Consumer Choice Under Limited Attention When Alternatives Have Different Information Costs*

Frank Huettner† ESMT Berlin
Tamer Boyacı ESMT Berlin
Yalçın Akçay Melbourne Business School

Author's accepted manuscript (To appear in Operations Research)
https://doi.org/10.1287/opre.2018.1828

Abstract

Consumers often do not have complete information about the choices they face and therefore have to spend time and effort acquiring information. Since information acquisition is costly, consumers trade off the value of better information against its cost, and make their final product choices based on imperfect information. We model this decision using the rational inattention approach and describe the rationally inattentive consumer's choice behavior when she faces alternatives with different information costs. To this end, we introduce an information cost function that distinguishes between direct and implied information. We then analytically characterize the optimal choice probabilities. We find that non-uniform information costs can have a strong impact on product choice, which gets particularly conspicuous when the product alternatives are otherwise very similar. There are significant implications on how a seller should provide information about its products and how changes to the product set impacts consumer choice. For example, non-uniform information costs can lead to situations where it is disadvantageous for the seller to provide easier access to information for a particular product, and to situations where the addition of an inferior (never chosen) product increases the market share of another existing product (i.e., failure of regularity). We also provide an algorithm to compute the optimal choice probabilities and discuss how our framework can be empirically estimated from suitable choice data.

*The authors thank the anonymous associate editor and the reviewers.
†Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 288880950.

Contents

1. Introduction

Facing an abundance of product choices and related information, but with only limited time and attention to evaluate them, consumers have to come to grips with how much and what type of information to acquire and to pay attention to (and what to ignore), and make product choice and purchase decisions based on this partial information. It is therefore quite possible that consumers make "wrong" choices, though this does not necessarily imply that they are irrational. Since the works of Simon (1955, 1979), *bounded rationality* acknowledges the fact that individuals make rational decisions, but subject to constraints. In an information-driven world, attention that can be allocated to a specific choice task is limited, which puts constraints on the amount and type of information that can be acquired. As information is "costly", rational consumers have to trade off the value of better information against its cost. *Rational inattention*[1] theory offers a compelling approach to capture this trade-off by *endogenizing* the information acquisition process. Specifically, the pioneering works of Sims (1998, 2003, 2006) propose a framework in which information is quantified as reduction in Shannon entropy such that utility-maximizing consumers optimally select not only the quantity but also the type of information they need, i.e., the consumer is free to choose the optimal information structure. This is in contrast to standard search models where the information structure is prescribed in advance (e.g., the consumer is assumed to receive signals with normal error).

Rational inattention theory has been applied to a broad spectrum of economic problems and has been a powerful construct in providing explanations of some observed market and macroeconomic phenomena such as price stickiness, business cycles and contractions, and consumption (Sims 2003, Maćkowiak and Wiederholt 2009, 2015, Tutino 2013). It is increasingly applied to microeconomics topics as well, especially pricing (e.g., Matějka and McKay 2012, Matějka 2015, Boyacı and Akçay 2018). A fundamental driver of these applications is the evolving understanding of how rational inattention influences choice behavior. In a recent paper, Matějka and McKay (2015) study the choice behavior of rationally inat-

[1] Throughout the paper, we use the terms "limited attention" and "rational inattention" interchangeably.

tentive consumers facing discrete choices with stochastic (payoff) values, assuming that the costs of acquiring and processing information is *identical* across choice alternatives. They establish that the optimal information processing strategy leads to a choice behaviour that can be characterized as *generalized* multinomial logit (GMNL). In particular, the choice probabilities depend not only on the true realizations of the choices as in the standard multinomial logit (MNL), but also on the prior beliefs of the consumer and the cost of information.

In this paper, we generalize the GMNL model and establish the choice model for a rationally inattentive consumer whose *information costs vary across the components of the state of the world* she studies. When the consumer picks an alternative, she achieves a utility that depends on the realized state. The structure of the state space reflects the composition of the "to-be-learned" information, which may be acquired at different effort levels (i.e., different unit costs of information). Hence, the composition of the state space can carry several connotations in various contexts:

(i) A consumer visits a store to buy a digital camera and is considering *multiple products*. A few of these cameras have display models in the store, while the rest do not. The level of effort required to evaluate each camera differs depending on the availability of a display model. In this setting, the state of the world echoes the valuations of the products under consideration. Each product corresponds to a different component of the state space and is learned about at a different effort level.

(ii) A consumer is deciding whether to purchase a smartphone with *multiple features*. The fact that some features (e.g., price, weight) are easier to evaluate than others (e.g., service quality, lifetime) is captured by different information costs. Each component of the state space corresponds to the value that a feature adds to the consumer's utility.

(iii) A consumer is deciding which movie to rent when her friend visits. In this case, the states of the world describe the *preference* (or *type*) of her friend, which drives her utility obtained from watching a movie. Effectively, the consumer is trying to learn her friend's type, and it is easier to learn about some type dimensions (e.g., age, gender), than

about others (e.g., attitude towards different genres, current mood). Each type dimension corresponds to a different component of the state space.

The framework developed in this paper is general, and can be applied to all of the contexts above (and their combinations). However, given the focus of this paper and also to ease exposition, we restrict our presentation in the sequel to the first context, i.e., consumer choice among multiple products.

There are three key reasons as to why rational inattention to discrete choice with different information costs is significant. These also constitute the cuneate contributions of our paper:

REALISM & APPLICABILITY. The uniform information cost assumption underpinning the GMNL characterization can be interpreted as the consumer acquiring and processing information through a common channel with a certain associated cost. Effectively, it means that the amount of effort (and hence cost) spent to obtain and process 1-byte of information about each alternative is the same. As exemplified above, this is not necessarily the case in reality. It is oftentimes easier to obtain information about some products than about others, by the very nature of the product. Or sometimes information is obtained from different sources with different levels of time-and-attention-efficiency (online, catalog, direct sales force etc). It can also depend on the assortment that is offered – it is easier to obtain information about products that are readily available to "touch and feel" compared to others that are not available and require extra effort to garner information.[2] These realities call for a choice model that allows the information cost to vary among the alternatives considered by the consumer. Such a choice model would form the essential building block for a variety of operations/marketing applications involving consumers with limited attention.

RELATION TO OTHER CHOICE MODELS. Discrete choice models under rational inattention are particularly promising because of their close relation to MNL choice models. By generalizing the GMNL model, we extend this relationship. The connection with MNL is particularly relevant in our

[2]Hoch and Ha (1986) and Hamilton and Thompson (2007) stress the importance of consumers' experience at the point of sale and consumers' struggle to judge the value of a product through abstract product description compared with direct user experience.

context because a rather common approach to modeling the bounded rationality of customers is to adopt the quantal choice model of Luce (1959), which leads to the MNL (see McKelvey and Palfrey 1995). We refer to Anderson et al. (1992) for a comprehensive coverage of MNL models in general and to Wierenga (2008) for their use in marketing science. MNL and its variations (e.g., nested logit) have been extensively studied to model consumer behavior in the operations management literature as well, in particular in the context of pricing, revenue management, and assortment planning (Hanson and Martin 1996, van Ryzin and Mahajan 1999, Dong et al. 2009, Zhang and Adelman 2009, Davis et al. 2014, among many others). In this stream, the need for richer and more general choice models has also been recognized and some propositions have been made, such as Talluri and van Ryzin (2004), Alptekinoğlu and Semple (2015), Blanchet et al. (2016), Srikanth and Rusmevichientong (2017). Our paper complements this literature by offering a new, general, and versatile choice model that is derived from an analysis of the optimal behaviour of an individual consumer.

INSIGHTS ON CHOICE BEHAVIOR. The detailed assessment of information costs in our framework allows insight into the attention allocation strategy of the consumer, which drives the ultimate choice behavior. By comparison with the case of uniform costs, we show that information cost differences among the alternatives have substantial impact on the optimal choice of consumers. Naturally, the consumer pays more attention to and processes more information about alternatives with lower costs. If the alternatives are otherwise identical, this implies a strict preference for the one with the lowest cost. However, in general, the information obtained from the "cheaper" alternatives can increase or decrease the likelihood of choosing other alternatives. In this sense, reducing the cost of an alternative does not always mean it will be selected more often, nor does it imply better choices for the consumer overall. As a matter of fact, it can lead to new decision biases, implying that a uniform information provisioning strategy can be preferable in such cases. In a similar vein, the addition of a new alternative to the choice set can steer demand to or away from existing alternatives, depending on the information costs and consumer beliefs. Moreover, our choice model has some desirable properties. For example, it does not suffer from IIA (independence of irrelevant alterna-

tives), duplicate (identical) alternatives are jointly processed as one. Furthermore, dominated alternatives are never selected. However, their presence can influence choice behavior. Specifically, we show that if a dominated alternative has a lower information cost, its addition to a set can result in the failure of regularity by increasing the choice probability of an existing alternative. By providing a precise description of rational choice behavior under limited attention and costly information, our model has the potential to guide product assortment and information provisioning strategies of firms.

A central element of our model is the derivation of the total cost associated with the consumer's information processing strategy. Quantifying the amount of information the consumer acquires when evaluating a particular alternative and accounting for its cost is an intricate task in the presence of a non-uniform information cost structure, and this gets even more pronounced when there are similarities (i.e., correlations) between the products. This is because, as the consumer learns about a product, she may also learn about another product (and vice versa). Accordingly, there are two forms of information acquired by the consumer: (i) direct information that the consumer obtains by studying the particular alternative, and (ii) implied information that the consumer acquires about the alternative by studying another alternative. Since the unit costs of these sources might differ, it becomes important to glean from the consumer's information processing strategy the amount of information acquired from each source. The consumer should prioritize cheaper sources of information and should not attempt to obtain information about an alternative directly if that information can already be inferred from previously studied alternatives. We develop an information cost function that quantifies separately the amount of implied and direct information, and generalizes the Shannon entropy based cost functions utilized in the rational inattention literature.

We formulate the consumer's discrete choice problem based on this information cost function, and then characterize the structure of the optimal solution. We show that the optimal choice behaviour can be described analytically. Our choice model generalizes the GMNL model in the sense that it reduces to GMNL when the cost of information is uniform across all alternatives. After establishing this result, we concentrate

on a number of limiting cases and then provide two auxiliary examples to illustrate the impact of multiple information channels and different costs on the choice behavior of consumers with limited attention. In addition, utilizing the necessary and sufficient conditions, we develop an algorithm to determine the optimal choice probabilities. Finally, we discuss empirical and experimental estimation and validation of our choice model.

2. Overview of Literature

Starting with the seminal work of Stigler (1961), the fact that customers need to exert costly effort in order to acquire and process information about different alternatives has been widely investigated in the context of information search models. In sequential search models, consumers gather information about the value of alternatives one-by-one (or gradually learn about a particular product, possibly one attribute at a time), and make the choice decision once they optimally decide to stop collecting more information (e.g., Weitzman 1979, Branco et al. 2012, Ke et al. 2016). Rational inattention models differ in that no assumption is made on the process by which the consumer gets informed nor on the type or quantity of information acquired , i.e., the information strategy is fully endogenized. In parallel search models, consumers first determine the fixed set of products (commonly referred to as the consideration set) about which to gather information and then make a choice among them (e.g., Roberts and Lattin 1991, Manzini and Mariotti 2014), whereas rationally inattentive consumers keep all alternatives on the table during the decision process. At optimality, it is possible that only a subset of alternatives are chosen with positive probability, the rest are not considered at all. In this sense, rational inattention leads to an endogenously defined consideration set among all available alternatives (Caplin et al. 2016a). Also it is worth remarking that in most sequential and parallel search models, upon paying "search" costs, consumers resolve the associated uncertainty. In contrast, in rational inattention models, some uncertainty always remains after the consumer incurs the cost of the information (which is optimally acquired).

As noted earlier, the theory of rational inattention belongs to the literature of bounded rationality and receives significant interest in economics

(Gabaix 2014) as well as psychology (Todd and Gigerenzer 2000). In extant related literature, models differ in the way limited information is acquired and processed. For example, in Reis (2006), available information is attended to only sporadically, while in Verrecchia (1982) the consumer decides on the degree of the precision to which she receives information (the variance of the signal). Further, Stoneman (1981) proposes a Bayesian learning framework to update consumers' expectations of product attributes based on noisy signals – this particular approach has been commonly used in the marketing literature to capture salient effects of word-of-mouth, usage experience, product trials, and advertising exposure (e.g., Erdem and Keane 1996, Ackerberg 2003, Narayanan et al. 2005, Ching et al. 2014). The models conceived by Sims (1998) and later adopted by many other researchers generalize these approaches, as they offer the consumer the opportunity to receive signals of any type and to improve her prior in every desirable way. The common feature is the modeling of the cost of information as a reduction of uncertainty with respect to the prior, where uncertainty is measured as Shannon entropy (Shannon 1948). Our paper follows this prominent approach to modeling rational inattention.

To our knowledge, Matějka and McKay (2015) is the first application of rational inattention to discrete choice. Closely related, yet with a stronger focus on the posterior beliefs induced by rationally inattentive choice, is the work of Caplin and Dean (2015). Testable behavioral implications are studied in Caplin and Dean (2013). Caplin et al. (2016b) provide a dynamic model where consumers observe previous market shares and refine their beliefs over time. In steady state, market shares coincide with choice probabilities and this connection enables estimation from market share data. We expand the above noted literature by incorporating a non-uniform information cost structure to the choice decision. Our work also contributes to the research agenda laid out by Sims (2006).

3. Choice Model Formulation

In this section, we develop the choice model for a rationally inattentive consumer with different information costs across different products. The general formulation is preceded by an introductory example, which we

revisit in §5.1 and §7.2. We refer the reader to Appendix A for a list of symbols.

3.1 Example 1: Tripartite Race

Consider a setting that resembles a "race" among three alternatives, namely $\{a, b, c\}$, the consumer is choosing from. There are also three equally probable states of the world – each representing the winning alternative, which provides a utility of $v_i > 0$, $i \in \{a, b, c\}$, to the consumer when chosen. The consumer receives zero payoff if she picks a wrong alternative. Specifically, the state space is

$$\boldsymbol{\Omega} = \Omega_a \times \Omega_b \times \Omega_c, \text{ where } \Omega_a = \{0, v_a\},\ \Omega_b = \{0, v_b\},\ \Omega_c = \{0, v_c\},$$

and only those states with exactly one positive value, $\bar{\boldsymbol{\Omega}} = \{(v_a, 0, 0), (0, v_b, 0), (0, 0, v_c)\}$, have non-zero probabilities according to the prior belief g:

$$g(v_a, 0, 0) = g(0, v_b, 0) = g(0, 0, v_c) = \frac{1}{3}, \text{ while } g(\boldsymbol{\omega}) = 0 \text{ for } \boldsymbol{\omega} \in \boldsymbol{\Omega} \setminus \bar{\boldsymbol{\Omega}}.$$

The utility of choosing i in state $\boldsymbol{\omega}$ is given by $u(i, \boldsymbol{\omega}) = \omega_i$.

The consumer can process information with the goal of sharpening her belief about the state of the world, and consequently improving her decision. She does so by choosing an information acquisition strategy f, according to which she receives signals, and conducts a Bayesian updating. Let $\mathbf{S}$ denote the signal space available to the consumer. Consistent with the theory of rational inattention, we place no restriction on the way the consumer learns and allow her to set up any joint distribution $f \in \Delta(\boldsymbol{\Omega} \times \mathbf{S})$ of states and signals, as long as it is consistent with her prior belief, i.e., the marginal of f with respect to $\boldsymbol{\Omega}$ must equal the prior g,

$$\sum_{\mathbf{s}} f(\boldsymbol{\omega}, \mathbf{s}) = g(\boldsymbol{\omega}) \qquad \text{for all } \boldsymbol{\omega} \in \boldsymbol{\Omega}. \tag{1}$$

In order to illustrate the benefits and costs associated with an information strategy, consider now a given strategy f. Suppose that f is to learn the value of alternative a for sure, but nothing more about alternative b and alternative c. This is implemented by a signal space containing two signals $\mathbf{s}'$ and $\mathbf{s}''$ that correspond to Ω_a being v_a or 0, respectively. Being consistent

with the prior belief, such an information strategy f assigns the following joint probabilities to states and signals:

$$f((v_a,0,0),\mathbf{s}')=f((0,v_b,0),\mathbf{s}'')=f((0,0,v_c),\mathbf{s}'')=\frac{1}{3},$$

i.e., the signals occur with marginal probabilities $f(\mathbf{s}')=\frac{1}{3}$ and $f(\mathbf{s}'')=\frac{2}{3}$, and the posterior beliefs are $f(0,0,v_a\mid\mathbf{s}')=1$, and $f(\cdot\mid\mathbf{s}'')=\frac{1}{2}$ for both of the states $(0,v_b,0)$ and $(0,0,v_c)$.

The consumer chooses the alternative that gives the highest expected value based on the updated belief. Hence, if signal $\mathbf{s}'$ is realized, she chooses a; otherwise she chooses b or c, depending on which alternative is better when winning. The expected payoff associated with this information strategy is $R(f)=\frac{1}{3}(v_a+\max\{v_b,v_c\})$. Acquiring and processing this information is costly though. Specifically, the information cost $C(f)$ associated with the strategy f depends on the extent of the reduction of uncertainty achieved, as measured by Shannon entropy H. In particular, the a priori entropy is

$$H_g(\mathbf{\Omega})=\sum_{\boldsymbol{\omega}} g(\boldsymbol{\omega})(-\log g(\boldsymbol{\omega}))\approx 1.099,$$

where $0\cdot\log 0:=0$. Receiving signal $\mathbf{s}'$ leaves no posterior entropy, $H_{f(\cdot|\mathbf{s}')}(\mathbf{\Omega})=0$, while receiving signal $\mathbf{s}''$ reduces the posterior entropy to $H_{f(\cdot|\mathbf{s}'')}(\mathbf{\Omega})\approx 0.693$. The expectation of this reduction over all signals is called the *mutual information* $I_f(\mathbf{\Omega},\mathbf{S})$ between $\mathbf{S}$ and $\mathbf{\Omega}$ under the joint distribution f,

$$I_f(\mathbf{\Omega},\mathbf{S})=H_g(\mathbf{\Omega})-\sum_{\mathbf{s}} f(\mathbf{s})H_{f(\cdot\,|\mathbf{s})}(\mathbf{\Omega}). \tag{2}$$

Accordingly, we have $I_f(\mathbf{\Omega},\mathbf{S})\approx 0.637$ for our example information strategy.

