Rational Inattention and Perceptual Distance

David Walker-Jones

University of Toronto

david.walker.jones@mail.utoronto.ca

Keywords: rational inattention, Shannon Entropy, perceptual distance.

JEL Classification : D83

February 21, 2020

Abstract

This paper uses an axiomatic foundation to create a new measure for the cost of learning that allows for multiple perceptual distances in a single choice environment so that some events can be harder to differentiate between than others. The new measure maintains the tractability of Shannon's classic measure but produces richer choice predictions and identifies a new form of informational bias significant for welfare and counterfactual analysis.[1]

1 Introduction

In many choice environments it is costly for agents to learn about the options that they face because it takes time to acquire and process information. Understanding how agents learn in such environments is crucial because partially informed choices have serious implications for revealed preference analysis, which makes welfare and

[1]Special thanks to Rahul Deb for all of the support. I would also like to thank Yoram Halevy, Marcin Peski, Carolyn Pitchik, and Colin Stewart, for their helpful advice.

counterfacutal analysis more difficult.

The standard technique for quantifying the cost of learning in models of rational inattention (RI) is Shannon Entropy (Sims, 2003). Shannon Entropy has an axiomatic foundation, is grounded in the optimal coding of information, and provides a tractable and flexible framework with which to study agent behavior (Shannon, 1948).

While Shannon Entropy has proven to be a valuable tool, it does have limitations in economic environments as they are not what it is designed for. It is natural to think that it should be, for instance, more difficult to differentiate between outcomes that are more similar. Differentiating between two types of black tea should be more difficult than differentiating between water and coffee. Shannon Entropy, however, does not allow for different outcomes to be more or less similar than each other. Without a mechanism to allow for what is referred to in the literature as 'perceptual distance,'[2] the choice behavior predicted by Shannon Entropy can differ from observed behavior, as is discussed in Example 1 in Section 2.1, which can limit the effectiveness of Shannon Entropy in empirical settings.

This paper proposes five axioms that are similar to Shannon's original axioms (1948) in that they focus on the cost of answering simple questions, questions that can be represented by partitions of the state space. Taken together, the five axioms in this paper are weaker than Shannon's axioms (1948) because they relax Shannon's assumption that all partitions are, what we refer to in this paper as, 'learning strategy invariant'. By allowing for some partitions to not be learning strategy invariant we incorporate perceptual distance into our new measure for the cost of information, which we call Multisource Shannon Entropy (MSSE).

Though the axioms in this paper discuss an agent learning through simple partitions of the state space, we need not constrain the agent to learn in such a fashion, and can use MSSE to measure the cost of information in a more general setting where the agent can choose any signal structure they desire, as is typical in the literature

[2]If two outcomes are more similar it is said that they have less perceptual distance between them.

on RI. This is because MSSE can be viewed as a measure of total uncertainty, and, as such, the cost of an arbitrary signal can simply be measured as the difference between the total uncertainty before and after the signal is realized, as is frequently done with Shannon Entropy in models of RI. This paper shows that, when used in such a fashion, MSSE maintains much of the desired tractability and flexibility of Shannon's classic measure when incorporated into a model of RI, but also predicts behavioral patterns that have been identified as problematic for Shannon Entropy.

MSSE further identifies an informational bias in random utility (RU) models that should be considered a natural consequence of different perceptual distances in the same choice environment, as is demonstrated by Example 2 in Section 2.2. While other papers study measures of information that feature perceptual distance (e.g., Hébert & Woodford, 2017), this paper is the first to identify an informational bias in RU models that is generated by the presence of different perceptual distances in the same choice environment. Unlike the informational bias identified with Shannon Entropy (Matějka & McKay, 2015), this type of informational bias cannot be identified in the unconditional choice probabilities of the agent, and thus presents a new challenge for welfare and counterfactual analysis.

1.1 Literature Review

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

The Shannon RI model has also led to a number of predictive successes. Acharya and Wee (2019) show that using Shannon Entropy to model firms as rationally inattentive results in a better fitting of labor market dynamics after the great depression. Dasgupta and Mondria (2018) show that using Shannon Entropy to model importers as rationally inattentive results in novel predictions that are supported by trade data. Ambuehl, Ockenfels, and Stewart (2019) experimentally verify predictions of Shannon Entropy in environments where agents are rationally inattentive to the consequences of participating in different transactions.

Perhaps as a response to the success Shannon Entropy has enjoyed, several recent papers have noted that Shannon Entropy may be a poor measure of the cost of acquiring information in some environments (Caplin, Dean, & Leahy, 2017; Morris & Yang, 2016) because it lacks what is called "perceptual distance" (Caplin et al., 2017, p. 39). As was alluded to previously, these papers argue that (i) more similar outcomes (outcomes that have less perceptual distance between them) should be more difficult to differentiate between, and (ii) when this property is missing, predicted behavior can differ signficantly from the type of behavior that it would seem natural to expect (Morris & Yang, 2016).

An ad hoc group of cost functions that generalize Shannon Entropy and allow for different perpetual distances is provided by Huettner, Boyacı, and Akçay (2019). In their paper, the different alternatives that the agent can choose between are allowed to differ in how costly they are to learn about, i.e., how much perceptual distance there is between realizations of alternatives' values can differ across alternatives. The group of costs functions developed by Huettner et al. (2019) are a strict subset of the cost functions that can be defined with MSSE, and though they allow for different perceptual distances, they are not capable of predicting the behavior we argue is intuitive in Example 1 in Section 2.1 and is predicted by MSSE.

To better understand the relationship between the cost of learning and agent behavior, a number of papers have studied axiomatic models of rational inattention.

Different papers, however, choose to focus their axioms on different aspects of the choice environment. Caplin et al. (2017), 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 MSSE 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 MSSE it is never optimal for the agent to choose an information strategy that is equivalent to a partition of the state space (Walker-Jones, 2019).[3]

Closer in nature to the work done in this paper, Pomatto, Strack, and Tamuz (2019) develop axioms that focus directly on the costs of information. Axioms that focus on costs of 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. MSSE, 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, as is discussed by (Walker-Jones, 2019). 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. (2019), which features axioms that are con-

[3]This is true whenever the agent does some learning, and they have a positive probability of a posterior that is different than their prior.

cerned with probabilistic experiments that can result in different outcomes in the same state of the world, this paper's axioms are concerned with deterministic experiments (questions) that always result in the same outcome in a given state of the world, and contradict the form of constant marginal cost assumed in their paper.

1.2 Organization of Paper

The remainder of the paper is organized as follows: Section 2 introduces Shannon Entropy, discusses models of RI, and provides motivating examples. Section 3 proposes five new axioms, and uses them to develop a more flexible cost of acquiring information, MSSE, which features perceptual distance. Section 4 uses MSSE as a benchmark with which to price inattentive information strategies in a model of RI, and discusses the resultant agent behavior. Section 5 discusses the relationship between RU models and the agent behavior found in Section 4, and revisits the motivating examples from Section 2.1 and Section 2.2. Section 6 concludes.

2 Rational Inattention and Shannon Entropy

What follows is intended to introduce the Shannon Entropy model of rational inattention to those that are not familiar with it. If you are familiar with said model, you can skip to Section 2.1.

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

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). We call $\mu : \mathcal{F} \to [0, 1]$, which assigns probabilities to events, the **prior** distribution of the agent.

Suppose that an agent who has stopped learning 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 $\mathbf{v}_n(\omega)$.

The agent's problem is to maximize the expected value of the selected option less the cost of learning. They do this by choosing an **information strategy** $F(s, \omega) \in \Delta(\mathbb{R} \times \Omega)$, which is a joint distribution between s, the observed **signal**, and the states of the world.[4] The only restriction on the information strategy is that the marginal, $F(\omega) : \mathcal{F} \to \mathbb{R}_+$, must equal the prior μ. Alternatively, an agent can select a probability measure $F(s|\omega) : \mathbb{R} \to \mathbb{R}_+$ for each $\omega \in \Omega$, which, combined with μ, determine both $F(s, \omega)$ and the posterior $F(\omega|s)$. It is a property of the cost function for information derived in this paper, as is true with Shannon Entropy, that if $F(s, \omega)$ is optimal, then the agent is done learning after a single signal s. After the signal is realized, the agent simply picks the action with the highest expected value:

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

Ignoring the cost of learning momentarily, the value to the agent of receiving a signal s, which induces posterior $F(\omega|s)$, is then:

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

Let the expected cost of a particular information strategy, given the agent's prior, be denoted $\mathbf{C}(F(s, \omega), \mu)$. We describe the form of the cost functions studied

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

in this paper in Section 4. The agent's problem can thus be written:

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

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

The choice behavior the agent exhibits depends on the cost function for information. Shannon Entropy is a measure of uncertainty with an axiomatic foundation that can be used to assign costs to information. If we are given a partition of the possible states of the world $\mathcal{P} = \{A_1,\,\ldots,\,A_m\}$, and probability measure μ over these events, the uncertainty about which event has occurred, as measured by **Shannon Entropy**, is defined:[5]

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

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

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

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

where $\mathcal{P} = \{\{\omega_1\},\,\{\omega_2\},\,\ldots,\,\{\omega_n\}\}$. Bayes rule, and the nature of Shannon Entropy, guarantee that every potential information strategy has a weakly positive cost.

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

Table 1: Example 1				
State:	ω_1	ω_2	ω_3	ω_4
Balls in State:	60 Blue & 40 Red	51 Blue & 49 Red	49 Blue & 51 Red	40 Blue & 60 Red
Probability of State:	1/4	1/4	1/4	1/4
Value of selecting option 1:	y	y	$-y$	$-y$
Value of selecting option 2:	0	0	0	0

2.1 Example 1: Perceptual Distance and Problems with Predictions

Caplin et al. (2017, p. 19) show that Shannon Entropy results in choice behavior that satisfies "invariance under compression." That is, when Shannon Entropy is used to measure information, if there are two states of the world, ω_1 and ω_2, across which payoffs are identical for each option ($\mathbf{v}_n(\omega_1) = \mathbf{v}_n(\omega_2)\ \forall n \in \mathcal{N}$), then the chance of each option being selected is the same in ω_1 and ω_2. The invariance under compression that is predicted by Shannon Entropy is, unfortunately, not found in many settings, as is shown by the work of Dean and Neligh (2019). The intuition for why invariance under compression may not be present in every choice environment is demonstrated by the following example.

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

The Shannon RI model, which imposes invariance under compression, predicts that the agent has the same chance of selecting option 1 when there are 40 red balls as when there are 49 red balls, and that the agent has the same chance of selecting option 1 when there are 60 red balls as when there are 51 red balls. This predicted behavior is not intuitive because it should be easier for the agent to differentiate between the

states that are more different (40 versus 60 red balls) than the states that are more similar (49 versus 51 red balls). One should instead expect that the chance that option 1 is selected is decreasing in the number of red balls, as is demonstrated by the experiments of Dean and Neligh (2019), because it should be easier to determine which color of ball constitutes the majority the more of that color ball there are.

Why does Shannon Entropy impose this type of behavior? In short, Shannon Entropy results in invariance under compression because of Shannon's third axiom (Shannon, 1948). In the context of Example 1, let $\mathcal{P} = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}\}$, and $\tilde{\mathcal{P}} = \{\{\omega_1 \cup \omega_2\}, \{\omega_3 \cup \omega_4\}\}$, be two partitions of the state space. Shannon's third axiom requires that total uncertainty about the state of the world, which is the uncertainty about which event in $\mathcal{P}$ has occurred, be equal to the uncertainty about which event in $\tilde{\mathcal{P}}$ has occurred, plus the expected amount of uncertainty that remains about which event in $\mathcal{P}$ has occurred after we have learned which event in $\tilde{\mathcal{P}}$ has occurred. This equality means that the reduction in uncertainty caused by a signal is equal to the reduction in uncertainty about which event in $\tilde{\mathcal{P}}$ has occurred, plus the expected reduction in uncertainty about which event in $\mathcal{P}$ has occurred given which event in $\tilde{\mathcal{P}}$ has occurred.

The agent is only concerned with which event in $\tilde{\mathcal{P}}$ has occurred, as this fully determines payoffs. Given which event in $\tilde{\mathcal{P}}$ has occurred, the agent does not care which event in $\mathcal{P}$ has occurred. If agent behavior is different in ω_1 compared to ω_2, or ω_3 compared to ω_4, so that their behavior does not satisfy invariance under compression, then the agent is, to an extent, differentiating between these states, and paying for information that does not benefit them, and their information strategy is thus not optimal.

While other information cost functions do not require that choice behavior satisfies invariance under compression (Caplin et al., 2017; Morris & Yang, 2016), they lack the tractability and flexibility of Shannon Entropy,[6] which limits the potential

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

for their application. This has led to the following open question: "what workable alternative models allow for the complex behavioral patterns identified in practice?" (Caplin et al., 2017, p. 2), a question that this paper attempts to answer.

2.2 Example 2: Perceptual Distance and Biases in Fitting

If different perceptual distances are present in the same choice environment, a RU model may be susceptible to a form of informational bias that has not previously been identified, as demonstrated by the following example. This is significant for those who wish to conduct welfare or counterfactual analysis because there are many economically significant examples where, for instance, one option is easier to learn about, as in Example 2.

Consider a choice environment where an agent has two options: option 1 and option 2, which can each be of high value H, or low value $L < H$, as is described in Table 2. Assume, contrary to what is possible with Shannon Entropy, that learning the value of option 1 is less costly than learning the value of option 2.[7] For example, perhaps the agent is interested in investing in one of two businesses that are *a priori* identical except for the fact that one is local and easier to learn about, while the other is foreign and harder to learn about. It is not difficult to come up with other similar examples.

