Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model[‡]

Filip Matějka[*] and Alisdair McKay[**]

First draft: February 2011
This draft: January 2013

Abstract

Individuals must often choose among discrete alternatives with imperfect information about their values. Before choosing, they have an opportunity to study the options, but doing so is costly. This creates new choices such as the number of and types of questions to ask. We model these situations using the rational inattention approach to information frictions. We find that the decision maker's optimal strategy results in choosing probabilistically in line with a generalized multinomial logit model, which depends both on options' true values as well as on prior beliefs.

Keywords: discrete choice, information, rational inattention, multinomial logit.

[†] We thank Levent Celik, Satyajit Chatterjee, Faruk Gul, Christian Hellwig, Stepan Jurajda, Bart Lipman, Tony Marley, Andreas Ortmann, Juan Ortner, Christopher Sims, Leonidas Spiliopoulos, Jakub Steiner, Jorgen Weibull, and Michael Woodford for helpful discussions as well as seminar participants at BU, CERGE, CREI, Harvard Institute for Quantitative Social Science, Princeton University, SED 2011, and the Toulouse School of Economics.

[‡]This research was funded by GAČR P402/11/P236, GDN project RRC 11+001 and by European Research Council under the European Community's Seventh Framework Programme FP7/2007-2013 grant agreement N. 263790

[*]CERGE-EI, A joint workplace of the Center for Economic Research and Graduate Education, Charles University, and the Economics Institute of the Academy of Sciences of the Czech Republic. filip.matejka@cerge-ei.cz

[**]Boston University. amckay@bu.edu

1 Introduction

Economists and psychologists have long known that scarce attention plays an important role in decision making (Simon, 1959; Kahneman, 1973). In this paper we study the discrete choice behavior of an agent who must allocate his limited attention to the available information about the choice situation.

It is not uncommon for one to be faced with a choice among discrete alternatives with imperfect information about the value of each alternative. Before making a choice, one may have an opportunity to study the options, however, in most cases it is too costly to investigate the options to the point where their values are known with certainty. As a result, some uncertainty about the values remains when one chooses among the options even if complete information were available in principle. Because of this uncertainty, the option that is ultimately chosen may not be the one that provides the highest utility to the decision maker (DM). Moreover, the noise in the decision process may lead identical individuals to make different choices. In this manner, imperfect information naturally leads choices to contain errors and be probabilistic as opposed to deterministic.

In this context, the DM faces choices of how much to study the options and what to investigate when doing so. That is, the DM must choose how to allocate his attention. For example, a firm might choose how long to spend interviewing candidates for a job and choose what to ask them during the interview. After completing the interview, the firm faces a discrete choice among the candidates.

We explore the optimal "information processing" behavior of a DM for whom acquiring information is costly and characterize the resulting choice behavior in this discrete choice context. As choices are probabilistic, our characterization involves describing the probability with which the DM selects a particular option in a particular choice situation. Specifically, we model the cost of acquiring and processing information using the rational inattention framework introduced by Sims (1998, 2003).

The major appeal of the rational inattention approach to information frictions is that it does not impose any particular assumptions on what agents learn or how they go about

learning it. Instead, the rational inattention approach derives the information structure from the utility-maximizing behavior of the agents for whom information is costly to acquire. As a result, rationally inattentive agents process information that they find useful and ignore information that is not worth the effort of acquiring and processing.

Our main finding is that the DM's optimal information processing strategy results in probabilistic choices that follow a logit model that reflects both the options' true values as well as the DM's prior beliefs. In a choice among N options with any form of prior beliefs, our modified logit formula takes the form

$$\frac{e^{(v_i+\alpha_i)/\lambda}}{\sum_{j=1}^{N} e^{(v_j+\alpha_j)/\lambda}}, \tag{1}$$

where v_i is the value of option i and λ is a parameter that scales the cost of information. The DM's prior knowledge and information processing strategy are incorporated into the choice probabilities through the weights, α_i, attached to each position in the choice set. These weights shift the choice probabilities towards those options that appeared to be good candidates *a priori* and they are completely independent of the actual values of the options. As the cost of information rises, the DM's choice becomes less sensitive to the actual values of the options and more sensitive to his prior beliefs.

When the *a priori* beliefs do not influence the choice our work provides a new foundation for the multinomial logit model . We show that whenever the options are exchangeable in the DM's prior, α_i is constant across i so equation (1) simplifies to the standard multinomial logit formula. Cases where the options are homogenous *a priori* arise naturally whenever the DM lacks specific knowledge that allows him to distinguish between the options before entering the choice situation.

The multinomial logit model is perhaps the most commonly used model of discrete choice and it has two canonical foundations.[1] According to the random utility derivation, the DM evaluates the options with some noise. If the noise in the evaluation is additively

[1]The logit model was first proposed for binary choices by Bradley and Terry (1952) and the multinomial logit was introduced by Luce (1959). Anderson et al. (1992), McFadden (2001), and Train (2009) present surveys of discrete choice theory and the multinomial logit model.

separable and independently distributed according to the extreme value distribution, then the multinomial logit model emerges.[2] To date, there is not a clear economic or psychological justification of why these disturbances should be extreme value distributed. The model has a second canonical foundation, namely, Luce's (1959) derivation from the independence of irrelevant alternatives (IIA) axiom, which states that ratios of choice probabilities are independent of the choice set.

We believe our findings are important for two reasons. First, while the logit model is sometimes used in situations in which information frictions are thought to be an important part of the choice environment, there has not previously been a fully-specified model of those information frictions that justifies the use of the multinomial logit. We fill that gap and demonstrate that the fully-specified model is not subject to some of the criticisms of the logit model. Second, most existing work with rational inattention has focussed on situations where the DM chooses from a continuous choice set. In this context, the model remains tractable if one assumes the agent is acquiring information about a normally-distributed quantity and the objective function is quadratic, as under these assumptions the DM chooses normally distributed signals. Beyond this situation, however, the continuous-choice rational inattention model must be solved numerically and even numerical solutions can be difficult to obtain. In contrast, we show here that the discrete-choice version of the rational inattention model is extremely tractable. Our results allow the rational inattention framework to be easily applied by building on a large body applied theoretical work that exploits the tractability of the multinomial logit.[3]

[2]Luce and Suppes (1965, p. 338) attribute this result to Holman and Marley (unpublished). See McFadden (1974) and Yellott (1977) for the proof that a random utility model generates the logit model only if the noise terms are extreme value distributed.

[3]The multinomial logit model is commonly used in the industrial organization and international trade literatures as a model of consumer demand, in political economy models of voting, and in experimental economics to capture an element of bounded rationality in subject behavior. See Anderson et al. (1992) for a survey of its use in industrial organization. The logit demand structure was introduced to international trade by Goldberg (1995) and Verboven (1996). The logit model was incorporated into probabilistic voting models by Lindbeck and Weibull (1987). Work following McKelvey and Palfrey (1995) uses logit response functions to capture randomness in the responses of experimental subjects playing a game. Matějka and McKay (2012) is an example of how the results of this paper can be integrated with existing results based on the multinomial logit model to study the optimal price-setting behavior of firms facing rationally inattentive consumers.

In our model, a choice set is a vector of true values and a prior distribution of beliefs about those values. Accordingly, there are two ways that a choice situation can change. First, the realized vector of true values can change. Second, the prior beliefs themselves can change. The generalized logit model describes how choice probabilities change when the realization of values changes. In this case, the position weights, α_i, are fixed across realizations.

The rationally inattentive agent's choices are context dependent where context is defined by prior beliefs that represent the immediately apparent characteristics of the options. As the DM's prior changes, his choice behavior changes both due to standard Bayesian updating and through endogenous changes in his information processing strategy. Changes in the information processing strategy can lead to results that are appealingly intuitive and to results that seem surprising. We show:

- The rationally inattentive agent ignores duplicate options.
- Adding an option to the choice set can increase the likelihood that an existing option is selected—an outcome that cannot occur in any random utility model. The addition of an option changes the choice set and therefore the DM's prior beliefs, which can induce him to process information sufficiently differently so as to increase the probability that he selects an existing option.
- Option i is more likely to be selected if the DM's beliefs about it improve in the sense of first-order-stochastic dominance. As we demonstrate, this monotonicity does not always hold under Bayesian updating with an exogenous information structure.

That the rationally inattentive agent ignores duplicate options stands in contrast to the standard logit model. Debreu (1960) criticized IIA and the logit model for having counter-intuitive implications for the treatment of duplicate options because the standard model treats each option as distinct and only allows them to be differentiated in one dimension (e.g. their values). Thus, there is no sense of similarity between options and adding a duplicate of one option will increase the probability that the option (or its duplicate) is selected. In our setting, however, we can introduce the concept of similarity through prior

knowledge. If two options are duplicates, then the DM knows *a priori* that they will have the same value even if this common value is unknown. We show that when a duplicate option is added to the choice set, the rationally inattentive agent will choose to ignore it. Therefore the model does not display the counter-intuitive behavior that Debreu (1960) criticized.

The paper is organized as follows: In the remainder of this section we review related work. Section 2 presents the choice setting and discusses the assumptions underlying the rational inattention approach to information frictions. Section 3 studies the DM's optimal strategy. Section 4 presents a characterization result that connects the optimal behavior of the rationally inattentive agent to weaker versions of Luce's IIA axiom. Section 5 demonstrates how the DM's prior knowledge influences his choice behavior. Finally, Section 6 concludes.

Related Literature Our work relates to the literature on rational inattention. Most existing work with rational inattention has focussed on situations where the DM chooses from a continuous choice set. Rational inattention has mostly been applied in macroeconomic contexts such as consumption-savings problems (Sims, 2006; Maćkowiak and Wiederholt, 2010) and price setting (Mackowiak and Wiederholt, 2009; Matějka, 2010a).[4] A few papers, however, consider applications with binary choice problems. Woodford (2009) was the first to do so in a study of a binary choice of whether to adjust a price, while Yang (2011) investigates a global game setting with the choice of whether to invest or not. Moreover, Matějka and Sims (2010) and Matějka (2010a) provide a connection between the continuous and discrete problems by showing that rationally inattentive agents can voluntarily constrain themselves to a discrete choice set even when the initial set of available options is continuous. We extend the existing literature by establishing a connection between rational inattention and the multinomial logit model and characterize the implications of rational inattention for discrete choice behavior.[5]

Closely related to our work is the work of Caplin and Martin (2011) who derive testable

[4]Other applications are Luo (2008); Luo and Young (2009); Tutino (2009); Van Nieuwerburgh and Veldkamp (2010); Mondria (2010); Matějka (2010b); Paciello and Wiederholt (2011).

[5]In an independent paper that is as of yet unfinished, Woodford (2008) notices the connection to the logit model in the context of a binary choice problem, but does not explore the connection in further detail.

implications from a model of choice under imperfect perception with rational expectations. Weibull et al. (2007) and Natenzon (2010) study discrete choice models with imperfect information in which the DM receives signals with an exogenously given structure. Masatlioglu et al. (2012) and Manzini and Mariotti (2012) model imperfect attention using a consideration set approach. Under this approach, choices occur in two stages: first, some of the available options are selected into a consideration set and then the utility maximizing option is chosen from the consideration set. Under this approach, the DM may overlook some of the available options while in our setting the DM is aware of all options, but may not be aware of their exact characteristics.

The rational inattention approach to information frictions uses information theoretic concepts to measure the amount of information processed by the DM and there is a mathematical connection between the entropy function, which is at the heart of information theory, and the multinomial logit. This connection has appeared in the context of statistical estimation (Anas, 1983) and in the context of an agent stabilizing a trembling hand (Stahl, 1990; Mattsson and Weibull, 2002). Here we are considering the decision problem of an agent who must acquire information about the values of the alternatives. In this context, the entropy function arises naturally.[6]

2 The model

In this section, we first describe the agent's decision problem, then we discuss the modeling choices of how the agent processes information. The DM is presented with a group of N options, from which he must choose one. The values of these options potentially differ and the agent wishes to select the option with the highest value. Therefore, let v_i denote the value of option $i \in \{1, \cdots, N\}$.

[6]In mathematical terms, our work is close to that of Shannon (1959) who derives the multinomial logit formula in an engineering application that is the dual to our problem in the *a priori* homogeneous case. Shannon's question is how quickly a message can be communicated through a limited-capacity channel, such as a telegraph wire, without distorting the message beyond a certain degree on average. We thank Michael Woodford for pointing us to this connection to Shannon's work.

The DM is rationally inattentive in the style of Sims (2003, 2006). He possesses some prior knowledge of the available options; this prior knowledge is described by a joint distribution $G(\mathbf{v})$, where $\mathbf{v} = (v_1, \cdots, v_N)$ is the vector of values of the N options. To refine his knowledge, he processes information about the options. One interpretation is that he asks questions about the values and each question comes at a cost. Finally, the DM chooses the option with the highest expected value.

At the heart of the model is a formulation of the cost of asking questions, submitting queries to a database, or otherwise gathering and processing information. We assume that all information about the N options is available to the DM, but processing the information is costly. If the DM could process information costlessly, he would select the option with the highest value with probability one. With costly information processing, the DM must choose the following:

(i) how much information to process, i.e. how much attention to pay,

(ii) what pieces of information to process, i.e. what to pay attention to,

(iii) what option to select conditional on the acquired posterior belief.

The novelty of rational inattention is that the DM is allowed to choose the optimal mechanism for processing information considering the cost of acquiring information. The first point to establish is that it is not necessary to model questions or signals explicitly: it is enough to consider the relationship between fundamentals, $\mathbf{v}$, and actions, the option i that is ultimately selected. In general, what knowledge outcomes can be generated by a specific mechanism of processing information, e.g. a series of questions, is fully described by a joint probability distribution of fundamentals and posterior beliefs about them (Blackwell, 1953). It follows that any information processing mechanism will induce a joint distribution between fundamentals and actions because each posterior belief will determine a particular choice from among the options. Therefore, it is this joint distribution between $\mathbf{v}$ and i that is the DM's strategy under the rational inattention approach. It is not necessary to model questions or signals, but nevertheless, the DM's strategy describes both the choice of how to

process information as well as the choice among the options conditional on the posteriors.[7]

Mathematically, we can describe the DM's strategy using the conditional probability $\mathcal{P}(i|\mathbf{v}) \in [0,1]$ for all i and $\mathbf{v}$. This is the probability of selecting option i when the realized values are $\mathbf{v}$. Let us denote this probability as $\mathcal{P}_i(\mathbf{v})$. Since the DM's prior on the values of all options is given by $G(\mathbf{v})$, the conditional distribution can be used to generate the full joint distribution.

The DM's strategy is thus a solution to the following problem:

$$\max_{\{\mathcal{P}_i(\mathbf{v})\}_{i=1}^N} \left(\sum_{i=1}^N \int_{\mathbf{v}} v_i \mathcal{P}_i(\mathbf{v})\, G(d\mathbf{v}) - \textit{ cost of information processing} \right), \tag{2}$$

subject to

$$\forall i: \qquad \mathcal{P}_i(\mathbf{v}) \geq 0 \qquad a.s., \tag{3}$$

$$\sum_{i=1}^N \mathcal{P}_i(\mathbf{v}) = 1 \qquad a.s. \tag{4}$$

The first term in (2) is the expected value of the selected option. What remains is to specify the cost of information.

