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

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

February 22, 2011

Abstract

We consider a discrete choice problem using the rational inattention approach to information frictions. The rationally inattentive agent chooses how to learn about the unknown values of the available options to maximize the expected value of the chosen option less an information cost. We solve the model analytically and find that if the agent views the options as equivalent *a priori*, then the agent chooses probabilistically according to the multinomial logit model, which is widely used to study discrete choices. When the options are not symmetric *a priori*, the agent incorporates prior knowledge of the options into the choice in a simple fashion that can be interpreted as a multinomial logit in which the *a priori* attractiveness of an options shifts its perceived value. Unlike the multinomial logit, this model predicts that duplicate options are treated as a single option.

[†] We thank Christopher Sims, Per Krusell, and Christian Hellwig for helpful discussions at various stages of this project.

[*]Center for Economic Research and Graduate Education, Prague. filip.matejka@cerge-ei.cz

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

1 Introduction

At times, individuals must choose among discrete alternatives with imperfect information about the value of each alternative. Before making a choice, one often has an opportunity to study the options. In most cases, however, it is too costly to investigate the options to the point where their values are known with certainty and, as a result, some uncertainty about the values remains when one chooses among the options. 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. That is, 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. For example, a firm might choose how long to spend interviewing candidates for a job and 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 a 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 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 they find useful and ignore information that is not worth the effort of acquiring and processing.

Our main result is that the resulting choice probabilities follow the multinomial logit formula according to which the DM selects option $i \in \{1, \cdots, N\}$ with probability

$$\frac{e^{q_i/\lambda}}{\sum_{j=1}^{N} e^{q_j/\lambda}},$$

where q_i is the value of option i and λ is a scale parameter. In the multinomial logit formula, the scale parameter controls the amount of noise in the choice probabilities. As λ goes to zero, the DM selects the option with the highest value with probability one. As λ goes to infinity, the DM selects all options with probability $1/N$. In our work, the scale parameter is exactly equal to the marginal cost of a unit of information.

The multinomial logit model is the most commonly used model of discrete choice behavior and is one of the principal tools of applied researchers studying discrete choices (McFadden, 1974). The multinomial logit is also used in industrial organization as a model of consumer demand (Anderson et al., 1992) and it is used in experimental economics to capture an element of bounded rationality in subject behavior (McKelvey and Palfrey, 1995). It is so widely used because it is particularly tractable both analytically and computationally and because it has a connection to consumer theory through a random utility model.

We distinguish between two cases depending on what the DM knows about the options before acquiring information. If the DM views the options symmetrically *a priori*, then the he or she chooses according to the multinomial logit formula above. In other cases, the DM might have some prior information that will inform his or her choice. We therefore also analyze the more general

model in which the DM incorporates prior knowledge of the options into the choice. In this context, we arrive at a model that can be interpreted as a multinomial logit in which the value of each option is shifted by an amount that reflects its *a priori* attractiveness. Importantly, these results do not depend on any distributional assumptions about the DM's prior knowledge.

There are two canonical derivations of the multinomial logit.[1] First, the multinomial logit can be derived from Luce's (1959) Choice Axiom, which we discuss below. Second, the multinomial logit can be derived from a random utility model. According to that derivation, the DM evaluates the options with some noise either due to randomness in his or her evaluation of the alternatives or due to some unobserved factor that is known to the agent, but unknown to the economic modeler. If the noise in the evaluation is additively separable and independently distributed according to the extreme value distribution, then the multinomial logit model emerges.[2] The extreme-value distributed noise is often interpreted in terms of idiosyncratic tastes that are not observable to the analyst, but it is also sometimes interpreted as errors of perception (e.g. McFadden, 1980, p. S15).

We believe our findings are important for two reasons. First, the wide popularity of the multinomial logit is in part due to its connection to the random utility model. This foundation allows researchers to interpret the choice probabilities in terms of optimizing behavior. 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 is not a fully specified model of those information frictions that justifies the use of the multinomial logit. We fill that gap. Second, most existing work with rational inattention has focussed on situations with a continuous action.[3] In this context, the model remains

[1] See McFadden (1976) and Anderson et al. (1992) for surveys.

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

[3] Rational inattention has mostly been applied in macroeconomic contexts. The major applications have been consumption-savings problems (Sims, 2006; Luo, 2008; Tutino, 2009), price setting (Mackowiak and Wiederholt, 2009; Woodford, 2009; Matejka, 2010a,b), and portfolio choices (Van Nieuwerburgh and Veldkamp, 2010; Mondria,

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. In contrast, we show here that the discrete-choice version of the rational inattention model is extremely tractable. Because discrete choices arise in many economic contexts, we expect that these results will be useful in applying the rational inattention framework to a number of questions.

Using our new derivation of the logit model we can view its features and flaws from a different perspective. For example, we can address Debreu's (1960) well-known criticism of the multinomial logit. Debreus's critique is best presented in the form of an example: The agent is confronted with a choice between a yellow bus and a train and selects each with probability 1/2. If a red bus is introduced to the choice set and the agent is thought to be indifferent between the two buses then it makes sense to think that each bus is equally likely to be selected. Then it follows from the multinomial logit that each of the buses and the train is selected with probability 1/3. Debreu argued that this is counterintuitive because duplicating one option should not materially change the choice problem. We formalize the notion of duplicate options as a scenario in which two options have perfectly correlated values. That is, the prior is asymmetric because the correlation between two options is higher than it is between other pairs of options. Given this asymmetry in the prior, we no longer expect the multinomial logit to hold exactly, but rather the values should be adjusted for their *a priori* attractiveness. We show that in this situation, the rationally inattentive agent does not display the counterintuitive behavior that Debreu criticized. In particular, we show that a rationally inattentive agent treats two duplicate options as a single option. In the bus example,

2010).

the DM chooses each bus with probability 1/4 and the train with probability 1/2.

When Debreu presented the critique that is now known as the bus problem, he was responding to Luce's Choice Axiom from which the multinomial logit follows. The central idea of the Choice Axiom is the property of independence of irrelevant alternatives (IIA), which states that the ratio of choice probabilities associated with two options should not depend on the other options that are available. Given that we have just claimed that the rationally inattentive agent violates this property in the bus problem, one might reasonably ask how our work still generates a multinomial logit. The answer is that when the prior is asymmetric the values of the options are adjusted for their *a priori* attractiveness. This adjustment breaks the IIA property.

The rational inattention framework uses information theory (Shannon, 1948) to measure the amount of information that the DM acquires about the choice environment. In information theory, a probability distribution is associated with an amount of entropy. If one acquires some information that leads to a reduction in uncertainty, the entropy of the posterior distribution is smaller than the entropy of the prior distribution. The reduction in entropy is a measure of the amount of information that has been acquired. We note that there is a close connection between Shannon's axiomatic derivation of entropy as a natural measure of information and Luce's (1959) Choice Axiom, which can be used to derive the multinomial logit. An equivalent statement of the Choice Axiom is that the choice probabilities should not change if one chooses through a two-stage procedure in which one first chooses a subset of the options and then makes a selection from that subset as opposed to a one-stage procedure in which one directly makes a selection from the full set of options. In Shannon's (1948) seminal paper, he actually discusses the amount of information conveyed by a choice. He places three requirements on the measure of information communicated by choosing an element from a set of discrete alternatives. Among these requirements is the requirement that

it should be irrelevant if one first chooses a subset of the options and then chooses from that subset.[4] The parallel between the Choice Axiom and Shannon's axiom is clear and Luce comments on this similarity soon after stating the Choice Axiom. As a result of this connection, our rationally inattentive agent will not gain anything from violating the Choice Axiom because the amount of information acquired is the same regardless of whether the agent makes the selection in two stages or one.

We are not the first to propose that information frictions can explain discrete choice behavior. Recently, Natenzon (2010) has proposed a model in which the DM has Gaussian priors on the utilities of the options and then receives Gaussian signals about the utilities. As more signals are collected, the DM updates his or her posterior and when forced to make a choice, selects the option with the highest posterior mean utility. Natenzon calls this model the Bayesian Probit. The difference between his approach and ours is that we consider an agent who is actively seeking out the most important signals about the alternatives while Natenzon's agent is responding to an exogenous flow of information.

Finally, while various connections between Shannon's concept of entropy and the multinomial logit model have been known for some time, we believe the link we make here is new. Specifically, the use of information theory to formulate an information flow constraint on an individual's decision problem does not appear before Sims's work on rational inattention and we are not aware of any previous attempts to consider discrete choices in this framework. Perhaps the closest line of reasoning is in the game theory literature where Stahl (1990) and Mattsson and Weibull (2002) consider the problem of a player who has difficulty implementing his or her selected strategy. Players must exert effort to steady their trembling hands and the amount of trembling is measured

[4]See Theorem 2 in Shannon (1948).

using information theory. A multinomial logit emerges when the players optimally allocate effort to avoiding the most costly trembles. Our work differs from these papers in that we explicitly link the agent's difficulty in implementing the "correct" decision to the cost of processing information about the options.[5]

We now turn to a presentation of the problem faced by the rationally inattentive agent. We then analyze the model's predictions for a generic prior in section 3. Section 4 considers the case when the options are *a priori* symmetric and provides the connection to the multinomial logit. In section 5, we discuss cases where the options are not symmetric *a priori*, including the possibility that two of the options are duplicates.