Because payoffs are symmetric, any knowledge about the value of option 1 has the same value to the agent as the same knowledge about option 2. Further, the cost of said information about option 1 is lower. As such, while the marginal benefit of information about option 1 or option 2 is the same, the marginal cost of information about option 1 is lower. We should thus expect research of a rational agent to be more attentive to option 1. If the agent was deciding between investing in two businesses

[7]With Shannon Entropy it is not possible for the cost of learning the value of option 1 to differ from the cost of learning the value of option 2. Each option realizes each of its two values with equal probabilities, and with Shannon Entropy it is not possible to have different perceptual distances in the same choice environment.

Table 2: Example 2				
State:	ω_1	ω_2	ω_3	ω_4
Probability of State:	1/4	1/4	1/4	1/4
Value of selecting option 1:	H	H	L	L
Value of selecting option 2:	H	L	H	L

that are *a priori* identical, except one is local and easier to learn about, while the other is foreign and harder to learn about, then we should expect the agent to be more attentive to the local business.

If both option 1 and option 2 have realized their high value H, we should expect that the agent is more likely to select option 1. Our intuition is that the agent should be more attentive to option 1, and thus should be more cognisant of option 1's high value, and more likely to select it. Similarly, if option 1 and option 2 have both realized their low value L, then we should expect that the agent is more likely to select option 2.[8]

Because of this, if an econometrician, who does not know that the two options have the same value distribution, tried to deduce the two values of option 1, H_1 and L_1, and the two values of option 2, H_2 and L_2, using a multinomial logit regression, they would decide that H_1 is more than the true value H, and that L_1 is less than the true value L (as is shown rigorously in Section 5). Fitting thus falls prey to an informational bias, undermining the value of any counterfactual or welfare analysis.

This type of bias has not previously been identified in the literature on RI. Let $\Pr(n|\omega)$ denote the probability that the agent selects option n in state ω. Let $\Pr(n) = \sum_\omega \Pr(n|\omega)\mu(\omega)$ denote the unconditional probability that option n is selected. Matějka and McKay (2015) show that fitting of multinomial logit results in the value of an option n to be biased by $\log(\Pr(n) \cdot N)$ in all states ω, where N is the number of available options. The bias found by Matějka and McKay (2015) can be identified by examining the unconditional choice probabilities of the agent

[8] Our intuition is that the agent should be more attentive to option 1, and thus should be more cognisant of option 1's low value, and less likely to select it.

because the driving mechanism is that the cost of learning causes the agent to be biased towards options that they have a higher chance of selecting *a priori*. The bias previously found by Matějka and McKay (2015) is fundamentally different than the bias demonstrated in this example because their bias does not allow for an option to be over valued in some states and under valued in others, which is in contrast with our setting where option 1 is over valued when it is of high value, and is undervalued when it is of low value.

An econometrician who observes equal unconditional choice probabilities in this environment, as is predicted in this setting by the model developed in this paper, might be tempted conclude, based on the previous literature, that their analysis is not susceptible to informational biases since each option has the same chance of being selected *a priori*, so the bias of option n is $\log(\Pr(n) \cdot N) = \log(\frac{1}{2} \cdot 2) = 0 \ \forall n$, and thus any counterfactual or welfare analysis that they conduct is valid. This conclusion may not be correct given the results in this paper.

Further, RU models and RI models with Shannon Entropy can both be rejected for RI with MSSE in this environment if we are able to alter the correlation between the values of the two options. If a RU model describes the agent, then changing the correlation between the values of the two options would not change the choice behavior of the agent. If the behavior of the agent is instead described by MSSE, then changing the correlation between the values of the two options would change the choice behavior of the agent in individual states. This effect is because the total information that can be acquired from learning the value of option 1 (the option that is easier to learn about) changes with the correlation of the options' values. Further, if the above MSSE specification is correct, the unconditional choice probabilities of the agent would remain constant when correlation is changed due to the symmetry of the environment, as long as the agent is doing some learning.[9] Finally, if the behavior of the agent is instead described by Shannon Entropy, then the choice behavior in the

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

individual states could only change if the unconditional choice probabilities changed, which is not the case with MSSE. With MSSE, since choice probabilities in a state can be impacted by choice probabilities that are conditioned on some larger subset of states, not only payoffs and unconditional choice probabilities, choice probabilities in a state can change even when payoffs and unconditional choice probabilities do not.

3 Multisource Shannon Entropy (MSSE)

In this section we use axioms to develop this paper's measure of uncertainty. The goal of our axioms are to measure the total amount of uncertainty, which is the expected cost to the agent of perfectly observing the state of the world. The measure of total uncertainty that we develop can then be used to study a rationally inattentive agent because the cost of a noisy information strategy can be taken to be the expected reduction in total uncertainty, as is frequently done with Shannon Entropy in models of RI. Thus, while this paper is interested in studying an inattentive agent that only partially learns about the state of the world, this section discusses an attentive agent that perfectly observes the state of the world.

3.1 Formal Setting

As was mentioned in Section 2, we are interested in an agent who is researching a measurable space $(\Omega, \mathcal{F})$. Ω is a finite set of possible states of the world. $\mathcal{F}$ is the set of events generated by Ω.

One natural way to think about an agent learning is through a series of questions that have answers that are uniquely determined by the state of the world. These are questions that you can answer if you know the state of the world. How do we model such questions? 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. Notice that our definition of a partition excludes trivial partitions that only contain a single event.

A question with multiple potential answers is thus equivalent to a partition whenever the answer to the question is deterministically determined by the state of the world. This equivalence occurs since every state space we consider has finite possible states of the world, so every such question must have a finite number of answers, and we can simply group states of the world based on the answer to the question they produce. Because we are concerned with questions that have answers that are deterministically determined by the state of the world, the words 'question' and 'partition' can be used interchangeably.

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 we 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' can thus be used interchangeably.

If $\omega \in \Omega$ is the state of the world, let the **realized event** of the partition $\mathcal{P} = \{A_1, \ldots, A_m\}$ be denoted by $\mathcal{P}(\omega)$, that is $\mathcal{P}(\omega) = A_i \in \{A_1, \ldots, A_m\}$ iff $\omega \in A_i$. Given a probability measure $\mu : \mathcal{F} \to \mathbb{R}_+$, and some partition $\mathcal{P}$, let $C(\mathcal{P}, \mu) \in \mathbb{R}_+$ denote the cost of learning the realized event $\mathcal{P}(\omega)$ of $\mathcal{P}$. $C(\mathcal{P}, \mu)$, 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 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 can be used interchangeably. If a learning strategy consists of only binary partitions, we 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. 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.

Define $C(S,\, \mu)$, 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 expected costs of each of the questions in S:

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

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

If $\mathcal{P} = \{A_1,\, \ldots,\, A_m\}$ is a partition, let $\sigma(\mathcal{P})$ denote the **σ-algebra generated by** $\mathcal{P}$, which is the smallest σ-algebra that contains all the events $A_1,\, \ldots,\, A_m$ in $\mathcal{P}$ (which is also the power set of the events in $\mathcal{P}$, since $\mathcal{P}$ is a partition). In general, if B is any collection of partitions, let $\sigma(B)$ denote the **σ-algebra generated by** B, which is the smallest σ-algebra containing all the events in each of the partitions in B. Since a learning strategy $S = (\mathcal{P}_1,\, \ldots,\, \mathcal{P}_n)$ is a collection of partitions, we thus use $\sigma(S)$ to denote the σ-algebra generated by S.

Sometimes a single question can be as informative as several questions. We say a learning strategy S is **equivalent** to a partition $\mathcal{P}$ if $\sigma(S) = \sigma(\mathcal{P})$, and we say that a series of questions is equivalent to a particular question if the learning strategy that represents the series of questions is equivalent to the partition that represents the particular question. 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 $\mathcal{P}$?'. We can thus read $\sigma(S) = \sigma(\mathcal{P})$ as saying

that, for all priors, S and $\mathcal{P}$ provide the same amount of information to the agent. Let $S(\mathcal{P})\{S|\sigma(S) = \sigma(\mathcal{P})\}$ denote the set of learning strategies that are equivalent to $\mathcal{P}$, and let $S^b(\mathcal{P}) = \{S^b|\sigma(S^b) = \sigma(\mathcal{P})\}$ denote the set of binary learning strategies that are equivalent to $\mathcal{P}$. We 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 $\tilde{\mathcal{P}}$. If a partition $\tilde{\mathcal{P}}$ is in a learning strategy S which is equivalent to $\mathcal{P}$, then notice that $\tilde{\mathcal{P}}$ must be coarser than $\mathcal{P}$.

3.2 Axioms

What form should a cost function for information take? This difficult question does not have an obvious answer, so this paper takes an axiomatic approach. The axioms make explicit the structure that is imposed on our cost function. Each axiom is meant to be normatively appealing, and 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 MSSE and standard Shannon Entropy.

When an agent learns in an inattentive fashion, and only acquires some of the available information, they reduce the amount that remains to be learned, and thus reduce the subsequent cost of learning the state of the world. The cost of the inattentive learning done by the agent can thus simple be measured as the reduction in the cost of learning the state of the world, as subsequent sections discuss,[10] as long as we can establish the cost of learning the state of the world for different probability measures.

Thus, while the learning of an agent is frequently inattentive, and this paper wishes to study environments where the agent only partially learns about the state of the world, this section discusses an attentive agent that tries to perfectly observes

[10]We argue later in the paper that the application of Shannon Entropy can be interpreted in this same fashion.

the state of the world. We do this because we want our axioms to be normatively appealing, and we find axioms about perfectly observing the state of the world to be a more intuitive, and hence easier to evaluate normatively, than axioms that focus directly on inattentive behavior, describing costs of different kinds of stochastic experiments. Another interpretation of this strategy is that, while the primitive of our model is the cost of learning the realized events of partitions, and the agent could choose to learn through such partitions of the state space, we do not constrain the agent's choice of information strategy so that they must learn through such partitions, and they can instead choose a noisy signal structure if they desire.

Before we introduce our axioms, we pause to discuss learning strategy invariance, a concept that helps us to make it explicit what we are assuming with our axioms, and that is the central pillar of Shannon's (1948) 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, and how they are ordered in S. 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. If a question has this strong property, we say it is learning strategy invariant. Formally, we 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 questions that are not learning strategy invariant. Consider the environment described in Example 2 in Section 2.2. In this context, let $A_1 = \{\omega_1, \omega_2\}$, $A_2 = \{\omega_1, \omega_3\}$, $\mathcal{P}_1^b = \{A_1, A_1^c\}$, and $\mathcal{P}_2^b = \{A_2, A_2^c\}$. Notice that observing the realized event of $\mathcal{P}_1^b$ is equivalent to learning the value of option 1, and observing the realized event of $\mathcal{P}_2^b$ is equivalent to learning the value of option 2. Now, let $\mathcal{P}_3 = \{\{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_4\}\}$ denote our partition of the state space. Notice that the learning strategy $S^b = (\mathcal{P}_1^b, \mathcal{P}_2^b)$ is equivalent to $\mathcal{P}_3$, because if we answer 'What is the value of option 1?', and then answer 'What is the value of option

2?', we have observed the state of the world.

Based on our discussion in Section 2.2, however, we should expect that $\mathcal{P}_3$ may not be learning strategy invariant. Consider $\tilde{S}^b = (\mathcal{P}_2^b, \mathcal{P}_1^b)$, which is also equivalent to $\mathcal{P}_3$. If the value of option 1 and option 2 were perfectly correlated, then observing the value of one of them would tell you the value of the other. The cost of S^b would then be the cost of observing the value of option 1, which we assumed to be less than the cost of observing the value of option 2, which is then the cost of $\tilde{S}^b$.

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

We now begin to state the five axioms required to achieve this paper's measure of uncertainty:

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 $\mu(A_2)$, and we can thus write $C(\mathcal{P}^b, \mu) = C(\mathcal{P}^b, \mu(A_1), \mu(A_2))$.

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

We begin to use our axioms by showing that if $\mathcal{P}$ is a leaning strategy invariant partition comprised of three or more events, then $C(\mathcal{P}, \mu)$ is constant with respect to permutations of the probability measure μ on $\mathcal{P}$. If $\mathcal{P} = \{A_1, \ldots, A_m\}$ is a leaning strategy invariant partition, we 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 $\forall i \in \{1, \ldots, m\}$, $\mu(A_i) = \tilde{\mu}(A_{\pi(i)})$.

Lemma 1. If a partition $\mathcal{P} = \{A_1, \ldots, A_m\}$ is learning strategy invariant, with $m \geq$

3, and C satisfies Axiom 1, then $C(\mathcal{P}, \mu)$ is fully determined by $\mu(A_1)$, $\mu(A_2)$, ..., and $\mu(A_m)$, and if $\tilde{\mu}$ is a permutation of μ on $\mathcal{P}$, then $C(\mathcal{P}, \mu) = C(\mathcal{P}, \tilde{\mu})$.

Proofs for results in Section 3 and can be found in Appendix 1

We next show that if a partition $\mathcal{P} = \{A_1, \ldots, A_m\}$ is learning strategy invariant with $m \geq 3$, structure is imposed onto $C(\mathcal{P}^b, \mu)$ for all $\mathcal{P}^b$ coarser than $\mathcal{P}$. As it turns out, this structure is quite helpful.