In macroeconomics, the literature following Sims uses information theory to measure the "amount" of information that has been used by an agent in making a decision. Under this approach, the cost of information is $\lambda\kappa$, where λ is the unit cost of information and κ the amount of information that the DM processes, which is measured by the expected reduction in the entropy of the distribution representing knowledge of $\mathbf{v}$. The amount of information processed, κ, is a function of the DM's strategy of how to process information, while λ is a given parameter.

There are no constraints on the DM's choice of strategy apart from requiring the proba-

[7]Signals do not appear in the final formulation of the problem since each posterior belief is associated with a single i that is selected given that belief. It would not be optimal to select an information structure that would generate two different forms of posterior knowledge leading to the same i, i.e. it would not be optimal to acquire information that is not ultimately used in the choice. This approach is used in Sims (2006) and also in Matějka (2010a), where it is discussed at more length.

bilities to be well defined. Rational inattention is a theory describing how the DM processes the information that is available. If the DM were willing to pay the cost, he could in principle acquire perfect information about the available options. However, suppose there were some uncertainty that the DM could not resolve through processing information. The model can accommodate such a situation if v_i is taken to be the expected value of the option where the expectation is taken over the uncertainty that cannot be resolved.

Entropy is a measure of the uncertainty associated with a random variable. In our case, the random variable is the vector $\mathbf{v}$ and acquiring better knowledge about the values, i.e. narrowing down the belief, is associated with a decrease in the entropy.

Our results depend crucially on the choice to model the cost of information using the reduction in entropy, but this is not an ad hoc modeling choice. Entropy provides the exact measure of the cost for a rationally inattentive agent who acquires information through a limited-capacity channel. Using a limited-capacity channel means the DM receives a sequence of symbols (e.g. a list of ones and zeros). The symbols can mean virtually anything: they can represent answers to questions the agent asks, pieces of text or digits he reads, etc. The more information the DM processes, i.e. the more symbols he receives, the tighter his posterior beliefs can be. The coding theorem of information theory (Shannon, 1948; Cover and Thomas, 2006) states that any joint distribution of source variables, i.e. fundamentals, and posterior beliefs is achievable by an information channel if and only if the expected decrease in the entropy of the knowledge is less than the amount of information processed, which is proportional to the number of symbols received.[8] Choosing how to process information is then equivalent to choosing how many questions to ask and what to ask about.[9]

The assumption that the cost of information processing is $\lambda\kappa$ can be interpreted as saying the cost is proportional to the expected number of questions asked. One could think of the coefficient λ as a shadow cost of allocating attention to this decision problem out of a larger

[8]The coding theorem relies on information being encoded optimally and this encoding can introduce delays. We, like other papers on rational inattention, do not consider the effects of these delays. If delays were penalized in some way, the DM's behavior could deviate from the results presented here.

[9]The amount of information per symbol depends on the physical properties of the channel. For instance, if the DM processes information by asking questions with yes or no answers, then the information per question is one bit.

budget of attention that the agent is allocating to many issues. By modeling the cost of information in terms of the number of questions that the DM asks or the number of symbols that he receives, we are modeling a world in which receiving answers to each question with the same number of possible answers is equally costly. The assumption that the cost is linear in the reduction in entropy could be modified without fundamentally changing the results. If the cost function were convex in κ, the equilibrium marginal cost of information would take the place of λ in our key result equation (9).

Mathematically, the entropy of a random variable X with a pdf $p(x)$ with respect to a probability measure σ is defined as:

$$H(X) \equiv -\int p(x) \log p(x)\, \sigma(dx). \tag{5}$$

In the DM's problem, the expected reduction in the entropy of $\mathbf{v}$ is the difference between the prior entropy of $\mathbf{v}$ and the expectation of the posterior entropy of $\mathbf{v}$ conditional on the chosen option i.[10] This quantity is also called the mutual information between $\mathbf{v}$ and i. For our purposes, it is convenient to use the symmetry of mutual information and express the amount of information processed as the expected reduction in the entropy of i conditional on $\mathbf{v}$: [11]

$$\begin{aligned}\kappa(\mathcal{P}, G) &= H(\mathbf{v}) - E_i\left[H(\mathbf{v}|i)\right] = H(i) - E_{\mathbf{v}}\left[H(i|\mathbf{v})\right] \\ &= -\sum_{i=1}^{N} \mathcal{P}_i^0 \log \mathcal{P}_i^0 + \int_{\mathbf{v}} \left(\sum_{i=1}^{N} \mathcal{P}_i(\mathbf{v}) \log \mathcal{P}_i(\mathbf{v}) \right) G(d\mathbf{v}),\end{aligned} \tag{6}$$

where $\mathcal{P} = \{\mathcal{P}_i(\mathbf{v})\}_{i=1}^N$ is the collection of conditional probabilities, and $\mathcal{P}_i^0$ is the unconditional probability of choosing option i defined as

$$\mathcal{P}_i^0 = \int_{\mathbf{v}} \mathcal{P}_i(\mathbf{v}) G(d\mathbf{v}). \tag{7}$$

[10] The pdf of the posterior with respect to the base measure $G(d\mathbf{v})$ is $\mathcal{P}_i(\mathbf{v}) / \int_{\mathbf{v}} \mathcal{P}_i(\mathbf{v}) G(d\mathbf{v})$.

[11] The mutual information between random variables X and Y is $H(X) - E_Y[H(X|Y)]$, which also equals $H[Y] - E_X[H[Y|X]]$, see Cover and Thomas (2006).

We can now state the DM's optimization problem.

Definition 1. *Let $G(\mathbf{v})$ be the DM's prior on the values of a finite number of options and let $\lambda \geq 0$ be the unit cost of information. The discrete choice strategy of the rationally inattentive DM is the collection of conditional probabilities $\mathcal{P} = \{\mathcal{P}_i(\mathbf{v})\}_{i=1}^N$ that solves the following optimization problem.*

$$\max_{\mathcal{P}=\{\mathcal{P}_i(\mathbf{v})\}_{i=1}^N} \sum_{i=1}^N \int_{\mathbf{v}} v_i \mathcal{P}_i(\mathbf{v})\, G(d\mathbf{v}) - \lambda\kappa(\mathcal{P}, G), \tag{8}$$

subject to (3) *and* (4)*, and where $\kappa(\mathcal{P}, G)$ denotes the right hand side of* (6)*.*

To summarize, the model allows the DM to investigate the values of the options at a cost that can be interpreted as proportional to the number of questions that he asks about their values. The DM is able to select both the number and content of these questions.

Finally, one might ask what types of situations is this model suited to? It is clear that our DM is solving a very sophisticated optimization problem and therefore the model is not a model of the random mistakes that a careless agent might make. Instead, the model captures the fact that it in many choice situations it is prohibitively costly to resolve all the uncertainty about the options. The key limitation is not computational, but rather stems from the difficulty of acquiring complete information about the options. In the case of a job interview, it is costly to divert the attention of the manager away from the operations of the firm to conduct an interview. Therefore, it is reasonable to think that the interview would focus on the features of the job candidates that are most important to the firm while less important features might go unexplored. While some of our assumptions, such as optimizing behavior, are stark, the model provides a natural benchmark for attention allocation and yields a convenient analytical form that is able to generate a rich set of behaviors.

3 Solving the model

3.1 Solving for choice probabilities

We begin our analysis of the model with a general analytical expression for the probability that the DM chooses a particular option conditional on the realized values of all the options.

Theorem 1. *If $\lambda > 0$, then the DM forms his strategy such that the conditional choice probabilities satisfy:*

$$\mathcal{P}_i(\mathbf{v}) = \frac{\mathcal{P}_i^0 e^{v_i/\lambda}}{\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda}}, \qquad a.s. \tag{9}$$

If $\lambda = 0$, then the DM selects the option(s) with the highest value with probability one.

Proof. See Appendix A. □

We can understand several properties of the DM's behavior from equation (9). The unconditional probabilities, $\mathcal{P}_i^0$, are by definition independent of a specific realization of the values $\mathbf{v}$. They are the marginal probabilities of selecting each option before the agent starts processing any information and they only depend on the prior knowledge $G(\mathbf{v})$ and the cost of information λ.

When the unconditional probabilities are uniform, $\mathcal{P}_i^0 = 1/N$ for all i, (9) becomes the usual multinomial logit formula. As we discuss in Section 3.2, this happens when G is invariant to permutations of its arguments. In other cases, the conditional choice probabilities are not driven just by $\mathbf{v}$, as in the logit case, but also by the unconditional probabilities of selecting each option, $\{\mathcal{P}_i^0\}_{i=1}^N$. These unconditional probabilities are monotonic transformations of the position weights that we referred to in the introduction with $\exp(\alpha_i/\lambda) = \mathcal{P}_i^0$. Using this transformation, equation (9) can be rewritten as equation (1) and the choice probabilities can be interpreted as a multinomial logit in which the value of option i is shifted by the term α_i. When an option seems very attractive *a priori*, then it has a relatively high probability of being selected even if its true value is low. As the cost of information, λ, rises, the less the DM finds out about the realization of $\mathbf{v}$ and the more he decides based on prior

knowledge of the options. In what follows, we find it more convenient to work directly with the unconditional probabilities, $\mathcal{P}_i^0$, rather than the transformed version, α_i.

The parameter λ converts bits of information to utils. Therefore, if one scales the values of all of the options by a constant c, while keeping the information cost, λ, fixed, the problem is equivalent to the one with the original values and the information cost scaled by $1/c$. By scaling up the values, one is scaling up the differences between the values and therefore raising the stakes for the DM. The DM chooses to process more information because more is at stake and thus is more likely to select the option that provides the highest utility. The DM behaves just as he would if the cost of information had fallen.

Equation (9) does not give a fully explicit expression for the choice probabilities because it depends on the $\mathcal{P}_i^0$ terms, which are themselves part of the DM's strategy although they are independent of the realized values, $\mathbf{v}$. We can substitute equation (9) into the objective function to arrive at the following formulation of the optimization problem that we will use to solve for the unconditional choice probabilities.

Lemma 1. ***Alternative formulation:*** *The agent's optimization problem in Definition 1 can be equivalently formulated as maximization over the unconditional probabilities,* $\{\mathcal{P}_i^0\}_{i=1}^N$,

$$\max_{\{\mathcal{P}_i^0\}_{i=1}^N} \int_{\mathbf{v}} \lambda \log \left(\sum_{i=1}^{N} \mathcal{P}_i^0 e^{v_i/\lambda} \right) G(d\mathbf{v}). \tag{10}$$

subject to

$$\forall i : \qquad \mathcal{P}_i^0 \geq 0, \tag{11}$$

$$\sum_i \mathcal{P}_i^0 = 1, \tag{12}$$

where the conditional probabilities $\mathcal{P}_i(\mathbf{v})$ *are given by* (9).

Proof. See Appendix A. □

This novel formulation is useful for two reasons. First, it allows for clearer insights into the properties of choice probabilities than the original problem and we use it extensively

in the proofs in Section 5. Second, it greatly reduces the complexity of the optimization problem and allows for more efficient computations. Rational inattention problems with continuous choice variables can also be formulated this way.[12] The first-order conditions of this problem give us

Corollary 1. ***Normalization condition:*** *For all i such that $\mathcal{P}_i^0 > 0$, the solution satisfies*

$$\int_{\mathbf{v}} \frac{e^{v_i/\lambda}}{\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda}} G(d\mathbf{v}) = 1. \tag{13}$$

Proof. See Appendix A. □

We call this the normalization condition because if one multiplies both sides of equation (13) by $\mathcal{P}_i^0$, the result ensures that the conditional choice probabilities in equation (9) integrate to the unconditional choice probability as in (7). The analysis in the next subsection and in section 5 is based on finding solutions to equations (9) and (13).

Appendix A establishes that a solution to the DM's maximization problem exists, however, the solution may not be unique. Cases with multiple solutions require a special structure for the uncertainty in which the values co-move in some very rigid way. For instance, when values of two options are equal in all states of the worlds, then the DM can relocate the choice probabilities between the two options arbitrarily and realize the same $E[U]$. Perhaps an illustrative interpretation of non-unique solutions is that when there are multiple solutions there always exists at least one option that can be eliminated from the choice set without reducing the expected utility that the for the DM can achieve. These eliminations can be repeated until the solution is unique. Appendix A provides the exact conditions that are necessary and sufficient for uniqueness.

[12] For a general problem under rational inattention, such as in Sims (2006), the alternative formulation takes the same form with v_i replaced by $U(X, Y)$, with the sum over i replaced an by integral over Y and $\mathbf{v}$ replaced by X, where X is the unknown and Y the action.

3.2 Multinomial logit

We now present conditions under which the behavior of the rationally inattentive agent follows the multinomial logit model. This connection to the logit model holds across different realizations of $\mathbf{v}$.

Let us assume that all the options are exchangeable in the prior G, i.e. the prior is invariant to all permutations of the entries of $\mathbf{v}$. We call such options *a priori homogeneous*.

Problem 1. *The DM chooses $i \in \{1, \cdots, N\}$, where the options are* a priori *homogeneous and take different values with positive probability.*

The options will be *a priori* homogeneous whenever the DM does not distinguish between them before he starts processing information so the position in the choice set does not provide any information. We view this as a plausible benchmark case as it will arise anytime the DM lacks specific knowledge of the options *a priori*.

Proposition 1. ***Logit:*** *In Problem 1, the probability of choosing option i as a function of the realized values of all of options is:*

$$\mathcal{P}_i(\mathbf{v}) = \frac{e^{v_i/\lambda}}{\sum_{j=1}^{N} e^{v_j/\lambda}}. \tag{14}$$

Proof. See Appendix A. □

This is the multinomial logit formula written in terms of exponentials as it is most often used in practice. We show that the *a priori* homogeneity of the options implies that the unconditional probabilities are uniform so that (9) then takes the form of the logit as the only thing that distinguishes options is the actual values. The assumption on the difference of values is needed so that the DM faces a meaningful solution and for uniqueness.

Let us emphasize that here $\mathcal{P}_i(\mathbf{v})$ does not depend on the prior G. As long as the options are *a priori* homogeneous, the resulting choice probabilities take the form of (14). This feature is particularly useful as it makes applications of the rational inattention framework very simple in this case. This result follows from the endogenous information structure in

the model. The optimal choice of information fixes the nature of the optimal posteriors and the DM selects what information to acquire so as to arrive at posteriors of that form. In contrast, with an exogenous information structure, changes in the prior lead directly to changes in the form of the posterior, with corresponding effects on the choice probabilities.