Extant literature assumes that the cost per unit of mutual information $\lambda\geq 0$ is uniform across all alternatives, and accordingly defines the total cost of an information strategy f as $C(f)=\lambda I_f(\mathbf{\Omega},\mathbf{S})$. The founding premise of our work is that the unit cost of information acquisition can vary among alternatives, which we denote as λ_i for alternative $i\in\{a,b,c\}$. Without loss of generality, suppose that $\lambda_a\leq\lambda_b\leq\lambda_c$. As evident from our example, in computing the total cost of information, it is necessary that we account for the fact that learning about one alternative implies learning

about the other alternatives as well. If signal s′ is realized, the consumer is also fully informed about the values of alternatives b and c, whereas if s″ is realized, the chance of alternative b (of alternative c) being the right alternative needs to be adjusted from $\frac{1}{3}$ to $\frac{1}{2}$. Specifically, this information strategy f removes all uncertainty $H_g(\Omega_a)$ concerning alternative a, such that $I_f(\Omega_a, \mathbf{S}) = H_g(\Omega_a) \approx 0.637$. On the other hand, the entropy of Ω_b and Ω_c is also expected to decrease, by about 0.174, respectively, i.e., $I_f(\Omega_b, \mathbf{S}) = I_f(\Omega_c, \mathbf{S}) = 0.174$. However, all that is learned about b and c in this case falls into the category of implied information. Note that the total information acquired does not equal the sum of the information obtained about the separate alternatives, i.e, $I_f(\mathbf{\Omega}, \mathbf{S}) \neq I_f(\Omega_a, \mathbf{S}) + I_f(\Omega_b, \mathbf{S}) + I_f(\Omega_c, \mathbf{S})$. This fact indeed prevents us from directly attributing unit cost λ_i to information acquired about alternative i.

We propose a total cost of information $C(f)$ that asserts the following: the information about the easiest-to-learn alternative a can be acquired at unit cost λ_a. What is implied from this about other alternatives does not come at additional information costs. All that is learned about alternatives a and b together, but not implied from information about alternative a, is direct information obtained about b and thus costs λ_b per unit. All information that goes beyond alternatives a and b is acquired at unit cost λ_c. Accordingly, we express $C(f)$ as

$$C(f) = \lambda_a I_f(\Omega_a, \mathbf{S}) + \lambda_b(I_f(\Omega_a \times \Omega_b, \mathbf{S}) - I_f(\Omega_a, \mathbf{S})) + \lambda_c(I_f(\mathbf{\Omega}, \mathbf{S}) - I_f(\Omega_a \times \Omega_b, \mathbf{S})).$$

This cost function is an intuitive generalization which ensures that information about an alternative is acquired from the cheapest available source, and that any implied information is not re-acquired and processed by the consumer. We further remark that the above example is a special case as there are only three states, each distinguishing the winning alternative. Intuitively, irrespective of the information strategy, nothing can be learned about alternative c beyond what will be implied through information acquired about alternatives a and b. Hence, in this example it must the that $I_f(\mathbf{\Omega}, \mathbf{S}) = I_f(\Omega_a \times \Omega_b, \mathbf{S})$ for any information strategy f, implying $C(f) = \lambda_a I_f(\Omega_a, \mathbf{S}) + \lambda_b(I_f(\Omega_a \times \Omega_b, \mathbf{S}) - I_f(\Omega_a, \mathbf{S}))$, which is indeed the case.

For the specific information strategy considered in our example, the information cost $C(f)$ simply equals $0.637\lambda_a$, since in this case all that is learned about alternatives b and c is implied from information obtained about a, i.e., $I_f(\mathbf{\Omega}, \mathbf{S}) - I_f(\Omega_a, \mathbf{S}) = 0$. Note the immediate connection with a traditional search model where the consumer chooses to learn the value of alternative a (upon paying a fixed search cost of $\lambda_a H_g(\Omega_a)$), or not. Needless to say, the consumer may prefer a very different information strategy in general, possibly relying on very complex signals. In fact, we allow the consumer to determine the optimal information strategy f that maximizes $R(f) - C(f)$, which typically is not fully informative (neither about a nor about b). It further turns out that at optimality, it suffices for the consumer to use as many signals as there are alternatives, each pointing to the corresponding alternative as being the best. This central connection enables the characterization of the optimal choice probabilities without referring to the information strategy. We formalize these observations for the general case in the next section.

3.2 General Choice Model Formulation

The consumer chooses an alternative i from a finite set A. The state of the world is a random variable $\mathbf{\Omega} = (\Omega_1 \times \ldots \times \Omega_k \times \ldots \times \Omega_n)$ taking values $\boldsymbol{\omega} \in \mathbb{R}^n$. Picking alternative i in state $\boldsymbol{\omega}$ yields finite utility $u(i, \boldsymbol{\omega}) \in \mathbb{R}$. The consumer has the prior belief $g \in \Delta(\mathbf{\Omega})$, where $\Delta(X)$ denotes the set of all probability distributions on X. The consumer processes information to sharpen her belief about the state of the world for an improved decision. Let λ_k denote the unit cost of acquiring information directly about the k^{th} component of the state space, i.e., about Ω_k. Without loss of generality, suppose that $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. Given the signal space $\mathbf{S} \subseteq \mathbb{R}^n$ that is available, the consumer sets up any joint distribution $f \in \Delta(\mathbf{\Omega} \times \mathbf{S})$ of states and signals that is consistent with her prior belief.

Clearly, depending on information strategy, i.e. the joint distribution f, the signal can be more or less beneficial (informative). In particular, given signal $\mathbf{s}$, the consumer creates an updated belief $f(\cdot \mid \mathbf{s}) \in \Delta(\mathbf{\Omega})$ over the state of the world, and chooses the alternative that yields the highest expected value based on this updated belief. The less noise that remains in the updated belief, the more promising this choice becomes. The result-

ing expected payoff $R(f)$ is then[3]

$$R(f)=\sum_{\boldsymbol{\omega}} g(\boldsymbol{\omega})\underbrace{\sum_{\mathbf{s}} f(\mathbf{s}\mid\boldsymbol{\omega})\underbrace{\max_{i\in A}\sum_{\boldsymbol{\omega}'} f\left(\boldsymbol{\omega}'\mid\mathbf{s}\right)u\left(i,\boldsymbol{\omega}'\right)}_{\substack{\text{expected utility after receiving}\\ \text{signal }\mathbf{s}\text{, holding belief } f(\cdot\mid\mathbf{s})}}}_{\text{anticipated utility if state }\boldsymbol{\omega}\text{ is realized}}, \tag{3}$$

where $f(\mathbf{s}\mid\boldsymbol{\omega})$ denotes the conditional probability (implied by the information strategy f) that the consumer observes the signal $\mathbf{s}$ if the state of the world is $\boldsymbol{\omega}$.

The most crucial element of the rationally inattentive consumer's choice framework with non-uniform information costs is the development of a total information cost function $C(f)$ based on mutual information $I_f(\boldsymbol{\Omega},\mathbf{S})$ that accounts for different costs across components of the state space. Conceptually, $I_f(\boldsymbol{\Omega},\mathbf{S})$ is generated from a series of queries and their responses. Practically, this is tantamount to consumers studying the states in some order, asking questions and updating beliefs accordingly. We do not specify the exact process by which information is acquired. However, as highlighted in the previous section, correlation across the components of the state space naturally implies that whenever something is learned about a particular component, information about other components is acquired as well. Indeed, even if the Ω_k's are independent according to the prior belief, it is typical that the rationally inattentive consumer designs her information strategy in such a way that they do become conditionally dependent on the signal. Consequently, the total information acquired does not equal the sum of the information acquired about the individual components, i.e., $\sum_{k=1}^{n} I_f(\Omega_k,\mathbf{S}) \neq I_f(\boldsymbol{\Omega},\mathbf{S})$. Nevertheless, it must be that for any Ω_k, the customer, whenever possible, should infer information acquired at lower unit costs, and only then attempt to acquire information directly at unit cost λ_k. This leads to the following total cost of information:

$$C(f)=\sum_{k=1}^{n}\lambda_k\left(I_f(\boldsymbol{\Omega}_{1\cdots k},\mathbf{S})-I_f(\boldsymbol{\Omega}_{1\cdots k-1},\mathbf{S})\right), \tag{4}$$

where $\boldsymbol{\Omega}_{1\cdots k}=\Omega_1\times\Omega_2\times\cdots\times\Omega_k$. We remark that $C(f)$ given by (4) can

[3]To simplify exposition, we adopt countable signal and state spaces in our notation. The results herein, however, can be generalized to continuous spaces.

be justified from first principals, using an axiomatic approach which postulates how the cost of two information strategies differ. In particular, the key assumption that generates (4) specifies that the gain in information (from switching strategies) inherent to any combination of Ω_k's comes at the maximal unit cost λ_k associated with this combination. It can then be established that $C(f)$ is the unique cost function satisfying this assumption. We omit these details in the interest of space.

Applying the chain rule of mutual information

$$I_f\left(\mathbf{\Omega}_{1..k}, \mathbf{S}\right) = \sum_{\ell=1}^{k} I_f\left(\Omega_\ell, \mathbf{S} \mid \mathbf{\Omega}_{1..\ell-1}\right)$$

on (4) results in the following more tractable representation of the total information cost.

Definition Let the unit costs of information associated with the components of the state space $\Omega_1, \Omega_2, \ldots, \Omega_n$ be given by $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$. The cost of information $C(f)$ of an information strategy $f \in \Delta\left(\mathbf{\Omega} \times \mathbf{S}\right)$ can be expressed as

$$C\left(f\right) = \sum_{i=1}^{n} \lambda_i I_f\left(\Omega_k, \mathbf{S} \mid \mathbf{\Omega}_{1..k-1}\right), \tag{5}$$

where the conditional mutual information $I_f\left(\Omega_k, \mathbf{S} \mid \mathbf{\Omega}_{1..k-1}\right)$ is the expected mutual information between Ω_k and $\mathbf{S}$ conditional on $\mathbf{\Omega}_{1..k-1}$,

$$I_f\left(\Omega_k, \mathbf{S} \mid \mathbf{\Omega}_{1..k-1}\right) = \sum_{\boldsymbol{\omega}_{1..k-1} \in \mathbf{\Omega}_{1..k-1}} f\left(\boldsymbol{\omega}_{1..k-1}\right) I_{f((\cdot,\cdot)|\boldsymbol{\omega}_{1..k-1})}\left(\Omega_k, \mathbf{S}\right).$$

Note that $\lambda_k = \lambda$ for all $i \in A$ implies $C\left(f\right) = \lambda I_f\left(\mathbf{\Omega}, \mathbf{S}\right)$ and the total information cost function reduces to that in extant literature. In this sense, the cost function (5) generalizes the Shannon entropy based information cost functions utilized in the literature to non-uniform information cost structures. With this cost function, the consumer can determine her optimal information strategy

Consumer's optimization problem Find an information strategy f that solves:

$$\max_{f \in \Delta(\mathbf{\Omega} \times \mathbf{S})} \quad R\left(f\right) - C\left(f\right) \tag{6}$$

$$\text{s. t.} \quad \sum_{\mathbf{s}} f\left(\mathbf{s}, \boldsymbol{\omega}\right) = g\left(\boldsymbol{\omega}\right) \qquad \text{for all } \boldsymbol{\omega}, \tag{7}$$

where $R\left(f\right)$ and $C\left(f\right)$ are given in (3) and (5), respectively.

3.3 Implications of an Optimal Information Strategy

The standard approach in the rational inattention literature, which has also been applied to the case of uniform information cost structure (Sims 2003, Matějka and McKay 2015), realizes that the optimal information strategy f may not resolve all uncertainty about the state of the world, and therefore would result in probabilistic choice behavior, captured by a joint probability distribution $p \in \Delta(\mathbf{\Omega} \times A)$, where A is the random variable that takes value $i \in A$ with the probability that alternative i is chosen. It is then shown that choosing the optimal information strategy is equivalent to a problem of selecting optimal choice probabilities. We now adapt and extend this argument to the case of non-uniform information cost structures.

In Appendix B, we argue that essentially a single signal for each chosen alternative is sufficient to implement an optimal information strategy (here, a "signal" can be complex in nature and represent an involved learning outcome). More specifically, Lemma 2 stipulates that under an optimal information strategy, for every signal that leads to the choice of a particular alternative i, the posterior belief about the state of the world conditional on the reception of this signal must be the same, i.e., $\mathbf{s}', \mathbf{s}'' \in \{\mathbf{s} \in \mathbf{S} \mid i = \arg\max_{j \in A} \sum_{\boldsymbol{\omega}} f(\boldsymbol{\omega} \mid \mathbf{s}) u(j, \boldsymbol{\omega})\}$ implies $f(\cdot \mid \mathbf{s}'') = f(\cdot \mid \mathbf{s}')$. The intuition is that if an information strategy leads to the choice of a particular alternative with distinct posterior beliefs about the state of the world, then the consumer would have processed "unnecessary" information and hence such a strategy would not be optimal. This follows from the fact that the choice has not improved, but the cost of the information strategy with distinct posteriors is higher. Eventually, the one-to-one relationship between the action and posterior belief means that the reception of the signal is as informative about the state of the world as is the observation of what alternative is chosen. Consequently, we can replace the mutual information between signal and state by the mutual information between action and state, $I_f(\Omega_k, \mathbf{S} \mid \mathbf{\Omega}_{1..k-1}) = I_p(\Omega_k, A \mid \mathbf{\Omega}_{1..k-1})$. This allows us to formulate the consumer's optimization problem in terms of choice probabilities.

Proposition 1 (Problem without reference to information strategy) *The set of conditional choice probabilities* $\{p(i \mid \boldsymbol{\omega})\}_{i\in A,\boldsymbol{\omega}\in\boldsymbol{\Omega}}$ *is implied by an optimal information strategy of an inattentive consumer, which optimally solves the problem in* (6)-(7)*, if and only if it is a solution to the following problem:*

$$\begin{aligned} \underset{\{p(i|\boldsymbol{\omega})\}_{i\in A,\boldsymbol{\omega}\in\boldsymbol{\Omega}}}{\text{maximize}} \quad & \sum_{i\in A}\sum_{\boldsymbol{\omega}\in\boldsymbol{\Omega}} u(i,\boldsymbol{\omega})\, p(i \mid \boldsymbol{\omega})\, g(\boldsymbol{\omega}) - \sum_{k=1}^{n} \lambda_k I_p(\Omega_k, A \mid \boldsymbol{\Omega}_{1..k-1}) \\ \text{subject to} \quad & p(i \mid \boldsymbol{\omega}) \geq 0 \qquad \textit{for all } i \in A \textit{ and } \boldsymbol{\omega} \in \boldsymbol{\Omega} \\ & \sum_{i\in A} p(i \mid \boldsymbol{\omega}) = 1 \qquad \textit{for all } \boldsymbol{\omega} \in \boldsymbol{\Omega}, \end{aligned}$$

where conditional mutual information is given by

$$\begin{aligned} & I_p(\Omega_k, A \mid \boldsymbol{\Omega}_{1..k-1}) \\ &= \sum_{i\in A}\sum_{\boldsymbol{\omega}_{1..k}\in\boldsymbol{\Omega}_{1..k}} p(i \mid \boldsymbol{\omega}_{1..k}) \left(\log p(i \mid \boldsymbol{\omega}_{1..k}) - \log p(i \mid \boldsymbol{\omega}_{1..k-1})\right) g(\boldsymbol{\omega}_{1..k}) \end{aligned} \tag{8}$$

and the (partial conditional) choice probabilities are given by

$$p(i \mid \boldsymbol{\omega}_{1..k}) = \sum_{\boldsymbol{\omega}_{k+1..n}\in\boldsymbol{\Omega}_{k+1..n}} p(i \mid \boldsymbol{\omega}_{1..k}\boldsymbol{\omega}_{k+1..n})\, g(\boldsymbol{\omega}_{k+1..n} \mid \boldsymbol{\omega}_{1..k}). \tag{9}$$

The advantage of Proposition 1 is the reduction of a very complex problem of finding the optimal information strategy to a more tractable problem of finding the implied optimal choice probabilities. Note the "as-if"-notion of this result: the consumer is not actually assumed to optimize choice probabilities, but using an optimal information strategy is *behaviorally equivalent* to the optimal choice probabilities.

4. Optimal Choice

We now solve the consumer's optimization problem stated in Proposition 1. To this end, we first study the necessary conditions and state the ensuing conditional choice probabilities. We then explore the consequences for some limiting cases. Finally, we discuss implications on posterior beliefs.

4.1 Necessary Conditions on Conditional Choice Probabilities

It can be verified that the consumer's problem in Proposition 1 is a concave optimization problem over a compact set. This follows from the fact that $I_p(\Omega_k, A \mid \boldsymbol{\Omega}_{1..k-1})$ is convex in the decision variables $\{p(i \mid \boldsymbol{\omega})\}_{i \in A}$, which can be established via Theorem 2.7.4 in Cover and Thomas (2006). Moreover, $p(i \mid \boldsymbol{\omega}) = 0$ only if $p(i \mid \boldsymbol{\omega}_{1..n-1}) = 0$. Treating this case separately allows us to obtain the structure of the optimal solution from the first order conditions of the Lagrangian. We formalize the result in the following theorem.

Theorem 1 (Necessary conditions) *For any information cost structure* $0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n < \infty$, *the consumer forms her information strategy such that the optimal conditional choice probabilities satisfy*

$$p(i \mid \boldsymbol{\omega}) = \frac{e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}} p(i)^{\frac{\lambda_1}{\lambda_n}} \prod_{k=1}^{n-1} p(i \mid \boldsymbol{\omega}_{1..k})^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}}{\sum_{j \in A} e^{\frac{u(j,\boldsymbol{\omega})}{\lambda_n}} p(j)^{\frac{\lambda_1}{\lambda_n}} \prod_{k=1}^{n-1} p(j \mid \boldsymbol{\omega}_{1..k})^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}} \tag{10}$$

for all $i \in A$ *and all* $\boldsymbol{\omega}$ *such that* $g(\boldsymbol{\omega}) > 0$, *where* $p(i) = \sum_{\boldsymbol{\omega}} p(i \mid \boldsymbol{\omega}) g(\boldsymbol{\omega})$ *is the unconditional probability and* $p(i \mid \boldsymbol{\omega}_{1..k})$ *given by* (9) *is the partial conditional probability of choosing* $i \in A$.

We remark that (10) is trivially true for unchosen alternatives, i.e., if $p(i) = 0$. It is also imperative to note that $p(i)$ and $p(i \mid \boldsymbol{\omega}_{1..k})$ are not exogenous parameters; they are implied by the $p(i \mid \boldsymbol{\omega})$'s, and as such, are part of the consumer's decision making strategy, capturing the effect of prior beliefs.

