)) \geq \frac{\alpha F^1(s)}{F(s)}\mathbb{H}(F^1(\cdot|s)) + \frac{(1-\alpha)F^2(s)}{F(s)}\mathbb{H}(F^2(\cdot|s))$$

$$\Rightarrow \mathbb{H}(F(\cdot|s))F(s) \geq \alpha\mathbb{H}(F^1(\cdot|s))F^1(s) + (1-\alpha)\mathbb{H}(F^2(\cdot|s))F^2(s).$$

So:

$$\mathbf{C}(F,\, \mu) \leq \alpha\mathbf{C}(F^1,\, \mu) + (1-\alpha)\mathbf{C}(F^2,\, \mu).$$

Further, if F is a solution to (1) subject to (2) then it is impossible that there are two distinct sets of signals $\mathcal{S}^1(n|F)$ and $\mathcal{S}^2(n|F)$ which are observed with strictly positive probability, both of which lead to the selection of n, and induce different posteriors: $F(\cdot|s_1) \neq F(\cdot|s_2)$ for all $s_1 \in \mathcal{S}^1(n|F)$ and $s_2 \in \mathcal{S}^2(n|F)$. This is because if the agent replaced their original information strategy F with a new information strategy $\tilde{F}$ which is identical to F except the signals in $\mathcal{S}^1(n|F)$ and $\mathcal{S}^2(n|F)$ are replaced by s_0 defined $\forall\omega \in \Omega$ by:

$$\tilde{F}(s_0|\omega) = \int\limits_{s\in\mathcal{S}^1(n|F)} F(ds|\omega) + \int\limits_{s\in\mathcal{S}^2(n|F)} F(ds|\omega),$$

then since $\mathbb{H}$ is strictly concave in μ, as established by Lemma 3, the agent would strictly reduce their cost of learning, but the expected value of the option selected by the agent is the same, so they do strictly better. The expected value of the option selected by the agent is the same because payoffs are linear, and the law of iterated expectations implies it is still optimal for the agent to select n after s_0 is realized since $\forall\, \nu \in \mathcal{N}$:

$$\begin{aligned}
\mathbb{E}_{\tilde{F}}[\mathbf{v}_n(\omega)|s_0] = {} & \frac{\sum\limits_{\omega\in\Omega}\int\limits_{s\in\mathcal{S}^1(n|F)} F(ds|\omega)\mu(\omega)}{\sum\limits_{\omega\in\Omega}\left(\int\limits_{s\in\mathcal{S}^1(n|F)} F(ds|\omega)\mu(\omega) + \int\limits_{s\in\mathcal{S}^2(n|F)} F(ds|\omega)\mu(\omega)\right)}\mathbb{E}_F[\mathbf{v}_n(\omega)|s \in \mathcal{S}^1(n|F)] \\
& + \frac{\sum\limits_{\omega\in\Omega}\int\limits_{s\in\mathcal{S}^2(n|F)} F(ds|\omega)\mu(\omega)}{\sum\limits_{\omega\in\Omega}\left(\int\limits_{s\in\mathcal{S}^1(n|F)} F(ds|\omega)\mu(\omega) + \int\limits_{s\in\mathcal{S}^2(n|F)} F(ds|\omega)\mu(\omega)\right)}\mathbb{E}_F[\mathbf{v}_n(\omega)|s \in \mathcal{S}^2(n|F)] \\
& \geq \frac{\sum\limits_{\omega\in\Omega}\int\limits_{s\in\mathcal{S}^1(n|F)} F(ds|\omega)\mu(\omega)}{\sum\limits_{\omega\in\Omega}\left(\int\limits_{s\in\mathcal{S}^1(n|F)} F(ds|\omega)\mu(\omega) + \int\limits_{s\in\mathcal{S}^2(n|F)} F(ds|\omega)\mu(\omega)\right)}\mathbb{E}_F[\mathbf{v}_\nu(\omega)|s \in \mathcal{S}^1(n|F)]
\end{aligned}$$

$$+\frac{\sum\limits_{\omega\in\Omega}\int\limits_{s\in\mathcal{S}^2(n|F)} F(ds|\omega)\mu(\omega)}{\sum\limits_{\omega\in\Omega}\left(\int\limits_{s\in\mathcal{S}^1(n|F)} F(ds|\omega)\mu(\omega)+\int\limits_{s\in\mathcal{S}^2(n|F)} F(ds|\omega)\mu(\omega)\right)}\mathbb{E}_F[\mathbf{v}_\nu(\omega)|s\in\mathcal{S}^2(n|F)]=\mathbb{E}_{\tilde{F}}[\mathbf{v}_\nu(\omega)|s_0].\blacksquare$$

Lemma 5. The cost of any information strategy F, which is a solution to (1) subject to (2) and produces behavior $\mathbb{P}$ based on equation (6), can be written:

$$\mathbf{C}(F,\,\mu)=\mathbf{C}(\mathbb{P},\,\mu)$$

$$\equiv\sum_{\omega\in\Omega}\mu(\omega)\sum_{n\in\mathcal{N}}\Big(-\lambda_1\Pr(n)\log(\Pr(n))-(\lambda_2-\lambda_1)\Pr(n|\mathcal{A}_1(\omega))\log(\Pr(n|\mathcal{A}_1(\omega)))$$

$$-(\lambda_3-\lambda_2)\Pr(n|\mathcal{A}_1(\omega)\cap\mathcal{A}_2(\omega))\log(\Pr(n|\mathcal{A}_1(\omega)\cap\mathcal{A}_2(\omega)))$$

$$-\ \ldots\ -(\lambda_M-\lambda_{M-1})\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))\log(\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega)))+\lambda_M\Pr(n|\omega)\log(\Pr(n|\omega))\Big),$$

and $\mathbf{C}(\mathbb{P},\,\mu)$, so defined, is convex in $\mathbb{P}$ that satisfy equations (10) and (11): if $\tilde{\mathbb{P}}$ (a $\tilde{\Pr}(n|\omega)$ for each option n and state ω) and $\hat{\mathbb{P}}$ (a $\hat{\Pr}(n|\omega)$ for each option n and state ω) both satisfy (10) and (11), then $\mathbb{P}$ defined by $\Pr(n|\omega)=\alpha\tilde{\Pr}(n|\omega)+(1-\alpha)\hat{\Pr}(n|\omega)$ for each option n and state ω satisfies (10) and (11) and $\mathbf{C}(\mathbb{P},\,\mu)\leq\alpha\mathbf{C}(\tilde{\mathbb{P}},\,\mu)+(1-\alpha)\mathbf{C}(\hat{\mathbb{P}},\,\mu)$.

Proof. Let $\mathcal{P}_s=(\mathcal{S}(1|F),\,\ldots,\,\mathcal{S}(N|F))$ denote a partition of the space of signals the agent may receive such that for each option n with $\Pr(n)>0$ each signal that results in the agent selecting option n is in $\mathcal{S}(n|F)$, and then as shown in Lemma 4, with probability one the s drawn from $\mathcal{S}(n|F)$ results in a particular posterior. Then (using the properties of $\mathcal{H}$):

$$\mathbf{C}(F,\,\mu)\equiv\mathbb{E}[\mathbb{H}(\mu)-\mathbb{H}(\mu(\cdot|s))]$$

$$=\mathbb{E}\Big[\lambda_1\Big(\mathcal{H}(\mathcal{A}_1,\,\mu)-\mathcal{H}(\mathcal{A}_1,\,\mu(\cdot|s))\Big)\tag{14}$$

$$+\ \ldots\ +\lambda_M\Big(\mathcal{H}(\mathcal{A}_M,\,\mu(\cdot|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega)))-\mathcal{H}(\mathcal{A}_M,\,\mu(\cdot|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega),\,s))\Big)\Big]$$

$$=\mathbb{E}\Big[\lambda_1\Big(\mathcal{H}(\mathcal{P}_s,\,F)-\mathcal{H}(\mathcal{P}_s,\,F(\cdot|\mathcal{A}_1(\omega)))\Big)\tag{15}$$

$$+\ \ldots\ +\lambda_M\Big(\mathcal{H}(\mathcal{P}_s,\,F(\cdot|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega)))-\mathcal{H}(\mathcal{P}_s,\,F(\cdot|\cap_{i=1}^{M}\mathcal{A}_i(\omega)))\Big)\Big]$$

$$=\mathbb{E}\Big[\lambda_1\mathcal{H}(\mathcal{P}_s,\,F)+(\lambda_2-\lambda_1)\mathcal{H}(\mathcal{P}_s,\,F(\cdot|\mathcal{A}_1(\omega)))$$

$$+ \ldots + (\lambda_M - \lambda_{M-1})\mathcal{H}(\mathcal{P}_s,\, F(\cdot| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega))) - \lambda_M \mathcal{H}(\mathcal{P}_s,\, F(\cdot| \cap_{i=1}^{M} \mathcal{A}_i(\omega)))\Big]$$

$$= \sum_{\omega\in\Omega} \mu(\omega) \sum_{n\in\mathcal{N}} \Big(- \lambda_1 \Pr(n) \log(\Pr(n)) - (\lambda_2 - \lambda_1)\Pr(n|\mathcal{A}_1(\omega)) \log(\Pr(n|\mathcal{A}_1(\omega)))$$

$$-(\lambda_3 - \lambda_2)\Pr(n|\mathcal{A}_1(\omega) \cap \mathcal{A}_2(\omega)) \log(\Pr(n|\mathcal{A}_1(\omega) \cap \mathcal{A}_2(\omega)))$$

$$- \ldots - (\lambda_M - \lambda_{M-1})\Pr(n| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega)) \log(\Pr(n| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega))) + \lambda_M \Pr(n|\omega) \log(\Pr(n|\omega))\Big).$$

The equality of (14) and (15) follows from the symmetry of mutual information, defined in Appendix 2. It is evident that if $\tilde{\mathbb{P}}$, $\hat{\mathbb{P}}$, and $\mathbb{P}$, are defined as in the statement of this lemma, then $\mathbb{P}$ satisfies (10) and (11). Convexity follows almost directly from Lemma 3 and what is shown above as any behavior $\mathbb{P}$ that satisfies (10) and (11) can be used to define an information strategy that satisfies (2)) by setting, for each state ω, $F(n,\, \omega) = \Pr(n|\omega)\mu(\omega)$ for each $n \in \mathcal{N}$ and $F(s,\, \omega) = 0$ otherwise. ■

Proof of Lemma 1. First, Lemma 5 implies the objective described by equation (9) is concave on the set of $\mathbb{P}$ that satisfy (10) and (11) as payoffs are linear and the cost of information is convex on the set of $\mathbb{P}$ that satisfy (10) and (11).

Given F that is a solution to (1) subject to (2), for each $n \in \mathcal{N}$, let s_n denote a signal in $\mathcal{S}(n|F)$ which results in the posterior generated by signals in $\mathcal{S}(n|F)$ with probability one (Lemma 4 shows this can be done). Then notice:

$$\sum_{\omega\in\Omega} \int_s V(s) F(ds|\omega)\mu(\omega) = \sum_{n\in\mathcal{N}} V(s_n) \int_{s\in\mathcal{S}_n} \sum_{\omega\in\Omega} F(ds|\omega)\mu(\omega)$$

$$= \sum_{n\in\mathcal{N}} V(s_n)\Pr(n) = \sum_{n\in\mathcal{N}} \sum_{\omega\in\Omega} \mathbf{v}_n(\omega) F(\omega|s_n) \Pr(n)$$

$$= \sum_{n\in\mathcal{N}} \sum_{\omega\in\Omega} \mathbf{v}_n(\omega) \Pr(n|\omega)\mu(\omega)$$

where the last step follows from the fact that $\Pr(X|Y)\Pr(Y) = \Pr(Y|X)\Pr(X)$. The rest of the proof proceeds with two proofs by contradiction. First, assume that F achieves expected utility U_1, and let $\mathbb{P}$ be the behavior induced by it. Assume that $\mathbb{P}$ is not a solution to (9) subject to (10) and (11), and thus there is a $\tilde{\mathbb{P}}$ which satisfies (10) and (11) and achieves expected utility $U_2 > U_1$. However, an information strategy $\tilde{F}$ and a choice of option for each signal can be created that generates $\tilde{\mathbb{P}}$. For instance, for each of the N distinct signals s_n, let the option selected by the agent after they observe s_n be denoted $\tilde{a}(\tilde{F}(\omega|s_n)) \equiv n$, and let $\tilde{F}(s_n,\, \omega) = \tilde{\Pr}(n|\omega)\mu(\omega)\ \ \forall\,\omega$ so that (2) is satisfied. This is impossible though as then $(\tilde{F},\, \tilde{a})$ achieves $U_2 > U_1$ and F cannot have been optimal.

Similarly, assume that $\mathbb{P}$ is a solution to (9) subject to (10) and (11), which achieves expected utility U_3, but is not induced by a solution to (1) subject to (2). That is, there is a $\tilde{F}$ which satisfies (2), produces a certain posterior with probability one after each option that is selected with a positive probability is selected (without loss given Lemma 4), and achieves $U_4 > U_3$. This means, however, that behavior defined for each option n and state ω by:

$$\tilde{\Pr}(n|\omega) = \int\limits_{s \in \mathcal{S}(n|\tilde{F})} \frac{\tilde{F}(s,\, \omega)}{\mu(\omega)},$$

also achieves U_4 by Lemma 4 and Lemma 5, which is impossible as $\mathbb{P}$ was supposedly optimal and $\tilde{\mathbb{P}}$ satisfies (10) and (11). ■

Proof of Theorem 1. The Lagrangian for the problem depicted in Lemma 1 can be written:

$$\mathcal{L} = \sum_{n\in\mathcal{N}} \sum_{\omega\in\Omega} \mathbf{v}_n(\omega)\Pr(n|\omega)\mu(\omega) - \mathbf{C}(\mathbb{P},\, \mu) + \sum_{n\in\mathcal{N}} \sum_{\omega\in\Omega} \xi_n(\omega)\Pr(n|\omega)\mu(\omega)$$

$$-\sum_{\omega\in\Omega} \gamma(\omega)\Big(\sum_{n\in\mathcal{N}} \Pr(n|\omega) - 1\Big)\mu(\omega).$$

$\xi_n(\omega) \geq 0$ are the multipliers for (10), and $\gamma(\omega)$ are the multipliers for (11).

The derivative of the Lagrangian with respect to $\Pr(n|\omega)$ is not well defined if $\Pr(n|\omega) = 0$, however, as $\log(0)$ is undefined, so it needs to be ensured that choice probabilities are non-zero to ensure differentiability. It can be shown, however, that if behavior is optimal and $\Pr(n) > 0$ then $\Pr(n|\omega)$ can be bound away from zero for all $\omega \in \Omega$. To show this, assume $\Pr(n) > 0$, for all ω it is the case that $\Pr(n|\omega) \geq \delta \geq 0$, and there exists $\hat{\omega}$ such that $\Pr(n|\hat{\omega}) = \delta$, and notice that there must be an option $m \in \mathcal{N}$ such that $\Pr(m|\hat{\omega}) \geq \frac{1}{|\mathcal{N}|}$. If δ is small enough it can be shown that the agent can do strictly better by increasing $\Pr(n|\hat{\omega})$ to strictly larger $\epsilon < \Pr(m|\hat{\omega})$, reduce $\Pr(m|\hat{\omega})$ by the same amount, keep all else equal, and denote the transformed behavior for each event B and option ν by $\tilde{\Pr}(\nu|B)$. If $\delta > 0$, let $\epsilon = 2\delta$. Optimality of the original behavior implies that the change in payoffs is weakly less than the change in learning costs:

$$\mu(\hat{\omega})(\epsilon - \delta)(\mathbf{v}_n(\hat{\omega}) - \mathbf{v}_m(\hat{\omega})) \leq$$

$$-\lambda_1\Big(\tilde{\Pr}(n)\log\tilde{\Pr}(n) - \Pr(n)\log\Pr(n) + \tilde{\Pr}(m)\log\tilde{\Pr}(m) - \Pr(m)\log\Pr(m)\Big)$$

$$-\sum_{\omega\in\mathcal{A}_1(\hat{\omega})}\mu(\omega)(\lambda_2-\lambda_1)\Big(\tilde{\Pr}(n|\mathcal{A}_1(\omega))\log\tilde{\Pr}(n|\mathcal{A}_1(\omega)) - \Pr(n|\mathcal{A}_1(\omega))\log\Pr(n|\mathcal{A}_1(\omega))$$

$$+\tilde{\Pr}(m|\mathcal{A}_1(\omega))\log\tilde{\Pr}(m|\mathcal{A}_1(\omega)) - \Pr(m|\mathcal{A}_1(\omega))\log\Pr(m|\mathcal{A}_1(\omega))\Big)$$

$$-\cdots-\sum_{\omega\in\cap_{i=1}^{M-1}\mathcal{A}_i(\hat{\omega})}\mu(\omega)(\lambda_M-\lambda_{M-1})\Big(\tilde{\Pr}(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))\log\tilde{\Pr}(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))$$

$$-\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))\log\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))$$

$$+\tilde{\Pr}(m|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))\log\tilde{\Pr}(m|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega)) - \Pr(m|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))\log\Pr(m|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))\Big)$$

$$+\lambda_M\mu(\hat{\omega})\Big(\epsilon\log\epsilon - \delta\log\delta + \tilde{\Pr}(m|\hat{\omega})\log\tilde{\Pr}(m|\hat{\omega}) - \Pr(m|\hat{\omega})\log\Pr(m|\hat{\omega})\Big),$$

but if both sides are divided by $(\epsilon-\delta)$ a contradiction is created as the left hand side of the inequality is finite while the right hand side is close to negative infinity when ϵ is close to zero (remember that the convention used with Shannon Entropy is that $0\log 0 = 0$). So, if optimal behavior features $\Pr(n) > 0$, it is necessary that for all $\omega\in\Omega$ that $\Pr(n|\omega) > 0$.

Given some candidate solution, $\mathbb{P}$, consider a transformed problem where for each $n\in\mathcal{N}$ if in the candidate solution $\Pr(n) = 0$ then it is now required that $\Pr(n) = 0$, while if $\Pr(n) > 0$ then remember that it has been shown above that it is necessary that $\Pr(n|\omega) > 0$ for all $\omega\in\Omega$ and then for some arbitrarily small $\delta > 0$ impose that it now must be that $\Pr(n|\omega)\geq\delta$ for all $\omega\in\Omega$, so that now for each such n the multipliers $\xi_n(\omega)$ correspond to the constraint $-\Pr(n|\omega)+\delta\leq 0$, so that in this transformed problem the first order conditions are necessary (Lange, 2013). If $\Pr(n) > 0$ in the candidate solution, and thus $\Pr(n|\omega) > \delta$ for each $\omega\in\Omega$, then the first order condition with respect to $\Pr(n|\omega)$ implies:

$$\mathbf{v}_n(\omega) + \lambda_1(1+\log\Pr(n)) + (\lambda_2-\lambda_1)(1+\log\Pr(n|\mathcal{A}_1(\omega)))$$

$$+\ldots+(\lambda_M-\lambda_{M-1})(1+\log\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))) - \lambda_M(1+\log\Pr(n|\omega)) = \gamma(\omega)-\xi_n(\omega).$$

Thus, since $\xi_n(\omega) = 0$, the first order condition implies:

$$\Pr(n|\omega) = \Pr(n)^{\frac{\lambda_1}{\lambda_M}}\Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ldots\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}e^{\frac{-\gamma(\omega)}{\lambda_M}} \tag{16}$$

Plugging (16) into (11), one can solve for $\gamma(\omega)$. Plugging $\gamma(\omega)$ back into (16) achieves the desired result. ■

The behavior described in Theorem 1 has many intuitive features. It is also a quite natural extension of the analogous result from Matějka and McKay (2015) for the Shannon RI model, which is described in equation (12).

The formula in (12) also has many intuitive features. If $\lambda_2 = \lambda$ grows (shrinks), which represents an increase (decrease) in the difficulty of learning, the value of each option in the realized state becomes less (more) significant for the determination of the selected option, and the significance of the agent's prior increases (decreases). If λ_2 approaches infinity, the realized values become insignificant, and the behavior of the agent approaches the behavior of the agent in the case where learning is impossible: they choose their option based on their prior. If λ_2 approaches zero the unconditional priors become insignificant, and the behavior of the agent approaches the behavior of the agent in the case where learning is costless: they choose the option with the highest realized value.

If it is instead assumed that the cost of information is measured with MASE and the agent may also learn about another attribute $\mathcal{A}_1$ with a lower associated multiplier λ_1, then if $\mathcal{A}_1 \neq \Omega$, and the agent has optimal behavior then in state $\omega \in \Omega$ they select option n from their set of options $\mathcal{N}$ with probability:

$$\Pr(n|\omega) = \frac{\Pr(n)^{\frac{\lambda_1}{\lambda_2}} \Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_2}} e^{\frac{\mathbf{v}_n(\omega)}{\lambda_2}}}{\sum\limits_{\nu\in\mathcal{N}} \Pr(\nu)^{\frac{\lambda_1}{\lambda_2}} \Pr(\nu|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_2}} e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_2}}}. \tag{17}$$

With MASE, as the formula in (17) indicates, the probability of the agent selecting an option n in a particular state of the world ω depends not only on the unconditional probabilities of the options being selected and the realized values of the options, but also on the realized value of $\mathcal{A}_1$. When option n is in general desirable in $\mathcal{A}_1(\omega)$ relative to the other options, then $\Pr(n|\mathcal{A}_1(\omega))$ is larger, and there may be a high probability of n being selected, even if $\Pr(n)$ is not that large, and $\mathbf{v}_n(\omega)$ is not that high.

The formula in (17) also has many intuitive features. It maintains the intuitive comparative statics for λ_2 that the formula in (12) have, and also features intuitive properties for $\Pr(n|\mathcal{A}_1(\omega))$ and λ_1.

If observing $\mathcal{A}_1(\omega)$ is completely uninformative about the value of the options, then it is optimal for the agent to select $\Pr(n|\mathcal{A}_1(\omega)) = \Pr(n)$ since $\mathbb{H}$ is strictly concave in μ. In this case

$\Pr(n)^{\frac{\lambda_1}{\lambda_2}}\Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_2}} = \Pr(n)$, and behavior is identical to that in (12). If the cheaper to learn about attribute is irrelevant it is thus ignored, and behavior collapses back to the environment described in Matějka and McKay (2015), as should be desired.

If λ_1 approaches λ_2 (the cheaper to learn about attribute becomes close to as expensive as the more expensive to learn about attribute) then behavior approaches that described in (12) since $\Pr(n)^{\frac{\lambda_1}{\lambda_2}}\Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_2}} \to \Pr(n)$. Thus, if an insignificantly cheaper to learn about attribute is introduced, behavior is changed in an insignificant fashion (see Figure 1). Again, this seems like a desirable property.

If λ_1 approaches zero then the role of the unconditional prior dissipates, and the exponent on $\Pr(n|\mathcal{A}_1(\omega))$ approaches one, meaning it replaces the unconditional prior from (12). This makes sense because if λ_1 goes to zero it means $\mathcal{A}_1(\omega)$ can essentially be viewed for free, in which case behavior within each $\mathcal{A}_1(\omega)$ should resemble that in the setting where there is only one attribute with multiplier λ_2 and a prior of $\mu(\cdot|\mathcal{A}_1(\omega))$.

New attributes can be added with new multipliers and the description of behavior in Theorem 1 maintains the intuitive properties described in the paragraphs above. RI with MASE is thus a quite natural extension of RI with Shannon Entropy. You can click here to return to Theorem 1.

If B is any collection of partitions, let $\sigma(B)$ denote the **σ-algebra generated by** B, which is the smallest σ-algebra containing all the events in each of the partitions in B. To help make the notation more compact, a group of partitions can be used to **generate** a finer partition: if $(\mathcal{P}_1, \dots, \mathcal{P}_m)$ is a group of partitions, let $\times\{\mathcal{P}_i\}_{i=1}^n$ denote the partition such that $\sigma(\times\{\mathcal{P}_i\}_{i=1}^n) = \sigma(\mathcal{P}_1, \dots, \mathcal{P}_n)$.