4 Characterization and IIA

In this section we show that the model provides a testable theory of choice. Thus far, we have been working with objects that are hard to observe or in principle unobservable. Now we characterize the choice behavior implied by the model, which we do for a given yet unobserved prior.

We use the term "object" to refer to the outcome that the DM has preferences over. The DM has prior beliefs over what object—and therefore what value—he will find in each position of the choice set. Let $\mathbf{x}$ be the choice set vector, which consists of an ordered set of objects. If the choice is indeed driven by rational inattention, then the key implication of Theorem 1 is that while the agent processes information about the unknown values, the choice probabilities are as if the DM attaches some weight, $\mathcal{P}_i^0$, to each position in the choice set and some weight $v_i/\lambda = v(x_i)$ to each object and then chooses among the position-object pairs probabilistically according to equation (9).

Notice that in our setting the domain of choice is larger than that considered in typical models of discrete choice as it is not just the objects that are offered to the agent, but object-position pairs, which is to say objects in the context of the agent's prior beliefs.

We now present necessary and sufficient conditions on choice probabilities that allow us to write them as functions of position weights and object weights as in Theorem 1. While the prior is fixed, we consider varying the actual realization of objects that appear in each position.

Slightly abusing notation, let $\mathcal{P}_i(\mathbf{x})$ be the probability of selecting the object in position i when the realization of objects is $\mathbf{x}$. Then we have the following axioms:

Axiom 1. Independence from irrelevant alternatives (objects)*: if* $\mathcal{P}_j(\mathbf{x}) > 0$*, then*

$$\frac{\mathcal{P}_i(\mathbf{x})}{\mathcal{P}_j(\mathbf{x})} = \frac{\mathcal{P}_i(\mathbf{y})}{\mathcal{P}_j(\mathbf{y})} \qquad \forall \mathbf{x}, \mathbf{y}, i, j \ \textit{s.t.} \ x_i = y_i \ \textit{and} \ x_j = y_j. \tag{15}$$

Axiom 2. Independence from irrelevant alternatives (positions)*: if* $\mathcal{P}_j(\mathbf{x}) > 0$*, then*

$$\frac{\mathcal{P}_i(\mathbf{x})}{\mathcal{P}_j(\mathbf{x})} = \frac{\mathcal{P}_i(\mathbf{y})}{\mathcal{P}_j(\mathbf{y})} \qquad \forall \mathbf{x}, \mathbf{y}, i, j \ \textit{s.t.} \ x_i = x_j \ \textit{and} \ y_i = y_j. \tag{16}$$

Our axioms are similar to the standard IIA axiom, which implies proportional scaling of choice probabilities: as the choice set changes, the probability of two given object scales proportionally leaving the ratio intact. Our axioms weaken the standard IIA axiom to only require proportional scaling in some cases. Axiom 1 states that proportional scaling holds if the the two objects remain in the same positions, but not necessarily otherwise.

Axiom 2 is a statement about proportional scaling of positions. Consider duplicate objects, which we can define as two objects that, when inserted into the same choice set lead to the same choice probabilities.[13] Standard IIA implies the choice probabilities of duplicate objects must be equal. Our axiom, however, does not require that their choice probabilities are equal, but that it is a constant ratio when the duplicate options remain in the same positions. Moreover, the axiom says this ratio is a function of the positions and not the nature of the objects except that they are duplicates. This seems reasonable as it is only the positions that distinguish these two object-position pairs. It is restrictive, however, as it limits the way objects and positions can interact in the choice probabilities and position weights cannot be functions of the objects. This is the key to the separation between object weights and position weights in Theorem 1.

Luce viewed the the proportional scaling implied by the standard IIA axiom as a natural benchmark. Following Luce's work, Debreu criticized IIA for generating counter-intuitive implications. We show in the next section that this criticism does not apply to the rational inattention model. One final note about the axioms, they imply that if $\mathcal{P}_j(\mathbf{x}) > 0$, then for

[13]This is the definition of duplicate objects used by Gul et al. (2012). In terms of rational inattention model, duplicate objects yield the same value to the DM.

a vector $\mathbf{y}$ that satisfies the conditions in the axioms we have $\mathcal{P}_j(\mathbf{y}) > 0$ as otherwise the ratios on the right-hand side are not well-defined.

Towards a representation result, we first establish the following:

Lemma 2. *If Axioms 1 and 2 hold for a choice among $N \geq 3$ options, then each position is either never selected for any $\mathbf{x}$ or it is selected with positive probability for all $\mathbf{x}$.*

Proof. See Appendix B. □

We will refer to positions that are always selected with some probability as "positive positions" and refer to those that are never selected as "zero positions." The main result of this section is:

Proposition 2. *Let there exist at least three positive positions. Then the choice probabilities satisfy Axioms 1 and 2 if and only if there exist non-negative constants $\{\mathcal{P}_i^0\}_{i=1}^N$ such that $\sum_{i=1}^N \mathcal{P}_i^0 = 1$ and also a function $v(x_i)$ from a space of objects to $\mathbb{R}$ such that for any i and any vector of objects $\mathbf{x}$, the probability of selecting the object in position i, is*

$$\mathcal{P}_i(\mathbf{x}) = \frac{\mathcal{P}_i^0 e^{v(x_i)}}{\sum_j \mathcal{P}_j^0 e^{v(x_j)}}. \tag{17}$$

Proof. See Appendix B. □

Some intuition for the result can be gained by considering the extreme case of a realization in which the same object appears in all positions. In this case, the choice of object-position pairs is reduced to a choice of positions because all objects are the same. Therefore, we can think of the positions as the choice set in a standard discrete choice setting and then the logit model follows from the fact that the choice probabilities satisfy Axiom 2, which is an analog of the standard IIA axiom, and thus it implies logit in positions, too.

The requirement that there are three positive positions is a technical convenience for the proof that the axioms are sufficient. The characterization of choice probabilities could be extended to the case of two positive positions by taking the limit as the probability of a third position goes to zero.

4.1 Relationship between entropy and IIA

Axioms 1 and 2 are useful for understanding the connection between rational inattention and the multinomial logit model. An important property of both models is that the choice procedure could equivalently be formulated as a two-stage process in which the DM first selects between sub-groups of options and then in a second stage selects from within the chosen sub-group. The standard IIA axiom implies that the choice probabilities are unaffected by this two-stage process and, indeed, the axiom is sometimes stated in this way. Under rational inattention, the DM can achieve the same final joint distribution between $\mathbf{v}$ and i in the two-stage procedure because he has complete flexibility in choosing his strategy in each of the two stages. The key is that the information cost he would incur in the two-stage procedure is the exact same as he would incur in the direct procedure in our model. This invariance to intermediate stages is in fact one of Shannon's (1948) axioms that characterize entropy as a measure of information.

Using the logic of a two-stage choice procedure, it becomes clear why the rationally inattentive agent's behavior satisfies axioms 1 and 2. Suppose that we required the agent to select between $\{1,2\}$ and $\{3,\ldots,N\}$ in the first stage. Whenever the DM selects $\{1,2\}$, he will enter the second stage with the same prior beliefs about the values of options 1 and 2. As such, the second stage will proceed independently of the values of options $3,\cdots,N$ although their values can influence the probability of $\{1,2\}$ in the first stage.

5 The role of prior beliefs

In this section, we present comparative static results that demonstrate how the choice probabilities depend on the DM's prior beliefs about the options. Mathematically, we show how changes in the prior distribution $G(\mathbf{v})$ affect the unconditional probabilities or in other words the position weights.

The solutions exhibit lots of structure due to the endogeneity of the DM's posterior beliefs. We begin with general results concerning two main types of modifications to the prior:

changes in the level of values and changes the co-movement of values across realizations. First, we show that anytime an option is improved in the prior then the DM is more likely to select it. We then show that if two options become more similar, defined as having values that co-move more strongly in the prior, then the probability that the DM selects either of these options falls. We connect this notion of similarity to the criticisms of the IIA property of the multinomial logit model and show that the rationally inattentive agent's behavior is not subject to the same criticisms. Finally, we demonstrate that the behavior of the rationally inattentive agent fails regularity as adding an additional option to the choice set can increase the probability that an existing option is selected. An implication of this failure of regularity is that the rationally inattentive agent cannot generally be viewed as maximizing a random utility function.

5.1 Monotonicity

The following proposition states that a change in the prior that makes one option more attractive *ceteris paribus* leads the DM to select that option with a higher unconditional probability and therefore a higher probability in all states of the world. Such a result does not generally hold under Bayesian updating if the information structure is exogenously given.[14] In this way, rational inattention places more structure on choice behavior than does incompleteness of information alone.

Proposition 3. *Assume $\lambda > 0$ and let $\{\mathcal{P}_i^0\}_{i=1}^N$ be the unique solution to the agent's maxi-*

[14] Here is a counterexample in which improving prior beliefs about one option leads it to be selected less often under Bayesian updating with an exogenous information structure. Imagine that there are two states of the world that each occur with probability 1/2. There are two available options that take the following values in the two states of the world

	state 1	state 2
v_1	-2	4
v_2	0	x.

Suppose the information structure is that the DM receives a signal $y = v_2 + \varepsilon$ where the noise ε is distributed uniformly on $(-1, 1)$. When $x = -2$, the signal perfectly distinguishes between the states of the world and so the DM selects option 1 in state 2 and option 2 in state 1. When option 2 is improved so that $x = 0$, the signal contains no information about the state of the world. In this case, the DM always selects option 1 as it has the larger expected value. So improving option 2 leads it to be selected less often.

mization problem with prior $G(\mathbf{v})$*. If* $\hat{G}(\hat{\mathbf{v}})$ *is generated from* $G(\mathbf{v})$ *by transforming* $\mathbf{v}$ *to* $\hat{\mathbf{v}}$ *such that* $\hat{v}_i = v_i$ *for all* $i > 1$*,* $\hat{v}_1 \geq v_1$ *for all* $\boldsymbol{v}$ *and* $\hat{v}_1 > v_1$ *on a set of positive measure, then* $\hat{\mathcal{P}}_1^0 \geq \mathcal{P}_1^0$*, where* $\left\{\hat{\mathcal{P}}_i^0\right\}_{i=1}^N$ *is the solution to the problem with prior* $\hat{G}$*. The last inequality holds strictly if* $\mathcal{P}_1^0 \in (0,1)$*.*

Proof. See Appendix C.1. □

That the monotonicity is strict only when $\mathcal{P}_1^0 \in (0,1)$ is intuitive. When $\mathcal{P}_1^0 = 1$, there is no scope for increasing its probability further. When $\mathcal{P}_1^0 = 0$, option 1 might be so unattractive to start with that the improvement does not lead the DM to select it with any probability.

In some special cases we can say even more. If $\lambda = 0$ so the agent is not inattentive at all then then the prior is irrelevant and the DM selects the highest value option in all states of the world. Conversely, if $\lambda = \infty$, the actual realization of $\mathbf{v}$ is irrelevant and the DM selects the option with the highest expected value according to the prior. The DM may also behave in this way for a finite λ if he *chooses* not to process any information. Finally, if one option is dominated by another in all states of the world, then it will never be selected.

5.2 Similarity and independence from irrelevant alternatives

While the previous proposition considers the effect of improving one option, suppose we leave the marginal distributions of the values of each option unchanged, but change the prior to increase the degree to which the values of two options co-move. What happens to the choice probabilities in this case? To motivate this investigation, consider Debreu's famous criticism of IIA, which is now known as the red-bus blue-bus problem. The well known example goes: The DM is pairwise indifferent between choosing a bus or a train, and selects each with probability 1/2. If a second bus of a different color is added to the choice set and the DM is indifferent to the color of the bus, then IIA—and therefore the multinomial logit, which can be derived from IIA—implies probabilities of 1/3, 1/3, 1/3. Debreu argued that this is counter-intuitive because duplicating one option should not materially change the choice problem.

What does it mean that the busses are duplicates? We will say that two options are duplicates if the prior asserts that the values of the two options are equal to one another in all states of the world: it is clear to the DM *a priori* that it is irrelevant which of the two options he selects. We will now show that the rationally inattentive agent does not display the counter-intuitive behavior that Debreu criticized and then expand these ideas to cases in which the options can be thought to be "similar" i.e. have values that are correlated across states of the world, but are not exact duplicates.

5.2.1 Duplicates

We study a generalized version of Debreu's bus problem to analyze how the rationally inattentive agent treats duplicate options. Duplicates carry the same value almost surely although this common value may be unknown.

Definition 2. *Options i and j are duplicates if and only if the probability that $v_i \neq v_j$ is zero.*

Problem 2. *The DM chooses $i \in \{1, \cdots, N+1\}$, where the options N and $N+1$ are duplicates.*

The following proposition states that duplicate options are treated as a single option. We compare the choice probabilities in two choice problems, where the second one is constructed from the first by duplicating one option. In the first problem, the DM's prior is $G(\mathbf{v})$, where $\mathbf{v} \in \mathbb{R}^N$. In the second problem, the DM's prior is $\hat{G}(\mathbf{u})$, where $\mathbf{u} \in \mathbb{R}^{N+1}$. $\hat{G}$ is generated from G by duplicating option N. This means that options N and $N+1$ satisfy Definition 2, and $G(\mathbf{v})$ is the marginal of $\hat{G}(\mathbf{u})$ with respect to u_{N+1}.

Proposition 4. *If $\{\mathcal{P}_i^0\}_{i=1}^N$ and $\{\mathcal{P}_i(\mathbf{v})\}_{i=1}^N$ are unconditional and conditional choice probabilities that are a solution to Problem 2, then $\{\hat{\mathcal{P}}_i(\mathbf{u})\}_{i=1}^{N+1}$ solve the corresponding problem*

with the added duplicate of the option N if and only if they satisfy the following:

$$\hat{\mathcal{P}}_i(\mathbf{u}) = \mathcal{P}_i(\mathbf{v}), \qquad \forall i < N \tag{18}$$

$$\hat{\mathcal{P}}_N(\mathbf{u}) + \hat{\mathcal{P}}_{N+1}(\mathbf{u}) = \mathcal{P}_N(\mathbf{v}), \tag{19}$$

where $\mathbf{v} \in \mathbb{R}^N$ and $\mathbf{u} \in \mathbb{R}^{N+1}$, and $v_k = u_k$ for all $k \leq N$. The analogous equalities hold for the unconditional probabilities.

Proof. See Appendix C.2. □

The implication of this proposition is that the DM treats duplicate options as though they were a single option. The behavior of the rationally inattentive agent does not always satisfy IIA and as a result is not subject to Debreu's critique. As we showed in section 4, a version of IIA holds for a fixed prior, but not when the prior changes.

5.2.2 Similar options

The case of exact duplicates is somewhat extreme as the DM knows *a priori* that the values of the two options are exactly equal. Here we consider options with values that are correlated, but not identical. We show that the probability that the DM selects either of two options (among three or more) decreases as those two options become more similar. By "more similar," we mean that we consider an alternative choice situation in which we increase the probability that the two options have the same value by shifting the prior probability from states of the world in which their values differ to states of the world where their values are the same. We have the following proposition.