In order to gain more intuition on the optimal choice probabilities, consider the special case when information costs are identical. Indeed, plugging $\lambda_1 = ... = \lambda_n = \lambda$ into (12) yields

$$p(i \mid \boldsymbol{\omega}) = \frac{e^{\frac{u(i,\boldsymbol{\omega})}{\lambda}} p(i)}{\sum_{j \in A} e^{\frac{u(j,\boldsymbol{\omega})}{\lambda}} p(j)}, \tag{11}$$

which is precisely the Generalized Multinomial Logit (GMNL)[4] as defined by Matějka and McKay (2015). Evidently, our choice model reduces to the

[4]When the consumer is a-priori indifferent to all alternatives (i.e., $g(\boldsymbol{\omega})$ is invariant to

GMNL specification when information costs are uniform. On the other hand, it is more general in the sense that when information costs differ, the conditional choice probabilities are swayed by prior beliefs $g(\boldsymbol{\omega})$ not only through the unconditional probabilities $p(i)$, but also via the *partially* conditional choice probabilities $p\left(i \mid \boldsymbol{\omega}_{1..k}\right)$'s of selecting each alternative. To substantiate our understanding of such implications, we rewrite (10) as

$$p\left(i \mid \boldsymbol{\omega}\right) = \frac{e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}+\alpha_i}}{\sum_{j\in A} e^{\frac{u(j,\boldsymbol{\omega})}{\lambda_n}+\alpha_j}}, \tag{12}$$

where we define α_i as

$$\alpha_i = \frac{\lambda_1}{\lambda_n}\log p(i) + \sum_{k=1}^{n-1}\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}\log p\left(i \mid \boldsymbol{\omega}_{1..k}\right). \tag{13}$$

Written this way, the conditional probabilities follow a formula similar to the standard MNL, with the payoff of each alternative shifted by the term α_i. For the GMNL in (11), α_i simply equals $\log p(i)$, implying that if an alternative is generally attractive across all states, i.e., if $p(i)$ is relatively high, it can still be chosen with high probability even if its true value $u(i,\boldsymbol{\omega})$ in a particular state $\boldsymbol{\omega}$ is low (Matějka and McKay 2015).

When the information costs are non-uniform, the consumer will typically know more about easier-to-learn alternatives, and this is reflected in the computation of how "attractive" an alternative is, beyond the utility it delivers in the realized state $\boldsymbol{\omega}$. Specifically, the shift term (13) becomes a weighted average of the log transformations of the partial conditional and the unconditional choice probabilities. The partial conditional probability $p\left(i \mid \boldsymbol{\omega}_{1..k}\right)$ represents the overall likelihood of choosing an alternative i, based on what is learned from studying the k most accessible alternatives. Hence, it is possible that a generally attractive alternative (with a relatively high $p(i)$) can be chosen with a low probability if the information obtained from studying the component with low information cost (say Ω_1) implies a low selection probability $p\left(i \mid \omega_1\right)$, even if the true value $u(i,\boldsymbol{\omega})$ is high. To what extend this happens depends also on the relative values of the

all permutations of the elements of $\boldsymbol{\omega}$), then $p(i) = \frac{1}{|A|}$ and (11) reduces to the standard MNL formula.

information costs. To clarify this, consider the simplest case with two alternatives ($\lambda_1 < \lambda_2$), where the shift term is

$$\alpha_i = \frac{\lambda_1}{\lambda_2} \log p(i) + \left(1 - \frac{\lambda_1}{\lambda_2}\right) \log p\left(i \mid \omega_1\right).$$

Suppose λ_1 is fixed but λ_2 increases. Then, the impact of the attractiveness of alternative i after accounting for the cheaper information $p\left(i \mid \omega_1\right)$ gains more weight relative to the impact of the overall attractiveness $p(i)$ that also comprises judgement about the more costly information. Furthermore, from (12), it is evident that the utility $u(i, \omega_1, \omega_2)$ obtained from choosing alternative i in the realized state (ω_1, ω_2) becomes less relevant compared to the attractiveness $p\left(i \mid \omega_1\right)$ of alternative i in just state ω_1. This indicates that the consumer relies less on the utility she can get from the particular state ω_2 and instead focuses on the utility she expects to get across all states ω_2.

4.2 Limiting Scenarios

In the previous section, we characterized the optimal behaviour of customers for the most general case involving distinct alternatives with non-zero and finite information costs. There are some limiting scenarios that do not directly follow from the conditional choice probability equation (10) in Theorem 1. In this section, we focus on four such scenarios – infinite and zero information cost for some alternatives, and duplicate and dominated alternatives. Delving into these limiting cases also sheds some light on how non-uniform information costs impact the choices of inattentive consumers.

ZERO INFORMATION COST (AND DETERMINISTIC ALTERNATIVES). Suppose that the consumer can freely process all information for the first alternative, i.e, $\lambda_1 = 0$. This could represent a product for which the customer can assign a true value very easily (e.g. a simple search good). Then, (12) becomes

$$\alpha_i = \frac{\lambda_2}{\lambda_n} \log p\left(i \mid \omega_1\right) + \sum_{k=2}^{n-1} \frac{\lambda_{k+1} - \lambda_k}{\lambda_n} \log p\left(i \mid \boldsymbol{\omega}_{1..k}\right).$$

Note that if the utility provided by the first alternative is deterministic, i.e., $\Omega_1 = \omega_1$ with probability 1, then $p(i) = p(i \mid \omega_1)$ irrespective of λ_1, and consequently (12) also reduces to the above expression. Accordingly, this expression applies also for deterministic alternatives (This could represent a product about which the consumer is well-informed due to past experience, or the no-purchase alternative with a particular reservation value.) We remark however that while the functional form of the conditional choice probabilities for a particular state $\boldsymbol{\omega}$ are the same, the choice behaviour for zero information costs is richer as it changes with ω_1.

INFINITE INFORMATION COST. Suppose that it is infinitely costly (or prohibitively expensive) for the consumer to process information about the last alternative, i.e., $\lambda_n = \infty$. This could represent a product for which the customer is not willing to acquire any direct information, or for which such information is not obtainable (e.g. product is not offered/available). In this case, Theorem 1 no longer applies and its proof needs to be extended by an argument that no information is processed for infinitely costly states beyond what can be inferred from feasible information. Specifically, when $\lambda_n = \infty$, it is necessary that $I_p(\Omega_n, A \mid \boldsymbol{\Omega}_{1..n-1}) = 0$ in optimum to avoid an infinite information processing cost. Accordingly, the consumer does *not* update her priors beyond the information obtained from alternatives $1...n-1$. Then, from (8), $p(i \mid \boldsymbol{\omega}_{1..n-1}, \omega_n) = p(i \mid \boldsymbol{\omega}_{1..n-1})$ for any ω_n. Accordingly, the utility obtained from the n^{th} alternative enters the consumer's decisions process only in terms of its expectation, i.e., $u(n, \boldsymbol{\omega})$ is replaced with $\sum_{\omega_n} u(n, \boldsymbol{\omega}) g(\omega_n \mid \boldsymbol{\omega}_{1..n-1})$ for all $\boldsymbol{\omega}_{1..n-1}$. This is intuitive; since obtaining information about this alternative is not feasible, its attractiveness only depends on the utility it is expected to provide across all ω_n. All in all, (10) then reads

$$\begin{aligned} p(i \mid \boldsymbol{\omega}) &= p(i \mid \boldsymbol{\omega}_{1..n-1}) \\ &= \frac{e^{\frac{\sum_{\omega_n} g(\omega_n \mid \boldsymbol{\omega}_{1..n-1}) u(i,\boldsymbol{\omega})}{\lambda_{n-1}}} p(i)^{\frac{\lambda_1}{\lambda_{n-1}}} \prod_{k=1}^{n-2} p(i \mid \boldsymbol{\omega}_{1..k})^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_{n-1}}}}{\sum_{j \in A} e^{\frac{\sum_{\omega_n} g(\omega_n \mid \boldsymbol{\omega}_{1..n-1}) u(j,\boldsymbol{\omega})}{\lambda_{n-1}}} p(j)^{\frac{\lambda_1}{\lambda_{n-1}}} \prod_{k=1}^{n-2} p(j \mid \boldsymbol{\omega}_{1..k})^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_{n-1}}}}. \end{aligned}$$

DUPLICATE ALTERNATIVES. It has been shown that when the information costs are uniform, the resulting choice behavior of rationally inattentive consumers (i.e. GMNL) does not suffer from the IIA property. Specifically,

Matějka and McKay (2015) establish that duplicate alternatives are jointly treated as one alternative. Two alternatives are referred to as "duplicates" if they take the same values in all states of the world according to the prior belief of the customer. Suppose now that a duplicate alternative with a different information cost is added to a choice set. Since the values of the duplicate products are perfectly correlated, the consumer will only process information about the cheaper cost alternative and this would yield the exact same information about the other alternative. Hence, even if the individual information costs might differ, the consumer remains indifferent between the original and its duplicate. Consequently, the probability of choosing the original *or* its duplicate among available alternatives *exactly equals* the choice probability of the original alternative, provided that it is *not* cheaper for the consumer to process information about the added duplicate. This requirement is critical – if it is easier to acquire information about the duplicate, the optimal choice may differ since more information is likely to be processed due to the availability of a cheaper information source.

DOMINATED ALTERNATIVES (AND STRONG FAILURE OF REGULARITY). A closely related notion to duplicates is dominated alternatives. A dominated alternative is one for which the value is lower than another alternative in all states of the world. Such alternatives are never selected when the cost of information is uniform across alternatives (Matějka and McKay 2015). It can be easily verified that this extends to the more general case of differentiated information costs (shifting the choice probability from the dominated to the dominating alternative would increase the consumer's objective function). This does not mean information is not processed about a dominated alternative. As a matter of fact, whether a dominated alternative is available or not in the choice set can become relevant in the case of non-uniform information costs since it might serve as a cheap channel to learn about other alternatives and thereby influence their choice probabilities. This can lead to a "failure" of the regularity condition put forth by Luce and Suppes (1965), which requires that adding a product to the choice set does *not* increase the market share of another product. Matějka and McKay (2015) show that a rationally inattentive consumer facing uniform information costs might fail the regularity condition, but only if the added product has a positive chance of being selected. This is because in-

troducing a new product can set incentives for the consumer to get information about the new product in a way that she is also informed about a previously "uninteresting" product. With this additional information, she might identify cases where she buys a previously uninteresting product. If the new product is inferior (i.e., dominated), however, the consumer would completely disregard the new product and also would not process any information about it. Hence, there is *no failure of regularity* under uniform information costs when the added product is dominated. In contrast, non-uniform information costs can induce failure of regularity even if the inclusion is an inferior, never-selected alternative (hence our usage of the term *strong* failure of regularity).

4.3 Implications on Posterior Beliefs

Recall from §3.3 that an optimal information strategy leads to a unique posterior belief for every chosen alternative. Applying the Bayes' rule to conditional choice probabilities $p(i \mid \boldsymbol{\omega})$ characterized in Theorem 1, we can obtain the posterior belief $p(\boldsymbol{\omega} \mid i)$ the consumer holds upon choosing i, i.e., the probability that the consumer attributes to state $\boldsymbol{\omega}$ when her choice is i. Going a step further, taking the ratio of the posterior beliefs held for different alternatives, we can relate them to the utilities associated with these alternatives (as in Caplin and Dean 2013, 2015).

Corollary 1 (Invariant ratio of posterior beliefs) *For any two alternatives i and j such that $p(i), p(j) > 0$, and for all $\boldsymbol{\omega}$, the posterior beliefs satisfy*

$$\frac{e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}}}{e^{\frac{u(j,\boldsymbol{\omega})}{\lambda_n}}} = \frac{\prod_{k=1}^{n} p\left(\boldsymbol{\omega}_{1..k} \mid i, \boldsymbol{\omega}_{1..k-1}\right)^{\frac{\lambda_k}{\lambda_n}}}{\prod_{k=1}^{n} p\left(\boldsymbol{\omega}_{1..k} \mid j, \boldsymbol{\omega}_{1..k-1}\right)^{\frac{\lambda_k}{\lambda_n}}}. \tag{14}$$

This ratio is useful in explicitly solving the optimal choice probabilities for small examples and also for empirical estimation purposes. Observe that when information costs are uniform, (14) boils down to the invariant likelihood ratio (ILR) $e^{(u(i,\boldsymbol{\omega})-u(j,\boldsymbol{\omega}))/\lambda} = p(\boldsymbol{\omega} \mid i)/p(\boldsymbol{\omega} \mid j)$ derived and discussed in Caplin and Dean (2013).

5. Choice Behavior

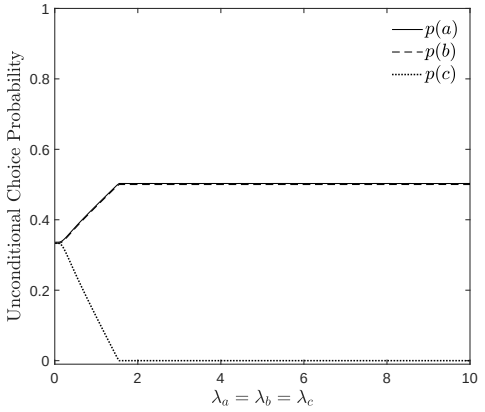
In this section, utilizing simple examples, we illustrate the impact of information costs on the optimal choice behavior of inattentive consumers. The first example is presumably the simplest setting with non-uniform information costs that can be solved in closed-form. The second example is the classic red-bus/blue-bus problem.

5.1 Example 1 (Revisited): Tripartite Race

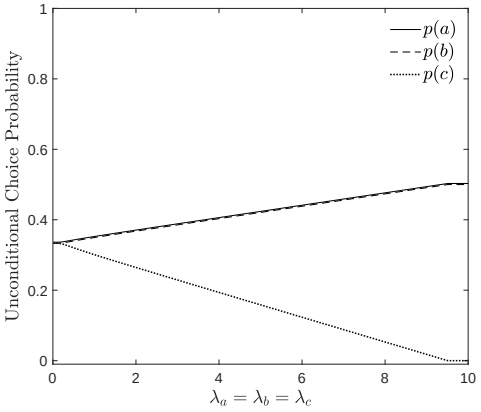
Reconsider the example from §3.1. We let $v_a = v_b = 1$ and $v_c \in \{0.75, 0.95\}$, capturing *distinct* versus *close* third alternative values – effectively, we consider two sets of possible states of the world, $\bar{\boldsymbol{\Omega}} = \{(1,0,0),(0,1,0),(0,0,0.75)\}$ and $\bar{\boldsymbol{\Omega}} = \{(1,0,0),(0,1,0),(0,0,0.95)\}$. Assuming $g(\boldsymbol{\omega}) = \frac{1}{3}$ for all $\boldsymbol{\omega} \in \bar{\boldsymbol{\Omega}}$, we then solve for $p(a)$, $p(b)$ and $p(c)$ in closed-form, utilizing (14) in Corollary 1. The derivations of the closed-form expressions can be found in Appendix C.

First, we study the case with uniform information costs ($\lambda_a = \lambda_b = \lambda_c = \lambda$). Since each alternative is the best only in one of the three states in $\bar{\boldsymbol{\Omega}}$, under full information ($\lambda = 0$), we have $p(i) = \frac{1}{3}$ for $i \in \{a,b,c\}$. In contrast, if the consumer does not process *any* information at all ($\lambda = \infty$), then she would choose either a or b, but never c, since its expected value is lower than the other two (actually she is indifferent to a and b, and here we assume she chooses each one with equal probability, i.e., $p(a) = p(b) = \frac{1}{2}$ and $p(c) = 0$). Figure 1 shows how the unconditional choice probabilities vary as a function of the information cost between these two extreme situations. We observe that the consumer can afford to process information at higher cost when c's value is close to that of a and b compared to the case when c's value is distinctly lower. Specifically, when $v_c = 0.75$, the consumer stops processing any information if $\lambda \geq 1.6$, and chooses between a and b with equal probability based on her priors, completely ignoring c due to its significantly lower value (Figure 1(a)). On the other hand, this only happens if $\lambda \geq 9.65$ when $v_c = 0.95$ (Figure 1(b)). Clearly, when the alternatives are more similar, the consumer keeps them on the table for a larger range of information costs and spends more effort to distinguish them. This suggests that a seller can steer demand towards a set of prod-

ucts by adding a somewhat less attractive product to the choice set, or divert demand away from the same set of products by adding a similarly attractive product to the choice set.



(a) $\bar{\mathbf{\Omega}} = \{(1,0,0),(0,1,0),(0,0,0.75)\}$



(b) $\bar{\mathbf{\Omega}} = \{(1,0,0),(0,1,0),(0,0,0.95)\}$

Figure 1: Unconditional choice probabilities for identical information costs ($\lambda_a = \lambda_b = \lambda_c$)

Next, we explore the case when information costs are non-uniform. Recall from §3.1 that for this three state problem, the consumer does not process any information about c. Hence, the unit cost λ_c does not matter, and we can assume $\lambda_a \leq \lambda_b = \lambda_c$ without loss of generality.

We initially focus on a limiting scenario and look into the unconditional choice probabilities as a function of λ_a when $\lambda_b = \lambda_c = \infty$. In these cases, the consumer never chooses c since in expectation it is inferior to b (and a); hence $p(c) = 0$, as illustrated in Figure 2. The only effect that is present is the asymmetry in the information costs between a and b. At a first glance, it might seem intuitively appealing that reducing the information cost of a should increase its choice probability, since the consumer would be able to more confidently assess it as the better alternative. However, this is not entirely correct – with reduced information cost, the consumer is also able to learn with more confidence the states in which a is *not* the best alternative. Therefore, $p(a)$ increases (monotonically from $\frac{1}{3}$ to $\frac{1}{2}$) and $p(b)$ decreases (monotonically from $\frac{2}{3}$ to $\frac{1}{2}$) with λ_a.

We now consider the general case with $\lambda_a \leq \lambda_b = \lambda_c < \infty$. Figures 3 and 4 depict the unconditional choice probabilities for $\lambda_a = 0.1$ and

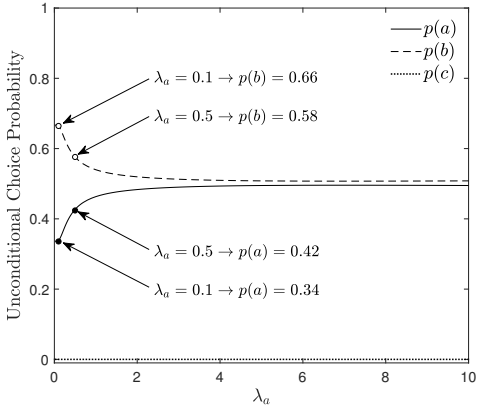
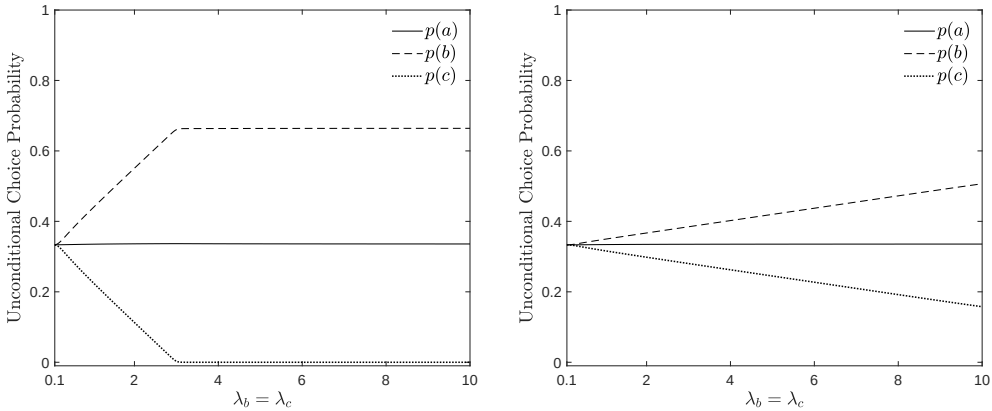


Figure 2: Unconditional choice probabilities when $\lambda_b = \lambda_c = \infty$

$\lambda_a = 0.5$, respectively, for $\bar{\mathbf{\Omega}} = \{(1,0,0),(0,1,0),(0,0,0.75)\}$ and $\bar{\mathbf{\Omega}} = \{(1,0,0),(0,1,0),(0,0,0.95)\}$. Note that in each figure, the first observation point coincides with the identical information costs case ($\lambda_a = \lambda_b = \lambda_c$). On the other hand, the limit values when $\lambda_b = \lambda_c = \infty$ follow directly from Figure 2 – $p(a) = 0.34$ and $p(b) = 0.66$ when $\lambda_a = 0.1$, and $p(a) = 0.42$ and $p(b) = 0.58$ when $\lambda_a = 0.5$.