Proof of Lemma 2. In this proof it is said behavior $\mathbb{P}$ satisfies Theorem 1 if: for each $n \in \mathcal{N}$ if $\Pr(n) > 0$ then for all $\omega \in \Omega$ it is the case that $\Pr(n|\omega) > 0$ and $\Pr(n|\omega)$ is described by equation (13). Given behavior $\mathbb{P}$ that satisfies Theorem 1, plug equation (13) into the last instance of $\Pr(n|\omega)$ in equation (9) after using Lemma 5, and notice that cancelling like terms then produces the new objective:

$$\sum_{n\in\mathcal{N}}\sum_{\omega\in\Omega}\Bigg(\mathbf{v}_n(\omega)\Pr(n|\omega) + \lambda_1\Pr(n)\log(\Pr(n)) + (\lambda_2-\lambda_1)\Pr(n|\mathcal{A}_1(\omega))\log(\Pr(n|\mathcal{A}_1(\omega)))$$

$$+\cdots+(\lambda_M-\lambda_{M-1})\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))\log(\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))) - \lambda_M\Pr(n|\omega)\log(\Pr(n|\omega))\Bigg)\mu(\omega)$$

$$
\begin{aligned}
&= \sum_{n\in\mathcal{N}}\sum_{\omega\in\Omega}\Bigg(\mathbf{v}_n(\omega)\Pr(n|\omega) + \lambda_1\Pr(n)\log(\Pr(n)) + (\lambda_2-\lambda_1)\Pr(n|\mathcal{A}_1(\omega))\log(\Pr(n|\mathcal{A}_1(\omega)))\\
&\qquad + \ldots + (\lambda_M-\lambda_{M-1})\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))\log(\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega)))\\
&\qquad -\lambda_M\Pr(n|\omega)\log\left(\Pr(n)^{\frac{\lambda_1}{\lambda_M}}\right) - \lambda_M\Pr(n|\omega)\log\left(\Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\right)\\
&\qquad - \ldots - \lambda_M\Pr(n|\omega)\log\left(\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}\right) - \lambda_M\Pr(n|\omega)\log\left(e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_M}}\right)\\
&\qquad +\lambda_M\Pr(n|\omega)\log\left(\sum_{\nu\in\mathcal{N}}\Pr(\nu)^{\frac{\lambda_1}{\lambda_M}}\Pr(\nu|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ldots\Pr(\nu|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_M}}\right)\Bigg)\mu(\omega)\\
&= \sum_{n\in\mathcal{N}}\sum_{\omega\in\Omega}\left(\lambda_M\Pr(n|\omega)\log\left(\sum_{\nu\in\mathcal{N}}\Pr(\nu)^{\frac{\lambda_1}{\lambda_M}}\Pr(\nu|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ldots\Pr(\nu|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_M}}\right)\right)\mu(\omega)\\
&= \sum_{\omega\in\Omega}\lambda_M\log\left(\sum_{\nu\in\mathcal{N}}\Pr(\nu)^{\frac{\lambda_1}{\lambda_M}}\Pr(\nu|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ldots\Pr(\nu|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_M}}\right)\mu(\omega).
\end{aligned}
$$

Call (9) the 'old objective' and call the objective in Lemma 2 produced immediately above the 'new objective.' What is shown above is that the old objective and the new objective have the same value if behavior $\mathbb{P}$ satisfies Theorem 1, and thus the maximal value of the new objective subject to the associated constraints is at least as large as the maximal value of the old objective subject to the associated constraints.

Next, note that the new objective is concave. This is true because the function:

$$
f(x_1,\ldots,x_M) = x_1^{\alpha_1}\cdot\ldots\cdot x_M^{\alpha_M} : \mathbb{R}_+^M \to \mathbb{R},
$$

a generalized Cobb-Douglas utility function, is known to be concave for positive x_i's and α_i's if $\alpha_1 + \ldots + \alpha_M \leq 1$. So, if for each $A \in \mathcal{F}$ and each $n \in \mathcal{N}$ it is true that $\Pr(n|A) = \alpha\hat{\Pr}(n|A) + (1-\alpha)\tilde{\Pr}(n|A)$ with $\alpha \in (0,1)$ and $\hat{\Pr}(n|A), \tilde{\Pr}(n|A) \geq 0$, then for each $\omega \in \Omega$ and $n \in \mathcal{N}$ (using the concavity of logarithms):

$$
\begin{aligned}
&\Pr(n)^{\frac{\lambda_1}{\lambda_M}}\Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ldots\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}\\
&\geq \alpha\hat{\Pr}(n)^{\frac{\lambda_1}{\lambda_M}}\hat{\Pr}(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ldots\hat{\Pr}(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}\\
&+(1-\alpha)\tilde{\Pr}(n)^{\frac{\lambda_1}{\lambda_M}}\tilde{\Pr}(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ldots\tilde{\Pr}(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}
\end{aligned}
$$

$$
\begin{aligned}
&\Rightarrow \log\left(\sum_{n=1}^{N}\Pr(n)^{\frac{\lambda_1}{\lambda_M}}\Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ldots\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}\right)\\
&\geq \log\left(\sum_{n=1}^{N}\left(\alpha\hat{\Pr}(n)^{\frac{\lambda_1}{\lambda_M}}\hat{\Pr}(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ldots\hat{\Pr}(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}\right.\right.\\
&\quad\left.\left.+(1-\alpha)\tilde{\Pr}(n)^{\frac{\lambda_1}{\lambda_M}}\tilde{\Pr}(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ldots\tilde{\Pr}(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}\right)e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}\right)\\
&\geq \alpha\log\left(\sum_{n=1}^{N}\hat{\Pr}(n)^{\frac{\lambda_1}{\lambda_M}}\hat{\Pr}(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ldots\hat{\Pr}(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}\right)\\
&\quad+(1-\alpha)\log\left(\sum_{n=1}^{N}\tilde{\Pr}(n)^{\frac{\lambda_1}{\lambda_M}}\tilde{\Pr}(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ldots\tilde{\Pr}(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}\right),
\end{aligned}
$$

and multiplying by $\lambda_M\mu(\omega)$ and summing over the $\omega\in\Omega$ shows the objective is concave.

Let $\mathcal{M}\subseteq\mathcal{N}$ denote a non-empty subset of options. If δ is set equal to zero, the Lagrangian for the problem described in Lemma 2 when the agent may only select options $n\in\mathcal{M}$ with a positive probability is:

$$
\begin{aligned}
\mathcal{L}=&\sum_{\omega\in\Omega}\left(\lambda_M\log\left(\sum_{n\in\mathcal{M}}\Pr(n)^{\frac{\lambda_1}{\lambda_M}}\Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ldots\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}\right)\mu(\omega)\right)\\
&+\sum_{A\in\times\{\mathcal{A}_i\}_{i=1}^{M-1}}\sum_{n\in\mathcal{M}}\xi_n(A)(\Pr(n|A)-\delta)-\sum_{A\in\times\{\mathcal{A}_i\}_{i=1}^{M-1}}\gamma(A)\left(\sum_{n\in\mathcal{M}}\Pr(n|A)-1\right),
\end{aligned}
$$

but the constraint of $\Pr(n|A)\geq\delta$ for arbitrarily small weakly positive δ is considered as the objective of this problem is concave and thus if $\delta>0$ then the first order conditions are both necessary and sufficient (Train, 2009) as the objective is then differentiable on the relevant set (the meaning of $\times\{\mathcal{A}_i\}_{i=1}^{M-1}$ is explained just before this proof). Assume $\mathbb{P}$ satisfies (13) and is such that $\Pr(n|\omega)>0$ for all options $n\in\mathcal{M}$ and states $\omega\in\Omega$, then the first order condition with respect to $\Pr(n|A)$ for any $A\in\times\{\mathcal{A}_i\}_{i=1}^{M-1}$ and $\tilde{\omega}\in A$ is then:

$$
\begin{aligned}
&\left(\sum_{\omega\in\Omega}\frac{\lambda_1\mu(\cap_{i=1}^{M-1}\mathcal{A}_i(\tilde{\omega}))}{\Pr(n)}\Pr(n|\omega)\mu(\omega)\right)+\left(\sum_{\omega\in\mathcal{A}_1(\tilde{\omega})}\frac{(\lambda_2-\lambda_1)\mu(\cap_{i=1}^{M-1}\mathcal{A}_i(\tilde{\omega}))}{\Pr(n|\mathcal{A}_1(\omega))\mu(\mathcal{A}_1(\tilde{\omega}))}\Pr(n|\omega)\mu(\omega)\right)\\
&+\cdots+\left(\sum_{\omega\in\cap_{i=1}^{M-1}\mathcal{A}_i(\tilde{\omega})}\frac{(\lambda_M-\lambda_{M-1})\mu(\cap_{i=1}^{M-1}\mathcal{A}_i(\tilde{\omega}))}{\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))\mu(\cap_{i=1}^{M-1}\mathcal{A}_i(\tilde{\omega}))}\Pr(n|\omega)\mu(\omega)\right)+\xi_n(A)=\gamma(A)
\end{aligned}
$$

but, $\xi_n(A) = 0$ if δ is small enough, and:

$$\sum_{\omega \in \Omega} \frac{\lambda_1 \mu(\cap_{i=1}^{M-1} \mathcal{A}_i(\tilde{\omega}))}{\Pr(n)} \Pr(n|\omega)\mu(\omega) = \lambda_1 \mu(A),$$

and for for each $m \in \{1, \ldots, M-1\}$:

$$\sum_{\omega \in \cap_{i=1}^{m} \mathcal{A}_i(\tilde{\omega})} \frac{(\lambda_{m+1} - \lambda_m)\mu(\cap_{i=1}^{M-1} \mathcal{A}_i(\tilde{\omega}))}{\Pr(n| \cap_{i=1}^{m} \mathcal{A}_i(\omega))\mu(\cap_{i=1}^{m} \mathcal{A}_i(\tilde{\omega}))} \Pr(n|\omega)\mu(\omega) = (\lambda_{m+1} - \lambda_m)\mu(A),$$

so if:

$$\gamma(A) = \mu(A)\Big(\lambda_1 + \lambda_2 - \lambda_1 + \cdots + \lambda_M - \lambda_{M-1}\Big) = \lambda_M \mu(A), \quad \forall A \in \times\{\mathcal{A}_i\}_{i=1}^{M-1},$$

then the first order conditions are all satisfied, and since δ can be chosen to be arbitrarily small and strictly positive, the first order conditions are both necessary and sufficient (Train, 2009) and $\mathbb{P}$ solves the problem described in Lemma 2 if the agent is further constraint so they can only select options from $\mathcal{M}$ with a positive probability as the objective is continuous.

What remains to be shown is that a solution to the problem described in Lemma 2 combined with (13) maximizes (9) subject to (10) and (11), and thus the solution to the problem described in Lemma 2 satisfies Theorem 1 when combined with (13).

Let x denote the maximal value of the old objective subject to the associated constraints. Suppose a maximizer of the new objective subject to the associated constraints assigns positive probabilities of selection to a subset of options $\mathcal{M} \subseteq \mathcal{N}$, namely a maximizer of the new objective features $\Pr(m) > 0$ iff $m \in \mathcal{M}$, and produces value y. Notice that $y \geq x$ given what is shown above. If the maximization of the old objective is then revisited (subject to the associated constraints) and it is further imposed that $\Pr(n) = 0$ if $n \notin \mathcal{M}$ and $\Pr(m) \geq \epsilon$ if $m \in \mathcal{M}$ for some arbitrarily small $\epsilon > 0$, then the solution to this problem produces a payoff for the agent of $z \leq x$ as more constraints have been imposed. As is shown in the proof of Theorem 1, as long as $\Pr(m) > 0$, it is optimal to have $\Pr(m|\omega) > 0$ for all $\omega \in \Omega$, so for an arbitrarily smaller $\delta > 0$ it can be further imposed that $\Pr(m|\omega) \geq \delta$ for all $\omega \in \Omega$ and $m \in \mathcal{M}$ when trying to maximize the old objective, and none of these new constraints bind. The first order condition for the Lagrangian associated with maximizing the old objective with respect to $\Pr(m|\omega)$ for $m \in \mathcal{M}$ is implied by our work in the proof of Theorem 1 to then be:

$$\mathbf{v}_m(\omega) + \lambda_1(1 + \log \Pr(m)) + (\lambda_2 - \lambda_1)(1 + \log \Pr(m|\mathcal{A}_1(\omega)))$$

$$+ \ldots + (\lambda_M - \lambda_{M-1})(1 + \log \Pr(m| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega))) - \lambda_M(1 + \log \Pr(m|\omega)) = \gamma(\omega) - \xi_m(\omega) - \beta_m,$$

with $\xi_m(\omega) = 0$ and β_m being the multiplier on the constraint that $\Pr(m) \geq \epsilon$. But then, given the solution, whatever the values of the multipliers $\beta_m \geq 0$ on the constraint:

$$-\Big(\sum_{\omega \in \Omega} \Pr(m|\omega)\mu(\omega)\Big) + \epsilon \leq 0,$$

are for each $m \in \mathcal{M}$, transform the payoff of each such m in each state ω to instead be $\tilde{\mathbf{v}}_m(\omega) = \mathbf{v}_m(\omega) + \beta_m$, remove the constraint that $\Pr(m) \geq \epsilon$, and, based on the work in the proof of Theorem 1, behavior then satisfies Theorem 1 when the transformed payoffs are used. This behavior then maximizes the new objective when the associated constraints are imposed, it is further imposed that $\Pr(n) = 0$ if $n \notin \mathcal{M}$, and the transformed payoffs are used, based on what is shown above, and thus the transformed maximized value of the new objective is some $q \geq y$. Thus, $q \geq y \geq x \geq z$. If $y > x$, since ϵ is arbitrarily small when the agent is solving the problem that produced q, it is thus the case that for some m that $\Pr(m) = \epsilon$ and β_m is arbitrarily large, but then there is a contradiction as the agent could do better by only selecting from some of the options with unconditional probability of being selected ϵ and arbitrarily high values, so it must be that $y = x$. What remains to be shown is that a maximizer of the new objective subject to the associated constraints, denoted $\Pr(n|A)$ for each option n and event $A \in \mathcal{F}$, plugged into equation (13) is self consistent, namely for each option n and event $A \in \mathcal{F}$:

$$\sum_{\omega \in A} \frac{\Pr(n)^{\frac{\lambda_1}{\lambda_M}} \Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2 - \lambda_1}{\lambda_M}} \ldots \Pr(n| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega))^{\frac{\lambda_M - \lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}}{\sum_{\nu \in \mathcal{N}} \Pr(\nu)^{\frac{\lambda_1}{\lambda_M}} \Pr(\nu|\mathcal{A}_1(\omega))^{\frac{\lambda_2 - \lambda_1}{\lambda_M}} \ldots \Pr(\nu| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega))^{\frac{\lambda_M - \lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_M}}} \mu(\omega|A) = \Pr(n|A).$$

This is not hard to do. If a maximizer of the new objective subject to the associated constraints is denoted $\tilde{\Pr}(n|A)$ for each option n and event $A \in \mathcal{F}$, since the strict concavity of logarithms and the concavity of the generalized Cobb-Douglas utility function implies the value of the objective in each state of the world is the same under the constrained maximizers of both the new and old objectives, and the denominator in (13) is uniquely determined by the value of the objective in each state of the world, and the numerator would have to be linear in mixtures of the two maximizers, and it is not then hard to show that for each option n there is a constant $\theta_n \geq 0$ such that $\tilde{\Pr}(n|A) = \theta_n \Pr(n|A)$ for each event $A \in \times\{\mathcal{A}_i\}_{i=1}^{M-1}$. ■

Proof of Theorem 2. A fixed effect interpretation of MASE follows easily from the optimal choice probabilities described in Theorem 1:

$$\Pr(n|\omega) = \frac{\Pr(n)^{\frac{\lambda_1}{\lambda_M}} \Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots \Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}}{\sum_{\nu\in\mathcal{N}} \Pr(\nu)^{\frac{\lambda_1}{\lambda_M}} \Pr(\nu|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots \Pr(\nu|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{N-1}}{\lambda_M}} e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_M}}}$$

$$= \frac{(N\Pr(n))^{\frac{\lambda_1}{\lambda_M}} (N\Pr(n|\mathcal{A}_1(\omega)))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots (N\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega)))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}}{\sum_{\nu\in\mathcal{N}} (N\Pr(\nu))^{\frac{\lambda_1}{\lambda_M}} (N\Pr(\nu|\mathcal{A}_1(\omega)))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots (N\Pr(\nu|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega)))^{\frac{\lambda_M-\lambda_{N-1}}{\lambda_M}} e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_M}}}$$

$$= \frac{e^{\frac{\mathbf{v}_n(\omega)+\lambda_1\alpha_n^0+(\lambda_2-\lambda_1)\alpha_n^1+\cdots+(\lambda_M-\lambda_{M-1})\alpha_n^{M-1}}{\lambda_M}}}{\sum_{\nu\in\mathcal{N}} e^{\frac{\mathbf{v}_\nu(\omega)+\lambda_1\alpha_\nu^0+(\lambda_2-\lambda_1)\alpha_\nu^1+\cdots+(\lambda_M-\lambda_{M-1})\alpha_\nu^{M-1}}{\lambda_M}}}$$

Where $\alpha_\nu^0 = \log(N\Pr(\nu))$, and for $m \in \{1, \dots, M-1\}$ define $\alpha_\nu^m = \log(N\Pr(\nu|\cap_{i=1}^m \mathcal{A}_i(\omega)))$. Normalizing the value of the options by λ_M, namely letting $\tilde{v}_n = \frac{\mathbf{v}_n(\omega)}{\lambda_M}$, and defining α_n appropriately, agent choice behavior described by RI with MASE can then be interpreted as a RU model where each option n has perceived value:

$$u_n = \tilde{v}_n + \tfrac{\lambda_1}{\lambda_M}\alpha_n^0 + \tfrac{\lambda_2-\lambda_1}{\lambda_M}\alpha_n^1 + \cdots + \tfrac{\lambda_M-\lambda_{M-1}}{\lambda_M}\alpha_n^{M-1} + \epsilon_n = \tilde{v}_n + \alpha_n + \epsilon_n$$

Such an RU model where ϵ_n is distributed iid according to a Gumbel distribution is consistent with the optimal choice probabilities described in Theorem 1 (Train, 2009). ■

Behavior that is consistent with Theorem 1 is not necessarily optimal because in many settings it is not optimal for the agent to consider all of the available options (choose them with positive probability), and though such a corner solution may be optimal, there are many corners that are consistent with Theorem 1 but are not optimal.

Given behavior $\mathbb{P}$, define the **consideration set** to be $\mathcal{C}(\mathbb{P}) \equiv \{n \in \mathcal{N}|\Pr(n) > 0\}$. An option n is said to be **considered** if $\Pr(n) > 0$. This definition of a consideration set has the advantage that it can be observed in the data and fits with the definition given by Caplin et al. (2018).

Suppose behavior $\mathbb{P}$ is consistent with Theorem 1 and is thus a candidate for optimal behavior.

To determine if it is in fact optimal for an option $n \in \mathcal{N}$ that is not consider under $\mathbb{P}$ to not be considered, n needs to be compared to a representative value of the options that are being considered under $\mathbb{P}$ and given a score in each state of the world. The agent would do better with n in the consideration set if it scores well enough across all states of the world, in which case $\mathbb{P}$ is not optimal even though it is consistent with Theorem 1.

Define the **score** of option n in state ω to be:

$$s_n(\omega|\mathbb{P}) = \frac{e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}}{\sum_{\nu \in \mathcal{N}} \Pr(\nu)^{\frac{\lambda_1}{\lambda_M}} (\Pr(\nu|\mathcal{A}_1(\omega)))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots (\Pr(\nu| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega)))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_M}}}.$$

Theorem 4. Behavior $\mathbb{P}$ is a solution to (9) subject to (10) and (11) iff for all $n \in \mathcal{C}(\mathbb{P})$ it is the case that $\Pr(n|\omega) > 0$ and $\Pr(n|\omega)$ is described by equation (13) for each state $\omega \in \Omega$, and for all $n \notin \mathcal{C}(\mathbb{P})$ it is the case that:

$$\mathbb{E}\Bigg[\mathbb{E}\Bigg[\dots \mathbb{E}\Bigg[\mathbb{E}\Bigg[s_n(\omega|\mathbb{P})| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega)\Bigg]^{\frac{\lambda_M}{\lambda_{M-1}}} | \cap_{i=1}^{M-2} \mathcal{A}_i(\omega)\Bigg]^{\frac{\lambda_{M-1}}{\lambda_{M-2}}} \dots |A_1(\omega)\Bigg]^{\frac{\lambda_2}{\lambda_1}}\Bigg] \leq 1.$$

Proof. Assume $\mathbb{P}$ is such that for all $n \in \mathcal{C}(\mathbb{P})$ and $\omega \in \Omega$ it is the case that $\Pr(n|\omega) > 0$ and $\Pr(n|\omega)$ is described by equation (13). Further, so that there remains something to be proven, assume there is at least one $n \notin \mathcal{C}(\mathbb{P})$. To figure out if the agent can do strictly better than $\mathbb{P}$, given Lemma 2, it needs to be determined if the agent could do strictly better by changing their behavior so that at least one of the options $n \notin \mathcal{C}(\mathbb{P})$ is chosen instead with a strictly positive probability.

First, revisit the problem of maximizing (9) subject to (10) and (11), except transform the problem into a problem where the solution to the transformed problem $\tilde{\mathbb{P}}$ is required to be such that $\tilde{\Pr}(n) = \epsilon$ for each $n \in \mathcal{N} \backslash \mathcal{C}(\mathbb{P})$ and $\tilde{\Pr}(m) \geq \epsilon$ for each $m \in \mathcal{C}(\mathbb{P})$, where ϵ is some arbitrarily small and strictly positive constant, and for all $n \in \mathcal{N}$ and $\omega \in \Omega$ it is imposed that $\tilde{\Pr}(n|\omega) \geq \delta$ where δ is some strictly positive constant that is arbitrarily smaller than ϵ. The objective for this transformed problem is differentiable on the set over which maximization is occurring, which is convex and closed, so a solution to the first order conditions exists, and the first order conditions are necessary (Lange, 2013). The Lagrangian for the transformed problem is:

$$\mathcal{L} = \sum_{n \in \mathcal{N}} \sum_{\omega \in \Omega} \mathbf{v}_n(\omega) \tilde{\Pr}(n|\omega) \mu(\omega) - \mathbf{C}(\tilde{\mathbb{P}}, \mu) + \sum_{n \in \mathcal{N}} \sum_{\omega \in \Omega} \xi_n(\omega)(\tilde{\Pr}(n|\omega) - \delta)\mu(\omega)$$

$$-\sum_{\omega\in\Omega}\gamma(\omega)\Big(\sum_{n\in\mathcal{N}}\tilde{\Pr}(n|\omega)-1\Big)\mu(\omega)-\sum_{n\in\mathcal{N}\backslash\mathcal{C}(\mathbb{P})}\beta_n\Big(\Big(\sum_{\omega\in\Omega}\tilde{\Pr}(n|\omega)\mu(\omega)\Big)-\epsilon\Big)$$

$$-\sum_{m\in\mathcal{C}(\mathbb{P})}\theta_m\Big(\Big(\sum_{\omega\in\Omega}-\tilde{\Pr}(m|\omega)\mu(\omega)\Big)+\epsilon\Big),$$

where $\xi_n(\omega)\geq 0$ are the multipliers for $-\tilde{\Pr}(n|\omega)+\delta\leq 0$ (remember that based on what is shown in the proof of Theorem 1 these constraints do not bind as long as δ is small enough relative to ϵ since if there is $\hat{\omega}$ and n such that $\tilde{\Pr}(n|\hat{\omega})=\delta$ then there is an option $m\in\mathcal{C}(\mathbb{P})$ such that $\tilde{\Pr}(m)>\epsilon$ and the agent would do better by increasing $\tilde{\Pr}(n|\hat{\omega})$, decreasing $\tilde{\Pr}(m|\hat{\omega})$ by the same amount, and adjusting $\tilde{\Pr}(n|\omega)$ and $\tilde{\Pr}(m|\omega)$ in some other state where $\tilde{\Pr}(n|\omega)$ and $\tilde{\Pr}(m|\omega)$ are arbitrarily higher than δ so that each of the other constraints is satisfied, and thus each $\xi_n(\omega)=0$), $\gamma(\omega)$ are the multipliers for (11), each β_n for each $n\in\mathcal{N}\backslash\mathcal{C}(\mathbb{P})$ is the multiplier for the constraint that $\sum_\omega\tilde{\Pr}(n|\omega)\mu(\omega)=\epsilon$, and each $\theta_m\geq 0$ for each $m\in\mathcal{C}(\mathbb{P})$ is the multiplier for the constraint that $(\sum_\omega-\tilde{\Pr}(m|\omega)\mu(\omega))+\epsilon\leq 0$. The important insight is that once a solution $\tilde{\mathbb{P}}$ to the transformed problem is found, a new problem can be considered where for each $\omega\in\Omega$ and $n\in\mathcal{N}\backslash\mathcal{C}(\mathbb{P})$ the value is instead altered to be $\tilde{\mathbf{v}}_n(\omega)=\mathbf{v}_n(\omega)-\beta_n$ and for each $\omega\in\Omega$ and $m\in\mathcal{C}(\mathbb{P})$ the value is instead altered to be $\tilde{\mathbf{v}}_m(\omega)=\mathbf{v}_m(\omega)+\theta_m$, and drop the constraints that $\sum_\omega\tilde{\Pr}(n|\omega)\mu(\omega)=\epsilon$ for $n\in\mathcal{N}\backslash\mathcal{C}(\mathbb{P})$ and $\sum_\omega\tilde{\Pr}(m|\omega)\mu(\omega)\geq\epsilon$ for $m\in\mathcal{C}(\mathbb{P})$. Given the work in the proof of Theorem 1, $\tilde{\mathbb{P}}$ satisfies (13) for all options n and states ω when the altered payoffs are used, and thus Lemma 2 implies that $\tilde{\mathbb{P}}$ is a solution to the problem of maximizing (9) subject to (10) and (11) when payoffs are altered in this way and thus maximizes the problem from Lemma 2 when payoffs are altered in this way.