2 The Model

The DM is presented with a group of N options. The values of these options potentially differ and the agent wishes to select the option with the highest value. Let q_i denote the value of the selected option $i \in \{1..N\}$. Initially,the agent possesses some knowledge about the available option and this prior knowledge can be described by a joint probability distribution with a pdf $g(\overrightarrow{q})$, where $\overrightarrow{q} = (q_1, .., q_N)$ is the vector of values of the N options.

Following the rational inattention approach to information frictions, we assume that information about the N options is available to the DM, but processing the information is costly. If the DM could process information costlessly, he or she would select the best available option. With costly information acquisition the DM must choose how much and which information to acquire and process. Formally, we follow Sims and quantify the amount of information processed using

[5] Anas (1983) made a more statistical connection between entropy and the multinomial logit. He showed that the estimation of a multinomial logit can be viewed as the solution to a problem of maximizing the entropy of the selection probabilities for the individual choices subject to the constraint that the aggregate behavior predicted by those probabilities match the observed aggregates.

information theory. A random variable is associated with a level of entropy, which measures the amount of information that is conveyed when the random variable is realized. In our setting, the prior g has a level of entropy and after processing information the DM has a posterior distribution over the values of the available options, call it g^*. On average, g^* has a smaller level of entropy than g because some uncertainty has been resolved through information processing. With the posterior distribution in mind, the DM makes his or her choice.

For a random variable X with density function f, the entropy is given by

$$H[f(X)] = -\int f(x) \log f(x) dx. \tag{1}$$

Now suppose the DM receives a signal, Y, about X. The DM's knowledge of X is now given by the conditional distribution $f(X|Y)$ and the entropy of this distribution is $H[f(X|Y)]$. The amount of information processed is given by the reduction in entropy $H[f(X)] - H[f(X|Y)] \equiv I(X;Y)$. The quantity $I(X;Y)$ is referred to as the mutual information between X and Y. If Y is informative about X, $I(X;Y)$ will be positive. While particular signals may make the DM less certain about X and therefore lead to a higher level of entropy, on average these signals will lead to more precise knowledge of X and lower levels of entropy.[6]

While there is an intuitive appeal to thinking of the DM as asking for a signal about the unknown values and then choosing an option conditional on that signal, the rational inattention approach abstracts from the signals and models a joint probability distribution between the true values and the DM's action. For example, the DM might receive a signal y about the values $\overrightarrow{q}$ and then implement a choice, i, as some function $h(y)$. As $h(\cdot)$ is a deterministic function of y, the joint distribution between y and $\overrightarrow{q}$ then generates a joint distribution between $\overrightarrow{q}$ and the choice, i.

[6]See Cover and Thomas (2006) for further discussion of these concepts.

The explicit treatment of signals, however, is not necessary and the rational inattention approach abstracts from signals and works with the joint distribution between $\overrightarrow{q}$ and i. We can describe this joint distribution by a collection $\{\mathcal{P}_i^0, f(\overrightarrow{q}|i)\}_{i=1}^N$, where $f(\overrightarrow{q}|i)$ is the distribution of the true values conditional on option i being selected and $\mathcal{P}_i^0$ is the marginal probability of selecting option i. Another way of looking at these probabilities is that $\mathcal{P}_i^0$ is the DM's subjective probability of selecting i before processing information and $f(\overrightarrow{q}|i)$ is the DM's posterior on $\overrightarrow{q}$ conditional on selecting option i. In total, this collection describes the joint distribution of the agent's choice and the vector $\overrightarrow{q}$.

Our DM faces a cost of processing information that is quantified in terms of the reduction in entropy. The decision making process can be thought of as a series of question that the DM asks. The number and types of questions that the DM asks and the accuracy with which the DM determines the answers to the questions generates a posterior distribution and a resulting entropy reduction. This formulation of the information processing cost is meant to capture the fact that it takes time and effort to carefully study the available options. The DM maximizes the expected value of the option that he or she selects less the quantity $\lambda I(\overrightarrow{q}; i)$, where λ is a scalar that controls the degree of the information friction. $I(\overrightarrow{q}; i)$ is the mutual information between the true values and the selected option. If the DM carefully studies the options before making a choice, then observing which option the DM selects, i, provides relatively more information about what the options are than it would if the DM acquired less information before choosing. As such, when the DM processes more information, the mutual information between i and $\overrightarrow{q}$ rises.

One might ask how the cost of information should be interpreted. Sims (2010) argues that a person has a finite amount of attention—or capacity for processing information—to devote to a number of things. As such, the parameter λ reflects the shadow cost of allocating attention to the

decision that we are considering.

We can now state the DM's optimization problem .

$$\max_{\{\mathcal{P}_i^0, f(\overrightarrow{q}|i)\}_{i=1}^N, \kappa} \sum_i \mathcal{P}_i^0 \int_{\overrightarrow{q}} q_i f(\overrightarrow{q}|i) d\overrightarrow{q} - \lambda\kappa, \tag{2}$$

subject to

$$\begin{aligned} I(\overrightarrow{q}; i) &\leq \kappa & (3)\\ \sum_i \mathcal{P}_i^0 f(\overrightarrow{q}|i) &= g(\overrightarrow{q}) \qquad \forall \overrightarrow{q} & (4)\\ \sum_i \mathcal{P}_i^0 &= 1, \mathcal{P}_i^0 \in [0,1] \quad . \end{aligned}$$

Equation (3) limits how much the agent can find out about the options by processing the selected amount of information, κ. Equation (4) states that posterior knowledge has to be consistent with the DM's prior. If this constraint were omitted, the DM could raise his or her expected utility by selecting a probability distribution that places a large weight on high values even if the agent knows (according to the prior) that this is not the case. Readers who are familiar with rational inattention will recognize this problem as a standard information-constrained, static optimization problem very similar to the generic example presented by Sims (2010, p. 162). The only difference is that here the DM is choosing over a discrete set of actions.

An alternative modeling assumption would be to assume that the DM has a fixed capacity for information processing to devote to the decision. In that case, κ would not be a choice variable but an exogenous parameter and λ would be the Lagrange multiplier associated with the constraint (3).

3 Solving the model

Let us study the case of $\lambda > 0$, solutions for $\lambda = 0$ are trivial since the perfectly attentive DM simply selects the option(s) of the highest value with the probability one. We show in Appendix A that the first order condition for the DM's choice of $f(\overrightarrow{q}|i)$ is

$$f(\overrightarrow{q}|i) = h(\overrightarrow{q})e^{\frac{q_i}{\lambda}} \qquad \forall i; \quad \mathcal{P}_i^0 > 0 \tag{5}$$

where h is a function of Lagrange multipliers on the prior, (4). Plugging (5) into (4) we get

$$h(\overrightarrow{q}) = \frac{g(\overrightarrow{q})}{\sum_{i=1}^{N} \mathcal{P}_i^0 e^{\frac{q_i}{\lambda}}}. \tag{6}$$

The first order condition (5) thus takes the following form.

$$f(\overrightarrow{q}|i) = \frac{g(\overrightarrow{q})e^{\frac{q_i}{\lambda}}}{\sum_{i=1}^{N} \mathcal{P}_i^0 e^{\frac{q_i}{\lambda}}}. \tag{7}$$

Since f is a pdf, it satisfies

$$\begin{aligned} 1 &= \int f(\overrightarrow{q}|i) d\overrightarrow{q}, \\ 1 &= \int \frac{g(\overrightarrow{q})e^{\frac{q_i}{\lambda}}}{\sum_{j=1}^{N} \mathcal{P}_j^0 e^{\frac{q_j}{\lambda}}} d\overrightarrow{q}, \end{aligned} \tag{8}$$

for all $\mathcal{P}_i^0 > 0$. As we demonstrate below, this normalization condition can be useful in characterizing the solution when the prior is asymmetric.

Given a set of values, the probability of selecting i is the following conditional probability

$$\mathcal{P}(i|\overrightarrow{q}) = \frac{\mathcal{P}_i^0 f(\overrightarrow{q}|i)}{g(\overrightarrow{q})}. \tag{9}$$

From now on, we denote $\mathcal{P}(i|\overrightarrow{q})$ as $\mathcal{P}_i(\overrightarrow{q})$. Plugging (7) into (9) we get

$$\mathcal{P}_i(\overrightarrow{q}) = \frac{\mathcal{P}_i^0 e^{q_i/\lambda}}{\sum_{j=1}^{N} \mathcal{P}_j^0 e^{q_j/\lambda}}, \tag{10}$$

We can now state the first result.