In these cases, a second effect comes into the picture – the need for the consumer to distinguish between b and c – and this can have intricate consequences on the overall choice behavior of the consumer. Observe that if λ_a is relatively low ($\lambda_a = 0.1$), the consumer can acquire enough information to essentially "know" when a is the best alternative. Hence, $p(a) \approx \frac{1}{3}$ in Figures 3(a) and 3(b). As λ_b increases, the consumer acquires less information about b (and infers less about c), but increasingly prefers it over c as her decisions become more based on prior beliefs. Accordingly, $p(b)$ increases and $p(c)$ decreases with λ_b and $p(b)+p(c) \approx \frac{2}{3}$. When c's value is significantly lower than that of a and b ($v_c = 0.75$ compared to $v_c = 0.95$), we also observe that the consumer drops the inferior alternative c from consideration much more rapidly with increasing λ_b.

Compare the above to the case when λ_a is relatively high ($\lambda_a = 0.5$). Since λ_a is no longer negligible, the consumer cannot be very confident about the value of a and this has a strong impact on the choice behavior. In particular, when c is a viable alternative for the consumer, choosing



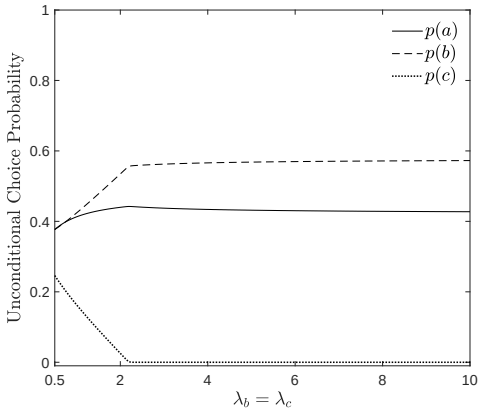
(a) $\bar{\mathbf{\Omega}} = \{(1,0,0),(0,1,0),(0,0,0.75)\}$ (b) $\bar{\mathbf{\Omega}} = \{(1,0,0),(0,1,0),(0,0,0.95)\}$

Figure 3: Unconditional choice probabilities for $\lambda_a = 0.1$

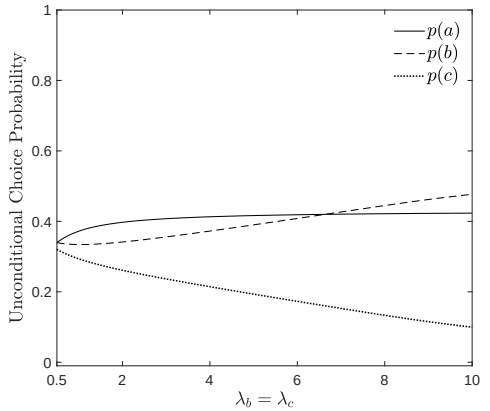
either b or c (i.e., not choosing a) becomes less attractive as λ_b increases, because the consumer would have to acquire additional information at a higher cost to learn which of the two is likely to be the better alternative. Hence, $p(b)+p(c)$ decreases and equivalently $p(a)$ increases. Furthermore, as discussed before, the consumer increasingly prefers b over c. If λ_b increases further, the consumer no longer considers c as an alternative, i.e., $p(c) = 0$. Then, the consumer has to select among the two same-valued alternatives (a and b). Since $\lambda_a < \lambda_b$, it is relatively easier for the consumer to learn when a is *not* the best alternative (as pointed out before). Accordingly, $p(a)$ decreases and $p(b)$ increases with λ_b[5]. Note that if $v_c = 0.75$, the consumer always chooses b over a more frequently. On the other hand, if $v_c = 0.95$, this happens only when λ_b is relatively high as we see in Figure 4(b).

In order to shed further light on the driving forces behind the key observations stated above, it is worth also looking at the *conditional* choice probabilities. These are depicted in Figure 5 for $\bar{\mathbf{\Omega}} = \{(1,0,0),(0,1,0),(0,0,0.$ when $\lambda_a = 0.5$ (these conditional probabilities yield the unconditional probabilities in Figure 4(a)). Observe that when either a or b is indeed the best alternative, the consumer is able to make the correct decision more

[5]This result is valid both when $v_c = 0.75$ and $v_c = 0.95$, even though for the latter case it is not directly observable in Figure 4(b) due to λ_b being limited to the range $[0, 10]$.



(a) $\bar{\mathbf{\Omega}} = \{(1,0,0),(0,1,0),(0,0,0.75)\}$

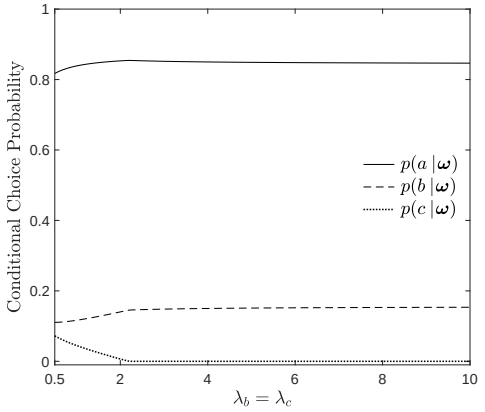


(b) $\bar{\mathbf{\Omega}} = \{(1,0,0),(0,1,0),(0,0,0.95)\}$

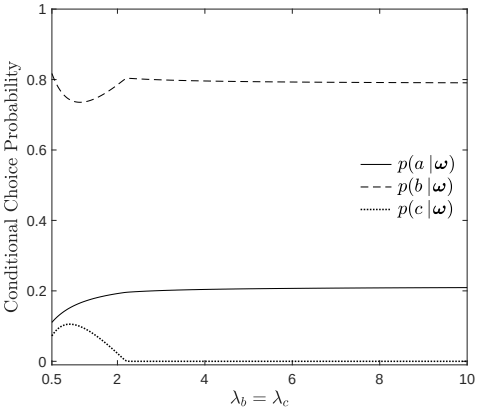
Figure 4: Unconditional choice probabilities for $\lambda_a = 0.5$

frequently. However, when c is the best choice, she makes the correct decision only when the information cost associated with this particular alternative is relatively low. We can glean further insights on information provisioning from these results:

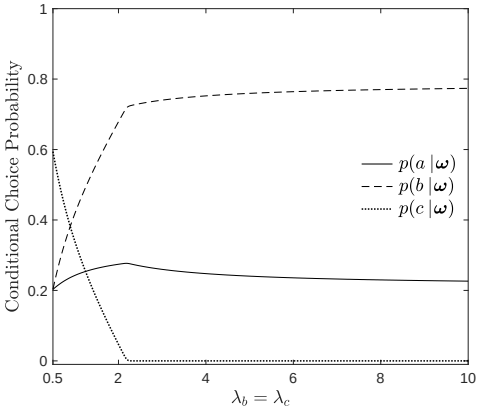
- When an a-priori attractive alternative is also easy to evaluate, it does not matter how difficult it is to gather information about other available alternatives (alternative a in Figure 5(a)).
- In contrast, when an a-priori less attractive alternative is also difficult to evaluate, it will be chosen with low probability even if in reality it is the best alternative (alternative c in Figure 5(c)). Making information easily accessible to the customer is extremely critical for such products.
- When an a-priori attractive alternative is difficult to evaluate, it will be chosen with high probability only if the information costs are very high (alternative b in Figure 5(b)). It is essential that the customers can single out such products. This can be done by making information easily accessible when the product is indeed the best alternative, or by making obscuring the information acquisition process so that the customer relies strictly on her prior beliefs.



(a) $\boldsymbol{\omega} = (1, 0, 0)$



(b) $\boldsymbol{\omega} = (0, 1, 0)$



(c) $\boldsymbol{\omega} = (0, 0, 0.75)$

Figure 5: Conditional choice probabilities for $\lambda_a = 0.5$ and $\bar{\boldsymbol{\Omega}} = \{(1, 0, 0), (0, 1, 0), (0, 0, 0.75)\}$

5.2 Example 2: RED BUS/BLUE BUS

The tripartite race example in the previous section highlights the impact different information costs have on choice behavior, but it does not provide a detailed account of how *correlations* among the alternatives shape this behavior. For this purpose, we turn our attention to the classic RED-BUS/BLUE-BUS problem, and adopt the primary setup in Matějka and McKay (2015). The consumer faces three alternatives – she may take the TRAIN (T), the BLUE BUS (B), or the RED BUS (R). Table 1 gives the four possible states of the world, $(\frac{1}{2},0,0),(\frac{1}{2},1,0),(\frac{1}{2},0,1),(\frac{1}{2},1,1)$, and the prior belief of the consumer about each state, where ρ is the correlation between the values of the two buses (values of 0 and 1 indicate that the particular bus is "slow" or "fast", respectively). The optimal choice probabilities for this example cannot be derived in closed-form, and have to be calculated numerically using either the algorithm we discuss in §6.2 or by convex optimization.

	State 1	*State* 2	*State* 3	*State* 4
TRAIN	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
RED BUS	0	1	0	1
BLUE BUS	0	0	1	1
$g(\boldsymbol{\omega})$	$(1+\rho)/4$	$(1-\rho)/4$	$(1-\rho)/4$	$(1+\rho)/4$

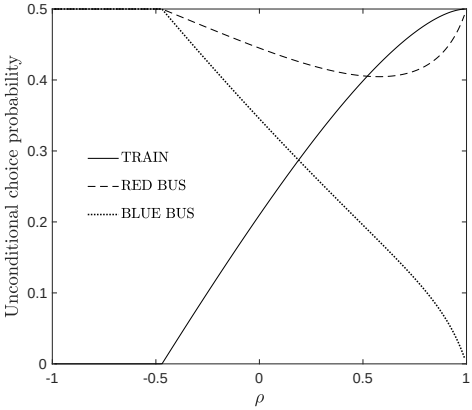
Table 1: Possible states and prior beliefs

Note that the speed of the TRAIN is deterministic, and the expected value of each of the three alternatives equals $\frac{1}{2}$. If the consumer were to choose an alternative *without* processing any information, she would be indifferent between the three alternatives, i.e., $p(T)=p(R)=p(B)=\frac{1}{3}$. On the other hand, if the consumer could process information freely, she would always choose the fastest alternative, resulting in unconditional choice probabilities $p(T)=(1+\rho)/4$ and $p(R)=p(B)=(3-\rho)/8$. Then, the buses are symmetric and they are chosen with equal probability, and this probability decreases as ρ increases (the consumer learns more often that both buses are slow).

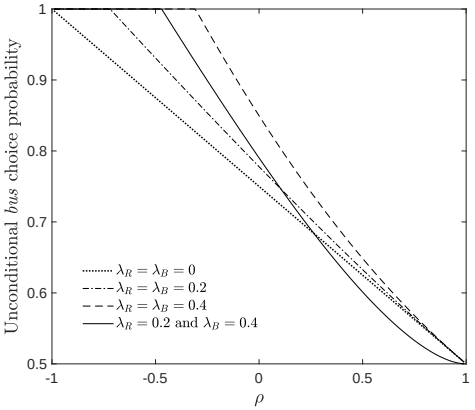
Suppose that acquiring and processing information about the RED BUS

is less expensive than about the BLUE BUS, i.e., $\lambda_R \leq \lambda_B$. Figure 6(a) shows the unconditional choice probabilities when $\lambda_R = 0.2$ and $\lambda_B = 0.4$ for varying correlation in the prior belief. We observe that the TRAIN is never selected if the consumer has sufficiently strong belief that the two bus speeds are negatively correlated (i.e., if she sufficiently believes that *one* of the buses must be fast). Given that information processing is costly, due to her beliefs that the TRAIN is unlikely to be the best alternative, she instead allocates all her time and attention to understand which bus is faster (in fact, she just learns about the RED BUS and if it isn't promising, opts for the BLUE BUS, such that each bus is chosen with 50% chance). Nevertheless, as the consumer's prior belief that the two buses are similar gets stronger with ρ, she also starts selecting the TRAIN. Interestingly, in this range she builds a stronger preference for the "cheap" RED BUS over the BLUE BUS. This is because she acquires more information about the RED BUS and has more confidence about its speed compared to the BLUE BUS. In particular, as ρ approaches 1, the consumer believes that the buses have identical speed. Consequently, whenever she decides to take a bus, she takes the RED BUS, on which she has more information. This signifies the importance of information provision for a seller in forming its product choice set. When the alternatives are very similar in the eyes of the consumer, even a slight improvement in the provision of information for one product can significantly shift demand towards it. This is particularly stark considering that when $\rho = 1$, the consumer treats duplicate alternatives jointly as one.

From the above discussion, it is clear that correlations in the alternatives enable the consumer to draw inferences, which in turn significantly influence her attention allocation and final choice. To elaborate on this fact, we depict in Figure 6(b), the conditional probability of taking "a bus" (red or blue). On one hand, when information is freely available, the consumer always makes the correct choice and hence takes the bus $(3-\rho)/4$ fraction of time. On the other hand, when the consumer faces uniform information costs ($\lambda_R = \lambda_B = 0.4$), she takes the bus too often (for the same reason discussed above). One could conjecture that reducing the information cost of even one alternative would increase the amount of information processed, so it should bring this probability close to the perfect information case ($\lambda_R = \lambda_B = 0$). From Figure 6(b), it is evident that



(a) Unconditional choice probabilities (when information costs are asymmetric $\lambda_R = 0.2$ and $\lambda_B = 0.4$)



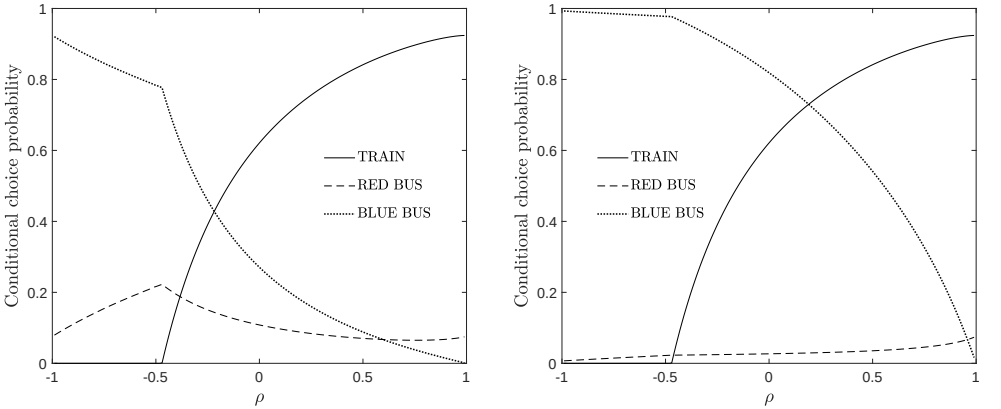
(b) Unconditional *bus* choice probabilities for various information cost constellations

Figure 6: Unconditional choice probabilities for the RED BUS/BLUE BUS problem

this is only partially correct. When $\lambda_R = 0.2$ and $\lambda_B = 0.4$, for negative correlation levels, the consumer more correctly identifies the TRAIN as the fastest alternative. However, at positive correlation levels a new decision bias is created. Since the consumer knows more about the bus with the lower cost, she starts drawing strong (and wrong) inferences about the other bus, and this time she ends up taking the TRAIN too often. Reducing the information cost of the BLUE BUS so that information costs are uniform again ($\lambda_R = \lambda_B = 0.2$) eliminates this decision bias and brings the conditional choice probability of choosing either bus closer to the perfect information case for all levels of ρ. This highlights the benefits that a seller can potentially earn from presenting information about different choice alternatives in a rather similar and uniform manner.

In order to deepen our understanding of how asymmetry in information costs, coupled with correlations, lead to more wrong/correct choices, we next focus on the conditional choice probabilities, given in Figure 7. As seen in Figure 7(a), even when the TRAIN is the best alternative, it is not selected by the consumer for sufficiently negative ρ, $\rho \leq -0.47$ (as previously explained). Moreover, in this range, her conditional belief for the RED BUS being slow but the BLUE BUS being fast is decreasing in ρ. Therefore, the

conditional choice probability of RED BUS (resp. BLUE BUS) increases (resp. decreases) in ρ. On the other hand, for $\rho > -0.47$, the TRAIN is also chosen and the consumer increasingly prefers the TRAIN and avoids the buses as ρ increases. Further she also learns that both buses are more likely to be slow mainly by processing direct information about the (cheaper) RED BUS. Hence, as long as $\rho \leq 0.60$, the consumer takes the RED BUS less often than the BLUE BUS. Interestingly, for high levels of ρ, when she erroneously learns that the RED BUS is fast, she also infers that the BLUE BUS must also be fast but since she is more informed about the RED BUS, she takes it. In this case, the BLUE BUS is rarely chosen.



(a) In state $(\frac{1}{2}, 0, 0)$, when TRAIN is best

(b) In state $(\frac{1}{2}, 0, 1)$, when BLUE BUS is best

Figure 7: Conditional choice probabilities for the RED BUS/BLUE BUS problem ($\lambda_R = 0.2$ and $\lambda_B = 0.4$)

Figure 7(b) shows that the consumer most often makes the right choice even when information processing is expensive for the fast bus (RED BUS) and cheap for the slow bus (BLUE BUS), provided that she has negatively correlated beliefs. When she increasingly believes the buses are similar (ρ increases), however, the likelihood of taking the BLUE BUS decreases sharply. This is because the consumer has more information about the slow RED BUS, and since the buses are very similar according to her beliefs, she draws the inference that the BLUE BUS must also be slow. She instead increasingly chooses the TRAIN (and makes the wrong decision). In the extreme case, $\rho \geq 0.95$ the likelihood of her taking the correct BLUE BUS is even less than the RED BUS.

