Rational Inattention with Multiple Attributes

David Walker-Jones*

University of Surrey

July 11, 2023

Abstract

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

JEL Classification: D83

1 Introduction

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

The standard technique for quantifying the cost of learning in models of rational inattention (RI) is Shannon Entropy (Shannon, 1948; Sims, 2003; Maćkowiak, Matějka, & Wiederholt, 2023). Shannon Entropy has an axiomatic foundation, is grounded in the optimal coding of information,

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

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

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

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

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

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

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

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

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

1.1 Organization of Paper

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

2 Rational Inattention and Shannon Entropy

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

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

In the rational inattention (RI) literature learning by the agent is typically modelled as the choice of a signal structure, which means the agent chooses the probability of receiving different

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

signals in the different states of the world. Receiving a signal updates the agent's belief about the state of the world, giving them a more informed posterior belief. More informative signal structures are more costly for the agent, but allow them to make a more informed decision about which option to select.

The agent's problem is thus to maximize the expected value of their selected option less the cost of learning. They do this by choosing an **information strategy** $F \in \Delta(S \times \Omega)$, which is a joint distribution between s, the observed **signal** from some arbitrarily large and finite set of signals S, and the states of the world.[3] The only restriction on the information strategy is that the marginal, $F(\omega) : \mathcal{F} \to [0, 1]$, must equal the prior μ.

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

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

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

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

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

The choice behavior the agent exhibits depends on the cost function for information. Shannon Entropy is a measure of total uncertainty that is frequently used to assign costs to information (Matějka & McKay, 2015; Maćkowiak et al., 2023). Given a partition (defined formally in Section 3) of the possible states of the world $\mathcal{P} = \{A_1, \ldots, A_m\}$, and a probability measure μ over these

[3] The specifics of S is rather unimportant. This is a richer signal space than is required in practice. It is shown that an optimal strategy only results in one of at most N different signals being observed. If σ is the set of events generated by S, then the agent is selecting a probability measure $F : \sigma \times \mathcal{F} \to \mathbb{R}_+$.

events, the uncertainty about which event has occurred, as measured by **Shannon Entropy**, is defined:[4]

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

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

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

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

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

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

2.1 Example 1: Multiple Attributes and Problems with Predictions

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

Consider the environment described in Table 1 where an agent is faced with a screen that shows 100 balls, each of which is either red or blue. The agent is offered a prize that they may either accept (option 1), or reject to get a payoff of zero (option 2). The agent is told that if the majority of the balls on the screen are blue then the prize is $y \in \mathbb{R}_{++}$, and if the majority of the

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

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

balls on the screen are red then the prize is $-y$. Suppose further that the agent is also told that there is a 1/4 chance of each of four different states of the world in which there are either 40, 49, 51, or 60 red balls.

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

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

While other information cost functions do not require that choice behavior satisfies invariance

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

under compression (Caplin et al., 2022; Morris & Yang, 2022), they lack the tractability and flexibility of Shannon Entropy,[5] which limits the potential for their application. This has led to the following open question: "what workable alternative models allow for the complex behavioral patterns identified in practice?" (Caplin, Dean, & Leahy, 2017, p. 2), a question that this paper attempts to answer. MASE solves the problem with predictions outlined in this example by allowing option 1 to have multiple attributes that differ in their learning cost, as is explained in Section 4.2.

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

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

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

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

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

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

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

This type of bias has not previously been identified in the literature on RI: Matějka and McKay (2015) show that fitting of multinomial logit results in the value of any option n being biased by the (weighted) average probability of it being selected over states ω. The bias found by Matějka and McKay (2015) can thus be identified by examining the average probabilities of the agent selecting each option because the driving mechanism is that the cost of learning causes the agent to be biased towards options that they have a higher probability of selecting *a priori*. The bias previously found by Matějka and McKay (2015) is fundamentally different than the bias demonstrated in this example because their bias does not allow for an option to be over valued in 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 analyst who observes equal average choice probabilities in this setting, as is predicted by MASE (this is shown rigorously in Section 4.3), might be tempted conclude, based on the previous literature, that their analysis is not susceptible to informational biases since each option has the same unconditional probability of being selected *a priori*, and thus any counterfactual or welfare analysis that they conduct is valid. This conclusion may not be correct given the results in this paper.

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

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

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

3 Inattentive Learning with MASE

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

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

A **partition** $\mathcal{P}$ of a state space Ω is a set of more than one disjoint events in $\mathcal{F}$ whose union is Ω.[8] For each event $A \in \mathcal{F}$, define the **complement** of the event, denoted A^c, to be the set of states that are not in A, so $A^c = \Omega \backslash A$, and thus $\{A, A^c\}$ forms a partition. If $\omega \in \Omega$ is the state of the world, let the **realized event** of the partition $\mathcal{P} = \{A_1, \ldots, A_m\}$ be denoted by $\mathcal{P}(\omega)$, that is $\mathcal{P}(\omega) = A_i \in \{A_1, \ldots, A_m\}$ iff $\omega \in A_i$.

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

Using these attributes and their associated multipliers, define **Multi-Attribute Shannon**

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

Entropy (MASE), $\mathbb{H} : \Delta(\Omega) \to \mathbb{R}_+$, to be the measure of total uncertainty:

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

where $\mathcal{H}$ is Shannon Entropy, which is defined in equation (3). This paper refers to $\mathbb{H}$ as a measure of total uncertainty because, given any probability measure over states, it describes the minimal expected cost of perfectly observing the state of the world, as is true with Shannon Entropy in the Shannon RI model. The formula for $\mathbb{H}$ can be interpreted as describing the agent as learning the state of the world by successively learning the realizations of the different attributes, learning about the less costly to learn about attributes first so as to minimize the cost of learning any information that the attributes share. A more detailed discussion of the rationale for this functional form can be found in the work of Walker-Jones (2023), which also provides conditions that describe when a dataset is sufficient for the unique identification of the MASE cost function for learning.

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

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

where $\mathbb{H}(\mu)$ is as defined in equation (4) and $\mu(\cdot|s) : \mathcal{F} \to [0,\, 1]$ is the distribution over states induced by the signal s, the prior μ, the information strategy F, and Bayes' Rule, which is to say $\mu(\omega|s) = F(\omega|s)$ for each state ω. This definition of the cost of learning is the same as in the standard Shannon model of RI studied by Matějka and McKay (2015) except Shannon Entropy is replaced by MASE. Because $\mathbb{H}(\mu)$ is shown by Lemma 3 to be a strictly concave function of μ, the work of Mensch (2018) indicates that $\mathbf{C}(F,\, \mu)$ is monotone in Blackwell (1953) informativeness.