Theorem 1. *Let a rationally inattentive agent be presented with N options and maximize the expected utility, which is the expected value of the selected option minus the cost of processing information, $g(\overrightarrow{q})$ be a pdf describing his prior knowledge of the options' values and $\lambda > 0$ be the unit cost of information. Then, the probability of choosing option i as a function of the realized values the options, is given by* (10)*. If $\lambda = 0$, then the perfectly attentive DM simply selects the option(s) with the highest value.*

What is left to fully solve the agent's problem is to find the unconditional probabilities of selecting each option, $\{\mathcal{P}_i^0\}_{i=1}^N$. These probabilities are independent of a specific realization of values $\overrightarrow{q}$, they are the marginal probabilities of selecting each option before the agent starts processing any information and they depend only on $g(\cdot)$ and λ.

If we omit the $\mathcal{P}_i^0$ terms from equation (10) we have the usual multinomial logit formula. The implication is that the relative probability of selecting i is not driven just by $e^{q_i/\lambda}$, as in the logit case, but also by the prior probability of selection option i, $\mathcal{P}_i^0$.

The prior probability on option i depends on the value of q_i relative to the values of other

options. For example, if option i is likely to have a high value, but sure to be dominated by another option, then $\mathcal{P}_i^0$ will be zero. Conversely, an option might have an extremely low expected value but with some probability have the highest value in the choice set and therefore have a positive prior probability.

The dependence of the model on the cost of information, λ, is very intuitive. As information processing becomes more costly the DM processes less and the selection probabilities depend less on the actual realization of values and more on the prior $\mathcal{P}_i^0$. Simply put, the less information is processed the more prior knowledge enters in the DM's decision. On the other hand, as λ falls, the DM processes more information and in the extreme, as $\lambda \to 0$, the DM selects the option with the highest value with probability one, which is to say that all uncertainty about which option is best is resolved.

A fairly obvious, but important, point is that λ converts bits of information to utils. Therefore, if one scales the utility function by a constant c, one must also scale λ by the same factor for consistency. Of course, if the the utility levels are scaled up because the stakes are higher (at a fixed λ) the selection probabilities change in a manner equivalent to a reduction in the information friction (scaling λ down by $1/c$). The reason is that the DM chooses to process more information when more is at stake and thus makes less error in selecting the best option.

Finally, we offer an alternative way of interpreting equation (10), which we can rewrite as

$$\mathcal{P}_i(\overrightarrow{q}) = \frac{e^{(q_i+v_i)/\lambda}}{\sum_{j=1}^{N} e^{(q_j+v_j)/\lambda}}, \tag{11}$$

where $v_i = \lambda \log\left(\mathcal{P}_i^0\right)$. Written this way, the selection probabilities can be interpreted as a multinomial logit in which the value of option i is shifted by the term v_i. v_i reflects the *a priori*

attractiveness of option i as measured by the prior probability that the option is selected. As the cost of information, λ, rises, the weight on the prior rises. Notice that the choice behavior generated by the multinomial logit does not depend on the location of utilities, but only the differences between utilities. Therefore, the relevant feature of the v_i terms is not their level, but how they differ across options.

3.1 Independence of Irrelevant Alternatives

Unlike the multinomial logit, the rationally inattentive agent's choice probabilities do not generally have the property of independence of irrelevant alternatives (IIA). IIA states that the ratio of the selection probabilities for two alternatives is independent of what other alternatives are included in the choice set. According to equation (10), the ratio of the selection probabilities of alternatives i and j is

$$\frac{\mathcal{P}_i(\overrightarrow{q})}{\mathcal{P}_j(\overrightarrow{q})} = \frac{\mathcal{P}_i^0 e^{q_i/\lambda}}{\mathcal{P}_j^0 e^{q_j/\lambda}}.$$

The reason that IIA does not hold here is that the prior selection probabilities, $\mathcal{P}_i^0$ and $\mathcal{P}_j^0$ can change in complex ways as new choices are added to the set of available alternatives. Section 5.2 provides an example of the failure IIA.

The multinomial logit's IIA property is closely related to its predictions for the way the DM will substitute across options as their values change. Suppose the value of option k increases. The DM will be more likely to select this options and less likely to select other options. The logit predicts that the probability of selecting all other options, $i \neq k$, will be reduced by the same proportion. This proportionate shifting is an implication of IIA in that this is the only way that

the ratio of selection probabilities can remain the same as the value option k changes. In the rational inattention model, there is a crucial distinction between changes in the value of an option that are known *a priori* and those that are not. Using equation (10), the proportionate change in the probability of selecting option i can be written

$$\frac{\mathcal{P}_i(\overrightarrow{q})}{\mathcal{P}_i(\overrightarrow{q})} = \frac{\hat{\mathcal{P}}_i^0}{\mathcal{P}_i^0}\frac{e^{\hat{q}_i/\lambda}}{e^{q_i/\lambda}}\frac{\sum_{j=1}^N \mathcal{P}_j^0 e^{q_j/\lambda}}{\sum_{j=1}^N \hat{\mathcal{P}}_j^0 e^{\hat{q}_j/\lambda}},$$

where a hat on a variable indicates the value after the value of option k has changed and variables without hats refer to the choice probabilities before the change. We assume that the value of option i has not changed, $q_i = \hat{q}_i$, so the second fraction drops out of the expression. In the multinomial logit case, the $\mathcal{P}_i^0$ are not present and it follows that this expression is the same for any $i \neq k$. Notice, that if the prior information is fixed and therefore the prior selection probabilities are the same before and after the change, $\hat{\mathcal{P}}_i^0 = \mathcal{P}_i^0$, then we arrive at the same conclusion. If, however, the DM is (even partially) aware of the change *a priori*, then the prior selection probabilities may change. In this case, the model can generate richer substitution patterns as the ratio $\hat{\mathcal{P}}_i^0/\mathcal{P}_i^0$ can vary across options.

3.2 Existence and Uniqueness of the Solution

In the optimization problem stated above, the objective function is continuous and the constraint set is compact so a solution exists by the extreme value theorem.[7] Whether or not the solution is unique, depends on whether the options are sufficiently different. Consider a case where two options have values that are perfectly equal in all states of the world. Call these options, option 1 and option 2. The DM is indifferent between the two options as he or she knows that selecting

[7] See Lemma 6 in Appendix B for details.

one is always equivalent to selecting the other. Therefore the objective function does not change as the DM increases $\mathcal{P}_1^0$ and reduces $\mathcal{P}_2^0$ as long as the sum of these probabilities is held fixed. In this case, the solution would not be unique unless $\mathcal{P}_1^0 = \mathcal{P}_2^0 = 0$. When we rule out cases such as this, the solution is indeed unique. The following assumptions are each sufficient conditions for a unique solution to exist.

Assumption 1. *The options are exchangeable in the prior in that, for any permutation, π, of the indices, the random vectors $\{q_1, q_2, \cdots, q_N\}$ and $\{q_{\pi_1}, q_{\pi_2}, \cdots, q_{\pi_N}\}$ are equal in distribution with respect to the prior and for any i and j in $\{1, \cdots, N\}$, q_i and q_j are not almost surely equal.*

Assumption 2. *$N = 2$ and the values of the two options are not almost surely equal.*

Assumption 3. *For all but at most one $k \in \{1..N\}$, there exist two sets $S_1 \subset \mathbb{R}^N, S_2 \subset \mathbb{R}^N$ with positive probability measures with respect to the prior, $g(\overrightarrow{q})$, such that for all $\overrightarrow{q}_1 \in S_1$ there exists $\overrightarrow{q}_2 \in S_2$ where $\overrightarrow{q}_1$ and $\overrightarrow{q}_2$ differ in k^{th} entry only.*

In assumption 1, the condition that the options are exchangeable in the prior is a formalization of the notion that they are viewed symmetrically *ex ante*. The second part of the assumption is that there is some positive probability that the options have different values. When $N = 2$, as in assumption 2, we do not need to assume symmetry. For $N > 2$ and *ex ante* asymmetric options, we have assumption 3. In words, this assumption says that there is independent variation in the value of all options except possibly one. Assumption 3 is satisfied if the values of the options are independently distributed and no more than one of their marginals is degenerate to a single point although the assumption is quite a bit weaker than independence as it just requires that there is not some form of perfect co-movement between the values. With these assumptions in hand, we can now state the result.

