Foundation and Identification of Multi-Attribute Shannon Entropy

David Walker-Jones*
University of Surrey

February 1, 2024

Abstract

By weakening Shannon's original axioms to allow for attributes of the choice environment to differ in their associated learning costs, this paper provides an axiomatic foundation for Multi-Attribute Shannon Entropy, a natural multi-parameter generalization of Shannon Entropy. Sufficient conditions are also provided for a simple dataset that identifies the Multi-Attribute Shannon Entropy cost function for information by analysing stochastic choice data produced by a rationally inattentive agent that is picking between pairs of options when relatively few states of the world have a positive probability of being realized.

1 Introduction

It is costly for an economic agent to learn about the options that they face because it takes time and effort to acquire and process information. The cost of information may result in agents not acquiring all of the relevant information before making a decision, which creates important caveats for standard economic analysis techniques. Both welfare and counterfactual analysis, for instance, are more difficult if an agent does not always pick the best available option due to incomplete information.

The standard tool for measuring the cost of information in the rational inattention (RI) literature, Shannon Entropy (Shannon, 1948; Sims, 2003; Maćkowiak, Matějka, & Wiederholt, 2023), has limitations in economic environments because it is a one-parameter model for the cost

*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. Earlier versions of parts of this paper were previously circulated under the title "Rational Inattention and Perceptual Distance."

of learning the state of the world. In economic choice settings, however, it is natural to think that some attributes of the choice environment may be easier for the agent to learn about than others. If Shannon Entropy, whose one parameter imposes that all attributes are the same difficulty to learn about, is used to model agent behavior in such settings then predicted behavior does not resemble the behavior that has been observed in experiments (Dean & Neligh, 2023).

It is not difficult to come up with examples where Shannon Entropy's single parameter imposes unrealistic structure. Pomatto, Strack, and Tamuz (2023) have a particularly good example in which a researcher is gathering information about the GDP per capita of a country. If the researcher has a uniform prior belief about which of an interval of integers is the realized GDP per capita, then Shannon Entropy imposes that the expected cost to the researcher of determining if the GDP per capita is an even or odd number is the same as the expected cost of determining if the GDP per capita is above or below the median integer from the interval of outcomes they believed to be possible. This, of course, does not make sense because the "attribute" of GDP per capita of being even or odd should, on average, be more costly to learn about.

These types of problems arise with Shannon Entropy because Shannon's original work, which features an axiomatic foundation for Shannon Entropy, assumes that the agent can learn the state of the world by answering a series of questions,[1] and, crucially, that the cost to the agent of learning the state of the world does not depend on the order in which the questions are answered. This is problematic, however, since if some attributes are easier for the agent to learn about and are more helpful for identifying the state of the world then the agent might be able to reduce their expected cost of learning the state of the world by first trying to determine the realizations of these "cheaper" to learn about attributes.

This paper proposes four axioms that are similar to Shannon's original axioms (Shannon, 1948) in that they focus on the cost of answering simple questions that can be represented by partitions of the state space. Taken together, the four axioms in this paper are weaker than Shannon's axioms because they relax Shannon's assumption that the set of simple questions that is used, and the order in which they are answered, cannot change the expected cost of learning the state of the world. By allowing for the set of questions that is used to learn the state of the world, and the order in which they are answered, to change the agent's expected learning cost, this paper's axioms provide a foundation for Multi-Attribute Shannon Entropy (MASE), a

[1] Shannon does not refer to questions, but what he studies is the analogue of the partitions of the state space in this paper that are eventually defined as questions.

multi-parameter generalization of Shannon Entropy.

MASE, which can be understood as a measure of the agent's uncertainty, can be used to study a rationally inattentive agent that optimally learns in a flexible fashion because the cost of any imprecise learning that the agent does can be measured as the expected reduction in uncertainty that the learning causes, as is typically done with Shannon Entropy in models of RI. Thus, while this paper proposes axioms that, like Shannon's original axioms (Shannon, 1948), discuss an attentive agent that perfectly observes the state of the world, the model that the axioms produce can be used to study an inattentive agent that can choose to learn in a quite flexible manner and, in general, only partially learns about the state of the world.

MASE maintains much of the coveted tractability of Shannon's classic measure when incorporated into such a model of RI because Walker-Jones (2023) provides the MASE analogues of the famous necessary conditions provided by Matějka and McKay (2015) and necessary and sufficient conditions provided by Caplin, Dean, and Leahy (2018) for optimal agent behavior in RI models that use Shannon Entropy. MASE is thus a natural and tractable multi-parameter generalization of Shannon Entropy.

