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

September 10, 2019

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 obtain information about the options that they face. Understanding how agents learn in such environments is crucial because partially informed choices have serious implications for revealed preference analysis, which makes welfare and counterfacutal analysis more difficult.

[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.

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 is not without limitations. It is natural to think that it should be more difficult to differentiate between outcomes that are more similar. Differentiating between two types of black tea, for instance, 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 demonstrated by Example 1 in Section 2.1, which can limit the effectiveness of Shannon Entropy in empirical settings.

This paper proposes five axioms that focus on the cost of asking simple questions, questions that can be represented by partitions of the outcome space. The axioms are used to create a new measure for the cost of learning that we call Multi-source Shannon Entropy (MSSE). MSSE features perceptual distance, maintains the desired tractability and flexibility of Shannon's classic measure when incorporated into a model of RI, and predicts behavioral patterns that have been identified as problematic for Shannon Entropy.

MSSE also identifies a previously undiscovered 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 informational biases in RU models generated by the presence of different perceptual distances in the same choice environment. Unlike the informational bias identified

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

with Shannon Entropy, 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 the labour 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).

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.

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 concerned 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. Further, Pomatto et al. (2019) imposes a type of constant marginal cost of

information that, interestingly enough, contradicts the form of constant marginal cost assumed in this 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 RU model that is analogous to 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

Economic agents frequently face choice environments that feature uncertainty about payoffs. In such environments, agents must decide how much information to acquire before choosing between available options. Models of Rational Inattention (RI) study cost functions for information, and how the trade-off agents face between the quality of information and the cost of learning affects their choice behavior.

In the 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 the measurable space $(\Omega, \mathcal{F})$, where Ω is the finite set of possible **states of the world** (the outcome

space), and $\mathcal{F}$ is the set of **events** (the power set of Ω). The agent is assumed to have a **prior** distribution, $\mu : \mathcal{F} \to [0, 1]$, over the potential states of the world.

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 **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** $F(s, \omega) \in \Delta(\mathbb{R}^N \times \Omega)$, which is a joint distribution between s, the N dimensional observed **signal**, and the states of the world.[3] The only restriction on the information strategy is that the marginal, $F(\omega) : \mathcal{F} \to \mathbb{R}_+$, must equal the prior μ. Alternatively, an agent can select a probability measure $F(s|\omega) : \mathbb{R}^N \to \mathbb{R}_+$ 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:

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

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

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

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 this cost function in Section

[3] The decision to allow s to be N dimensional is rather arbitrary. This is a much richer signal space than is required in practice. We show later that s only need be one dimensional in our setting.

4. The agent's problem can thus be written:

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

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

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:[4]

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

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 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 of the agent has a weakly positive cost.

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

Table 1: Example 1				
State:	ω_1	ω_2	ω_3	ω_4
Balls in State:	60 Blue & 40 Red	51 Blue & 49 Red	49 Blue & 51 Red	40 Blue & 60 Red
Probability of State:	1/4	1/4	1/4	1/4
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 ($\mathbf{v}_n(\omega_1) = \mathbf{v}_n(\omega_2)\ \forall n$), 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 (2018). 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 four 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 (2018), 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 outcome 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,[5] which limits the potential

[5]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.[6] 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

[6]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.[7]

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 in all states ω to be biased by $\log(\Pr(n) \cdot N)$, 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

[7] 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.[8] Finally, if the behavior of the agent is instead described by Shannon Entropy, then the choice behavior in the

[8]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. 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 the measurable space $(\Omega, \mathcal{F})$. Ω is the finite set of possible states of the world. $\mathcal{F}$ is the set of events.

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 Ω is a set of multiple disjoint events in $\mathcal{F}$ whose union is Ω. 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 there are 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 $C(\mathcal{P}, \mu)$ 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, and inefficient for a learnng strategy to begin with them. 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.

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.

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 can be separately evaluated in different contexts, either empirically or through introspection, to determine how appropriate it is.

Again, 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 perfectly observes the state of the world. We focus our axioms on attentive learning as we find this to be a more intuitive exercise than imposing axioms directly on inattentive behavior. 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 then be measured as the reduction in the cost of attentively learning, as subsequent sections discuss, as long as we can establish a cost of attentively learning the state of the world.[9] Our axioms are thus concerned with the cost of questions, and series of questions, whose answers are deterministically determined by the state of the world.