Theorem 2. *If any of assumptions 1, 2, or 3 holds, then the solution to the DM's optimization problem is unique.*

Proof. See corollaries 8, 9, and 10 in Appendix B. □

4 Ex ante symmetric options: the multinomial logit

In this section, we assume that all the options seem identical to the buyer *a priori* so the values are exchangeable in the prior g. That is, the DM finds differences between the options only after he or she starts processing information. We also assume that there are some states of the world in which the options take different values. If this is not the case, the DM does not face a meaningful choice. These assumptions are stated as assumption 1 in the previous section.

Under these assumption, 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[8]. However, Theorem 2 tells us that there is a unique solution. Using $\mathcal{P}_i^0 = 1/N$ in equation (10), we arrive at the following result.

Theorem 3. *Let a rationally inattentive agent be presented with N options with $g(\overrightarrow{q})$ symmetric with respect to permutations of its arguments. The options are ex ante identical. Then, the probability of choosing an option i as a function of realized values of all of options, is given by*

$$\mathcal{P}_i(\overrightarrow{q}) = \frac{e^{q_i/\lambda}}{\sum_{j=1}^{N} e^{q_j/\lambda}}, \tag{12}$$

which is the multinomial logit formula.

[8] Appendix *B*

It is worth mentioning that $\mathcal{P}_i(\overrightarrow{q})$ does not depend on the prior g. Moreover the DM always chooses to process some information, which is not necessarily the case when the prior is asymmetric. Here the marginal expected value of additional information is initially infinite and then decreasing with more information processed so the DM chooses to process some positive amount of information as long as λ is finite.

5 Asymmetric options

In the previous section, we provided an analytic solution for the case where the prior is symmetric with the result that the selection probabilities are given by the multinomial logit. For an asymmetric prior, the selection probabilities are given by equation (10), which depends on the prior probabilities $\{\mathcal{P}_i^0\}_{i=1}^N$, which in turn depend on the specifics of the prior. We now explore how these prior probabilities are formed.

If the cost of information is sufficiently high, the DM may not process any information, in which case he or she simply selects the option with the highest expected value according to the prior. Notice that these expected values only depend on the marginal distributions of the values. When the DM does process information, choices depend on the full joint distribution of the values.

If an option has higher expectation than another one, then it is often more likely to be selected even when both options take the same values. The option with a higher expected value is simply a safer bet and the rational inattentive agent is aware of his limits to processing information. However, it does not always need to be the case. Imagine a situation where the DM chooses from 101 different options. Option 1 takes value 0.99 with certainty, while all the other options take the value 0 with the probability 99% and the value 1 otherwise. If the DM processed little information, then he or she would most certainly choose option 1. The DM would often choose the first option

even if after processing quite a bit of information simply because all other options' realized values would equal zero, or because of uncertainty about whether a certain option's realized value equals 1 and thus going for $q = 0.99$ with certainty would be a good choice. If the values of options 2 through 101 are independent of one another, option 1 will be selected with some positive probability. However, the situation changes drastically if the values of options 2 to 101 co-move in such a way that exactly one of them takes the value 1, while all others 0. In this case, the DM knows there is one better option than option 1. If the information is costly, the DM will always choose option 1. If it is very cheap, the DM will never choose option 1, although its expected value is 0.99 compared to 0.01 for the other options.

We now provide several examples of how a rationally inattentive agent would behave for different specifications of his or her prior knowledge of the options. In doing so, there are two main points that we would like to convey. First, these examples demonstrate how one can solve for the prior selection probabilities $\mathcal{P}_i^0$ when the options are asymmetric. It is important to find these probabilities because they are needed to compute the conditional selection probabilities $\mathcal{P}_i(\overrightarrow{q})$ as shown in equation (10). Second, we demonstrate how the IIA property of the multinomial logit fails when the options are asymmetric.

5.1 Simple asymmetric case

In this subsection we consider a simple example in which the prior is asymmetric. In this example 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 1. *The DM chooses* $i \in \{1, 2\}$. *The value of option* 1 *is distributed as* $q_1 = 0$ *with the probability* g_0 *and* $q_1 = 1$ *with the probability* $1 - g_0$. *Option* 2 *carries the value* $q_2 = R \in (0, 1)$ *with*

certainty. The cost of information is λ.

To solve the problem, we must find $\{\mathcal{P}_i^0\}_{i=1}^2$. To do so we use the normalization conditions on the distribution of $\overrightarrow{q}$ conditional on each choice $i \in \{1,2\}$, equation (8), which take the following form

$$1 = \frac{g_0}{\mathcal{P}_1^0 + \mathcal{P}_2^0 e^{\frac{R}{\lambda}}} + \frac{(1-g_0)e^{\frac{1}{\lambda}}}{\mathcal{P}_1^0 e^{\frac{1}{\lambda}} + \mathcal{P}_2^0 e^{\frac{R}{\lambda}}} \tag{13}$$

$$1 = \frac{g_0 e^{\frac{R}{\lambda}}}{\mathcal{P}_1^0 + \mathcal{P}_2^0 e^{\frac{R}{\lambda}}} + \frac{(1-g_0)e^{\frac{R}{\lambda}}}{\mathcal{P}_1^0 e^{\frac{1}{\lambda}} + \mathcal{P}_2^0 e^{\frac{R}{\lambda}}}. \tag{14}$$

These are two equations in the unknowns $\{\mathcal{P}_i^0\}_{i=1}^2$ although if $\mathcal{P}_i^0 = 0$ then the equation for the corresponding choice of i need not hold. Solutions to the system of equations generated by the normalization conditions will always satisfy $\sum_i \mathcal{P}_i^0 = 1$.[9] There are three solutions to this system,

$$\begin{aligned} \mathcal{P}_1^0 &\in \left\{ 0, 1, -\frac{e^{\frac{R}{\lambda}}\left(-e^{\frac{1}{\lambda}} + e^{\frac{R}{\lambda}} - g_0 + g_0 e^{\frac{1}{\lambda}}\right)}{\left(e^{\frac{1}{\lambda}} - e^{\frac{R}{\lambda}}\right)\left(-1 + e^{\frac{R}{\lambda}}\right)} \right\} \\ \mathcal{P}_2^0 &= 1 - \mathcal{P}_1^0. \end{aligned} \tag{15}$$

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 utility is then R with certainty. The second solution, $\mathcal{P}_1^0 = 1$, results in the *a priori* selection of option 1, expected utility equals $(1 - g_0)$. The third solution describes the case when the DM chooses to process a positive amount of information.

Problem 1 satisfies assumption 2 as there are just two options and they never take the same

[9] It follows from equation (8) that

$$\sum_{i=1}^N \mathcal{P}_i^0 = \sum_{i=1}^N \mathcal{P}_i^0 \int \frac{g(\overrightarrow{q}) e^{\frac{q_i}{\lambda}}}{\sum_{j=1}^N \mathcal{P}_j^0 e^{\frac{q_j}{\lambda}}} d\overrightarrow{q}.$$

Then exchanging the order of summation and integration and noting that the prior integrates to one yields the result.

values. Therefore, theorem 2 establishes that the solution to the DM's optimization problem must be unique. In fact, there is an alternative way to see that the solution must unique. Following Appendix B, any convex linear combination of two solutions needs to be a solution too. Since $\mathcal{P}_1^0$ satisfies the normalization condition at three different values only, never on an entire interval, the solution to the DM's problem has to be unique.

Given that there must be a unique solution, not all three solutions to the system of equations (13) and (14) can be solutions to the DM's optimization problem. 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

$$\mathcal{P}_1^0 = \max\left(0, \min\left(1, -\frac{e^{\frac{R}{\lambda}}\left(-e^{\frac{1}{\lambda}} + e^{\frac{R}{\lambda}} - g_0 + g_0 e^{\frac{1}{\lambda}}\right)}{\left(e^{\frac{1}{\lambda}} - e^{\frac{R}{\lambda}}\right)\left(-1 + e^{\frac{R}{\lambda}}\right)}\right)\right). \tag{16}$$

For a given set of parameters, $\mathcal{P}_1^0$ as a function of R is shown in Figure 1. For R close to 0 or to 1, the DM decides to process no information and selects one of the options with certainty. In the middle range however, the DM does processes information and the selection of option 1 is less and less probable as R increases, since option 2 is more and more appealing.

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 (16) with

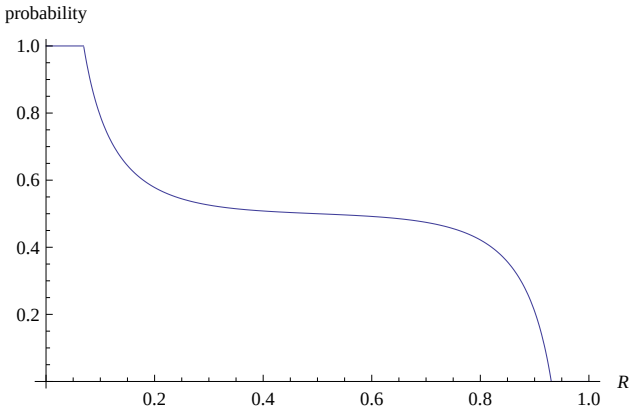

Figure 1: $\mathcal{P}_1^0$ as a function of R and $\lambda = 0.1, g_0 = 0.5$.