This paper also provides conditions that describe when a dataset is sufficient for the unique identification of the MASE cost function for information. 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 both a set of attributes and their associated learning costs that fully determines the cost of differentiating between outcomes when any set 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 provides an axiomatic foundation for MASE that weakens Shannon's original axioms (Shannon, 1948). Section 3 introduces a model of rational inattention that uses MASE to measure the cost of information and provides conditions for a dataset generated by a rationally inattentive agent that are sufficient for the unique identification of the agent's MASE cost function for information. Section 4 provides a literature review, and Section 5 concludes.

2 Axioms and MASE

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.

2.1 Learning Strategies

One natural way to model 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.[2] 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 paper.

Formally, 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.[3] 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. 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 $\omega \in A_i$.

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 paper.

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 $\mu(\cdot|\mathcal{P}(\omega))$.[4] $C(\mathcal{P}, \mu)$, 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 paper.

A **learning strategy**, $S = (\mathcal{P}_1, \ldots, \mathcal{P}_n)$, is a list of partitions whose realized events are

[2] 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.

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

[4] Where $\mu(\cdot|\mathcal{P}(\omega))$ is the distribution over states given the realization of partition $\mathcal{P}(\omega)$ and Bayes' rule.

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 paper. 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.[5]

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:

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

The definition of $C(S, \mu)$ thus imposes that, in a sense, 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. 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 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.

[5]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.

2.2 Axioms

What form should a cost function for information take? This difficult question does not have an obvious answer, so this paper 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. 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).

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 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 partitions that are not learning strategy invariant, however. In a setting where it seems that the different attributes of the state space should differ in their associated learning costs, and the agent may be able to lower their expected learning cost by changing the order in which they learn about them, there are partitions that one should not expect to be learning strategy invariant. Allowing for some partitions to not be learning strategy invariant is the key difference between the work in this paper and the work of Shannon (1948), who imposes that all partitions are learning strategy invariant.

A set of partitions that are certainly learning strategy invariant 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 lemma shows, 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 quite 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 1. 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 $p_1 + p_2 + p_3 = 1$:

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

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

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

Proofs for all the results in this paper can be found in Appendix 1.

In plainer language, Lemma 1 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. What I instead desire is 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.

Table 1: Firm Investment Example

State:	ω_1	ω_2
Value of investing in firm:	H	L
Value of not investing in firm:	M	M

To make this more concrete, consider a choice environment where an agent has two options: option 1, to invest in a firm that is either of high value H or low value L, or option 2, to not invest in the firm and receive a payoff of M with $L < M < H$, as is described in Table 1. Now, 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 firms that are essentially identical as the only way that they can differ is in their probability of being of high value, and the realization of Ω_i determines the value of investing in firm i for $i \in \{1, 2, 3\}$. Suppose $\mathcal{P}^b$ is the binary partition of Ω and that $\mathcal{P}_i^b$ is the equivalent binary partition of Ω_i for each i, so $\mathcal{P}_i^b$ is the question "Does firm i have high value?". Thus, $\mathcal{P}_1^b$, $\mathcal{P}_2^b$, and $\mathcal{P}_3^b$, are as similar as partitions can be by construction as they ask the values of three essentially identical firms. Now, suppose that the agent knows that the answer to one of the questions, $\mathcal{P}_1^b$, $\mathcal{P}_2^b$, or $\mathcal{P}_3^b$, is 'yes,' while the other two have the answer 'no,' which means that exactly one of the three firms has high value.[6] Denote the probability of $\mathcal{P}_i^b$ having the answer 'yes,' the probability of firm i having high value, by $p_i \in [0, 1)$ for $i \in \{1, 2, 3\}$.[7] Suppose the agent begins by learning about the realized event of $\mathcal{P}_i^b$. If the agent learns the answer to $\mathcal{P}_i^b$ 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 $\mathcal{P}_i^b$ is 'no' then their belief is updated using Bayes' Rule so that the probability of the answer to $\mathcal{P}_j^b$ 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 $\mathcal{P}_j^b$, 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 $C(\mathcal{P}_i^b, p, 1-p) = C(\mathcal{P}^b, p, 1-p)$ 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 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:

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

Another way of understanding Axiom 2 is that it extends the conclusion of Lemma 1 to any binary partition, not just those that are coarser than a learning strategy invariant partition.