Proposition 5. *Assume $\lambda > 0$ and let $\{\mathcal{P}_i^0\}_{i=1}^N$ be the unique solution to the agent's maximization problem with prior $G(\mathbf{v})$. Let the prior $G(\cdot)$ be such that the there exist two states of positive probabilities P_1 and P_2 with $(v_1, v_2) = (H, L)$ in state 1, and $(v_1, v_2) = (L, H)$ in state 2, and values of any other option be equal in both states.*

If $\hat{G}(\hat{\mathbf{v}})$ is generated from $G(\cdot)$ by relocating probability mass $\Pi \leq \min(P_1, P_2)$ from state 1 to state 3, where $(v_1, v_2) = (L, L)$, and relocating probability mass Π from state 2 to state

4, where $(v_1, v_2) = (H, H)$*, then* $\hat{\mathcal{P}}_1^0 + \hat{\mathcal{P}}_2^0 < \mathcal{P}_1^0 + \mathcal{P}_2^0$*, where* $\left\{\hat{\mathcal{P}}_i^0\right\}_{i=1}^N$ *is the solution to the problem with prior* $\hat{G}$*.*

Proof. See Appendix C.3. □

Intuitively, a higher degree of co-movement in the manner described in the proposition means that the event $(v_1 = H) \cup (v_2 = H)$ has lower probability and this is the event in which the DM is most interested in selecting option 1 or 2. In the next section, we present an example of this type of effect in which increasing the correlation between the values of two options decreases their cumulative probability.

5.3 Examples

We bring our analysis to a close with two examples. First, we demonstrate the effect of the degree of similarity among options. We then show that adding an additional option can increase the probability that an existing option is selected.

5.3.1 Co-movement

We now explore a choice among three options, where two options have positively or negatively correlated values. Even though all three options have the same *a priori* expected value, in some cases the DM will ignore one of the options completely. This example demonstrates that when the allocation of attention is endogenous, the DM can choose to investigate different options in different levels of detail.

Problem 3. *The DM chooses from the set* $\{$*red bus, blue bus, train*$\}$*. The DM knows the value of the train exactly,* $v_t = 1/2$*. The buses each take one of two values, either 0 or 1, with expected values 1/2 for each, the correlation coefficient between their values is* ρ*. The*

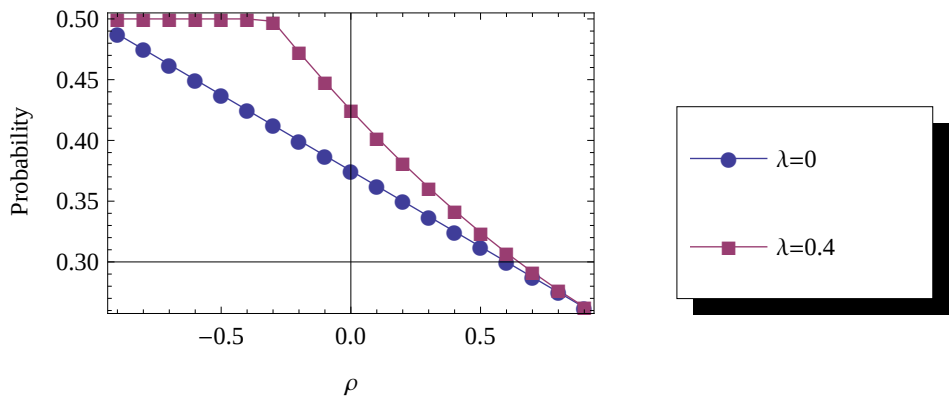

Figure 1: Unconditional probability of selecting a bus for various values of λ and ρ. The probability is the same for both the red and blue buses.

joint distribution of the values of all three options is:

$$\begin{aligned} g(0,0,1/2) &= \tfrac{1}{4}(1+\rho) \\ g(1,0,1/2) &= \tfrac{1}{4}(1-\rho) \\ g(0,1,1/2) &= \tfrac{1}{4}(1-\rho) \\ g(1,1,1/2) &= \tfrac{1}{4}(1+\rho). \end{aligned} \tag{20}$$

In Appendix D we describe how to solve the problem analytically. Figure 1 illustrates the behavior of the model for various values of ρ and λ. The figure shows the unconditional probability that the DM selects a bus of a given color (the probability is the same for both buses). As the correlation between the values of the buses decreases, the probability that a bus carries the largest value among the three options increases and the unconditional probability of choosing either bus increases, too. If the buses' values are perfectly correlated, then the sum of their probabilities is 0.5, they are effectively treated as one option, i.e. they become duplicates in the limit. On the other hand, if $\rho = -1$, then the unconditional probability of either bus is 0.5 and thus the train is never selected.

For $\lambda > 0$ and $\rho \in (-1, 1)$, the probability that a bus is selected is larger than it is in the perfect information case ($\lambda = 0$). With a larger cost of information, the DM economizes

on information by paying more attention to choosing among the buses and less to assessing their values relative to the reservation value 1/2.

The choice probabilities strongly reflect the endogeneity of the information structure in this case. As the correlation decreases, the DM knows that the best option is more likely to be one of the buses. As a result, the DM focusses more of his attention on choosing between the buses and eventually ignores the train completely. Notice that this can happen even when there is some chance that the train is actually the best option.

5.3.2 Failure of regularity

Random utility models, such as the standard multinomial logit model, have the feature that adding an additional option to the choice set will not increase the probability that an existing option is selected (Luce and Suppes, 1965, p. 342). However, the following example demonstrates that the behavior of the rationally inattentive agent does not always satisfy this regularity condition as adding an additional option can raise the probability that an existing option is selected.

Problem 4. *Suppose there are three options and two states of the world. The options take the following values in the two states of the world*

	state 1	*state 2*
option 1	0	1
option 2	1/2	1/2
option 3	Y	$-Y$

States 1 and 2 have prior probabilities $g(1)$ *and* $g(2)$*, respectively.*

First, consider a variant of this choice situation in which only options 1 and 2 are available. In Appendix D we show that there exists $g(1) \in (0,1)$ large enough that the DM will not process information and will select option 2 with probability 1 in all states of the world so $\mathcal{P}_1^0 = 0$. Now add option 3 to the choice set. For a large enough value of Y and the given $g(1) \in (0,1)$, the DM will find it worthwhile to process information about the state of the

world in order to determine whether option 3 should be selected. Given that the DM will now have information about the state of the world, if state 2 is realized, the DM might as well select option 1. From an *a priori* perspective, there is a positive probability of selecting option 1 so $\mathcal{P}_1^0 > 0$. The choice probabilities conditional on the realization of the state of the world are given by equation (9), which implies that the probability of selecting option 1 is zero if $\mathcal{P}_1^0 = 0$ and positive if $\mathcal{P}_1^0 > 0$ and all options have finite values. So we have the following.

Proposition 6. *For $\lambda > 0$, there exist $g(1) \in (0, 1)$ and $Y > 0$ such that adding option 3 to the choice set in Problem 4 increases the probability that option 1 is selected in all states of the world.*

Proof. See Appendix D. □

Corollary 2. *The behavior of a rationally inattentive agent cannot always be described by a random utility model.*

Obviously there are cases, such as the standard logit case, when the rationally inattentive agent's behavior *can* be described by a random utility model.

6 Conclusion

This paper builds on the literature that treats scarce attention as an important factor in choice behavior. Rational inattention is an appealing model of attention allocation because it does not depend on assumptions about how the DM allocates his attention to the available information except in the form of a well-founded cost function that measures the DM's information processing effort. While appealing, the rational inattention model in its original continuous-action form is considered by some to be intractable. In this paper, however, we have shown that the discrete choice behavior of a rationally inattentive agent has a simple analytical structure providing a new foundation for the multinomial logit model. The standard logit model emerges in cases where the options are homogeneous *a priori*. More

generally, choices depend on the context formed by prior beliefs. While the DM's prior is a complicated infinite-dimensional object, its effect on choice probabilities is captured by a simple vector of position weights. The resulting tractability allows us to establish results on monotonicity, co-movement, and uniqueness, which are general and fundamental properties of rationally inattentive behavior. The discrete choice framework presented here facilitates the application of rational inattention to new questions.

References

Anas, A. (1983). Discrete choice theory, information theory and the multinomial logit and gravity models. *Transportation Research Part B: Methodological*, 17(1):13–23.

Anderson, S. P., de Palma, A., and Thisse, J.-F. (1992). *Discrete Choice Theory of Product Differentiation.* MIT Press, Cambridge, MA.

Blackwell, D. (1953). Equivalent comparison of experiments. *Annals of Mathematical Statistics*, (24).

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):pp. 324–345.

Caplin, A. and Martin, D. (2011). A testable theory of imperfect perception. NBER Working Paper 17163.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley, Hoboken, NJ.

Debreu, G. (1960). Review of individual choice behavior by R. D. Luce. *American Economic Review*, 50(1).

Goldberg, P. K. (1995). Product differentiation and oligopoly in international markets: The case of the u.s. automobile industry. *Econometrica*, 63(4).

Gul, F., Natenzon, P., and Pesendorfer, W. (2012). Random choice as behaviorial optimization. Princeton University Working Paper.

Kahneman, D. (1973). *Attention and Effort*. Prentice Hall, New Jersey.

Lindbeck, A. and Weibull, J. W. (1987). Balanced-budget redistribution as the outcome of political competition. *Public Choice*, 52.

Luce, R. D. (1959). *Individual Choice Behavior: a Theoretical Analysis*. Wiley, New York.

Luce, R. D. and Suppes, P. (1965). Preference, utility, and subjective probability. In Luce, R. D.; Bush, R. and Galanter, E., editors, *Handbook of Mathematical Psychology*, volume 3, pages 249–410. Wiley, New York.

Luo, Y. (2008). Consumption dynamics under information processing constraints. *Review of Economic Dynamics*, 11(2).

Luo, Y. and Young, E. R. (2009). Rational inattention and aggregate fluctuations. *The B.E. Journal of Macroeconomics*, 9(1).

Mackowiak, B. and Wiederholt, M. (2009). Optimal sticky prices under rational inattention. *The American Economic Review*, 99.

Maćkowiak, B. A. and Wiederholt, M. (2010). Business cycle dynamics under rational inattention. CEPR Discussion Papers 7691, C.E.P.R. Discussion Papers.

Manzini, P. and Mariotti, M. (2012). Stochastic choice and consideration sets.

Masatlioglu, Y., Nakajima, D., and Ozbay, E. Y. (2012). Revealed attention. *The American Economic Review*, 102(5):2183–2205.

Mattsson, L.-G. and Weibull, J. W. (2002). Probabilistic choice and procedurally bounded rationality. *Games and Economic Behavior*, 41(1).

Matějka, F. (2010a). Rationally inattentive seller: Sales and discrete pricing. CERGE-EI Working Papers wp408.

Matějka, F. (2010b). Rigid pricing and rationally inattentive consumer. CERGE-EI Working Papers wp409.

Matějka, F. and McKay, A. (2012). Simple market equilibria with rationally inattentive consumers. *American Economic Review*, 102(3):24 – 29.

Matějka, F. and Sims, C. A. (2010). Discrete actions in information-constrained tracking problems. Technical report, Princeton University.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*. Academic Press, New York.

McFadden, D. (2001). Economic choices. *The American Economic Review*, 91(3):pp. 351–378.

McKelvey, R. D. and Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1).

Mondria, J. (2010). Portfolio choice, attention allocation, and price comovement. *Journal of Economic Theory*, 145(5).

Natenzon, P. (2010). Random choice and learning. Working paper, Princeton University.

Paciello, L. and Wiederholt, M. (2011). Exogenous information, endogenous information and optimal monetary policy. Working paper, Northwestern University.

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

Shannon, C. E. (1959). Coding theorems for a discrete source with a fidelity criterion. In *Institute of Radio Engineers, International Convention Record*, volume 7.

Simon, H. A. (1959). Theories of decision-making in economics and behavioral science. *The American Economic Review*, 49(3):pp. 253–283.

Sims, C. A. (1998). Stickiness. *Carnegie-Rochester Conference Series on Public Policy*, 49.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50(3).

Sims, C. A. (2006). Rational inattention: Beyond the linear-quadratic case. *The American Economic Review*, 96(2).

Stahl, D. O. (1990). Entropy control costs and entropic equilibria. *International Journal of Game Theory*, 19(2).

Train, K. (2009). *Discrete Choice Methods with Simulation.* Cambridge University Press, Cambridge, U.K.

Tutino, A. (2009). The rigidity of choice: Lifetime savings under information-processing constraints. Working paper.

Van Nieuwerburgh, S. and Veldkamp, L. (2010). Information acquisition and under-diversification. *Review of Economic Studies*, 77(2).

Verboven, F. (1996). International price discrimination in the european car market. *RAND Journal of Economics*, 27(2).

Weibull, J. W., Mattsson, L.-G., and Voorneveld, M. (2007). Better may be worse: Some monotonicity results and paradoxes in discrete choice under uncertainty. *Theory and Decision*, 63.

Woodford, M. (2008). Inattention as a source of randomized discrete adjustment.

Woodford, M. (2009). Information-constrained state-dependent pricing. *Journal of Monetary Economics*, 56(Supplement 1).

Yang, M. (2011). Coordination with rational inattention. Princeton University Working Paper.

Yellott, J. I. (1977). The relationship between luce's choice axiom, thurstone's theory of comparative judgment, and the double exponential distribution. *Journal of Mathematical Psychology*, 15(2).

A Proof for Section 3

Proof of Theorem 1. The case of $\lambda = 0$ is trivial. When $\lambda > 0$, then the Lagrangian of the DM's problem formulated in Definition 1 is:

$$\mathcal{L}(\mathcal{P}) = \sum_{i=1}^{N} \int_{\mathbf{v}} v_i \mathcal{P}_i(\mathbf{v}) G(d\mathbf{v}) - \lambda \left(- \sum_{i=1}^{N} \mathcal{P}_i^0 \log \mathcal{P}_i^0 + \sum_{i=1}^{N} \int_{\mathbf{v}} \mathcal{P}_i(\mathbf{v}) \log \mathcal{P}_i(\mathbf{v}) G(d\mathbf{v}) \right) + \int_{\mathbf{v}} \xi_i(\mathbf{v}) \mathcal{P}_i(\mathbf{v}) G(d\mathbf{v}) - \int_{\mathbf{v}} \mu(\mathbf{v}) \left(\sum_{i=1}^{N} \mathcal{P}_i(\mathbf{v}) - 1 \right) G(d\mathbf{v}),$$

where $\xi_i(\mathbf{v}) \geq 0$ are Lagrange multipliers on (3) and $\mu(\mathbf{v})$ are the multipliers on (4).