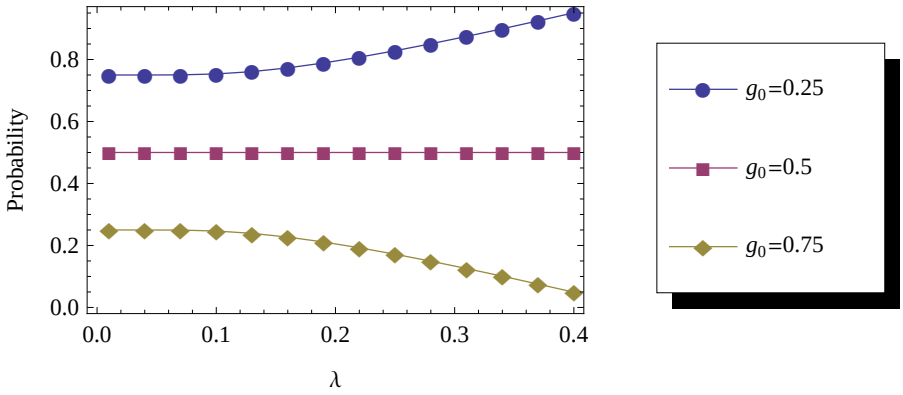

Figure 2: $\mathcal{P}_1^0$ as a function of λ evaluated at various values of of g_0 and $R = 0.5$.

respect to R we find $\partial \mathcal{P}_1^0 / \partial R < 0$, the function is non-increasing .[10] Similarly, one would expect the unconditional probability of selecting option 1 to fall as g_0 rises, as it is more likely to have a low value. Again, the intuition can be confirmed from differentiating the non-constant part of (16) with respect to g_0. The dependence of the model on the cost of processing information, λ, is more difficult to characterize analytically. Figure 2 plots $\mathcal{P}_1^0$ for three values of the prior, g_0. When processing information is cheap—low values of λ—$\mathcal{P}_1^0$ is just equal to $1 - g_0$ because the DM will always learn 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 0.5 because the DM no longer learns as much about the value of option 1 and eventually just selects the option with the highest value according to the prior. For $g_0 = 1/2$ and $R = 1/2$, $\mathcal{P}_1^0$ simplifies to $1/2$. In this case the DM is *a priori* indifferent between the two options and even for high values of λ, the DM will process at least a small amount of information in order to break the tie.[11]

5.2 Duplicate options

The previous subsection studied a case where the options differ in the marginal distributions of their values. Options may also differ in other features of the joint distribution of their values. For example, the values of two options may be more highly correlated with each other than they are with a third option. In the extreme, two options might be exact duplicates. The multinomial logit has well known difficulties when some options are similar or duplicates. These difficulties were illustrated in the introduction with Debreu's bus paradox. Debreu's logic was that duplicating an option does not fundamentally change the choice facing the DM and so should not have a

[10] Verifying this inequality requires a few steps and details are available upon request.

[11] To see that some information is always processed, notice that the conditional probability of selection option 1 is $e^{q_1/\lambda} / \left(e^{q_1/\lambda} + e^{1/(2\lambda)}\right)$, which is never equal to the unconditional probability $\mathcal{P}_1^0 = 1/2$ for $q_1 \in \{0, 1\}$ and a finite λ.

substantial impact on the choice probabilities. In this section, we begin by showing that the rationally inattentive DM treats duplicate options as a single option. We then extend this idea in section 5.3 to consider a case where two options are similar, but not exact duplicates.

We use a version of the bus problem to analyze how the rationally inattentive agent treats duplicate options. In our framework there are two sets of selection probabilities: the probabilities of selecting each option conditional on the true values of the options and the prior probabilities of selecting each option that would describe the DM's anticipated actions before he or she begins processing information. The notion that the values of the available options are uncertain and believed to be distributed according to a prior distribution is a particular feature of our framework so it is reasonable to think that the conditional probabilities are closer to what Debreu and the subsequent literature have in mind. Nevertheless, the rationally inattentive DM treats duplicate options as a single option both in terms of prior probabilities and in terms of conditional probabilities.

To show that the DM treats duplicate options as a single option we state two choice problems. In the first, the DM chooses from the set $\{\text{yellow bus}, \text{train}\}$ and in the second the DM chooses from $\{\text{yellow bus}, \text{red bus}, \text{train}\}$. When the buses are exact duplicates—a notion that we formalize below in assumption 4—the probability of choosing a bus (of any color) is the same in both of these choice problems. We now state the two choice problems formally.

Problem 2. *The DM chooses from the set $\{$yellow bus, train$\}$. The prior distribution for the values of the two options is $g^1(q_y, q_t)$, where q_y is the value of the yellow bus and q_t is the value of the train. q_y and q_t are not a.s. equal. The cost of information is λ_1.*

Problem 3. *The DM chooses from the set $\{$yellow bus, red bus, train$\}$. The prior distribution for the values of the options is $g^2(q_y, q_r, q_t)$, where q_r is the value of the red bus. The cost of information is λ_2.*

We now introduce our assumptions. The first assumption formalizes the notion that the buses are duplicates. We assume that the two buses are duplicates in that the prior places no weight on their values being different. The meaning of this assumption is that the DM knows the two buses are identical before processing any information although does not know what their (joint) value is. It is also natural to assume that the joint distribution of a bus and the train is the same as in the one-bus case.

Assumption 4. *The prior for the two-bus case satisfies*

$$g^2(q_y, q_r, q_t) = \begin{cases} g^1(q_y, q_t) & \text{if } q_y = q_r, \\ 0 & \text{if } q_y \neq q_r. \end{cases}$$

Our second assumption is simply that the cost of a bit of information is the same in the two problems.

Assumption 5. $\lambda_1 = \lambda_2$.

Before we state the proposition, we must introduce some notation to describe the solutions to these problems. Let $\mathcal{P}_b^0$ be the prior probability of selecting the (yellow) bus in problem 2 and $\mathcal{P}_b(\overrightarrow{q})$ be the probability of selecting the bus conditional on a realization of $\overrightarrow{q}$ in problem 2. For problem 3 we use analogous notation with the subscript y to denote probabilities of selecting the yellow bus and subscript r to denote probabilities of selecting the red bus.

Our first proposition is that the prior probability of selecting a bus is the same in both problems.

Proposition 4. *If assumptions 4 and 5 hold, then* $\mathcal{P}_b^0 = \mathcal{P}_y^0 + \mathcal{P}_r^0$.

Proof. See Appendix C. □

A corollary to this proposition is that the probability of selecting a bus conditional on realized values of the options is the same in both problems. As the two buses are duplicates it is natural to restrict attention to realizations of the vector $\overrightarrow{q}$ for which $q_y = q_r$.

Corollary 5. *If assumptions 4 and 5 hold, then for any* $\overrightarrow{q} = (q_y, q_r, q_t)$ *that satisfies* $q_y = q_r$ *we have* $\mathcal{P}_b(\overrightarrow{q}) = \mathcal{P}_y(\overrightarrow{q}) + \mathcal{P}_r(\overrightarrow{q})$.

Proof. See Appendix C. □

5.3 Correlated values

The previous subsection considered the case where two options are known to be exactly identical. In this subsection we explore the behavior of the rationally inattentive agent as the co-movement of two options varies. We do so in the context of a choice among three options for which we can make some progress analytically.

Problem 4. *The DM chooses from the set* $\{$*yellow bus, red bus, train*$\}$. *The DM knows the quality of the train exactly,* $q_t = R \in (0, 1)$. *The buses each take one of two values, either 0 or 1, with expected values 1/2 for each. The joint distribution of the values of all three options is*

$$\begin{aligned} g(0,0,R) &= 1/2 - g_1 \\ g(1,0,R) &= g_1 \\ g(0,1,R) &= g_1 \\ g(1,1,R) &= 1/2 - g_1. \end{aligned} \tag{17}$$

The DM can process information about the values of the busses at a cost λ.

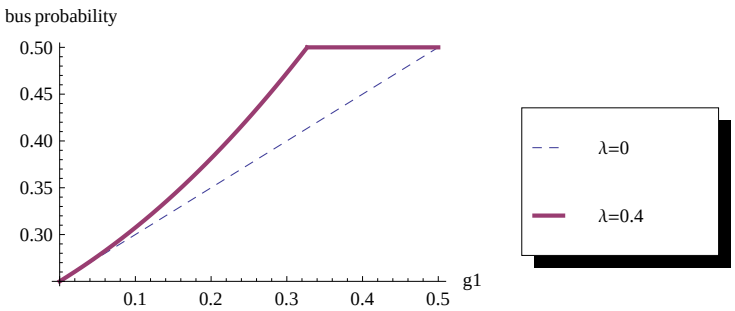

Figure 3: $\mathcal{P}_y^0$ for various values of λ and g_1 and $R = 1/2$.