6. Solving the Choice Model

The necessary conditions (10) presented in Theorem 1 are also sufficient if in optimality all alternatives are selected with positive probability. Clearly, this is not always the case; there may be unchosen alternatives in a given setting. In this section, we first give both necessary *and* sufficient conditions of optimality, and subsequently propose an algorithm to solve for the choice probabilities.

6.1 A Characterization of Optimal Choice Probabilities

The necessary conditions in (10) are silent for unchosen alternatives. This calls for conditions of optimality not restricted to interior points. These conditions constitute the basis of our algorithm. In due course, we first substitute (10) into the consumer's choice problem in Proposition 1 to obtain a more simplified formulation.

Lemma 1 (Alternative Formulation) *The conditional choice probabilities* $(p(i \mid \boldsymbol{\omega}))_{i\in A,\boldsymbol{\omega}\in\boldsymbol{\Omega}}$, *solve the problem in Proposition 1 if and only if they are calculated by (10) from a collection of partial conditional choice probabilities* $\mathbf{p} = \{p(i \mid \boldsymbol{\omega}_{1..n-1})\}_{i\in A,\boldsymbol{\omega}_{1..n-1}\in\boldsymbol{\Omega}_{1..n-1}}$, *that solve*

$$\begin{aligned}
\max_{\mathbf{p}} \quad & W(\mathbf{p}) = \lambda_n \sum_{\boldsymbol{\omega}} g(\boldsymbol{\omega}) \log\left(\sum_{i\in A} e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}} p(i)^{\frac{\lambda_1}{\lambda_n}} \prod_{k=1}^{n-1} p(i \mid \boldsymbol{\omega}_{1..k})^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}\right) \\
\text{s. t.} \quad & p(i \mid \boldsymbol{\omega}_{1..n-1}) \geq 0 \text{ for all } i \in A \text{ and all } \boldsymbol{\omega}_{1..n-1} \in \boldsymbol{\Omega}_{1..n-1} \\
& \sum_{i\in A} p(i \mid \boldsymbol{\omega}_{1..n-1}) = 1 \text{ for all } \boldsymbol{\omega}_{1..n-1} \in \boldsymbol{\Omega}_{1..n-1}
\end{aligned}$$

Remember that the partial conditional probabilities yield the unconditional choice probabilities via $p(i) = \sum_{\boldsymbol{\omega}_{1..n-1}} p(i \mid \boldsymbol{\omega}_{1..n-1})\, g(\boldsymbol{\omega}_{1..n-1})$. For uniform information costs, the alternative formulation boils down to finding unconditional choice probabilities $p(i)$ that maximize

$$\lambda \sum_{\boldsymbol{\omega}} g(\boldsymbol{\omega}) \log \sum_{i\in A} e^{u(i,\boldsymbol{\omega})/\lambda} p(i),$$

which is tantamount to maximizing a log-sum expectation with applications in other fields as well. In particular, it is equivalent to finding a so-

called log-optimal portfolio in finance (cf. Cover 1984). The alternative formulation is a concave problem on a compact set, albeit not differentiable on the boundaries. Thus, the optimality conditions for interior solutions are more expressive than those on the boundaries:

Theorem 2 (Sufficient Conditions) *The partial conditional choice probabilities* $\mathbf{p}$ *solve the problem in Lemma 1 only if*

$$p\left(i \mid \boldsymbol{\omega}_{1\cdot\cdot n-1}\right) = \sum_{\omega_n \in \Omega_n} \frac{g\left(\omega_n \mid \boldsymbol{\omega}_{1\cdot\cdot n-1}\right) e^{\frac{u\left(i,\boldsymbol{\omega}_{1\cdot\cdot n-1}\omega_n\right)}{\lambda_n}} p\left(i\right)^{\frac{\lambda_1}{\lambda_n}} \prod_{k=1}^{n-1} p\left(i \mid \boldsymbol{\omega}_{1\cdot\cdot k}\right)^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}}{\sum_{j\in A} e^{\frac{u\left(i,\boldsymbol{\omega}_{1\cdot\cdot n-1}\omega_n\right)}{\lambda_n}} p\left(j\right)^{\frac{\lambda_1}{\lambda_n}} \prod_{k=1}^{n-1} p\left(j \mid \boldsymbol{\omega}_{1\cdot\cdot k}\right)^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}} \tag{15}$$

for all $i \in A$ *and all* $\boldsymbol{\omega} \in \boldsymbol{\Omega}_{1\cdot\cdot n-1}$ *such that* $p\left(i \mid \boldsymbol{\omega}_{1\cdot\cdot n-1}\right) > 0$.

If $\mathbf{p}$ *is interior, i.e., if* $p\left(i \mid \boldsymbol{\omega}_{1\cdot\cdot n-1}\right) > 0$ *for all* $i \in A$ *and all* $\boldsymbol{\omega} \in \boldsymbol{\Omega}_{1\cdot\cdot n-1}$, *then* (15) *is sufficient.*

In general, the following are necessary and sufficient conditions:

$$\sum_{\boldsymbol{\omega}\in\boldsymbol{\Omega}} g\left(\boldsymbol{\omega}\right) \frac{e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}}}{\sum_{j\in A} e^{\frac{u\left(i,\boldsymbol{\omega}\right)}{\lambda_n}} p\left(j\right)^{\frac{\lambda_1}{\lambda_n}} \prod_{k=1}^{n-1} p\left(j \mid \boldsymbol{\omega}_{1\cdot\cdot k}\right)^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}} \leq 1 \quad \forall i \in A. \tag{16}$$

Note that (15) is implied by the necessary condition (10) in Theorem 1 (refer to Appendix B for details). Sufficiency of (15) is established by showing that (15) guarantees satisfaction of the KKT conditions. We also remark that Theorem 2 generalizes the necessary and sufficient conditions for the uniform information cost case specified in Caplin et al. (2016a, Proposition 1) and in Cover and Thomas (2006, Theorem 16.2.1) in the context of portfolio optimization.

6.2 Algorithm to Find Optimal Choice Probabilities

Lemma 1 offers a significant simplification over the formulation in Proposition 1. Nevertheless, solving for the optimal partial probabilities $\mathbf{p}$ can still be quite challenging when there are many alternatives and possible realizations of the values. To mitigate this problem, we propose an iterative algorithm inspired by Cover (1984) that exploits the optimality conditions in Theorem 2:

Algorithm (optimal partial conditional choice probabilities)

STEP 1: Start with a vector $\mathbf{p}^0 \in \mathbb{R}_{++}^{|A\times\boldsymbol{\Omega}_{1..n-1}|}$.

STEP 2: While $W\left(\mathbf{p}^t\right) > W\left(\mathbf{p}^{t-1}\right)$, calculate $\mathbf{p}^{t+1} = \left(p^{t+1}\left(i \mid \boldsymbol{\omega}_{1..n-1}\right)\right)_{i\in A,\boldsymbol{\omega}_{1..n-1}\in\boldsymbol{\Omega}_{1..n-1}}$ as

$$p^{t+1}\left(i \mid \boldsymbol{\omega}_{1..n-1}\right) = \sum_{\omega_n\in\Omega_n} \frac{g\left(\omega_n \mid \boldsymbol{\omega}_{1..n-1}\right) e^{\frac{u(i,\boldsymbol{\omega}_{1..n-1},\omega_n)}{\lambda_n}} p\left(i\right)^{\frac{\lambda_1}{\lambda_n}} \prod_{k=1}^{n-1} p^t\left(i \mid \boldsymbol{\omega}_{1..k}\right)^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}}{\sum_{j\in A} e^{\frac{u(j,\boldsymbol{\omega}_{1..n-1},\omega_n)}{\lambda_n}} p\left(j\right)^{\frac{\lambda_1}{\lambda_n}} \prod_{k=1}^{n-1} p^t\left(j \mid \boldsymbol{\omega}_{1..k}\right)^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}}. \tag{17}$$

STEP 3: Check if $\mathbf{p}^{t+1}$ satisfies conditions (16). If $\mathbf{p}^{t+1}$ satisfies conditions (16), abort with $\mathbf{p}^{t+1}$ as the solution. Otherwise, slightly increase the probability of the alternative for which the violation of (16) is strongest; i.e., denote this alternative by $\hat{\imath}$,

$$\hat{\imath} \in \arg\max_i \sum_{\boldsymbol{\omega}} g\left(\boldsymbol{\omega}\right) \frac{e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}}}{\sum_{j\in A} e^{\frac{u(j,\boldsymbol{\omega})}{\lambda_n}} p\left(i\right)^{\frac{\lambda_1}{\lambda_n}} \prod_{k=0}^{n-1} p^t\left(j \mid \boldsymbol{\omega}_{1..k}\right)^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}},$$

and let $\hat{\mathbf{p}}$ be given by $\hat{p}\left(\hat{\imath} \mid \boldsymbol{\omega}_{1..n-1}\right) = 1$ for all $\boldsymbol{\omega}_{1..n-1}$. Set $\mathbf{p}^{t+1} = (1-\varepsilon)\mathbf{p}^{t+1} + \varepsilon\hat{\mathbf{p}}$, where ε is sufficiently small, e.g. as given in (30) in Appendix B, and go to STEP 2.

Note that the updating process in (17) produces feasible vectors that satisfy $\sum_i p^{t+1}(i \mid \boldsymbol{\omega}_{1..n-1}) = 1$. Moreover, we establish in the next proposition that it improves the consumer's objective. Furthermore, when the objective converges in STEP 2, the optimality condition (15) is satisfied. If the optimality condition (16) is violated, then we perturb $\mathbf{p}^{t+1}$ in STEP 3 in a way that improves the objective, and repeat the updating process.

Proposition 2 *The updating described in (17) weakly improves the objective, i.e.,* $W\left(\mathbf{p}^{t+1}\right) \geq W\left(\mathbf{p}^{t}\right)$ *. In particular, we have*

$$\begin{aligned} & W\left(\mathbf{p}^{t+1}\right)-W\left(\mathbf{p}^{t}\right) \\ \geq \quad & \lambda_{1} D_{KL}\left(p^{t+1}(\cdot) \| p^{t}(\cdot)\right) \\ & +\sum_{k=1}^{n-1}\left(\lambda_{k+1}-\lambda_{k}\right) \sum_{\boldsymbol{\omega}_{1 \cdot \cdot k} \in \boldsymbol{\Omega}_{1 \cdot \cdot k}} g\left(\boldsymbol{\omega}_{1 \cdot \cdot k}\right) D_{KL}\left(p^{t+1}\left(\cdot \mid \boldsymbol{\omega}_{1 \cdot \cdot k}\right) \| p^{t}\left(\cdot \mid \boldsymbol{\omega}_{1 \cdot \cdot k}\right)\right), \end{aligned}$$

where D_{KL} *denotes the Kullback-Leibler divergence defined as* $D_{KL}(p\|q) = \sum_x p(x) \frac{\log p(x)}{\log q(x)}$.

Finally, we establish that the algorithm finds optimal partial conditional choice probabilities.

Theorem 3 *Algorithm 6.2 finds an optimal solution to the optimization problem of the consumer.*

7. On Empirical Estimation and Validation

We now explore the connection between our model and data. To this end, we first present a model that motivates the usage of market share data in order to infer the utility of rationally inattentive consumers. We then sketch parallels with a second stream of literature that aims at testing the model from an experimental researcher's perspective. In order to avoid confusion with the notion of "state" in empirical studies, which often alludes to a consumer's purchase history, we refer to the state of the world as "type" hereon.

7.1 Inference from Market Share Data

Equating observed market shares with the choice probabilities induced by a rationally inattentive consumer rests on the assumption that consumers have correct prior beliefs, i.e., they know the true distribution of consumer types, which determines the potential utility they might gain from products. This assumption is motivated recently by Caplin et al. (2016b), who develop a model where rationally inattentive consumers freely observe

past product market shares, and then acquire costly additional information about the utility they derive from products. In this model there are two sources of uncertainty: (i) each consumer is unsure about her type, which reflects her preferences and which can be learned in a rationally inattentive fashion; (ii) initially, consumers are not sure about the distribution of these types in the population. Over time, the possible distributions of consumer types are refined until a steady-state is reached.

The fundamental result is that steady-state market shares of chosen products are equal to those that would follow from correct knowledge of the true type distribution in the population, and that the market shares subsequently reflect the choice probabilities induced by the GMNL formula. This central connection paves the way for empirical work when sufficiently rich type dependent choice data exists. In what follows, we establish a generalization and demonstrate that detailed market share data can be used when information costs are non-uniform among the type characteristics (refer to Appendix D for details).

Consider a dynamic variation of our problem indexed by time t. Consumers neither know their actual type $\boldsymbol{\omega} \in \Omega$, nor the true type distribution $g^\star \in \Delta(\Omega)$, where $\Delta(\Omega)$ denotes the set of distributions over types. A consumer's type determines the utility $u(i, \boldsymbol{\omega})$ she gains from choosing one of the available alternatives $i \in A$. In each period t, each consumer observes all past realized partial-type dependent market shares $\{M_{t'}(i \mid \boldsymbol{\omega}_{1..n-1})$ and then enters the decision making process with the belief $\mu_t \in \Delta(\Omega)$, the average of all type distributions deemed possible at time t. This is followed by an optimal information acquisition and choice making in the spirit of rational inattention with non-uniform information costs, resulting in conditional choice probabilities $p_t(i \mid \boldsymbol{\omega})$. The realized partial-type dependent market shares $M_t(i \mid \boldsymbol{\omega}_{1..n-1})$ are determined by type dependent choice probabilities $p_t(i \mid \boldsymbol{\omega})$ and the true distribution $g^\star$, i.e.,

$$M_t(i \mid \boldsymbol{\omega}_{1..n-1}) \quad = \quad \sum_{\omega_n} g^\star(\omega_n \mid \boldsymbol{\omega}_{1..n-1})\, p_t(i \mid \boldsymbol{\omega})\,. \tag{18}$$

Consequently, in each period, consumers eliminate a belief $\breve{\mu}$ from the set of possible type distributions if the observed partial-type dependent market shares cannot be generated from it, i.e., if (18) is not satisfied when $g^\star$ is replaced by $\breve{\mu}$.

Using an analogous approach to Caplin et al. (2016b), we can show the following:

1. Consumer beliefs refine and converge to a steady-state after a finite number of periods.

2. In steady-state, overall market shares satisfy $M(i) = p(i)$ and partial-type dependent market shares satisfy $M\left(i \mid \boldsymbol{\omega}_{1..k}\right) = p\left(i \mid \boldsymbol{\omega}_{1..k}\right)$ for $k = 1, .., n-1$.

3. Assuming further that the consumers learn independently, i.e., $M\left(i \mid \boldsymbol{\omega}\right) = p\left(i \mid \boldsymbol{\omega}\right)$, we obtain

$$M\left(i \mid \boldsymbol{\omega}\right) = \frac{e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}} M\left(i\right)^{\frac{\lambda_1}{\lambda_n}} \prod_{k=1}^{n-1} M(i \mid \boldsymbol{\omega}_{1..k})^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}}{\sum_{j\in A} e^{\frac{u(j,\boldsymbol{\omega})}{\lambda_n}} M\left(j\right)^{\frac{\lambda_1}{\lambda_n}} \prod_{k=1}^{n-1} M\left(j \mid \boldsymbol{\omega}_{1..k}\right)^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}}. \tag{19}$$

As type dependent choice and market shares follow the (generalized) logit model, known estimation techniques from extant literature can be used. To see this more explicitly, take the seminal model in McFadden (1974), also discussed in Caplin et al. (2016b). In this model, consumer l of type $\boldsymbol{\omega}$ derives utility

$$U_{li}^{\boldsymbol{\omega}} = u(i, \boldsymbol{\omega}) + \sigma\epsilon_{li}$$

from choosing alternative i. Here, $u(i, \boldsymbol{\omega})$ is the intrinsic utility, ϵ_{li} is a shock following an extreme value distributed, and $\sigma > 0$ is a scalar. The type dependent market shares are given by the logit formula as

$$M(i \mid \boldsymbol{\omega}) = \frac{e^{\sigma u(i,\boldsymbol{\omega})}}{\sum_{j\in A} e^{\sigma u(j,\boldsymbol{\omega})}}.$$

In our model, type dependent market shares are of similar form, i.e.,

$$M(i \mid \boldsymbol{\omega}) = \frac{e^{\frac{1}{\lambda_n}u(i,\boldsymbol{\omega})+\alpha(i,\boldsymbol{\omega}_{1..n-1})}}{\sum_{j\in A} e^{\frac{1}{\lambda_n}u(j,\boldsymbol{\omega})+\alpha(j,\boldsymbol{\omega}_{1..n-1})}}, \tag{20}$$

where $\alpha(i, \boldsymbol{\omega}_{1..n-1}) = \frac{\lambda_1}{\lambda_n} \log M\left(i\right) + \sum_{k=1}^{n-1} \frac{\lambda_{k+1}-\lambda_k}{\lambda_n} \log M\left(i \mid \boldsymbol{\omega}_{1..k}\right)$.

Based on the above, we can state the following equivalency: Cross-sectional consumer choices in our model are *observationally equivalent*

to choices in a random-utility model with

$$U_{li}^{\boldsymbol{\omega}} = (u(i, \boldsymbol{\omega}) + \lambda_n \alpha(i, \boldsymbol{\omega}_{1..n-1})) + \tfrac{1}{\lambda_n}\epsilon_{li},$$

where ϵ_{li} are extreme-value distributed.

In the uniform case with $\lambda_1 = \ldots = \lambda_n = \lambda$, the term $\lambda_n \alpha(i, \boldsymbol{\omega}_{1..n-1})$ simply becomes $\lambda \log M(i)$. Then, as argued by Caplin et al. (2016b), if one uses the standard logit model to infer utility, one would actually estimate $u(i, \boldsymbol{\omega}) + \lambda \log M(i)$ instead of $u(i, \boldsymbol{\omega})$. They suggest to determine $1/\lambda$ in the same way as σ, and then subtract $\lambda \log M(i)$ to obtain estimates of the intrinsic utility $u(i, \boldsymbol{\omega})$.

Note that similar to the scalar σ, which "normalizes" the explained utility $u(i, \boldsymbol{\omega})$ relative to the unobserved shock $\sigma\epsilon_{li}$, the scalar λ_n "normalizes" the intrinsic utility relative to the unit cost of information in a rational inattention model. Indeed, dividing utilities $u(i, \boldsymbol{\omega})$ and unit costs of information λ_k with a positive constant in the objective (6) does not alter the behaviour of the consumer, which indicates that utilities and cost of information can only be determined up to a scalar. Thus, setting $\lambda_n = 1$ just means that the utilities and the other unit costs of information are measured in terms of the highest unit cost of information.

Estimating the intrinsic utility is slightly more involved in the non-uniform case than in the uniform case, since $\lambda_n \alpha(i, \boldsymbol{\omega}_{1..n-1})$ contains the additional terms $\sum_{k=1}^{n} \lambda_k (\log M(i \mid \boldsymbol{\omega}_{1..k-1}) - \log M(i \mid \boldsymbol{\omega}_{1..k}))$, which are linear in the observed partial-type dependent log market shares. Therefore, one also needs to obtain estimates of λ_k for $k = 1, ..., n-1$. Together with λ_n, all information cost parameters would then be specified, and one could account for the information cost to obtain the intrinsic utility $u(i, \boldsymbol{\omega})$.