If $\mathcal{P}_i^0 > 0$, then the first order condition with respect to $\mathcal{P}_i(\mathbf{v})$ is:

$$v_i + \xi_i(\mathbf{v}) - \mu(\mathbf{v}) + \lambda \Big(\log \mathcal{P}_i^0 + 1 - \log \mathcal{P}_i(\mathbf{v}) - 1 \Big) = 0. \tag{21}$$

This implies that if $\mathcal{P}_i^0 > 0$ and $v_i > -\infty$, then $\mathcal{P}_i(\mathbf{v}) > 0$ almost surely. To see this, suppose to the contrary that $\mathcal{P}_i(\mathbf{v}) = 0$ on a set of positive measure with respect to G. Since $\xi_i(\mathbf{v}) \geq 0$ and since we also assume that $\mathcal{P}_i^0 > 0$, and thus $\log \mathcal{P}_i^0 > -\infty$, it would have to be $\mu(\mathbf{v})$ going to infinity that would balance $\log \mathcal{P}_i(\mathbf{v}) = -\infty$ to make the first order condition hold. However, if $\mu(\mathbf{v}) = \infty$ on a set of positive measure, then for all such $\mathbf{v}$ in order for (21) to hold then for all j either $\mathcal{P}_j(\mathbf{v}) = 0$ or $\xi_j(\mathbf{v}) = \infty$. But $\xi_j(\mathbf{v}) > 0$ only if $\mathcal{P}_j(\mathbf{v}) = 0$, when (3) is binding. Therefore, if there exists i such that $\mathcal{P}_i(\mathbf{v}) = 0$, then $\mathcal{P}_j(\mathbf{v}) = 0$ for all j. This is not possible, since then $\sum_{j=1}^{N} \mathcal{P}_j(\mathbf{v}) = 0$, which must sum up to 1 and hence a contradiction.

Therefore, whenever $\mathcal{P}_i^0$ is positive, then the conditional probability $\mathcal{P}_i(\mathbf{v})$ is also positive as long as the realized value v_i is not minus infinity. As (3) does not bind, we have $\xi_i(\mathbf{v}) = 0$ and the first order condition can be rearranged to

$$\mathcal{P}_i(\mathbf{v}) = \mathcal{P}_i^0 e^{(v_i - \mu(\mathbf{v}))/\lambda}. \tag{22}$$

Plugging (22) into (4), we find:

$$e^{\mu(\mathbf{v})/\lambda} = \sum_i P_i^0 e^{v_i/\lambda},$$

which we again use in (22) to arrive at equation (9). Finally, notice that the theorem holds even for $\mathcal{P}_i^0 = 0$, as otherwise equation (7) could not hold.

Proof of Lemma 1. Substitute equation (9) into the objective function to obtain

$$\sum_{i=1}^{N} \int_{\mathbf{v}} v_i \mathcal{P}_i(\mathbf{v})\, G(d\mathbf{v}) + \lambda \left\{ \sum_{i=1}^{N} \mathcal{P}_i^0 \log \mathcal{P}_i^0 - \int_{\mathbf{v}} \left[\sum_{i=1}^{N} \mathcal{P}_i(\mathbf{v}) \log \left(\frac{\mathcal{P}_i^0 e^{v_i/\lambda}}{\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda}} \right) \right] G(d\mathbf{v}) \right\}$$

and rearrange to obtain

$$\begin{aligned}
&\int_{\mathbf{v}} \sum_{i=1}^{N} \mathcal{P}_i(\mathbf{v}) \left[v_i - \lambda \log \left(\frac{\mathcal{P}_i^0 e^{v_i/\lambda}}{\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda}} \right) \right] G(d\mathbf{v}) + \lambda \sum_{i=1}^{N} \mathcal{P}_i^0 \log \mathcal{P}_i^0 \\
&\quad = \int_{\mathbf{v}} \sum_{i=1}^{N} \mathcal{P}_i(\mathbf{v}) \left[v_i - v_i - \lambda \log \left(\mathcal{P}_i^0 \right) + \lambda \log \left(\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda} \right) \right] G(d\mathbf{v}) + \lambda \sum_{i=1}^{N} \mathcal{P}_i^0 \log \mathcal{P}_i^0 \\
&\quad = \int_{\mathbf{v}} \sum_{i=1}^{N} \mathcal{P}_i(\mathbf{v}) \lambda \log \left(\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda} \right) G(d\mathbf{v}) - \lambda \sum_{i=1}^{N} \underbrace{\int_{\mathbf{v}} \mathcal{P}_i(\mathbf{v})\, G(d\mathbf{v})}_{=\mathcal{P}_i^0} \log \mathcal{P}_i^0 + \lambda \sum_{i=1}^{N} \mathcal{P}_i^0 \log \mathcal{P}_i^0 \\
&\quad = \int_{\mathbf{v}} \left[\sum_{i=1}^{N} \mathcal{P}_i(\mathbf{v}) \right] \lambda \log \left(\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda} \right) G(d\mathbf{v}) \\
&\quad = \int_{\mathbf{v}} \lambda \log \left(\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda} \right) G(d\mathbf{v}),
\end{aligned}$$

where the last line follows from the fact that $\mathcal{P}_i(\mathbf{v})$ is the conditional probability of selecting i given $\mathbf{v}$ and so the sum is equal to one. □

Proof of Corollary 1. For $\mathcal{P}_i^0 > 0$, the first order condition on (10) with respect to $\mathcal{P}_i^0$, is

$$\lambda \int_{\mathbf{v}} \frac{e^{v_i/\lambda} - e^{v_N/\lambda}}{\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda}} G(d\mathbf{v}) = 0, \tag{23}$$

where $\mathcal{P}_N^0$ denotes $1 - \sum_{i=1}^{N-1} \mathcal{P}_i^0$.

For $i \in \{1, \cdots, N-1\}$, we can write

$$\int_{\mathbf{v}} \frac{e^{v_i/\lambda}}{\sum_{i=j}^{N} \mathcal{P}_j^0 e^{v_j/\lambda}} G(d\mathbf{v}) = \int_{\mathbf{v}} \frac{e^{v_N/\lambda}}{\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda}} G(d\mathbf{v}) \equiv \mu.$$

Notice that $\mu = 1$ because

$$\begin{aligned}
\sum_{i=1}^{N} \mathcal{P}_i^0 \mu &= \sum_{i=1}^{N} \mathcal{P}_i^0 \int_{\mathbf{v}} \frac{e^{v_i/\lambda}}{\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda}} G(d\mathbf{v}) \\
&= \int_{\mathbf{v}} \frac{\sum_{i=1}^{N} \mathcal{P}_i^0 e^{v_i/\lambda}}{\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda}} G(d\mathbf{v}) \\
&= \int_{\mathbf{v}} G(d\mathbf{v}) = 1
\end{aligned}$$

so $\mu \sum_{i=1}^{N} \mathcal{P}_i^0 = 1$, but as $\{\mathcal{P}_i^0\}_{i=1}^N$ are probabilities we know $\sum_{i=1}^{N} \mathcal{P}_i^0 = 1$. Therefore, $\mu = 1$. Equation (23) then becomes (13). □

Lemma 3. *The DM's optimization problem in Definition 1 always has a solution.*

Proof. Since (9) is a necessary condition for the maximum, then the collection $\{\mathcal{P}_i^0\}_{i=1}^N$ determines the whole solution. However, the objective is a continuous function of $\{\mathcal{P}_i^0\}_{i=1}^N$, since $\{\mathcal{P}_i(\mathbf{v})\}_{i=1}^N$ is also a continuous function of $\{\mathcal{P}_i^0\}_{i=1}^N$. Moreover, the admissible set for $\{\mathcal{P}_i^0\}_{i=1}^N$ is compact. Therefore, the maximum always exists. □

Uniqueness Concerning uniqueness, there can be special cases where the DM is indifferent between processing more information in order to generate a higher $E[v_i]$ and processing less information and conserving on information costs. However, a rigid co-movement of values is required for these cases to arise. Without this structure, if the DM were indifferent between two different strategies then their convex combination would be preferred as the entropy cost is convex in strategies, $\{\mathcal{P}_i(\mathbf{v})\}_{i=1}^N$, while $E[v_i]$ is linear.

Lemma 4. ***Uniqueness:*** *If the random vectors $e^{v_j/\lambda}$ are linearly independent with unit*

scaling, i.e. if there does not exist a set $\{a_j\}_{j=2}^{N-1}$ *such that* $\sum_{k=2}^{N} a_k = 1$ *and*

$$e^{v_i/\lambda} = \sum_{k=2}^{N} a_k e^{v_j/\lambda} \qquad a.s., \tag{24}$$

then the agent's problem has a unique solution. Conversely, if the agent's problem has a unique solution, then the random vectors $e^{v_j/\lambda}$ *for all* j *s.t.* $\mathcal{P}_j^0 > 0$ *are linearly independent with unit scaling.*

Proof. Let us first study an interior optimum of (10), where the boundary constraint (11) is not binding. The first order conditions take the form of (13) for $i < N$ and denote $\mathcal{P}_N^0 = 1 - \sum_{k=1}^{N-1} \mathcal{P}_k^0$ to satisfy the constraint (12). The $(N-1)$ dimensional Hessian is

$$H_{ij} = -\int_{\mathbf{v}} \Big(\frac{e^{v_i/\lambda} - e^{v_N/\lambda}}{\sum_{k=1}^{N} \mathcal{P}_k^0 e^{v_k/\lambda}} \Big) \Big(\frac{e^{v_j/\lambda} - e^{v_N/\lambda}}{\sum_{k=1}^{N} \mathcal{P}_k^0 e^{v_k/\lambda}} \Big) G(d\mathbf{v}) \qquad \forall i, j < N. \tag{25}$$

The hessian H is thus (-1) times a Gramian matrix, which is a matrix generated from inner products of random vectors $\frac{e^{v_i/\lambda} - e^{v_N/\lambda}}{\sum_{k=1}^{N} \mathcal{P}_k^0 e^{v_k/\lambda}}$. H is thus negative semi-definite at all interior points $\{\mathcal{P}_i^0\}_{i=1}^{N}$, not just at the optimum only. This implies that an interior optimum of the objective (10) is a unique maximum if and only if the Hessian is negative definite at the optimum. From (25) we see that for $N = 2$ the Hessian is not negative-definite, i.e. the objective is not a strictly concave function of $\mathcal{P}_1^0$, only when $e^{v_1} = e^{v_2}$ almost surely, i.e. when the options are identical. Moreover for $N > 2$, we can use the fact that Gramian matrices have a zero eigenvalue if and only if the generating vectors are linearly dependent, which means that there exist i and a set $\{a_j\}_{j=1,\neq i}^{N-1}$ such that

$$\frac{e^{v_i/\lambda} - e^{v_N/\lambda}}{\sum_{k=1}^{N} \mathcal{P}_k^0 e^{v_k/\lambda}} = \sum_{j=1,\neq i}^{N-1} a_j \frac{e^{v_j/\lambda} - e^{v_N/\lambda}}{\sum_{k=1}^{N} \mathcal{P}_k^0 e^{v_k/\lambda}} \qquad a.s. \tag{26}$$

Since the equality needs to hold almost surely, we can get rid of the denominators, which are the same on both sides. By denoting $a_N = 1 - \sum_{j=1,\neq i}^{N-1} a_j$, this implies the following sufficient and necessary condition for non-uniqueness in the interior for $N \geq 1$: there exists

a set $\{a_j\}_{j=1,\neq i}^{N}$ such that $\sum_{k=2}^{N} a_k = 1$, and

$$e^{v_i/\lambda} = \sum_{k=1,\neq i}^{N} a_k e^{v_j/\lambda} \qquad a.s. \tag{27}$$

Now, we extend the proof to non-interior solutions. In the the agent's optimization problem formulated in Definition 1, the objective is a weakly concave function of $\{\mathcal{P}_i(\mathbf{v})\}_i$ on a convex set. Therefore, any convex linear combination of two solutions must be a solution, too. This implies that there always exists a solution with $\mathcal{P}_i^0 > 0$ for all $i \in S$, where S is a set of indices i for which there exists a solution with $\mathcal{P}_i^0 > 0$. For example, if there exist two distinct solutions such that $\mathcal{P}_1^0 > 0$ for one and $\mathcal{P}_2^0 > 0$ for the other, then there exists a solution with both $\mathcal{P}_1^0$ and $\mathcal{P}_2^0$ positive. Therefore, there always exists an interior solution on the subset S of options that are ever selected in some solution. Moreover, we can create many such solutions by taking different convex combinations. However, if the conditions of the proposition are satisfied, then there can only be one interior solution and hence the multiple boundary solutions leads to a contradiction. For the options that are not selected in any solution, the solution is unique with $\mathcal{P}_i^0 = 0$. □

Corollary 3. *If any one of the following conditions is satisfied, then the solution is unique.*

(1) $N = 2$ and the values of the two options are not equal almost surely.

(2) The prior is symmetric and values of the options are not equal almost surely.

Proof of Proposition 1. The solution to the DM's problem is unique due to Corollary 3, point (2).

The DM forms a strategy such that $\mathcal{P}_i^0 = 1/N$ for all i. If there were a solution with non-uniform $\mathcal{P}_i^0$, then any permutation of the set would necessarily be a solution too, but the solution is unique. Using $\mathcal{P}_i^0 = 1/N$ in equation (9), we arrive at the result. □

B Proofs for Section 4

Proof of Lemma 2. Let there exist an $\mathbf{x}$ such that $\mathcal{P}_i(\mathbf{x}) > 0$. Let without the loss of generality $i = 1$. We prove the statement by showing that the position 1 is selected with positive probability for all vectors $\mathbf{y}$.

Let us first generate a vector $\mathbf{x^{11}}$ by copying x_1 to position 2: $\mathbf{x^{11}} = (x_1, x_1, x_3, \ldots, x_N)$. We find that $\mathcal{P}_1(\mathbf{x^{11}}) > 0$, which is due to Axiom 1 when it is applied to $\mathbf{x}$ and $\mathbf{x^{11}}$ with $i = 3$ and $j = 1$: the axiom implies that $\frac{\mathcal{P}_3(\mathbf{x})}{\mathcal{P}_1(\mathbf{x})} = \frac{\mathcal{P}_3(\mathbf{x^{11}})}{\mathcal{P}_1(\mathbf{x^{11}})}$. Now, Axiom 2 implies that as long as there is the same object in both positions 1 and 2, then the probability of position 1 is positive independently of what the object is and what objects in the other positions are.

Finally, we show that $\mathcal{P}_1(\mathbf{y}) > 0$, where $\mathbf{y}$ is an arbitrary vector of objects. This is due to the fact that $\mathcal{P}_1(\mathbf{y^{11}}) > 0$, which we just showed in the paragraph above, where $\mathbf{y^{11}} = (y_1, y_1, y_3, \ldots, y_N)$, and due to Axiom 1 when it is applied to $\mathbf{y^{11}}$, $\mathbf{y}$, $i = 3$ and $j = 1$. The Axiom implies that if $\mathcal{P}_1(\mathbf{y^{11}}) > 0$, then $\mathcal{P}_1(\mathbf{y}) > 0$ too, since $\mathbf{y}$ and $\mathbf{y}^{11}$ differ in position 2 only.