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

[6] 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 answers to the other questions.

[7] 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 2, which further shows that the cost function is permutation invariant on binary partitions.

Lemma 2. 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 $p \in [0, 1]$.

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 3. 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 exists $\theta > 0$ such that if $0 < \gamma < \theta$ then $C(\mathcal{P}^b, p, 1-p) \leq C(\mathcal{P}^b, p+\gamma, 1-p-\gamma)$.

I now show that the cost of learning the realized event of a learning strategy invariant partition is dictated by Shannon Entropy, which needs to be defined. Given a partition 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:[8]

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

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

Lemma 4. 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 (1).

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 1 in the next subsection. 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

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

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.[9]

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

Lemma 4 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: $C(\mathcal{P}^b,\, \mu) = \lambda(\mathcal{P}^b)\mathcal{H}(\mathcal{P}^b,\, \mu)$.

2.3 Total Uncertainty

This subsection defines MASE using $M \geq 1$ attributes. The number of attributes required for modelling the learning of the agent, M, is determined by the environment and, in particular, is the number of different associated multipliers for the binary partition used when the agent efficiently learns the state of the world using binary partitions, as is described in the following paragraphs.

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 (learning the realization of only one attribute is always sufficient for learning the state of the world).[10] 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 $\sigma(\{\mathcal{P}_i^{b,\lambda_1}\}_{i=1}^{n_1},\, \{\mathcal{P}_i^{b,\lambda_2}\}_{i=1}^{n_2},\, \{\mathcal{P}_i^{b,\lambda_3}\}_{i=1}^{n_3}) \neq \sigma(\{\mathcal{P}_i^{b,\lambda_1}\}_{i=1}^{n_1},\, \{\mathcal{P}_i^{b,\lambda_2}\}_{i=1}^{n_2})$.

[9] 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.

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

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, 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 $j \in \{1, \ldots, M\}$,[11] 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 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 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 $\sigma(S^b) = \sigma(\Omega) = \mathcal{F}$.

Theorem 1. If C satisfies all four axioms then the attributes (partitions) $\mathcal{A}_1, \ldots, \mathcal{A}_M$, with associated multipliers (constants) $0 < \lambda_1 < \ldots < \lambda_M$, are 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 (1).

In plain language, Theorem 1 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 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 1 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

[11] M is defined in the preceding paragraphs.

much more general signal structures that do not, in general, result in the agent perfectly observing the state of the world.

Theorem 1 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 (see Section 3 for more details).

The $\mathcal{P}_{\lambda_i}$'s that are used to generate the attributes in Theorem 1 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 1 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 1, which would constitute the unique finest representation of the partitions that could be used to define the attributes.

Using the attributes, their associated multipliers, and Theorem 1, I define **Multi-Attribute Shannon Entropy** (MASE), $\mathbb{H} : \Delta(\Omega) \to \mathbb{R}_+$, to be the measure of total uncertainty:

$$\mathbb{H}(\mu) \equiv \min_{S \in S^b(\Omega)} C(S, \mu)$$

$$= \lambda_1 \mathcal{H}\Big(\mathcal{A}_1, \mu\Big) + \mathbb{E}\Big[\lambda_2 \mathcal{H}\Big(\mathcal{A}_2, \mu(\cdot|\mathcal{A}_1(\omega))\Big) + \cdots + \lambda_M \mathcal{H}\Big(\mathcal{A}_M, \mu(\cdot|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))\Big)\Big], \quad (2)$$

where $\mathcal{H}$ is Shannon Entropy, which is defined in equation (1). 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 typically done with Shannon Entropy when it is used in RI models (Matějka & McKay, 2015).

3 Inattentive Learning with MASE and Identification

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 $\mathbf{v}_n(\omega) \in \mathbb{R}$. Informally, the agent's problem is to maximize the expected value of their selected option less the cost of their learning. The more different their behavior is in different states, i.e. the more their chances of selecting different options varies across states, the more expensive their information gathering is because it requires more information to have behavior that is more different from state to state.

We follow Matějka and McKay (2015) and Walker-Jones (2023) and write the agent's problem directly in terms of the choice probabilities of the agent. Denote the probability of option n being selected conditional on event $A \in \mathcal{F}$ to be $\Pr(n|A)$ and, as a minor abuse of notation, define the **unconditional probability** of option n being selected to be the probability of n being selected conditional on the event $A = \Omega$: $\Pr(n) \equiv \Pr(n|\Omega)$. 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**.