We now state the five axioms required to achieve this paper's measure of uncertainty, the cost of learning the state of the world:

Axiom 1: $C(S, \mu)$, the expected cost of a learning strategy $S = (\mathcal{P}_1, \ldots, \mathcal{P}_n)$, given a probability measure μ, is:

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

Axiom 1 asserts that the cost of a learning strategy S is simply the sum of the costs of learning the realizations of each of the partitions in S, given the agent's belief before observing each realization. Axiom 1 is a form of constant marginal cost because over the course of learning the agent does not fatigue, nor do they gain experience with research and become better at learning.

Axiom 2: Given a partition $\mathcal{P}$, for all probability measures μ:

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

$$\text{such that: } S^b \text{ is equivalent to } \mathcal{P}.$$

Axiom 2 asserts that the agent can always learn about their environment in an

[9] When we discuss an agent that attentively learns the state of the world, we refer to an agent that perfectly observes the state of the world.

efficient fashion using simple yes or no questions. This claim is supported by research in the psychology and psychophysics literatures. Eye tracking analysis shows that frequently 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?'. This assumption is made 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). This research thus supports this section's decision to model agents as learning through a series of yes or no questions.

Before we introduce the rest of our axioms, we pause to discuss learning strategy invariance, a concept that helps us to make it explicit what we are assuming with the rest of our 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 outcome 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 is then 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 be the cost of $\tilde{S}^b$. A set of partitions that are certainly learning strategy invariant, in contrast, is the set of binary partitions:

Lemma 1. If $\mathcal{P}^b$ is a binary partition, then $\mathcal{P}^b$ is learning strategy invariant.

Proofs for Section 3 can be found in Appendix 1.

Our three remaining axioms are concerned with the costs of questions that are learning strategy invariant. These axioms are rather weak in nature, only imposing structure onto the costs of a particular kind of question.

Axiom 3: Suppose $\mathcal{P}$ is a learning strategy invariant partition. Suppose μ is a probability measure such that n events $\{A_i\}_{i=1}^n \subset \mathcal{P}$ are given probability $1/n$. Suppose $\tilde{\mu}$ is a different probability measure such that $n+1$ events $\{B_j\}_{j=1}^{n+1} \subseteq \mathcal{P}$ are given probability $1/(n+1)$. The expected cost of observing the realized event of $\mathcal{P}$ is lower when the probability measure is μ than when the probability measure is $\tilde{\mu}$: $C(\mathcal{P}, \mu) < C(\mathcal{P}, \tilde{\mu})$.

Axiom 3 makes intuitive sense because there are more events that are realized with positive probability under $\tilde{\mu}$ than there are under μ, and each of these events is less likely to occur. Axiom 3 essentially states that differentiating between more

events that are less likely should be more expensive than differentiating between fewer events that are more likely.

Axiom 4: Suppose there are two probability measures μ and $\tilde{\mu}$ that assign positive probability to the same events in the learning strategy invariant partition $\mathcal{P}$. If we let μ_α be defined for $\alpha \in [0, 1]$ so that $\mu_\alpha(\omega) = \alpha\mu(\omega) + (1 - \alpha)\tilde{\mu}(\omega)$ for all $\omega \in \Omega$, $C(\mathcal{P}, \mu_\alpha)$ changes continuously in α.

Axiom 4 asserts that if the events that are given a positive probability of occurring in a learning strategy invariant partition $\mathcal{P}$ do not change, then the cost of learning which event in $\mathcal{P}$ has occurred, $C(\mathcal{P}, \mu)$, should change continuously with respect to μ. This property for C is intuitive since small changes in the chances of events occurring should not result in a large change in the cost of a question that differentiates between said events if which events are possible does not change.

Axiom 5: There is a function c such that for any learning strategy invariant partition $\mathcal{P} = \{A_1, \ldots, A_m\}$, there is a constant, $\lambda(\mathcal{P}) > 0$, such that: $C(\mathcal{P}, \mu) = \lambda(\mathcal{P})c(\mu(A_1), \ldots, \mu(A_m))$ for all probability measures $\mu : \mathcal{F} \to \mathbb{R}_+$.