3.1 Selecting Optimal Choice Probabilities

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

$A \in \mathcal{F}$ to be:

$$\Pr(n|A) = \sum_{\omega \in A} \left(\sum_{s \in \mathcal{S}(n|F)} F(s|\omega) \right) \mu(\omega|A). \tag{6}$$

Then, as a minor abuse of notation, define the **unconditional probability** of option n being selected to be the probability of n being selected conditional on the event $A = \Omega$: $\Pr(n) \equiv \Pr(n|\Omega)$. Denote the collection of $\Pr(n|\omega)$ for each $n \in \mathcal{N}$ and $\omega \in \Omega$ by $\mathbb{P}$, which is referred to as the agent's observable **behavior**. Using this notation, the agent's problem can be re-written:

Lemma 1. Behavior $\mathbb{P}$, and the resultant information strategy F constructed from $\mathbb{P}$ such that for each $n \in \mathcal{N}$ there is a single signal s^n that results in option n being selected and $F(s^n|\omega) = \Pr(n|\omega)$ for all $\omega \in \Omega$, is a solution to the agent's problem in (1) subject to (2) iff it solves:

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

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

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

Further, the objective described by equation (7) is concave on the set of $\mathbb{P}$ that satisfy (8) and (9). Proofs for results in Section 3 and Section 4 can be found in Appendix 1.

The new problem outlined in Lemma 1, where the agent selects their conditional choice behavior $\mathbb{P}$, is substantially easier to solve than the problem where the agent picks their information strategy F. If behavior solves (7) subject to (8) and (9) then it is referred to as **optimal**.

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

$$\Pr(n|\omega) = \frac{\Pr(n) e^{\frac{\mathbf{v}_n(\omega)}{\lambda}}}{\sum_{\nu \in \mathcal{N}} \Pr(\nu) e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda}}}. \tag{10}$$

In the Shannon RI model the probability of the agent selecting an option in any given state thus

depends on both the unconditional probabilities of the options being selected and the realized values of the options in said state.

3.2 Behavior of a Rationally Inattentive Agent with MASE

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

Theorem 1.

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

$$\Pr(n|\omega) = \frac{\Pr(n)^{\frac{\lambda_1}{\lambda_M}} \Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \ldots \Pr(n|\cap_{i=1}^{M-1} \mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}}{\sum_{\nu\in\mathcal{N}} \Pr(\nu)^{\frac{\lambda_1}{\lambda_M}} \Pr(\nu|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \ldots \Pr(\nu|\cap_{i=1}^{M-1} \mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_M}}}. \tag{11}$$

Those familiar with the work of Matějka and McKay (2015) will recognize the above formula as the MASE analogue of the necessary condition for optimal behavior stated in Matějka and McKay (2015)'s Theorem 1 and depicted in equation (10) for the Shannon RI model. When there is only one attribute to learn about and $\lambda_1 = \lambda_2 = \cdots = \lambda_M = \lambda$, the above formula collapses to the one from Matějka and McKay (2015)'s Theorem 1.

For $m \in \{1, \ldots, M-1\}$, $\Pr(n|\cap_{i=1}^{m} \mathcal{A}_i(\omega))$ is the average probability of n being selected given the realizations of the m easiest to learn about attributes. With MASE, as the above formula indicates, the chance of the agent selecting an option n in a particular state of the world ω depends on the the unconditional probability of n being selected and its realized value, $\Pr(n)$ and $\mathbf{v}_n(\omega)$, as well as the unconditional probabilities of the other options being selected and their realized values, $\Pr(\nu)$ and $\mathbf{v}_\nu(\omega)$ for each $\nu \in \mathcal{N}$ with $\nu \neq n$, as is the case with Shannon Entropy. But, with MASE, the chance of the agent selecting an option n further depends on the realizations of the attributes that are easier to learn about. It makes sense that when easier to observe pieces of information indicate that an option n is likely of above average value, that the agent should select option n with a higher probability, even if the above average value has not been realized.

The behavior described in Theorem 1 has many intuitive features. It is also a quite natural extension of the analogous result from Matějka and McKay (2015) for the Shannon RI model, which is described in equation (10) and also has many intuitive features. If $\lambda_2 = \lambda$ grows (shrinks) in (10), which represents an increase (decrease) in the difficulty of learning, the value of each option in the realized state becomes less (more) significant for the determination of the selected option, and the significance of the agent's prior increases (decreases). If λ_2 approaches infinity, the realized values become insignificant, and the behavior of the agent approaches the behavior of the agent in the case where learning is impossible: they choose their option based on their prior. If λ_2 approaches zero the unconditional priors become insignificant, and the behavior of the agent approaches the behavior of the agent in the case where learning is costless: they choose the option with the highest realized value.

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

$$\Pr(n|\omega) = \frac{\Pr(n)^{\frac{\lambda_1}{\lambda_2}} \Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_2}} e^{\frac{\mathbf{v}_n(\omega)}{\lambda_2}}}{\sum_{\nu\in\mathcal{N}} \Pr(\nu)^{\frac{\lambda_1}{\lambda_2}} \Pr(\nu|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_2}} e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_2}}}. \tag{12}$$

With MASE, as the formula in (12) indicates, the probability of the agent selecting an option n in a particular state of the world ω depends not only on the unconditional probabilities of the options being selected and the realized values of the options, but also on the realized value of $\mathcal{A}_1$. When option n is in general desirable in $\mathcal{A}_1(\omega)$ relative to the other options, then $\Pr(n|\mathcal{A}_1(\omega))$ is larger, and there may be a high probability of n being selected, even if $\Pr(n)$ is not that large, and $\mathbf{v}_n(\omega)$ is not that high.

The formula in (12) also has many intuitive features. It maintains the intuitive comparative statics for λ_2 that the formula in (10) has, and also features intuitive properties for $\Pr(n|\mathcal{A}_1(\omega))$ and λ_1.

If observing $\mathcal{A}_1(\omega)$ is completely uninformative about the value of the options, then it is optimal for the agent to select $\Pr(n|\mathcal{A}_1(\omega)) = \Pr(n)$ since $\mathbb{H}$ is strictly concave in μ. In this case $\Pr(n)^{\frac{\lambda_1}{\lambda_2}} \Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_2}} = \Pr(n)$, and behavior is identical to that in (10). If the cheaper to learn about attribute is irrelevant it is thus ignored, and behavior collapses back to the environment

described in Matějka and McKay (2015), as should be desired.

If λ_1 approaches λ_2 (the cheaper to learn about attribute becomes close to as expensive as the more expensive to learn about attribute) then behavior approaches that described in (10) since $\Pr(n)^{\frac{\lambda_1}{\lambda_2}}\Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_2}} \to \Pr(n)$. Thus, if an insignificantly cheaper to learn about attribute is introduced, behavior is changed in an insignificant fashion (see Figure 1). Again, this seems like a desirable property.