The agent's problem is to maximize the expected value of their selected option less the cost of learning. Define the expected cost of the agent's behavior to be the expected reduction in total uncertainty caused by $\mathbb{P}$ and the observation of the selected option as measured by $\mathbb{H}$:

$$\mathbf{C}(\mathbb{P},\, \mu) \equiv \sum_{n \in \mathcal{N}} \Pr(n) \Big[\mathbb{H}(\mu) - \mathbb{H}(\mu(\cdot|n)) \Big]$$

where $\mathbb{H}(\mu)$ is as defined in equation (2) and $\mu(\cdot|n) : \Omega \to \mathbb{R}_+$ is the posterior belief of the agent after option n is selected given the prior μ, behavior $\mathbb{P}$, and Bayes' Rule. 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. The agent's problem can thus be written:

$$\max_{\mathbb{P}} \sum_{n \in \mathcal{N}} \sum_{\omega \in \Omega} \mathbf{v}_n(\omega) \Pr(n|\omega) \mu(\omega) - \mathbf{C}(\mathbb{P},\, \mu), \tag{3}$$

$$\text{such that: } \forall\, n \in \mathcal{N},\ \Pr(n|\omega) \geq 0,\ \forall\, \omega \in \Omega, \tag{4}$$

$$\text{and } \sum_{n \in \mathcal{N}} \Pr(n|\omega) = 1\ \forall\, \omega \in \Omega. \tag{5}$$

The objective described by equation (3) is shown by Walker-Jones (2023) to be concave on the set of $\mathbb{P}$ that satisfy (4) and (5). If behavior $\mathbb{P}$ solves (3) subject to (4) and (5) then it is referred to as **optimal**. Necessary and sufficient conditions for optimal behavior are provided by Walker-Jones (2023) and are re-stated in Appendix 1 for the reader's convenience.

3.1 Identification of the Cost of Learning

It is natural to want to fit a costly learning model to data. (Denti, 2022) demonstrates that sufficiently rich data can be used to uniquely identify any posterior separable cost function for information in a non-parametric manner. Typical datasets, however, feature stochastic choice data from finitely many choice problems. One advantage of fitting a MASE cost function is that the parametric form, though less flexible compared to the more general class of posterior separable cost functions, is that estimation may require a less rich dataset.

What data is required to uniquely identify $\mathbb{H} : \Delta(\Omega) \to \mathbb{R}_+$? Theorem 2 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 $\mathbb{H}$ and, importantly, sufficient for determining if said certain properties are satisfied.

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 (3) subject to (4) and (5) 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 $\mathbb{H}(\mu) = \lambda(\omega_i, \omega_j)(-\log(\frac{1}{2}))$. 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 $\mathbf{C}(\mathbb{P}, \mu)$. While some of the conditions in Theorem 2 mention the multipliers, they do not assume anything about the multipliers, whether or not the multipliers satisfy the conditions is identifiable with the assumed data.

Theorem 2: 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 a finite subset of the $\mathbb{P}^*(\mathcal{M}, \mu)$ uniquely identifies $\mathbb{H}$ out of the set of MASE cost functions for information. Further, for each such pair of states, whether or not at least one of the following conditions **(i)**-**(v)** are satisfied is observable given the assumed dataset.

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 2 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 2 does not say that behavior uniquely identifies the attributes, as there can be different sets of attributes that produce the same $\mathbb{H}$.

The intuition behind the proof of Theorem 2 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 it can be shown 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 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 2 is constructive in the sense that, if a pair of states satisfies one of the five conditions, the proof of Theorem 2 demonstrates how the data indicates which of the five conditions is satisfied, 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 $\mathbb{H}$.

4 Literature Review

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, Dean, and Leahy (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 et al. (2023) 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. (2023), 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 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, for which Mensch (2018) provides an axiomatic characterization, and are, in particular, uniformly posterior separable (Caplin et al., 2022; Denti, 2022) and a strict subset of the neighborhood-based cost functions described by Hébert and Woodford (2021). Walker-Jones (2023) studies the optimal behavior of a rationally inattentive agent that pays for information according to a MASE cost function.

This paper complements the literature in math and information theory on axiomatic characterizations of information measures. For a survey of this literature, see the work of Csiszár (2008).