Axiom 5 asserts that if $\mathcal{P}$ is learning strategy invariant, then two things should be true about $C(\mathcal{P}, \mu)$. First, $C(\mathcal{P}, \mu)$ should be measurable with respect to $\sigma(\mathcal{P})$. We are thus asserting that the expected cost of asking the question represented by $\mathcal{P}$ should be fully determined by the chance that each of its answers occurs. Second, if $\tilde{\mathcal{P}}$ is another learning strategy invariant partition, then the functions $C(\mathcal{P}, \mu)$ and $C(\tilde{\mathcal{P}}, \mu)$ should differ only by a multiplicative constant. We want to allow for learning strategy invariant questions to differ from each other with regard to how difficult they are to answer, but we do not wish for them to differ in a more fundamental way.

3.3 Minimal Cost of Learning

Learning the realized event of a partition can be done through different learning strategies, but what we seek is the minimal expected cost of doing so. Given some partition $\mathcal{P}$ of Ω, and probability measure μ, define:

$$C^*(\mathcal{P}, \mu) = \min_S C(S, \mu),$$

$$\text{such that: } S \text{ is equivalent to } \mathcal{P}.$$

$C^*(\mathcal{P}, \mu)$ tells us the minimal expected cost of a learning strategy that is equivalent to $\mathcal{P}$, the minimal expected cost of a learning strategy that always results in the same posterior as asking 'What is the realized event of the partition $\mathcal{P}$?', given a prior μ. Since Ω is a partition of itself, as a slight abuse of notation, we write $C^*(\Omega, \mu)$ to denote the minimal expected cost of observing the state of the world given the agent's prior μ.

Lemma 2. If partition $\mathcal{P}$ is learning strategy invariant, and C satisfies our five axioms, 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 defined as in equation (1).

While Shannon (1948) imposes learning strategy invariance onto all partitions of the possible states of the world, we instead allow for some partitions to be learning strategy invariant, and for some to not be. When learning is focused on a particular learning strategy invariant partition, learning costs are analogous to Shannon's (1948) original work, as is demonstrated by Lemma 2.

Lemma 1 and Lemma 2 together tell 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.[10] If not, let λ_2 denote the multiplier associated with all binary partitions, denoted $\{\mathcal{P}_i^{b,\lambda_2}\}_{i=1}^{n_2}$, with the second lowest multiplier.

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 $j \in \{1, \ldots, M\}$,[11] let $\mathcal{P}_{\lambda_j} = \times\{\mathcal{P}_i^{b,\lambda_j}\}_{i=1}^{n_j}$.

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 $M > 1$).

To establish our measure of uncertainty we want to describe the cost according to C^* of observing which state of the world has been realized. Again, since Ω is a partition of itself, we can, as a minor abuse of notation, write this cost as $C^*(\Omega, \mu)$: the minimal cost given μ of a learning strategy S such that $\sigma(S) = \sigma(\Omega) = \mathcal{F}$.

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

[11] M is defined in the proceeding paragraphs.

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 $\lambda_1 < \ldots < \lambda_M$, such that for any probability measure μ on $\mathcal{F}$:

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

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 learning the state of the world: $C^*(\Omega, \mu) - C^*(\Omega, \tilde{\mu})$, 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 axiomatic derivation of the cost benchmark in this paper requires a discrete outcome space for the state of the world, as is the case with Shannon Entropy. If a continuous outcome space is desired for the state of the world, however, a measure of uncertainty for a continuous outcome space can be defined in an analogous manner to the measure of uncertainty defined in Theorem 1 for a discrete outcome 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:

$$\mathbf{C}(F(s,\,\omega),\,\mu) = \mathbb{E}[C^*(\Omega,\,\mu) - C^*(\Omega,\,\mu(\cdot|s))].$$

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

The cost function defined above lies in the class of uniformly posterior-separable cost functions described by Caplin et al. (2017). The behavior generated in static settings by 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).

4.1 Rationally Inattentive Agent's Problem

As was discussed in Section 2, 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:

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

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

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 $S_n = \{s \in \mathbb{R}^N : a(s) = n\}$, 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:

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

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

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

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

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

Proofs for 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 3, 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.

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 3, 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 3 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 3.

Any choice behavior that complies with Corollary 1 and Theorem 2 is optimal. Optimal choice behavior may not be unique, however. 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 RU model that is analogous to RI with MSSE, and then revisit the two motivating examples, Example 1 and Example 2, from Section 2.