It is important to note here the nature of data needed to conduct this estimation. Simply observing unconditional choices is not enough (for both uniform and non-uniform information cost cases). What is needed instead is type dependent choice data, which captures both the consumer's choice and the actual realization of the consumer's type. An observation in this data set thus contains the realized state of the world which the consumer tried to identify and the choice she made in that state. Clearly, one assumption made here is that we (as the econometricians) are not subject

to the same information constraints as the consumer, e.g. the consumer may not know whether shipping cost are included in the price or not but we know. We refer the reader to Caplin et al. (2017) for more details, as well as examples of such type dependent choice data.

It is possible to go a step further than just estimating the intrinsic utility $u(i, \boldsymbol{\omega})$. We can parameterize the utility $u(i, \boldsymbol{\omega})$ and then estimate these parameters along with the information cost parameters λ_k's. As an example, reconsider the RED BUS/BLUE BUS/TRAIN-problem in §5.2. Suppose that the correlation between the buses, ρ, is known, and with this, the distribution of types (the true prior) is known by the consumer. Here, a consumer's type specifies which transportation mode is fastest for her. Further, assume that we have type dependent choice data, i.e., we know for each consumer l, her choice i, and we know the actual fastest transportation mode that she tried to figure out and to select. Denote the observed type of consumer l by $\boldsymbol{\omega}_l = (\omega_{lT}, \omega_{lR}, \omega_{lB}) \in \{1/2\} \times \{0, 1\} \times \{0, 1\}$. For example, $\boldsymbol{\omega}_l = (1/2, 0, 1)$ means that the BLUE BUS was the fastest when consumer l made her pick. The utility of a bus is given by $u(i, \boldsymbol{\omega}_l) = \beta_{RB}\omega_{li}$ for $i \in \{R, B\}$, while the train always yields $u(T, \boldsymbol{\omega}_l) = \beta_T/2$. Here, β_{RB} and β_T are taste parameters that shall be estimated along with the unit costs of information λ_R and λ_B (in §5.2, we use $\beta_{RB} = \beta_T = 1$). As discussed above, we can set $\lambda_B = 1$, i.e., we are interested in the parameters λ_R, β_{RB}, and β_T in terms of the higher unit cost of information. Define $\mathbf{x}_{li}$ as a vector of observed variables that relate to alternative i as faced by consumer l, as follows:

$$\begin{aligned}
\mathbf{x}_{lT} &= (\omega_{lT}, \quad 0, \log M(T) - \log M(T \mid 1/2, \omega_{lR})), \\
\mathbf{x}_{lR} &= (\ 0, \omega_{lR}, \log M(R) - \log M(R \mid 1/2, \omega_{lR})), \\
\mathbf{x}_{lB} &= (\ 0, \omega_{lB}, \log M(B) - \log M(B \mid 1/2, \omega_{lR})).
\end{aligned}$$

With the coefficients denoted by $\boldsymbol{\beta} = (\beta_T, \beta_{RB}, \lambda_R)$, it is easy to verify that written this way, we get

$$u(i, \boldsymbol{\omega}) + \lambda_n \alpha(i, \boldsymbol{\omega}_{1..n-1}) = \boldsymbol{\beta}' \mathbf{x}_{li} + \log M(i \mid \omega_{lR}),$$

i.e., we have to estimate a logit model with linear utility up to a constant in order to determine both the utility generating coefficients and the unit cost of information.

In summary, when sufficiently rich type dependent choice data exists, the rationally inattentive choice with non-uniform information costs can be estimated using logit models of inference. This would also reveal the potential biases that exist in standard estimation models that overlook the possibility of information frictions (or limited time and attention capacity of consumers), and show how these are shaped by the asymmetry in obtaining and processing information about different alternatives, attributes, or states of the world in general.

7.2 Experimental Validation and Estimation

In other recent work, Caplin and Dean (2013, 2015) describe a method to identify whether a consumer is rationally inattentive or not from type dependent choice data. The method is tested using data collected from an experiment where the subject observes 100 balls appearing on a computer screen, some of which are red and the remaining are blue, e.g. 51 red balls and 49 blue balls or the other way around. Subjects are told that both states are equally likely. The subject has to determine whether a majority of the balls are red or blue, receiving $\$10$ when her answer is correct, and nothing when she is incorrect. Repetition of the experiment yields type dependent choice frequencies that can be interpreted as conditional choice probabilities. These probabilities imply posterior beliefs that have to satisfy the condition (14), which rearranged for the unit cost of information λ becomes

$$\lambda = u(\$10)\left(\log \frac{p(51 \text{ balls are red} \mid \text{red is guessed to be the majority})}{p(51 \text{ balls are red} \mid \text{blue is guessed to be the majority})}\right)^{-1},$$

where we set $u(\$0)=0$. Note that a variation of prizes (e.g. $\$2, \$10, \$30$) now has strict implications on the posterior probabilities since λ remains fixed.

A variation of this experiment can be conducted in the case of nonuniform information costs:

> 100 circles with continuous, dashed, or dashed-dotted outline appear on the screen.
>
> Subject have to figure out which outline appears most often and receive $\$10$ when guessing right.

We suppose that it is easier to distinguish continuous circles (see Figure 8). Hence, verifying whether these prevail is easier than verifying whether the other types prevail. In this sense, the setup is equivalent to the Tripartite Race example we studied earlier in §3.1 and §5.1.

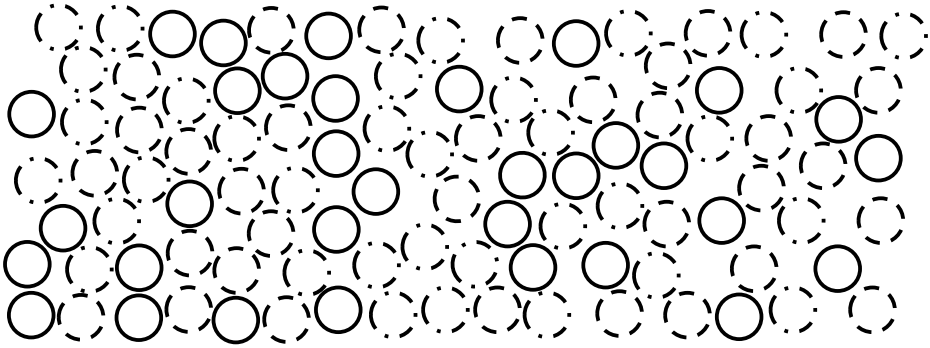


Figure 8: Experiment with $100 = 33 + 33 + 34$ circles. Subjects guess which type appears most (here: dash-dotted)

Repetition of the experiment yields choice frequencies that can be interpreted as conditional choice probabilities, and which in turn imply posterior beliefs. Using these empirical posteriors, we can check for *qualitative* predictions of the model, which relate to the fact that subjects should prioritize cheap information: Subjects should first count the continuous circles (which is easiest), and guess continuous when sufficiently confident. Otherwise they proceed to decide between dashed and dashed-dotted. This procedure – as does the model, see (46) in Appendix C – implies

$$p(34 \text{ circles have dashed-dotted outline} \mid \text{continuous is guessed to be the majority}) = p(34 \text{ circles have dotted outline} \mid \text{continuous is guessed to be the majority}),$$

since (erroneously) guessing continuous happens without inspecting dash-dotted or dotted circles and therefore is not affected by whether there are more dashed or dashed-dotted circles. Moreover, having erroneously rejected continuous should not impact the chance of choosing dashed or dashed-dotted. This implies that the chance attached to having erroneously rejected continuous should be the same irrespectively of whether dashed or dashed-dotted is chosen ((41) in Appendix C):

$$p(34 \text{ circles have continuous outline} \,|\, \text{dashed-dotted is guessed to be the majority})$$
$$= p(34 \text{ circles have continuous outline} \,|\, \text{dotted is guessed to be the majority}).$$

Note that these equations constitute additional conditions that go beyond the conditions already imposed by the model with uniform information costs (e.g. NIAS and NIAC conditions in Caplin and Dean (2013)). Restricting oneself to subjects that conform with these conditions, we can further elicit the information costs (see (40), (42) in Appendix C) from

$$\frac{u(\$10)}{\lambda_1} = \log \frac{p(34 \text{ circles have continuous outline} | \text{continuous is guessed to be the majority})}{p(34 \text{ circles have continuous outline} | \text{dashed is guessed to be the majority})},$$

$$\frac{u(\$10)}{\lambda_2} = \log \frac{p(34 \text{ circles have dashed outline} | \text{dashed is guessed to be the majority})}{p(34 \text{ circles have dashed-dotted outline} | \text{dashed-dotted is guessed to be the majority})},$$

which then has testable *quantitative* implications on subjects' reactions to payoff changes.

In a similar vein, Oliveira et al. (2016) introduce a method to elicit preferences and to estimate the information cost function of a rationally inattentive consumer. They consider an environment where the consumer is asked to make two choices: first, she has to choose a choice set, i.e., form a "choice menu". Then, she selects in a rationally inattentive fashion from the alternatives belonging to her choice menu. Their insights are generated from analyzing the consumer's selection of menus – a rationally inattentive consumer anticipates the cost of information that she will process later on in order to pick an alternative from within the previously selected menu in a particular way. Their setup works for what they call canonical information cost functions, to which our generalized Shannon information cost function belongs. Thus, their insights can be applied to our cost function in order to elicit the cost of information from the menu-selection behaviour of consumers.

8. Concluding Remarks

In this paper, we develop a consumer choice model where rationally inattentive customers choose among a given set of alternatives. Our novel contribution is the incorporation of information costs that differ among

the alternatives. This captures the notion that it might be inherently (or by seller design) more difficult to learn about some alternatives than about others. We develop an information cost function that distinguishes between direct and implied information obtained by the consumer by studying each alternative, and that prioritizes the use of cheaper sources in the acquisition and processing of information. This conditional mutual information based function generalizes the Shannon cost functions commonly utilized in the rational inattention literature. We analyze the choice problem of the consumer and show that the optimal choice behavior can be characterized analytically. When the unit cost of acquiring information is the same across all alternatives, the choice behavior reduces to the GMNL choice studied by Matějka and McKay (2015). According to the optimal choice behavior, the conditional choice probability associated with each alternative depends on realized values of the alternatives, their information costs, and prior beliefs. Although the exact relationship is non-trivial, essentially the relative "attractiveness" of each alternative is adjusted by the fact that the consumer learns more about the alternatives with lower information costs. Accordingly, if the information obtained by these alternatives imply a higher (or lower) likelihood of selecting a particular alternative, it is weighed into the attractiveness of that alternative appropriately.

We study a number of limiting scenarios and typical examples to illustrate the optimal choice behavior, and show that non-uniform information costs can induce complex consumer behaviour. Accordingly, the consequences for the seller depend on the particular situation. Although an asymmetric reduction of information costs yields a better-informed consumer overall, the consumer's beliefs can become strongly biased by focusing on a particular information channel. Perhaps surprisingly, there are situations where the market share of a product may increase when it becomes harder to learn about it. Our characterization enables us to verify if such changes (perhaps due to alterations in the information provision strategy of the seller) would lead to more correct (or incorrect) choices for the consumer, and can be used to evaluate the benefits (or losses) to the seller. We identify two scenarios where creating a difference in information costs leads to a striking change in relative market share. One is the case when two products are very similar in nature. In such contexts, the

consumer mainly relies on the information about the product with low information cost and forms her belief about the product with high information cost based on implied information. As both products are similar in quality, she strongly prefers the product of which she is more confident about. Another is the case where there is hard-to-evaluate product that is also believed to be less attractive. Slight improvement in the provision of information for such products can significantly impact upon demand. In addition to the above, our framework provides an explanation for why adding an unattractive – even a dominated, never selected – product may increase the market share of another product (known as failure of regularity). This may occur if the newly introduced inferior product facilitates an easier access to information about existing products.

Most of the decisions that consumers have to make require time, attention and cognitive effort, all of which are limited resources. Our model offers a micro-founded description of how such choices are made when the consumers trade off the value of better information against the costs, in a context where information can be acquired about the alternatives with different rates of time-and-attention-efficiency. As noted earlier, this choice behavior and the resulting description of demand is a crucial input to many practical operational problems. As a concluding example, consider an online firm like airbnb.com or booking.com. When consumers search for a particular accommodation, there are usually a large number of potential hits. It is well-known (e.g., De los Santos et al. 2012) that people do not have the time and attention span to go through all pages. What is often displayed on the first page (or even a subset of this page) is where most attention is directed, while choices listed on the following pages require additional effort to evaluate. From the seller's revenue management perspective, it is extremely important to decide on the order at which alternatives are displayed. Determining this requires a consumer demand model that describes how choices are going to be made when the cost of information differs among the alternatives and the consumer is rational and efficient when evaluating her alternatives. Going a step of further, such sellers face the trade-off between displaying more alternatives on the same page with less related information (high information costs) versus less alternatives but with more available information (low information costs). Our choice model has the potential to serve as the building block of

such product assortment, ordering and strategic information provisioning decisions.

It is important to recognize that real-life practical applications of our consumer choice model would benefit from developments in two key areas. We make advances in both of these areas. The first one pertains to the empirical validation of rational inattention and estimation of the choice model. Complementing the fast growing recent economics literature on the empirical dimension of rational inattention, we demonstrate how market share data can be used to infer the utility of customers with limited attention using methods from logit models, and also discuss how preferences and information costs can be elicited using data from experiments. The second development that is needed pertains to solution methods. In order to solve realistically-sized practical problems, it is necessary to develop an efficient algorithm to first solve the consumer choice model, so that the algorithm can be readily embedded in subsequent firm decision problems such as pricing and assortment optimization. We develop an iterative algorithm for this purpose that exploits the necessary and sufficient conditions derived in the paper.

A. List of Symbols

A : Finite set of alternatives; also the r.v. taking values $i \in A$ with the probability with which i is chosen

$\mathbf{\Omega} = (\Omega_1 \times \ldots \times \Omega_n)$: The state of the world, i.e., a r.v. taking values $\boldsymbol{\omega} \in \mathbb{R}^n$

$u(i, \boldsymbol{\omega})$: The utility obtained from taking alternative i in state $\boldsymbol{\omega}$

$\mathbf{\Omega}_{1..k} = (\Omega_1 \times \ldots \times \Omega_k)$: k-dim. subspace of $\mathbf{\Omega}$, taking values $\boldsymbol{\omega}_{1..k} \in \mathbb{R}^k$ of the k easiest to learn components

$\mathbf{S}$: The signal space, i.e., a r.v. taking values $\mathbf{s} \in \mathbb{R}^n$

$\Delta(X)$: The set of all probability distributions on r.v. X

$g \in \Delta(\mathbf{\Omega})$: Prior belief

$f \in \Delta(\mathbf{\Omega} \times \mathbf{S})$: Joint probability distribution over $\mathbf{\Omega}$ and $\mathbf{S}$ that captures the information strategy

$f(\cdot \mid \mathbf{s})$: Posterior belief given the reception of signal $\mathbf{s}$

$f(\mathbf{s} \mid \boldsymbol{\omega})$: Conditional probability of receiving signal $\mathbf{s}$ if the state of the world is $\boldsymbol{\omega}$

$f(\mathbf{s})$: Probability of receiving signal $\mathbf{s}$ (marginal probability distribution of f)

$R(f) \in \mathbb{R}$: The expected gross payoff implied by information strategy f (ignoring the cost of information)

$C(f) \in \mathbb{R}_+$: Cost of information implied by the information strategy f

$H_f(X) \in \mathbb{R}_+$: Entropy of the (marginal) probability distribution f of the random variable X

$I_f(X,Y) \in \mathbb{R}_+$: Mutual information between X and Y under the joint distribution $f \in \Delta(X \times Y)$

$I_f(X,Y \mid Z) \in \mathbb{R}_+$: Conditional mutual information between X and Y given Z where $f \in \Delta(X \times Y \times Z)$

$\lambda_k \in \mathbb{R}_+$: The unit cost of information associated to learning about Ω_k; w.l.o.g. $\lambda_1 \leq \cdots \leq \lambda_n$

$p \in \Delta(\mathbf{\Omega} \times A)$: Joint probability distribution over $\mathbf{\Omega}$ and A that captures state related choice behavior

$p(i \mid \boldsymbol{\omega})$: Conditional probability of choosing alternative $i \in A$ when the state is $\boldsymbol{\omega}$

$p(i|\boldsymbol{\omega}_{1..k-1})$: Partial conditional probability of choosing $i \in A$ when $\boldsymbol{\omega} \in \{\boldsymbol{\omega}_{1..k-1}\} \times \Omega_k \times \ldots \times \Omega_n$

$p(i)$: Unconditional probability of choosing alternative $i \in A$

$\alpha_i \in \mathbb{R}$: Shift term for the utility in the GMNL model for alternative $i \in A$

$W(\mathbf{p})$: Objective function value of the consumer for a given vector $\mathbf{p} = (p(i \mid \boldsymbol{\omega}_{1..n-1}))_{i \in A, \boldsymbol{\omega}_{1..n-1} \in \mathbf{\Omega}_{1..n-1}}$

$M(i), M(i \mid \boldsymbol{\omega}_{1..n-1}), M(i \mid \boldsymbol{\omega})$: Unconditional and (partial) conditional market shares of alternative i

$g^\star$: True distribution of consumer types

μ_t : Average of consumer type distributions being feasible in period t, prior to RI choice in period t

$M_t(i), M_t(i \mid \boldsymbol{\omega}_{1..n-1}), M_t(i \mid \boldsymbol{\omega})$: Market shares of alternative i after t periods

$U_{li}^{\boldsymbol{\omega}}$: Utility of consumer l of type $\boldsymbol{\omega}$ from choosing alternative i

B. Proofs

For ease of notation, we set $\lambda_0 = 0$ and write unconditionals by conditioning on nothing, e.g. $p(i) = p(i \mid \boldsymbol{\omega}_{1..0})$. We further write $\Pi(\mathbf{p}_i, \boldsymbol{\omega}_{1..n-1})$, which depends on the i-components of $\mathbf{p}$ and $\boldsymbol{\omega}_{1..n-1}$, and is given by

$$\Pi(\mathbf{p}_i, \boldsymbol{\omega}_{1..n-1}) \quad = \quad p(i)^{\frac{\lambda_1}{\lambda_n}} \prod_{k=1}^{n} p(i \mid \boldsymbol{\omega}_{1..k})^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}.$$

Proof of Proposition 1

For a given information strategy f, define the set of signals that lead to the choice of alternative i as

$$S_i(f) := \quad \left\{ \mathbf{s} \in \mathbf{S} \mid i \in \arg\max_{i \in A} \sum_{\boldsymbol{\omega}} f(\boldsymbol{\omega} \mid \mathbf{s}) u(i, \boldsymbol{\omega}) \right\}.$$

Accordingly, we calculate the conditional choice probability for alternative i given state $\boldsymbol{\omega}$ as $p(i \mid \boldsymbol{\omega}) := \sum_{\mathbf{s} \in S_i} f(\mathbf{s} \mid \boldsymbol{\omega})$. This defines the joint probability $p \in \Delta(\boldsymbol{\Omega} \times A)$ via $p(\boldsymbol{\omega}, i) = p(i \mid \boldsymbol{\omega}) g(\boldsymbol{\omega})$ and the unconditional probability of choosing i as $p(i) := \sum_{\boldsymbol{\omega}} p(i \mid \boldsymbol{\omega}) g(\boldsymbol{\omega})$. We next invoke a lemma stating that the posterior beliefs are identical for signals that induce the same action.