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

5 Conclusion

This paper introduces four axioms that are similar to Shannon's original axioms (Shannon, 1948) in that they focus on the cost of answering simple questions that can be represented by partitions of the state space. Taken together, the four axioms in this paper are weaker than Shannon's axioms because they relax Shannon's "learning strategy invariance" assumption that imposes that the set of simple questions that is used, and the order in which they are answered, cannot change the expected cost of learning the state of the world. By allowing for the set of questions that is used to learn the state of the world, and the order in which they are answered, to change the agent's expected learning cost, this paper's axioms provide a foundation for Multi-Attribute Shannon Entropy (MASE), a multi-parameter generalization of Shannon Entropy. MASE allows for attributes of the choice environment to differ in their associated learning costs, and it is shown that learning about the less costly to observe attributes first, i.e. learning by answering questions about the realizations of the attributes in the order of their associated learning costs, always minimizes the expected cost, no matter the distribution over states.

Several redundancies in Shannon's original axioms are also identified by the work in this paper. For instance, if a binary partition is coarser than a learning strategy invariant partition then the function that maps probabilities of the two answers onto costs must either be continuous everywhere or discontinuous everywhere, so assuming it is continuous at a single point is sufficient for developing MASE. It is thus only substantive to relax Shannon's"learning strategy invariance" property to allow for the order of questions to change the agent's expected learning cost.

MASE, which can be understood as a measure of the agent's uncertainty about the state of the world, can be used to study a rationally inattentive agent that optimally learns in a flexible fashion because the cost of any imprecise learning that the agent does can be measured as the expected reduction in uncertainty that it causes, as is typically done with Shannon Entropy in models of RI. Thus, while this paper proposes axioms that, like Shannon's original axioms (Shannon, 1948), discuss an attentive agent that perfectly observes the state of the world, the model that the axioms produce can be used to study an inattentive agent that can choose to learn in a quite flexible manner and, in general, only partially learns about the state of the world.

MASE maintains much of the coveted tractability of Shannon's classic measure when incorporated into such a model of RI because Walker-Jones (2023) 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 is thus a natural and tractable multi-parameter generalization of Shannon Entropy.

This paper also provides conditions that describe when a dataset is sufficient for the unique identification of the MASE cost function for information. 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 both a set of attributes and their associated learning costs that fully determines the cost of differentiating between outcomes when any set of the potential states of the world occur with a positive probability. This identification is made possible through variation of the prior belief of the agent and the set of options that are available to them, and builds upon the work of Walker-Jones (2023).

References

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

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.

Csiszár, I. (2008). Axiomatic characterizations of information measures. *Entropy*, *10*(3), 261–273.

Dean, M., & Neligh, N. (2023). Experimental tests of rational inattention. *Journal of Political Economy*, *131*(12), 3415–3461.

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.

Maćkowiak, B., Matějka, F., & Wiederholt, M. (2023). Rational inattention: A review. *Journal of Economic Literature*, *61*(1), 226–273.

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*.

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. (2023). The cost of information: The case of constant marginal costs. *American Economic Review*, *113*(5), 1360–1393.

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.

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

Walker-Jones, D. (2023). Rational inattention with multiple attributes. *Journal of Economic Theory*, 105688.

Appendix 1

Before proving Lemma 1, I pause to introduce and prove some other useful results. All of the proofs for all of the results in this paper are contained in this appendix.

Lemma 5. 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 5 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 6 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 6. 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 7 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 1.

Lemma 7. 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 $\mathcal{P} = \{A_1,\, \dots,\, A_m\}$ is a learning strategy invariant partition of the state space Ω with $m \geq 3$. Axiom 1 imposes that knowing $\mu(A_1),\, \dots,$ 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 $\mu(A_1),\, \mu(A_2)\, \dots,$ and $\mu(A_m)$.

If I then show that for any $i,\, j \in \{1,\, \dots,\, m\}$ 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 5. 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 $\hat{\mathcal{P}} = \{A_1 \cup A_2,\, A_3,\, \dots,\, A_m\}$ that is also learning strategy invariant based on Lemma 5, then Lemma 6 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 $i \in \{3,\, \dots,\, m\}$, and otherwise so that $\hat{\mu}(A_1) = 1$. Now, let $\mathcal{P}^b_1 = \{A_1, A_1^c\}$, $\mathcal{P}^b_2 = \{A_2, A_2^c\}$, and $\mathcal{P}^b_3 = \{A_1 \cup A_2,\, (A_1 \cup A_2)^c\}$. Notice $\mathcal{P}^b_1$, $\mathcal{P}^b_2$ and $\mathcal{P}^b_3$, are all coarser than $\tilde{\mathcal{P}}$. Then, since $\tilde{\mathcal{P}}$ is learning strategy invariant:

$$C(\tilde{\mathcal{P}},\, \mu) = C(\mathcal{P}^b_3,\, \mu) + \mathbb{E}[C(\mathcal{P}^b_1,\, \mu(\cdot|\mathcal{P}^b_3(\omega))],\text{ and } C(\tilde{\mathcal{P}},\, \tilde{\mu}) = C(\mathcal{P}^b_3,\, \tilde{\mu}) + \mathbb{E}[C(\mathcal{P}^b_1,\, \tilde{\mu}(\cdot|\mathcal{P}^b_3(\omega))].$$