5.1 Analogous 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 (2018), 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, 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:

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

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 $\mathcal{P}_{\lambda_2} = \{\{\omega_1 \cup \omega_2\}, \{\omega_3 \cup \omega_4\}\}$.

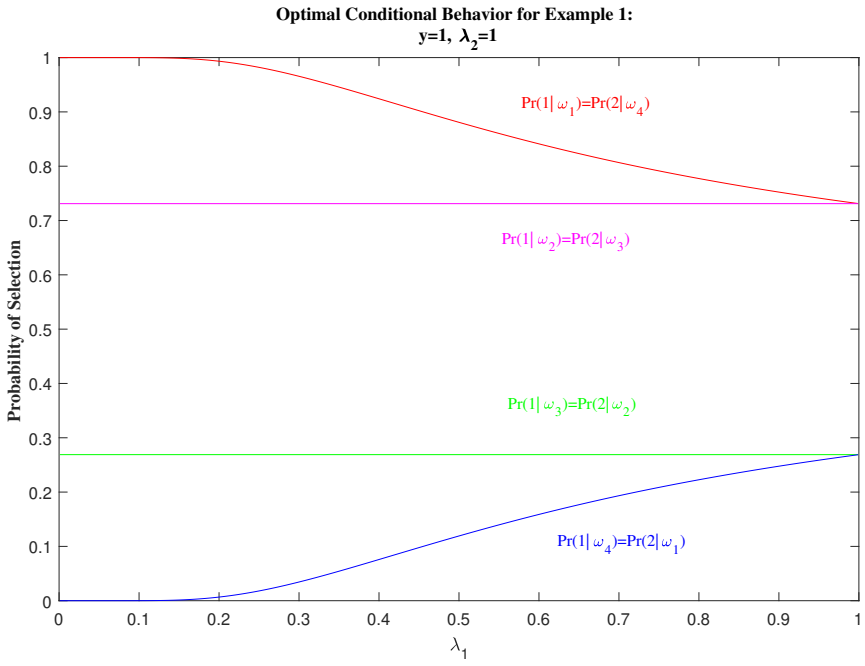
Solutions to Corollary 1 combined with Theorem 2 mean that the chance of the agent selecting option 1 is increasing in the number of blue balls, as can be seen in Figure 1, which depicts optimal behavior in each state of the world for a range of λ_1. When λ_1 is small relative to λ_2 the agent chooses 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 (2018), 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 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.

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

Figure 1:

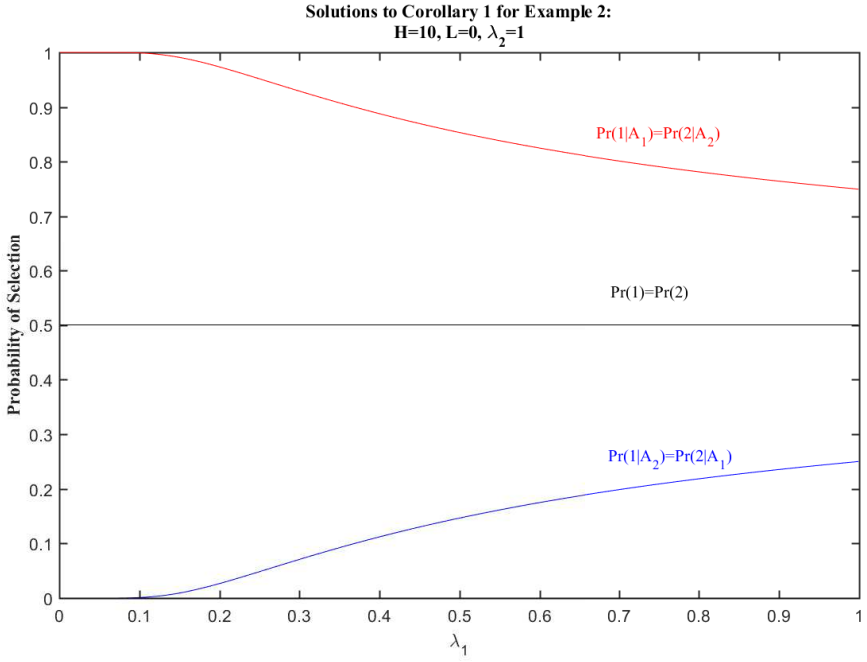


expected cost to the agent than the question 'Is option 2 of value H?'. We can thus assume: $\mathcal{P}_{\lambda_1} = \{A_1, A_2\} = \{\{\omega_1 \cup \omega_2\}, \{\omega_3 \cup \omega_4\}\}$, and $\mathcal{P}_{\lambda_2} = \{\{\omega_1 \cup \omega_3\}, \{\omega_2 \cup \omega_4\}\}$.