Lemma 2 *Let f be optimal and let $i \in A$ be such that $p(i \mid \boldsymbol{\omega}) > 0$. Then, for all signals $\mathbf{s}', \mathbf{s}'' \in S_i$ the posterior beliefs are identical, i.e., $f(\cdot \mid \mathbf{s}') = f(\cdot \mid \mathbf{s}'')$.*

Proof of Lemma 2

The proof is by contradiction. Suppose that $i \in A$ is such that $p_f(i, \boldsymbol{\omega}) > 0$, and that there exist $S', S'' \subseteq S_i$, satisfying $\sum_{\boldsymbol{\omega} \in \boldsymbol{\Omega}} \sum_{\mathbf{s}' \in S'} f(\mathbf{s}', \boldsymbol{\omega}) > 0$, $\sum_{\boldsymbol{\omega} \in \boldsymbol{\Omega}} \sum_{\mathbf{s}'' \in S''} f(\mathbf{s}'', \boldsymbol{\omega}) > 0$, and $f(\cdot \mid \mathbf{s}') \neq f(\cdot \mid \mathbf{s}'')$ for all $\mathbf{s}' \in S', \mathbf{s}'' \in S''$. We can construct a better information strategy h as follows. Pick some $\hat{\mathbf{s}} \in S' \cup S''$. Define h by setting for all $\boldsymbol{\omega}$:

$$h(\mathbf{s}, \boldsymbol{\omega}) \quad : \quad = f(\mathbf{s}, \boldsymbol{\omega}) \text{ for all } \mathbf{s} \notin \left(S' \cup S''\right), \tag{21}$$

$$h(\hat{\mathbf{s}} \mid \boldsymbol{\omega}) \quad : \quad = \sum_{\mathbf{s}' \in S'} f(\mathbf{s}' \mid \boldsymbol{\omega}) + \sum_{\mathbf{s}'' \in S''} f(\mathbf{s}'' \mid \boldsymbol{\omega}), \text{ and} \tag{22}$$

$$h(\mathbf{s}, \boldsymbol{\omega}) \quad : \quad = 0 \text{ for all } \mathbf{s} \in \left(S' \cup S''\right) \setminus \{\hat{\mathbf{s}}\}. \tag{23}$$

Note that h is consistent with g. The consumer chooses i under h whenever observing some signal $\mathbf{s} \in S' \cup S''$ under f; other choices remain unaffected such that h yields the same revenue as f. Thus, it suffices to show that the information costs of h is lower than of f. Since the difference between the mutual information of h and f stems from where the distributions differ, it is helpful to make use of the the probability distributions restricted to this domain. More precisely, we construct a probability distribution $f|_{S' \cup S'', \mathbf{\Omega}_{1..k}}$ on the restricted domain $\mathcal{D} := (S' \cup S'') \times \mathbf{\Omega}_{1..k}$ from f by rescaling to $f(\mathcal{D}) = \sum_{(\mathbf{s}, \boldsymbol{\omega}_{1..k}) \in \mathcal{D}} f\left(\mathbf{s}, \boldsymbol{\omega}_{1..k}\right)$. Formally, let

$$f|_{\mathcal{D}}\left(\mathbf{s}, \boldsymbol{\omega}_{1..k}\right) = f\left(\mathcal{D}\right)^{-1} f\left(\mathbf{s}, \boldsymbol{\omega}_{1..k}\right) \qquad \text{for all } \left(\mathbf{s}, \boldsymbol{\omega}_{1..k}\right) \in \mathcal{D}.$$

Analogously, we define $h|_{\mathcal{D}}$. Now, dividing the following equation by $f\left(\mathcal{D}\right)$,

$$\begin{aligned}
& I_f\left(\Omega_k, \mathbf{S} \mid \mathbf{\Omega}_{1..k-1}\right) - I_h\left(\Omega_k, \mathbf{S} \mid \mathbf{\Omega}_{1..k-1}\right) \\
= \quad & \sum_{\mathbf{s} \in \mathbf{S}} \sum_{\boldsymbol{\omega}_{1..k}} f\left(\mathbf{s}, \boldsymbol{\omega}_{1..k}\right) \left(\log \frac{f\left(\cdot \mid \boldsymbol{\omega}_{1..k}\right)}{f\left(\cdot \mid \boldsymbol{\omega}_{1..k-1}\right)} \right) \\
& - \sum_{\mathbf{s} \in \mathbf{S}} \sum_{\boldsymbol{\omega}_{1..k}} h\left(\mathbf{s}, \boldsymbol{\omega}_{1..k}\right) \left(\log \frac{h\left(\cdot \mid \boldsymbol{\omega}_{1..k}\right)}{h\left(\cdot \mid \boldsymbol{\omega}_{1..k-1}\right)} \right) \\
\overset{(21)\text{-}(23)}{=} \quad & \sum_{(\mathbf{s}, \boldsymbol{\omega}_{1..k}) \in \mathcal{D}} f\left(\mathbf{s}, \boldsymbol{\omega}_{1..k}\right) \left(\log \frac{f\left(\cdot \mid \boldsymbol{\omega}_{1..k}\right)}{f\left(\cdot \mid \boldsymbol{\omega}_{1..k-1}\right)} \right) \\
& - \sum_{\boldsymbol{\omega}_{1..k}} h\left(\hat{\mathbf{s}}, \boldsymbol{\omega}_{1..k}\right) \left(\log \frac{h\left(\hat{\mathbf{s}} \mid \boldsymbol{\omega}_{1..k}\right)}{h\left(\hat{\mathbf{s}} \mid \boldsymbol{\omega}_{1..k-1}\right)} \right),
\end{aligned}$$

we get

$$\begin{aligned}
& f\left(\mathcal{D}\right)^{-1} \cdot \left[I_f\left(\Omega_k, \mathbf{S} \mid \mathbf{\Omega}_{1..k-1}\right) - I_h\left(\Omega_k, \mathbf{S} \mid \mathbf{\Omega}_{1..k-1}\right) \right] \\
& \quad = \sum_{(\mathbf{s}, \boldsymbol{\omega}_{1..k}) \in \mathcal{D}} f|_{\mathcal{D}}\left(\mathbf{s}, \boldsymbol{\omega}_{1..k}\right) \log \frac{f|_{\mathcal{D}}\left(\mathbf{s} \mid \boldsymbol{\omega}_{1..k}\right)}{f|_{\mathcal{D}}\left(\mathbf{s} \mid \boldsymbol{\omega}_{1..k-1}\right)} \\
& \qquad - \sum_{\boldsymbol{\omega}_{1..k}} h|_{\mathcal{D}}\left(\hat{\mathbf{s}}, \boldsymbol{\omega}_{1..k}\right) \log \frac{h|_{\mathcal{D}}\left(\hat{\mathbf{s}} \mid \boldsymbol{\omega}_{1..k}\right)}{h|_{\mathcal{D}}\left(\hat{\mathbf{s}} \mid \boldsymbol{\omega}_{1..k-1}\right)},
\end{aligned}$$

since the ratio of the conditionals within the $\log$ terms remains the same when restricting the domain to $\mathcal{D}$. Applying $\Pr\left(x, y\right) = \Pr\left(x \mid y\right) \cdot \Pr\left(y\right)$, we obtain

$$
\begin{aligned}
&\frac{I_f\left(\Omega_k, \mathbf{S} \mid \mathbf{\Omega}_{1..k-1}\right) - I_h\left(\Omega_k, \mathbf{S} \mid \mathbf{\Omega}_{1..k-1}\right)}{f\left(\mathcal{D}\right)} \\
= &\sum_{(\mathbf{s},\boldsymbol{\omega}_{1..k})\in\mathcal{D}} f|_{\mathcal{D}}\left(\mathbf{s},\boldsymbol{\omega}_{1..k}\right)\log\left(\frac{f|_{\mathcal{D}}\left(\mathbf{s},\boldsymbol{\omega}_{1..k}\right)}{f|_{\mathcal{D}}\left(\mathbf{s},\boldsymbol{\omega}_{1..k-1}\right)} - \frac{f|_{\mathcal{D}}\left(\boldsymbol{\omega}_{1..k}\right)}{f|_{\mathcal{D}}\left(\boldsymbol{\omega}_{1..k-1}\right)}\right) \\
&- \sum_{\boldsymbol{\omega}_{1..k}} h|_{\mathcal{D}}\left(\hat{\mathbf{s}},\boldsymbol{\omega}_{1..k}\right)\log\left(\frac{h|_{\mathcal{D}}\left(\hat{\mathbf{s}} \mid \boldsymbol{\omega}_{1..k}\right)}{h|_{\mathcal{D}}\left(\hat{\mathbf{s}} \mid \boldsymbol{\omega}_{1..k-1}\right)} - \frac{h|_{\mathcal{D}}\left(\boldsymbol{\omega}_{1..k}\right)}{h|_{\mathcal{D}}\left(\boldsymbol{\omega}_{1..k-1}\right)}\right)
\end{aligned}
$$

By the construction of h, (21)-(23), this can be further simplified to

$$
\begin{aligned}
&f\left(\mathcal{D}\right)^{-1}\cdot\left[I_f\left(\Omega_k, \mathbf{S} \mid \mathbf{\Omega}_{1..k-1}\right) - I_h\left(\Omega_k, \mathbf{S} \mid \mathbf{\Omega}_{1..k-1}\right)\right] \\
= &\sum_{(\mathbf{s},\boldsymbol{\omega}_{1..k})\in\mathcal{D}} f|_{\mathcal{D}}\left(\mathbf{s},\boldsymbol{\omega}_{1..k}\right)\left(\log\frac{f|_{\mathcal{D}}\left(\mathbf{s},\boldsymbol{\omega}_{1..k}\right)}{f|_{\mathcal{D}}\left(\mathbf{s},\boldsymbol{\omega}_{1..k-1}\right)}\right) \\
&- \sum_{\boldsymbol{\omega}_{1..k}\in\mathbb{R}^k} h|_{\mathcal{D}}\left(\hat{\mathbf{s}},\boldsymbol{\omega}_{1..k}\right)\left(\log\frac{h|_{\mathcal{D}}\left(\hat{\mathbf{s}},\boldsymbol{\omega}\right)}{h|_{\mathcal{D}}\left(\hat{\mathbf{s}},\boldsymbol{\omega}_{1..k-1}\right)}\right) \\
= &\; H_{f|_{\mathcal{D}}}(\mathbf{S},\mathbf{\Omega}_{1..k-1}) - H_{f|_{\mathcal{D}}}(\mathbf{S},\mathbf{\Omega}_{1..k}) - \left[H_{f|_{\mathcal{D}}}\left(\mathbf{\Omega}_{1..k-1}\right) - H_{f|_{\mathcal{D}}}(\mathbf{\Omega}_{1..k})\right] \\
= &\; I_{f|_{\mathcal{D}}}\left(\mathbf{S},\mathbf{\Omega}_{1..k} \mid \mathbf{\Omega}_{1..k-1}\right) \geq 0.
\end{aligned}
$$

Since we have multiple posteriors on $S' \cup S''$, it cannot be true that for all $k = 1,\ldots,n$, $\mathbf{S}$ and $\mathbf{\Omega}_{1..k}$ are independent conditional on $\mathbf{\Omega}_{1..k-1}$. Hence, there exists k for which the last term is strictly positive. But then

$$I_f\left(\Omega_k, \mathbf{S} \mid \mathbf{\Omega}_{1..k-1}\right) - I_h\left(\Omega_k, \mathbf{S} \mid \mathbf{\Omega}_{1..k-1}\right) > 0,$$

implying that h is strictly cheaper than f; a contradiction.

Proof of Proposition 1 (Continued)

According to Lemma 2, an optimal information strategy f necessitates a single posterior for all signals $\mathbf{s} \in S_i(f)$. Denote by p the probabilistic choice induced by f. With,

$$
\begin{aligned}
&\sum_{\boldsymbol{\omega}}\sum_{i\in A}\sum_{\mathbf{s}\in S_i} f(\mathbf{s} \mid \boldsymbol{\omega})g(\boldsymbol{\omega})\sum_{\boldsymbol{\omega}'} f(\boldsymbol{\omega}' \mid \mathbf{s})u(i,\boldsymbol{\omega}') \\
= &\sum_{i\in A}\left(\sum_{\boldsymbol{\omega}'} f(\boldsymbol{\omega}' \mid \mathbf{s}\in S_i)u(i,\boldsymbol{\omega}')\right)\sum_{\boldsymbol{\omega}}\sum_{\mathbf{s}\in S_i} f(\mathbf{s} \mid \boldsymbol{\omega})g(\boldsymbol{\omega}) \\
= &\sum_{i\in A}\left(\sum_{\boldsymbol{\omega}'} p(\boldsymbol{\omega}' \mid i)u(i,\boldsymbol{\omega}'))\right)p(i),
\end{aligned}
$$

we infer

$$R(f) = \sum_{i \in A} \sum_{\boldsymbol{\omega}} u(i, \boldsymbol{\omega}) p(i \mid \boldsymbol{\omega}) g(\boldsymbol{\omega}). \tag{24}$$

Turning to the information cost, denote with A the random variable that takes the value i with the probability that i is chosen. We argue that the distribution of $\boldsymbol{\Omega}$ conditional on $\mathbf{S}$ is the same as the distribution of $\boldsymbol{\Omega}$ conditional on A, hence knowing A is as informative about $\boldsymbol{\Omega}$ as knowing $\mathbf{S}$, i.e.,

$$I_f\left(\Omega_\ell, \mathbf{S} \mid \boldsymbol{\Omega}_{1..\ell-1}\right) = I_p\left(\Omega_\ell, A \mid \boldsymbol{\Omega}_{1..\ell-1}\right). \tag{25}$$

Formally, this can be seen as follows: $p(\cdot \mid i, \boldsymbol{\omega}_{1..k-1}) = f(\cdot \mid \mathbf{s}, \boldsymbol{\omega}_{1..k-1}) \in \Delta(\Omega_k)$ for $\mathbf{s} \in S_i$ implies $H_{p(\cdot|i,\boldsymbol{\omega}_{1..k-1})}(\Omega_k) = H_{f(\cdot|\mathbf{s},\boldsymbol{\omega}_{1..k-1})}(\Omega_k)$ for $\mathbf{s} \in S_i$. Thus,

$$\begin{aligned}
I_f\left(\Omega_k, \mathbf{S} \mid \boldsymbol{\Omega}_{1..k-1}\right) &= \sum_{\boldsymbol{\omega}_{1..k-1}} f\left(\boldsymbol{\omega}_{1..k-1}\right) H_{f(\cdot|\boldsymbol{\omega}_{1..k-1})}\left(\Omega_k\right) \\
&\quad - \sum_{\boldsymbol{\omega}_{1..k-1}} \sum_{i} \sum_{\mathbf{s} \in S_i} f\left(\mathbf{s}\right) f\left(\boldsymbol{\omega}_{1..k-1} \mid \mathbf{s}\right) H_{f(\cdot|\mathbf{s},\boldsymbol{\omega}_{1..k-1})}\left(\Omega_k\right) \\
&= \sum_{\boldsymbol{\omega}_{1..k-1}} f\left(\boldsymbol{\omega}_{1..k-1}\right) H_{f(\cdot|\boldsymbol{\omega}_{1..k-1})}\left(\Omega_k\right) \\
&\quad - \sum_{\boldsymbol{\omega}_{1..k-1}} \sum_{i} p\left(i\right) p\left(\boldsymbol{\omega}_{1..k-1} \mid a\right) H_{p(\cdot|a,\boldsymbol{\omega}_{1..k-1})}\left(\Omega_k\right) \\
&= I_p\left(\Omega_k, A \mid \boldsymbol{\Omega}_{1..k-1}\right).
\end{aligned}$$

The cost of information can now be written as $\sum_{k=1}^{n} \lambda_k \cdot I_p\left(\Omega_k, A \mid \boldsymbol{\Omega}_{1..k-1}\right)$. Using $H_{p(\cdot|\boldsymbol{\omega}_{1..k})}(A) = -\sum_{i \in A} p\left(i \mid \boldsymbol{\omega}_{1..k}\right) \log p\left(i \mid \boldsymbol{\omega}_{1..k}\right)$ and $I_p\left(\Omega_k, A \mid \boldsymbol{\Omega}_{1..k-1}\right) H_{p(\cdot|\boldsymbol{\omega}_{1..k})}\left(A\right) - H_{p(\cdot|\boldsymbol{\omega}_{1..k-1})}\left(A\right)$, we can write the objective in terms of $\left\{p\left(i \mid \boldsymbol{\omega}_{1..k}\right)\right\}_{i \in A, \boldsymbol{\omega}_{1..k} \in \boldsymbol{\Omega}_{1..k}}$. Thus, every optimal information strategy f induces choice probabilities p that also maximize the objectives of Proposition 1.

On the other hand, let $\left\{p^*\left(i \mid \boldsymbol{\omega}_{1..k}\right)\right\}_{i \in A, \boldsymbol{\omega}_{1..k} \in \boldsymbol{\Omega}_{1..k}}$ solve the problem in the Proposition 1. Select $|A|$ distinct signals $\left\{\bar{\mathbf{s}}_i\right\}_{i \in A}$. Define f^* by

$$f^*\left(\mathbf{s}, \boldsymbol{\omega}\right) = \begin{cases} p\left(i \mid \boldsymbol{\omega}\right) p\left(\boldsymbol{\omega}\right), & \text{if } \mathbf{s} = \bar{\mathbf{s}} \\ 0, & \text{otherwise.} \end{cases}$$

Then, f^* is consistent with the prior belief, i.e., it satisfies (7). Suppose there were another information strategy $\hat{f}$ that achieves a higher objective(6) than f^*. Then, (24) and (25) imply that $\hat{f}$ induces choice probabilities that do better in the problem of Proposition 1 than $\{p^*(i \mid \boldsymbol{\omega}_{1..k})\}_{i\in A,\boldsymbol{\omega}_{1..k}\in\boldsymbol{\Omega}_{1..k}}$, a contradiction.