Lemma 2. If a partition $\mathcal{P} = \{A_1, \ldots, A_m\}$ is learning strategy invariant with $m \geq 3$, and $\mathcal{P}^b$ is a binary partition coarser than $\mathcal{P}$, then if 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 $p_1 + p_2 + p_3 = 1$:

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

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

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

Our next axiom, Axiom 2, is concerned with the subjective nature of the state space. In practice, 'the' state space Ω is determined by the researcher and the application, and referring to it as 'the' state space is typically a misnomer. Sometimes the researcher uses just enough states so that the payoff function is measurable, as in Example 2, and other times the researcher includes more states than are required for measuring the payoff function because it is deemed relevant to the agent, as in Example 1. In Example 1, however, the four 'states' we describe are not actually states, they are events in some richer underlying state space. Two realizations of the 'state' of the world in which 51 blue balls appear may differ from each other because the balls may appear in a different order, which may be relevant for the agent's cost of learning.[11] Further, a change in the payoff function may necessitate the description

[11] Imagine that the 100 balls from Example 1 are displayed on the screen in ten rows of ten, and

of a richer state space.

While our description of the state space can change, the reality of the agent does not, and the cost of asking certain questions should not change based on the subjective modelling decisions of the researcher. Thus, when we consider the cost of learning the outcomes of a binary partition $\mathcal{P}^b$, if these costs do not satisfy the equations outlined in Lemma 2, then we are ruling out that a finer state space could be defined with a learning strategy invariant partition with three or more states that our binary partition $\mathcal{P}^b$ is coarser than. This notion is formalized in Axiom 2.

Axiom 2 (Subdivision): 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 $p_1 + p_2 + p_3 = 1$, we assume C is such that it allows for there to be a partition comprised of three of more events which $\mathcal{P}^b$ is coarser than, which is to say:

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

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

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

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

that 51 are blue. If the position of the red and blue balls appears random, it would seem intuitive that it is more costly for the agent to learn whether or not the the majority of the ball are blue compared to the setting where the top five rows all consist of ten blue balls, and in the bottom ten rows there is one blue ball and the rest are red.

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 our first three axioms, it is continuous everywhere, as is formalized by Lemma 3, which further shows that the cost function is permutation invariant on binary partitions.

Lemma 3. 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 are not the only helpful properties imposed onto our cost function by our axioms. On binary partitions, our cost function is also non-decreasing if the chance of whichever event is less likely increases.

Lemma 4. 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 with small increases in p.

Our first three axioms do not rule out that learning with a binary partition can be costless. We, however, wish to study a costly learning environment, so any time answering a question changes the agent's beliefs we think it should be costly to the agent.

Axiom 4 (Costly Learning): Given a binary partition $\mathcal{P}^b$, if $p \in (0, 1)$, then $C(\mathcal{P}^b, p, 1-p) > 0$.

We are now ready to show that the cost of learning with a learning strategy invariant partition is dictated by Shannon Entropy.

Lemma 5. If a partition $\mathcal{P}$ is learning strategy invariant, and C satisfies Axiom 1, Axiom 2, Axiom 3, and Axiom 4, 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 information source that allows the agent to differentiate between the events that comprise the partition. Shannon (1948) imposes learning strategy invariance onto all partitions of Ω, 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 from a single information source that allows them to differentiate between the different states of the world. With MSSE, 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 different pieces of information from different sources depending on which source allows them to acquire the information at the lowest costs, as is formalized by Theorem 1. This interpretation is how MSSE gets its name.

Shannon's (1948) key axiom, his third axiom, assumes that all partitions of the state space are learning strategy invariant, and further, that the cost function derived is defined for vectors of arbitrary length, even though Shannon also uses a finite state space to derive his cost function. In addition to this 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). 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 leaning with binary partitions is assumed to be costly when there is uncertainty about their realized

event (as we do in Axiom 4). Removing his first axiom only has an impact if we 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 always costly, it must be Shannon's third axiom that is weakened.

Our last axiom, Axiom 5, asserts that it is without loss for us to think about an agent learning through binary partitions. Since binary partitions are learning strategy invariant, it is thus without loss for us to consider the agent learning through learning strategy invariant partitions, and Lemma 5 becomes quite useful.

Axiom 5 (Efficient Yes or No Questions): Given a partition $\mathcal{P}$, for all probability measures μ:

$$C(\mathcal{P}, \mu) \geq \min_{S^b \in S^b(\mathcal{P})} C(S^b, \mu).$$

In plain language, Axiom 5 says that for any question $\mathcal{P}$, there are a series of yes or no question S^b, that provide the same amount of information as $\mathcal{P}$, and can be asked instead for the same cost or less. This assertion allows us to focus on series of yes or no questions without loss when we try to determine the cost to the agent of learning the state of the world, which is supported by research in the psychology and psychophysics literatures.

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).

3.3 Total Uncertainty

Lemma 5 tells us that for each binary partition $\mathcal{P}^b$, there is an **associated multiplier**, $\lambda(\mathcal{P}^b) \in \mathbb{R}_{++}$, such that for all probability measures μ: $C(\mathcal{P}^b, \mu) = \lambda(\mathcal{P}^b)\mathcal{H}(\mathcal{P}^b, \mu)$. Since there are a finite number of binary partitions of Ω, we 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 we let M=1.[12] 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.

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 we let $M = 2$. 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.

Continue in this fashion until we let λ_M denote 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 our notation more compact, we can use a group of partitions 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 $j \in \{1, \ldots, M\}$,[13] let $\mathcal{P}_{\lambda_j} = \times\{\mathcal{P}_i^{b,\lambda_j}\}_{i=1}^{n_j}$.

[12] If M=1, then MSSE collapses to standard Shannon Entropy.
[13] M is defined in the proceeding paragraphs.

MSSE incorporates different perceptual distances because it allows for different events to be different distances from each other. Events in $\mathcal{P}_{\lambda_1}$, for instance, have greater perceptual distances between them than events in $\mathcal{P}_{\lambda_M}$ (assuming $M > 1$).

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

Theorem 1. If C satisfies all five axioms, then there exist partitions $\mathcal{P}_{\lambda_1}, \ldots, \mathcal{P}_{\lambda_M}$ as defined above, and constants strictly positive constants $\lambda_1 < \ldots < \lambda_M$, such that for any probability measure μ on $\mathcal{F}$

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

where $\mathcal{H}$ is defined as in equation (1).

In plain language, Theorem 1 says that if the cost of learning satisfies all five axioms, then the cheapest way (in expectation) to learn the state of the world always involves first asking all the yes or no questions with the lowest associated multiplier (in any order), then asking all the yes or no questions with the second lowest multiplier, and continuing in this fashion until the state of the world has been realized.

Theorem 1 generates the more flexible measure of uncertainty that we desired for studying inattentive behavior. If the agent starts with a prior μ, and does optimal learning that reaches a posterior $\tilde{\mu}$, then we let the cost of this inattentive research be the reduction in the cost of perfectly learning the state of the world, as is discussed in the next section.

In terms of Shannon's original context, this paper's model can be thought of as describing learning of information from M sources, where source i, for $i \in \{1, 2, \ldots, M\}$, is capable of providing information about $\mathcal{P}_{\lambda_i}(\omega)$. Shannon's original axioms, in contrast, impose that all partitions $\mathcal{P}$ are learning strategy invariant, which is analogous to all binary partitions having the lowest multiplier, and there only being one information source relevant for learning.

The $\mathcal{P}_{\lambda_i}$'s that could be used in Theorem 1 are not unique, with the exception of $\mathcal{P}_{\lambda_1}$. The versions described in the paragraphs preceding Theorem 1 are the unique coarsest partitions that could be used in the statement of the theorem. For $i \in \{2, \ldots, M\}$, $\mathcal{P}_{\lambda_i}$ could, for instance, be replaced by $\tilde{\mathcal{P}}_{\lambda_i} = \times\{\mathcal{P}_{\lambda_j}\}_{j=1}^{i}$ in the statement of Theorem 1, which would constitute the unique finest representation of the partitions.

The axiomatic derivation of the cost benchmark in this paper requires a discrete state space for the state of the world, as is the case with Shannon Entropy. If a continuous state space is desired for the state of the world, however, a measure of uncertainty for a continuous state space can be defined in an analogous manner to the measure of uncertainty defined in Theorem 1 for a discrete state space, which is similar to what is done by Shannon (1948) to apply Shannon Entropy in a continuous setting.

4 Inattentive Learning with MSSE

The following section introduces and solves a model of RI that uses MSSE to measure the cost of acquiring information. We establish that our new more flexible measure of uncertainty can still be incorporated tractably into a model of RI, which is not an obvious result. Apart from the use of MSSE instead of Shannon Entropy for the measurement of uncertainty, this section follows the work of Matějka and McKay (2015) closely so as to aid comparison between the two models.

Given our result in Theorem 1, we take the expected cost of a particular information strategy to be defined as:

$$\mathbf{C}(F(s, \omega), \mu) = \mathbb{E}\left[\min_{S \in S(\Omega)} C(S, \mu) - \min_{S \in S(\Omega)} C(S, \mu(\cdot|s))\right].$$

A noisy information strategy reduces the total amount of uncertainty, and we thus measure the cost of such a noisy information strategy as the expected reduction in

total uncertainty. This interpretation can also be applied to RI models that use Shannon Entropy to measure the cost of noisy information structures. Shannon Entropy is a measure of total uncertainty derived from axioms about the cost of successively learning the realized events of partitions, and in such models the cost of a noisy signal is simply taken to be the reduction in total uncertainty, as measured by Shannon Entropy.

The cost functions that can be defined as above with MSSE are in the class of uniformly posterior-separable cost functions described by Caplin et al. (2017). The behavior generated in static settings by such posterior-separable cost functions has been shown to be equivalent to the behavior generated by sequential information sampling in some dynamic contexts (Hébert & Woodford, 2017; Morris & Strack, 2019). In particular, Hébert and Woodford (2017) show that a class of static cost functions, which they call 'neighborhood-based' cost functions, can be micro-founded in this way. The cost functions explored in this paper that measure the reduction in MSSE are a strict subset of the neighborhood-based cost functions described in their paper, and thus the cost functions in this paper are micro-founded in two ways, directly through the axioms in this paper, and indirectly through the dynamic analysis conducted by Hébert and Woodford (2017). While symmetry imposes a unique set of partitions in Example 1 when MSSE is used, there are numerous representations that can be used when a neighborhood-based cost function is assumed. Hébert and Woodford (2017) suggest two ways of modelling the neighborhoods in such a setting, one of which is fitted by Dean and Neligh (2019), and neither of which is equivalent to the partitions suggested by MSSE.

Huettner et al. (2019), in turn, create an ad hoc group of cost functions that are also a generalization of Shannon Entropy, but are a strict subset of the cost functions studied in this paper that measure reduction in MSSE. The cost functions studied by Huettner et al. (2019) allow for multiple perceptual distances, but are not capable of predicting the behavior we argued was intuitive in Example 1, since in Example 1

their cost functions collapses to standard Shannon Entropy.

4.1 Rationally Inattentive Agent's Problem

As was discussed in Section 2, when the agent faces a probability space $(\Omega, \mathcal{F}, \mu)$ and a set of options N, the agent's problem is to maximize the expected value of the option they select less the cost of learning by choosing an optimal information strategy, and subsequently selecting an option based on the signal produced by their information strategy. The agent's problem can thus be written:

$$\max_{F \in \Delta(\mathbb{R} \times \Omega)} \sum_{\omega \in \Omega} \int_s V(s|F) F(ds|\omega) \mu(\omega) - \mathbf{C}(F(s, \omega), \mu), \tag{2}$$

$$\text{such that } \forall \omega \in \Omega : \int_s F(ds, \omega) = \mu(\omega). \tag{3}$$

The above problem is complicated and not particularly tractable, so we follow Matějka and McKay (2015) and re-write this problem directly in terms of the choice probabilities of the agent. This process requires the development of some new notation. Define $S(n|F) = \{s \in \mathbb{R} : F(s) > 0, a(s|F) = n\}$, to be the set of signals that result in the agent selecting option n. Next, as was done in Section 2, define the chance of option n being selected conditional on the state of the world to be:

$$\Pr(n|\omega) = \int_{s \in S(n|F)} F(ds|\omega), \tag{4}$$

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

$$\Pr(n|A) = \sum_{\omega \in A} \Pr(n|\omega) \mu(\omega|A). \tag{5}$$

Define the **unconditional choice probability** of option n to be:

$$\Pr(n) = \sum_{\omega \in \Omega} \Pr(n|\omega)\mu(\omega). \tag{6}$$

Denote the collection $\{\Pr(n|\omega)\}_{n=1}^{N}$ by $\mathbb{P}$. Using this notation, we can re-write the agent's problem:

Lemma 6. Choice probabilities $\mathbb{P}$ are the outcome of a solution to the agent's problem in (2) subject to (3) iff they solve:

$$\max_{\mathbb{P}} \sum_{n \in \mathcal{N}} \sum_{\omega \in \Omega} \mathbf{v}_n(\omega)\Pr(n|\omega)\mu(\omega) - \mathbf{C}(\mathbb{P},\, \mu), \tag{7}$$

$$\text{such that: } \forall\, n \in \mathcal{N},\ \Pr(n|\omega) \geq 0,\ \forall\, \omega \in \Omega, \tag{8}$$

$$\text{and } \sum_{n \in \mathcal{N}} \Pr(n|\omega) = 1\ \forall\, \omega \in \Omega, \tag{9}$$

where $\mathbf{C}(\mathbb{P},\, \mu)$ is as defined in Lemma 14.