Next, it is shown that by picking arbitrarily small ϵ, the value of the objective from Lemma 2 when behavior is $\mathbb{P}$ and the original payoffs are used, call it x, can be made arbitrarily close to the value of the objective from Lemma 2 when behavior is $\tilde{\mathbb{P}}$ and the altered payoffs are used, call it y. It is evident that $y\geq x$ for each such ϵ, so it needs to ruled out that there is $c>0$ such that $y>x+c$ for all arbitrarily small ϵ. This is also evident as $\tilde{\mathbb{P}}$ is a solution to (9) subject to (10) and (11) when the altered payoffs are used, and if the agent chose the same signal structure as with $\tilde{\mathbb{P}}$, except whenever they get a signal that should lead them to choose one of the options n with $\tilde{\Pr}(n)=\epsilon$ they instead randomized with equal probabilities over the $m\in\mathcal{C}(\mathbb{P})$ with $\tilde{\Pr}(m)>\epsilon$, they would then get some payoff of $z\leq x$. Thus, given $c>0$ it cannot be that $y>x+c$ for an arbitrarily small ϵ, as then some options that are being selected with unconditional probability ϵ have altered payoffs that are arbitrarily high, and thus the agent could do better by picking from a subset of these options and $\tilde{\mathbb{P}}$ is not maximal.

Next, it is shown that by picking arbitrarily small ϵ, for each $\omega \in \Omega$:

$$\sum_{n=1}^{N} \Pr(n)^{\frac{\lambda_1}{\lambda_M}} \Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots \Pr(n|\cap_{i=1}^{M-1} \mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}$$

and

$$\sum_{n=1}^{N} \tilde{\Pr}(n)^{\frac{\lambda_1}{\lambda_M}} \tilde{\Pr}(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots \tilde{\Pr}(n|\cap_{i=1}^{M-1} \mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\tilde{\mathbf{v}}_n(\omega)}{\lambda_M}}$$

can be made arbitrarily close. If not, for some $\hat{\omega} \in \Omega$ there is arbitrarily small ϵ such that the former is larger than the latter by some $\rho > 0$, then consider behavior $\hat{\mathcal{P}}$ such that $\hat{\Pr}(n|\omega) = \frac{1}{2}\tilde{\Pr}(n|\omega) + \frac{1}{2}\Pr(n|\omega)$ for each $n \in \mathcal{N}$ and $\omega \in \Omega$. Strict concavity of logs and concavity of the generalized Cobb-Douglas utility function (see the proof of Lemma 2) then implies that there is $c > 0$ such that:

$$\log\left(\sum_{n=1}^{N} \hat{\Pr}(n)^{\frac{\lambda_1}{\lambda_M}} \hat{\Pr}(n|\mathcal{A}_1(\hat{\omega}))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots \hat{\Pr}(n|\cap_{i=1}^{M-1} \mathcal{A}_i(\hat{\omega}))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\tilde{\mathbf{v}}_n(\hat{\omega})}{\lambda_M}}\right)\mu(\hat{\omega})$$

$$\geq \frac{1}{2}\log\left(\sum_{n=1}^{N} \tilde{\Pr}(n)^{\frac{\lambda_1}{\lambda_M}} \tilde{\Pr}(n|\mathcal{A}_1(\hat{\omega}))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots \tilde{\Pr}(n|\cap_{i=1}^{M-1} \mathcal{A}_i(\hat{\omega}))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\tilde{\mathbf{v}}_n(\hat{\omega})}{\lambda_M}}\right)\mu(\hat{\omega})$$

$$+\frac{1}{2}\log\left(\sum_{n=1}^{N} \Pr(n)^{\frac{\lambda_1}{\lambda_M}} \Pr(n|\mathcal{A}_1(\hat{\omega}))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots \Pr(n|\cap_{i=1}^{M-1} \mathcal{A}_i(\hat{\omega}))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\tilde{\mathbf{v}}_n(\hat{\omega})}{\lambda_M}}\right)\mu(\hat{\omega}) + c,$$

since for each $x > 0$ it is true that $\log(x + \frac{1}{2}\rho) - \frac{1}{2}\log(x) - \frac{1}{2}\log(x + \rho)$ is increasing in ρ for $\rho > 0$. A contradiction has thus been created as concavity indicates that:

$$\sum_{\omega\in\Omega}\left(\log\left(\sum_{n=1}^{N} \hat{\Pr}(n)^{\frac{\lambda_1}{\lambda_M}} \hat{\Pr}(n|\mathcal{A}_1(\hat{\omega}))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots \hat{\Pr}(n|\cap_{i=1}^{M-1} \mathcal{A}_i(\hat{\omega}))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\tilde{\mathbf{v}}_n(\hat{\omega})}{\lambda_M}}\right)\mu(\hat{\omega})\right)$$

$$\geq \sum_{\omega\in\Omega}\left(\frac{1}{2}\log\left(\sum_{n=1}^{N} \tilde{\Pr}(n)^{\frac{\lambda_1}{\lambda_M}} \tilde{\Pr}(n|\mathcal{A}_1(\hat{\omega}))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots \tilde{\Pr}(n|\cap_{i=1}^{M-1} \mathcal{A}_i(\hat{\omega}))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\tilde{\mathbf{v}}_n(\hat{\omega})}{\lambda_M}}\right)\mu(\hat{\omega})\right.$$

$$\left.+\frac{1}{2}\log\left(\sum_{n=1}^{N} \Pr(n)^{\frac{\lambda_1}{\lambda_M}} \Pr(n|\mathcal{A}_1(\hat{\omega}))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots \Pr(n|\cap_{i=1}^{M-1} \mathcal{A}_i(\hat{\omega}))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\tilde{\mathbf{v}}_n(\hat{\omega})}{\lambda_M}}\right)\mu(\hat{\omega})\right) + c,$$

and it cannot then be that $\tilde{\mathbb{P}}$ is optimal.

Remember that $s_n(\omega|\mathbb{P})$ is defined just before the statement of Theorem 4, and use the altered payoffs to analogously define $\tilde{s}_n(\omega|\mathbb{P})$ for each state $\omega \in \Omega$ and option $n \in \mathcal{N}$ as follows:

$$\tilde{s}_n(\omega|\mathbb{P}) = \frac{e^{\frac{\tilde{\mathbf{v}}_n(\omega)}{\lambda_M}}}{\sum\limits_{\nu \in \mathcal{N}} \Pr(\nu)^{\frac{\lambda_1}{\lambda_M}} (\Pr(\nu|\mathcal{A}_1(\omega)))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots (\Pr(\nu| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega)))^{\frac{\lambda_M - \lambda_{M-1}}{\lambda_M}} e^{\frac{\tilde{\mathbf{v}}_\nu(\omega)}{\lambda_M}}}.$$

Since $\tilde{\mathbb{P}}$ satisfies (13) for all options n and states ω when the altered payoffs are used, for each $n \in \mathcal{N} \backslash \mathcal{C}(\mathbb{P})$ and $\tilde{\omega} \in \Omega$:

$$\tilde{\Pr}(n| \cap_{i=1}^{M-1} \mathcal{A}_i(\tilde{\omega})) = \sum_{\omega \in \cap_{i=1}^{M-1} \mathcal{A}_i(\tilde{\omega})} \tilde{\Pr}(n|\omega) \mu(\omega| \cap_{i=1}^{M-1} \mathcal{A}_i(\tilde{\omega}))$$

$$\Rightarrow \tilde{\Pr}(n| \cap_{i=1}^{M-1} \mathcal{A}_i(\tilde{\omega}))$$

$$= \tilde{\Pr}(n)^{\frac{\lambda_1}{\lambda_{M-1}}} \Pr(n|\mathcal{A}_1(\tilde{\omega}))^{\frac{\lambda_2-\lambda_1}{\lambda_{M-1}}} \dots \Pr(n| \cap_{i=1}^{M-2} \mathcal{A}_i(\hat{\omega}))^{\frac{\lambda_{M-1}-\lambda_{M-2}}{\lambda_{M-1}}} \mathbb{E}\Big[\tilde{s}_n(\omega|\mathbb{P})| \cap_{i=1}^{M-1} \mathcal{A}_i(\tilde{\omega})\Big]^{\frac{\lambda_M}{\lambda_{M-1}}},$$

$$\tilde{\Pr}(n| \cap_{i=1}^{M-2} \mathcal{A}_i(\tilde{\omega})) = \sum_{\omega \in \cap_{i=1}^{M-2} \mathcal{A}_i(\tilde{\omega})} \tilde{\Pr}(n| \cap_{i=1}^{M-1} \mathcal{A}_i(\tilde{\omega})) \mu(\cap_{i=1}^{M-1} \mathcal{A}_i(\tilde{\omega})| \cap_{i=1}^{M-2} \mathcal{A}_i(\tilde{\omega}))$$

$$\Rightarrow \tilde{\Pr}(n| \cap_{i=1}^{M-2} \mathcal{A}_i(\tilde{\omega})) = \tilde{\Pr}(n)^{\frac{\lambda_1}{\lambda_{M-2}}} \Pr(n|\mathcal{A}_1(\tilde{\omega}))^{\frac{\lambda_2-\lambda_1}{\lambda_{M-2}}} \dots \Pr(n| \cap_{i=1}^{M-3} \mathcal{A}_i(\hat{\omega}))^{\frac{\lambda_{M-2}-\lambda_{M-3}}{\lambda_{M-2}}}$$

$$\cdot \mathbb{E}\Big[\mathbb{E}\Big[\tilde{s}_n(\omega|\tilde{\mathbb{P}})| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega)\Big]^{\frac{\lambda_M}{\lambda_{M-1}}} | \cap_{i=1}^{M-2} \mathcal{A}_i(\tilde{\omega})\Big]^{\frac{\lambda_{M-1}}{\lambda_{M-2}}},$$

$$\dots$$

$$\tilde{\Pr}(n|\mathcal{A}_1(\tilde{\omega})) = \sum_{\omega \in \cap_{i=1}^{2} \mathcal{A}_i(\tilde{\omega})} \tilde{\Pr}(n| \cap_{i=1}^{2} \mathcal{A}_i(\omega)) \mu(\cap_{i=1}^{2} \mathcal{A}_i(\tilde{\omega})|\mathcal{A}_1(\tilde{\omega}))$$

$$\tilde{\Pr}(n|\mathcal{A}_1(\tilde{\omega})) = \tilde{\Pr}(n)^{\frac{\lambda_1}{\lambda_1}}$$

$$\cdot \mathbb{E}\Big[\dots \mathbb{E}\Big[\mathbb{E}\Big[\tilde{s}_n(\omega|\tilde{\mathbb{P}})| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega)\Big]^{\frac{\lambda_M}{\lambda_{M-1}}} | \cap_{i=1}^{M-2} \mathcal{A}_i(\tilde{\omega})\Big]^{\frac{\lambda_{M-1}}{\lambda_{M-2}}} \dots |A_1(\omega)\Big]^{\frac{\lambda_2}{\lambda_1}},$$

$$\Rightarrow 1 = \mathbb{E}\Big[\mathbb{E}\Big[\dots \mathbb{E}\Big[\mathbb{E}\Big[\tilde{s}_n(\omega|\tilde{\mathbb{P}})| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega)\Big]^{\frac{\lambda_M}{\lambda_{M-1}}} | \cap_{i=1}^{M-2} \mathcal{A}_i(\omega)\Big]^{\frac{\lambda_{M-1}}{\lambda_{M-2}}} \dots |A_1(\omega)\Big]^{\frac{\lambda_2}{\lambda_1}}\Big],$$

and since it has been shown that the denominator of $s_n(\omega|\mathbb{P})$ is arbitrarily close to the denominator of $\tilde{s}_n(\omega|\tilde{\mathbb{P}})$ for each ω:

$$\mathbb{E}\Big[\mathbb{E}\Big[\dots \mathbb{E}\Big[\mathbb{E}\Big[s_n(\omega|\mathbb{P})| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega)\Big]^{\frac{\lambda_M}{\lambda_{M-1}}} | \cap_{i=1}^{M-2} \mathcal{A}_i(\omega)\Big]^{\frac{\lambda_{M-1}}{\lambda_{M-2}}} \dots |A_1(\omega)\Big]^{\frac{\lambda_2}{\lambda_1}}\Big] e^{\frac{-\beta_n}{\lambda_1}} \approx 1.$$

This means:

$$\mathbb{E}\Big[\mathbb{E}\Big[\dots\mathbb{E}\Big[\mathbb{E}\Big[s_n(\omega|\mathbb{P})|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega)\Big]^{\frac{\lambda_M}{\lambda_{M-1}}}|\cap_{i=1}^{M-2}\mathcal{A}_i(\omega)\Big]^{\frac{\lambda_{M-1}}{\lambda_{M-2}}}\dots|A_1(\omega)\Big]^{\frac{\lambda_2}{\lambda_1}}\Big] > 1$$

for an option $n \in \mathcal{N}\backslash\mathcal{C}(\mathbb{P})$ iff behavior $\mathbb{P}$ is not optimal as the agent could do better by including such an n in the consideration set, while:

$$\mathbb{E}\Big[\mathbb{E}\Big[\dots\mathbb{E}\Big[\mathbb{E}\Big[s_n(\omega|\mathbb{P})|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega)\Big]^{\frac{\lambda_M}{\lambda_{M-1}}}|\cap_{i=1}^{M-2}\mathcal{A}_i(\omega)\Big]^{\frac{\lambda_{M-1}}{\lambda_{M-2}}}\dots|A_1(\omega)\Big]^{\frac{\lambda_2}{\lambda_1}}\Big] \leq 1$$

for all options $n \in \mathcal{N}\backslash\mathcal{C}(\mathbb{P})$ iff the behavior $\mathbb{P}$ is optimal as the consideration set is then optimal. This is true because if the consideration set is not optimal then the agent can do strictly better by including some option n in the consideration set and they could still do strictly better by including that same option n in the consideration set even if its payoffs in each state were made slightly lower. ■

Proof of Theorem 3. The proof begins by showing that if for each pair of states ω_i and ω_j, with $\omega_i \neq \omega_j$, one of the five conditions is satisfied, then this can be identified with the known set of optimal behavior and the payoff functions for the different options, and $\lambda(\omega_i, \omega_j)$ is identified. In this proof it is assumed that two states are the same iff they have the same subscript.

If condition **(i)** is satisfied, so $\mathbf{v}_n(\omega_i)-\mathbf{v}_m(\omega_i) > 0$ and $\mathbf{v}_m(\omega_j)-\mathbf{v}_n(\omega_j) > 0$, then there exists μ with $\mu(\omega_i)+\mu(\omega_j) = 1$ such that any $\mathbb{P}^*(\{n, m\}, \mu)$ features a positive probability of both n and m being selected by the agent. This is true because for any $c > 0$ (and in particular $c = \lambda(\omega_i, \omega_j)$) there is a μ with $\mu(\omega_i) + \mu(\omega_j) = 1$ such that:

$$\sum_{\omega\in\{\omega_i,\,\omega_j\}} \frac{e^{\frac{\mathbf{v}_n(\omega)}{c}}}{e^{\frac{\mathbf{v}_m(\omega)}{c}}}\mu(\omega) > 1 \text{ and } \sum_{\omega\in\{\omega_i,\,\omega_j\}} \frac{e^{\frac{\mathbf{v}_m(\omega)}{c}}}{e^{\frac{\mathbf{v}_n(\omega)}{c}}}\mu(\omega) > 1,$$

as this is true when

$$\frac{1-\dfrac{e^{\frac{\mathbf{v}_n(\omega_j)}{c}}}{e^{\frac{\mathbf{v}_m(\omega_j)}{c}}}}{\dfrac{e^{\frac{\mathbf{v}_n(\omega_i)}{c}}}{e^{\frac{\mathbf{v}_m(\omega_i)}{c}}}-\dfrac{e^{\frac{\mathbf{v}_n(\omega_j)}{c}}}{e^{\frac{\mathbf{v}_m(\omega_j)}{c}}}} < \mu(\omega_i) < \frac{\dfrac{e^{\frac{\mathbf{v}_m(\omega_j)}{c}}}{e^{\frac{\mathbf{v}_n(\omega_j)}{c}}}-1}{\dfrac{e^{\frac{\mathbf{v}_m(\omega_j)}{c}}}{e^{\frac{\mathbf{v}_n(\omega_j)}{c}}}-\dfrac{e^{\frac{\mathbf{v}_m(\omega_i)}{c}}}{e^{\frac{\mathbf{v}_n(\omega_i)}{c}}}},$$

and it is not hard to show

$$0 < \frac{1 - \dfrac{e^{\frac{\mathbf{v}_n(\omega_j)}{c}}}{e^{\frac{\mathbf{v}_m(\omega_j)}{c}}}}{\dfrac{e^{\frac{\mathbf{v}_n(\omega_i)}{c}}}{e^{\frac{\mathbf{v}_m(\omega_i)}{c}}} - \dfrac{e^{\frac{\mathbf{v}_n(\omega_j)}{c}}}{e^{\frac{\mathbf{v}_m(\omega_j)}{c}}}} < \frac{\dfrac{e^{\frac{\mathbf{v}_m(\omega_j)}{c}}}{e^{\frac{\mathbf{v}_n(\omega_j)}{c}}} - 1}{\dfrac{e^{\frac{\mathbf{v}_m(\omega_j)}{c}}}{e^{\frac{\mathbf{v}_n(\omega_j)}{c}}} - \dfrac{e^{\frac{\mathbf{v}_m(\omega_i)}{c}}}{e^{\frac{\mathbf{v}_n(\omega_i)}{c}}}} < 1,$$

thus, Theorem 4 indicates that both options are selected with a positive probability when such a μ is the prior, and therefore Theorem 1 indicates that $\lambda(\omega_i, \omega_j)$ solves:

$$\Pr(n|\omega_i) = \frac{\Pr(n) e^{\frac{\mathbf{v}_n(\omega_i)}{\lambda(\omega_i,\, \omega_j)}}}{\sum\limits_{\nu \in \{n,\, m\}} \Pr(\nu) e^{\frac{\mathbf{v}_\nu(\omega_i)}{\lambda(\omega_i,\, \omega_j)}}} = \frac{1}{1 + \dfrac{\Pr(m)}{\Pr(n)} e^{\frac{\mathbf{v}_m(\omega_i) - \mathbf{v}_n(\omega_i)}{\lambda(\omega_i,\, \omega_j)}}},$$

which clearly has a unique solution that some simple algebra produces a closed-form solution for.

If condition **(ii)** is satisfied, so $\mathbf{v}_n(\omega_i) - \mathbf{v}_m(\omega_i) > 0$, $\mathbf{v}_n(\omega_i) - \mathbf{v}_m(\omega_i) \neq \mathbf{v}_n(\omega_j) - \mathbf{v}_m(\omega_j) > 0$, and $\mathbf{v}_m(\omega_k) - \mathbf{v}_n(\omega_k) > 0$, then, based on what is shown in the previous paragraph, there is a prior that only assigns positive probabilities to ω_i and ω_k such that optimal behavior features a positive probability of both n and m being selected by the agent, and such behavior uniquely identifies $\lambda(\omega_i, \omega_k)$. Similarly, $\lambda(\omega_j, \omega_k)$ is uniquely identified by an almost identical logic. Then, since attributes are partitions of the state space, if $\lambda(\omega_i, \omega_k) \neq \lambda(\omega_j, \omega_k)$ then $\lambda(\omega_i, \omega_j) = \min(\lambda(\omega_i, \omega_k), \lambda(\omega_j, \omega_k))$ (and thus $\lambda(\omega_i, \omega_j)$ is identified), while if $\lambda(\omega_i, \omega_k) = \lambda(\omega_j, \omega_k)$ then $\lambda(\omega_i, \omega_j) \geq \lambda(\omega_i, \omega_k)$, but more work needs to be done. If $\lambda(\omega_i, \omega_j) \geq \lambda(\omega_i, \omega_k)$ then, based on what is shown in the previous paragraph, there exists μ with $\mu(\omega_i) + \mu(\omega_k) = 1$ such that:

$$\sum_{\omega \in \{\omega_i, \omega_k\}} \frac{e^{\frac{\mathbf{v}_n(\omega)}{\lambda(\omega_i,\, \omega_j)}}}{e^{\frac{\mathbf{v}_m(\omega)}{\lambda(\omega_i,\, \omega_j)}}} \mu(\omega) > 1 \text{ and } \sum_{\omega \in \{\omega_i, \omega_k\}} \frac{e^{\frac{\mathbf{v}_m(\omega)}{\lambda(\omega_i,\, \omega_j)}}}{e^{\frac{\mathbf{v}_n(\omega)}{\lambda(\omega_i,\, \omega_j)}}} \mu(\omega) > 1.$$

Thus, if $\lambda(\omega_i, \omega_j) \geq \lambda(\omega_i, \omega_k)$, for small enough $\epsilon > 0$, it must be that if $\tilde{\mu}$ is defined so that $\tilde{\mu}(\omega_k) = \mu(\omega_k)$, $\tilde{\mu}(\omega_i) = \mu(\omega_i) - \epsilon$, and $\tilde{\mu}(\omega_j) = \epsilon$, then:

$$\left(\frac{e^{\frac{\mathbf{v}_n(\omega_k)}{\lambda(\omega_i,\, \omega_j)}}}{e^{\frac{\mathbf{v}_m(\omega_k)}{\lambda(\omega_i,\, \omega_j)}}}\right)^{\frac{\lambda(\omega_i,\, \omega_j)}{\lambda(\omega_i,\, \omega_k)}} \mu(\omega_k) + \left(\frac{e^{\frac{\mathbf{v}_n(\omega_i)}{\lambda(\omega_i,\, \omega_j)}}}{e^{\frac{\mathbf{v}_m(\omega_i)}{\lambda(\omega_i,\, \omega_j)}}} \frac{\tilde{\mu}(\omega_i)}{\mu(\omega_i)} + \frac{e^{\frac{\mathbf{v}_n(\omega_j)}{\lambda(\omega_i,\, \omega_j)}}}{e^{\frac{\mathbf{v}_m(\omega_j)}{\lambda(\omega_i,\, \omega_j)}}} \frac{\tilde{\mu}(\omega_j)}{\mu(\omega_i)}\right)^{\frac{\lambda(\omega_i,\, \omega_j)}{\lambda(\omega_i,\, \omega_k)}} \mu(\omega_i) > 1,$$

$$\left(\frac{e^{\frac{\mathbf{v}_m(\omega_k)}{\lambda(\omega_i,\,\omega_j)}}}{e^{\frac{\mathbf{v}_n(\omega_k)}{\lambda(\omega_i,\,\omega_j)}}}\right)^{\frac{\lambda(\omega_i,\,\omega_j)}{\lambda(\omega_i,\,\omega_k)}}\mu(\omega_k)+\left(\frac{e^{\frac{\mathbf{v}_m(\omega_i)}{\lambda(\omega_i,\,\omega_j)}}}{e^{\frac{\mathbf{v}_n(\omega_i)}{\lambda(\omega_i,\,\omega_j)}}}\frac{\tilde{\mu}(\omega_i)}{\mu(\omega_i)}+\frac{e^{\frac{\mathbf{v}_m(\omega_j)}{\lambda(\omega_i,\,\omega_j)}}}{e^{\frac{\mathbf{v}_n(\omega_j)}{\lambda(\omega_i,\,\omega_j)}}}\frac{\tilde{\mu}(\omega_j)}{\mu(\omega_i)}\right)^{\frac{\lambda(\omega_i,\,\omega_j)}{\lambda(\omega_i,\,\omega_k)}}\mu(\omega_i)>1,$$

by Jensen's inequality, and Theorem 4 thus implies any $\mathbb{P}^*(\{n,\, m\},\, \tilde{\mu})$ features both options being selected with a positive probability, and therefore Theorem 1 and some algebra indicates that $\Pr(n|\omega_i)$ and $\Pr(n|\omega_j)$ from $\mathbb{P}^*(\{n,\, m\},\, \tilde{\mu})$ are such that $\lambda(\omega_i,\, \omega_j)$ solves:

$$\left(\frac{1}{\Pr(n|\omega_i)}-1\right)\frac{e^{\frac{\mathbf{v}_n(\omega_i)}{\lambda(\omega_i,\,\omega_j)}}}{e^{\frac{\mathbf{v}_m(\omega_i)}{\lambda(\omega_i,\,\omega_j)}}}\frac{e^{\frac{\mathbf{v}_m(\omega_j)}{\lambda(\omega_i,\,\omega_j)}}}{e^{\frac{\mathbf{v}_n(\omega_j)}{\lambda(\omega_i,\,\omega_j)}}}=\left(\frac{1}{\Pr(n|\omega_i)}-1\right)\left(\frac{e^{\frac{\mathbf{v}_n(\omega_i)-\mathbf{v}_m(\omega_i)}{1}}}{e^{\frac{\mathbf{v}_n(\omega_j)-\mathbf{v}_m(\omega_j)}{1}}}\right)^{\frac{1}{\lambda(\omega_i,\,\omega_j)}}=\frac{1}{\Pr(n|\omega_j)}-1,$$

which clearly has a unique solution that some simple algebra produces a closed-form solution for.