To establish that if $\mathcal{P}_i(\mathbf{x}) = 0$ then $\mathcal{P}_i(\mathbf{y}) = 0$ for all $\mathbf{y}$ is straightforward: suppose $\mathcal{P}_i(\mathbf{y}) > 0$, then the argument above implies $\mathcal{P}_i(\mathbf{x}) > 0$. □

Proof of Proposition 2. Assume w.l.o.g. that positions 1, 2, and N are positive. Fix a vector of objects $\mathbf{x}$ and define

$$v(a) \equiv \log\left(\frac{\mathcal{P}_1(a, x_2, x_3, \cdots, x_N)}{\mathcal{P}_N(a, x_2, x_3, \cdots, x_N)}\right). \tag{28}$$

So the value of object a is defined in terms of the probability that it is selected when it is inserted into the first position of a particular vector of objects. Also define

$$\xi_k \equiv \frac{\mathcal{P}_k(\mathbf{x^k})}{\mathcal{P}_1(\mathbf{x^k})},$$

where $\mathbf{x^k}$ is defined as $(x_k, x_2, x_3, \cdots, x_N)$, which is $\mathbf{x}$ with the first element replaced by a second instance of the object in the k^{th} position. Notice that if k is a zero position, then

$\xi_k = 0$. By Axiom 2, we have

$$\xi_k = \frac{\mathcal{P}_k(\mathbf{y^k})}{\mathcal{P}_1(\mathbf{y^k})},$$

where $\mathbf{y^k}$ is generated from an arbitrary vector of objects $\mathbf{y}$ in the same manner that $\mathbf{x^k}$ was generated from $\mathbf{x}$.

Consider a vector of objects $\mathbf{y}$ that shares the N^{th} entry with the generating vector $\mathbf{x}$, such that $y_N = x_N$. We will show

$$\mathcal{P}_i(\mathbf{y}) = \frac{\xi_i e^{v(y_i)}}{\sum_{j=1}^{N} \xi_j e^{v(y_j)}}, \tag{29}$$

for all i. If i is a zero position, then (29) holds trivially so we will suppose that i is a positive position. As the choice probabilities must sum to one, we proceed as follows

$$\begin{aligned} 1 = \sum_j \mathcal{P}_j(\mathbf{y}) = \mathcal{P}_i(\mathbf{y}) \sum_j \frac{\mathcal{P}_j(\mathbf{y})}{\mathcal{P}_i(\mathbf{y})} = \mathcal{P}_i(\mathbf{y}) \sum_j \frac{\mathcal{P}_j(\mathbf{y})/\mathcal{P}_N(\mathbf{y})}{\mathcal{P}_i(\mathbf{y})/\mathcal{P}_N(\mathbf{y})} \\ \mathcal{P}_i(\mathbf{y}) = \frac{\mathcal{P}_i(\mathbf{y})/\mathcal{P}_N(\mathbf{y})}{\sum_j \mathcal{P}_j(\mathbf{y})/\mathcal{P}_N(\mathbf{y})}. \end{aligned} \tag{30}$$

Now, by Axiom 1:

$$\frac{\mathcal{P}_k(\mathbf{y})}{\mathcal{P}_N(\mathbf{y})} = \frac{\mathcal{P}_k(\mathbf{y^k})}{\mathcal{P}_N(\mathbf{y^k})}$$

so (30) becomes

$$\begin{aligned} \mathcal{P}_i(\mathbf{y}) &= \frac{\mathcal{P}_i(\mathbf{y^i})/\mathcal{P}_N(\mathbf{y^i})}{\sum_j \mathcal{P}_j(\mathbf{y^j})/\mathcal{P}_N(\mathbf{y^j})} \\ &= \frac{\xi_i \mathcal{P}_1(\mathbf{y^i})/\mathcal{P}_N(\mathbf{y^i})}{\sum_j \xi_j \mathcal{P}_1(\mathbf{y^j})/\mathcal{P}_N(\mathbf{y^j})}. \end{aligned} \tag{31}$$

For any k, as $y_N = x_N$ in the case we are considering, by Axiom 1 and definition of $v(\cdot)$ in

(28):

$$\frac{\mathcal{P}_1(\mathbf{y^k})}{\mathcal{P}_N(\mathbf{y^k})} = \frac{\mathcal{P}_1(y_k, x_2, x_3, \cdots, x_N)}{\mathcal{P}_N(y_k, x_2, x_3, \cdots, x_N)}$$
$$= e^{v(y_k)}.$$

Therefore (31) becomes (29).

We will now consider an arbitrary $\mathbf{y}$ allowing for $y_N \neq x_N$ as well. We will show that (29) still holds by using the axioms and some intermediate vectors to connect $\mathbf{y}$ to $\mathbf{x}$. Let $\mathbf{y^{wizj}}$ be the vector generated from $\mathbf{y}$ by replacing the first element of $\mathbf{y}$ with w_i and the last element of $\mathbf{y}$ with z_j for given vectors $\mathbf{w}$ and $\mathbf{z}$. For example:

$$\mathbf{y^{y1xN}} = (y_1, y_2, y_3, \cdots, y_{N-1}, x_N)$$
$$\mathbf{y^{x1xN}} = (x_1, y_2, y_3, \cdots, y_{N-1}, x_N)$$
$$\mathbf{y^{xNxN}} = (x_N, y_2, y_3, \cdots, y_{N-1}, x_N)$$
$$\mathbf{y^{yNyN}} = (y_N, y_2, y_3, \cdots, y_{N-1}, y_N).$$

Consider $\mathbf{y^{yNyN}}$. For any $i < N$, by Axiom 1:

$$\mathcal{P}_i(\mathbf{y^{yNyN}}) = \mathcal{P}_1(\mathbf{y^{yNyN}})\frac{\mathcal{P}_i(\mathbf{y^{yNxN}})}{\mathcal{P}_1(\mathbf{y^{yNxN}})}$$
$$= \mathcal{P}_1(\mathbf{y^{yNyN}})\frac{\xi_i e^{v(y_i)}}{\xi_1 e^{v(y_N)}}, \tag{32}$$

where the second equality follows from the fact that (29) holds for $\mathbf{y} = \mathbf{y^{yNxN}}$ as its N^{th} entry is x_N and we have already established that (29) holds for vectors $\mathbf{y}$ for which $y_n = x_N$. For $i = N$ we have, by Axiom 2:

$$\mathcal{P}_N(\mathbf{y^{yNyN}}) = \mathcal{P}_1(\mathbf{y^{yNyN}})\frac{\mathcal{P}_N(\mathbf{y^{xNxN}})}{\mathcal{P}_1(\mathbf{y^{xNxN}})}$$
$$= \mathcal{P}_1(\mathbf{y^{yNyN}})\frac{\xi_N e^{v(x_N)}}{\xi_1 e^{v(x_N)}} = \mathcal{P}_1(\mathbf{y^{yNyN}})\frac{\xi_N}{\xi_1}. \tag{33}$$

Combining (32) and (33),

$$\frac{\mathcal{P}_i(\mathbf{y}^{\mathbf{yNyN}})}{\mathcal{P}_N(\mathbf{y}^{\mathbf{yNyN}})} = \frac{\xi_i e^{v(y_i)}}{\xi_N e^{v(y_N)}} \tag{34}$$

for all i. As the probabilities sum to one, we arrive at

$$\mathcal{P}_N(\mathbf{y}^{\mathbf{yNyN}}) = \frac{\xi_N e^{v(y_N)}}{\sum_j \xi_j e^{v(y_j)}} \tag{35}$$

and (29) for $\mathbf{y} = \mathbf{y}^{\mathbf{yNyN}}$ follows from (34) and (35).

Finally, we turn our attention to the arbitrary $\mathbf{y}$. For any $j < N$, we use Axiom 1 to write

$$\frac{\mathcal{P}_j(\mathbf{y})}{\mathcal{P}_2(\mathbf{y})} = \frac{\mathcal{P}_j(\mathbf{y}^{\mathbf{y1xN}})}{\mathcal{P}_2(\mathbf{y}^{\mathbf{y1xN}})} = \frac{\xi_j e^{v(y_j)}}{\xi_2 e^{v(y_2)}}, \tag{36}$$

where the second equality follows from the fact that (29) has already been established for $\mathbf{y} = \mathbf{y}^{\mathbf{y1xN}}$. For $j = N$, by Axiom 1 we can write

$$\frac{\mathcal{P}_N(\mathbf{y})}{\mathcal{P}_2(\mathbf{y})} = \frac{\mathcal{P}_N(\mathbf{y}^{\mathbf{yNyN}})}{\mathcal{P}_2(\mathbf{y}^{\mathbf{yNyN}})} = \frac{\xi_N e^{v(y_N)}}{\xi_2 e^{v(y_2)}}, \tag{37}$$

where the second equality follows from the fact that (29) has already been established for $\mathbf{y} = \mathbf{y}^{\mathbf{yNyN}}$. Using $\sum_j \mathcal{P}_j(\mathbf{y}) = 1$ we arrive at

$$\mathcal{P}_2(\mathbf{y}) = \frac{\xi_2 e^{v(y_2)}}{\sum_j \xi_j e^{v(y_j)}}$$

and then (29) follows from (36) and (37).

To complete the proof, we apply the normalization $\mathcal{P}_i^0 = \xi_i / \sum_j \xi_j$ for all i. □

C Proofs for Section 5

C.1 Monotonicity

Proof of Proposition 3. The agent's objective function, (10), can be rewritten to include the constraint $\sum_{i=1}^{N} \mathcal{P}_i^0 = 1$

$$\int_{\mathbf{v}} \lambda \log \left[\sum_{i=1}^{N-1} \mathcal{P}_i^0 e^{v_i/\lambda} + \left(1 - \sum_{i=1}^{N-1} \mathcal{P}_i^0 \right) e^{v_N/\lambda} \right] G(d\mathbf{v}).$$

Written in this way, the agent is maximizing over $\{\mathcal{P}_i^0\}_{i=1}^{N-1}$ subject to (11). Let us first assume that the constraint (11) is not binding and later on we show the statement holds in general.

The first order condition with respect to $\mathcal{P}_1^0$ is

$$\lambda \int_{\mathbf{v}} \frac{e^{v_1/\lambda} - e^{v_N/\lambda}}{\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda}} G(d\mathbf{v}) = 0, \tag{38}$$

where $\mathcal{P}_N^0$ denotes $1 - \sum_{i=1}^{N-1} \mathcal{P}_i^0$.

$\hat{G}(\cdot)$ is generated from $G(\cdot)$ by increasing the values of option 1 and this change can be implemented using a function $f(\mathbf{v}) \geq 0$, where $\int f(\mathbf{v}) G(d\mathbf{v}) > 0$, which describes the increase in v_1 in various states. Let v be transformed such that $e^{\hat{v}_1/\lambda} = e^{v_1/\lambda}(1 + f(\mathbf{v}))$ and with $\hat{v}_j = v_j$ for all $\mathbf{v}$ and $j = 2, \cdots, N$. Under the new prior, $\hat{G}(\cdot)$, the left-hand side of (38) becomes

$$\Delta_1 \equiv \lambda \int_{\mathbf{v}} \frac{e^{v_1/\lambda}(1 + f(\mathbf{v})) - e^{v_N/\lambda}}{\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda} + \mathcal{P}_1^0 e^{v_1/\lambda} f(\mathbf{v})} G(d\mathbf{v}). \tag{39}$$

Notice that $[\Delta_1, \cdots, \Delta_{N-1}]$ is the gradient of the agent's objective function under the new prior evaluated at the optimal strategy under the original prior. We now consider a marginal improvement in the direction of $f(\mathbf{v})$. In particular, consider an improvement of $\varepsilon f(\mathbf{v})$ for

some $\varepsilon > 0$. Equation (39) becomes

$$\Delta_1 = \lambda \int_{\mathbf{v}} \frac{e^{v_1/\lambda}\left(1+\varepsilon f(\mathbf{v})\right) - e^{v_N/\lambda}}{\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda} + \mathcal{P}_1^0 e^{v_1/\lambda}\varepsilon f(\mathbf{v})} G(d\mathbf{v}). \tag{40}$$

Differentiating with respect to ε at $\varepsilon = 0$ leads to

$$\left.\frac{\partial \Delta_1}{\partial \varepsilon}\right|_{\varepsilon=0} = \lambda \int_{\mathbf{v}} \frac{e^{v_1/\lambda} f(\mathbf{v}) \sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda} - \left(e^{v_1/\lambda} - e^{v_N/\lambda}\right) \mathcal{P}_1^0 e^{v_1/\lambda} f(\mathbf{v})}{\left(\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda}\right)^2} G(d\mathbf{v}) \tag{41}$$

$$= \lambda \int_{\mathbf{v}} e^{v_1/\lambda} f(\mathbf{v}) \frac{\sum_{j=2}^{N} \mathcal{P}_j^0 e^{v_j/\lambda} + e^{v_N/\lambda} \mathcal{P}_1^0}{\left(\sum_{j=1}^{N} \mathcal{P}_j^0 e^{v_j/\lambda}\right)^2} G(d\mathbf{v}) > 0. \tag{42}$$

This establishes that at the original optimum, $\{\mathcal{P}_i^0\}_{i=1}^{N-1}$, the impact of a marginal improvement in option 1 is to increase the gradient of the new objective function with respect to the probability of the first option. Therefore the agent will increase $\mathcal{P}_1^0$ in response to the marginal improvement. Notice that this holds for a marginal change derived from any total $f(\mathbf{v})$. Therefore, if the addition of the total $f(\mathbf{v})$ were to decrease $\mathcal{P}_1^0$, then by regularity there would have to be a marginal change of the prior along $\nu f(\mathbf{v})$, where $\nu[0,1]$, such that $\mathcal{P}_1^0$ decreases due to this marginal change too. However, we showed that the marginal change never decreases $\mathcal{P}_1^0$.

We conclude the proof by addressing the cases when the (11) can be binding. We already know that monotonicity holds everywhere in the interior, therefore the only remaining case that could violate the monotonicity is if $\mathcal{P}_1^0 = 1$, while $\hat{\mathcal{P}}_1^0 = 0$. In other words, if after an increase of value in some states the option comes from being selected with probability one to never being selected. However, this is not possible since the expected value of the option 1 increases. If $\mathcal{P}_1^0 = 1$, the agent processes no information and thus the expected utility equals the expectation of v_1. After the transformation of values of option 1, under $\hat{G}$, each strategy that ignores option 1 delivers the same utility as before the transformation, but the expected utility from selecting the option 1 with probability one is higher than before. Therefore, $\hat{\mathcal{P}}_1^0 = 1$. Strict monotonicity holds in the interior and weak monotonicity on the

boundaries. □.