Proofs for results in Section 4 and Section 5 can be found in Appendix 2

This new problem, where the agent selects their conditional choice behavior $\mathbb{P}$, is substantially easier to solve than the problem where the agent picks their information strategy $F(s,\, \omega)$.

4.2 Behavior of a Rationally Inattentive Agent

Using Lemma 6, we can establish a necessary condition for the optimal behavior of the agent with Theorem 2, and then use said necessary condition to simplify the maximization problem undertaken by the agent with Corollary 1.

Theorem 2:

If $\mathbb{P}$ is the solution to (7) subject to (8) and (9), then $\forall\, n \in \mathcal{N}$, and $\forall\, \omega \in \Omega$, the probability that option n is selected in state w satisfies:

$$\Pr(n|\omega) = \frac{\Pr(n)^{\frac{\lambda_1}{\lambda_M}} \Pr(n|\mathcal{P}_{\lambda_1}(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \ldots \Pr(n|\cap_{i=1}^{M-1} \mathcal{P}_{\lambda_i}(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}}{\sum_{\nu\in\mathcal{N}} \Pr(\nu)^{\frac{\lambda_1}{\lambda_M}} \Pr(\nu|\mathcal{P}_{\lambda_1}(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \ldots \Pr(\nu|\cap_{i=1}^{M-1} \mathcal{P}_{\lambda_i}(\omega))^{\frac{\lambda_M-\lambda_{N-1}}{\lambda_M}} e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_M}}}. \tag{10}$$

Those familiar with the work of Matějka and McKay (2015) will recognize the above formula as the MSSE analogue of Matějka and McKay (2015)'s Theorem 1. When all partitions are learning strategy invariant, $\lambda_1 = \lambda_2 = \cdots = \lambda_M$, and the above formula collapses to Matějka and McKay (2015)'s Theorem 1.

With standard Shannon Entropy, the chance that the agent selects an option thus depends only on the unconditional chances of the options being selected, and the realized values of the options. With MSSE, in contrast, as the above formula indicates, the chance that the agent selects an option n in a particular state of the world ω depends on the unconditional chances of the options being selected, $\Pr(n)$, the realized values of the options $\mathbf{v}_n(\omega)$, as well as the probabilities of the options being selected in similar states of the world. Here 'similar states of the world' refers to states that induce the same realization of partitions with associated multipliers smaller than λ_M. It makes sense that when easier to observe pieces of information indicate that an option n is likely of above average value, that the agent should select option n with a higher probability, even if the above average value has not been realized. For a more complete discussion of the intuitive properties of the choice behavior described in Theorem 2, please see Appendix 3.

Behavior that is consistent with Theorem 2 is not necessarily optimal because in many settings it is not optimal for the agent to consider all of the available options (choose them with positive probability), and though such a corner solution may be optimal, there are many corners that are consistent with Theorem 2 but are not

optimal. For instance, for any $n \in \mathcal{N}$, if the agent selects n with probability one in all states of the world, then their behavior is consistent with Theorem 2, but it is easy to come up with examples where this would not be optimal for any n.

Corollary 1:

Conditional and unconditional choice probabilities described in (5) and (6) are a solution to (7) subject to (8) and (9) iff they comply with Theorem 2 and solve:

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

such that:

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

Corollary 1 is helpful because it reduces the number of choice variables faced by the agent, which means it is easier for the researcher to find optimal agent behavior. When solving the problem described in Lemma 6, the agent must choose $\Pr(n|\omega)$ for all n and ω. When solving the problem in Corollary 1, the agent must only choose $\Pr(n|A)$ for all n and $A \in \times\{\mathcal{P}_{\lambda_i}\}_{i=1}^{M-1}$, which is a coarser partition. In Example 2, for instance, if the agent tries to solve Lemma 6 they must pick their probabilities of selecting option 1 and option 2 in four different states of the world, while if they solve the problem in Corollary 1 they must only pick their probabilities of selecting option 1 and option 2 in two events, and then Theorem 2 dictates their choice probabilities in each state of the world. This reduction makes finding optimal behavior of the agent easier for the researcher because there are thus half as many choice variables when analysing Example 2 if Corollary 1 is used instead of Lemma 6.

Any choice behavior that complies with Corollary 1 and Theorem 2 is optimal. This paper does not provide conditions for optimal behavior that are both necessary and sufficient, however, as is done by Caplin, Dean, and Leahy (2018) in a setting with standard Shannon Entropy. This may cause some to view finding optimal behavior

with MSSE quite daunting. The necessary and sufficient conditions given by Caplin et al. (2018) in the setting with Shannon Entropy are appealing because they verify if behavior that satisfies the necessary conditions are in fact optimal, and provide insight into the formation of the agent's optimal consideration set, which is interesting in and of itself. In the more complicated setting studied in this paper, the necessary and sufficient conditions are less appealing. The reality is that almost any problem in this more complicated setting requires a computer for finding optimal behavior, but that is true in the standard setting as well, even with the conditions derived by Caplin et al. (2018). The good news is that the optimization problem that needs to be solved involves maximization of a strictly concave function over a compact domain, which is differentiable everywhere on the interior. Thus, standard steepest accent algorithms work well for solving the problem described in Corollary 1, even when the number of options in $\mathcal{N}$ and the number of events in $\times\{\mathcal{P}_{\lambda_i}\}_{i=1}^{M-1}$ are large. For those that are interested in a discussion of how MSSE changes the formation of optimal consideration sets, please see Walker-Jones (2019). Further, while Huettner et al. (2019) do attempt to provide necessary and sufficient conditions for optimal behavior in their setting, the conditions are incorrect, as is also discussed by Walker-Jones (2019).

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

5 Comparisons with the Standard Model

In this section we compare and contrast the choice behavior that is produced by RI with Shannon Entropy and the choice behavior produced by the RI model developed in Section 4 that uses the MSSE measure developed in Section 3. We first discuss the relationship between RU models and RI with MSSE, and then revisit the two motivating examples, Example 1 and Example 2, from Section 2.

5.1 Comparison with Random Utility Model

It is standard practice to use a RU model to describe discrete choice settings. In such a model, the agent picks the option with the largest sum $u_n = v_n + \epsilon_n$ over all options $n \in \mathcal{N}$. Generally, u_n represents the value of the option to the agent, v_n represents the average value of the option across agents, and ϵ_n represents an idiosyncratic value to the agent. The role ϵ_n plays is up to interpretation, however, and is determined by the researchers specification (Train, 2009). In a setting where agents are thought to be rationally inattentive, the above terms are interpreted in a different way because the agent's noisy behavior is generated by perceptual error instead of idiosyncratic differences in taste. In such settings, u_n represents the perceived value to the agent, v_n represents the true value to the agent, and ϵ_n is interpreted as an unobservable perceptual error that results from the noisy information strategy selected by the agent. Woodford (2014) argues that this latter interpretation is necessary in many contexts due to the stochastic responses observed in perceptual discrimination tasks such as those administered by Dean and Neligh (2019), which are akin to our Example 1 in Section 2.1. While the interpretation of ϵ_n is relevant for welfare analysis, it is inconsequential for the description of choice behavior. How then can MSSE be interpreted in terms of an RU framework, and what insights may be provided about the fitting of RU models?

Matějka and McKay (2015) point out that choice probabilities predicted by RI

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

Rational inattention with MSSE takes this one step further, as is shown by Theorem 3, allowing the shift in perceived value to also depend on easier to observe information sources (binary partitions associated multipliers that are less than λ_M). This flexibility seems natural in many real world environments. Consider an agent that is trying to select a restaurant to go to. One may expect that the chance of the agent selecting a given option to increase not only with the quality of the restaurant, and their prior impression of it, but also with easy to observe information such as on-line ratings the restaurant may have received.

Theorem 3:

The choice behavior described by $\mathbb{P}$, a solution to (7) subject to (8) and (9), is identical to the behavior produced by an RU model where each option $n \in \mathcal{N}$ has perceived value:

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

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

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

Theorem 3 is meant to provide insight into the outcome of attempting to fit a RU model in an environment where agents are rationally inattentive with a cost function for information described by MSSE. Theorem 3 does not say that a model of RI with MSSE is equivalent to a RU model. Even if choice data from a given choice problem cannot be used to reject one for the other, across choice problems MSSE produces behavior that can reject the hypothesis of a RU model. With MSSE, for instance, as with standard Shannon Entropy, adding an option can increase the

chance of an existing option being selected, which is not possible with a RU model.

Also, it is worth mentioning that since optimal behavior may result in some options being selected with probability zero, Theorem 3 implicitly defines each α_n on the extended reals so that $\alpha_n = -\infty$ if $\Pr(n) = 0$.[14]

5.2 Example 1 Revisited

We now revisit Example 1 from Section 2.1, which is described in Table 1. It seems natural that it should be easier for the agent to answer the question 'Are 60 of the balls blue?', than it is for them to answer 'Are 51 or more of the balls blue?'. Similarly, it seems natural that it should be easier for the agent to answer the question 'Are 60 of the balls red?', than it is for them to answer 'Are 51 or more of the balls red?'. Symmetry also means that the questions 'Are 60 of the balls blue?' and 'Are 60 of the balls red?' should have the same expected cost, and the questions 'Are 51 or more of the balls blue?' and 'Are 51 or more of the balls red?' should have the same expected cost. We can thus assume $\mathcal{P}_{\lambda_1} = \{A_1, A_2, A_3\} = \{\{\omega_1\}, \{\omega_2 \cup \omega_3\}, \{\omega_4\}\}$, and $\mathcal{P}_{\lambda_2} = \{\{\omega_1 \cup \omega_2\}, \{\omega_3 \cup \omega_4\}\}$.

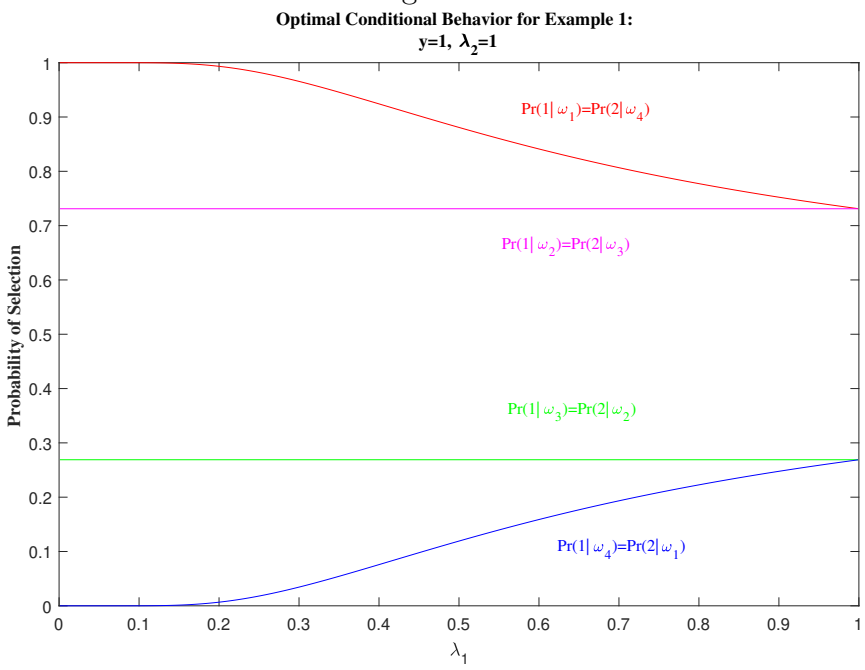
Solutions to Corollary 1 combined with Theorem 2 mean that the chance of the agent selecting option 1 is increasing in the number of blue balls, as can be seen in Figure 1, which depicts optimal behavior in each state of the world for a range of λ_1. When λ_1 is small relative to λ_2 the agent chooses option 1 in state ω_1 with a high probability, and choose option 2 in state ω_4 with a high probability. The agent is thus better able to discern the state of the world when there are 40 of one color ball and 60 of the other than when there are 49 of one color and 51 of the other. This is supported by the experimental work of Dean and Neligh (2019), and is in contrast with the behavior predicted by a model of RI that uses Shannon Entropy.

Morris and Yang (2016) identify a related issue with Shannon Entropy's lack

[14] It can be shown that if optimal behavior results in $\Pr(n) > 0$, then $\Pr(n|\omega) > 0 \ \forall \omega \in \Omega$. See (Walker-Jones, 2019).

of perceptual distance, and warn against its use in some continuous settings because it predicts discontinuous changes in behavior at places where payoffs change discontinuously. In the limit, as the number of different perceptual distances is allowed to grow, MSSE can be used to produce the kind of continuous behavior that Morris and Yang (2016) desire.