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 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

Figure 2:



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, depend-

ing 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

Proof of Lemma 1. If $\mathcal{P}^b$ is a binary partition, then 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 $C(S, \mu) = C(\mathcal{P}^b, \mu)$. ■

Proof of Lemma 2. Since $\mathcal{P}_i = (A_1, \ldots, A_m)$ is learning strategy invariant, if $\mathcal{P}_j = (A_1^j, \ldots, A_{m_j}^j)$ is another partition such that $\mathcal{P}_j \neq \mathcal{P}_i$, and $\sigma(\mathcal{P}_j, \mathcal{P}_i) = \sigma(\mathcal{P}_i)$, then $\mathcal{P}_j$ is also learning strategy invariant due to Axiom 1. Further, Axiom 1 and Axiom 5 tell us:

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

$$\implies \lambda_j c(\mu(A_1^j), \ldots, \mu(A_{m_j}^j)) + \mathbb{E}[\lambda_i c(\mu(A_1|\mathcal{P}_j(\omega)), \ldots, \mu(A_m|\mathcal{P}_j(\omega)))]$$

$$= \lambda_i c(\mu(A_1), \ldots, \mu(A_m)) + \mathbb{E}[\lambda_j c(\mu(A_1^j|\mathcal{P}_i(\omega)), \ldots, \mu(A_{m_j}^j|\mathcal{P}_i(\omega)))]$$

Since this is true for all μ, it is true if μ assigns probability $1/2$ to two events in each of $\mathcal{P}_i$ and $\mathcal{P}_j$ such that knowing the realized event of one of $\mathcal{P}_i$ and $\mathcal{P}_j$ tells you the realized event of the other. Thus:

$$\lambda_i c(1/2, 1/2) + \lambda_j c(1) = \lambda_j c(1/2, 1/2) + \lambda_i c(1).$$

Then, since Axiom 3 implies $c(1) < c(1/2, 1/2)$, it must be that $\lambda_i = \lambda_j$. Since this is true for all such $\mathcal{P}_j$, the function c satisfies Shannon's (1948) Axiom 3, as well as Axiom 2, Axiom 3, Axiom 4, and Axiom 5.

The rest of the proof follows the work of Shannon (1948) closely. Define h so $h(n) \equiv c(1/n, \ldots, 1/n)$. Shannon's (1948) Axiom 3 implies $h(s^m) = m \cdot h(s)$, which is reminiscent of logarithms. Given arbitrarily small $\epsilon > 0$, and integers s and t, pick

n and m so that $2/n < \epsilon$, and $s^m \leq t^n < s^{m+1}$. So:

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

Axiom 3 then tell us:

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

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

All of this tells us:

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

and thus $h(n) = K\log(n)$, where K must be a positive constant to satisfy Axiom 3. In our context it is without loss to assume $K = 1$, since we can just rescale the associated multiplier λ_i for each learning strategy invariant partition $\mathcal{P}_i$.

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

$$p_i = \frac{n_i}{\sum_{j=1}^{m} n_j}.$$

Our interpretation is that we have a uniform distribution over $\sum_i n_i$ equally likely states, and the chance of the event which happens with probability p_i is the chance of one of the n_i associated states occurring. Then using Shannon's Axiom 3:

$$c\left(\frac{1}{\sum_i n_i}, \ldots, \frac{1}{\sum_i n_i}\right) = h\left(\sum_{i=1}^{m} n_i\right) = \log\left(\sum_{i=1}^{m} n_i\right) = c(p_1, \ldots, p_m) + \sum_{i=1}^{m} p_i \log(n_i),$$

$$\Longrightarrow c(p_1, \ldots, p_m) = \log\left(\sum_{i=1}^{m} n_i\right) - \sum_{i=1}^{m} p_i \log(n_i) = \sum_{j=1}^{m}\left(p_j \log\left(\sum_{i=1}^{m} n_i\right)\right) - \sum_{i=1}^{m} p_i \log(n_i)$$