C.2 Duplicates

Proof of Proposition 4. Let us consider a problem with $N+1$ options, where the options N and $N+1$ are duplicates. Let $\{\hat{\mathcal{P}}_i^0(\mathbf{u})\}_{i=1}^{N+1}$ be the unconditional probabilities in the solution to this problem. Since u_N and u_{N+1} are almost surely equal, then we can substitute u_N for u_{N+1} in the first order condition (9) to arrive at:

$$\hat{\mathcal{P}}_i(\mathbf{u}) = \frac{\hat{\mathcal{P}}_i^0 e^{u_i/\lambda}}{\sum_{j=1}^{N-1} \hat{\mathcal{P}}_j^0 e^{u_j/\lambda} + (\hat{\mathcal{P}}_N^0 + \hat{\mathcal{P}}_{N+1}^0) e^{u_N/\lambda}} \qquad a.s., \forall i < N \tag{43}$$

$$\hat{\mathcal{P}}_N(\mathbf{u}) + \hat{\mathcal{P}}_{N+1}^0(\mathbf{u}) = \frac{(\hat{\mathcal{P}}_N^0 + \hat{\mathcal{P}}_{N+1}^0) e^{u_i/\lambda}}{\sum_{j=1}^{N-1} \hat{\mathcal{P}}_j^0 e^{u_j/\lambda} + (\hat{\mathcal{P}}_N^0 + \hat{\mathcal{P}}_{N+1}^0) e^{u_N/\lambda}} \qquad a.s. \tag{44}$$

Therefore, the right hand sides do not change when only $\hat{\mathcal{P}}_N^0$ and $\hat{\mathcal{P}}_{N+1}^0$ change if their sum stays constant. Inspecting (43)-(44), we see that any such strategy produces the same expected value as the original one. Moreover, the amount of processed information is also the same for both strategies. To show this we use (9) to rewrite (6) as:[15]

$$\kappa = \int \sum_{i=1}^{N+1} \hat{\mathcal{P}}_i(\mathbf{u}) \log \frac{\hat{\mathcal{P}}_i(\mathbf{u})}{\hat{\mathcal{P}}_i^0} G(d\mathbf{u}) = \int \sum_{i=1}^{N+1} \hat{\mathcal{P}}_i(\mathbf{u}) \log \frac{e^{u_i/\lambda}}{\sum_{j=1}^{N-1} \hat{\mathcal{P}}_j^0 e^{u_j/\lambda} + (\hat{\mathcal{P}}_N^0 + \hat{\mathcal{P}}_{N+1}^0) e^{u_N/\lambda}} G(d\mathbf{u}). \tag{45}$$

Therefore, the achieved objective in (8) is the same for any such strategy as for the original strategy, and all of them solve the DM's problem.

Finally, even the corresponding strategy with $\hat{\mathcal{P}}_{N+1}^0 = 0$ is a solution. Moreover, this implies that the remaining $\{\hat{\mathcal{P}}_i^0\}_{i=1}^{N}$ is the solution to the problem without the duplicate option $N+1$, which completes the proof. □

[15] Here we use the fact that the mutual information between random variables X and Y can be expressed as $E_{p(x,y)}\left[\log \frac{p(x,y)}{p(x)p(y)}\right]$. See Cover and Thomas (2006, p. 20).

C.3 Similar options

Proof of Proposition 5. We proceed similarly as in the proof of Proposition 3 by showing that $\Delta_2 \equiv \frac{\partial E[U]}{\partial \mathcal{P}_1^0} + \frac{\partial E[U]}{\partial \mathcal{P}_2^0}$ decreases at all points $\{P_i^0\}_{i=1}^N$ after a marginal change of prior in the direction of interest. Notice that Δ_2 is a scalar product of the gradient of $E[U]$ and the vector $(1, 1, 0, \ldots, 0)$. We thus show that at each point, the gradient passes through each plane of constant $\mathcal{P}_1^0 + \mathcal{P}_2^0$ more in the direction of the negative change of $\mathcal{P}_1^0 + \mathcal{P}_2^0$ than before the change of the prior.

The analog to equation (39), after relocating $\epsilon\Pi$ probability from state 1 to 3 and from state 2 to 4, the sum of the left hand sides of the first order conditions for $i = 1$ and $i = 2$, becomes:

$$\begin{aligned}\Delta_2 =&\lambda \int_{\mathbf{v}} \frac{e^{v_1/\lambda} + e^{v_2/\lambda} - e^{v_N/\lambda}}{\sum_{j=1}^N \mathcal{P}_j^0 e^{v_j/\lambda}} G(d\mathbf{v}) \\ &+ \lambda\Pi\Big(\frac{(1-\epsilon)(e^{H/\lambda} + e^{L/\lambda} - e^{v_N/\lambda})}{\mathcal{P}_1^0 e^{H/\lambda} + \mathcal{P}_2^0 e^{L/\lambda} + a} + \frac{(1-\epsilon)(e^{H/\lambda} + e^{L/\lambda} - e^{v_N/\lambda})}{\mathcal{P}_1^0 e^{L/\lambda} + \mathcal{P}_2^0 e^{H/\lambda} + a} \\ &+ \frac{\epsilon(2e^{H/\lambda} - e^{v_N/\lambda})}{\mathcal{P}_1^0 e^{H/\lambda} + \mathcal{P}_2^0 e^{H/\lambda} + a} + \frac{\epsilon(2e^{L/\lambda} - e^{v_N/\lambda})}{\mathcal{P}_1^0 e^{L/\lambda} + \mathcal{P}_2^0 e^{L/\lambda} + a}\Big),\end{aligned} \tag{46}$$

where $a = \sum_{j=3}^N \mathcal{P}_j^0 e^{v_j/\lambda}$ is constant across the states 1-4, since for $j > 2$, v_j is constant there.

The analog to equation (41) when we differentiate $\Delta_2 = \frac{\partial E[U]}{\partial \mathcal{P}_1^0} + \frac{\partial E[U]}{\partial \mathcal{P}_2^0}$ with respect to ϵ is:

$$\begin{aligned}\left.\frac{\partial \Delta_2}{\partial \varepsilon}\right|_{\varepsilon=0} =&\lambda\Big(-\frac{e^{H/\lambda} + e^{L/\lambda} - e^{v_N/\lambda}}{\mathcal{P}_1^0 e^{H/\lambda} + \mathcal{P}_2^0 e^{L/\lambda} + a} - \frac{e^{H/\lambda} + e^{L/\lambda} - e^{v_N/\lambda}}{\mathcal{P}_1^0 e^{L/\lambda} + \mathcal{P}_2^0 e^{H/\lambda} + a} \\ &+ \frac{2e^{H/\lambda} - e^{v_N/\lambda}}{\mathcal{P}_1^0 e^{H/\lambda} + \mathcal{P}_2^0 e^{H/\lambda} + a} + \frac{2e^{L/\lambda} - e^{v_N/\lambda}}{\mathcal{P}_1^0 e^{L/\lambda} + \mathcal{P}_2^0 e^{L/\lambda} + a}\Big).\end{aligned} \tag{47}$$

Multiplying the right hand side by the positive denominators, the resulting expression

can be re-arranged to

$$
\begin{aligned}
&-\lambda(e^{H/\lambda}-e^{L/\lambda})^2 \\
&\Big[a^2(\mathcal{P}_1^0+\mathcal{P}_2^0)+e^{H/\lambda}e^{L/\lambda}(\mathcal{P}_1^0-\mathcal{P}_2^0)^2(\mathcal{P}_1^0+\mathcal{P}_2^0)+a(e^{H/\lambda}+e^{L/\lambda})((\mathcal{P}_1^0)^2+(\mathcal{P}_2^0)^2) \\
&+e^{v_N/\lambda}\mathcal{P}_1^0\mathcal{P}_2^0(2a+e^{H/\lambda}\mathcal{P}_1^0+e^{L/\lambda}\mathcal{P}_1^0+e^{H/\lambda}\mathcal{P}_2^0+e^{L/\lambda}\mathcal{P}_2^0)\Big]
\end{aligned}
$$

which is negative, and thus $\frac{\partial\Delta_2}{\partial\varepsilon}\big|_{\varepsilon=0}$ is negative, too. After the marginal relocation of probabilities that makes options 1 and 2 co-move more closely, the optimal $\mathcal{P}_1^0+\mathcal{P}_2^0$ decreases. The treatment of the boundary cases is analogous to that in the proof of Proposition 3. □

D Derivations for examples (For online publication)

D.1 Auxillary example

This is perhaps the simplest example of how rational inattention can be applied to a discrete choice situation. We present it here principally because this analysis forms the basis of our proof of Proposition 6 in Appendix D.3, but also because it provides some additional insight into the workings of the model.

Suppose there are two options, one of which has a known value while the other takes one of two values. One interpretation is that the known option is an outside option or reservation value.

Problem 5. *The DM chooses $i \in \{1,2\}$. The value of option 1 is distributed as $v_1 = 0$ with the probability g_0 and $v_1 = 1$ with the probability $1-g_0$. Option 2 carries the value $v_2 = R \in (0,1)$ with certainty.*

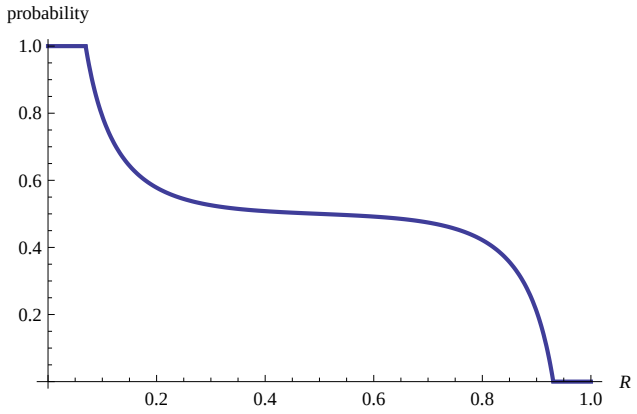

Figure 2: $\mathcal{P}_1^0$ as a function of R and $\lambda = 0.1, g_0 = 0.5$.

To solve the problem, we must find $\{\mathcal{P}_i^0\}_{i=1}^2$. We show below that the solution is:

$$\mathcal{P}_1^0 = \max\left(0, \min\left(1, -\frac{e^{\frac{R}{\lambda}}\left(-e^{\frac{1}{\lambda}} + e^{\frac{R}{\lambda}} - g_0 + g_0 e^{\frac{1}{\lambda}}\right)}{\left(e^{\frac{1}{\lambda}} - e^{\frac{R}{\lambda}}\right)\left(-1 + e^{\frac{R}{\lambda}}\right)}\right)\right) \tag{48}$$

$$\mathcal{P}_2^0 = 1 - \mathcal{P}_1^0.$$

For a given set of parameters, the unconditional probability $\mathcal{P}_1^0$ as a function of R is shown in Figure 2. For R close to 0 or to 1, the DM decides not to process information and selects one of the options with certainty. In the middle range however, the DM does process information and the selection of option 1 is less and less probable as the reservation value, R, increases, since option 2 is more and more appealing. For $g_0 = 1/2$ and $R = 1/2$, solutions take the form of the multinomial logit, i.e. $\mathcal{P}_1^0 = \mathcal{P}_2^0 = 1/2$. If the DM observed the values, he would choose option 1 with the probability $(1 - g_0) = 1/2$ for any reservation value R. However, the rationally inattentive agent chooses option 1 with higher probability when R is low.

Figure 3 again shows the dependance on R, but this time it presents the probability of selecting the first option *conditional* on the realized value $v_1 = 1$, it is $\mathcal{P}_1(1, R)$. Since $R < 1$, it would be optimal to always select the option 1 when its value is 1. The DM obviously does not choose to do that because he is not sure what the realized value is. When R is high, the DM processes less information and selects a low $\mathcal{P}_1^0$. As a result, $\mathcal{P}_1(1, R)$ is low.

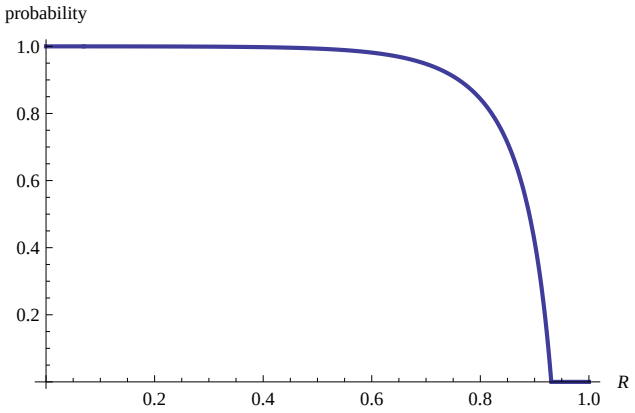

Figure 3: $\mathcal{P}_1(1, R)$ as a function of R and $\lambda = 0.1, g_0 = 0.5$.

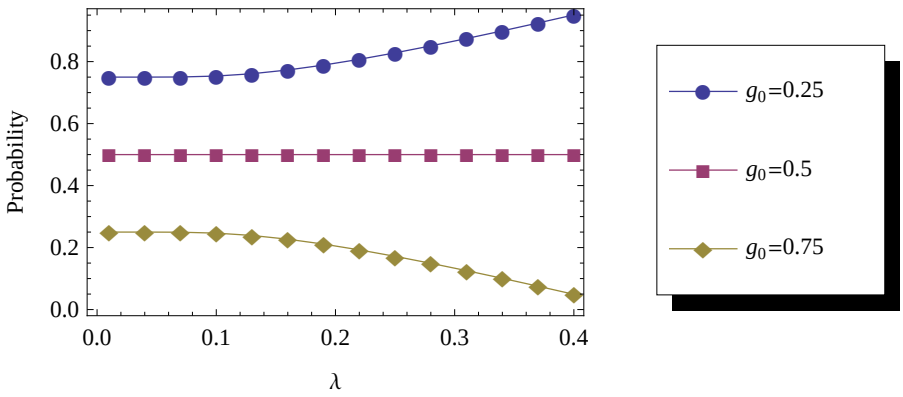

Figure 4: $\mathcal{P}_1^0$ as a function of λ evaluated at various values of g_0 and $R = 0.5$.

In general, one would expect that as R increases, the DM would be more willing to reject option 1 and receive the certain value R. Indeed, differentiating the non-constant part of (48) one finds that the function is non-increasing in R. Similarly, the unconditional probability of selecting option 1 falls as g_0 rises, as it is more likely to have a low value. Moreover, we see from equation (48) that, for $R \in (0,1)$, $\mathcal{P}_1^0$ equals 1 for g_0 in some neighborhood of 0 and it equals 0 for g_0 close to 1.[16] For these parameters, the DM chooses not to process information.

The following Proposition summarizes the immediate implications of equation (48). Moreover, the findings hold for any values of the uncertain option $\{a, b\}$ such that $R \in (a, b)$.