If condition **(iii)** is satisfied, so $\mathbf{v}_n(\omega_i)-\mathbf{v}_m(\omega_i)>\mathbf{v}_n(\omega_j)-\mathbf{v}_m(\omega_j)=0<\mathbf{v}_m(\omega_k)-\mathbf{v}_n(\omega_k)$ and $\lambda(\omega_i,\,\omega_j)\neq\lambda(\omega_j,\,\omega_k)$, then, based on what is shown in the previous paragraphs, there is belief μ such that $\mathbb{P}(\{n,\, m\},\, \mu)$ features a positive probability of both n and m being selected by Theorem 4 because for any $c>0$ there is such a μ with $\mu(\omega_i)+\mu(\omega_k)=1$ and $\mu(\omega_i)\in(0,\,1)$ such that:

$$\sum_{\omega\in\{\omega_i,\,\omega_k\}}\frac{e^{\frac{\mathbf{v}_n(\omega)}{c}}}{e^{\frac{\mathbf{v}_m(\omega)}{c}}}\mu(\omega)>1\text{ and }\sum_{\omega\in\{\omega_i,\,\omega_k\}}\frac{e^{\frac{\mathbf{v}_m(\omega)}{c}}}{e^{\frac{\mathbf{v}_n(\omega)}{c}}}\mu(\omega)>1,$$

so $\lambda(\omega_i,\,\omega_k)$ is identified using the logic from condition **(i)**. Further, if the prior is $\tilde{\mu}$ such that $\tilde{\mu}(\omega_j)=2\epsilon$, $\tilde{\mu}(\omega_i)=\mu(\omega_i)-\epsilon$, and $\tilde{\mu}(\omega_k)=\mu(\omega_k)-\epsilon$, for arbitrarily small $\epsilon>0$, then Jensen's inequality implies that for any non-trivial partition $\mathcal{P}$ of $\{\omega_i,\,\omega_j,\,\omega_k\}$, comprised of events denoted A_t, that for $d\in(0,\,c]$:

$$\sum_{A_t\in\mathcal{P}}\left(\sum_{\omega\in A_t}\frac{e^{\frac{\mathbf{v}_n(\omega)}{c}}}{e^{\frac{\mathbf{v}_m(\omega)}{c}}}\tilde{\mu}(\omega|A_t)\right)^{\frac{c}{d}}\tilde{\mu}(A_t)>1$$

$$\text{and }\sum_{A_t\in\mathcal{P}}\left(\sum_{\omega\in A_t}\frac{e^{\frac{\mathbf{v}_m(\omega)}{c}}}{e^{\frac{\mathbf{v}_n(\omega)}{c}}}\tilde{\mu}(\omega|A_t)\right)^{\frac{c}{d}}\tilde{\mu}(A_t)>1,$$

so, letting $c=\max(\lambda(\omega_i,\,\omega_j),\,\lambda(\omega_i,\,\omega_k),\,\lambda(\omega_j,\,\omega_k))$ and $d=\min(\lambda(\omega_i,\,\omega_j),\,\lambda(\omega_i,\,\omega_k),\,\lambda(\omega_j,\,\omega_k))$ (noticing that $\lambda(\omega_i,\,\omega_j)$, $\lambda(\omega_i,\,\omega_k)$, and $\lambda(\omega_j,\,\omega_k)$, can feature at most two unique values due to the nature of partitions, more on this below), Theorem 4 indicates that $\mathbb{P}(\{n,\, m\},\, \tilde{\mu})$ features a positive probability of both n and m being selected, and thus Theorem 1 indicates that each of these options is selected with a positive probability in each of the three states that occur with a positive probability. For the remainder of the consideration of condition **(iii)** assume that n and m are the only available options and the prior is the $\tilde{\mu}$ that is constructed immediately above. Notice that, since attributes are partitions of the state space, only one of three cases is possible, either $\lambda(\omega_i, \omega_j) = \lambda(\omega_j, \omega_k)$ and then $\lambda(\omega_i, \omega_j) = \lambda(\omega_j, \omega_k) \leq \lambda(\omega_i, \omega_k)$, or $\lambda(\omega_i, \omega_j) > \lambda(\omega_j, \omega_k)$ and then $\lambda(\omega_i, \omega_j) > \lambda(\omega_j, \omega_k) = \lambda(\omega_i, \omega_k)$, or $\lambda(\omega_i, \omega_j) < \lambda(\omega_j, \omega_k)$ then $\lambda(\omega_j, \omega_k) > \lambda(\omega_i, \omega_j) = \lambda(\omega_i, \omega_k)$, so regardless of which of the three cases is realized, at most two attributes (non-trivial partitions) are required to model learning when the prior is restricted to ω_i, ω_j, and ω_k, call them $\mathcal{A}_1$ and $\mathcal{A}_2$ with associated multipliers λ_1 and λ_2 ($\lambda_2 \geq \lambda_1$, and $\lambda_2 = \lambda_1$ and $\mathcal{A}_2 = \mathcal{A}_1$ iff only one attributes is required since $\lambda(\omega_i, \omega_j) = \lambda(\omega_j, \omega_k) = \lambda(\omega_i, \omega_k)$). Notice that which of these three cases is realized can be inferred from optimal behavior. If $\lambda(\omega_i, \omega_j) = \lambda(\omega_j, \omega_k) \leq \lambda(\omega_i, \omega_k)$, so $\mathcal{A}_1(\omega_j) = \{\omega_j\}$, then Theorem 1 implies:

$$\Pr(n|\omega_j) = \frac{\Pr(n)^{\frac{\lambda_1}{\lambda_2}} \Pr(n|\mathcal{A}_1(\omega_j))^{\frac{\lambda_2-\lambda_1}{\lambda_2}} e^{\frac{\mathbf{v}_n(\omega_j)}{\lambda_2}}}{\sum\limits_{\nu \in \{n,m\}} \Pr(\nu)^{\frac{\lambda_1}{\lambda_2}} \Pr(\nu|\mathcal{A}_1(\omega_j))^{\frac{\lambda_2-\lambda_1}{\lambda_2}} e^{\frac{\mathbf{v}_\nu(\omega_j)}{\lambda_2}}}$$

$$= \frac{\Pr(n)^{\frac{\lambda_1}{\lambda_2}} \Pr(n|\omega_j)^{\frac{\lambda_2-\lambda_1}{\lambda_2}}}{\sum\limits_{\nu \in \{n,m\}} \Pr(\nu)^{\frac{\lambda_1}{\lambda_2}} \Pr(\nu|\omega_j)^{\frac{\lambda_2-\lambda_1}{\lambda_2}}}$$

$$\iff \Pr(n)^{\frac{\lambda_1}{\lambda_2}} \Pr(n|\omega_j)^{\frac{\lambda_2-\lambda_1}{\lambda_2}} + \Pr(m)^{\frac{\lambda_1}{\lambda_2}} \Pr(m|\omega_j)^{\frac{\lambda_2-\lambda_1}{\lambda_2}} = \left(\frac{\Pr(n)}{\Pr(n|\omega_j)}\right)^{\frac{\lambda_1}{\lambda_2}}$$

$$\iff \Pr(n|\omega_j)\left(\frac{\Pr(n)}{\Pr(n|\omega_j)}\right)^{\frac{\lambda_1}{\lambda_2}} + \Pr(m|\omega_j)\left(\frac{\Pr(m)}{\Pr(m|\omega_j)}\right)^{\frac{\lambda_1}{\lambda_2}} = \left(\frac{\Pr(n)}{\Pr(n|\omega_j)}\right)^{\frac{\lambda_1}{\lambda_2}}$$

$$\iff \left(\frac{\Pr(n)}{\Pr(n|\omega_j)}\right)^{\frac{\lambda_1}{\lambda_2}} + \Pr(m|\omega_j)\left(\left(\frac{\Pr(m)}{\Pr(m|\omega_j)}\right)^{\frac{\lambda_1}{\lambda_2}} - \left(\frac{\Pr(n)}{\Pr(n|\omega_j)}\right)^{\frac{\lambda_1}{\lambda_2}}\right) = \left(\frac{\Pr(n)}{\Pr(n|\omega_j)}\right)^{\frac{\lambda_1}{\lambda_2}}$$

$$\iff \frac{\Pr(m)}{\Pr(m|\omega_j)} = \frac{\Pr(n)}{\Pr(n|\omega_j)},$$

and since $\Pr(n|\omega_j) = \Pr(n)$ and $\Pr(m|\omega_j) = \Pr(m)$ satisfy that last equality, and $\Pr(n|\omega_j) + \Pr(m|\omega_j) = 1$ and $\Pr(n) + \Pr(m) = 1$, the only solution is $\Pr(n|\omega_j) = \Pr(n)$ and $\Pr(m|\omega_j) = \Pr(m)$. If, instead, $\lambda(\omega_i, \omega_j) > \lambda(\omega_j, \omega_k) = \lambda(\omega_i, \omega_k)$, so $\mathcal{A}_1(\omega_j) = \{\omega_i, \omega_j\}$, then Theorem 1 implies:

$$\Pr(n|\omega_j) = \frac{\Pr(n)^{\frac{\lambda_1}{\lambda_2}} \Pr(n|\mathcal{A}_1(\omega_j))^{\frac{\lambda_2-\lambda_1}{\lambda_2}}}{\sum\limits_{\nu \in \{n,m\}} \Pr(\nu)^{\frac{\lambda_1}{\lambda_2}} \Pr(\nu|\mathcal{A}_1(\omega_j))^{\frac{\lambda_2-\lambda_1}{\lambda_2}}},$$

and $\Pr(n|\mathcal{A}_1(\omega_j)) > \Pr(n)$ since $\Pr(m|\mathcal{A}_1(\omega_k)) = \Pr(m|\omega_k) > \Pr(m)$ (the last inequality is not hard to show with Theorem 1), so $\Pr(n|\omega_j) > \Pr(n)$. Finally, if $\lambda(\omega_j, \omega_k) > \lambda(\omega_i, \omega_j) = \lambda(\omega_i, \omega_k)$, so $\mathcal{A}_1(\omega_j) = \{\omega_j, \omega_k\}$, then Theorem 1 similarly implies $\Pr(n|\mathcal{A}_1(\omega_j)) < \Pr(n)$ since $\Pr(n|\mathcal{A}_1(\omega_i)) = \Pr(n|\omega_i) > \Pr(n)$ (the last inequality is not hard to show with Theorem 1), so $\Pr(n|\omega_j) < \Pr(n)$. Thus, if $\Pr(n|\omega_j) = \Pr(n)$ then $\lambda(\omega_i, \omega_j) = \lambda(\omega_j, \omega_k) \leq \lambda(\omega_i, \omega_k)$, if $\Pr(n|\omega_j) > \Pr(n)$ then $\lambda(\omega_i, \omega_j) > \lambda(\omega_j, \omega_k) = \lambda(\omega_i, \omega_k)$, and if $\Pr(n|\omega_j) < \Pr(n)$ then $\lambda(\omega_j, \omega_k) > \lambda(\omega_i, \omega_j) = \lambda(\omega_i, \omega_k)$. Whether or not condition **(iii)** is satisfied can thus be inferred from the set of optimal behavior, and if it is satisfied then $\Pr(n|\omega_j) \neq \Pr(n)$, and, thus, there are two cases to deal with: $\Pr(n|\omega_j) > \Pr(n)$ and $\Pr(n|\omega_j) < \Pr(n)$. If $\Pr(n|\omega_j) > \Pr(n)$, then $\lambda(\omega_i, \omega_j) > \lambda(\omega_j, \omega_k) = \lambda(\omega_i, \omega_k)$ and $\mathcal{A}_1(\omega_j) = \{\omega_i, \omega_j\}$, remember that $\lambda(\omega_i, \omega_k)$ is known, and Theorem 1 implies $\lambda(\omega_i, \omega_j)$ solves:

$$\Pr(n|\omega_j) = \frac{\Pr(n)^{\frac{\lambda(\omega_i,\omega_k)}{\lambda(\omega_i,\omega_j)}} \Pr(n|\mathcal{A}_1(\omega_j))^{\frac{\lambda(\omega_i,\omega_j)-\lambda(\omega_i,\omega_k)}{\lambda(\omega_i,\omega_j)}} e^{\frac{\mathbf{v}_n(\omega_j)}{\lambda(\omega_i,\omega_j)}}}{\sum\limits_{\nu\in\{n,m\}} \Pr(\nu)^{\frac{\lambda(\omega_i,\omega_k)}{\lambda(\omega_i,\omega_j)}} \Pr(\nu|\mathcal{A}_1(\omega_j))^{\frac{\lambda(\omega_i,\omega_j)-\lambda(\omega_i,\omega_k)}{\lambda(\omega_i,\omega_j)}} e^{\frac{\mathbf{v}_\nu(\omega_j)}{\lambda(\omega_i,\omega_j)}}}$$

$$= \frac{1}{1 + \left(\frac{\Pr(m)}{\Pr(n)}\right)^{\frac{\lambda(\omega_i,\omega_k)}{\lambda(\omega_i,\omega_j)}} \left(\frac{\Pr(m|\mathcal{A}_1(\omega_j))}{\Pr(n|\mathcal{A}_1(\omega_j))}\right)^{\frac{\lambda(\omega_i,\omega_j)-\lambda(\omega_i,\omega_k)}{\lambda(\omega_i,\omega_j)}}},$$

$$= \frac{1}{1 + \frac{\Pr(m|\mathcal{A}_1(\omega_j))}{\Pr(n|\mathcal{A}_1(\omega_j))} \left(\frac{\Pr(m)}{\Pr(m|\mathcal{A}_1(\omega_j))} \frac{\Pr(n|\mathcal{A}_1(\omega_j))}{\Pr(n)}\right)^{\frac{\lambda(\omega_i,\omega_k)}{\lambda(\omega_i,\omega_j)}}},$$

which clearly has a unique solution since $\Pr(n|\mathcal{A}_1(\omega_j)) > \Pr(n)$ and $\Pr(m|\mathcal{A}_1(\omega_j)) < \Pr(m)$. If, instead, $\Pr(n|\omega_j) < \Pr(n)$, then $\mathcal{A}_1(\omega_j) = \{\omega_j, \omega_k\}$, $\lambda(\omega_j, \omega_k) > \lambda(\omega_i, \omega_j) = \lambda(\omega_i, \omega_k)$, $\lambda(\omega_i, \omega_k)$ is known, and thus $\lambda(\omega_i, \omega_j)$ is identified, as is $\lambda(\omega_j, \omega_k)$ since Theorem 1 implies it solves:

$$\Pr(n|\omega_j) = \frac{\Pr(n)^{\frac{\lambda(\omega_i,\omega_k)}{\lambda(\omega_j,\omega_k)}} \Pr(n|\mathcal{A}_1(\omega_j))^{\frac{\lambda(\omega_j,\omega_k)-\lambda(\omega_i,\omega_k)}{\lambda(\omega_j,\omega_k)}} e^{\frac{\mathbf{v}_n(\omega_j)}{\lambda(\omega_j,\omega_k)}}}{\sum\limits_{\nu\in\{n,m\}} \Pr(\nu)^{\frac{\lambda(\omega_i,\omega_k)}{\lambda(\omega_j,\omega_k)}} \Pr(\nu|\mathcal{A}_1(\omega_j))^{\frac{\lambda(\omega_j,\omega_k)-\lambda(\omega_i,\omega_k)}{\lambda(\omega_j,\omega_k)}} e^{\frac{\mathbf{v}_\nu(\omega_j)}{\lambda(\omega_j,\omega_k)}}}$$

$$= \frac{1}{1 + \left(\frac{\Pr(m)}{\Pr(n)}\right)^{\frac{\lambda(\omega_i,\omega_k)}{\lambda(\omega_j,\omega_k)}} \left(\frac{\Pr(m|\mathcal{A}_1(\omega_j))}{\Pr(n|\mathcal{A}_1(\omega_j))}\right)^{\frac{\lambda(\omega_j,\omega_k)-\lambda(\omega_i,\omega_k)}{\lambda(\omega_j,\omega_k)}}},$$

$$= \frac{1}{1 + \frac{\Pr(m|\mathcal{A}_1(\omega_j))}{\Pr(n|\mathcal{A}_1(\omega_j))} \left(\frac{\Pr(m)}{\Pr(m|\mathcal{A}_1(\omega_j))} \frac{\Pr(n|\mathcal{A}_1(\omega_j))}{\Pr(n)}\right)^{\frac{\lambda(\omega_i,\omega_k)}{\lambda(\omega_j,\omega_k)}}},$$

which clearly has a unique solution that some simple algebra produces a closed-form solution for as $\Pr(n|\mathcal{A}_1(\omega_j)) < \Pr(n)$ and $\Pr(m|\mathcal{A}_1(\omega_j)) > \Pr(m)$.

If condition **(iv)** is satisfied, so $\mathbf{v}_n(\omega_i) - \mathbf{v}_m(\omega_i) = \mathbf{v}_n(\omega_j) - \mathbf{v}_m(\omega_j) > 0 < \mathbf{v}_m(\omega_k) - \mathbf{v}_n(\omega_k)$ and $\lambda(\omega_i, \omega_k) \neq \lambda(\omega_j, \omega_k)$, then Theorem 4 implies there is μ with $\mu(\omega_i) + \mu(\omega_k) = 1$ and $\mu(\omega_i) \in (0, 1)$ such that $\mathbb{P}(\{n, m\}, \mu)$ features a positive probability of both n and m being selected as for any $c > 0$ there is such a μ with:

$$\sum_{\omega \in \{\omega_i, \omega_k\}} \frac{e^{\frac{\mathbf{v}_n(\omega)}{c}}}{e^{\frac{\mathbf{v}_m(\omega)}{c}}} \mu(\omega) > 1 \text{ and } \sum_{\omega \in \{\omega_i, \omega_k\}} \frac{e^{\frac{\mathbf{v}_m(\omega)}{c}}}{e^{\frac{\mathbf{v}_n(\omega)}{c}}} \mu(\omega) > 1,$$

so $\lambda(\omega_i, \omega_k)$ is identified using the logic from condition **(i)**. Similarly, $\lambda(\omega_j, \omega_k)$ is identified, and, if $\lambda(\omega_i, \omega_k) \neq \lambda(\omega_j, \omega_k)$, as is the case when condition **(iv)** is satisfied, then it is evident from the set of optimal beahvior as a result, and $\lambda(\omega_i, \omega_j) = \min(\lambda(\omega_i, \omega_k), \lambda(\omega_j, \omega_k))$ due to the nature of partitions.

If condition **(v)** is satisfied, so $\mathbf{v}_n(\omega_i) - \mathbf{v}_m(\omega_i) = 0$, $\mathbf{v}_n(\omega_k) - \mathbf{v}_m(\omega_k) > 0 < \mathbf{v}_m(\omega_r) - \mathbf{v}_n(\omega_r)$, and $\lambda(\omega_i, \omega_k) \neq \lambda(\omega_i, \omega_r)$, then the work done in the consideration of condition **(iii)** above indicates that this is observable and $\lambda(\omega_i, \omega_k)$ and $\lambda(\omega_i, \omega_r)$ are both identified by the set of optimal choice behavior. Similarly, if condition **(v)** is satisfied, so $\mathbf{v}_n(\omega_j) - \mathbf{v}_m(\omega_j) = 0$, $\mathbf{v}_n(\omega_k) - \mathbf{v}_m(\omega_k) > 0 < \mathbf{v}_m(\omega_r) - \mathbf{v}_n(\omega_r)$, and $\lambda(\omega_j, \omega_k) \neq \lambda(\omega_j, \omega_r)$, then the work done in the consideration of condition **(iii)** above indicates that this is observable and $\lambda(\omega_j, \omega_k)$ and $\lambda(\omega_j, \omega_r)$ are both identified by the set of optimal choice behavior. Further, if condition **(v)** is satisfied then either $\lambda(\omega_i, \omega_k) \neq \lambda(\omega_j, \omega_k)$ and $\lambda(\omega_i, \omega_j) = \min(\lambda(\omega_i, \omega_k), \lambda(\omega_j, \omega_k))$ due to the nature of partitions, or $\lambda(\omega_i, \omega_r) \neq \lambda(\omega_j, \omega_r)$ and $\lambda(\omega_i, \omega_j) = \min(\lambda(\omega_i, \omega_r), \lambda(\omega_j, \omega_r))$ due to the nature of partitions, so either way $\lambda(\omega_i, \omega_j)$ is identified.

What remains to be shown is that if for each pair of ω_i and ω_j in Ω, if $\lambda(\omega_i, \omega_j)$ is known, then $\mathbb{H}$ is known. First, organise all the $\lambda(\omega_i, \omega_j)$ into groups so that two such λs are in the same group iff they have the same value, and number the groups so that groups with lower numbers have lower values. Then λ_1 must be equal to the value of the members of group 1, λ_2 must be equal to the value of the members of group 2, and continuing in this way, λ_M must be the value of the members of the highest group, so the multipliers $\lambda_M > \cdots > \lambda_1 > 0$ have been identified. Next, notice that $\mathcal{A}_1(\omega_i) = \mathcal{A}_1(\omega_j)$ iff $\lambda(\omega_i, \omega_j) \neq \lambda_1$, so the events that constitute $\mathcal{A}_1$ are known. Further, for each ω_i and ω_j such that $\mathcal{A}_1(\omega_i) = \mathcal{A}_1(\omega_j)$, $\cap_{k=1}^2 \mathcal{A}_k(\omega_i) = \cap_{k=1}^2 \mathcal{A}_k(\omega_j)$ iff $\lambda(\omega_i, \omega_j) \neq \lambda_2$, so the events that constitute $\cap_{k=1}^2 \mathcal{A}_k$ are known. Similarly, for each $m \in \{1, \ldots, M-1\}$ and ω_i and ω_j such that $\cap_{k=1}^{m}\mathcal{A}_k(\omega_i) = \cap_{k=1}^{m}\mathcal{A}_k(\omega_j)$, $\cap_{k=1}^{m+1}\mathcal{A}_k(\omega_i) = \cap_{k=1}^{m+1}\mathcal{A}_k(\omega_j)$ iff $\lambda(\omega_i,\, \omega_j) \neq \lambda_{m+1}$, so the events that constitute $\cap_{k=1}^{m+1}\mathcal{A}_k$ are known. Thus, while the attributes themselves are not identified, $\mathbb{H}$ is identified. ■

Appendix 2

In this appendix axioms are used to develop this paper's measure of uncertainty, MASE, which can be understood as the expected cost to the agent of perfectly observing the state of the world. MASE can be used to study a rationally inattentive agent because the cost of any imprecise learning can be taken to be the expected reduction in uncertainty it causes, as is done with Shannon Entropy in the Shannon RI model. Thus, while this paper is interested in studying an inattentive agent that might only partially learn about the state of the world, the axioms in this appendix, like Shonnon's original axioms (1948), discuss an attentive agent that perfectly observes the state of the world. Before the axioms are introduced, some notation and terminology is required.

Learning Strategies

One natural way to think about an agent learning about the state of the world is through a series of questions that have answers that are determined by the state of the world.[13] I use partitions to model such question because a question with multiple potential answers is equivalent to a partition of the state space whenever the answer to the question is determined by the state of the world. This equivalence occurs since I can simply group states of the world based on the answer to the question they produce. The words 'question' and 'partition' are thus used interchangeably in this appendix.

The simplest kind of question in this setting is a yes or no question. A yes or no question is equivalent to a **binary partition** $\mathcal{P}^b$ of Ω, which I define as a set of two events, $\mathcal{P}^b = \{A_1,\, A_2\}$, such that $A_1 \cup A_2 = \Omega$, and $A_1 \cap A_2 = \emptyset$. The two phrases 'binary partition' and 'yes or no question' are thus used interchangeably in this appendix.

Given a prior μ, and some partition $\mathcal{P}$, let $C(\mathcal{P},\, \mu) \in \mathbb{R}_+$ denote the (expected) cost of learning the realized event $\mathcal{P}(\omega)$ of $\mathcal{P}$, that is, the agent's expected cost of changing their belief from μ to $\mu(\cdot|\mathcal{P}(\omega))$. $C(\mathcal{P},\, \mu)$, the cost of answering 'What is the realized event of $\mathcal{P}$?' given the agent's prior belief, is the basic building block of this appendix.

[13] A question's answer is said to be determined by the state of the world if knowing the state indicates the answer to the question with certainty.

A **learning strategy**, $S = (\mathcal{P}_1, \ldots, \mathcal{P}_n)$, is a list of partitions whose realized events are successively observed by the agent such that if $\mathcal{P}_i, \mathcal{P}_j \in S$, and $i \neq j$, then $\mathcal{P}_i \neq \mathcal{P}_j$. A 'learning strategy' is thus 'a series of questions' and the two phrases are used interchangeably in this appendix. When the agent selects a learning strategy of this form it may seem that the agent is being restricted to selecting 'history-independent' learning strategies in the sense that it seems like they cannot select the second partition based on the realization of the first partition, but this is not really the case. When the agent selects the second partition for their learning strategy they are essentially choosing a (perhaps trivial) partition of each of the potential realized events of the first partition, and thus their learning strategy is effectively 'history-dependent;' they are effectively choosing what to learn next based on what they have already learned.

If a learning strategy consists of only binary partitions, I call it a **binary learning strategy**, and denote it $S^b = (\mathcal{P}_1^b, \ldots, \mathcal{P}_n^b)$. The order of the questions in a learning strategy is important, and changing the order results in a different learning strategy. If, for instance, some questions are more costly for the agent to answer, and help to identify states that are seldom observed, then it may seem efficient for a learning strategy to leave these questions towards the end.[14]

I define $C(S, \mu)$, which is the (expected) cost of a learning strategy $S = (\mathcal{P}_1, \ldots, \mathcal{P}_n)$ given a probability measure μ, to be the sum of the costs of each of the questions in S:

$$C(S, \mu) = C(\mathcal{P}_1, \mu) + \mathbb{E}\Big[C\Big(\mathcal{P}_2, \mu(\cdot|\mathcal{P}_1(\omega))\Big) + \cdots + C\Big(\mathcal{P}_n, \mu(\cdot|\cap_{i=1}^{n-1}\mathcal{P}_i(\omega))\Big)\Big].$$

The definition of $C(S, \mu)$ thus imposes a form of constant marginal cost onto learning strategies because over the course of their learning strategy the agent does not fatigue, nor do they gain experience with research and become better at learning: all that matters for determining the cost of each question are the beliefs of the agent immediately before the question is answered, and not how much has previously been learned.