We are going to illustrate how the choice probabilities vary with the correlation of the values of the two buses. Given the joint distribution above, the correlation between q_y and q_r is $1 - 4g_1$.

Notice that when g_1 is greater than zero, the conditions of assumption 3 are satisfied as it is possible to vary each bus value while holding the values of the other options constant. When g_1 equals zero, this problem resolves to the duplicates case and assumption 3.

As before, to find the solution to the DM's optimization problem we must solve for $\{\mathcal{P}_y^0, \mathcal{P}_r^0, \mathcal{P}_t^0\}$. The normalization condition on choosing the first option is

$$\begin{aligned} 1 = & \frac{1/2 - g_1}{\mathcal{P}_y^0 + \mathcal{P}_r^0 + (1 - \mathcal{P}_y^0 - \mathcal{P}_r^0)e^{R/\lambda}} + \frac{g_1 e^{1/\lambda}}{\mathcal{P}_y^0 e^{1/\lambda} + \mathcal{P}_r^0 + (1 - \mathcal{P}_y^0 - \mathcal{P}_r^0)e^{R/\lambda}} \\ + & \frac{g_1}{\mathcal{P}_2^0 e^{1/\lambda} + \mathcal{P}_y^0 + (1 - \mathcal{P}_y^0 - \mathcal{P}_r^0)e^{R/\lambda}} + \frac{(1/2 - g_1)e^{1/\lambda}}{\mathcal{P}_y^0 e^{1/\lambda} + \mathcal{P}_r^0 e^{1/\lambda} + (1 - \mathcal{P}_y^0 - \mathcal{P}_r^0)e^{R/\lambda}} \end{aligned} \quad (18)$$

Due to the symmetry between the buses, we know $\mathcal{P}_y^0 = \mathcal{P}_r^0$. This problem can be solved analytically using the same technique as in the previous sections, the resulting expression is however too complicated to include here.[12] Instead, we illustrate the behavior of the model for $R = 1/2$ and various values of g_1 and λ in Figure 3. As g_1 increases, and the correlation between the values of the buses decreases, the unconditional probability of choosing either bus increases. If they are

[12] They can be provided upon request.

perfectly correlated, then their collective probability decreases to 0.5, they are effectively treated as one option. To see that they are treated as a single option, recall from Section 5.1 that when values of zero and one were equally likely ($g_0 = 1/2$) and the reservation value was equal to 1/2, we found $\mathcal{P}_y^0 = 1/2$ for all λ. That case corresponds to the situation here with just a single bus in the choice set. So with two buses in the choice set we find that the sum of their selection probabilities is equal to 1/2 as section 5.2 tells us we should.

In the perfect information case, the probability of choosing the yellow bus, $\mathcal{P}_y^0$, equals $1/4+g_1/2$, if we assume that ties are broken at random. As the correlation between the values of the buses decreases, the probability that option 3 has the highest value decreases, and thus $\mathcal{P}_y^0$ increases. This effect persists when $\lambda > 0$. The more similar the two buses are, the lower is the probability of either of them being selected. This is the extension of the duplicates results.

For $\lambda > 0$, however, $\mathcal{P}_y^0$ is larger than it is in the perfect information case. If the DM does not possess perfect information, then he or she considers that *a priori* it is more likely that either of the buses, rather than the train, possesses the highest value among the three options. With $g_1 > 0$ and increasingly costly information, the DM would shift his or her attention to which one of the two buses to select rather than whether to select the train, since the buses values are more likely to be the highest.

6 Concluding Remarks

In this paper, we have studied the optimal behavior of a rationally inattentive agent who faces a discrete choice problem and shown that this model gives rise to the multinomial logit model when the options are *a priori* symmetric. This finding opens the door for future research to combine the rational inattention framework with our existing knowledge of the implications of the multinomial

logit.

We have also analyzed the way in which prior knowledge of the available options affects choice behavior when information costs result in the DM choosing with incomplete information. The incorporation of this prior knowledge can lead to more intuitive predictions than those that arise out of the standard multinomial logit. For example, the rationally inattentive agent treat's duplicate options as a single option while the multinomial logit's IIA property implies that they are treated as distinct options.

A Derivation of the first order condition

Here we derive the first order condition, equation (5). Using equation 2.35 in Cover and Thomas, we can write the mutual information as

$$I(\overrightarrow{q};i) = \sum_i \mathcal{P}_i^0 \int_{\overrightarrow{q}} f(\overrightarrow{q}\,|i) \log \frac{\mathcal{P}_i^0 f(\overrightarrow{q}\,|i)}{\mathcal{P}_i^0 g(\overrightarrow{q}\,)} d\overrightarrow{q}$$

replacing the prior with equation (4) and canceling two $\mathcal{P}_i^0$ terms

$$I(\overrightarrow{q};i) = \sum_i \mathcal{P}_i^0 \int_{\overrightarrow{q}} f(\overrightarrow{q}\,|i) \log \frac{f(\overrightarrow{q}\,|i)}{\sum_j \mathcal{P}_j^0 f(\overrightarrow{q}\,|j)} d\overrightarrow{q}\,. \tag{19}$$

The Lagrangian is then

$$\begin{aligned}\mathcal{L} = &\sum_i \mathcal{P}_i^0 \int_{\overrightarrow{q}} q_i f(\overrightarrow{q}\,|i) d\overrightarrow{q} - \lambda\kappa \\ &- \chi \left[\sum_i \mathcal{P}_i^0 \int_{\overrightarrow{q}} f(\overrightarrow{q}\,|i) \log \frac{f(\overrightarrow{q}\,|i)}{\sum_j \mathcal{P}_j^0 f(\overrightarrow{q}\,|j)} d\overrightarrow{q} - \kappa \right] - \int_{\overrightarrow{q}} \mu(\overrightarrow{q}\,) \left[\sum_i \mathcal{P}_i^0 f(\overrightarrow{q}\,|i) - g(\overrightarrow{q}\,) \right] d\overrightarrow{q}\,,\end{aligned}$$

where $\chi \in \mathbb{R}$ and $\mu \in L^\infty(\mathbb{R}^N)$ are Lagrange multipliers. The first order condition with respect to κ is simply $\lambda = \chi$. The first order condition with respect to $f(\overrightarrow{q}\,|i)$ is

$$\begin{aligned}\mathcal{P}_i^0 q_i - \chi \mathcal{P}_i^0 \log \frac{f(\overrightarrow{q}\,|i)}{\sum_j \mathcal{P}_j^0 f(\overrightarrow{q}\,|j)} - \chi \mathcal{P}_i^0 f(\overrightarrow{q}\,|i) \frac{\sum_j \mathcal{P}_j^0 f(\overrightarrow{q}\,|j)}{f(\overrightarrow{q}\,|i)} \frac{1}{\sum_j \mathcal{P}_j^0 f(\overrightarrow{q}\,|j)} \\ + \chi \sum_k \mathcal{P}_k^0 f(\overrightarrow{q}\,|k) \frac{\sum_j \mathcal{P}_j^0 f(\overrightarrow{q}\,|j)}{f(\overrightarrow{q}\,|k)} \frac{f(\overrightarrow{q}\,|k)\mathcal{P}_i^0}{\left[\sum_j \mathcal{P}_j^0 f(\overrightarrow{q}\,|j)\right]^2} - \mu(\overrightarrow{q}\,)\mathcal{P}_i^0 = 0.\end{aligned}$$

We can now cancel a number of terms and replace $\sum_j \mathcal{P}_j^0 f(\overrightarrow{q}|j)$ with $g(\overrightarrow{q})$ to arrive at

$$\mathcal{P}_i^0 \left(q_i - \chi \log \frac{f(\overrightarrow{q}|i)}{g(\overrightarrow{q})} - \mu(\overrightarrow{q}) \right) = 0.$$

If $\mathcal{P}_i^0 > 0$ and $\lambda > 0$, solving for $f(\overrightarrow{q}|i)$ we obtain

$$f(\overrightarrow{q}|i) = \exp\left(\frac{q_i}{\chi}\right) \exp\left(-\frac{\mu(\overrightarrow{q})}{\chi}\right) g(\overrightarrow{q})$$

using $\lambda = \chi$ and defining $h(\overrightarrow{q}) \equiv \exp\left(-\frac{\mu(\overrightarrow{q})}{\lambda}\right) g(\overrightarrow{q})$ produces equation (5).