Figure 1:

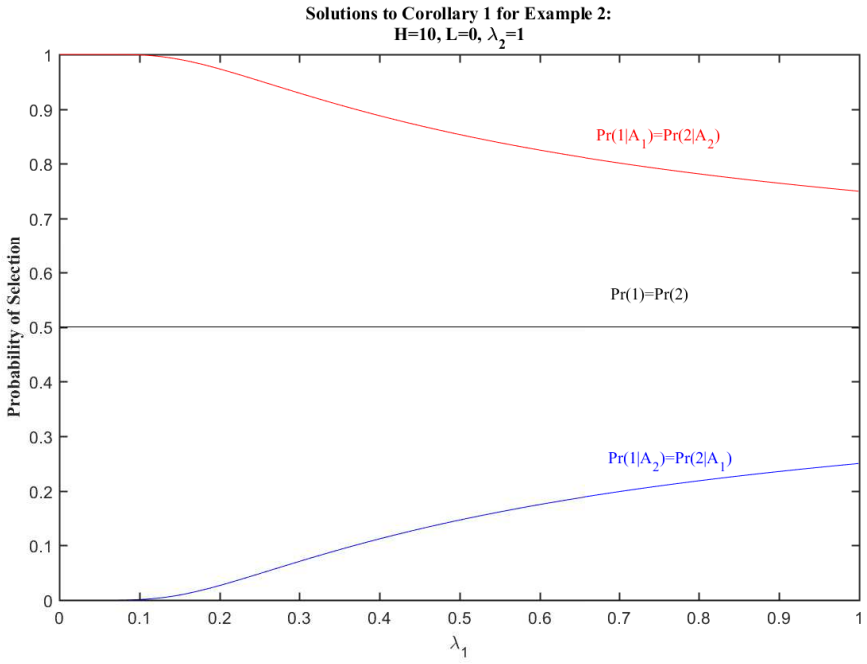


5.3 Example 2 Revisited

We now revisit Example 2 from Section 2.2, which is described in Table 2. We assumed that learning the value of option 1 is less costly than learning the value of option 2. That is to say, answering the question 'Is option 1 of value H?' has a lower expected cost to the agent than the question 'Is option 2 of value H?'. We can thus assume: $\mathcal{P}_{\lambda_1} = \{A_1, A_2\} = \{\{\omega_1 \cup \omega_2\}, \{\omega_3 \cup \omega_4\}\}$, and $\mathcal{P}_{\lambda_2} = \{\{\omega_1 \cup \omega_3\}, \{\omega_2 \cup \omega_4\}\}$.

Solutions to Corollary 1 in this environment for a range of λ_1 can be found in Figure 2, which shows that when λ_1 is small compared to λ_2, the agent selects option

Figure 2:



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

Solutions to Corollary 1 combined with Theorem 3 mean that if an econometrician tries to fit this environment with a multinomial logit model that their estimate of H_1, the high value of option 1, is biased upwards by $\frac{\lambda_2-\lambda_1}{\lambda_2}\log(2\Pr(1|A_1))$, which is greater than zero since $\Pr(1|A_1) > 1/2$, and their estimate of L_1, the low value of option 1, is biased downwards by $\frac{\lambda_2-\lambda_1}{\lambda_2}\log(2\Pr(1|A_2))$, which is less than zero since $\Pr(1|A_2) < 1/2$. These biases are despite the fact that the unconditional chance of

either option being selected is the same: $\Pr(1) = \Pr(2) = 1/2$. As such, the econometrician may have believed their analysis was not susceptible to informational biases if they had used Shannon Entropy to model the environment.

6 Conclusion

Rational inattention models that use Shannon Entropy to measure the cost of learning demonstrate that informational biases in random utility models can be significant for welfare and counterfactual analysis. The biases that have previously been identified in the literature are independent of the realized state of the world, depending only on the agent's prior about the environment. These previously identified biases manifest themselves in the unconditional choice probabilities of the agent.

This paper contributes to the literature by proposing and axiomatizing a new measure of uncertainty that features perceptual distance, maintains much of the tractability of Shannon's standard measure, and identifies a new kind of informational bias. The new form of bias can be present even when the agent has the same unconditional chance of selecting each option, which may seem to indicate an unbiased environment based on the previous literature.

Appendix 1

Before we prove Lemma 1, we show some other useful results:

Lemma 7. 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 $\tilde{\mathcal{P}} = \mathcal{P}$ we are done.

If $\tilde{\mathcal{P}} \neq \mathcal{P}$, then the definition of learning strategy invariance tells us 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:

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

and,

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

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 8. 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 $A_i \in \{A_1, \ldots, A_m\}$, then $C(\mathcal{P}, \mu) = 0$.

Proof. Suppose $\mathcal{P} = \{A_1, \ldots, A_m\}$ is a learning strategy invariant partition of the state space Ω, with $m \geq 3$, and there is an $A_i \in \{A_1, \ldots, A_m\}$ 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 tells us $C(S_1, \mu) = C(S_2, \mu)$, so $C(\mathcal{P}, \mu) = 0$.■

Proof of Lemma 1. Suppose $\mathcal{P} = \{A_1, \ldots, A_m\}$ is a learning strategy invariant

partition of the state space Ω with $m \geq 3$. The definition of learning strategy invariance implies $C(\mathcal{P}, \mu)$ is fully determined by expected learning costs on binary partitions coarser than $\mathcal{P}$. Axiom 1 tells us knowing $\mu(A_1)$, ..., and $\mu(A_m)$ is more than enough to compute expected learning costs on binary partitions coarser than $\mathcal{P}$, and thus $C(\mathcal{P}, \mu)$ is fully determined by $\mu(A_1)$, ..., and $\mu(A_m)$.

If we show that for any $i, j \in \{1, \ldots, m\}$ such that $i \neq j$, $C(\mathcal{P}, \mu) = C(\mathcal{P}, \tilde{\mu})$ if $\mu(A_k) = \tilde{\mu}(A_k)$ for $k \notin \{i, j\}$, $\mu(A_i) = \tilde{\mu}(A_j)$, and $\mu(A_j) = \tilde{\mu}(A_i)$, then the desired result holds, since a series of pairwise switches like this can be used to create any permutation desired. It is without loss to assume $i = 1$ and $j = 2$. Define $\tilde{\mathcal{P}} = \{A_1, A_2, (A_1 \cup A_2)^c\}$. Notice that $\tilde{\mathcal{P}}$ must be learning strategy invariant based on Lemma 7. Further, if we 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 we define $\hat{\mathcal{P}} = \{A_1 \cup A_2, A_3, \ldots, A_m\}$, which is also learning strategy invariant based on Lemma 7, then Lemma 8 tell 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 we define probability measure $\hat{\mu}$ so that $\hat{\mu}(A_1) = \hat{\mu}(A_2) = 0$, and for $i \in \{3, \ldots, m\}$ we have $\hat{\mu}(A_i) = \mu(A_i)/(1 - \mu(A_1 \cup A_2))$. Now, let $\mathcal{P}_1^b = \{A_1, A_1^c\}$, $\mathcal{P}_2^b = \{A_2, A_2^c\}$, and $\mathcal{P}_3^b = \{A_1 \cup A_2, (A_1 \cup A_2)^c\}$. Notice $\mathcal{P}_1^b$, $\mathcal{P}_2^b$ and $\mathcal{P}_3^b$, are all coarser than $\tilde{\mathcal{P}}$. Then, since $\tilde{\mathcal{P}}$ is learning strategy invariant:

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

and,

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

Notice that Axiom 1 tells us $C(\mathcal{P}_3^b, \mu) = C(\mathcal{P}_3^b, \tilde{\mu})$. So, all that remains to show is that if the probability measure $\tilde{\nu}$ is a permutation of the probability measure ν on

$\mathcal{P}_1^b$, then $C(\mathcal{P}_1^b, \nu) = C(\mathcal{P}_1^b, \tilde{\nu})$. Fix arbitrary $\nu(A_1) = x \in [0, 1]$. Now consider the probability measures q_1, q_2, q_3, such that:

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

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

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

Notice that q_3 is a permutation of q_1 on $\mathcal{P}_1^b$. So then, using Axiom 1, the definition of learning strategy invariance, and Lemma 8, all repeatedly:

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

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

and we are done.■

Proof of Lemma 2. 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 $\mathcal{P}_i = \{A_1, \ldots, A_m\}$ is learning strategy invariant with $m \geq 3$, and $\tilde{\mathcal{P}}_i$ is another learning strategy invariant partition such that $\tilde{\mathcal{P}}_i \neq \mathcal{P}_i$, and $\tilde{\mathcal{P}}_i$ is coarser than $\mathcal{P}_i$. Lemma 1 tells us that $C(\mathcal{P}_i, \mu)$ and $C(\tilde{\mathcal{P}}_i, \mu)$ are fully determined by $\mu(\mathcal{P}_i)$ and $\mu(\tilde{\mathcal{P}}_i)$ respectively, and if the strictly positive entries of $\mu(\mathcal{P}_i)$ and $\mu(\tilde{\mathcal{P}}_i)$ are the same (up to a permutation), then Lemma 8 and the definition of learning strategy invariant partitions tell us $C(\mathcal{P}_i, \mu) = C(\tilde{\mathcal{P}}_i, \mu)$. This is true for any $\tilde{\mathcal{P}}_i$ that is coarser than $\mathcal{P}_i$ since we can pick μ so that uncertainty about which even 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 $\mathcal{P}_i$, $C(\tilde{\mathcal{P}}_i, \mu) = c_i(\mu(\tilde{\mathcal{P}}_i)) \equiv C(\mathcal{P}_i, \mu)$.

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 7 tells us $\tilde{\mathcal{P}}_i$ 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:

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

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

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

and we are done.■

We 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 we prove Lemma 3, we pause to show another useful result.

Lemma 9. Given a binary partition $\mathcal{P}^b$, if we 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 of Lemma 9. 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. We proceed with an inductive argument, beginning by showing $c_{\mathcal{P}^b}(p, 1-p)$ is constant with respect to permutations. 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:

$$c_{\mathcal{P}^b}(0, 1) + c_{\mathcal{P}^b}(q_1, q_2) = c_{\mathcal{P}^b}(q_1, q_2) + p_2 c_{\mathcal{P}^b}(1, 0),$$

so since $c_{\mathcal{P}^b}(0, 1) = 0$, we know $c_{\mathcal{P}^b}(1, 0) = 0 = c_{\mathcal{P}^b}(0, 1)$, and combined with our previous two equalities above we 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 $c_{\mathcal{P}^b}(1, 0) = 0$, if we show $c_{\mathcal{P}^b}$ is constant with respect to permutations of vectors of arbitrary length (greater or equal to two), then 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 we 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 $c_{\mathcal{P}^b}(1, 0) = c_{\mathcal{P}^b}(0, 1) = 0$, tells us $c_{\mathcal{P}^b}(p_1, p_2, p_3) = c_{\mathcal{P}^b}(p_1, p_3, p_2)$. Thus, if $c_{\mathcal{P}^b}(p_1, p_2, p_3) = c_{\mathcal{P}^b}(p_2, p_1, p_3)$, 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, as can be shown, since if $p_1 = 1$ or $p_2 = 1$ we know this is true, and otherwise with Axiom 2 we 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 we next show $c_{\mathcal{P}^b}$ is constant with respect to permutations on vectors of length $n+1$, and we are done. If $p_1 + p_2 = 1$ we are done. If not, notice that $c_{\mathcal{P}^b}(p_1, \ldots, p_{n+1}) = c_{\mathcal{P}^b}(p_1, 1-p_1) + (1-p_1)c_{\mathcal{P}^b}\Big(\frac{p_2}{1-p_1}, \ldots, \frac{p_{n+1}}{1-p_1}\Big)$, for $n \geq 2$ whenever $p_1 \neq 1$, so we 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 10. 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 9, 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 (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_js with each p_j used (once) in

the sum of one q_i. Let us 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 of Lemma 10. 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 9, and $(q_1, \ldots, q_m)$ and $(p_1, \ldots, p_n)$ are described as in the statement of Lemma 10 (including the re-naming of the p_js). All vectors discussed in this proof are assumed to sum to one and have at least length two, and we 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 9 we showed $c_{\mathcal{P}^b}$ is constant with respect to permutations of vectors of arbitrary length (greater or equal to two). Thus, all we 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 $q_m = p_m + \ldots + p_n > 0$. This is of course true. 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 $p_n > 0$):

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

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

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

$$+ \cdots + (1-p_1-\ldots-p_{n_1}) C\Big(\mathcal{P}^b, \frac{p_n}{1-p_1-\ldots-p_{n_1}}, \frac{1-p_1-\ldots-p_n}{1-p_1-\ldots-p_{m-1}}\Big)$$

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

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 9, 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 9 tells us $C(\mathcal{P}^b, p, 1-p) = C(\mathcal{P}^b, 1-p, p)$, for each $p \in [0, 1]$, and we thus only wish to show $c_{\mathcal{P}^b}(p, 1-p)$ is continuous for $p \in [0, 1]$. 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 $c_{\mathcal{P}^b}(p, 1-p)$ is discontinuous at $p = 0$ iff it is discontinuous at a $\tilde{p} \in (0, \frac{1}{2}]$, because Axiom 2 tells us that for small $\delta > 0$: $c_{\mathcal{P}^b}(\delta, \frac{1}{2} - \frac{\delta}{2}, \frac{1}{2} - \frac{\delta}{2}) = c_{\mathcal{P}^b}(\delta, 1-\delta) + (1-\delta)c_{\mathcal{P}^b}(1/2, 1/2) = c_{\mathcal{P}^b}(\frac{1}{2} - \frac{\delta}{2}, \frac{1}{2} + \frac{\delta}{2}) + (\frac{1}{2} + \frac{\delta}{2})c_{\mathcal{P}^b}(\frac{2\delta}{1+\delta}, \frac{1-\delta}{1+\delta})$.