If B is a collection of partitions, let $\sigma(B)$ denote the **σ-algebra generated by** B, which is the smallest σ-algebra containing all the events in each of the partitions in B, as is done in Appendix 1. Since a learning strategy S is a collection of partitions, I use $\sigma(S)$ to denote the σ-algebra generated by S.

Sometimes a single question can be as informative as several questions. I say a learning strategy S is **equivalent** to a partition $\mathcal{P}$ if $\sigma(S) = \sigma(\mathcal{P})$. What $\sigma(S) = \sigma(\mathcal{P})$ means intuitively is

[14] The order of the events in a partition, in contrast, is not important, and switching the order in which the events in a partition are listed does not result in a different partition.

that, for any prior probability measure $\mu : \mathcal{F} \to \mathbb{R}_+$, observing the answers to the series of questions in S always leads to the same posterior as observing the answer to the question 'what is the realized event of the partition $\mathcal{P}$?', and thus, for all priors, S and $\mathcal{P}$ provide the same information.

Axioms

What form should a cost function for information take? This difficult question does not have an obvious answer, so this appendix provides axioms that help illustrate the structure imposed by MASE. Each axiom can be separately evaluated in different contexts, either empirically, or through introspection, to determine how appropriate it is. Further, the axioms help demonstrate to those that are familiar with Shannon's original axioms (1948) the differences between MASE and standard Shannon Entropy.

Axiom 1 (Measurement): Given a binary partition $\mathcal{P}^b = \{A_1, A_2\}$, $C(\mathcal{P}^b, \mu)$ is determined by $\mu(A_1)$ and $\mu(A_2)$: if μ and $\tilde{\mu}$ are two probability measures on Ω with $\mu(A_1) = \tilde{\mu}(A_1)$ (and hence $\mu(A_2) = \tilde{\mu}(A_2)$), then $C(\mathcal{P}^b, \mu) = C(\mathcal{P}^b, \tilde{\mu})$, and notationally I can thus replace $C(\mathcal{P}^b, \mu)$ with $C(\mathcal{P}^b, \mu(A_1), \mu(A_2))$.

In plain language, Axiom 1 says that the expected cost of learning the answer to the yes or no question represented by $\mathcal{P}^b$ should be determined by the probability of the answer being yes and the probability of the answer being no. If I know the yes or no question being asked, and the probability of each of its answers, then I know the expected cost of answering the question, I do not require any additional information.[15]

I am now going to introduce learning strategy invariance, a concept that is the central pillar of Shannon's (1948) axioms and helps to make it explicit what I am assuming with this paper's axioms. In general, a particular question $\mathcal{P}$ and an equivalent series of questions S may produce different expected costs depending on what questions are selected to be in S and how they are ordered. A given question $\mathcal{P}$, however, may have the peculiar property that, given any prior, all series of questions that are equivalent to it have the same expected cost, in which case I say it is learning strategy invariant. Formally, I say a partition $\mathcal{P}$ is **learning strategy invariant**, if for

[15] The axioms focus on learning with yes or no questions for a number of reasons. Eye tracking analysis shows that when agents are faced with multiple options, they successively compare pairs of the options along a single attribute dimension (Noguchi & Stewart, 2014, 2018). This suggests that, in practice, agents are breaking their learning into a number of smaller queries. Further, in the psychology literature these pairwise comparisons are frequently modelled as ordinal in nature (Noguchi & Stewart, 2018), equivalent to questions with binary outcomes, e.g. 'Is option a better than option b in dimension x?', instead of more complicated questions, e.g. 'How much better is option a than option b in dimension x?', because findings in the field of psychophysics suggest that agents are good at discriminating stimuli, but are not good at determining the magnitude of the same stimuli (Stewart, Chater, & Brown, 2006).

each probability measure μ, the expected cost $C(S, \mu)$ is the same for every learning strategy S that is equivalent to $\mathcal{P}$.

In many environments there are questions that are not learning strategy invariant. Consider the environment described in Example 2 in Section 2.2 and let $A_1 = \{\omega_1, \omega_2\}$, $A_2 = \{\omega_1, \omega_3\}$, $\mathcal{P}_1^b = \{A_1, A_1^c\}$, and $\mathcal{P}_2^b = \{A_2, A_2^c\}$. Notice that observing the realized event of $\mathcal{P}_1^b$ is equivalent to learning the value of option 1, and observing the realized event of $\mathcal{P}_2^b$ is equivalent to learning the value of option 2. Now, let $\mathcal{P}_3 = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}\}$ denote the partition of the state space. Notice that the learning strategy $S^b = (\mathcal{P}_1^b, \mathcal{P}_2^b)$ is equivalent to $\mathcal{P}_3$, because if I answer 'What is the value of option 1?', and then answer 'What is the value of option 2?', I have observed the state of the world. Based on the discussion in Section 2.2, however, $\mathcal{P}_3$ should not be learning strategy invariant if it is assumed that it is costless to answer a question if the belief of the agent assigns a probability of one to one of its events. If the distribution over the states described in Section 2.2 is altered so that $\mu(\omega_1) = \mu(\omega_4) = \frac{1}{2}$, then observing the value of one of the options would tell you the value of the other. The cost of S^b would then be the cost of observing the value of option 1, which I assumed to be less than the cost of observing the value of option 2, which is then the cost of $\tilde{S}^b = (\mathcal{P}_2^b, \mathcal{P}_1^b)$, which is also equivalent to $\mathcal{P}_3$, so $\mathcal{P}_3$ should not be learning strategy invariant. This is why the Shannon RI model makes strange predictions in Example 2, because Shannon Entropy imposes that all partitions are learning strategy invariant.

A set of partitions that are certainly learning strategy invariant, in contrast, is the set of binary partitions. If $\mathcal{P}^b$ is a binary partition, then $\mathcal{P}^b$ is learning strategy invariant because the only learning strategy S such that $\sigma(S) = \sigma(\mathcal{P}^b)$, is $S = (\mathcal{P}^b)$. Thus, for any μ, all learning strategies S such that $\sigma(S) = \sigma(\mathcal{P}^b)$ have the same expected cost $C(S, \mu) = C(\mathcal{P}^b, \mu)$.

As the next four lemmas show, quite a bit of structure is imposed onto C when it is applied to learning strategy invariant partitions. In particular, structure is imposed onto C when it is applied to any partition that is coarser than a learning strategy invariant partition, and this structure ends up being useful. I say a partition $\mathcal{P}$ of a state space Ω is **coarser** than a partition $\tilde{\mathcal{P}}$ of the same state space Ω, if each event in $\mathcal{P}$ corresponds to a union of events in $\tilde{\mathcal{P}}$.

Lemma 6. If a partition $\tilde{\mathcal{P}}$ is coarser than a learning strategy invariant partition $\mathcal{P}$, then $\tilde{\mathcal{P}}$ is also learning strategy invariant.

Proof. Suppose $\mathcal{P}$ is a learning strategy invariant partition, and $\tilde{\mathcal{P}}$ is coarser than $\mathcal{P}$. If $\tilde{\mathcal{P}} = \mathcal{P}$ I am done, so assume $\tilde{\mathcal{P}} \neq \mathcal{P}$. The definition of learning strategy invariance then indicates that for any learning strategy $\tilde{S} = (\mathcal{P}_1, \ldots, \mathcal{P}_n)$ such that $\sigma(\tilde{S}) = \sigma(\tilde{P})$, and any μ:

$$C(\mathcal{P}, \mu) = C(\tilde{\mathcal{P}}, \mu) + \mathbb{E}[C(\mathcal{P}, \mu(\cdot|\tilde{\mathcal{P}}(\omega)))] = C(\tilde{S}, \mu) + \mathbb{E}[C(\mathcal{P}, \mu(\cdot|\tilde{\mathcal{P}}(\omega)))].$$

Thus, $C(\tilde{\mathcal{P}}, \mu) = C(\tilde{S}, \mu)$ for all such $\tilde{S}$, and any μ, so $\tilde{\mathcal{P}}$ is also learning strategy invariant. ■

Lemma 6 makes sense because, if a partition $\tilde{\mathcal{P}}$ is coarser than a learning strategy invariant partition $\mathcal{P}$, the way the realised event of $\tilde{\mathcal{P}}$ is learnt cannot impact the expected cost of learning it as then the cost of learning the realized event of $\mathcal{P}$ could differ depending on how the realised event of $\tilde{\mathcal{P}}$ is learnt. Lemma 7 also makes a lot of sense because if the agent assigns a probability of one to a particular event in a partition then they already know the realized event with certainty and 'learning' the realized event should be costless.

Lemma 7. If $\mathcal{P} = \{A_1, \ldots, A_m\}$ is a learning strategy invariant partition with $m \geq 3$, and probability measure μ assigns a probability of one to an event $A_i \in \mathcal{P}$, then $C(\mathcal{P}, \mu) = 0$.

Proof. Suppose $\mathcal{P} = \{A_1, \ldots, A_m\}$ is a learning strategy invariant partition with $m \geq 3$ and there is an $A_i \in \mathcal{P}$ such that $\mu(A_i) = 1$. It is without loss to further assume $i = 1$. Let $\tilde{\mathcal{P}} = \{A_1, A_1^c\}$, $\hat{\mathcal{P}} = \{A_1 \cup A_2, A_3, \ldots, A_m\}$, $S_1 = (\tilde{\mathcal{P}}, \hat{\mathcal{P}})$, and $S_2 = (\tilde{\mathcal{P}}, \hat{\mathcal{P}}, \mathcal{P})$. The definition of learning strategy invariance indicates that $C(S_1, \mu) = C(S_2, \mu)$, so $C(\mathcal{P}, \mu) = 0$ if μ assigns a probability of one to an event in $\mathcal{P}$. ■

Before introducing the next lemma, I require another definition. If $\mathcal{P} = \{A_1, \ldots, A_m\}$ is a learning strategy invariant partition, I say that $\tilde{\mu}$ is a **permutation** of μ on $\mathcal{P}$ if there is a bijection $\pi : \{1, \ldots, m\} \to \{1, \ldots, m\}$ such that $\forall i \in \{1, \ldots, m\}$, $\mu(A_i) = \tilde{\mu}(A_{\pi(i)})$. The result in Lemma 8 is perhaps more surprising, but it speaks to the amount of structure imposed onto the cost of learning the realised event of a partition by learning strategy invariance, as is further demonstrated by Lemma 9.

Lemma 8. If a partition $\mathcal{P} = \{A_1, \ldots, A_m\}$ is learning strategy invariant, with $m \geq 3$, and C satisfies Axiom 1, then if $\tilde{\mu}$ is a permutation of μ on $\mathcal{P}$, then $C(\mathcal{P}, \mu) = C(\mathcal{P}, \tilde{\mu})$.

Proof. Suppose $\mathcal{P} = \{A_1, \ldots, A_m\}$ is a learning strategy invariant partition of the state space Ω with $m \geq 3$. Axiom 1 imposes that knowing $\mu(A_1), \ldots,$ and $\mu(A_m)$ is enough information to compute the expected learning costs of binary partitions coarser than $\mathcal{P}$, and thus, given $\mathcal{P}$, $C(\mathcal{P}, \mu)$ is determined by $\mu(A_1), \mu(A_2) \ldots,$ and $\mu(A_m)$.

If I then show that for any $i, j \in \{1, \ldots, m\}$ with $i \neq j$, and probability measures μ and $\tilde{\mu}$ with $\mu(A_k) = \tilde{\mu}(A_k)$ for $k \notin \{i, j\}$, $\mu(A_i) = \tilde{\mu}(A_j)$, and $\mu(A_j) = \tilde{\mu}(A_i)$, that $C(\mathcal{P}, \mu) = C(\mathcal{P}, \tilde{\mu})$, then the desired result holds since a series of pairwise switches like this can be used to create any permutation desired. Assume that μ and $\tilde{\mu}$ satisfy the conditions from the previous sentence. It is without loss to assume $i = 1$ and $j = 2$. Define $\tilde{\mathcal{P}} = \{A_1, A_2, (A_1 \cup A_2)^c\}$ (it is fine if $\tilde{\mathcal{P}} = \mathcal{P}$). Notice that $\tilde{\mathcal{P}}$ must be learning strategy invariant based on Lemma 6. Further, if I show that $C(\tilde{\mathcal{P}}, \mu) = C(\tilde{\mathcal{P}}, \tilde{\mu})$ then $C(\mathcal{P}, \mu) = C(\mathcal{P}, \tilde{\mu})$ since, if I define $\hat{\mathcal{P}} = \{A_1 \cup A_2, A_3, \ldots, A_m\}$ that is also learning strategy invariant based on Lemma 6, then Lemma 7 and the definition of learning strategy invariance tells us:

$$C(\mathcal{P}, \mu) = C(\tilde{\mathcal{P}}, \mu) + (1 - \mu(A_1 \cup A_2))C(\hat{\mathcal{P}}, \hat{\mu})$$

$$= C(\tilde{\mathcal{P}}, \tilde{\mu}) + (1 - \mu(A_1 \cup A_2))C(\hat{\mathcal{P}}, \hat{\mu}) = C(\mathcal{P}, \tilde{\mu}),$$

if I define probability measure $\hat{\mu}$ so that if $\mu(A_1 \cup A_2)) < 1$ then $\hat{\mu}(A_1) = \hat{\mu}(A_2) = 0$ and $\hat{\mu}(A_i) = \mu(A_i)/(1 - \mu(A_1 \cup A_2))$ for $i \in \{3, \ldots, m\}$, and otherwise so that $\hat{\mu}(A_1) = 1$. Now, let $\mathcal{P}_1^b = \{A_1, A_1^c\}$, $\mathcal{P}_2^b = \{A_2, A_2^c\}$, and $\mathcal{P}_3^b = \{A_1 \cup A_2, (A_1 \cup A_2)^c\}$. Notice $\mathcal{P}_1^b$, $\mathcal{P}_2^b$ and $\mathcal{P}_3^b$, are all coarser than $\tilde{\mathcal{P}}$. Then, since $\tilde{\mathcal{P}}$ is learning strategy invariant:

$$C(\tilde{\mathcal{P}}, \mu) = C(\mathcal{P}_3^b, \mu) + \mathbb{E}[C(\mathcal{P}_1^b, \mu(\cdot|\mathcal{P}_3^b(\omega))], \text{ and } C(\tilde{\mathcal{P}}, \tilde{\mu}) = C(\mathcal{P}_3^b, \tilde{\mu}) + \mathbb{E}[C(\mathcal{P}_1^b, \tilde{\mu}(\cdot|\mathcal{P}_3^b(\omega))].$$

Notice that Axiom 1 imposes that $C(\mathcal{P}_3^b, \mu) = C(\mathcal{P}_3^b, \tilde{\mu})$ since both μ and $\tilde{\mu}$ assign the same probability to the events $A_1 \cup A_2$ and $(A_1 \cup A_2)^c$. So, all that remains to be shown is that if the probability measure $\tilde{\nu}$ is a permutation of the probability measure ν on $\mathcal{P}_1^b$, then $C(\mathcal{P}_1^b, \nu) = C(\mathcal{P}_1^b, \tilde{\nu})$. Fix arbitrary $\nu(A_1) = x \in [0, 1]$. Now consider the probability measures q_1, q_2, q_3, such that:

$$q_1(A_1) = x,\ q_1(A_2) = 0,\ q_1((A_1 \cup A_2)^c) = 1 - x,$$

$$q_2(A_1) = 0,\ q_2(A_2) = x,\ q_2((A_1 \cup A_2)^c) = 1 - x,$$

$$q_3(A_1) = 1 - x,\ q_3(A_2) = x,\ q_3((A_1 \cup A_2)^c) = 0.$$

Notice that q_3 is a permutation of q_1 on $\mathcal{P}_1^b$. So then, using Axiom 1, the definition of learning strategy invariance, and Lemma 7, all repeatedly:

$$C(\mathcal{P}_1^b, q_1) = C(\tilde{\mathcal{P}}, q_1) = C(\mathcal{P}_3^b, q_1) = C(\mathcal{P}_3^b, q_2)$$

$$= C(\tilde{\mathcal{P}}, q_2) = C(\mathcal{P}_2^b, q_2) = C(\mathcal{P}_2^b, q_3) = C(\tilde{\mathcal{P}}, q_3) = C(\mathcal{P}_1^b, q_3). \blacksquare$$

Lemma 9. If a partition $\mathcal{P} = \{A_1, \ldots, A_m\}$ is learning strategy invariant with $m \geq 3$, $\mathcal{P}^b$ is a binary partition that is coarser than $\mathcal{P}$, and C satisfies Axiom 1, then for all (p_1, p_2, p_3) such that $p_1, p_2, p_3 \in [0, 1)$ and $p_1 + p_2 + p_3 = 1$:

$$C(\mathcal{P}^b, p_1, 1 - p_1) + (1 - p_1)C\Big(\mathcal{P}^b, \frac{p_2}{p_2 + p_3}, \frac{p_3}{p_2 + p_3}\Big)$$

$$= C(\mathcal{P}^b, p_2, 1 - p_2) + (1 - p_2)C\Big(\mathcal{P}^b, \frac{p_1}{p_1 + p_3}, \frac{p_3}{p_1 + p_3}\Big)$$

$$= C(\mathcal{P}^b, p_3, 1 - p_3) + (1 - p_3)C\Big(\mathcal{P}^b, \frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\Big).$$

Proof. For all partitions $\mathcal{P} = \{A_1, \ldots, A_m\}$ and probability measures μ defined on $\mathcal{P}$, define the vector $\mu(\mathcal{P}) = (\mu(A_1), \ldots, \mu(A_m))$.

Suppose C satisfies Axiom 1, that $\mathcal{P}_i = \{A_1, \ldots, A_m\}$ is a learning strategy invariant with $m \geq 3$, and $\tilde{\mathcal{P}}_i$ is another learning strategy invariant partition that is coarser than $\mathcal{P}_i$. Lemma 8 indicates that $C(\mathcal{P}_i, \mu)$ is determined by $\mu(\mathcal{P}_i)$, and if the strictly positive entries of $\mu(\mathcal{P}_i)$ and $\mu(\tilde{\mathcal{P}}_i)$ are the same (up to a permutation), then the addition of Lemma 7 and the definition of learning strategy invariant partitions indicates that $C(\mathcal{P}_i, \mu) = C(\tilde{\mathcal{P}}_i, \mu)$ since I can pick μ so that uncertainty about which event in $\mathcal{P}_i$ has been realized is fully determined by the realized event of $\tilde{\mathcal{P}}_i$. What does this mean? This means that there is a function which maps from vectors of probabilities onto the reals, $c_i : \cup_{j=1}^{m-1} \triangle^j \to \mathbb{R}$, where $\triangle^j$ is the j simplex, such that for any learning strategy invariant partition $\tilde{\mathcal{P}}_i$ coarser than $\mathcal{P}_i$, if the strictly positive entries of $\mu(\mathcal{P}_i)$ and $\mu(\tilde{\mathcal{P}}_i)$ are the same (up to a permutation) then $C(\tilde{\mathcal{P}}_i, \mu) = c_i(\mu(\tilde{\mathcal{P}}_i)) = c_i(\mu(\mathcal{P}_i)) \equiv C(\mathcal{P}_i, \mu)$.

So, for any binary partition $\mathcal{P}^b$ coarser than $\mathcal{P}_i$, $C(\mathcal{P}^b, \mu) = c_i(\mu(\mathcal{P}^b))$ (notice that this means that $C(\mathcal{P}^b, \mu)$ is constant with respect to permutations of μ on $\mathcal{P}^b$ for all such $\mathcal{P}^b$ since $C(\mathcal{P}, \mu)$ is constant with respect to permutations of μ on $\mathcal{P}$). Now pick $\tilde{\mathcal{P}}_i = \{B_1, B_2, B_3\}$ so that it is coarser than $\mathcal{P}_i$ and it has three elements. Lemma 6 indicates that $\tilde{\mathcal{P}}_i$ is learning strategy invariant, and it is easy to show each binary partition which is coarser than $\tilde{\mathcal{P}}_i$ is coarser than $\mathcal{P}_i$. Thus, for all probability measures μ on $\tilde{\mathcal{P}}_i$ such that $\mu(B_1)$, $\mu(B_2)$, and $\mu(B_3)$ are all strictly less than one, the definition of learning strategy invariance tells us:

$$C(\tilde{\mathcal{P}}_i, \mu) = c_i(\mu(B_1), 1 - \mu(B_1)) + (1 - \mu(B_1))c_i\Big(\frac{\mu(B_2)}{\mu(B_2) + \mu(B_3)}, \frac{\mu(B_3)}{\mu(B_2) + \mu(B_3)}\Big)$$

$$= c_i(\mu(B_2),\, 1 - \mu(B_2)) + (1 - \mu(B_2))c_i\Big(\frac{\mu(B_1)}{\mu(B_1) + \mu(B_3)},\, \frac{\mu(B_3)}{\mu(B_1) + \mu(B_3)}\Big)$$

$$= c_i(\mu(B_3),\, 1 - \mu(B_3)) + (1 - \mu(B_3))c_i\Big(\frac{\mu(B_1)}{\mu(B_1) + \mu(B_2)},\, \frac{\mu(B_2)}{\mu(B_1) + \mu(B_2)}\Big). \blacksquare$$

In plainer language, Lemma 9 says that if the cost of learning satisfies Axiom 1 and $\mathcal{P}^b$ is a binary partition that is coarser than a learning strategy invariant partition with at least three events, then for $p_1, p_2, p_3 \in [0, 1)$ and $p_1 + p_2 + p_3 = 1$, the cost of learning the realized event of $\mathcal{P}^b$ when they occur with probabilities p_1 and $1 - p_1$ plus $1 - p_1$ times the cost of learning the realized event of $\mathcal{P}^b$ when the events occur with probabilities $\frac{p_2}{p_2+p_3}$ and $\frac{p_3}{p_2+p_3}$, is equal to the cost of learning the realized event of $\mathcal{P}^b$ when the events occur with probabilities p_2 and $1 - p_2$ plus $1 - p_2$ times the cost of learning the realized event of $\mathcal{P}^b$ when the events occur with probabilities $\frac{p_1}{p_1+p_3}$ and $\frac{p_3}{p_1+p_3}$, which is also equal to the cost of learning the realized event of $\mathcal{P}^b$ when the events occur with probabilities p_3 and $1 - p_3$ plus $1 - p_3$ times the cost of learning the realized event of $\mathcal{P}^b$ when the events occur with probabilities $\frac{p_1}{p_1+p_2}$ and $\frac{p_2}{p_1+p_2}$. This all means that Axiom 1 imposes a staggering amount of structure onto the cost of learning the realized event of a binary partition whenever it is coarser than some other learning strategy invariant partition.

One limitation of Shannon's (1948) original work, at least when applied in economic settings, is that he assumes that permuting the order of questions does not change the expected cost of learning the state of the world. This might not make sense, however, if different attributes of the choice environment have different learning costs associated with them. If some attributes are less expensive to learn about then it might make sense to learn about these attributes first, as is argued in this subsection in the context of Example 2.

What I instead desire is that that permuting the order of questions does not change the expected learning cost only if the questions provide 'similar' information about the state of the world. The most succinct and objective way to discuss a partition providing 'similar' information to another partition is with a product space, as is explained in the next paragraph. This is because, if I ignore the distribution over states and the product is taken over replicas of the same set of states, then a question about the realization of one of the elements of the product is essentially identical to the same question about the realization of one of the other elements of the product.

To make this more concrete, consider the state space Ω and associated choice environment from Example 2 and replicate the state space three times so that the new state space is $\tilde{\Omega} = \Omega_1 \times \Omega_2 \times \Omega_3$ with $\Omega_1 = \Omega_2 = \Omega_3 = \Omega$, but do not yet fix any distribution over states. A natural interpretation of this newly constructed choice environment is that there are three *a priori* identical local firms, three *a priori* identical foreign firms, and the realization of Ω_i determines the value of investing in both local firm i and foreign firm i for $i \in \{1, 2, 3\}$. Suppose $\mathcal{P}^b$ is a binary partition of Ω and that $\mathcal{P}^b_i$ is the equivalent binary partition of Ω_i for each i. For simplicity, one can assume $\mathcal{P}^b_i$ is the question "Is local firm i of high quality?". Thus, $\mathcal{P}^b_1$, $\mathcal{P}^b_2$, and $\mathcal{P}^b_3$, are as similar as partitions can be by construction as they ask the values of three *a priori* identical firms. Now, suppose that the agent knows that the answer to one of the questions, $\mathcal{P}^b_1$, $\mathcal{P}^b_2$, or $\mathcal{P}^b_3$, is 'yes,' while the other two have the answer 'no,' which means that exactly one of the three local firms has high value.[16] Denote the probability of $\mathcal{P}^b_i$ having the answer 'yes,' the probability of local firm i having high value, by $p_i \in [0, 1)$ for $i \in \{1, 2, 3\}$.[17] Suppose the agent begins by learning about the realized event of $\mathcal{P}^b_i$. If the agent learns the answer to $\mathcal{P}^b_i$ is 'yes' they have also learned the answers to the other two partitions as only one has the answer 'yes,' while if they instead learn the answer to $\mathcal{P}^b_i$ is 'no' then their belief is updated using Bayes' Rule so that the probability of the answer to $\mathcal{P}^b_j$ being 'yes' for $j \in \{1, 2, 3\}\backslash\{i\}$ is $\frac{p_j}{p_j+p_k}$, where $k \in \{1, 2, 3\}\backslash\{i, j\}$, and after they learn the answer to $\mathcal{P}^b_j$, no matter the answer, they know the realization of all three partitions as exactly one has the answer 'yes.' What Axiom 2 imposes is that, if $C(\mathcal{P}^b_i, p, 1-p) = C(\mathcal{P}^b, p, 1-p)$ for each $p \in [0, 1]$ and $i \in \{1, 2, 3\}$ and the answers feature the relationship outlined in this paragraph, the order in which the agent answers these three ostensibly identical questions is irrelevant to their expected learning cost: the order that the agent learns about the three firms does not change the expected cost of learning which of the three local firms is of high value.