Notice that Axiom 1 imposes that $C(\mathcal{P}^b_3,\, \mu) = C(\mathcal{P}^b_3,\, \tilde{\mu})$ 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 $\mathcal{P}^b_1$, then $C(\mathcal{P}^b_1,\, \nu) = C(\mathcal{P}^b_1,\, \tilde{\nu})$. 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 6, all repeatedly:

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

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

Proof of Lemma 1. 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 7 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 (meaning that if the vectors $\mu(\mathcal{P}_i)$ and $\mu(\tilde{\mathcal{P}}_i)$ were both permuted so that the probabilities in them are weakly decreasing then the strictly positive entries of the permuted vectors would be identical), then the addition of Lemma 6 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 5 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:

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

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

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

Proof of Lemma 2. 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 2 I show two technical results, Lemma 8 and Lemma 9, that are helpful for proving Lemma 2.

Lemma 8. 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 9. 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 8, 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 $q_i = p_1^i + \ldots p_{n_i}^i$. Then it is true that:

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

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 8, and $(q_1,\, \ldots,\, q_m)$ and $(p_1,\, \ldots,\, p_n)$ are defined as in the statement of Lemma 9 (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 8 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 $p_n > 0$):

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

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

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

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

$$= c_{\mathcal{P}^b}(q_1,\, \ldots,\, q_m) + q_m c_{\mathcal{P}^b}\Big(\frac{p_m}{q_m},\, \ldots,\, \frac{p_n}{q_m},\, 0\Big). \blacksquare$$

I am now ready to resume the proof of Lemma 2. 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 8, 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 8 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 9 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 $c_{\mathcal{P}^b}(p_1+p_2, p_3+p_4)+(p_1+p_2)c_{\mathcal{P}^b}\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right)+(p_3+p_4)c_{\mathcal{P}^b}\left(\frac{p_3}{p_3 + p_4}, \frac{p_4}{p_3 + p_4}\right)$ goes to zero, so discontinuities at either side of p_d must offset each other so the change in $c_{\mathcal{P}^b}(p_1+p_4, p_2+p_3)+(p_1+p_4)c_{\mathcal{P}^b}\left(\frac{p_1}{p_1 + p_4}, \frac{p_4}{p_1 + p_4}\right)+(p_2+p_3)c_{\mathcal{P}^b}\left(\frac{p_2}{p_2 + p_3}, \frac{p_3}{p_2 + p_3}\right)$ 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 $c_{\mathcal{P}^b}\left(\frac{p_1}{p_1 + p_4}, \frac{p_4}{p_1 + p_4}\right)$ 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 9 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. ■

Proof of Lemma 3. 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 8, and suppose C satisfies Axiom 1, Axiom 2, and Axiom 3. Remember Lemma 8 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 $p \in (0,\, 1/2)$. Further, remember that Lemma 2 implies $c_{\mathcal{P}^b}$ is continuous.

If it is not the case that for all $p \in [0,\, \frac{1}{2})$ there exists $\theta > 0$ such that if $0 < \gamma < \theta$ then $c_{\mathcal{P}^b}(p,\, 1-p) \leq c_{\mathcal{P}^b}(p+\gamma,\, 1-p-\gamma)$, then (since $c_{\mathcal{P}^b}(0,\, 1) = 0$ and $c_{\mathcal{P}^b}$ is a weakly positive function) $\exists p \in (0,\, \frac{1}{2})$ such that for all $\theta > 0$ there is $\gamma < \theta$ with $\gamma > 0$ such that $c_{\mathcal{P}^b}(p,\, 1-p) > c_{\mathcal{P}^b}(p+\gamma,\, 1-p-\gamma)$. But, the Extreme Value Theorem implies that there is at least one point $p_d \in [p,\, p+\gamma)$ at which $c_{\mathcal{P}^b}$ attains its maximum value over the range $[p,\, p+\gamma]$, and since $c_{\mathcal{P}^b}$ is continuous and $c_{\mathcal{P}^b}(p+\gamma,\, 1-p-\gamma) < c_{\mathcal{P}^b}(p,\, 1-p) \leq c_{\mathcal{P}^b}(p_d,\, 1-p_d)$, there is a last (highest) $p_d \in [p,\, p+\gamma)$ at which $c_{\mathcal{P}^b}$ attains its maximum value over this range. This all implies that there is a point $p_d \in (0,\, 1/2)$ such that $c_{\mathcal{P}^b}(p_d,\, 1-p_d)$ is decreasing for small increases in p_d.