B Existence and Uniqueness of Solutions to the DM's Problem

Lemma 6. *The DM's optimization* (2)-(4) *problem always has a solution.*

Proof: Since (10) 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 $f(\overrightarrow{q}|i)$ is also a continuous function of $\{\mathcal{P}_i^0\}_{i=1}^N$. Moreover, the admissible set for $\{\mathcal{P}_i^0\}_{i=1}^N$ given by $\sum_i \mathcal{P}_i^0 = 1$ and $\mathcal{P}_k^0 \geq 0 \quad \forall k$, is compact. Therefore, the maximum always exists.

Lemma 7. *If* $S = \{\mathcal{P}_i^0, f(\overrightarrow{q}|i)\}_{i=1}^N$ *and* $\hat{S} = \{\hat{\mathcal{P}}_i^0, \hat{f}(\overrightarrow{q}|i)\}_{i=1}^N$ *are two distinct solutions to the DM's optimization problem, then*

$$\sum_i (\mathcal{P}_i^0 - \hat{\mathcal{P}}_i^0) e^{q_i/\lambda} = 0 \qquad a.s. \tag{20}$$

Proof: Mutual information is a convex function of the joint distribution of the two variables. The objective (2) is thus a concave functional: the first term is linear and the second is concave. Moreover, the admissible set of $\{\mathcal{P}_i^0, f(\overrightarrow{q}|i)\}_{i=1}^N$, satisfying the constraints is convex: (3) is a concave constraint and all other are linear. Therefore, any convex linear combination $\tilde{S}(\xi)$ of the solutions

S and $\hat{S}$

$$\tilde{\mathcal{P}}_i^0(\xi) = \mathcal{P}_i^0 + \xi\left(\hat{\mathcal{P}}_i^0 - \mathcal{P}_i^0\right) \qquad \xi \in [0,1], \forall i. \tag{21}$$

is also a solution. This solution thus needs to satisfy (8) for all $\xi \in [0,1]$. The right hand side of (8) has to be a constant as a function of ξ. However, its second derivative with respect to ξ at $\xi = 0$, which has to equal zero, is

$$\int \frac{g(\overrightarrow{q})e^{\frac{q_i}{\lambda}}\left(\sum_{j=1}^N(\hat{\mathcal{P}}_j^0 - \mathcal{P}_j^0)e^{\frac{q_j}{\lambda}}\right)^2}{\left(\sum_{j=1}^N \tilde{\mathcal{P}}_j^0 e^{\frac{q_j}{\lambda}}\right)^3} d\overrightarrow{q}. \tag{22}$$

Therefore, for the two solutions to exist, (20) has to hold.

Corollary 8. *If assumption 1 holds then the solution to the DM's optimization problem is unique.*

Proof: Let the solution be non-unique. (20) thus needs to hold. According to the corollary's assumption, there exists $S_1 \subset \mathbb{R}^N$ with a positive measure w.r.t. g, such that all $\overrightarrow{q}$ in S_1 are non-constant vectors. Since the solution is non-unique, (20) holds almost surely and S_1 has a positive mass, then there surely exist $\overrightarrow{q}$ and $\overrightarrow{q}'$ generated from $\overrightarrow{q}$ only by switching entries i and j, where that $q_i \neq q_j$, satisfying (20) point-wise. By subtracting the equations for $\overrightarrow{q}$ and $\overrightarrow{q}'$ we get

$$(\Delta_i - \Delta_j)\left(e^{\frac{q_i}{\lambda}} - e^{\frac{q_j}{\lambda}}\right) = 0, \tag{23}$$

where Δ_i denotes $(\mathcal{P}_i^0 - \hat{\mathcal{P}}_i^0)$. We get $\Delta_i = \Delta_j$. However, since we can reshuffle the entries arbitrarily, Δ_i equals a constant Δ for all i in $\{1..N\}$. Moreover, $\Delta = 0$ since $\sum_i \Delta_i = 0$. The solution must be unique. □.

Corollary 9. *If assumption 2 holds then the solution to the DM's optimization problem is unique.*

Proof: for $N = 2$, $(\mathcal{P}_1^0 - \hat{\mathcal{P}}_1^0) = -(\mathcal{P}_2^0 - \hat{\mathcal{P}}_2^0)$ (20) takes the form:

$$(\mathcal{P}_1^0 - \hat{\mathcal{P}}_1^0)(e^{\frac{q_1}{\lambda}} - e^{\frac{q_2}{\lambda}}) = 0 \qquad a.s. \tag{24}$$

If q_1 and q_2 are not equal almost surely, then $\mathcal{P}_1^0 = \hat{\mathcal{P}}_1^0$. □

The following corollary of Lemma 7 is more directly linked to the analytical structure of (20), and its assumptions are perhaps less intuitive. Let $(\mathcal{P}_k^0 - \hat{\mathcal{P}}_k^0) \neq 0$. It is clear that if (20) holds for $\overrightarrow{q}_1$ than it can not hold for any $\overrightarrow{q}_2$ that differs from $\overrightarrow{q}_1$ in the k^{th} entry only.

Corollary 10. *If assumption 3 holds then the solution to the DM's optimization problem is unique.*

Proof: The Corollary follows directly from Lemma 7 and the discussion right above it, adjusted to satisfy the requirement of "almost sure" equality. What we need to explain is why it does not matter for the uniqueness if there exists exactly one k that does not satisfy the assumptions. If $\{\mathcal{P}_i^0\}_i$ and $\{\hat{\mathcal{P}}_i^0\}_i$ are two different solutions then there exist at least two different k's s.t. $(\mathcal{P}_k^0 - \hat{\mathcal{P}}_k^0) \neq 0$. Therefore, being able to vary only one of these two entries suffices for the uniqueness. □

Corollary 10 applies to setups in Sections 5.1 and 5.3. In these cases, one option takes value R with certainty, its marginal is degenerate. That is why we explicitly allowed for one entry not satisfying the assumptions. However, the corollary does not apply to the case of pure duplicates in Section 5.2 and $g_1 = 0$ in 5.3. In these cases, we can not vary values corresponding to both of the duplicates independently of each other. But this is fine, the solution is not unique. If two options are exactly equivalent in all realizations, then the DM chooses only the sum of their probabilities.

C Proofs for section 5.2

Proof of proposition 4. The normalization conditions for problem 2 can be written as[13]

$$1 = \int\int \frac{e^{q_y/\lambda}}{\mathcal{P}_b^0 e^{q_y/\lambda} + (1-\mathcal{P}_b^0)e^{q_t/\lambda}} g^1(q_y, q_t) dq_y dq_t \tag{25}$$

$$1 = \int\int \frac{e^{q_t/\lambda}}{\mathcal{P}_b^0 e^{q_y/\lambda} + (1-\mathcal{P}_b^0)e^{q_t/\lambda}} g^1(q_y, q_t) dq_y dq_t \tag{26}$$

Both of these equations must hold for $\mathcal{P}_b^0 \in (0,1)$. For $\mathcal{P}_b^0 = 1$ only equation (25) must hold and for $\mathcal{P}_b^0 = 0$ only equation (26) must hold. This system of equations has two or three possible solutions: $\mathcal{P}_b^0 = \{0, P^*, 1\}$ where P^* is an interior solution that may or may not exist. To see that there cannot be multiple interior solutions, subtract equation (26) from (25) and rearrange to arrive at

$$0 = \int\int \frac{1 - e^{(q_t - q_y)/\lambda}}{\mathcal{P}_b^0 + (1-\mathcal{P}_b^0)e^{(q_t-q_y)/\lambda}} g^1(q_y, q_t) dq_y dq_t. \tag{27}$$

Define the random variable $z \equiv e^{(q_t - q_y)/\lambda}$ so we can write equation (27) as

$$0 = E\left[\frac{1-z}{\mathcal{P}_b^0 + (1-\mathcal{P}_b^0)z}\right]. \tag{28}$$

The right-hand side of this equation is decreasing in $\mathcal{P}_b^0$, so there can be at most one solution. To see this, differentiate w.r.t. $\mathcal{P}_b^0$

$$\frac{\partial}{\partial \mathcal{P}_b^0} E\left[\frac{1-z}{\mathcal{P}_b^0 + (1-\mathcal{P}_b^0)z}\right] = E\left[-\frac{(1-z)^2}{\left[\mathcal{P}_b^0 + (1-\mathcal{P}_b^0)z\right]^2}\right]$$

[13] Note that previously we have written $\int g(\overrightarrow{q}) d\overrightarrow{q}$ to indicate integration with respect to the prior, but in this section we find it useful to distinguish between the dimensions over which we are integrating.

and on the right-hand side, the expectation is taken over a function that is negative for all $z \neq 1$ and zero for $z = 1$. Given that we have q_y is no a.s. equal to q_t, the expectation places some weight of $z \neq 1$.