If λ_1 approaches zero then the role of the unconditional prior dissipates, and the exponent on $\Pr(n|\mathcal{A}_1(\omega))$ approaches one, meaning it replaces the unconditional prior from (10). This makes sense because if λ_1 goes to zero it means $\mathcal{A}_1(\omega)$ can essentially be viewed for free, in which case behavior within each $\mathcal{A}_1(\omega)$ should resemble that in the setting where there is only one attribute with multiplier λ_2 and a prior of $\mu(\cdot|\mathcal{A}_1(\omega))$.

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

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

Lemma 2.

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

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

such that:

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

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

numbers of appropriate dimension.

Lemma 2 is helpful for two main reasons. First, Lemma 2 indicates that if behavior is such that all options are selected with a positive probability and Theorem 1 is satisfied, then it is optimal. Second, Lemma 2 reduces the number of choice variables faced by the agent, which means it is easier for the researcher to find optimal agent behavior, and the optimization problem in Lemma 2 is a standard concave maximization program that can be easily solved numerically.

Theorem 3, which can be be found in Appendix 1, provides necessary and sufficient conditions for optimal behavior in settings where MASE is used to measure the cost of information. Theorem 3 thus establishes the MASE analogue of Caplin et al. (2018)'s Proposition 1, their central proposition.

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

4 Comparisons with Standard Models

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

4.1 Comparison with Random Utility Model

RU models are frequently used to fit behavior in discrete choice settings. In such a model, the agent picks the option $n \in \mathcal{N}$ with the largest sum $u_n = v_n + \epsilon_n$. Generally, u_n represents the value of the option to the agent, v_n represents the average value of the option across agents, and ϵ_n represents an idiosyncratic value to the agent. The role ϵ_n plays is up to interpretation, however, and is determined by the researchers specification (Train, 2009). In a setting where agents are thought to be rationally inattentive, the above terms are interpreted in a different way because the agent's noisy behavior is generated by perceptual error instead of idiosyncratic differences in

taste. In such settings, u_n represents the perceived value to the agent, v_n represents the true value to the agent, and ϵ_n is interpreted as an unobservable perceptual error that results from the noisy information strategy selected by the agent. Woodford (2014) argues that this latter interpretation is necessary in many contexts due to the stochastic responses observed in perceptual discrimination tasks such as those administered by Dean and Neligh (2022), which are akin to Example 1 in Section 2.1. While the interpretation of ϵ_n is relevant for welfare analysis, it is inconsequential for the description of choice behavior. How then can MASE be interpreted in terms of a RU framework, and what insights may be provided about the fitting of RU models?

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

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

Theorem 2:

If the choice behavior $\mathbb{P}$ is optimal then it is identical to the behavior produced by a RU model where each option $n \in \mathcal{N}$ has perceived value:

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

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

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

Theorem 2 is meant to provide insight into the outcome of attempting to fit a RU model in an environment where agents are rationally inattentive with a cost function for information described

by MASE. If an analyst that is trying to infer the different $\tilde{v}_n(\omega)$s, the payoffs of the agent from different options in different states of the world (normalized by λ_M), does so by performing a multinomial logit regression then their estimate $u_n(\omega)$ of the (normalized) value of an option n in state ω would be biased by $\alpha_n(\omega)$.[9] For an example of this type of bias, and how it depends on the realized state of the world and the relative values of the different λs, see Section 4.3 in which Example 2 is revisited.

Theorem 2 does not say that a model of RI with MASE is equivalent to a RU model. Even if choice data from a given choice problem cannot be used to reject one for the other, across choice problems MASE produces behavior that can reject the hypothesis of a RU model. With MASE, for instance, as with standard Shannon Entropy, adding an option can increase the probability of an existing option being selected, which is not possible with a RU model.

Also, it is worth mentioning that since optimal behavior may result in some options being selected with probability zero, Theorem 2 implicitly defines each $\alpha_n(\omega)$ on the extended reals so that $\alpha_n = -\infty$ if $\Pr(n) = 0$.[10]

4.2 Example 1 Revisited

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

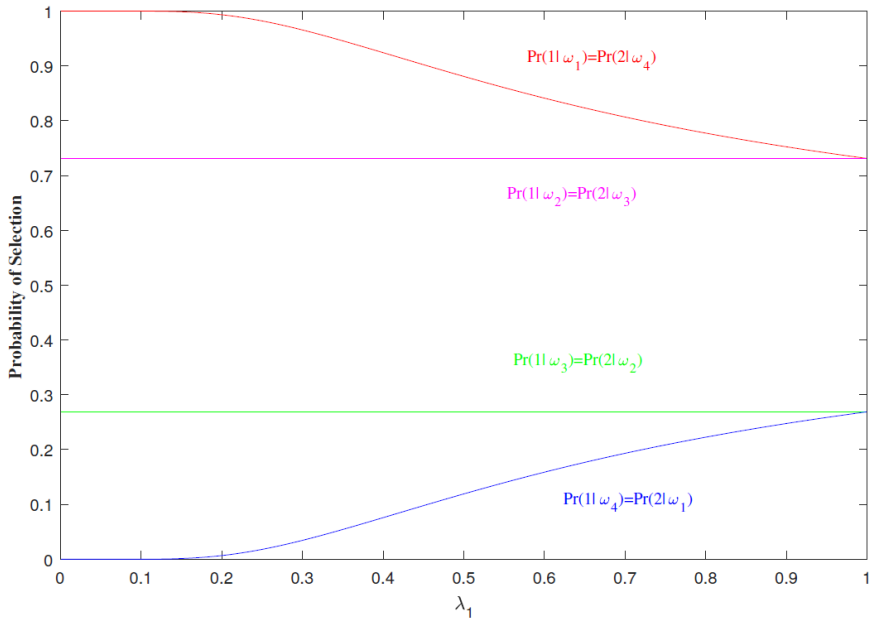
Solutions to Lemma 2 combined with Theorem 1 mean that the probability of the agent selecting option 1 is increasing in the number of blue balls, as can be seen in Figure 1, which depicts optimal behavior in each state of the world for a range of λ_1. When λ_1 is small relative to

[9] Whether or not the analyst knows μ is inconsequential.

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

λ_2 the agent chooses option 1 in state ω_1 with a high probability, and choose option 2 in state ω_4 with a high probability. The agent is thus better able to discern the state of the world when there are 40 of one color ball and 60 of the other than when there are 49 of one color and 51 of the other. This is supported by the experimental work of Dean and Neligh (2022), and is in contrast with the behavior predicted by the Shannon RI model.