Further, if $c_{\mathcal{P}^b}(p, 1-p)$ is discontinuous at $p = \frac{1}{2}$ then it is discontinuous at $p \in \{\frac{1}{4}, \frac{1}{3}\}$ because Axiom 2 tells us that for small $\delta > 0$: $c_{\mathcal{P}^b}(\frac{1}{2} - \delta, \frac{1}{3} + \frac{2\delta}{3}, \frac{1}{6} + \frac{\delta}{3}) = c_{\mathcal{P}^b}(\frac{1}{2} - \delta, \frac{1}{2} + \delta) + (\frac{1}{2} + \delta)c_{\mathcal{P}^b}(\frac{1}{3}, \frac{2}{3}) = c_{\mathcal{P}^b}(\frac{1}{3} + \frac{2\delta}{3}, \frac{2}{3} - \frac{2\delta}{3}) + (\frac{2}{3} - \frac{2\delta}{3})c_{\mathcal{P}^b}((\frac{1}{6} + \frac{\delta}{3})/(\frac{2}{3} - \frac{2\delta}{3}), (\frac{1}{2} - \delta)/(\frac{2}{3} - \frac{2\delta}{3}))$. It is thus without loss to assume $c_{\mathcal{P}^b}(p, 1-p)$ is discontinuous at $p = 0$, and $p_d \in (0, \frac{1}{2})$. Axiom 3 then requires there is $p_c \in (0, 1/2]$ such that $c_{\mathcal{P}^b}(p, 1-p)$ is continuous at $p = p_c$. But this is not possible, and we can reach a contradiction. Pick (p_1, p_2, p_2, p_4) such that they sum to one and:

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

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

Then notice Lemma 10 tells us:

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

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

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

We begin by pointing out that $c_{\mathcal{P}^b}$ is discontinuous from both sides at p_d, since we could increase p_1 and p_3 by a small $\delta > 0$, and decrease p_2 and p_4 by the same δ. As δ is taken to zero, continuity at p_c implies the change in $c_{\mathcal{P}^b}(p_1 + p_2,\, p_3 + p_4) + (p_1 + p_2)c_{\mathcal{P}^b}\Big(\frac{p_1}{p_1+p_2},\, \frac{p_2}{p_1+p_2}\Big) + (p_3 + p_4)c_{\mathcal{P}^b}\Big(\frac{p_3}{p_3+p_4},\, \frac{p_4}{p_3+p_4}\Big)$ goes to zero, so discontinuities at either side of p_d must offset each other so the change in $c_{\mathcal{P}^b}(p_1 + p_4,\, p_2 + p_3) + (p_1 + p_4)c_{\mathcal{P}^b}\Big(\frac{p_1}{p_1+p_4},\, \frac{p_4}{p_1+p_4}\Big) + (p_2 + p_3)c_{\mathcal{P}^b}\Big(\frac{p_2}{p_2+p_3},\, \frac{p_3}{p_2+p_3}\Big)$ goes to zero.

Next, we show there must be a $p_d \in (0,\, \frac{1}{2})$ where $c_{\mathcal{P}^b}(p,\, 1-p)$ drops at $p = p_d$ ($\exists p_d \in (0,\, \frac{1}{2})$, $\epsilon > 0$ such that $\forall \delta > 0$ there is p such that $p \in (p_d, p_d + \delta)$ and $c_{\mathcal{P}^b}(p_d,\, 1-p_d) - c_{\mathcal{P}^b}(p,\, 1-p) \geq \epsilon$). If not, since $c_{\mathcal{P}^b}$ is discontinuous from both sides at p_d, there is an $\epsilon > 0$ such that $\forall \delta > 0$, $\exists p \in (p_d, p_d + \delta)$ such that $c_{\mathcal{P}^b}(p,\, 1-p) - c_{\mathcal{P}^b}(p_d,\, 1-p_d) \geq \epsilon$. Now, fix particular ϵ, $\delta > 0$ such that if $p \in (p_d, p_d + \delta)$ then $c_{\mathcal{P}^b}(p,\, 1-p) - c_{\mathcal{P}^b}(p_d,\, 1-p_d) \in [\frac{4\epsilon}{5},\, \frac{5\epsilon}{4}]$. Given this ϵ, pick $\tilde{\delta} > 0$ such that $\tilde{\delta} < \delta$, and if $|p - p_c| \leq \frac{\tilde{\delta}}{p_d}$), then $|c_{\mathcal{P}^b}(p,\, 1-p) - c_{\mathcal{P}^b}(p_c,\, 1-p_c)| \leq \frac{\epsilon}{1000}$, and $\tilde{\delta}c_{\mathcal{P}^b}(p,\, 1-p) \leq \frac{\epsilon}{1000}$. Now, pick $\hat{\delta} > 0$ such that $\frac{\hat{\delta}}{p_c} \leq \tilde{\delta}$. If we think about increasing p_1 by $\hat{\delta}$, and decreasing p_4 by $\hat{\delta}$, keeping p_2 and p_3 constant (where $(p_1,\, p_2,\, p_2,\, p_4)$ are picked as in the previous paragraph), then it is evident a contradiction has been created since, using Lemma 10 as in the previous paragraph:

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

$$+p_d c_{\mathcal{P}^b}(p_c,\, 1-p_c) - (p_d + \hat{\delta})c_{\mathcal{P}^b}\Big(\frac{p_1+\hat{\delta}}{p_1+\hat{\delta}+p_2},\, \frac{p_2}{p_1+\hat{\delta}+p_2}\Big)$$

$$+(1-p_d)c_{\mathcal{P}^b}(1-p_c,\, p_c) - (1-p_d-\hat{\delta})c_{\mathcal{P}^b}\Big(\frac{p_3}{p_3+p_4-\hat{\delta}},\, \frac{p_4-\hat{\delta}}{p_3+p_4-\hat{\delta}}\Big)$$

$$= p_c\Big(c_{\mathcal{P}^b}(p_d,\, 1-p_d) - c_{\mathcal{P}^b}\Big(p_d + \frac{\hat{\delta}}{p_c},\, 1-p_d-\frac{\hat{\delta}}{p_c}\Big)\Big),$$

so,

$$c_{\mathcal{P}^b}(p_d,\, 1-p_d) - c_{\mathcal{P}^b}(p_d+\hat{\delta},\, 1-p_d-\hat{\delta}) - \frac{\epsilon}{250} \leq p_c\Big(c_{\mathcal{P}^b}(p_d,\, 1-p_d) - c_{\mathcal{P}^b}\Big(p_d+\frac{\hat{\delta}}{p_c},\, 1-p_d-\frac{\hat{\delta}}{p_c}\Big)\Big)$$

$$\leq c_{\mathcal{P}^b}(p_d,\, 1-p_d) - c_{\mathcal{P}^b}(p_d+\hat{\delta},\, 1-p_d-\hat{\delta}) + \frac{\epsilon}{250}$$

$$\implies \frac{4\epsilon}{5} - \frac{\epsilon}{250} \leq p_c\Big(c_{\mathcal{P}^b}\Big(p_d+\frac{\hat{\delta}}{p_c},\, 1-p_d-\frac{\hat{\delta}}{p_c}\Big) - c_{\mathcal{P}^b}(p_d,\, 1-p_d)\Big) \leq \frac{5\epsilon}{4} + \frac{\epsilon}{250},$$

but this is not possible since $p_c \leq \frac{1}{2}$, so then we would require:

$$c_{\mathcal{P}^b}\Big(p_d+\frac{\hat{\delta}}{p_c},\, 1-p_d-\frac{\hat{\delta}}{p_c}\Big) - c_{\mathcal{P}^b}(p_d,\, 1-p_d) \geq \frac{398\epsilon}{250} > \frac{5\epsilon}{4}.$$

Thus, $\exists p_d \in (0,\, \frac{1}{2})$, and $\epsilon > 0$, such that $\forall \delta > 0$, there is p such that $p \in (p_d, p_d+\delta)$ and $c_{\mathcal{P}^b}(p_d,\, 1-p_d) - c_{\mathcal{P}^b}(p,\, 1-p) \geq \epsilon$. Fix such a p_d and ϵ, letting $c_{\mathcal{P}^b}(p_d,\, 1-p_d) = k > 0$. We next show that the x, such that $\forall \delta > 0$ there is $p \in (p_d, p_d+\delta)$ such that $c_{\mathcal{P}^b}(p_d,\, 1-p_d) - c_{\mathcal{P}^b}(p,\, 1-p) \geq x$, is unbounded, which causes a contradiction, since it is then true for some $x > k$, and $c_{\mathcal{P}^b}(p,\, 1-p)$ cannot be negative. If this were not true, then there would be $x > 0$ such that $\forall \delta > 0$ there is $p \in (p_d, p_d+\delta)$ such that $c_{\mathcal{P}^b}(p_d,\, 1-p_d) - c_{\mathcal{P}^b}(p,\, 1-p) \geq x$, but $\exists \delta > 0$ such that $\nexists p \in (p_d, p_d+\delta)$ such that $c_{\mathcal{P}^b}(p_d,\, 1-p_d) - c_{\mathcal{P}^b}(p,\, 1-p) \geq \frac{3x}{2}$. Pick any small $\delta > 0$ such that if $|p-p_c| \leq \frac{\delta}{p_d}$, then $|c_{\mathcal{P}^b}(p,\, 1-p) - c_{\mathcal{P}^b}(p_c,\, 1-p_c)| \leq \frac{x}{1000}$, and $\delta c_{\mathcal{P}^b}(p,\, 1-p) \leq \frac{x}{1000}$. Now, pick $\hat{\delta} > 0$ such that $\frac{\hat{\delta}}{p_c} < \delta$, and $c_{\mathcal{P}^b}(p_d,\, 1-p_d) - c_{\mathcal{P}^b}(p_d+\hat{\delta},\, 1-p_d-\hat{\delta}) \geq x$. Now consider increasing p_1 by $\hat{\delta}$, and decreasing p_4 by $\hat{\delta}$, keeping p_2 and p_3 constant (where $(p_1,\, p_2,\, p_2,\, p_4)$ are picked as above in this proof). Then, using Lemma 10 as in the previous paragraphs:

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

$$+p_d c_{\mathcal{P}^b}(p_c,\, 1-p_c) - \Big(p_d+\hat{\delta}\Big)c_{\mathcal{P}^b}\Big(\frac{p_1+\hat{\delta}}{p_1+\hat{\delta}+p_2},\, \frac{p_2}{p_1+\hat{\delta}+p_2}\Big)$$

$$+(1-p_d)c_{\mathcal{P}^b}(1-p_c,\, p_c) - (1-p_d-\frac{\hat{\delta}}{p_c})c_{\mathcal{P}^b}\Big(\frac{p_3}{p_3+p_4-\hat{\delta}},\, \frac{p_4-\hat{\delta}}{p_3+p_4-\hat{\delta}}\Big)$$

$$= p_c\Big(c_{\mathcal{P}^b}(p_d,\, 1-p_d) - c_{\mathcal{P}^b}\Big(p_d+\frac{\hat{\delta}}{p_c},\, 1-p_d-\frac{\hat{\delta}}{p_c}\Big)\Big),$$

$$\implies p_c\Big(c_{\mathcal{P}^b}(p_d,\, 1-p_d) - c_{\mathcal{P}^b}\Big(p_d+\frac{\hat{\delta}}{p_c},\, 1-p_d-\frac{\hat{\delta}}{p_c}\Big)\Big) \geq c_{\mathcal{P}^b}(p_d,\, 1-p_d) - c_{\mathcal{P}^b}\Big(p_d+\hat{\delta},\, 1-p_d-\hat{\delta}\Big) - \frac{x}{250},$$

but then, since $p_c \leq \frac{1}{2}$:

$$c_{\mathcal{P}^b}(p_d,\, 1-p_d) - c_{\mathcal{P}^b}\Big(p_d+\frac{\hat{\delta}}{p_c},\, 1-p_d-\frac{\hat{\delta}}{p_c}\Big)\Big) \geq 2x - \frac{x}{125} > \frac{3x}{2}.$$

Since this strategy can be employed for any arbitrarily small $\delta > 0$, our contradiction is achieved, and discontinuity at a point would imply $\exists p \in [0,\, 1]$ such that $c_{\mathcal{P}^b}(p,\, 1-p) < 0$, and thus $c_{\mathcal{P}^b}(p,\, 1-p)$ must be continuous $\forall p \in [0,\, 1]$.■

Proof of Lemma 4.

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 9, and suppose C satisfies Axiom 1, Axiom 2, and Axiom 3. Remember $C(\mathcal{P}^b,\, \mu) = c_{\mathcal{P}^b}(\mu(A_1),\, \mu(A_2)) = c_{\mathcal{P}^b}(\mu(A_2),\, \mu(A_1))$ for all probability measures μ, so we only need to show $c_{\mathcal{P}^b}(p,\, 1-p)$ is non-decreasing for small increases to $p \in (0,\, 1/2)$ since Lemma 3 shows $c_{\mathcal{P}^b}(p,\, 1-p)$ is continuous, and Lemma 9 $c_{\mathcal{P}^b}(0,\, 1) = 0$, and so before each p where $c_{\mathcal{P}^b}(p,\, 1-p)$ is decreasing in p, there must be a smaller p where $c_{\mathcal{P}^b}(p,\, 1-p)$ is increasing in p. We proceed by assuming there is a $p_d \in (0,\, 1/2)$ such that $c_{\mathcal{P}^b}(p_d,\, 1-p_d)$ is decreasing for small increases in p_d, and create a contradiction. Notice that then 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 we can

pick (p_1, p_2, p_2, p_4) such that:

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

and then notice Lemma 10 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 $p = p_1/(p_1 + p_4) < p_d$. This all means $c_{\mathcal{P}^b}(p, 1-p)$ has dense local maxima and minima for p close to zero.