Proposition 7. *Solutions to Problem 5 have the following properties:*

1. *The unconditional probability of option 1, $\mathcal{P}_1^0$, is a non-increasing function of g_0 and the value R of the other option.*

2. *For all $R \in (0,1)$ and $\lambda > 0$, there exist g_m and g_M in $(0,1)$ such that if $g_0 \leq g_m$, the DM does not process any information and selects option 1 with probability one. Similarly, if $g_0 \geq g_M$, the DM processes no information and selects option 2 with probability one.*

Figure 4 plots $\mathcal{P}_1^0$ as a function of the information cost λ for three values of the prior, g_0. When $\lambda = 0$, $\mathcal{P}_1^0$ is just equal to $1 - g_0$ because the DM will have perfect knowledge of the value of option 1 and choose it when it has a high value, which occurs with probability $1 - g_0$. As λ increases, $\mathcal{P}_1^0$ fans out away from $1 - g_0$ because the DM no longer possesses perfect knowledge about the value of option 1 and eventually just selects the option with the higher expected value according to the prior.

Solving for the choice probabilities in Problem 5. To solve the problem, we must find $\mathcal{P}_1^0$, while $\mathcal{P}_2^0 = 1 - \mathcal{P}_1^0$. These probabilities must satisfy the normalization condition, equation

[16] The non-constant argument on the right-hand side of (48) is continuous and decreasing in g_0, and it is greater than 1 at $g_0 = 0$ and negative at $g_0 = 1$.

(13):

$$1 = \frac{g_0}{\mathcal{P}_1^0 + \mathcal{P}_2^0 e^{\frac{R}{\lambda}}} + \frac{(1-g_0)e^{\frac{1}{\lambda}}}{\mathcal{P}_1^0 e^{\frac{1}{\lambda}} + \mathcal{P}_2^0 e^{\frac{R}{\lambda}}} \qquad \text{if } \mathcal{P}_1^0 > 0, \tag{49}$$

$$1 = \frac{g_0 e^{\frac{R}{\lambda}}}{\mathcal{P}_1^0 + \mathcal{P}_2^0 e^{\frac{R}{\lambda}}} + \frac{(1-g_0)e^{\frac{R}{\lambda}}}{\mathcal{P}_1^0 e^{\frac{1}{\lambda}} + \mathcal{P}_2^0 e^{\frac{R}{\lambda}}} \qquad \text{if } \mathcal{P}_2^0 > 0. \tag{50}$$

There are three solutions to this system,

$$\begin{aligned} \mathcal{P}_1^0 &\in \left\{ 0, 1, -\frac{e^{\frac{R}{\lambda}}\left(-e^{\frac{1}{\lambda}} + e^{\frac{R}{\lambda}} - g_0 + g_0 e^{\frac{1}{\lambda}}\right)}{\left(e^{\frac{1}{\lambda}} - e^{\frac{R}{\lambda}}\right)\left(-1 + e^{\frac{R}{\lambda}}\right)} \right\} \\ \mathcal{P}_2^0 &= 1 - \mathcal{P}_1^0. \end{aligned} \tag{51}$$

Now, we make an argument using the solution's uniqueness to deduce the true solution to the DM's problem. The first solution to the system, $\mathcal{P}_1^0 = 0$, corresponds to the case when the DM chooses option 2 without processing any information. The realized value is then R with certainty. The second solution, $\mathcal{P}_1^0 = 1$, results in the *a priori* selection of option 1 so the expected value equals $(1 - g_0)$. The third solution describes the case when the DM chooses to process a positive amount of information.

In Problem 5, there are just two options and they do not take the same values with probability one. Therefore, Corollary 3 establishes that the solution to the DM's optimization problem must be unique.

Since the expected utility is a continuous function of $\mathcal{P}_1^0, R, \lambda$ and g_0, then the optimal $\mathcal{P}_1^0$ must be a continuous function of the parameters. Otherwise, there would be at least two solutions at the point of discontinuity of $\mathcal{P}_1^0$. We also know that, when no information is processed, option 1 generates higher expected utility than option 2 for $(1 - g_0) > R$, and vice versa. So for some configurations of parameters $\mathcal{P}_1^0 = 0$ is the solution and for some configurations of parameters $\mathcal{P}_1^0 = 1$ is the solution. Therefore, the solution to the DM's problem has to include the non-constant branch, the third solution. To summarize this, the

only possible solution to the DM's optimization problem is

$$\mathcal{P}_1^0 = \max\left(0, \min\left(1, -\frac{e^{\frac{R}{\lambda}}\left(-e^{\frac{1}{\lambda}} + e^{\frac{R}{\lambda}} - g_0 + g_0 e^{\frac{1}{\lambda}}\right)}{\left(e^{\frac{1}{\lambda}} - e^{\frac{R}{\lambda}}\right)\left(-1 + e^{\frac{R}{\lambda}}\right)}\right)\right). \tag{52}$$

D.2 Problem 3

To find the solution to Problem 3 we must solve for $\{\mathcal{P}_r^0, \mathcal{P}_b^0, \mathcal{P}_t^0\}$. The normalization condition $\mathcal{P}_r^0 = \int_{\mathbf{v}} \mathcal{P}_r(\mathbf{v}) G(d\mathbf{v})$ yields:

$$\begin{aligned} 1 = & \frac{\frac{1}{4}(1+\rho)}{\mathcal{P}_r^0 + \mathcal{P}_b^0 + (1 - \mathcal{P}_r^0 - \mathcal{P}_b^0)e^{1/2\lambda}} + \frac{\frac{1}{4}(1-\rho)\, e^{1/\lambda}}{\mathcal{P}_r^0 e^{1/\lambda} + \mathcal{P}_b^0 + (1 - \mathcal{P}_r^0 - \mathcal{P}_b^0)e^{1/2\lambda}} \\ + & \frac{\frac{1}{4}(1-\rho)}{\mathcal{P}_r^0 + \mathcal{P}_b^0 e^{1/\lambda} + (1 - \mathcal{P}_r^0 - \mathcal{P}_b^0)e^{1/2\lambda}} + \frac{\frac{1}{4}(1+\rho)\, e^{1/\lambda}}{\mathcal{P}_r^0 e^{1/\lambda} + \mathcal{P}_b^0 e^{1/\lambda} + (1 - \mathcal{P}_r^0 - \mathcal{P}_b^0)e^{1/2\lambda}} \end{aligned} \tag{53}$$

Due to the symmetry between the buses, we know $\mathcal{P}_r^0 = \mathcal{P}_b^0$. This makes the problem one equation with one unknown, $\mathcal{P}_r^0$. The problem can be solved analytically using the same arguments as in Appendix D.1. The resulting analytical expression is:

$$\mathcal{P}_r^0 = \max\left(0, \min\left(0.5, \frac{\left(\begin{array}{l} e^{\frac{1}{2\lambda}} - 8e^{\frac{1}{\lambda}} + 14e^{\frac{3}{2\lambda}} - 8e^{2/\lambda} + e^{\frac{5}{2\lambda}} \\ +\frac{1}{2}e^{\frac{1}{2\lambda}}(1-\rho) - e^{\frac{3}{2\lambda}}(1-\rho) + \frac{1}{2}e^{\frac{5}{2\lambda}}(1-\rho) \\ +e^{\frac{1}{2\lambda}}\left(-1 + e^{\frac{1}{\lambda}}\right)x \end{array}\right)}{2\left(4e^{\frac{1}{2\lambda}} - 16e^{\frac{1}{\lambda}} + 24e^{\frac{3}{2\lambda}} - 16e^{2/\lambda} + 4e^{\frac{5}{2\lambda}}\right)}\right)\right),$$

where

$$x = \sqrt{\begin{array}{l} 2 - 2e^{\frac{1}{\lambda}} + e^{2/\lambda} - 8e^{\frac{1}{2\lambda}}(1-\rho) + 14e^{\frac{1}{\lambda}}(1-\rho) \\ -8e^{\frac{3}{2\lambda}}(1-\rho) + e^{2/\lambda}(1-\rho) + \frac{1}{4}(1-\rho)^2 \\ -\frac{1}{2}e^{\frac{1}{\lambda}}(1-\rho)^2 + \frac{1}{4}e^{2/\lambda}(1-\rho)^2 - \rho \end{array}}.$$

D.3 Inconsistency with a random utility model

This appendix establishes that the behavior of the rationally inattentive agent is not consistent with a random utility model. The argument is based on the counterexample described in section 5.3.2. Let Problem A refer to the choice among options 1 and 2 and Problem B refer to the choice among all three options. For simplicity, $\mathcal{P}_i(s)$ denotes the probability of selecting option i conditional on the state s, and $g(s)$ is the prior probability of state s.

Lemma 5. *For all $\epsilon > 0$ there exists Y s.t. the DM's strategy in Problem B satisfies*

$$\mathcal{P}_3(1) > 1 - \epsilon, \qquad \mathcal{P}_3(2) < \epsilon.$$

Proof: For $Y > 1$, an increase of $\mathcal{P}_3(1)$ (decrease of $\mathcal{P}_3(2)$) and the corresponding relocation of the choice probabilities from (to) other options increases the agent's expected payoff. The resulting marginal increase of the expected payoff is larger than $(Y-1)\min(g(1), g(2))$. Selecting Y allows us to make the marginal increase arbitrarily large and therefore the marginal value of information arbitrarily large.

On the other hand, with λ being finite, the marginal change in the cost of information is also finite as long as the varied conditional probabilities are bounded away from zero. See equation (6), the derivative of entropy with respect to $\mathcal{P}_i(s)$ is finite at all $\mathcal{P}_i(s) > 0$. Therefore, for any ϵ there exists high enough Y such that it is optimal to relocate probabilities from options 1 and 2 unless $\mathcal{P}_3(1) > 1 - \epsilon$, and to options 1 and 2 unless $\mathcal{P}_3(2) < \epsilon$. □

Proof of proposition 6. We will show that there exist $g(1) \in (0, 1)$ and $Y > 0$ such that option 1 has zero probability of being selected in Problem A, while the probability is positive in both states in Problem B. Let us start with Problem A. According to Proposition 7, there exists a sufficiently high $g(1) \in (0, 1)$, call it g_M, such that the DM processes no information and $\mathcal{P}_1(1) = \mathcal{P}_1(2) = 0$. We will show that for $g(1) = g_M$ there exists a high enough Y, such the choice probabilities of option 1 are positive in Problem B.

Let $\mathcal{P} = \{\mathcal{P}_i(s)\}_{i=1,s=1}^{3,2}$ be the solution to Problem B. We now show that the optimal choice probabilities of options 1 and 2, $\{\mathcal{P}_i(s)\}_{i=1,s=1}^{2,2}$, solve a version of Problem A with

modified prior probabilities. The objective function for Problem B is

$$\max_{\{\mathcal{P}_i(s)\}_{i=1,s=1}^{3,2}} \sum_{i=1}^{3}\sum_{s=1}^{2} v_i(s)\mathcal{P}_i(s)g(s) - \lambda\left[-\sum_{s=1}^{2} g(s)\log g(s) + \sum_{i=1}^{3}\sum_{s=1}^{2}\mathcal{P}_i(s)g(s)\log\frac{\mathcal{P}_i(s)g(s)}{\sum_{s'}\mathcal{P}_i(s')g(s')}\right], \tag{54}$$

where we have written the information cost as $H(s)-E[H(s|i)]$.[17] If $\mathcal{P}_3(1)$ and $\mathcal{P}_3(2)$ are the conditional probabilities of the solution to Problem B, the remaining conditional probabilities solve the following maximization problem.

$$\max_{\{\mathcal{P}_i(s)\}_{i=1,s=1}^{2,2}} \sum_{i=1}^{2}\sum_{s=1}^{2} v_i(s)\mathcal{P}_i(s)g(s) - \lambda\left[\sum_{i=1}^{2}\sum_{s=1}^{2}\mathcal{P}_i(s)g(s)\log\frac{\mathcal{P}_i(s)g(s)}{\sum_{s'}\mathcal{P}_i(s')g(s')}\right], \tag{55}$$

subject to $\mathcal{P}_1(s)+\mathcal{P}_2(s) = 1-\mathcal{P}_3(s)$, $\forall s$. Equation (55) is generated from (54) by omitting the terms independent of $\{\mathcal{P}_i(s)\}_{i=1,s=1}^{2,2}$. Now, we make the following substitution of variables.

$$\mathcal{R}_i(s) = \mathcal{P}_i(s)/\Big(1-\mathcal{P}_3(s)\Big) \tag{56}$$

$$\hat{g}(s) = Kg(s)\Big(1-\mathcal{P}_3(s)\Big) \tag{57}$$

$$1/K = \sum_{s=1}^{2} g(s)\Big(1-\mathcal{P}_3(s)\Big). \tag{58}$$

where K, which is given by (58), is the normalization constant that makes the new prior, $\hat{g}(s)$, sum up to 1.

The maximization problem (55) now takes the form:

$$\max_{\{\mathcal{R}_i(s)\}_{i=1,s=1}^{1,2}} \sum_{i=1}^{2}\sum_{s=1}^{2} v_i(s)\mathcal{R}_i(s)\hat{g}(s) - \lambda\sum_{i=1}^{2}\sum_{s=1}^{2}\mathcal{R}_i(s)\hat{g}(s)\log\frac{\mathcal{R}_i(s)\hat{g}(s)}{\sum_{s'}\mathcal{R}_i(s')\hat{g}(s')}, \tag{59}$$

[17] Recall that $H(Y|X) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x)$ (Cover and Thomas, 2006, p. 17).

subject to

$$\mathcal{R}_1(s) + \mathcal{R}_2(s) = 1 \quad \forall s. \tag{60}$$

The objective function of this problem is equivalent to (55) up to a factor of K, which is a positive constant. The optimization problem (59) subject to (60) is equivalent to Problem A with the prior modified to from $g(s)$ to $\hat{g}(s)$, let us call it Problem C.[18]

According to Proposition 7, there exists $\hat{g}_m \in (0,1)$ such that the DM always selects option 1 in Problem C for all $\hat{g}(1) \leq \hat{g}_m$. From equations (57) and (58) we see that for any $\hat{g}_m > 0$ and $g(1)$, $g(2) \in (0,1)$ there exists $\epsilon > 0$ such that if $\mathcal{P}_3(1) > 1-\epsilon$ and $\mathcal{P}_3(2) < \epsilon$, then $\hat{g}(1) < \hat{g}_m$.[19] Moreover, Lemma 5 states that for any such $\epsilon > 0$ there exists Y such that $\mathcal{P}_3(1) > 1-\epsilon$ and $\mathcal{P}_3(2) < \epsilon$. Therefore there is a Y such that in Problem C, option 1 is selected with positive probability in both states, which also implies it is selected with positive probabilities in Problem B, see equation (56). □

[18] To see the equivalence to Problem A, observe that this objective function has the same form as (54) except for a) the constant corresponding to $H(s)$ and b) we only sum over $i = 1, 2$.

[19] $\hat{g}(1) = \frac{g(1)(1-\mathcal{P}_3(1))}{\sum_s g(s)(1-\mathcal{P}_3(s))} < \frac{g(1)(1-\mathcal{P}_3(1))}{g(2)(1-\mathcal{P}_3(2))} < \frac{g(1)\epsilon}{g(2)(1-\epsilon)}$.