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



On the horizontal axis λ_1, the cost parameter associated with learning if there are 60 or more of one color of ball present, varies from 0 to 1 while the cost of learning that there are 51 of a certain color of ball, λ_2, is held constant at 1. Both states ω_1 and ω_2 feature a value of 1 for option 1, but behavior differs across the states since ω_1 is easier to identify for the agent as in it there are more of the color of ball that constitutes the majority. Both states ω_3 and ω_4 feature a value of -1 for option 1, but behavior differs across the states since ω_4 is easier to identify for the agent as in it there are more of the color of ball that constitutes the majority. The value of selecting option 2 is always 0.

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

4.3 Example 2 Revisited

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

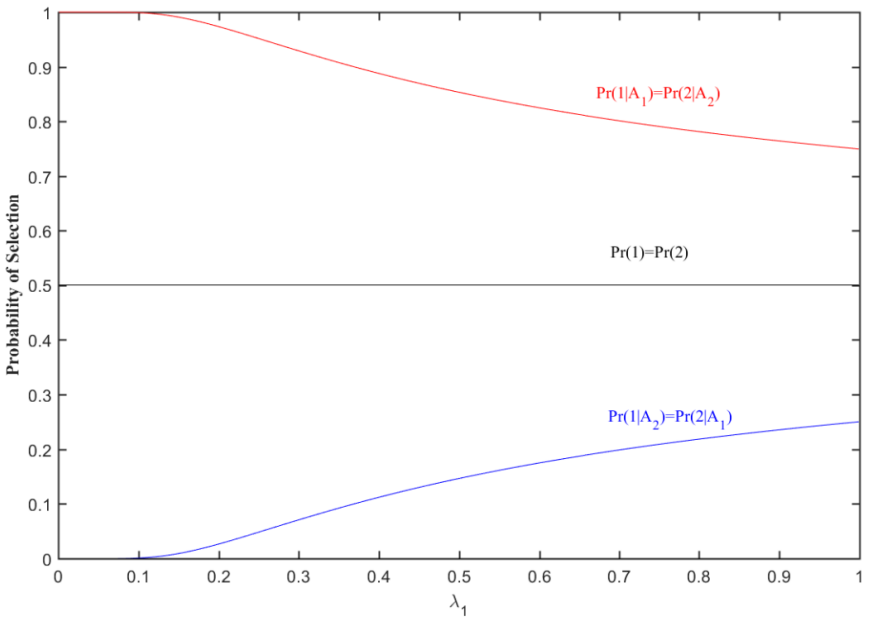
Solutions to Lemma 2 in this environment for a range of λ_1 can be found in Figure 2, which shows that when λ_1 is small compared to λ_2, the agent selects option 1 with a high probability when it is of value H, and selects option 2 with a high probability when option 1 is of value L. As λ_1 increases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

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

Further, as is mentioned in Section 2, RU models and the Shannon RI model can both be rejected for RI with MASE if it is possible to alter the correlation between the values of the two options while holding the marginal distributions over values fixed for each option by changing the distribution over states in this example so it is no longer uniform.[11] If a RU model describes the agent, then changing the correlation between the values of the two options would not change the choice behavior of the agent in any state. If the behavior of the agent is instead described by MASE, then changing the correlation between the values of the two options would change the choice

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

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



On the horizontal axis λ_1, the cost parameter associated with learning the value the easier to learn about option 1, varies from 0 to 1 while the cost parameter associated with learning option 2, λ_2, is held constant at 1. A_1 is the event in which option 1 has high value H and A_2 is the event in which option 1 has low value L. So, $\Pr(1|A_1)$ is thus the average chance of option 1 being selected when it has high value (weighted averaged across the states where option 1 has a high value), while $\Pr(2|A_2)$ is the average chance of option 2 being selected when option 1 has low value. $\Pr(1)$ and $\Pr(2)$ are the average chances of option 1 and 2 being selected (weighted averaged across all states).

behavior of the agent in individual states because the total information that can be acquired from learning the value of option 1 (the option that is easier to learn about) changes with the correlation of the options' values. Further, if the above MASE specification is correct, the unconditional choice probabilities of the agent would remain constant when correlation is changed due to the symmetry of the environment, as long as the agent is doing some learning.[12] Finally, if the behavior of the agent is instead described by Shannon Entropy, in contrast, then the choice behavior in the individual states could only change if the unconditional choice probabilities changed.

5 Literature Review

Shannon Entropy has been used in several contexts to demonstrate informational biases in RU models. Matějka and McKay (2015) use the Shannon RI model to demonstrate the potential for informational biases in multinomial logit, while Steiner, Stewart, and Matějka (2017) use Shannon

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

Entropy in a model of RI to demonstrate the potential for a similar bias in dynamic Logit. These results are significant for those who wish to fit RU models because, while observational data may coincide with the assumptions of a fitted RU model, informational biases can potentially invalidate counterfactual and welfare analysis, two common goals of such a fitting.

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

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

A number of recent papers, such as the work of Pomatto, Strack, and Tamuz (2023), feature models that, like MASE, allow for perceptual distance. Unlike the work of Pomatto et al. (2023), which features axioms that are concerned with probabilistic experiments that can result in different outcomes in the same state of the world, this paper's cost of information is based on axioms that can be found in the work of Walker-Jones (2023) and are concerned with deterministic experiments that always result in the same outcome in a given state of the world, and contradicts the form of constant marginal cost assumed in their paper.

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

The cost functions explored in this paper that measure the reduction in MASE are a strict subset of the neighborhood-based cost functions described by Hébert and Woodford (2021). While

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

Huettner, Boyacı, and Akçay (2019), in turn, create an ad hoc group of cost functions that are also a generalization of Shannon Entropy, but are a subset of the cost functions studied in this paper that measure reduction in MASE. The cost functions studied by Huettner et al. (2019) allow different options to have different learning costs associated with them, but are not capable of predicting the behavior that is argued to be intuitive in Example 1 without additional states of the world being introduced. Further, the sufficient conditions for optimal behavior provided by Huettner et al. (2019) contradict the sufficient conditions provided by this paper's Theorem 3.

Walker-Jones (2023) provides conditions that describe if a dataset is sufficient for the unique identification of the MASE cost function for learning. Such a dataset features observed behavior from simple choice problems, choice problems where two options are available and only a few states of the world occur with a positive probability, and identifies the MASE cost function, both a set of attributes and their associated learning costs, that determines the cost of differentiating between outcomes when any subset of the potential states of the world occur with a positive probability.