In problem 3, the normalization condition for the yellow bus is

$$1 = \int\int\int \frac{e^{q_y/\lambda}}{\mathcal{P}_y^0 e^{q_y/\lambda} + \mathcal{P}_r^0 e^{q_r/\lambda} + (1 - \mathcal{P}_y^0 - \mathcal{P}_r^0)e^{q_t/\lambda}} g^2(q_y, q_r, q_t) dq_y dq_r dq_t.$$

This normalization condition involves integrating over a three-dimensional space, but the assumption that the buses are exact duplicates means that the prior only places weight on a two dimensional subspace. Therefore, we can think of integrating just over q_y and using the fact that $q_r = q_y$. Doing so yields

$$1 = \int\int \frac{e^{q_y/\lambda}}{\mathcal{P}_y^0 e^{q_y/\lambda} + \mathcal{P}_r^0 e^{q_y/\lambda} + (1 - \mathcal{P}_y^0 - \mathcal{P}_r^0)e^{q_t/\lambda}} g^2(q_y, q_y, q_t) dq_y dq_t \tag{29}$$

$$= \int\int \frac{e^{q_y/\lambda}}{\left(\mathcal{P}_y^0 + \mathcal{P}_r^0\right) e^{q_y/\lambda} + \left[1 - \left(\mathcal{P}_y^0 + \mathcal{P}_r^0\right)\right] e^{q_t/\lambda}} g^1(q_y, q_t) dq_y dq_t. \tag{30}$$

Now notice that the sum of $\mathcal{P}_y^0$ and $\mathcal{P}_r^0$ enters equation (30) just as $\mathcal{P}_b^0$ enters equation (25) and otherwise the two equations are the same. Using similar steps, we find that the three normalization conditions for problem 3 reduce to the same equations as (25) and (26) with $\mathcal{P}_y^0 + \mathcal{P}_r^0$ replacing $\mathcal{P}_b^0$. Therefore, any pair of $\mathcal{P}_y^0$ and $\mathcal{P}_r^0$ that sum to a $\mathcal{P}_b^0$ that solves equations (25) and (26) satisfies the normalization conditions for problem 3.

The final step of the proof is to establish that the particular solution to the normalization conditions for problem 2 that yields the highest value to the DM is also the one that yields the highest value to the DM in problem 3. Suppose the DM in problem 2 chooses $\mathcal{P}_b^0 = 1$ or $\mathcal{P}_b^0 = 0$.

Then there is no reason to process any information and the objective function value is just the expected value of the bus or train, respectively. The exact same holds in problem 3 if the DM selects $\mathcal{P}_y^0 + \mathcal{P}_r^0 = 0$, $\mathcal{P}_y^0 = 1$, $\mathcal{P}_r^0 = 1$, or $\mathcal{P}_y^0 + \mathcal{P}_r^0 = 1$. For interior solutions in problem 3, the expected value of the selected option, which differs from the objective function by the information cost, can be written

$$\int\int\int \frac{\mathcal{P}_y^0 e^{q_y/\lambda} q_y + \mathcal{P}_r^0 e^{q_r/\lambda} q_r + \left(1 - \mathcal{P}_y^0 - \mathcal{P}_r^0\right) e^{q_t/\lambda} q_t}{\mathcal{P}_y^0 e^{q_y/\lambda} + \mathcal{P}_r^0 e^{q_r/\lambda} + \left(1 - \mathcal{P}_y^0 - \mathcal{P}_r^0\right) e^{q_t/\lambda}} g^2(q_y, q_r, q_t) dq_y dq_r dq_t.$$

Restricting attention to cases where $q_y = q_r$, we have

$$\int\int \frac{\left(\mathcal{P}_y^0 + \mathcal{P}_r^0\right) e^{q_y/\lambda} q_y + \left(1 - \mathcal{P}_y^0 - \mathcal{P}_r^0\right) e^{q_t/\lambda} q_t}{\left(\mathcal{P}_y^0 + \mathcal{P}_r^0\right) e^{q_y/\lambda} + \left(1 - \mathcal{P}_y^0 - \mathcal{P}_r^0\right) e^{q_t/\lambda}} g^1(q_y, q_t) dq_y dq_t,$$

which is the same as the expected value of the selected option in problem 2 if $\mathcal{P}_y^0 + \mathcal{P}_r^0 = \mathcal{P}_b^0$. What is left is to establish that the information flow is the same in both problems. Plug equation (7) into equation (19) to find the mutual information between the DM's choice and the vector $\overrightarrow{q}$

$$I(\overrightarrow{q}; i) = \sum_i \mathcal{P}_i^0 \int_{\overrightarrow{q}} \frac{e^{q_i/\lambda}}{\sum_j \mathcal{P}_j^0 e^{q_j/\lambda}} \log \frac{e^{q_i/\lambda}}{\sum_j \mathcal{P}_j^0 e^{q_j/\lambda}} g(\overrightarrow{q}) d\overrightarrow{q}.$$

Consider this equation for problem 3, using the same logic as above, we can restrict attention to those cases where $q_y = q_r$. In that case, the term $\sum_{j \in \{r,y,t\}} \mathcal{P}_j^0 e^{q_j/\lambda}$ only depends on the sum of $\mathcal{P}_y^0$ and $\mathcal{P}_r^0$ and takes the same value as in problem 2 if $\mathcal{P}_y^0 + \mathcal{P}_r^0 = \mathcal{P}_b^0$. And similarly for the outer sum over the $\mathcal{P}_i^0$. This establishes that the information flow is the same in the two problems. As the expected value of the selected option and the information flow are the same in the two problems, the objective function takes the same value across problems for each of the two or three candidate

solutions. So whichever yields the highest value in one problem will yield the highest value in the other problem. This establishes that the DM treats the two identical buses as a single option in terms of prior probabilities. □

Proof of corollary 5. From equation (10) we have

$$\begin{aligned}
\mathcal{P}_y(\overrightarrow{q}) + \mathcal{P}_r(\overrightarrow{q}) &= \frac{\mathcal{P}_y^0 e^{q_y/\lambda} + \mathcal{P}_r^0 e^{q_r/\lambda}}{\mathcal{P}_y^0 e^{q_y/\lambda} + \mathcal{P}_r^0 e^{q_r/\lambda} + \left(1 - \mathcal{P}_y^0 - \mathcal{P}_r^0\right) e^{q_t/\lambda}} \\
&= \frac{\left(\mathcal{P}_y^0 + \mathcal{P}_r^0\right) e^{q_y/\lambda}}{\left(\mathcal{P}_y^0 + \mathcal{P}_r^0\right) e^{q_y/\lambda} + \left[1 - \left(\mathcal{P}_y^0 + \mathcal{P}_r^0\right)\right] e^{q_t/\lambda}} \\
&= \frac{\mathcal{P}_b^0 e^{q_y/\lambda}}{\mathcal{P}_b^0 e^{q_y/\lambda} + \left[1 - \mathcal{P}_b^0\right] e^{q_t/\lambda}} \\
&= \mathcal{P}_b(\overrightarrow{q}),
\end{aligned}$$

where the second equality follows from $q_y = q_r$ and the third follows from proposition 4. □

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.

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):186–188.

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.

Luo, Y. (2008). Consumption dynamics under information processing constraints. *Review of Economic Dynamics*, 11(2):366 – 385.

Mackowiak, B. and Wiederholt, M. (2009). Optimal sticky prices under rational inattention. *The American Economic Review*, 99:769–803(35).

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

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

Mattsson, L.-G. and Weibull, J. W. (2002). Probabilistic choice and procedurally bounded rationality. *Games and Economic Behavior*, 41(1):61 – 78.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York.

McFadden, D. (1976). Quantal choice analysis: A survey. *Annals of Economic and Social Measurement*, 5(4):363 – 390.

McFadden, D. (1980). Econometric models for probabilistic choice among products. *The Journal of Business*, 53(3):pp. S13–S29.

McKelvey, R. D. and Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6 – 38.

Mondria, J. (2010). Portfolio choice, attention allocation, and price comovement. *Journal of Economic Theory*, 145(5):1837 – 1864.

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

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

Sims, C. A. (1998). Stickiness. *Carnegie-Rochester Conference Series on Public Policy*, 49:317 – 356.

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

Sims, C. A. (2006). Rational inattention: Beyond the linear-quadratic case. *The American Economic Review*, 96(2):pp. 158–163.

Sims, C. A. (2010). Rational inattention and monetary economics. In Friedman, B. M. and Woodford, M., editors, *Handbook of Monetary Economics*, volume 3A, pages 155 – 181. Elsevier.

Stahl, D. O. (1990). Entropy control costs and entropic equilibria. *International Journal of Game Theory*, 19(2):129 – 138.

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):779–805.

Woodford, M. (2009). Information-constrained state-dependent pricing. *Journal of Monetary Economics*, 56(Supplement 1):S100 – S124.

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):109 – 144.