$$= -\sum_{i=1}^{m} p_i \log\left(\frac{n_i}{\sum_j n_j}\right) = -\sum_{i=1}^{m} p_i \log(p_i) = \mathcal{H}(\mathcal{P}_i,\, \mu),$$

where $\mathcal{H}$ is defined as in equation (1). If any of the p_i are irrational, then the density of the rationals and Axiom 4 can be used to get the same result. The learning strategy invariance of $\mathcal{P}_i$, and Axiom 5, thus give us:

$$C^*(\mathcal{P}_i,\, \mu) = C(\mathcal{P}_i,\, \mu) = \lambda_i c(\mu(A_1^i),\, \ldots,\, \mu(A_{m_i}^i)) = \lambda_i \mathcal{H}(\mathcal{P}_i,\, \mu).\blacksquare$$

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 4.

Lemma 4. 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 $C(S^b, \mu) = C^*(\Omega, \mu)$. We know such binary learning strategy exists whenever C satisfies Axiom 2. 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 4.

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 $C(\tilde{S}^b, \mu) = C^*(\Omega, \mu)$. This is true since each appended binary partition has an expected cost of zero, since $\sigma(S^b) = \mathcal{F}$. Lemma 4 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: $C(\hat{S}^b, \mu) = C^*(\Omega, \mu)$. 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:

$$C^*(\Omega, \mu) = C(S^b, \mu)$$

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

Appendix 2

Proof of Lemma 3. In Lemma 3, 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 5, we show that $C^*(\Omega,\, \mu)$ is a strictly concave function of μ. This is a commonly known property of Shannon Entropy, but needs to be established for C^*. This implies that $\mathbf{C}$ is strictly convex. We then show, in Lemma 6, 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 7, we show that we may rewrite the cost function in terms of the choice probabilities in equations (4), (5), and (6).

Lemma 5. $C^*(\Omega,\, \mu)$ 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:

$$C^*(\Omega,\, \mu) > \alpha C^*(\Omega,\, \mu_a) + (1-\alpha)C^*(\Omega,\, \mu_b).$$

Proof. For each probability measure μ, $i \in \{1,\, \ldots,\, M\}$, and realization of $\cap_{j=1}^{i-1}\mathcal{P}_{\lambda_j}(\omega)$, the strict concavity of Shannon Entropy (Matějka & McKay, 2015; Caplin et al., 2017) implies:

$$\mathcal{H}(\mathcal{P}_{\lambda_i},\, \mu(\cdot|\cap_{j=1}^{i-1}\mathcal{P}_{\lambda_j}(\omega))) \geq \alpha\mathcal{H}(\mathcal{P}_{\lambda_i},\, \mu_a(\cdot|\cap_{j=1}^{i-1}\mathcal{P}_{\lambda_j}(\omega)))+(1-\alpha)\mathcal{H}(\mathcal{P}_{\lambda_i},\, \mu_b(\cdot|\cap_{j=1}^{i-1}\mathcal{P}_{\lambda_j}(\omega))).$$

The inequality is also strict for at least one $i \in \{1,\, \ldots,\, M\}$ since $\mu_a \neq \mu_b$. The desired result thus follows from Theorem 1.■

Lemma 6. 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 $s_2 \in S_n^2$. C^* is strictly concave, 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 $\tilde{F}(s_0|\omega) = \int\limits_{s \in S_n^1} F(s|\omega) + \int\limits_{s \in S_n^2} F(s|\omega)$. 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 $\forall\, \nu \in \mathcal{N}$:

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

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

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

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

.

Lemma 7. 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 then have:

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

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

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

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

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

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

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

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

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

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

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

We now resume our proof of Lemma 3. For each $n \in \mathcal{N}$, let s_n denote some

signal in S_n. 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 $m \in \{1, \ldots, M-1\}$, then $\Pr(n|\omega) = 0\ \forall \omega$. 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:

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

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

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

Where $\alpha_\nu^0 = \log(N\Pr(\nu))$, and for $m \in \{1, \ldots, M-1\}$ 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:

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

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 alternatives, 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 C^* is concave. 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.

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

Dean, M., & Neligh, N. L. (2018). 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.

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.

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 similar-

ity 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, September). *Rational inattention and non-compensatory choice.*

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