6 Conclusion

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

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

3 establishes the MASE analogue of Caplin et al. (2018)'s necessary and sufficient conditions for optimal behavior in the context of Shannon Entropy.

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Pomatto, L., Strack, P., & Tamuz, O. (2023). The cost of information: The case of constant marginal costs. *American Economic Review*, *113*(5), 1360–1393.

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

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

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

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

Walker-Jones, D. (2023). Foundation and identification of multi-attribute shannon entropy. Retrieved from https://www.dwalkerjones.com

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

Appendix 1

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\mu(\omega) = \alpha\sum_{s\in S} F^1(s, \omega) + (1-\alpha)\sum_{s\in S} F^2(s, \omega) = \sum_{s\in S} F(s, \omega).$$

Next, notice that for each s:

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

and for each s with $F(s) > 0$ and ω:

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

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

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

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

So:

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

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

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

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

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

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

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

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

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

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

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

and $\mathbf{C}(\mathbb{P},\,\mu)$, so defined, is convex in $\mathbb{P}$ that satisfy equations (8) and (9): if $\tilde{\mathbb{P}}$ (a $\tilde{\Pr}(n|\omega)$ for each option n and state ω) and $\hat{\mathbb{P}}$ (a $\hat{\Pr}(n|\omega)$ for each option n and state ω) both satisfy (8) and (9), then $\mathbb{P}$ defined by $\Pr(n|\omega)=\alpha\tilde{\Pr}(n|\omega)+(1-\alpha)\hat{\Pr}(n|\omega)$ for each option n and state ω satisfies (8) and (9) and $\mathbf{C}(\mathbb{P},\,\mu)\leq\alpha\mathbf{C}(\tilde{\mathbb{P}},\,\mu)+(1-\alpha)\mathbf{C}(\hat{\mathbb{P}},\,\mu)$.

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

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

$$=\mathbb{E}\Big[\lambda_1\Big(\mathcal{H}(\mathcal{A}_1,\,\mu)-\mathcal{H}(\mathcal{A}_1,\,\mu(\cdot|s))\Big)\tag{13}$$

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

$$=\mathbb{E}\Big[\lambda_1\Big(\mathcal{H}(\mathcal{P}_s,\,F)-\mathcal{H}(\mathcal{P}_s,\,F(\cdot|\mathcal{A}_1(\omega)))\Big)\tag{14}$$

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

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

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

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

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

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

The equality of (13) and (14) follows from the symmetry of mutual information. It is evident that if $\tilde{\mathbb{P}}$, $\hat{\mathbb{P}}$, and $\mathbb{P}$, are defined as in the statement of this lemma, then $\mathbb{P}$ satisfies (8) and (9). Convexity follows almost directly from Lemma 3 and what is shown above as any behavior $\mathbb{P}$ that satisfies (8) and (9) can be used to define an information strategy that satisfies (2)) by setting, for each state ω, $F(s^n,\,\omega) = \Pr(n|\omega)\mu(\omega)$ for each $n\in\mathcal{N}$ and $F(s,\,\omega) = 0$ otherwise. ■

Proof of Lemma 1. First, Lemma 5 implies the objective described by equation (7) is concave on the set of $\mathbb{P}$ that satisfy (8) and (9) as payoffs are linear and the cost of information is convex on the set of $\mathbb{P}$ that satisfy (8) and (9).

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

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

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

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

where the last step follows from the fact that $\Pr(X|Y)\Pr(Y) = \Pr(Y|X)\Pr(X)$. The rest of the proof proceeds with two proofs by contradiction. First, assume that F achieves expected utility U_1, and let $\mathbb{P}$ be the behavior induced by it. Assume that $\mathbb{P}$ is not a solution to (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, an information strategy $\tilde{F}$ and a choice of option for each signal can be created that generates $\tilde{\mathbb{P}}$. For instance, for each of the N distinct signals s^n, let the option selected by the agent after they observe s^n be denoted $\tilde{a}(\tilde{F}(\omega|s^n)) \equiv n$, and let $\tilde{F}(s^n,\,\omega) = \tilde{\Pr}(n|\omega)\mu(\omega)\ \ \forall\,\omega$ so that (2)

is satisfied. This is impossible though as then $(\tilde{F}, \tilde{a})$ achieves $U_2 > U_1$ and F cannot have been optimal.

Similarly, assume that $\mathbb{P}$ is a solution to (7) subject to (8) and (9), which achieves expected utility U_3, but is not induced by a solution to (1) subject to (2). That is, there is a $\tilde{F}$ which satisfies (2), produces a certain posterior with probability one after each option that is selected with a positive probability is selected (without loss given Lemma 4), and achieves $U_4 > U_3$. This means, however, that behavior defined for each option n and state ω by:

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

also achieves U_4 by Lemma 4 and Lemma 5, which is impossible as $\mathbb{P}$ was supposedly optimal and $\tilde{\mathbb{P}}$ satisfies (8) and (9). ■

Proof of Theorem 1. The Lagrangian for the problem depicted in Lemma 1 can be written (given the result in Lemma 5 and the definition of $\mathbf{C}(\mathbb{P}, \mu)$ therein):

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

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

$\xi_n(\omega) \geq 0$ are the multipliers for (8), and $\gamma(\omega)$ are the multipliers for (9).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\Pr(n|\omega)=\Pr(n)^{\frac{\lambda_1}{\lambda_M}}\Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ldots\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}e^{\frac{-\gamma(\omega)}{\lambda_M}}\qquad(15)$$

Plugging (15) into (9), one can solve for $\gamma(\omega)$. Plugging $\gamma(\omega)$ back into (15) achieves the desired result. ■

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

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

To help make the notation more compact, a group of partitions can be used to **generate** a finer partition: if $(\mathcal{P}_1, \ldots, \mathcal{P}_m)$ is a group of partitions, let $\times\{\mathcal{P}_i\}_{i=1}^n$ denote the partition such that for all $\omega \in \Omega$: $\times\{\mathcal{P}_i\}_{i=1}^n(\omega) = \cap_{i=1}^n \mathcal{P}_i(\omega)$.