Next we 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 small increases in p_1. Given such a small $\epsilon > 0$, 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 $p_1 \leq 1/4$, we know $p_1 < p_3 < p_2$. Now consider increasing p_1 by ϵ and decreasing p_2 by ϵ. 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 10 tells us:

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

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

Thus:

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

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

so,

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

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

Thus, at

$$\frac{p_1}{p_1+p_2} > p_1,$$

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

$$\frac{p_1}{p_1+p_2} \text{ and } \frac{p_1+\epsilon}{p_1+p_2},$$

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 is p_1 close to $1/4$, if we pick p_2 and p_3 as above, keeping our small ϵ in mind, we 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) $p_s(\epsilon) \in (1/4,\, 1/2 - \epsilon)$, given what we just showed.

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

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

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

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

and finally let $p_1 = 1 - p_2 - p_3$, noticing $p_1 > 1/4$ so that:

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

Lemma 10 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 we increase p_3 by ϵ, and decrease p_2 by ϵ, holding p_1 constant:

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

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

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

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

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

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

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

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 we 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 $\delta=\dfrac{\epsilon}{p_1+p_3+\epsilon}\dfrac{p_1}{p_1+p_3}\in\Big[\epsilon,\,\dfrac{\epsilon}{p_1+p_3+\epsilon}\Big]$:

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

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

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

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

Proof of Lemma 5. Given learning strategy invariant partition $\mathcal{P} = \{A_1, \ldots, A_m\}$, if $m \geq 3$, pick any binary partition $\mathcal{P}^b$ coarser than $\mathcal{P}$, and if $m = 2$ take $\mathcal{P}^b = \mathcal{P}$, and with this $\mathcal{P}^b$, 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 9, and suppose C satisfies Axiom 1, Axiom 2, and Axiom 3, and Axiom 4. By definition of $c_{\mathcal{P}^b}$, $C(\mathcal{P}, \mu) = c_{\mathcal{P}^b}(\mu(A_1), \ldots, \mu(A_m))$. Define h so for $n \in \mathbb{N}$, $h(n) \equiv c_{\mathcal{P}^b}(1/n, \ldots, 1/n, 0)$. Notice that Axiom 4 implies$h(2) > h(1) = 0$, and in general $h(n) > 0$ if $n > 1$. It is also easy to show $h(n+1) \geq h(n)$ for all $n \geq 2$ using Lemma 10 and Lemma 4:

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

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

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

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

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

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

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

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:

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

The work we did above then tells then tells us:

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

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

All of this tells us:

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

which can be shown to be true $\forall \epsilon > 0$, and thus $h(n) = \lambda \log(n)$, where λ must be a positive constant to satisfy Axiom 4.

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 $k \in \{1, \ldots, m\}$ we have:

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

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

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

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

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

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

where $\mathcal{H}$ is defined as in equation (1). If any of the p_i are irrational, then the density of the rationals and Lemma 3 can be used to get the same result. Thus:

$$C(\mathcal{P}, \mu) = c_{\mathcal{P}^b}(\mu(A_1), \ldots, \mu(A_m)) = \lambda \mathcal{H}(\mathcal{P}, \mu). \blacksquare$$

Mutual Information

Consider two partitions $\mathcal{P}_1$ and $\mathcal{P}_2$. Given some probability measure μ, define the **mutual information** between $\mathcal{P}_1$ 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:

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

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

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

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 we think of an agent that is trying to acquire information in an efficient fashion, we should always envision them acquiring mutual information from the cheapest source, 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 11.

Lemma 11. If C satisfies our five axioms, and $S^b = \{\mathcal{P}_1^b, \ldots, \mathcal{P}_i^b, \mathcal{P}_{i+1}^b, \ldots, \mathcal{P}_m^b\}$ and $\tilde{S}^b = \{\mathcal{P}_1^b, \ldots, \mathcal{P}_{i+1}^b, \mathcal{P}_i^b, \ldots, \mathcal{P}_m^b\}$ are two binary learning strategies such that $\mathcal{P}_i^b$ and $\mathcal{P}_{i+1}^b$'s associated multipliers are ordered $\lambda_i \geq \lambda_{i+1}$, then for all probability measures μ:

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

Proof. For all realizations of $\cap_{j=1}^{i-1}\mathcal{P}_j^b(\omega)$:

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

$$= \lambda_i\mathcal{H}(\mathcal{P}_i^b, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}_j^b(\omega)))+\lambda_{i+1}\Big(\mathcal{H}(\mathcal{P}_{i+1}^b, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}_j^b(\omega)))-I(\mathcal{P}_i^b, \mathcal{P}_{i+1}^b, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}_j^b(\omega))\Big)$$

$$\geq \lambda_i\Big(\mathcal{H}(\mathcal{P}_i^b, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}_j^b(\omega)))-I(\mathcal{P}_i^b, \mathcal{P}_{i+1}^b, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}_j^b(\omega))\Big)+\lambda_{i+1}\mathcal{H}(\mathcal{P}_{i+1}^b, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}_j^b(\omega)))$$

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

$$= C((\mathcal{P}_{i+1}^b, \mathcal{P}_i^b), \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}_j^b(\omega))).$$

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

Proof of Theorem 1. Given some probability measure μ, suppose S^b is a binary

learning strategy such that $\sigma(S^b) = \mathcal{F}$, and

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

We know such binary learning strategy exists whenever C satisfies Axiom 5. We may assume that if $\mathcal{P}_i^b$ and $\mathcal{P}_{i+1}^b$ 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 11.

If for any $j \in \{1, \ldots, M\}$, multiplier λ_j's associated binary partitions $\mathcal{P}_i^b, \ldots, \mathcal{P}_{i+k}^b$ in S^b are such that $\sigma(\mathcal{P}_i^b, \ldots, \mathcal{P}_{i+k}^b) \neq \sigma(\mathcal{P}_{\lambda_j}^b)$, then there are binary partitions $\mathcal{P}_{m+1}^b, \ldots, \mathcal{P}_{m+l}^b$ with associated multiplier λ_j, such that $\sigma(\mathcal{P}_i^b, \ldots, \mathcal{P}_{i+k}^b, \mathcal{P}_{m+1}, \ldots, \mathcal{P}_{m+l}^b) = \sigma(\mathcal{P}_{\lambda_j}^b)$. $\mathcal{P}_{m+1}^b, \ldots, \mathcal{P}_{m+l}^b$ 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 11 then implies that if we reorder $\tilde{S}^b$ so that the new learning strategy $\hat{S}$'s binary partitions are ordered by their multipliers, then:

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

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

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

$$\mathbb{E}[C((\mathcal{P}_i^b, \ldots, \mathcal{P}_{i+k}^b), \mu(\cdot | \cap_{t=1}^{i-1} \mathcal{P}_t^b(\omega)))] = \mathbb{E}\Big[\sum_{l=i}^{i+k} \lambda_j \mathcal{H}(\mathcal{P}_l^b, \mu(\cdot | \cap_{t=1}^{l-1} \mathcal{P}_t^b(\omega)))\Big]$$

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

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

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

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

Appendix 2

Proof of Lemma 6. In Lemma 6, we show that we can rewrite the agent's problem in terms of selecting the choice probabilities described in equations (4), (5), and (6). To do this, we first establish several other lemmas. In Lemma 12, we show that: $\min_{S \in S(\Omega)} C(S,\, \mu)$ is a strictly concave function of μ. This is a commonly known property of Shannon Entropy, but needs to be established for in our context. This implies that $\mathbf{C}$ is strictly convex. We then show, in Lemma 13, that, given the convexity of $\mathbf{C}$, any selected action is associated with a particular posterior probability. This is desirable because it allows us to reduce the strategies considered to recommendation strategies. That is, we are able to focus on signals that are simply a recommendation of an option. In Lemma 14, we show that we may rewrite the cost function in terms of the choice probabilities in equations (4), (5), and (6).

Lemma 12. If C satisfies all five axioms then $\min_{S \in S(\Omega)} C(S,\, \mu)$ is a strictly concave function of μ. Namely, if there are probability measures μ_a, and μ_b, such that $\mu = \alpha\mu_a + (1-\alpha)\mu_b$ for some $\alpha \in (0,1)$, and $\mu_a \neq \mu_b$, then:

$$\min_{S \in S(\Omega)} C(S,\, \mu) > \alpha\Big(\min_{S \in S(\Omega)} C(S,\, \mu_a)\Big) + (1-\alpha)\Big(\min_{S \in S(\Omega)} C(S,\, \mu_b)\Big).$$

Proof. For each such μ_a, μ_b, $\alpha \in (0,1)$, and μ, the strict concavity of Shannon

Entropy (Matějka & McKay, 2015; Caplin et al., 2017) implies:

$$\mathcal{H}(\mathcal{P}_{\lambda_1},\,\mu) \geq \alpha\mathcal{H}(\mathcal{P}_{\lambda_1},\,\mu_a) + (1-\alpha)\mathcal{H}(\mathcal{P}_{\lambda_1},\,\mu_b).$$

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

$$\mathcal{H}(X,\,\nu) = \sum_X \nu(x)\log(\nu(x)),\;\; \mathcal{H}(\mathcal{P}_{\lambda_i},\,X,\,\nu) = \sum_{A\in\mathcal{P}_{\lambda_i}}\sum_X \nu(A,x)\log(\nu(A,x)).$$

Then, for each such μ_a, μ_b, $\alpha \in (0,1)$, μ, and $i \in \{2,\,\ldots,\,M\}$, the properties of Shannon Entropy tell us:

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

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

$$\begin{aligned}
\implies \mathbb{E}\Big[\mathcal{H}(\mathcal{P}_{\lambda_i},\,\mu(\cdot|\cap_{j=1}^{i-1}\,\mathcal{P}_{\lambda_j}(\omega)))\Big] &= \mathbb{E}\Big[\mathcal{H}(\mathcal{P}_{\lambda_i},\,\mu(\cdot|\cap_{j=1}^{i-1}\,\mathcal{P}_{\lambda_j}(\omega),\,X))\Big] \\
&+\mathbb{E}\Big[\mathcal{H}(X,\,\mu(\cdot|\cap_{j=1}^{i-1}\,\mathcal{P}_{\lambda_j}(\omega)))\Big] - \mathbb{E}\Big[\mathcal{H}(X,\,\mu(\cdot|\cap_{j=1}^{i}\,\mathcal{P}_{\lambda_j}(\omega)))\Big] \\
&\geq \mathbb{E}\Big[\mathcal{H}(\mathcal{P}_{\lambda_i},\,\mu(\cdot|\cap_{j=1}^{i-1}\,\mathcal{P}_{\lambda_j}(\omega),\,X))\Big] \\
= \mathbb{E}\Big[\alpha\mathcal{H}(\mathcal{P}_{\lambda_i},\,\mu_a(\cdot|\cap_{j=1}^{i-1}\,\mathcal{P}_{\lambda_j}(\omega))) &+ (1-\alpha)\mathcal{H}(\mathcal{P}_{\lambda_i},\,\mu_b(\cdot|\cap_{j=1}^{i-1}\,\mathcal{P}_{\lambda_j}(\omega)))\Big].
\end{aligned}$$

The above inequality is strict for at least one $i \in \{2,\,\ldots,\,M\}$ if the inequality from the previous paragraph is not strict, since $\mu_a \neq \mu_b$ and $\mathcal{H}$ is strictly concave. The desired result thus follows from Theorem 1.■

Lemma 13. If action $n \in \mathcal{N}$ is selected with positive probability, $\Pr(n) > 0$, as the outcome of information strategy F with is a solution to (2) subject to (3), then there

exists a posterior belief B_n such that $F(\omega|s) = B_n$ with probability one whenever n is selected.

Proof. It is impossible that there are two distinct sets of signals S_n^1 and S_n^2 which are observed with strictly positive probability, both of which lead to the selection of n, and induce different posteriors $F(\omega|s_1) \neq F(\omega|s_2)$ for $s_1 \in S_n^1$ and $s_2 \in S_n^2$. $\min_{S \in S(\Omega)} C(S, \mu)$ is strictly concave in μ, as shown in Lemma 12, so the agent could thus do better by replacing their original information strategy F with a new information strategy $\tilde{F}$ which is identical to F except the signals in S_n^1 and S_n^2 are replaced by s_0: $\forall \omega \in \Omega$ let $\tilde{F}(s_0|\omega) = \int_{s \in S_n^1} F(s|\omega) + \int_{s \in S_n^2} F(s|\omega)$. This is true because payoffs are linear, and the law of iterated expectations implies the agent still picks n after s_0 is realized since $\forall\, \nu \in \mathcal{N}$:

$$\mathbb{E}_{\tilde{F}}[\mathbf{v}_n(\omega)|s_0] = \frac{\sum_{\omega \in \Omega} \int_{s \in S_n^1} F(s|\omega)\mu(\omega)}{\sum_{\omega \in \Omega} \left(\int_{s \in S_n^1} F(s|\omega)\mu(\omega) + \int_{s \in S_n^2} F(s|\omega)\mu(\omega) \right)} \mathbb{E}_F[\mathbf{v}_n(\omega)|s \in S_n^1]$$

$$+ \frac{\sum_{\omega \in \Omega} \int_{s \in S_n^2} F(s|\omega)\mu(\omega)}{\sum_{\omega \in \Omega} \left(\int_{s \in S_n^2} F(s|\omega)\mu(\omega) + \int_{s \in S_n^2} F(s|\omega)\mu(\omega) \right)} \mathbb{E}_F[\mathbf{v}_n(\omega)|s \in S_n^2]$$

$$\geq \frac{\sum_{\omega \in \Omega} \int_{s \in S_n^1} F(s|\omega)\mu(\omega)}{\sum_{\omega \in \Omega} \left(\int_{s \in S_n^1} F(s|\omega)\mu(\omega) + \int_{s \in S_n^2} F(s|\omega)\mu(\omega) \right)} \mathbb{E}_F[\mathbf{v}_\nu(\omega)|s \in S_n^1]$$

$$+ \frac{\sum_{\omega \in \Omega} \int_{s \in S_n^2} F(s|\omega)\mu(\omega)}{\sum_{\omega \in \Omega} \left(\int_{s \in S_n^2} F(s|\omega)\mu(\omega) + \int_{s \in S_n^2} F(s|\omega)\mu(\omega) \right)} \mathbb{E}_F[\mathbf{v}_\nu(\omega)|s \in S_2] = \mathbb{E}_{\tilde{F}}[\mathbf{v}_\nu(\omega)|s_0]. \blacksquare$$

.