Axiom 2 (Self-Similarity): Given a binary partition $\mathcal{P}^b$, and a vector of probabilities (p_1, p_2, p_3) such that $p_1, p_2, p_3 \in [0, 1)$ and $p_1 + p_2 + p_3 = 1$, C is such that:

$$\begin{aligned} &C(\mathcal{P}^b, p_1, 1-p_1) + (1-p_1)C\Big(\mathcal{P}^b, \frac{p_2}{p_2+p_3}, \frac{p_3}{p_2+p_3}\Big) \\ &= C(\mathcal{P}^b, p_2, 1-p_2) + (1-p_2)C\Big(\mathcal{P}^b, \frac{p_1}{p_1+p_3}, \frac{p_3}{p_1+p_3}\Big) \\ &= C(\mathcal{P}^b, p_3, 1-p_3) + (1-p_3)C\Big(\mathcal{P}^b, \frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\Big). \end{aligned}$$

The reader may notice that Axiom 2 implies that $C(\mathcal{P}^b, p, 1-p)$ is not constant in p (unless

[16] If the answer to one question does not contain information about the answers to the other questions, then assuming that the order in which they are answered does not impact expected costs is a vacuous assumption. The assumption made here is perhaps the simplest way to ensure the answer to one question provides information about the answer to the other questions.

[17] The open upper bound on the p_i ensures the agent does not already know the realization of the three partitions.

the cost is always zero) because, revisiting the example from the paragraph before Axiom 2, if the cost of learning the value of a firm is constant for $p \in (0,\, 1)$ then the agent could lower learning costs by learning about firms that have higher probabilities of being of high value first as this reduces the expected number of firms the agent must learn about. The intuition for why the cost of learning the value of essentially identical firms may differ is that the agent may posses different pieces of information about them, and thus what remains to be learnt about each firm may differ. Axiom 2 makes more sense if the belief of the agent is taken to be a parsimonious representation of the information the agent possesses: beginning by learning the value of a firm that the agent believes has a very low probability of being high value might not be a bad strategy if the low probability is indicative of the agent already possessing a lot of information about the firm and as a result it is in expectation not costly for them to rule out that it is of high value.

Next I make a quite weak assumption about the continuity of the cost function on binary partitions. As such, the axioms do not explicitly rule out discontinuities in the cost function, but, later results show that the cost function is continuous on binary partitions. This is because the properties described in Axiom 1 and Axiom 2 are only compatible with a cost function that is either continuous or discontinuous at every point for each binary partition.

Axiom 3 (Weak continuity): Given a binary partition $\mathcal{P}^b$, there is a probability $p \in [0,\, 1]$ such that C is continuous at $(p,\, 1-p)$ when applied to $\mathcal{P}^b$.

As was alluded to, a cost function on binary partitions only satisfies Axiom 1 and Axiom 2 if it is either continuous everywhere or discontinuous everywhere. Thus, if a cost function on binary partitions satisfies the first three axioms, it is continuous everywhere, as is formalized by Lemma 10, which further shows that the cost function is permutation invariant on binary partitions.

Lemma 10. If C satisfies Axiom 1, Axiom 2, and Axiom 3, then for each binary partition $\mathcal{P}^b$, $C(\mathcal{P}^b,\, p,\, 1-p)$ is continuous in p, and $C(\mathcal{P}^b,\, p,\, 1-p) = C(\mathcal{P}^b,\, 1-p,\, p)$, for each $p \in [0,\, 1]$.

Proof. I say that the vector $(q_1,\, \ldots,\, q_n)$ is a **permutation** of the vector $(p_1,\, \ldots,\, p_n)$ if there is a bijection $\pi : \{1,\, \ldots,\, n\} \to \{1,\, \ldots,\, n\}$ such that $\forall i \in \{1,\, \ldots,\, n\}$, $q_i = p_{\pi(i)})$. Before I prove Lemma 10 I show two technical results, Lemma 11 and Lemma 12, that are helpful for proving Lemma 10.

Lemma 11. Given a binary partition $\mathcal{P}^b$, if I define $c_{\mathcal{P}^b} : \cup_{j=1}^{\infty} \triangle^j \to \mathbb{R}$, where $\triangle^j$ is the j simplex, such that (for $n \geq 2$): $c_{\mathcal{P}^b}(p_1,\, \ldots,\, p_n) = C(\mathcal{P}^b,\, p_1,\, 1-p_1)$ if $p_1 + p_2 = 1$, and otherwise:

$$c_{\mathcal{P}^b}(p_1,\, \ldots,\, p_n) = C(\mathcal{P}^b,\, p_1,\, 1-p_1) + (1-p_1) C\Big(\mathcal{P}^b,\, \frac{p_2}{1-p_1},\, \frac{1-p_1-p_2}{1-p_1}\Big)$$

$$+ \ldots + (1 - p_1 - \ldots - p_{m-1}) C\Big(\mathcal{P}^b, \frac{p_m}{1 - p_1 - \ldots - p_{m-1}}, \frac{1 - p_1 - \ldots - p_m}{1 - p_1 - \ldots - p_{m-1}}\Big),$$

where m is the lowest integer such that $p_1 + \ldots + p_m = 1$, then if $(q_1, \ldots, q_n)$ is a permutation of $(p_1, \ldots, p_n)$, and C satisfies Axiom 1, and Axiom 2, then: $c_{\mathcal{P}^b}(q_1, \ldots, q_n) = c_{\mathcal{P}^b}(p_1, \ldots, p_n)$, and further if $(p_1, \ldots, p_n)$ is a vector $(n \geq 2)$ with one entry of value one, and the rest zero $c_{\mathcal{P}^b}(p_1, \ldots, p_n) = 0$.

Proof. Given a binary partition $\mathcal{P}^b$, suppose C satisfies Axiom 1, and Axiom 2, and that $c_{\mathcal{P}^b}$ is defined as above. All vectors discussed in this proof are assumed to sum to one and contain only non-negative constants. I proceed with an inductive argument, beginning by showing $c_{\mathcal{P}^b}(p, 1-p)$ satisfies the desired properties. Consider $c_{\mathcal{P}^b}(p_1, p_2, p_3)$ when $p_1, p_3 > 0$, and $p_2 = 0$. Axiom 2 tells us:

$$c_{\mathcal{P}^b}(p_1, 1-p_1) + (1-p_1)c_{\mathcal{P}^b}(0, 1) = c_{\mathcal{P}^b}(0, 1) + c_{\mathcal{P}^b}(p_1, 1-p_1) = c_{\mathcal{P}^b}(p_3, 1-p_3) + (1-p_3)c_{\mathcal{P}^b}(1, 0).$$

The first equality implies $c_{\mathcal{P}^b}(0, 1) = 0$. Now consider $c_{\mathcal{P}^b}(q_1, q_2, q_3)$ when $q_1, q_2 > 0$, and $q_3 = 0$. Axiom 2 tells us:

$$c_{\mathcal{P}^b}(q_1, q_2) + (1 - q_1)c_{\mathcal{P}^b}(1, 0) = c_{\mathcal{P}^b}(0, 1) + c_{\mathcal{P}^b}(q_1, q_2),$$

so since $c_{\mathcal{P}^b}(0, 1) = 0$, I know $c_{\mathcal{P}^b}(1, 0) = 0 = c_{\mathcal{P}^b}(0, 1)$, and combined with the previous two equalities above I know:

$$c_{\mathcal{P}^b}(p_1, 1 - p_1) = c_{\mathcal{P}^b}(p_3, 1 - p_3) + (1 - p_3)c_{\mathcal{P}^b}(1, 0) = c_{\mathcal{P}^b}(1 - p_1, p_1).$$

Thus, $c_{\mathcal{P}^b}(p, 1 - p) = c_{\mathcal{P}^b}(1 - p, p)$ for all $p \in [0, 1]$. Since $c_{\mathcal{P}^b}(1, 0) = 0$, when I show $c_{\mathcal{P}^b}$ is constant with respect to permutations of vectors of arbitrary length (greater or equal to two), it establishes that if $(p_1, \ldots, p_n)$ is a vector $(n \geq 2)$ with one entry of value one, and the rest zero, then $c_{\mathcal{P}^b}(p_1, \ldots, p_n) = 0$.

Next I show $c_{\mathcal{P}^b}(p_1, p_2, p_3)$ is constant with respect to permutations. Since $c_{\mathcal{P}^b}$ is constant with respect to permutation on vectors of length two, the definition of $c_{\mathcal{P}^b}$, and the fact that $c_{\mathcal{P}^b}(1, 0) = c_{\mathcal{P}^b}(0, 1) = 0$, implies $c_{\mathcal{P}^b}(p_1, p_2, p_3) = c_{\mathcal{P}^b}(p_1, p_3, p_2)$. Thus, if I show for any probability vector of length three that $c_{\mathcal{P}^b}(p_1, p_2, p_3) = c_{\mathcal{P}^b}(p_2, p_1, p_3)$, then $c_{\mathcal{P}^b}(p_1, p_2, p_3)$ is constant with respect to permutations since combinations of these two different pairwise permutations can achieve any permutation desired. This is easy to show since if $p_1 = 1$, or $p_2 = 1$, or $p_1 = p_2 = 0$, then I know this is true, and otherwise with Axiom 2 I know:

$$c_{\mathcal{P}^b}(p_1,\, p_2,\, p_3) = c_{\mathcal{P}^b}(p_1,\, 1-p_1) + (1-p_1)c_{\mathcal{P}^b}\Big(\frac{p_2}{1-p_1},\, \frac{1-p_1-p_2}{1-p_1}\Big)$$

$$= c_{\mathcal{P}^b}(p_2,\, 1-p_2) + (1-p_2)c_{\mathcal{P}^b}\Big(\frac{p_1}{1-p_2},\, \frac{1-p_1-p_2}{1-p_2}\Big) = c_{\mathcal{P}^b}(p_2,\, p_1,\, p_3).$$

Now assume that $c_{\mathcal{P}^b}$ is constant with respect to permutations on vectors of length $n \geq 3$, and I next show $c_{\mathcal{P}^b}$ is constant with respect to permutations on vectors of length $n+1$, and the proof is finished. If $p_1 + p_2 = 1$, then I am done. If not, notice that $c_{\mathcal{P}^b}(p_1,\, \ldots,\, p_{n+1}) = c_{\mathcal{P}^b}(p_1,\, 1-p_1) + (1-p_1)c_{\mathcal{P}^b}\Big(\frac{p_2}{1-p_1},\, \ldots,\, \frac{p_{n+1}}{1-p_1}\Big)$, whenever $p_1 \neq 1$, and as part of the inductive argument I assumed $c_{\mathcal{P}^b}$ was constant with respect to permutations on vectors of length n, so I only need to show $c_{\mathcal{P}^b}(p_1,\, p_2,\, \ldots,\, p_{n+1}) = c_{\mathcal{P}^b}(p_2,\, p_1,\, \ldots,\, p_{n+1})$, which is true:

$$c_{\mathcal{P}^b}(p_1,\, p_2,\, \ldots,\, p_{n+1}) = c_{\mathcal{P}^b}(p_1,\, 1-p_1) + (1-p_1)c_{\mathcal{P}^b}\Big(\frac{p_2}{1-p_1},\, \ldots,\, \frac{p_{n+1}}{1-p_1}\Big)$$

$$= c_{\mathcal{P}^b}(p_1,\, 1-p_1)+(1-p_1)c_{\mathcal{P}^b}\Big(\frac{p_2}{1-p_1},\, \frac{1-p_1-p_2}{1-p_1}\Big)+(1-p_1-p_2)c_{\mathcal{P}^b}\Big(\frac{p_3}{1-p_1-p_2},\, \ldots,\, \frac{p_{n+1}}{1-p_1-p_2}\Big)$$

$$= c_{\mathcal{P}^b}(p_1,\, p_2,\, 1-p_1-p_2) + (1-p_1-p_2)c_{\mathcal{P}^b}\Big(\frac{p_3}{1-p_1-p_2},\, \ldots,\, \frac{p_{n+1}}{1-p_1-p_2}\Big)$$

$$= c_{\mathcal{P}^b}(p_2,\, p_1,\, 1-p_1-p_2) + (1-p_1-p_2)c_{\mathcal{P}^b}\Big(\frac{p_3}{1-p_1-p_2},\, \ldots,\, \frac{p_{n+1}}{1-p_1-p_2}\Big)$$

$$= c_{\mathcal{P}^b}(p_2,\, 1-p_2)+(1-p_2)c_{\mathcal{P}^b}\Big(\frac{p_1}{1-p_2},\, \frac{1-p_1-p_2}{1-p_2}\Big)+(1-p_1-p_2)c_{\mathcal{P}^b}\Big(\frac{p_3}{1-p_1-p_2},\, \ldots,\, \frac{p_{n+1}}{1-p_1-p_2}\Big)$$

$$= c_{\mathcal{P}^b}(p_2,\, 1-p_2) + (1-p_2)c_{\mathcal{P}^b}\Big(\frac{p_1}{1-p_2},\, \ldots,\, \frac{p_{n+1}}{1-p_2}\Big) = c_{\mathcal{P}^b}(p_2,\, p_1,\, \ldots,\, p_{n+1}). \blacksquare$$

Lemma 12. Given a binary partition $\mathcal{P}^b$, define $c_{\mathcal{P}^b} : \cup_{j=1}^{\infty} \triangle^j \to \mathbb{R}$, where $\triangle^j$ is the j simplex, as in the statement of Lemma 11, and suppose C satisfies Axiom 1, and Axiom 2, then if $(q_1,\, \ldots,\, q_m)$ and $(p_1,\, \ldots,\, p_n)$ are two probability vectors (vectors of weakly positive numbers that sum to one with $1 < m < n$), such that each q_i is strictly positive, and can be written as the sum of one or more p_j with each p_j used once in the sum of only one q_i. Rename the p_j(s) assigned to each q_i so that $q_i = p^i_1 + \ldots p^i_{n_i}$. Then it is true that:

$$c_{\mathcal{P}^b}(p_1,\, \ldots,\, p_n) = c_{\mathcal{P}^b}(q_1,\, \ldots,\, q_m) + \sum_{i=1}^{m} q_i c_{\mathcal{P}^b}\Big(\frac{p^i_1}{q_i},\, \ldots,\, \frac{p^i_{n_i}}{q_i},\, 0\Big).$$

Proof. Given a binary partition $\mathcal{P}^b$, suppose C satisfies Axiom 1, and Axiom 2, that $c_{\mathcal{P}^b}$ is defined as in the statement of Lemma 11, and $(q_1, \ldots, q_m)$ and $(p_1, \ldots, p_n)$ are defined as in the statement of Lemma 12 (including the renaming of each p_j). I use the fact that the definition of $c_{\mathcal{P}^b}$ implies $c_{\mathcal{P}^b}(p_1, \ldots, p_n) = c_{\mathcal{P}^b}(p_1, \ldots, p_n, 0)$, and $c_{\mathcal{P}^b}(1, 0) = 0$, without reference. In Lemma 11 I showed $c_{\mathcal{P}^b}$ is constant with respect to permutations of vectors of arbitrary length (greater or equal to two). Thus, all I need to do is show:

$$c_{\mathcal{P}^b}(p_1, \ldots, p_{m-1}, p_m, \ldots, p_n) = c_{\mathcal{P}^b}(q_1, \ldots, q_m) + q_m c_{\mathcal{P}^b}\Big(\frac{p_m}{q_m}, \ldots, \frac{p_n}{q_m}, 0\Big),$$

where for $i \in \{1, \ldots m-1\}$ $q_i = p_i$, $1 < m < n$, and $q_m = p_m + \ldots + p_n > 0$. If $m = 2$, or $q_m = p_m$, this is trivially true. If $m > 2$ and $q_m > p_m$, then it is still true given the definition of $c_{\mathcal{P}^b}$ since (assuming without loss that $p_n > 0$):

$$\begin{aligned}
c_{\mathcal{P}^b}(p_1, \ldots, p_{m-1}, p_m, \ldots, p_n) &= C(\mathcal{P}^b, p_1, 1-p_1) + (1-p_1)C\Big(\mathcal{P}^b, \frac{p_2}{1-p_1}, \frac{1-p_1-p_2}{1-p_1}\Big) \\
&+ \ldots + (1-p_1-\ldots-p_{m-1})C\Big(\mathcal{P}^b, \frac{p_m}{1-p_1-\ldots-p_{m-1}}, \frac{1-p_1-\ldots-p_m}{1-p_1-\ldots-p_{m-1}}\Big) \\
&+(1-p_1-\ldots-p_m)C\Big(\mathcal{P}^b, \frac{p_{m+1}}{1-p_1-\ldots-p_m}, \frac{1-p_1-\ldots-p_m}{1-p_1-\ldots-p_{m-1}}\Big) \\
&+\cdots+(1-p_1-\ldots-p_{n-1})C\Big(\mathcal{P}^b, \frac{p_n}{1-p_1-\ldots-p_{n-1}}, \frac{1-p_1-\ldots-p_n}{1-p_1-\ldots-p_{m-1}}\Big) \\
&= c_{\mathcal{P}^b}(q_1, \ldots, q_m) + q_m c_{\mathcal{P}^b}\Big(\frac{p_m}{q_m}, \ldots, \frac{p_n}{q_m}, 0\Big). \blacksquare
\end{aligned}$$

I am now ready to resume the proof of Lemma 10. Given a binary partition $\mathcal{P}^b = \{A_1, A_2\}$, define $c_{\mathcal{P}^b} : \cup_{j=1}^{\infty} \triangle^j \to \mathbb{R}$, where $\triangle^j$ is the j simplex, as in the statement of Lemma 11, and suppose C satisfies Axiom 1, Axiom 2, and Axiom 3. Remember $C(\mathcal{P}^b, \mu) = c_{\mathcal{P}^b}(\mu(A_1), \mu(A_2))$ for all probability measures μ so Lemma 11 implies that $C(\mathcal{P}^b, p, 1-p) = C(\mathcal{P}^b, 1-p, p)$, for each $p \in [0, 1]$, and I thus only wish to show $c_{\mathcal{P}^b}(p, 1-p)$ is continuous for $p \in [0, 1]$. I proceed with a proof by contradiction: Suppose not, and $c_{\mathcal{P}^b}(p, 1-p)$ is discontinuous at some point $p = p_d \in [0, 1]$. Since $c_{\mathcal{P}^b}(p, 1-p) = c_{\mathcal{P}^b}(1-p, p)$, it is without loss to assume $p_d \in [0, \frac{1}{2}]$.

First, notice that if $c_{\mathcal{P}^b}(p, 1-p)$ is continuous at $p = 0$ then it is continuous at $p = \frac{1}{2}$: this is because Axiom 2 imposes that for small $\delta > 0$: $c_{\mathcal{P}^b}(\delta, \frac{1}{2} - \frac{\delta}{2}, \frac{1}{2} - \frac{\delta}{2}) = c_{\mathcal{P}^b}(\delta, 1-\delta) + (1-\delta)c_{\mathcal{P}^b}(1/2, 1/2) = c_{\mathcal{P}^b}(\frac{1}{2} - \frac{\delta}{2}, \frac{1}{2} + \frac{\delta}{2}) + (\frac{1}{2} + \frac{\delta}{2})c_{\mathcal{P}^b}(\frac{2\delta}{1+\delta}, \frac{1-\delta}{1+\delta})$. Since Axiom 3 requires that there is some $p_c \in [0, \frac{1}{2}]$ such that $c_{\mathcal{P}^b}(p, 1-p)$ is continuous at p_c, it is thus without loss to assume

$c_{\mathcal{P}^b}(p,\, 1-p)$ is continuous at $p_c \in (0,\, \frac{1}{2}]$.

Second, notice that it is not possible that the only $p \in [0,\, \frac{1}{2}]$ at which $c_{\mathcal{P}^b}(p,\, 1-p)$ is discontinuous is $p = 0$, because, if so, Axiom 2 once again imposes that for small $\delta > 0$: $c_{\mathcal{P}^b}(\delta,\, \frac{1}{2} - \frac{\delta}{2},\, \frac{1}{2} - \frac{\delta}{2}) = c_{\mathcal{P}^b}(\delta,\, 1-\delta) + (1-\delta)c_{\mathcal{P}^b}(1/2,\, 1/2) = c_{\mathcal{P}^b}(\frac{1}{2} - \frac{\delta}{2},\, \frac{1}{2} + \frac{\delta}{2}) + (\frac{1}{2} + \frac{\delta}{2})c_{\mathcal{P}^b}(\frac{2\delta}{1+\delta},\, \frac{1-\delta}{1+\delta})$, and either:

$$\limsup_{p \downarrow 0} c_{\mathcal{P}^b}(p,\, 1-p) = H < \infty \text{ (with } H > 0\text{) or } \limsup_{p \downarrow 0} c_{\mathcal{P}^b}(p,\, 1-p) = \infty.$$

If the former is true, then I can pick arbitrarily small $\delta \in (0,\, \frac{1}{4})$ to ensure that $c_{\mathcal{P}^b}(\delta,\, 1-\delta)$ is arbitrarily close to H, $c_{\mathcal{P}^b}(\frac{2\delta}{1+\delta},\, \frac{1-\delta}{1+\delta})$ is less than H or arbitrarily close to it, and $|(1-\delta)c_{\mathcal{P}^b}(1/2,\, 1/2) - c_{\mathcal{P}^b}(\frac{1}{2} - \frac{\delta}{2},\, \frac{1}{2} + \frac{\delta}{2})| < \frac{1}{8}H$, which creates a contradiction. If, instead, the latter is true, then I can pick arbitrarily small $\delta \in (0,\, \frac{1}{4})$ so that $c_{\mathcal{P}^b}(\delta,\, 1-\delta) \geq c_{\mathcal{P}^b}(p,\, 1-p) \forall\, p \in [\delta,\, \frac{1}{2}]$, and so that $|(1-\delta)c_{\mathcal{P}^b}(1/2,\, 1/2) - c_{\mathcal{P}^b}(\frac{1}{2} - \frac{\delta}{2},\, \frac{1}{2} + \frac{\delta}{2})| < \frac{1}{8}c_{\mathcal{P}^b}(\delta,\, 1-\delta)$, which again creates a contradiction as $\delta < \frac{2\delta}{1+\delta}$.

Third, if $c_{\mathcal{P}^b}(p,\, 1-p)$ is discontinuous at $p = \frac{1}{2}$ then it is discontinuous at a $p \in \{\frac{1}{4},\, \frac{1}{3}\}$ because Axiom 2 imposes that for small δ: $c_{\mathcal{P}^b}(\frac{1}{2} - \delta,\, \frac{1}{3} + \frac{2\delta}{3},\, \frac{1}{6} + \frac{\delta}{3}) = c_{\mathcal{P}^b}(\frac{1}{2} - \delta,\, \frac{1}{2} + \delta) + (\frac{1}{2} + \delta)c_{\mathcal{P}^b}(\frac{1}{3},\, \frac{2}{3}) = c_{\mathcal{P}^b}(\frac{1}{3} + \frac{2\delta}{3},\, \frac{2}{3} - \frac{2\delta}{3}) + (\frac{2}{3} - \frac{2\delta}{3})c_{\mathcal{P}^b}((\frac{1}{6} + \frac{\delta}{3})/(\frac{2}{3} - \frac{2\delta}{3}),\, (\frac{1}{2} - \delta)/(\frac{2}{3} - \frac{2\delta}{3}))$. Thus it is without loss to assume $c_{\mathcal{P}^b}(p,\, 1-p)$ is discontinuous at $p_d \in (0,\, \frac{1}{2})$ (given second and third point).