Proof of Lemma 2. In this proof it is said behavior $\mathbb{P}$ satisfies Theorem 1 if: for each $n \in \mathcal{N}$ if $\Pr(n) > 0$ then for all $\omega \in \Omega$ it is the case that $\Pr(n|\omega) > 0$ and $\Pr(n|\omega)$ is described by equation (11). Given behavior $\mathbb{P}$ that satisfies Theorem 1, plug equation (11) into the last instance of $\Pr(n|\omega)$ in equation (7) after using Lemma 5 to replace $\mathbf{C}(F, \mu)$ with $\mathbf{C}(\mathbb{P}, \mu)$, and notice that cancelling like terms then produces the new objective:

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

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

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

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

$$-\lambda_M\Pr(n|\omega)\log\left(\Pr(n)^{\frac{\lambda_1}{\lambda_M}}\right) - \lambda_M\Pr(n|\omega)\log\left(\Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\right)$$

$$-\ldots-\lambda_M\Pr(n|\omega)\log\left(\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}\right) - \lambda_M\Pr(n|\omega)\log\left(e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_M}}\right)$$

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

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

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

Call (7) with $\mathbf{C}(F,\,\mu)$ replaced by $\mathbf{C}(\mathbb{P},\,\mu)$ the 'old objective' and call the objective in Lemma 2 produced immediately above the 'new objective.' What is shown above is that the old objective and the new objective have the same value if behavior $\mathbb{P}$ satisfies Theorem 1, and thus the maximal value of the new objective subject to the associated constraints is at least as large as the maximal value of the old objective subject to the associated constraints.

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

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

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

$$\Pr(n)^{\frac{\lambda_1}{\lambda_M}}\Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ldots\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}$$

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

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

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

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

$$\left.\left.+(1-\alpha)\tilde{\Pr}(n)^{\frac{\lambda_1}{\lambda_M}}\tilde{\Pr}(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}}\ldots\tilde{\Pr}(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}\right)e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}\right)$$

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

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

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

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

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

$$+\sum_{A\in\times\{\mathcal{A}_i\}_{i=1}^{M-1}}\sum_{n\in\mathcal{M}}\xi_n(A)(\Pr(n|A)-\delta)-\sum_{A\in\times\{\mathcal{A}_i\}_{i=1}^{M-1}}\gamma(A)\left(\sum_{n\in\mathcal{M}}\Pr(n|A)-1\right),$$

but the constraint of $\Pr(n|A)\geq\delta$ for arbitrarily small weakly positive δ is considered as the objective of this problem is concave and thus if $\delta>0$ then the first order conditions are both necessary and sufficient (Train, 2009) as the objective is then differentiable on the relevant set (the meaning of $\times\{\mathcal{A}_i\}_{i=1}^{M-1}$ is explained just before this proof). Assume $\mathbb{P}$ satisfies (11) and is such that $\Pr(n|\omega)>0$ for all options $n\in\mathcal{M}$ and states $\omega\in\Omega$, then the first order condition with respect to $\Pr(n|A)$ for any $A\in\times\{\mathcal{A}_i\}_{i=1}^{M-1}$ and $\tilde{\omega}\in A$ is then:

$$\left(\sum_{\omega\in\Omega}\frac{\lambda_1\mu(\cap_{i=1}^{M-1}\mathcal{A}_i(\tilde{\omega}))}{\Pr(n)}\Pr(n|\omega)\mu(\omega)\right)+\left(\sum_{\omega\in\mathcal{A}_1(\tilde{\omega})}\frac{(\lambda_2-\lambda_1)\mu(\cap_{i=1}^{M-1}\mathcal{A}_i(\tilde{\omega}))}{\Pr(n|\mathcal{A}_1(\omega))\mu(\mathcal{A}_1(\tilde{\omega}))}\Pr(n|\omega)\mu(\omega)\right)$$

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

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

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

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

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

so if:

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

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

What remains to be shown is that a solution to the problem described in Lemma 2 combined with (11) maximizes (7) subject to (8) and (9), and thus the solution to the problem described in Lemma 2 satisfies Theorem 1 when combined with (11).

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

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

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

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

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

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

What remains to be shown is that a maximizer of the new objective subject to the associated constraints, denoted $\Pr(n|A)$ for each option n and event $A \in \mathcal{F}$ (denote the collection of these $\mathbb{P}(\mathcal{F})$, and as a slight abuse of notation let $\mathcal{C}(\mathbb{P}(\mathcal{F})) \equiv \{n \in \mathcal{N} | \Pr(n) > 0\}$), satisfies Theorem 1 and so for each option n and event $A \in \times\{\mathcal{A}_i\}_{i=1}^{M-1}$ if $\Pr(n|A) > 0$ then:

$$\sum_{\omega \in A} \left(\frac{\Pr(n)^{\frac{\lambda_1}{\lambda_M}} \Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2 - \lambda_1}{\lambda_M}} \ldots \Pr(n| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega))^{\frac{\lambda_M - \lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_n(\omega)}{\lambda_M}}}{\sum_{\nu \in \mathcal{N}} \Pr(\nu)^{\frac{\lambda_1}{\lambda_M}} \Pr(\nu|\mathcal{A}_1(\omega))^{\frac{\lambda_2 - \lambda_1}{\lambda_M}} \ldots \Pr(\nu| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega))^{\frac{\lambda_M - \lambda_{M-1}}{\lambda_M}} e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_M}}} \right) \mu(\omega|A) = \Pr(n|A). \tag{16}$$

To do this, consider a maximizer of the old objective subject to the associated constraints denoted $\tilde{\Pr}(n|\omega)$ for each option n and ω (denote this collection $\tilde{\mathbb{P}}$) such that $\mathcal{C}(\tilde{\mathbb{P}}) \subseteq \mathcal{C}(\mathbb{P}(\mathcal{F}))$ (such a maximizer exists given what is shown above). The strict concavity of logarithms and the concavity of the generalized Cobb-Douglas utility function implies that the value of the term inside of the parenthesis in the new objective is the same in each state of the world with the constrained maximizers of the new objective, old objective, or a mixture of the two, inputted, and is in each state of the world equal to the denominator in (11), so for each n and ω the numerator of the fraction in equation (16) would have to be linear in mixtures of the two maximizers. Let $\lambda_0 = 0$, and for a generic option n and state ω let $\delta \equiv \tilde{\Pr}(n)^{\frac{\lambda_1}{\lambda_M}} \tilde{\Pr}(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2 - \lambda_1}{\lambda_M}} \ldots \tilde{\Pr}(n| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega))^{\frac{\lambda_M - \lambda_{M-1}}{\lambda_M}} - \Pr(n)^{\frac{\lambda_1}{\lambda_M}} \Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2 - \lambda_1}{\lambda_M}} \ldots \Pr(n| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega))^{\frac{\lambda_M - \lambda_{M-1}}{\lambda_M}}$, let $\Pr(n| \cap_{j=1}^{0} \mathcal{A}_j(\omega)) \equiv \Pr(n)$, for $i \in \{0, 1, \ldots, M-1\}$ let $\delta_i \equiv \tilde{\Pr}(n| \cap_{j=1}^{i} \mathcal{A}_j(\omega)) - \Pr(n| \cap_{j=1}^{i} \mathcal{A}_j(\omega))$,