I thus proceed by assuming that 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. 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:

$$p_1+p_2 = p_d,\ \frac{p_1}{p_1+p_2} = p_d,\ \frac{p_3}{p_3+p_4} = p_d,\ \text{ so that } \frac{p_1}{p_1+p_4} = p_d\frac{p_d}{p_d^2+(1-p_d)^2} < p_d,$$

and then notice Lemma 9 tells us:

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

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

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

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 $p = p_1/(p_1 + p_4) < p_d$. Further, since $p/(p^2 + (1+p)^2)$ is increasing in p, there must be dense p near 0 where $c_{\mathcal{P}^b}(p, 1-p)$ decreases for small increases to p.

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 8 and Lemma 9 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 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 9 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 $c_{\mathcal{P}^b}(p_1,\,p_2,\,p_3)$:

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

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

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

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

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

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. ■

Proof of Lemma 4. Assume C satisfies Axiom 1, Axiom 2, and Axiom 3. Given learning strategy invariant partition $\mathcal{P} = \{A_1, \dots, A_m\}$ 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 8 so that, by Lemma 7, $C(\mathcal{P}, \mu) = c_{\mathcal{P}^b}(\mu(A_1), \dots, \mu(A_m))$.

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 8 and Lemma 9 imply that: $c_{\mathcal{P}^b}(p_1, p_2, p_3) =$

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

This and Lemma 2 and Lemma 3 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 2 and Lemma 3 since $\frac{p_2}{p_2+p_3} > p_1$, so:

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

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

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 $n \in \mathbb{N}$, $h(n) \equiv c_{\mathcal{P}^b}(1/n, \dots, 1/n, 0)$. 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 9 and Lemma 3:

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

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

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

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

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

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

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:

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

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

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

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

where $\mathcal{H}$ is defined as in equation (1). If any of the p_i are irrational, then the density of the rationals and Lemma 2 can be used to get the same result. Thus, $C(\mathcal{P}, \mu) = \lambda \mathcal{H}(\mathcal{P}, \mu)$. ■

Total Uncertainty

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:

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

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

$$\underset{\mathcal{H}(\mathcal{P}_1,\mu)-I(\mathcal{P}_1,\mathcal{P}_2,\mu)}{||} \qquad\qquad \underset{\mathcal{H}(\mathcal{P}_2,\mu)-I(\mathcal{P}_1,\mathcal{P}_2,\mu)}{||}$$

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

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 10, and leads almost directly to the result in Theorem 1.

Lemma 10. If C satisfies all four axioms, and $S^b = \{\mathcal{P}^b_1,\, \ldots,\, \mathcal{P}^b_i,\, \mathcal{P}^b_{i+1},\, \ldots,\, \mathcal{P}^b_m\}$ and $\tilde{S}^b = \{\mathcal{P}^b_1,\, \ldots,\, \mathcal{P}^b_{i+1},\, \mathcal{P}^b_i,\, \ldots,\, \mathcal{P}^b_m\}$ are two binary learning strategies such that $\mathcal{P}^b_i$ and $\mathcal{P}^b_{i+1}$'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 4 indicates:

$$C((\mathcal{P}^b_i,\, \mathcal{P}^b_{i+1}),\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))) = \lambda_i \mathcal{H}(\mathcal{P}^b_i,\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))) + \lambda_{i+1}\mathbb{E}[\mathcal{H}(\mathcal{P}^b_{i+1},\, \mu(\cdot|\cap_{j=1}^{i}\mathcal{P}^b_j(\omega)))]$$

$$= \lambda_i \mathcal{H}(\mathcal{P}^b_i,\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))) + \lambda_{i+1}\Big(\mathcal{H}(\mathcal{P}^b_{i+1},\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))) - I(\mathcal{P}^b_i,\, \mathcal{P}^b_{i+1},\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))\Big)$$