Lemma 14. The cost of information for a given strategy in equation (2) can be

written:

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

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

$$-(\lambda_3-\lambda_2)\Pr(n|\mathcal{P}_{\lambda_1}(\omega)\cap\mathcal{P}_{\lambda_2}(\omega))\log(\Pr(n|\mathcal{P}_{\lambda_1}(\omega)\cap\mathcal{P}_{\lambda_2}(\omega)))$$

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

Proof. Let $\mathcal{P}_s = (S_1,\,\ldots,\,S_n)$ denote a partition of the space of signals the agent may receive. We showed in Lemma 13 that for each S_i if s in S_i then with probability one s results in a particular posterior. We then have:

$$\mathbf{C}(F(s,\,\omega),\,\mu) = \mathbb{E}[\min_{S\in S(\Omega)} C(S,\,\mu) - \min_{S\in S(\Omega)} C(S,\,\mu(\cdot|s))]$$

$$= \mathbb{E}\Big[\lambda_1\Big(\mathcal{H}(\mathcal{P}_{\lambda_1},\,\mu) - \mathcal{H}(\mathcal{P}_{\lambda_1},\,\mu(\cdot|s))\Big) \tag{11}$$

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

$$= \mathbb{E}\Big[\lambda_1\Big(\mathcal{H}(\mathcal{P}_s,\,F(s)) - \mathcal{H}(\mathcal{P}_s,\,F(s|\mathcal{P}_{\lambda_1}(\omega)))\Big) \tag{12}$$

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

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

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

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

$$-(\lambda_3-\lambda_2)\Pr(n|\mathcal{P}_{\lambda_1}(\omega)\cap\mathcal{P}_{\lambda_2}(\omega))\log(\Pr(n|\mathcal{P}_{\lambda_1}(\omega)\cap\mathcal{P}_{\lambda_2}(\omega)))$$

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

The equality of (11) and (12) follows from the symmetry of mutual information, defined in Appendix 1. ■

We now resume our proof of Lemma 6. First notice that Lemma 14 establishes $\mathbf{C}(\mathbb{P},\,\mu)$. For each $n\in\mathcal{N}$, let s_n denote a signal in S_n which results in the posterior generated by signals in S_n with probability one (in Lemma 13 we showed we can do this). Then notice:

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

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

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

Where the last step follows from the fact that $\Pr(X|Y)\Pr(Y)=\Pr(Y|X)\Pr(X)$. We now proceed with two proofs by contradiction. First, assume that $(F,\,a)$ is a solution to (2) subject to (3), which achieves expected utility U_1, and let $\mathbb{P}$ be the choice probabilities induced by it. Assume that $\mathbb{P}$ is not a solution to (7) subject to (8) and (9), and thus there is a $\tilde{\mathbb{P}}$ which satisfies (8) and (9) and achieves expected utility $U_2>U_1$. However, a strategy pairing $(\tilde{F},\,\tilde{a})$ can be created that generates $\tilde{\mathbb{P}}$. For instance, for each of N distinct signals s_n, let $\tilde{a}(\tilde{F}(\omega|s_n))\equiv n$, and let $\tilde{F}(s_n,\,\omega)=\tilde{\Pr}(n|\omega)\mu(\omega)\;\forall\,\omega$ so that (3) is satisfied. This is impossible though as then $(\tilde{F},\,\tilde{a})$ achieves $U_2>U_1$ and $(F,\,a)$ cannot have been optimal.

Similarly, assume that $\mathbb{P}$ is a solution to (7) subject to (8) and (9), which achieves expected utility U_3 and but is not induced by a solution to 2 subject to (3).

That is there is a $\tilde{F}$ which satisfies (3) and achieves $U_4 > U_3$. This means, however, that $\tilde{\Pr}(n|\omega) = \dfrac{\tilde{F}(s_n,\, \omega)}{\mu(\omega)}$ also achieves U_4, which is impossible as $\mathbb{P}$ was supposedly optimal and $\tilde{\mathbb{P}}$ satisfies (8) and (9). ■

Proof of Theorem 2. The Lagrangian for the above problem can be written:

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

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

Where $\xi_n(\omega) \geq 0$ are the Lagrange multipliers for (8), and $\gamma(\omega)$ are the multipliers for (9). If $\Pr(n) = 0$, then $\Pr(n|\omega) = 0\ \forall\omega \in \Omega$. If $\Pr(n|\cap_{i=1}^{m}\mathcal{P}_{\lambda_i}(\omega)) = 0$ for some $m \in \{1,\ldots,M-1\}$ and ω, then $\Pr(n|\omega) = 0$. If $\Pr(n) > 0$, and $\Pr(n|\cap_{i=1}^{m}\mathcal{P}_{\lambda_i}(\omega)) > 0,\ \forall m \in \{1,\ldots,M-1\}$, then the first order condition with respect to $\Pr(n|\omega)$ implies:

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

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

which then implies $\Pr(n|\omega) > 0$ and $\xi_n(\omega) = 0$, because if not, and $\Pr(n|\omega) = 0$, then since $\xi_n(\omega) \geq 0$, equality of the first order condition then necessitates $\gamma(\omega) = \infty$. This is impossible, however, since then $\forall\,\nu \in \mathcal{N}$ their respective first order conditions holding necessitates $\Pr(\nu|\omega) = 0$. This being true $\forall\,\nu \in \mathcal{N}$ of course then violates (9). Thus, the first order condition implies:

$$\Pr(n|\omega) = \Pr(n)^{\frac{\lambda_1}{\lambda_M}}\Pr(n|\mathcal{P}_{\lambda_1}(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ \ldots\ \Pr(n|\cap_{i=1}^{M-1}\mathcal{P}_{\lambda_i}(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}} e^{\frac{-\gamma(\omega)}{\lambda_M}} \tag{13}$$

Plugging (13) into (9), one can solve for $\gamma(\omega)$. Plugging $\gamma(\omega)$ back into (13) achieves the desired result.■

Proof of Corollary 1. Plug equation (10) into equation (7). ■

Proof of Theorem 3. A fixed effect interpretation of MSSE follows easily from the optimal choice probabilities described in Theorem 2:

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

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

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

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

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

The only kind of RU model consistent with this behavior is one where ϵ_n is distributed iid according to a Gumbel distribution (Train, 2009). ■

Appendix 3

The behavior described in Theorem 2 has many intuitive features. It is also a quite natural extension of the analogous result from Matějka and McKay (2015), which is described in equation (14). If we assume the agent has prior μ, and all

partitions are learning strategy invariant (the environment studied in Matějka and McKay (2015)) and have associated multiplier λ_2, then if the agent does optimal research in state $\omega \in \Omega$, they select option n from their set of options $\mathcal{N}$ with probability:

$$\Pr(n|\omega) = \frac{\Pr(n)e^{\frac{\mathbf{v}_n(\omega)}{\lambda_2}}}{\sum_{\nu\in\mathcal{N}} \Pr(\nu)e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_2}}}. \tag{14}$$

One major takeaway from the formula in (14) is that when Shannon Entropy is used to measure uncertainty the chance of the agent selecting an option n in a particular state of the world ω is fully determined by the unconditional chances of the options being selected, $\Pr(n)$, and the realized values of the options in that state of the world. Beyond this takeaway, the formula in (14) also has many intuitive features. If λ_2 grows, which represents an increase in the difficulty of learning, the value of each option in the realized state becomes less significant for the determination of the selected option, and the significance of the agent's prior increases. Similarly, if λ_2 shrinks, the agent's prior becomes less significant, and the realized values of the options becomes more significant. If λ_2 approaches infinity, the realized values become insignificant, and the behavior of the agent approaches the behavior of the agent in the case where learning is impossible: they choose their option based on their prior. If λ_2 approaches zero the unconditional priors become insignificant, and the behavior of the agent approaches the behavior of the agent in the case where learning is costless: they choose the option with the highest realized value.

If we instead assume that the agent may also learn through a partition with a lower multiplier λ_1, that can convey information about the realization $\mathcal{P}_{\lambda_1}(\omega)$ of a partition $\mathcal{P}_{\lambda_1}$ of Ω, then if $\mathcal{P}_{\lambda_1} \neq \Omega$, and the agent does optimal research in state $\omega \in \Omega$, they select option n from their set of options $\mathcal{N}$ with probability:

$$\Pr(n|\omega) = \frac{\Pr(n)^{\frac{\lambda_1}{\lambda_2}} \Pr(n|\mathcal{P}_{\lambda_1}(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_2}} e^{\frac{\mathbf{v}_n(\omega)}{\lambda_2}}}{\sum_{\nu\in\mathcal{N}} \Pr(\nu)^{\frac{\lambda_1}{\lambda_2}} \Pr(\nu|\mathcal{P}_{\lambda_1}(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_2}} e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_2}}}. \tag{15}$$

With MSSE, as the formula in (15) indicates, the chance of the agent selecting an option n in a particular state of the world ω depends not only on the unconditional chances of the options being selected and the realized values of the options, but also on the values that the options take in similar states of the world, states that result in the same realization of $\mathcal{P}_{\lambda_1}$. When option n is in general desirable in $\mathcal{P}_{\lambda_1}(\omega)$ relative to the other options, then $\Pr(n|\mathcal{P}_{\lambda_1}(\omega))$ is larger, and there may be a high chance of n being selected, even if $\Pr(n)$ is not that large, and $\mathbf{v}_n(\omega)$ is not that high.

The formula in (15) also has many intuitive features. It maintains the intuitive comparative statistics for λ_2 that the formula in (14) had, and also features intuitive properties for $\Pr(n|\mathcal{P}_{\lambda_1}(\omega))$ and λ_1. If observing $\mathcal{P}_{\lambda_1}(\omega)$ is completely uninformative about the value of the options, then it is optimal for the agent to select $\Pr(n|\mathcal{P}_{\lambda_1}(\omega)) = \Pr(n)$ since $\min_{S\in S(\Omega)} C(S, \mu)$ is strictly concave in μ. In this case $\Pr(n)^{\frac{\lambda_1}{\lambda_2}}\Pr(n|\mathcal{P}_{\lambda_1}(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_2}} = \Pr(n)$, and behavior is identical to that in (14). If the cheaper information source contains irrelevant information it is thus ignored, and behavior collapses back to the environment described in Matějka and McKay (2015), as we should desire. If λ_1 approaches λ_2 (the cheaper information source becomes close to as expensive as the more expensive information source) then behavior approaches that described in (14) since $\Pr(n)^{\frac{\lambda_1}{\lambda_2}}\Pr(n|\mathcal{P}_{\lambda_1}(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_2}} \to \Pr(n)$. Thus, if an insignificantly cheaper information source is introduced behavior is changed in an insignificant fashion. Again, this seems like a desirable property. If λ_1 approaches zero then the role of the unconditional priors dissipates, and exponent on $\Pr(n|\mathcal{P}(\omega))$ approaches one, meaning it replaces the unconditional prior from (14). This makes sense because if λ_1 goes to zero it means $\mathcal{P}_{\lambda_1}(\omega)$ can essentially be viewed for free, in which case behavior within each $\mathcal{P}_{\lambda_1}(\omega)$ should resemble that in the setting where there is only one information source with multiplier λ_2 and a prior of $\mu(\cdot|\mathcal{P}_{\lambda_1}(\omega))$.

We can continue adding as many new partitions with new associated multipliers as we desire and the description of behavior in Theorem 2 maintains the sorts of intuitive properties described in the paragraphs above. RI with MSSE is thus a very

natural extension of RI with Shannon Entropy.

References

Acharya, S., & Wee, S. L. (2019). Rational inattention in hiring decisions. *FRB of New York Staff Report*(878).

Ambuehl, S., Ockenfels, A., & Stewart, C. (2019). Attention and selection effects. *Rotman School of Management Working Paper*(3154197).

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

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

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

Dean, M., & Neligh, N. L. (2019). Experimental tests of rational inattention.

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.

Folland, G. B. (2013). *Real analysis: modern techniques and their applications*. John Wiley & Sons.

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. (2017). *Rational inattention and sequential information sampling* (Tech. Rep.). National Bureau of Economic Research.

Huettner, F., Boyacı, T., & Akçay, Y. (2019). Consumer choice under limited atten-

tion when alternatives have different information costs. *Operations Research*.

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.

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

Morris, S., & Yang, M. (2016). Coordination and continuous choice. *Working paper*.

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. (2019). The cost of information.

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

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

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

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

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

Walker-Jones, D. (2019, December). *Rational inattention and non-compensatory choice.* Retrieved from https://www.dwalkerjones.com

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