and let $Y \equiv \Pr(n)^{\frac{\lambda_1}{\lambda_M}} \Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \ldots \Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}$. Then for all $\alpha \in (0,\,1)$:

$$(\Pr(n)+\alpha\delta_0)^{\frac{\lambda_1-\lambda_0}{\lambda_M}} (\Pr(n|\mathcal{A}_1(\omega))+\alpha\delta_1)^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \ldots (\Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))+\alpha\delta_{M-1})^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}$$

$$= \Pr(n)^{\frac{\lambda_1}{\lambda_M}} \Pr(n|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \ldots \Pr(n|\cap_{i=1}^{M-1}\mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}} + \alpha\delta,$$

and, assuming $Y > 0$, taking the derivative of both sides with respect to α:

$$\sum_{i=0}^{M-1} \frac{\delta_i(\lambda_{i+1}-\lambda_i)}{\lambda_M} \frac{Y+\alpha\delta}{\Pr(n|\cap_{j=1}^{i}\mathcal{A}_j(\omega))+\alpha\delta_i} = \delta,$$

and then taking the derivative of both sides with respect to α again:

$$\sum_{i=0}^{M-1} \frac{\delta_i(\lambda_{i+1}-\lambda_i)}{\lambda_M} \frac{\delta\Pr(n|\cap_{j=1}^{i}\mathcal{A}_j(\omega)) - Y\delta_i}{(\Pr(n|\cap_{j=1}^{i}\mathcal{A}_j(\omega))+\alpha\delta_i)^2} = 0, \tag{17}$$

and then taking the derivative of both sides with respect to α twice more:

$$\sum_{i=0}^{M-1} \frac{\delta_i(\lambda_{i+1}-\lambda_i)}{\lambda_M} \frac{\delta\Pr(n|\cap_{j=1}^{i}\mathcal{A}_j(\omega)) - Y\delta_i}{(\Pr(n|\cap_{j=1}^{i}\mathcal{A}_j(\omega))+\alpha\delta_i)^2} \frac{6}{\left(\frac{\Pr(n|\cap_{j=1}^{i}\mathcal{A}_j(\omega))}{\delta_i}+\alpha\right)^2} = 0. \tag{18}$$

Equations (17) and (18) then imply, for all $i \in \{0,\,1,\,\ldots,\,M-1\}$:

$$\frac{\delta}{Y} = \frac{\delta_i}{\Pr(n|\cap_{j=1}^{i}\mathcal{A}_j(\omega))} \text{ (proportional change is constant across events)},$$

because:

$$\delta\Pr(n|\cap_{j=1}^{i}\mathcal{A}_j(\omega)) > Y\delta_i \Rightarrow \frac{6}{\left(\frac{\Pr(n|\cap_{j=1}^{i}\mathcal{A}_j(\omega))}{\delta_i}+\alpha\right)^2} < \frac{6}{\left(\frac{Y}{\delta}+\alpha\right)^2}$$

and

$$\delta\Pr(n|\cap_{j=1}^{i}\mathcal{A}_j(\omega)) < Y\delta_i \Rightarrow \frac{6}{\left(\frac{\Pr(n|\cap_{j=1}^{i}\mathcal{A}_j(\omega))}{\delta_i}+\alpha\right)^2} > \frac{6}{\left(\frac{Y}{\delta}+\alpha\right)^2}.$$

To show there is a maximizer of the old objective subject to the associated constraints such that $\mathcal{C}(\tilde{\mathbb{P}}) = \mathcal{C}(\mathbb{P}(\mathcal{F}))$, assume the maximizer of the old objective subject to the associated constraints is instead such that $\mathcal{C}(\tilde{\mathbb{P}}) \subset \mathcal{C}(\mathbb{P}(\mathcal{F}))$, and let m denote an option such that $\Pr(m) < \tilde{\Pr}(m)$. As I move along the line from the old maximizer to the new maximizer (notice that for each event $A \in \mathcal{F}$ the sum of conditional probabilities on this line is one), denote the probability of m being

selected in each event $A \in \mathcal{F}$ by $\hat{\Pr}(m|A)$. Since:

$$\hat{\Pr}(m)^{\frac{\lambda_1}{\lambda_M}} \hat{\Pr}(m|\mathcal{A}_1(\omega))^{\frac{\lambda_2-\lambda_1}{\lambda_M}} \dots \hat{\Pr}(m| \cap_{i=1}^{M-1} \mathcal{A}_i(\omega))^{\frac{\lambda_M-\lambda_{M-1}}{\lambda_M}}$$

is changing linearly for each $A \in \mathcal{F}$, it implies that if I continue along the line from the old maximizer through the new maximizer I eventually hit the corner where $\hat{\Pr}(m|A) = 0$ for each $A \in \mathcal{F}$, and I can assume m is (one of) the first option(s) to reach its corner, and this corner also maximizes the new objective since as I moved along the line the new objective remained constant in each state of the word, and thus there is a different maximizer of the old objective (subject to the associated constraints) that does not assign a positive probability of selection to m and thus only assigns a positive probability of selection to a different strict subset of $\mathcal{C}(\mathbb{P}(\mathcal{F}))$. Since payoffs are linear and Lemma 3 indicates that $\mathbb{H}$ is strictly concave, it means that the different maximizer of the old objective results in the same distribution over posteriors as $\tilde{\mathbb{P}}$, and since the different maximizer of the old objective satisfies Theorem 1, there is another option (or a composite option that is a mixture over options) that has identical normalized payoffs to m ($e^{\frac{\mathbf{v}_m(\omega)}{\lambda_M}} = e^{\frac{\mathbf{v}_\nu(\omega)}{\lambda_M}} \ \forall \omega \in \Omega$), and I can change $\tilde{\mathbb{P}}$ so that $\tilde{\Pr}(m) = \Pr(m)$, re-distributing weight in $\tilde{\mathbb{P}}$ from m to ν while maintaining the same distribution over posteriors, and the result is still a maximizer of the old objective subject to the associated constraints (since payoffs are linear and Lemma 3 indicates that $\mathbb{H}$ is strictly concave). I can repeat this process until $\mathcal{C}(\tilde{\mathbb{P}}) = \mathcal{C}(\mathbb{P}(\mathcal{F}))$. Next, notice that for each $n \in \mathcal{C}(\mathbb{P}(\mathcal{F}))$ there is an $A \in \times\{\mathcal{A}_i\}_{i=1}^{M-1}$ such that $\Pr(n|A) > 0$, and for each $n \in \mathcal{C}(\mathbb{P}(\mathcal{F})) = \mathcal{C}(\tilde{\mathbb{P}})$ and all $A \in \times\{\mathcal{A}_i\}_{i=1}^{M-1}$ Theorem 1 implies $\tilde{\Pr}(n|A) > 0$, and thus by moving along the line from $\mathbb{P}(\mathcal{F})$ to $\tilde{\mathbb{P}}$, but starting arbitrarily close to $\mathbb{P}(\mathcal{F})$ as opposed to at $\mathbb{P}(\mathcal{F})$ so that the analogue Y defined as above with this new starting position is strictly positive for all events, and so equations (17) and (18) can be used to reach a contradiction unless for each $n \in \mathcal{C}(\mathbb{P}(\mathcal{F}))$ and all $A \in \times\{\mathcal{A}_i\}_{i=1}^{M-1}$ it is the case that $\Pr(n|A) > 0$. So, for each option $n \in \mathcal{C}(\mathbb{P}(\mathcal{F}))$ and for each event $A \in \times\{\mathcal{A}_i\}_{i=1}^{M-1}$, $\tilde{\Pr}(n|A) > 0$ and there is a constant $\theta_n > 0$ such that $\Pr(n|A) = \theta_n \tilde{\Pr}(n|A)$, and I am done. ■