It is not possible for $c_{\mathcal{P}^b}(p,\, 1-p)$ to be continuous at $p_c \in (0,\, \frac{1}{2}]$ and discontinuous at $p_d \in (0,\, \frac{1}{2})$, however, as if I assume this is the case I can reach a contradiction, beginning by picking $(p_1,\, p_2,\, p_3,\, p_4)$ such that they sum to one and:

$$p_1 + p_2 = p_d,\ \frac{p_1}{p_1 + p_2} = p_c,\ \text{ and } \frac{p_4}{p_3 + p_4} = p_c,$$

$$\text{so that as a result } p_1 + p_4 = p_c,\ \frac{p_1}{p_1 + p_4} = p_d,\ \text{ and } \frac{p_2}{p_2 + p_3} = p_d.$$

How these four probabilities are selected is quite important, and this is where a lot of the magic happens. Now, notice Lemma 12 tells us:

$$c_{\mathcal{P}^b}(p_1,\, p_2,\, p_3,\, p_4)$$

$$= c_{\mathcal{P}^b}(p_1 + p_2,\, p_3 + p_4) + (p_1 + p_2)c_{\mathcal{P}^b}\left(\frac{p_1}{p_1 + p_2},\, \frac{p_2}{p_1 + p_2}\right) + (p_3 + p_4)c_{\mathcal{P}^b}\left(\frac{p_3}{p_3 + p_4},\, \frac{p_4}{p_3 + p_4}\right)$$

$$= c_{\mathcal{P}^b}(p_1 + p_4,\, p_2 + p_3) + (p_1 + p_4)c_{\mathcal{P}^b}\left(\frac{p_1}{p_1 + p_4},\, \frac{p_4}{p_1 + p_4}\right) + (p_2 + p_3)c_{\mathcal{P}^b}\left(\frac{p_2}{p_2 + p_3},\, \frac{p_3}{p_2 + p_3}\right).$$

Substituting in terms using the definitions of the four probabilities it is then clear that:

$$c_{\mathcal{P}^b}(p_d,\, 1-p_d) + (p_d)c_{\mathcal{P}^b}(p_c,\, 1-p_c) + (1-p_d)c_{\mathcal{P}^b}(1-p_c,\, p_c)$$

$$= c_{\mathcal{P}^b}(p_c,\, 1-p_c) + (p_c)c_{\mathcal{P}^b}(p_d,\, 1-p_d) + (1-p_c)c_{\mathcal{P}^b}(p_d,\, 1-p_d).$$

Next, $c_{\mathcal{P}^b}$ is discontinuous from both sides at p_d if it is discontinuous at p_d since I can increase p_1 and p_3 by a small $\delta > 0$, and decrease p_2 and p_4 by the same δ, and as δ is taken to zero, continuity at p_c implies the change in $c_{\mathcal{P}^b}(p_1+p_2,\, p_3+p_4)+(p_1+p_2)c_{\mathcal{P}^b}\Big(\dfrac{p_1}{p_1+p_2},\, \dfrac{p_2}{p_1+p_2}\Big)+(p_3+p_4)c_{\mathcal{P}^b}\Big(\dfrac{p_3}{p_3+p_4},\, \dfrac{p_4}{p_3+p_4}\Big)$ goes to zero, so discontinuities at either side of p_d must offset each other so the change in $c_{\mathcal{P}^b}(p_1+p_4,\, p_2+p_3)+(p_1+p_4)c_{\mathcal{P}^b}\Big(\dfrac{p_1}{p_1+p_4},\, \dfrac{p_4}{p_1+p_4}\Big)+(p_2+p_3)c_{\mathcal{P}^b}\Big(\dfrac{p_2}{p_2+p_3},\, \dfrac{p_3}{p_2+p_3}\Big)$ goes to zero with δ.

Next, I show that it cannot be that:

$$\limsup_{p\downarrow p_d} c_{\mathcal{P}^b}(p,\, 1-p) = H > c_{\mathcal{P}^b}(p_d,\, 1-p_d).$$

There are two cases of interest, and in both I create a contradiction. In case one $H < \infty$. Case one is not possible, however, since I can choose arbitrarily small $\delta > 0$ and add it to p_1 and subtract it from p_4 so that $c_{\mathcal{P}^b}(p_1+p_2,\, p_3+p_4)$ is arbitrarily close to H, while $c_{\mathcal{P}^b}\Big(\dfrac{p_1}{p_1+p_4},\, \dfrac{p_4}{p_1+p_4}\Big)$ is less than H or arbitrarily close to H, and all other terms remain essentially constant, creating a contradiction. In case two $H = \infty$. Case two is also not possible, however, since I can choose arbitrarily small $\delta > 0$ and add it to p_1 and p_3 and subtract it from p_2 and p_4 so that $c_{\mathcal{P}^b}\Big(\dfrac{p_1}{p_1+p_4},\, \dfrac{p_4}{p_1+p_4}\Big)$ is arbitrarily close to ∞, while, other than $c_{\mathcal{P}^b}\Big(\dfrac{p_2}{p_2+p_3},\, \dfrac{p_3}{p_2+p_3}\Big)$, all other terms remain essentially constant. This then implies that $c_{\mathcal{P}^b}\Big(\dfrac{p_2}{p_2+p_3},\, \dfrac{p_3}{p_2+p_3}\Big)$ drops by an arbitrarily large amount, which is not possible since it is positive by definition. Thus, discontinuity on both sides of p_d requires:

$$\liminf_{p\downarrow p_d} c_{\mathcal{P}^b}(p,\, 1-p) = L < c_{\mathcal{P}^b}(p_d,\, 1-p_d).$$

I am now ready for the final contradiction as L must be positive. Increase p_1 and decrease p_4 by an arbitrarily small $\delta > 0$, keeping p_2 and p_3 constant, so that $c_{\mathcal{P}^b}(p_1+p_2,\, p_3+p_4)$ is arbitrarily close to L. Then it is easy to see the contradiction using Lemma 12 as in the previous paragraphs since $c_{\mathcal{P}^b}\Big(\dfrac{p_1}{p_1+p_4},\, \dfrac{p_4}{p_1+p_4}\Big)$ is more than L or arbitrarily close to it, and all other terms remain essentially constant. ■

Continuity and symmetry (invariance with respect to permutations) are not the only helpful properties imposed onto the cost function by the axioms. On binary partitions, the cost function is also non-decreasing if the probability of whichever event is less likely increases.

Lemma 13. If C satisfies Axiom 1, Axiom 2, and Axiom 3, then for each binary partition $\mathcal{P}^b$, and for each $p \in [0, \frac{1}{2})$, $C(\mathcal{P}^b, p, 1-p)$ is non-decreasing for small increases in p, which means that there exits $\theta > 0$ such that if $\gamma < \theta$ then $C(\mathcal{P}^b, p, 1-p) \leq C(\mathcal{P}^b, p+\gamma, 1-p-\gamma)$.

Proof. Given a binary partition $\mathcal{P}^b = \{A_1, A_2\}$, define $c_{\mathcal{P}^b} : \cup_{j=1}^{\infty} \triangle^j \to \mathbb{R}$, where $\triangle^j$ is the j simplex, as in the statement of Lemma 11, and suppose C satisfies Axiom 1, Axiom 2, and Axiom 3. Remember Lemma 11 implies that $c_{\mathcal{P}^b}(0, 1) = 0$, so I only need to show $c_{\mathcal{P}^b}(p, 1-p)$ is non-decreasing for small increases to $p \in (0, 1/2)$.

I proceed by assuming there is a $p_d \in (0, 1/2)$ such that $c_{\mathcal{P}^b}(p_d, 1-p_d)$ is decreasing for small increases in p_d and create a contradiction. First, notice that since Lemma 10 shows $c_{\mathcal{P}^b}(p, 1-p)$ is continuous and Lemma 11 shows $c_{\mathcal{P}^b}(0, 1) = 0$ that it must be that before any p where $c_{\mathcal{P}^b}(p, 1-p)$ is locally decreasing in p there must be a smaller p where $c_{\mathcal{P}^b}(p, 1-p)$ is locally increasing in p. Second, notice that there must be infinitely many $p \in (0, 1/2)$ where $c_{\mathcal{P}^b}(p, 1-p)$ decreases for small increases to p because if $p_d \in (0, 1/2)$ is such that $c_{\mathcal{P}^b}(p_d, 1-p_d)$ decreases for small increases to p_d I can pick (p_1, p_2, p_3, p_4) such that:

$$p_1 + p_2 = p_d, \ \frac{p_1}{p_1+p_2} = p_d, \ \frac{p_3}{p_3+p_4} = p_d, \text{ so that } \frac{p_1}{p_1+p_4} < p_d,$$

and then notice Lemma 12 tells us:

$$\begin{aligned} &c_{\mathcal{P}^b}(p_1, p_2, p_3, p_4) \\ &= c_{\mathcal{P}^b}(p_1+p_2, p_3+p_4) + (p_1+p_2)c_{\mathcal{P}^b}\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right) + (p_3+p_4)c_{\mathcal{P}^b}\left(\frac{p_3}{p_3+p_4}, \frac{p_4}{p_3+p_4}\right) \\ &= c_{\mathcal{P}^b}(p_1+p_4, p_2+p_3) + (p_1+p_4)c_{\mathcal{P}^b}\left(\frac{p_1}{p_1+p_4}, \frac{p_4}{p_1+p_4}\right) + (p_2+p_3)c_{\mathcal{P}^b}\left(\frac{p_2}{p_2+p_3}, \frac{p_3}{p_2+p_3}\right), \end{aligned}$$

and then consider increasing p_1 a small amount and decreasing p_4 by the same small amount, while keeping p_2 and p_3 constant, and notice this implies $c_{\mathcal{P}^b}(p, 1-p)$ decreases for small increases to $p = p_1/(p_1+p_4) < p_d$. This all means $c_{\mathcal{P}^b}(p, 1-p)$ has dense local maxima and minima for p close to zero.

Next, I show that the largest reduction in $c_{\mathcal{P}^b}(p, 1-p)$ from an increase in p of any particular small $\epsilon > 0$ must be at achieved at a $p > 1/4$. Pick $p_1 \leq 1/4$ such that $c_{\mathcal{P}^b}$ is decreasing there for an increases in p_1 of $\epsilon > 0$. Given $\epsilon > 0$, pick p_2 and p_3 so that $p_1 + p_2 + p_3 = 1$, and so:

$$\frac{p_3}{p_2 + p_3} = \frac{p_2 - \epsilon}{p_2 - \epsilon + p_3}.$$

Since ϵ is small and $p_1 \leq 1/4$, I know $p_1 < p_3 < p_2$. Pick $k \geq 0$ so:

$$k = c_{\mathcal{P}^b}\Big(\frac{p_3}{p_2 + p_3},\, 1 - \frac{p_3}{p_2 + p_3}\Big) = c_{\mathcal{P}^b}\Big(\frac{p_2 - \epsilon}{p_2 - \epsilon + p_3},\, 1 - \frac{p_2 - \epsilon}{p_2 - \epsilon + p_3}\Big).$$

Lemma 11 and Lemma 12 tell us:

$$c_{\mathcal{P}^b}(p_1,\, p_2,\, p_3) = c_{\mathcal{P}^b}(p_3,\, 1 - p_3) + (1 - p_3)c_{\mathcal{P}^b}\Big(\frac{p_1}{p_1 + p_2},\, \frac{p_2}{p_1 + p_2}\Big)$$

$$= c_{\mathcal{P}^b}(p_1,\, 1 - p_1) + (1 - p_1)c_{\mathcal{P}^b}\Big(\frac{p_2}{p_2 + p_3},\, \frac{p_3}{p_2 + p_3}\Big).$$

So, if I increase p_1 by ϵ and decrease p_2 by ϵ, the change in $c_{\mathcal{P}^b}(p_1,\, p_2,\, p_3)$ is:

$$(1 - p_3)\Bigg(c_{\mathcal{P}^b}\Big(\frac{p_1 + \epsilon}{p_1 + p_2},\, \frac{p_2 - \epsilon}{p_1 + p_2}\Big) - c_{\mathcal{P}^b}\Big(\frac{p_1}{p_1 + p_2},\, \frac{p_2}{p_1 + p_2}\Big)\Bigg)$$

$$= c_{\mathcal{P}^b}(p_1 + \epsilon,\, 1 - (p_1 + \epsilon)) - c_{\mathcal{P}^b}(p_1,\, 1 - p_1) - \epsilon k < 0.$$

This implies:

$$\frac{c_{\mathcal{P}^b}\Big(\frac{p_1}{p_1 + p_2} + \frac{\epsilon}{p_1 + p_2},\, \frac{p_2}{p_1 + p_2} - \frac{\epsilon}{p_1 + p_2}\Big) - c_{\mathcal{P}^b}\Big(\frac{p_1}{p_1 + p_2},\, \frac{p_2}{p_1 + p_2}\Big)}{\frac{\epsilon}{p_1 + p_2}}$$

$$\leq \frac{c_{\mathcal{P}^b}(p_1 + \epsilon,\, 1 - (p_1 + \epsilon)) - c_{\mathcal{P}^b}(p_1,\, 1 - p_1)}{\epsilon} < 0$$

Thus, at

$$\frac{p_1}{p_1 + p_2} > p_1 \,(\text{notice that for small } \epsilon : \frac{p_1}{p_1 + p_2} < \frac{1}{2}),$$

$c_{\mathcal{P}^b}$ is averaging a weakly steeper descent over a longer range, and thus there must be a point between

$$\frac{p_1}{p_1 + p_2} \text{ and } \frac{p_1 + \epsilon}{p_1 + p_2} \,(\text{notice that for small } \epsilon : \frac{p_1 + \epsilon}{p_1 + p_2} < \frac{1}{2}),$$

where the decrease of $c_{\mathcal{P}^b}$ over the next ϵ is as large as the decrease $c_{\mathcal{P}^b}(p_1 + \epsilon,\, 1 - (p_1 + \epsilon)) - c_{\mathcal{P}^b}(p_1,\, 1 - p_1)$. When p_1 is close to $1/4$, if I pick p_2 and p_3 as above, keeping our small ϵ in mind, I have:

$$\frac{p_1}{p_1 + p_2} > \frac{1}{4}.$$

$c_{\mathcal{P}^b}$ is a continuous function, so for all small $\epsilon > 0$, $f(p) = c_{\mathcal{P}^b}(p + \epsilon,\, 1 - (p + \epsilon)) - c_{\mathcal{P}^b}(p,\, 1 - p)$, defined for compact domain $p \in [0,\, \frac{1}{2} - \epsilon]$, is continuous, and has a minimizer (perhaps not unique) $p_s(\epsilon) \in (\frac{1}{4},\, \frac{1}{2} - \epsilon)$, given what I just showed.

I am now ready to create the desired contradiction. I begin by finding a local maximum, denote it p_m, such that $p_m \in (0,\, 1/1000)$, and an $\epsilon \in (0,\, 1/1000)$, such that if $\delta \in [0,\, \epsilon]$, then:

$$c_{\mathcal{P}^b}(p_m,\, 1 - p_m) > c_{\mathcal{P}^b}(p_m + 4\delta,\, 1 - (p_m + 4\delta)).$$

Now, let $p_2 = p_s(\epsilon) + \epsilon > 1/4 + \epsilon$, and let:

$$p_3 = \frac{p_2}{1 - p_m} p_m < p_m, \text{ so that } \frac{p_3}{p_2 + p_3} = p_m.$$

Finlly, let $p_1 = 1 - p_2 - p_3$, noticing $p_1 > 1/4$, so:

$$\frac{p_3}{p_1 + p_3} + \frac{\epsilon}{p_1 + p_3 + \epsilon} < \frac{1}{2}.$$

Lemma 12 tells us:

$$c_{\mathcal{P}^b}(p_1,\, p_2,\, p_3) = c_{\mathcal{P}^b}(p_1,\, 1 - p_1) + (1 - p_1) c_{\mathcal{P}^b}\Big(\frac{p_2}{p_2 + p_3},\, \frac{p_3}{p_2 + p_3}\Big)$$

$$= c_{\mathcal{P}^b}(p_2,\, 1 - p_2) + (1 - p_2) c_{\mathcal{P}^b}\Big(\frac{p_1}{p_1 + p_3},\, \frac{p_3}{p_1 + p_3}\Big).$$

This means, since $p_2 + p_3 > 1/4$, if I increase p_3 by ϵ, and decrease p_2 by ϵ, holding p_1 constant, and consider the change in $c_{\mathcal{P}^b}(p_1,\, p_2,\, p_3)$:

$$0 > (1 - p_1)\Big(c_{\mathcal{P}^b}\Big(\frac{p_3 + \epsilon}{p_2 + p_3},\, \frac{p_2 - \epsilon}{p_2 + p_3}\Big) - c_{\mathcal{P}^b}\Big(\frac{p_3}{p_2 + p_3},\, \frac{p_2}{p_2 + p_3}\Big)\Big)$$

$$= c_{\mathcal{P}^b}(p_2 - \epsilon,\, 1 - (p_2 - \epsilon)) - c_{\mathcal{P}^b}(p_2,\, 1 - p_2)$$

$$+ (p_1 + p_3 + \epsilon) c_{\mathcal{P}^b}\Big(\frac{p_3 + \epsilon}{p_1 + p_3 + \epsilon},\, \frac{p_1}{p_1 + p_3 + \epsilon}\Big) - (p_1 + p_3) c_{\mathcal{P}^b}\Big(\frac{p_3}{p_1 + p_3},\, \frac{p_1}{p_1 + p_3}\Big)$$

$$\geq c_{\mathcal{P}^b}(p_2 - \epsilon,\, 1 - (p_2 - \epsilon)) - c_{\mathcal{P}^b}(p_2,\, 1 - p_2)$$

$$+(p_1+p_3+\epsilon)\Big(c_{\mathcal{P}^b}\Big(\frac{p_3}{p_1+p_3+\epsilon}+\frac{\epsilon}{p_1+p_3+\epsilon},\,\frac{p_1}{p_1+p_3+\epsilon}\Big)-c_{\mathcal{P}^b}\Big(\frac{p_3}{p_1+p_3},\,\frac{p_1}{p_1+p_3}\Big)\Big).$$

This implies:

$$0>\frac{c_{\mathcal{P}^b}(p_s(\epsilon)+\epsilon,\,1-(p_s(\epsilon)+\epsilon))-c_{\mathcal{P}^b}(p_s(\epsilon),\,1-p_s(\epsilon))}{\epsilon}$$

$$>\frac{c_{\mathcal{P}^b}\Big(\frac{p_3}{p_1+p_3+\epsilon}+\frac{\epsilon}{p_1+p_3+\epsilon},\,\frac{p_1}{p_1+p_3+\epsilon}\Big)-c_{\mathcal{P}^b}\Big(\frac{p_3}{p_1+p_3},\,\frac{p_1}{p_1+p_3}\Big)}{\frac{\epsilon}{p_1+p_3+\epsilon}}.$$

But remember, the way I picked $p_s(\epsilon)$ implies for all $\delta\in\Big[\epsilon,\,\frac{\epsilon}{p_1+p_3+\epsilon}\Big]$:

$$\frac{c_{\mathcal{P}^b}(p_s(\epsilon)+\epsilon,\,1-(p_s(\epsilon)+\epsilon))-c_{\mathcal{P}^b}(p_s(\epsilon),\,1-p_s(\epsilon))}{\epsilon}$$

$$\leq\frac{c_{\mathcal{P}^b}\Big(\frac{p_3}{p_1+p_3}+\delta,\,\frac{p_1}{p_1+p_3}-\delta\Big)-c_{\mathcal{P}^b}\Big(\frac{p_3}{p_1+p_3},\,\frac{p_1}{p_1+p_3}\Big)}{\delta},$$

so letting $\delta=\frac{\epsilon}{p_1+p_3+\epsilon}\frac{p_1}{p_1+p_3}\in\Big[\epsilon,\,\frac{\epsilon}{p_1+p_3+\epsilon}\Big]$:

$$\frac{c_{\mathcal{P}^b}(p_s(\epsilon)+\epsilon,\,1-(p_s(\epsilon)+\epsilon))-c_{\mathcal{P}^b}(p_s(\epsilon),\,1-p_s(\epsilon))}{\epsilon}$$

$$\leq\frac{c_{\mathcal{P}^b}\Big(\frac{p_3}{p_1+p_3}+\frac{\epsilon}{p_1+p_3+\epsilon}\frac{p_1}{p_1+p_3},\,\frac{p_1}{p_1+p_3}-\frac{\epsilon}{p_1+p_3+\epsilon}\frac{p_1}{p_1+p_3}\Big)-c_{\mathcal{P}^b}\Big(\frac{p_3}{p_1+p_3},\,\frac{p_1}{p_1+p_3}\Big)}{\frac{\epsilon}{p_1+p_3+\epsilon}\frac{p_1}{p_1+p_3}}$$

$$=\frac{c_{\mathcal{P}^b}\Big(\frac{p_3}{p_1+p_3+\epsilon}+\frac{\epsilon}{p_1+p_3+\epsilon},\,\frac{p_1+\epsilon}{p_1+p_3+\epsilon}-\frac{\epsilon}{p_1+p_3+\epsilon}\Big)-c_{\mathcal{P}^b}\Big(\frac{p_3}{p_1+p_3},\,\frac{p_1}{p_1+p_3}\Big)}{\frac{\epsilon}{p_1+p_3+\epsilon}\frac{p_1}{p_1+p_3}}$$

$$<\frac{c_{\mathcal{P}^b}\Big(\frac{p_3}{p_1+p_3+\epsilon}+\frac{\epsilon}{p_1+p_3+\epsilon},\,\frac{p_1}{p_1+p_3+\epsilon}\Big)-c_{\mathcal{P}^b}\Big(\frac{p_3}{p_1+p_3},\,\frac{p_1}{p_1+p_3}\Big)}{\frac{\epsilon}{p_1+p_3+\epsilon}},$$

which establishes the desired contradiction. ■

I now show that the cost of learning the realized event of a learning strategy invariant partition is dictated by Shannon Entropy.

Lemma 14. If a partition $\mathcal{P}$ is learning strategy invariant, and C satisfies Axiom 1, Axiom 2, and Axiom 3, then there exists a multiplier $\lambda(\mathcal{P})\in\mathbb{R}_+$, such that for all probability measures μ: $C(\mathcal{P},\,\mu)=\lambda(\mathcal{P})\mathcal{H}(\mathcal{P},\,\mu)$, where $\mathcal{H}$ is Shannon's standard measure of entropy (1948), defined in equation (3).

Proof. Assume C satisfies Axiom 1, Axiom 2, and Axiom 3. Given learning strategy invariant partition $\mathcal{P} = \{A_1, \ldots, A_m\}$ pick any binary partition $\mathcal{P}^b$ coarser than $\mathcal{P}$ and define $c_{\mathcal{P}^b} : \cup_{j=1}^{\infty} \triangle^j \to \mathbb{R}$, where $\triangle^j$ is the j simplex, as in the statement of Lemma 11 so that, by Lemma 8, $C(\mathcal{P}, \mu) = c_{\mathcal{P}^b}(\mu(A_1), \ldots, \mu(A_m))$.

I begin by showing that if there is a $p \in (0, \frac{1}{2})$ such that $c_{\mathcal{P}^b}(p, 1-p) = 0$, then $c_{\mathcal{P}^b}(p, 1-p) = 0 \, \forall p \in (0, \frac{1}{2}]$. Assume there is $p \in [0, \frac{1}{2})$ that is the largest number less than $\frac{1}{2}$ such that $c_{\mathcal{P}^b}(p, 1-p) = 0$ (so $c_{\mathcal{P}^b}(\frac{1}{2}, \frac{1}{2}) > 0$), let $p_1 = p_2 = p$, and let $p_3 = 1 - p_1 - p_2$. Lemma 11 and Lemma 12 imply that: $c_{\mathcal{P}^b}(p_1, p_2, p_3) =$

$$c_{\mathcal{P}^b}(p_1, 1-p) + (1-p_1)c_{\mathcal{P}^b}\left(\frac{p_2}{p_2+p_3}, \frac{p_3}{p_2+p_3}\right) = c_{\mathcal{P}^b}(p_3, 1-p_3) + (1-p_3)c_{\mathcal{P}^b}\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2}\right).$$

This and Lemma 10 and Lemma 13 imply that $p_3 \geq \frac{1}{3}$. But if $p_1 > 0$, then decreasing p_1 and increase p_2 by the same arbitrarily small $\epsilon > 0$ results in a contradiction by Lemma 10 and Lemma 13 since $\dfrac{p_2}{p_2+p_3} > p_1$, so:

$$c_{\mathcal{P}^b}(p_1 - \epsilon, 1-(p_1-\epsilon)) + (1-(p_1-\epsilon))c_{\mathcal{P}^b}\left(\frac{p_2+\epsilon}{p_2+\epsilon+p_3}, \frac{p_3}{p_2+\epsilon+p_3}\right)$$

$$> c_{\mathcal{P}^b}(p_3, 1-p_3) + (1-p_3)c_{\mathcal{P}^b}\left(\frac{p_1-\epsilon}{p_1+p_2}, \frac{p_2+\epsilon}{p_1+p_2}\right).$$

Thus, p_1 cannot be strictly positive, and it must be that $c_{\mathcal{P}^b}(p, 1-p) > 0$ for $p \in (0, \frac{1}{2})$ if $c_{\mathcal{P}^b}(\frac{1}{2}, \frac{1}{2}) > 0$. So, if $\exists p \in (0, \frac{1}{2}]$ such that $c_{\mathcal{P}^b}(p, 1-p) = 0$, then: $C(\mathcal{P}, \mu) = 0 = 0\mathcal{H}(\mathcal{P}, \mu)$.