$$\geq \lambda_i\Big(\mathcal{H}(\mathcal{P}^b_i,\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))) - I(\mathcal{P}^b_i,\, \mathcal{P}^b_{i+1},\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))\Big) + \lambda_{i+1}\mathcal{H}(\mathcal{P}^b_{i+1},\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega)))$$

$$= \lambda_{i+1}\mathcal{H}(\mathcal{P}^b_{i+1},\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega))) + \lambda_i\mathbb{E}[\mathcal{H}(\mathcal{P}^b_i,\, \mu(\cdot|(\cap_{j=1}^{i-1}\mathcal{P}^b_j(\omega)) \cap \mathcal{P}^b_{i+1}(\omega)))]$$

$$= C((\mathcal{P}^b_{i+1}, \mathcal{P}^b_i), \mu(\cdot| \cap_{j=1}^{i-1} \mathcal{P}^b_j(\omega))).$$

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. ■

Proof of Theorem 1. 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 10.

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 10 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. ■

Identifying the Cost of Information

The proof of Theorem 2 builds upon the necessary and sufficient condotions for optimal behavior provided by Walker-Jones (2023), which are described by Theorem 1 and Theorem 3 from that paper. Theorem 1 and Theorem 3 from the work of Walker-Jones (2023) are presented below with small amendments so that they correspond to the correct equations in this paper and do not require any new notation.

Theorem 1 from the work of Walker-Jones (2023).

If $\mathbb{P}$ is optimal 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:

$$\Pr(n|\omega) = \frac{\Pr(n)^{\frac{\lambda_1}{\lambda_M}} \Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \ldots \Pr(n|\cap_{i=1}^{M-1} \mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}}{\sum_{\nu \in \mathcal{N}} \Pr(\nu)^{\frac{\lambda_1}{\lambda_M}} \Pr(\nu|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \ldots \Pr(\nu|\cap_{i=1}^{M-1} \mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_M}}}. \tag{6}$$

Theorem 3 from the work of Walker-Jones (2023). Behavior $\mathbb{P}$ is optimal iff for all $n \in \mathcal{N}$ with $\Pr(n) > 0$ it is the case that $\Pr(n|\omega) > 0$ and $\Pr(n|\omega)$ is described by equation (6) for each state $\omega \in \Omega$, and for all $n \in \mathcal{N}$ with $\Pr(n) = 0$ it is the case that:

$$\mathbb{E}\Big[\mathbb{E}\Big[\ldots \mathbb{E}\Big[\mathbb{E}\Big[s_n(\omega|\mathbb{P})|\cap_{i=1}^{M-1} \mathcal{A}_i(\omega)\Big]^{\frac{\lambda_M}{\lambda_{M-1}}} |\cap_{i=1}^{M-2} \mathcal{A}_i(\omega)\Big]^{\frac{\lambda_{M-1}}{\lambda_{M-2}}} \ldots |\mathcal{A}_1(\omega)\Big]^{\frac{\lambda_2}{\lambda_1}}\Big] \leq 1.$$

Proof of Theorem 2. 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

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

and it is not hard to show

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

thus, Theorem 3 from the work of Walker-Jones (2023) indicates that both options are selected with a positive probability when such a μ is the prior, and therefore Theorem 1 from the work of Walker-Jones (2023) indicates that $\lambda(\omega_i, \omega_j)$ solves:

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

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 3 from the work of Walker-Jones (2023) thus implies any $\mathbb{P}^*(\{n, m\}, \tilde{\mu})$ features both options being selected with a positive probability, and therefore Theorem 1 from the work of Walker-Jones (2023) 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 3 from the work of Walker-Jones (2023) 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 3 from the work of Walker-Jones (2023) indicates that $\mathbb{P}(\{n, m\}, \tilde{\mu})$ features a positive probability of both n and m being selected, and thus Theorem 1 from the work of Walker-Jones (2023) 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 from the work of Walker-Jones (2023) implies:

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

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

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

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

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

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

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 from the work of Walker-Jones (2023) implies:

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

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 from the work of Walker-Jones (2023)), 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 from the work of Walker-Jones (2023) 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 from the work of Walker-Jones (2023)), 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 from the work of Walker-Jones (2023) implies $\lambda(\omega_i, \omega_j)$ solves:

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

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

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

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 from the work of Walker-Jones (2023) implies it solves:

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

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

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

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 3 from the work of Walker-Jones (2023) 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. ■