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

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

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

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

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

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

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

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

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

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

Theorem 3. Behavior $\mathbb{P}$ is optimal iff for all $n \in \mathcal{C}(\mathbb{P})$ it is the case that $\Pr(n|\omega) > 0$ and $\Pr(n|\omega)$ is described by equation (11) for each state $\omega \in \Omega$, and for all $n \notin \mathcal{C}(\mathbb{P})$ it is the case that:

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

Proof. Assume $\mathbb{P}$ is such that for all $n \in \mathcal{C}(\mathbb{P})$ and $\omega \in \Omega$ it is the case that $\Pr(n|\omega) > 0$ and

$\Pr(n|\omega)$ is described by equation (11). Further, so that there remains something to be proven, assume there is at least one $n \notin \mathcal{C}(\mathbb{P})$. To figure out if the agent can do strictly better than $\mathbb{P}$, given Lemma 2, it needs to be determined if the agent could do strictly better by changing their behavior so that at least one of the options $n \notin \mathcal{C}(\mathbb{P})$ is chosen instead with a strictly positive probability.

First, revisit the problem of maximizing (7) subject to (8) and (9), except transform the problem into a problem where the solution to the transformed problem $\tilde{\mathbb{P}}$ is required to be such that $\tilde{\Pr}(n) = \epsilon$ for each $n \in \mathcal{N}\backslash\mathcal{C}(\mathbb{P})$ and $\tilde{\Pr}(m) \geq \epsilon$ for each $m \in \mathcal{C}(\mathbb{P})$, where ϵ is some arbitrarily small and strictly positive constant, and for all $n \in \mathcal{N}$ and $\omega \in \Omega$ it is imposed that $\tilde{\Pr}(n|\omega) \geq \delta$ where δ is some strictly positive constant that is arbitrarily smaller than ϵ. The objective for this transformed problem is differentiable on the set over which maximization is occurring, which is convex and closed, so a solution to the first order conditions exists, and the first order conditions are necessary (Lange, 2013). The Lagrangian for the transformed problem is:

$$\mathcal{L} = \sum_{n\in\mathcal{N}}\sum_{\omega\in\Omega} \mathbf{v}_n(\omega)\tilde{\Pr}(n|\omega)\mu(\omega) - \mathbf{C}(\tilde{\mathbb{P}}, \mu) + \sum_{n\in\mathcal{N}}\sum_{\omega\in\Omega} \xi_n(\omega)(\tilde{\Pr}(n|\omega) - \delta)\mu(\omega)$$
$$- \sum_{\omega\in\Omega} \gamma(\omega)\Big(\sum_{n\in\mathcal{N}} \tilde{\Pr}(n|\omega) - 1\Big)\mu(\omega) - \sum_{n\in\mathcal{N}\backslash\mathcal{C}(\mathbb{P})} \beta_n\Big(\Big(\sum_{\omega\in\Omega} \tilde{\Pr}(n|\omega)\mu(\omega)\Big) - \epsilon\Big)$$
$$- \sum_{m\in\mathcal{C}(\mathbb{P})} \theta_m\Big(\Big(\sum_{\omega\in\Omega} -\tilde{\Pr}(m|\omega)\mu(\omega)\Big) + \epsilon\Big),$$

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

satisfies (11) for all options n and states ω when the altered payoffs are used, and thus Lemma 2 implies that $\tilde{\mathbb{P}}$ is a solution to the problem of maximizing (7) subject to (8) and (9) when payoffs are altered in this way and thus maximizes the problem from Lemma 2 when payoffs are altered in this way.

Next, it is shown that by picking arbitrarily small ϵ, the value of the objective from Lemma 2 when behavior is $\mathbb{P}$ and the original payoffs are used, call it x, can be made arbitrarily close to the value of the objective from Lemma 2 when behavior is $\tilde{\mathbb{P}}$ and the altered payoffs are used, call it y. It is evident that $y \geq x$ for each such ϵ, so it needs to ruled out that there is $c > 0$ such that $y > x + c$ for all arbitrarily small ϵ. This is also evident as $\tilde{\mathbb{P}}$ is a solution to (7) subject to (8) and (9) when the altered payoffs are used, and if the agent chose the same signal structure as with $\tilde{\mathbb{P}}$, except whenever they get a signal that should lead them to choose one of the options n with $\tilde{\Pr}(n) = \epsilon$ they instead randomized with equal probabilities over the $m \in \mathcal{C}(\mathbb{P})$ with $\tilde{\Pr}(m) > \epsilon$, they would then get some payoff of $z \leq x$. Thus, given $c > 0$ it cannot be that $y > x + c$ for an arbitrarily small ϵ, as then some options that are being selected with unconditional probability ϵ have altered payoffs that are arbitrarily high, and thus the agent could do better by picking from a subset of these options and $\tilde{\mathbb{P}}$ is not maximal.

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

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

and

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

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

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

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

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

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

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

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

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

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

Remember that $s_n(\omega|\mathbb{P})$ is defined just before the statement of Theorem 3, and use the altered payoffs to analogously define $\tilde{s}_n(\omega|\mathbb{P})$ for each state $\omega\in\Omega$ and option $n\in\mathcal{N}$ as follows:

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

Since $\tilde{\mathbb{P}}$ satisfies (11) for all options n and states ω when the altered payoffs are used, for each $n\in\mathcal{N}\backslash\mathcal{C}(\mathbb{P})$ and $\tilde{\omega}\in\Omega$:

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

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

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

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

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

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

$$\dots$$

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

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

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

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

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

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

This means:

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

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

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

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