For the rest of the proof I assume $c_{\mathcal{P}^b}(p, 1-p) > 0 \; \forall p \in (0, \frac{1}{2}]$. Define h so that for $n \in \mathbb{N}$, $h(n) \equiv c_{\mathcal{P}^b}(1/n, \ldots, 1/n, 0)$. Since I assumed, $c_{\mathcal{P}^b}(p, 1-p) > 0 \; \forall p \in (0, \frac{1}{2}]$, $h(2) > h(1) = 0$, and in general $h(n) > 0$ if $n > 1$. It is also easy to show $h(n+1) > h(n)$ for all $n \geq 2$ using Lemma 12 and Lemma 13:

$$h(n) = c_{\mathcal{P}^b}(1/n, \ldots, 1/n, 0) = c_{\mathcal{P}^b}(1/n, \ldots, 1/n) + \left(\frac{1}{n}\right)c_{\mathcal{P}^b}\left(\frac{1/n}{1/n}, \frac{0}{1/n}\right)$$

$$< c_{\mathcal{P}^b}(1/n, \ldots, 1/n) + \left(\frac{1}{n}\right)c_{\mathcal{P}^b}\left(\frac{\frac{1}{(n+1)}}{\frac{1}{n}}, \frac{\frac{1}{n(n+1)}}{\frac{1}{n}}\right)$$

$$= c_{\mathcal{P}^b}(1/n, \ldots, 1/n, 1/n, 1/(n+1), 1/(n(n+1))) = c_{\mathcal{P}^b}(1/n, \ldots, 1/n, 1/(n+1), 1/n, 1/(n(n+1)))$$

$$= c_{\mathcal{P}^b}(1/n, \ldots, 1/n, 1/(n+1), (1/n) + 1/(n(n+1))) + \frac{n+2}{n(n+1)}c_{\mathcal{P}^b}\left(\frac{\frac{1}{n}}{\frac{n+2}{n(n+1)}}, \frac{\frac{1}{n(n+1)}}{\frac{n+2}{n(n+1)}}\right)$$

$$\leq c_{\mathcal{P}^b}(1/n,\, \ldots,\, 1/n,\, 1/(n+1),\, (1/n) + 1/(n(n+1))) + \frac{n+2}{n(n+1)} c_{\mathcal{P}^b}\Big(\tfrac{\frac{1}{n+1}}{\frac{n+2}{n(n+1)}},\, \tfrac{\frac{2}{n(n+1)}}{\frac{n+2}{n(n+1)}}\Big)$$

$$\leq \cdots \leq c_{\mathcal{P}^b}(1/(n+1),\, \ldots,\, 1/(n+1),\, 0) = h(n+1).$$

The rest of the proof follows the work of Shannon (1948) closely. Notice $h(s^r) = r \cdot h(s)$, which is reminiscent of logarithms, and is some nice foreshadowing for the rest of the proof. Given arbitrarily small $\epsilon > 0$, and integers $s > 1$ and $t > 1$, pick n and r so that $2/n < \epsilon$, and $s^r \leq t^n < s^{r+1}$. So:

$$r\log(s) \leq n\log(t) < (r+1)\log(s) \Longrightarrow \frac{r}{n} \leq \frac{\log(t)}{\log(s)} < \frac{r+1}{n} \Longrightarrow \Big|\frac{r}{n} - \frac{\log(t)}{\log(s)}\Big| < \frac{1}{n}.$$

The work I did above then tells us:

$$h(s^r) \leq h(t^n) \leq h(s^{r+1}) \Longrightarrow r \cdot h(s) \leq n \cdot h(t) \leq (r+1)h(s)$$

$$\Longrightarrow \frac{r}{n} \leq \frac{h(t)}{h(s)} \leq \frac{r+1}{n} \Longrightarrow \Big|\frac{r}{n} - \frac{h(t)}{h(s)}\Big| \leq \frac{1}{n}.$$

All of this tells us:

$$\Big|\frac{h(t)}{h(s)} - \frac{\log(t)}{\log(s)}\Big| < \epsilon,$$

which can be shown to be true $\forall \epsilon > 0$, and thus $h(n) = \lambda \log(n)$, where λ must be a positive constant.

Let $p_k = \mu(A_k)$ for each $A_k \in \mathcal{P}$. Suppose, for now, that each p_k is a rational number. Then there exists integers $n_1, \ldots, n_m$, such that for all $k \in \{1, \ldots, m\}$ I have:

$$p_k = \frac{n_k}{\sum\limits_{j=1}^{m} n_j}.$$

The interpretation is that I have a uniform distribution over $\sum\limits_j n_j$ equally likely states, and the probability of the event which happens with probability p_k is the probability of one of the n_k associated states occurring. Then using the definition of learning strategy invariance:

$$c_{\mathcal{P}^b}\left(\frac{1}{\sum\limits_j n_j},\, \ldots,\, \frac{1}{\sum\limits_j n_j}\right) = h\left(\sum_{j=1}^{m} n_j\right) = \lambda \log\left(\sum_{j=1}^{m} n_j\right) = c_{\mathcal{P}^b}(p_1,\, \ldots,\, p_m) + \sum_{j=1}^{m} p_j \lambda_i \log(n_j),$$

$$\implies c_{\mathcal{P}^b}(p_1, \ldots, p_m) = \lambda \log\left(\sum_{j=1}^{m} n_j\right) - \sum_{j=1}^{m} p_j \lambda \log(n_j)$$

$$= \sum_{k=1}^{m}\left(p_k \lambda \log\left(\sum_{j=1}^{m} n_j\right)\right) - \sum_{j=1}^{m} p_j \lambda \log(n_j)$$

$$= -\sum_{k=1}^{m} p_k \lambda \log\left(\frac{n_k}{\sum_j n_j}\right) = -\lambda \sum_{k=1}^{m} p_k \log(p_k) = \lambda \mathcal{H}(\mathcal{P},\, \mu),$$

where $\mathcal{H}$ is defined as in equation (3). If any of the p_i are irrational, then the density of the rationals and Lemma 10 can be used to get the same result. Thus, $C(\mathcal{P},\, \mu) = \lambda \mathcal{H}(\mathcal{P},\, \mu)$. ■

Underlying each learning strategy invariant partition is some attribute of the choice environment. Shannon (1948) imposes learning strategy invariance onto all partitions of Ω with his third axiom, which implies that all partitions have the same costs associated with them (there is a $\lambda > 0$ such that $\lambda(\mathcal{P}) = \lambda$ for all partitions $\mathcal{P}$ of Ω), and so it is without loss to think of the agent as learning about a single attribute that allows them to differentiate between the different states of the world. With MASE, in contrast, different learning strategy invariant partitions are allowed to have different costs associated with them ($\lambda(\mathcal{P})$ may differ depending on the learning strategy invariant partition $\mathcal{P}$), and thus it is natural to think of the agent as learning about different attributes of the choice environment depending on which attribute allows them to acquire the information at the lowest costs, as is formalized by Theorem 5. This interpretation is how MASE gets its name.

In addition to his learning strategy invariance axiom, Shannon has two other axioms, one of which imposes continuity onto his cost function (his axiom 1), and another that deals with the cost of differentiating between a greater number of equally likely states (his axiom 2) (Shannon, 1948). As it turns out, there is a great deal of redundancy in Shannon's axioms, as is demonstrated by this paper's axioms.

As a result, Shannon's third axiom is the only axiom that it is substantive to relax. Shannon's second axiom does not have any impact as long as learning with binary partitions is assumed to be costly when there is uncertainty about their realized event. Removing his first axiom only has an impact if I allow for a cost function that is discontinuous at every point when applied to a binary partition, which would render it too complex and intractable for practical application. As a result, if one wishes to generalize Shannon Entropy to achieve a more flexible but still tractable tool with which to study an environment where learning is costly, it must be Shannon's third axiom that is weakened.

I wish to study a costly learning environment so, to ease exposition slightly, Axiom 4 imposes that answering yes or no questions is costly to the agent.[18]

Axiom 4 (Costly Learning): Given a binary partition $\mathcal{P}^b$, $C(\mathcal{P}^b,\, \frac{1}{2},\, \frac{1}{2}) > 0$.

Total Uncertainty

Lemma 14 and Axiom 4 together imply that for each binary partition $\mathcal{P}^b$, there is an **associated multiplier**, $\lambda(\mathcal{P}^b) \in \mathbb{R}_{++}$, such that for all probability measures μ: $C(\mathcal{P}^b,\, \mu) = \lambda(\mathcal{P}^b)\mathcal{H}(\mathcal{P}^b,\, \mu)$. Since there are a finite number of binary partitions of Ω, I can order the binary partitions by their associated multipliers. Let λ_1 denote the multiplier associated with all binary partitions, denoted $\{\mathcal{P}_i^{b,\lambda_1}\}_{i=1}^{n_1}$, with the lowest multiplier.

If the agent can always learn the state of the world by asking questions with multiplier λ_1, then $\sigma(\{\mathcal{P}_i^{b,\lambda_1}\}_{i=1}^{n_1}) = \mathcal{F}$, and M=1.[19] If not, let λ_2 denote the multiplier associated with all binary partitions, denoted $\{\mathcal{P}_i^{b,\lambda_2}\}_{i=1}^{n_2}$, with the second lowest multiplier such that $\sigma(\{\mathcal{P}_i^{b,\lambda_1}\}_{i=1}^{n_1},\, \{\mathcal{P}_i^{b,\lambda_2}\}_{i=1}^{n_2}) \neq \sigma(\{\mathcal{P}_i^{b,\lambda_1}\}_{i=1}^{n_1})$.

If the agent can always learn the state of the world by asking questions with multipliers λ_1 or λ_2, then $\sigma(\{\mathcal{P}_i^{b,\lambda_1}\}_{i=1}^{n_1},\, \{\mathcal{P}_i^{b,\lambda_2}\}_{i=1}^{n_2}) = \mathcal{F}$, and $M = 2$. If not, let λ_3 denote the multiplier associated with all binary partitions, denoted $\{\mathcal{P}_i^{b,\lambda_3}\}_{i=1}^{n_3}$, with the third lowest multiplier such that $\sigma(\{\mathcal{P}_i^{b,\lambda_1}\}_{i=1}^{n_1},\, \{\mathcal{P}_i^{b,\lambda_2}\}_{i=1}^{n_2},\, \{\mathcal{P}_i^{b,\lambda_3}\}_{i=1}^{n_3}) \neq \sigma(\{\mathcal{P}_i^{b,\lambda_1}\}_{i=1}^{n_1},\, \{\mathcal{P}_i^{b,\lambda_2}\}_{i=1}^{n_2})$.

Continue in this fashion until λ_M denotes the multiplier associated with all binary partitions, denoted $\{\mathcal{P}_i^{b,\lambda_M}\}_{i=1}^{n_M}$, with the lowest multiplier such that the state of the world is always revealed when all questions with equal or lower associated multipliers are asked, that is, the lowest M such that: $\sigma(\{\mathcal{P}_i^{b,\lambda_1}\}_{i=1}^{n_1},\, \ldots,\, \{\mathcal{P}_i^{b,\lambda_M}\}_{i=1}^{n_M}) = \mathcal{F}$.

To help make the notation more compact, as is done in Appendix 1, a group of partitions can be used to **generate** a finer partition: if $(\mathcal{P}_1,\, \ldots,\, \mathcal{P}_m)$ is a group of partitions, let $\times\{\mathcal{P}_i\}_{i=1}^{n}$ denote the partition such that $\sigma(\times\{\mathcal{P}_i\}_{i=1}^{n}) = \sigma(\mathcal{P}_1,\, \ldots,\, \mathcal{P}_n)$. Then, for $j \in \{1,\, \ldots,\, M\}$,[20] let $\mathcal{P}_{\lambda_j} = \times\{\mathcal{P}_i^{b,\lambda_j}\}_{i=1}^{n_j}$.

The partitions described in the preceding paragraphs are the foundation for the different attributes of the choice environment used to define MASE in the body of this paper. More specifically, the **attributes** $\mathcal{A}_j \equiv \mathcal{P}_{\lambda_j}$ for $j \in \{1,\, \ldots,\, M\}$ are just specific partitions of the state space

[18] Allowing for costless learning is not difficult theoretically, but it does make exposition slightly more cumbersome. It can be shown that if free information is available then it is optimal for the agent to acquire that information, and then given its realization, choose an optimal learning strategy as described by the results in this paper.

[19] If M=1, then MASE collapses to standard Shannon Entropy.

[20] M is defined in the preceding paragraphs.

since the different outcomes for each attribute divide the state space into events. That is, $\forall\omega \in \Omega$ the **realization of the attribute** $\mathcal{A}_j$ is defined $\mathcal{A}_j(\omega) \equiv \mathcal{P}_{\lambda_j}(\omega) \in \mathcal{F}$.

Finally, since Ω is a partition of itself, one can, as a minor abuse of notation, let $S^b(\Omega) = \{S^b | \sigma(S^b) = \mathcal{F}\}$ denote the set of binary learning strategies such that $\sigma(S^b) = \sigma(\Omega) = \mathcal{F}$.

Given some probability measure μ, define the **mutual information** between two partitions $\mathcal{P}_1$ and $\mathcal{P}_2$, denoted $I(\mathcal{P}_1,\, \mathcal{P}_2,\, \mu)$, to be:

$$I(\mathcal{P}_1,\, \mathcal{P}_2,\, \mu) = \sum_{a_1 \in \mathcal{P}_1} \sum_{a_2 \in \mathcal{P}_2} \mu(a_1 \cap a_2) \log\left(\frac{\mu(a_1 \cap a_2)}{\mu(a_1)\mu(a_2)}\right)$$

Then, as is well known in the literature:

$$\mathcal{H}(\times\{\mathcal{P}_i\}_{i=1}^2,\, \mu) = \mathcal{H}(\mathcal{P}_1,\, \mu) + \mathcal{H}(\mathcal{P}_2,\, \mu) - I(\mathcal{P}_1,\, \mathcal{P}_2,\, \mu)$$

$$= \underset{\mathcal{H}(\mathcal{P}_1,\mu)-I(\mathcal{P}_1,\mathcal{P}_2,\mu)}{\underset{\parallel}{\mathbb{E}[\mathcal{H}(\mathcal{P}_1,\, \mu(\cdot|\mathcal{P}_2(\omega)))]}} + I(\mathcal{P}_1,\, \mathcal{P}_2,\, \mu) + \underset{\mathcal{H}(\mathcal{P}_2,\mu)-I(\mathcal{P}_1,\mathcal{P}_2,\mu)}{\underset{\parallel}{\mathbb{E}[\mathcal{H}(\mathcal{P}_2,\, \mu(\cdot|\mathcal{P}_1(\omega)))]}}$$

$$= \mathcal{H}(\mathcal{P}_1,\, \mu) + \mathbb{E}[\mathcal{H}(\mathcal{P}_2,\, \mu(\cdot|\mathcal{P}_1(\omega)))] = \mathcal{H}(\mathcal{P}_2,\, \mu) + \mathbb{E}[\mathcal{H}(\mathcal{P}_1,\, \mu(\cdot|\mathcal{P}_2(\omega)))],$$

and note that the strict concavity of $\mathcal{H}$ means that $I(\mathcal{P}_1,\, \mathcal{P}_2,\, \mu) \geq 0$.

Mutual information can be thought of as the information that is double counted if one were to compute the total uncertainty about the outcome of $\mathcal{P}_1$ and $\mathcal{P}_2$ by simply adding up the uncertainty about the outcome of $\mathcal{P}_1$ and the uncertainty about the outcome of $\mathcal{P}_2$. When the mutual information increases and the individual uncertainty about the outcome of $\mathcal{P}_1$ and the outcome of $\mathcal{P}_2$ are held constant the total uncertainty about the outcome of $\mathcal{P}_1$ and $\mathcal{P}_2$ decreases because the amount that remains to be learned after observing one of the outcomes of either $\mathcal{P}_1$ or $\mathcal{P}_2$ decreases.

Mutual information can be acquired by learning the value of either $\mathcal{P}_1$ or $\mathcal{P}_2$. When I think of an agent that is trying to acquire information in an efficient fashion, I should always envision them acquiring mutual information from the cheapest attribute, by learning about whichever of $\mathcal{P}_1$ and $\mathcal{P}_2$ has the lowest associated multiplier. This logic is formalized by the result in Lemma 15, and leads almost directly to the result in Theorem 5.

Lemma 15. If C satisfies all four axioms, and $S^b = \{\mathcal{P}_1^b,\, \dots,\, \mathcal{P}_i^b,\, \mathcal{P}_{i+1}^b,\, \dots,\, \mathcal{P}_m^b\}$ and $\tilde{S}^b = \{\mathcal{P}_1^b,\, \dots,\, \mathcal{P}_{i+1}^b,\, \mathcal{P}_i^b,\, \dots,\, \mathcal{P}_m^b\}$ are two binary learning strategies such that $\mathcal{P}_i^b$ and $\mathcal{P}_{i+1}^b$'s associated multipliers are ordered $\lambda_i \geq \lambda_{i+1}$, then for all probability measures μ:

$$C(S^b,\, \mu) \geq C(\tilde{S}^b,\, \mu).$$

Proof. Assume $\mathcal{P}^b_i$ and $\mathcal{P}^b_{i+1}$'s associated multipliers are ordered $\lambda_i \geq \lambda_{i+1}$ and that C satisfies all four axioms. For all realizations of $\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega)$ (if $i > 1$), Lemma 14 indicates:

$$\begin{aligned}
C((\mathcal{P}^b_i,\, \mathcal{P}^b_{i+1}),\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))) &= \lambda_i\mathcal{H}(\mathcal{P}^b_i,\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))) + \lambda_{i+1}\mathbb{E}[\mathcal{H}(\mathcal{P}^b_{i+1},\, \mu(\cdot|\cap_{j=1}^{i}\mathcal{P}^b_j(\omega)))]\\
&= \lambda_i\mathcal{H}(\mathcal{P}^b_i,\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))) + \lambda_{i+1}\Big(\mathcal{H}(\mathcal{P}^b_{i+1},\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))) - I(\mathcal{P}^b_i,\, \mathcal{P}^b_{i+1},\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))\Big)\\
&\geq \lambda_i\Big(\mathcal{H}(\mathcal{P}^b_i,\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))) - I(\mathcal{P}^b_i,\, \mathcal{P}^b_{i+1},\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))\Big) + \lambda_{i+1}\mathcal{H}(\mathcal{P}^b_{i+1},\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega)))\\
&= \lambda_{i+1}\mathcal{H}(\mathcal{P}^b_{i+1},\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))) + \lambda_i\mathbb{E}[\mathcal{H}(\mathcal{P}^b_i,\, \mu(\cdot|(\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))\cap\mathcal{P}^b_{i+1}(\omega)))]\\
&= C((\mathcal{P}^b_{i+1},\, \mathcal{P}^b_i),\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))).
\end{aligned}$$

Thus, it is weakly cheaper in expectation to have $\mathcal{P}_{i+1}$ before $\mathcal{P}_i$ as switching their order does not change the expected cost of the binary partitions before or after the pair. ■

Theorem 5. If C satisfies all four axioms then there exists attributes (partitions) $\mathcal{A}_1, \ldots, \mathcal{A}_M$ and unique constants $0 < \lambda_1 < \ldots < \lambda_M$ such that for any probability measure μ on $\mathcal{F}$:

$$\min_{S\in S^b(\Omega)} C(S,\, \mu) = \lambda_1\mathcal{H}\Big(\mathcal{A}_1,\, \mu\Big) + \mathbb{E}\Big[\lambda_2\mathcal{H}\Big(\mathcal{A}_2,\, \mu(\cdot|\mathcal{A}_1(\omega))\Big) + \cdots + \lambda_M\mathcal{H}\Big(\mathcal{A}_M,\, \mu(\cdot|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))\Big)\Big],$$

where $\mathcal{H}$ is Shannon Entropy, defined in equation (3).

Proof. Assume C satisfies all four axioms. Given some probability measure μ, suppose S^b is a binary learning strategy such that $\sigma(S^b) = \mathcal{F}$, and

$$C(S^b,\, \mu) = \min_{S^b\in S^b(\Omega)} C(S^b,\, \mu).$$

I may assume that if $\mathcal{P}^b_i$ and $\mathcal{P}^b_{i+1}$ are in S^b with associated multipliers λ_i and λ_{i+1}, that $\lambda_i \leq \lambda_{i+1}$. If not, then their order can be reversed and the resultant strategy is weakly less costly, as is shown in Lemma 15.

If for any $j \in \{1, \ldots, M\}$, multiplier λ_j's associated binary partitions $\mathcal{P}^b_i, \ldots, \mathcal{P}^b_{i+k}$ in S^b are such that $\sigma(\mathcal{P}^b_i, \ldots, \mathcal{P}^b_{i+k}) \neq \sigma(\mathcal{P}^b_{\lambda_j})$, then there are binary partitions $\mathcal{P}^b_{m+1}, \ldots, \mathcal{P}^b_{m+l}$ with associated multiplier λ_j, such that $\sigma(\mathcal{P}^b_i, \ldots, \mathcal{P}^b_{i+k}, \mathcal{P}_{m+1}, \ldots, \mathcal{P}^b_{m+l}) = \sigma(\mathcal{P}^b_{\lambda_j})$. $\mathcal{P}^b_{m+1}, \ldots, \mathcal{P}^b_{m+l}$ can be appended to the end of S^b, and the resultant strategy $\tilde{S}^b$ is also such that:

$$C(\tilde{S}^b, \mu) = \min_{S^b \in S^b(\Omega)} C(S, \mu).$$

This is true since each appended binary partition has an expected cost of zero, since $\sigma(S^b) = \mathcal{F}$. Lemma 15 then implies that if I reorder $\tilde{S}^b$ so that the new learning strategy $\hat{S}$'s binary partitions are ordered by their multipliers, then:

$$C(\hat{S}^b, \mu) = \min_{S^b \in S^b(\Omega)} C(S, \mu).$$

I can thus assume that S^b is such that for any $j \in \{1, \ldots, M\}$ multiplier λ_j's associated binary partitions $\mathcal{P}^b_i, \ldots, \mathcal{P}^b_{i+k}$ in S^b are such that $\sigma(\mathcal{P}^b_i, \ldots, \mathcal{P}^b_{i+k}) = \sigma(\mathcal{P}_{\lambda_j})$.

For each $j \in \{1, \ldots, M\}$ I thus have that if all binary partitions $\mathcal{P}^b_i, \ldots, \mathcal{P}^b_{i+k}$ in S^b with multiplier λ_j are taken together that:

$$\mathbb{E}[C((\mathcal{P}^b_i, \ldots, \mathcal{P}^b_{i+k}), \mu(\cdot| \cap_{t=1}^{i-1} \mathcal{P}^b_t(\omega)))] = \mathbb{E}\Big[\sum_{l=i}^{i+k} \lambda_j \mathcal{H}(\mathcal{P}^b_l, \mu(\cdot| \cap_{t=1}^{l-1} \mathcal{P}^b_t(\omega)))\Big]$$

$$= \mathbb{E}[\lambda_j \mathcal{H}(\mathcal{P}_{\lambda_j}, \mu(\cdot| \cap_{t=1}^{i-1} \mathcal{P}^b_t(\omega)))] = \mathbb{E}[\lambda_j \mathcal{H}(\mathcal{P}_{\lambda_j}, \mu(\cdot| \cap_{t=1}^{j-1} \mathcal{P}_{\lambda_t}(\omega)))],$$

where the second equality holds due to the properties of $\mathcal{H}$. This procedure can be carried out for all μ. Thus:

$$C(S^b, \mu) = \min_{S^b \in S^b(\Omega)} C(S, \mu).$$

$$= \lambda_1 \mathcal{H}\Big(\mathcal{P}_{\lambda_1}, \mu\Big) + \mathbb{E}\Big[\lambda_2 \mathcal{H}\Big(\mathcal{P}_{\lambda_2}, \mu(\cdot|\mathcal{P}_{\lambda_1}(\omega))\Big) + \cdots + \lambda_M \mathcal{H}\Big(\mathcal{P}_{\lambda_M}, \mu(\cdot|\cap_{i=1}^{M-1}\mathcal{P}_{\lambda_i}(\omega))\Big)\Big].$$

This is equivalent to the equation in the statement of the theorem due to the definition of the attributes. ■

In plain language, Theorem 5 says that if the cost of learning satisfies all four axioms, then the minimal cost (in expectation) to learn the state of the world with a binary learning strategy is equal to the cost of learning the realization of the attribute $\mathcal{A}_1$, the cheapest attribute to learn about, then learning the realization of attribute $\mathcal{A}_2$, the second cheapest attribute to learn about, and continuing in this fashion until the state of the world has been realized. This is optimal precisely because it minimizes the cost of acquiring the mutual information between the partitions.

In Theorem 5 the agent is minimizing their expected cost of learning the state of the world by selecting a sequence of binary partitions. This is different from the sequential optimization that is the focus of the work of Bloedel and Zhong (2021) as they allow agents to select a sequence of much more general signal structures that do not, in general, result in the agent perfectly observing the state of the world.

Theorem 5 generates the more flexible measure of uncertainty that I desired for studying inattentive behavior. If the agent starts with a prior μ, and does optimal learning that reaches a posterior $\tilde{\mu}$, then I let the cost of this inattentive research be measured by the reduction in the minimal expected cost of learning the state of the world with a binary learning strategy.

The $\mathcal{P}_{\lambda_i}$'s that are used to generate the attributes in Theorem 5 are not unique, with the exception of $\mathcal{P}_{\lambda_1}$, and thus the attributes are not unique. The versions described in the paragraphs preceding Theorem 5 can be used to define the attributes in the statement of the theorem, but, for $i \in \{2, \ldots, M\}$ the partition $\mathcal{P}_{\lambda_i}$ could, for instance, be replaced by $\tilde{\mathcal{P}}_{\lambda_i} = \times\{\mathcal{P}_{\lambda_j}\}_{j=1}^{i}$ for generating $\mathcal{A}_i$ in the statement of Theorem 5, which would constitute the unique finest representation of the partitions that could be used to define the attributes.