Proof of Theorem 1

The Lagrangian with variables $(p(i \mid \boldsymbol{\omega}))_{i\in A,\boldsymbol{\omega}\in\boldsymbol{\Omega}}$ and multipliers

$$(\xi(i \mid \boldsymbol{\omega}))_{i\in A,\boldsymbol{\omega}\in\boldsymbol{\Omega}} \geq 0$$

on inequalities and multipliers $(\varphi(\boldsymbol{\omega}))_{\boldsymbol{\omega}\in\boldsymbol{\Omega}}$ on equalities is given by

$$\begin{aligned}
&\sum_{i\in A}\sum_{\boldsymbol{\omega}\in\boldsymbol{\Omega}} u(i,\boldsymbol{\omega})\, p(i \mid \boldsymbol{\omega})\, g(\boldsymbol{\omega}) \\
&-\sum_{k=1}^{n}\lambda_k \sum_{\boldsymbol{\omega}_{1..k}} g(\boldsymbol{\omega}_{1..k}) \sum_{j\in A} p(j \mid \boldsymbol{\omega}_{1..k}) \log p(j \mid \boldsymbol{\omega}_{1..k}) \\
&+\sum_{k=1}^{n}\lambda_k \sum_{\boldsymbol{\omega}_{1..k}} g(\boldsymbol{\omega}_{1..k-1}) \sum_{j\in A} p(j \mid \boldsymbol{\omega}_{1..k-1}) \log p(j \mid \boldsymbol{\omega}_{1..k-1}) \\
&+\sum_{i\in A}\sum_{\boldsymbol{\omega}\in\boldsymbol{\Omega}} \xi(i \mid \boldsymbol{\omega})\, p(i \mid \boldsymbol{\omega})\, g(\boldsymbol{\omega}) + \sum_{\boldsymbol{\omega}\in\boldsymbol{\Omega}} \varphi(\boldsymbol{\omega}) \sum_{i\in A} (1 - p(i \mid \boldsymbol{\omega}))\, g(\boldsymbol{\omega})
\end{aligned}$$

If $p(i \mid \boldsymbol{\omega}_{1..n-1}) > 0$, then $p(i \mid \boldsymbol{\omega}_{1..k}) > 0$ for $k < n$ and the first-order condition with respect to $p(i \mid \boldsymbol{\omega})$ is

$$0 = u(i,\boldsymbol{\omega}) + \sum_{k=1}^{n}\lambda_k \left(\log p(i \mid \boldsymbol{\omega}_{1..k-1}) - \log p(i \mid \boldsymbol{\omega}_{1..k})\right) + \xi(i \mid \boldsymbol{\omega}) - \varphi(\boldsymbol{\omega})$$

for $g(\boldsymbol{\omega}) > 0$. This implies $p(i \mid \boldsymbol{\omega}) > 0$: Suppose $p(i \mid \boldsymbol{\omega}) = 0$. Then, $\log p(i \mid \boldsymbol{\omega}_{1..n-1}) - \log p(i \mid \boldsymbol{\omega}) = +\infty$, implying $\varphi(\boldsymbol{\omega}) = +\infty$. However, $\varphi(\boldsymbol{\omega}) = +\infty$ requires $\xi(j \mid \boldsymbol{\omega}) = +\infty$ for j such that $p(j \mid \boldsymbol{\omega}) > 0$, a contradiction. Consequently, $\xi(i \mid \boldsymbol{\omega}) = 0$, giving

$$e^{\frac{\varphi(\boldsymbol{\omega})}{\lambda_n}} p(i \mid \boldsymbol{\omega}) = e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}} \prod_{k=0}^{n-1} p(i \mid \boldsymbol{\omega}_{1..k})^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}. \tag{26}$$

Summing up over $j \in A$ gives

$$e^{\frac{\varphi(\boldsymbol{\omega})}{\lambda_n}} = \sum_{j \in A} e^{u(j,\boldsymbol{\omega})} \prod_{k=0}^{n-1} p\left(j \mid \boldsymbol{\omega}_{1..k}\right)^{\lambda_{k+1}-\lambda_k} \tag{27}$$

and insertion of (27) into (26) yields (10). If $p\left(i \mid \boldsymbol{\omega}_{1..n-1}\right) = 0$, the claim is trivially true.

Proof of Lemma 1

Insertion of the necessary conditions (10) into the objective of Proposition 1 yields

$$\begin{aligned}
&\sum_{i \in A} \sum_{\boldsymbol{\omega} \in \boldsymbol{\Omega}} u\left(i, \boldsymbol{\omega}\right) p\left(i \mid \boldsymbol{\omega}\right) g\left(\boldsymbol{\omega}\right) \\
&+ \sum_{k=1}^{n} \lambda_k \sum_{\boldsymbol{\omega}_{1..k}} g\left(\boldsymbol{\omega}_{1..k-1}\right) \sum_{j \in A} p\left(j \mid \boldsymbol{\omega}_{1..k-1}\right) \log p\left(j \mid \boldsymbol{\omega}_{1..k-1}\right) \\
&- \sum_{k=1}^{n-1} \lambda_k \sum_{\boldsymbol{\omega}_{1..k}} g\left(\boldsymbol{\omega}_{1..k}\right) \sum_{j \in A} p\left(j \mid \boldsymbol{\omega}_{1..k}\right) \log p\left(j \mid \boldsymbol{\omega}_{1..k}\right) \\
&- \lambda_n \sum_{\boldsymbol{\omega}} g\left(\boldsymbol{\omega}\right) \sum_{j \in A} p\left(j \mid \boldsymbol{\omega}\right) \log \frac{e^{\left(\frac{u(i,\boldsymbol{\omega})}{\lambda_n}\right)} p\left(i\right)^{\frac{\lambda_1}{\lambda_n}} \prod_{k=1}^{n} p\left(i \mid \boldsymbol{\omega}_{1..k}\right)^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}}{\sum_{j \in A} e^{\left(\frac{u(j,\boldsymbol{\omega})}{\lambda_n}\right)} p\left(j\right)^{\frac{\lambda_1}{\lambda_n}} \prod_{k=1}^{n} p\left(j \mid \boldsymbol{\omega}_{1..k}\right)^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}}.
\end{aligned}$$

Straightforward application of logarithm calculation and cancelation of terms gives $W(\mathbf{p})$. Now, if $\mathbf{p}^*$ solves the program of Lemma 1, calculating conditionals $\{p^*(i \mid \boldsymbol{\omega})\}_{i,\boldsymbol{\omega}}$ via (10) yields conditionals that solve the objective of Proposition 1 (otherwise, there would exist conditionals $\{p^{**}(i \mid \boldsymbol{\omega})\}_{i,\boldsymbol{\omega}}$ with partial conditionals $\mathbf{p}^{**}$ that relate to another via (10) and for which $W(\mathbf{p}^{**}) > W(\mathbf{p}^*)$, a contradiction to the optimality of $\mathbf{p}^*$). By the same argument, solving Proposition 1 induces a solution of the objective of Lemma 1.

Proof of Theorem 2

We first show that the objective of the *alternative formulation is a concave*. To establish concavity of W, let $\mathbf{p} = \delta \mathbf{p}' + (1-\delta)\,\mathbf{p}''$, for some $\mathbf{p}'$, $\mathbf{p}''$, and

$\delta \in (0,1)$. By (9), $p(i \mid \boldsymbol{\omega}_{1..k}) = \delta p'(i \mid \boldsymbol{\omega}_{1..k}) + (1-\delta) p''(i \mid \boldsymbol{\omega}_{1..k})$ for $k < n$. Applying Jensen's inequality gives

$$\begin{aligned}&\left(\delta p'(i \mid \boldsymbol{\omega}_{1..k}) + (1-\delta) p''(i \mid \boldsymbol{\omega}_{1..k})\right)^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}} \\ &\qquad \geq \delta p'(i \mid \boldsymbol{\omega}_{1..k})^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}} + (1-\delta) p''(i \mid \boldsymbol{\omega}_{1..k})^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}},\end{aligned}$$

Since all terms are non-negative, we get $\Pi(\mathbf{p}_i, \boldsymbol{\omega}_{1..n-1}) \geq \delta \Pi(\mathbf{p}'_i, \boldsymbol{\omega}_{1..n-1}) + (1-\delta)\Pi(\mathbf{p}''_i, \boldsymbol{\omega}_{1..n-1})$, i.e.,

$$\begin{aligned}&\prod_{k=0}^{n-1}\left(\delta p'(i \mid \boldsymbol{\omega}_{1..k})^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}} + (1-\delta) p''(i \mid \boldsymbol{\omega}_{1..k})^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}\right) \\ &\qquad \geq \delta \prod_{k=0}^{n-1} p'(i \mid \boldsymbol{\omega}_{1..k})^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}} + (1-\delta)\prod_{k=0}^{n-1} p''(i \mid \boldsymbol{\omega}_{1..k})^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}} \qquad (28)\end{aligned}$$

This establishes the next inequality; the second to the next inequality again follows with Jensen's inequality:

$$\begin{aligned}&\log\left(\sum_{i\in A} e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}}\Pi(\mathbf{p}_i, \boldsymbol{\omega}_{1..n-1})\right) \\ &\geq \log\left(\delta\sum_{i\in A} e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}}\Pi(\mathbf{p}'_i, \boldsymbol{\omega}_{1..n-1}) + (1-\delta)\sum_{i\in A} e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}}\Pi(\mathbf{p}''_i, \boldsymbol{\omega}_{1..n-1})\right) \\ &\geq \delta\log\left(\sum_{i\in A} e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}}\Pi(\mathbf{p}'_i, \boldsymbol{\omega}_{1..n-1})\right) + (1-\delta)\log\left(\sum_{i\in A} e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}}\Pi(\mathbf{p}''_i, \boldsymbol{\omega}_{1..n-1})\right).\end{aligned}$$

Taking the expectation over $\boldsymbol{\omega}$ on both sides establishes the concavity of W.

Next we show that (10) implies (15); i.e., we establish that (15) *is necessary* for $\mathbf{p}^*$ being optimal. Let $\mathbf{p}^*$ be derived via (9) from optimal conditionals $\{p^*(i \mid \boldsymbol{\omega})\}_{i\in A, \boldsymbol{\omega}\in\boldsymbol{\Omega}}$ that satisfy (10). Then, $\mathbf{p}^*$ satisfies (15):

$$\begin{aligned}&p^*(i \mid \boldsymbol{\omega}_{1..n-1}) \\ \overset{(9)}{=}\ &\sum_{\omega_n\in\Omega_n} g(\omega_n \mid \boldsymbol{\omega}_{1..n-1})\, p^*(i \mid \boldsymbol{\omega}) \\ \overset{(10)}{=}\ &\sum_{\omega_n\in\Omega_n} g(\omega_n \mid \boldsymbol{\omega}_{1..n-1})\,\frac{e^{\frac{u(i,\boldsymbol{\omega}_{1..n-1}\omega_n)}{\lambda_n}}\Pi(\mathbf{p}^*_i, \boldsymbol{\omega}_{1..n-1})}{\sum_{j\in A} z(j, \boldsymbol{\omega}_{1..n-1}\omega_n)\Pi(\mathbf{p}^*_j, \boldsymbol{\omega}_{1..n-1})}.\end{aligned}$$

For *sufficiency of* (15), note that the FOC of the Lagrangian wrt $p(i \mid \boldsymbol{\omega}_{1..n-1})$

is $\frac{\partial W(\mathbf{p})}{\partial p(i|\boldsymbol{\omega}_{1\cdot\cdot n-1})} - \varphi(\boldsymbol{\omega}_{1\cdot\cdot n-1}) = 0$, where $\{\varphi(\boldsymbol{\omega}_{1\cdot\cdot n-1})\}_{\boldsymbol{\omega}_{1\cdot\cdot n-1}\in\boldsymbol{\Omega}_{1\cdot\cdot n-1}}$ are the multipliers on the equality constraints. We will show that (15) yields $\partial W(\mathbf{p})/\partial p(i \mid \boldsymbol{\omega}_{1\cdot\cdot n-1}) = \lambda_n g(\boldsymbol{\omega}_{1\cdot\cdot n-1})$, establishing the KKT conditions for an interior point. Note that with

$$\frac{\partial p(j \mid \boldsymbol{\omega}_{1\cdot\cdot k})}{\partial p(i \mid \tilde{\boldsymbol{\omega}}_{1\cdot\cdot n-1})} = \begin{cases} g(\tilde{\boldsymbol{\omega}}_{k+1\cdot\cdot n-1} \mid \tilde{\boldsymbol{\omega}}_{1\cdot\cdot k}), & \text{if } \tilde{\boldsymbol{\omega}}_{1\cdot\cdot k} = \boldsymbol{\omega}_{1\cdot\cdot k} \text{ and } j = i, \\ 0, & \text{otherwise.} \end{cases}$$

and using the generalized product rule

$$\frac{d}{dx}\left[\prod_{i=1}^{k} g_i(x)\right] = \left(\prod_{j=1}^{k} g_j(x)\right)\left(\sum_{i=1}^{k} \frac{g_i'(x)}{g_i(x)}\right),$$

we obtain

$$\frac{\partial \Pi(\mathbf{p}_i, \boldsymbol{\omega}_{1\cdot\cdot n-1})}{\partial p(i \mid \tilde{\boldsymbol{\omega}}_{1\cdot\cdot n-1})} = \Pi(\mathbf{p}_i, \boldsymbol{\omega}_{1\cdot\cdot n-1}) \sum_{\ell=0}^{c(\boldsymbol{\omega}_{1\cdot\cdot\ell}, \tilde{\boldsymbol{\omega}}_{1\cdot\cdot n-1})} \frac{\lambda_{\ell+1} - \lambda_\ell}{\lambda_n} \frac{g(\tilde{\boldsymbol{\omega}}_{\ell+1\cdot\cdot n-1} \mid \boldsymbol{\omega}_{1\cdot\cdot\ell})}{p(i \mid \tilde{\boldsymbol{\omega}}_{1\cdot\cdot\ell})},$$

where $c(\boldsymbol{\omega}_{1\cdot\cdot\ell}, \tilde{\boldsymbol{\omega}}_{1\cdot\cdot n-1}) = \max\{k \mid \boldsymbol{\omega}_{1\cdot\cdot k} = \tilde{\boldsymbol{\omega}}_{1\cdot\cdot k}\}$ denotes the maximal index upon which $\boldsymbol{\omega}_{1\cdot\cdot\ell}$ and $\tilde{\boldsymbol{\omega}}_{1\cdot\cdot n-1}$ coincide. Hence, we have

$$\begin{aligned} &\frac{\partial W(\mathbf{p})}{\partial p(i \mid \tilde{\boldsymbol{\omega}}_{1\cdot\cdot n-1})} \\ &= \lambda_n \sum_{\boldsymbol{\omega}} g(\boldsymbol{\omega}) \frac{e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}_i, \boldsymbol{\omega}_{1\cdot\cdot n-1}) \sum_{\ell=0}^{c(\boldsymbol{\omega}_{1\cdot\cdot\ell}, \tilde{\boldsymbol{\omega}}_{1\cdot\cdot n-1})} \frac{\lambda_{\ell+1}-\lambda_\ell}{\lambda_n} \frac{g(\tilde{\boldsymbol{\omega}}_{\ell+1\cdot\cdot n-1} \mid \tilde{\boldsymbol{\omega}}_{1\cdot\cdot\ell})}{p(i \mid \tilde{\boldsymbol{\omega}}_{1\cdot\cdot\ell})}}{\sum_{j\in A} e^{\frac{u(j,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}_j, \boldsymbol{\omega}_{1\cdot\cdot n-1})} \end{aligned}$$

This gives the following formula of the gradient for $p(i \mid \tilde{\boldsymbol{\omega}}_{1\cdot\cdot n-1}) > 0$,

$$\begin{aligned} &\nabla_{p(i|\tilde{\boldsymbol{\omega}}_{1\cdot\cdot n-1})} W \\ &= \sum_{c=0}^{n-1} (\lambda_{c+1} - \lambda_c) \frac{g(\tilde{\boldsymbol{\omega}}_{c+1\cdot\cdot n-1} \mid \tilde{\boldsymbol{\omega}}_{1\cdot\cdot c})}{p(i \mid \tilde{\boldsymbol{\omega}}_{1\cdot\cdot c})} \sum_{\boldsymbol{\omega} \text{ s.t. } \boldsymbol{\omega}_{1\cdot\cdot c} = \tilde{\boldsymbol{\omega}}_{1\cdot\cdot c}} \frac{g(\boldsymbol{\omega}) e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}_i, \boldsymbol{\omega}_{1\cdot\cdot n-1})}{\sum_{j\in A} e^{\frac{u(j,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}_j, \boldsymbol{\omega}_{1\cdot\cdot n-1})}, \end{aligned} \tag{29}$$

where $g(\tilde{\boldsymbol{\omega}}_{0+1\cdot\cdot n-1} \mid \tilde{\boldsymbol{\omega}}_{1\cdot\cdot 0}) = g(\tilde{\boldsymbol{\omega}}_{1\cdot\cdot n-1})$ and $g(\tilde{\boldsymbol{\omega}}_{n\cdot\cdot n-1} \mid \tilde{\boldsymbol{\omega}}_{1\cdot\cdot n-1}) = 1$. Ap-

plying (10) to (29) reduces to

$$\begin{aligned}
& \nabla_{p(i|\tilde{\boldsymbol{\omega}}_{1..n-1})} W(\mathbf{p}) \\
= & \sum_{c=0}^{n-1} (\lambda_{c+1} - \lambda_c) \frac{g(\tilde{\boldsymbol{\omega}}_{c+1..n-1} \mid \tilde{\boldsymbol{\omega}}_{1..c})}{p(i \mid \tilde{\boldsymbol{\omega}}_{1..c})} \sum_{\boldsymbol{\omega}_{1..n-1} \text{ s.t. } \boldsymbol{\omega}_{1..c}=\tilde{\boldsymbol{\omega}}_{1..c}} g(\boldsymbol{\omega}_{1..n-1}) p(i|\boldsymbol{\omega}_{1..n-1}) \\
= & \lambda_n g(\tilde{\boldsymbol{\omega}}_{1..n-1}).
\end{aligned}$$

Now we turn to the second part. For *sufficiency of* (16), suppose that $\mathbf{p}^*$ satisfies (16). Then, for all $\mathbf{p}$,

$$\begin{aligned}
& W(\mathbf{p}) - W(\mathbf{p}^*) \\
= \lambda_n & \sum_{\boldsymbol{\omega}} g(\boldsymbol{\omega}) \log \left(\frac{\sum_{i \in A} e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}_i, \boldsymbol{\omega}_{1..n-1})}{\sum_{j \in A} e^{\frac{u(j,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}^*_j, \boldsymbol{\omega}_{1..n-1})} \right) \\
\leq \lambda_n & \sum_{\boldsymbol{\omega}_{1..n-1}} g(\boldsymbol{\omega}_{1..n-1}) \log \sum_{\omega_n} g(\omega_n \mid \boldsymbol{\omega}_{1..n-1}) \left(\frac{\sum_{i \in A} e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}_i, \boldsymbol{\omega}_{1..n-1})}{\sum_{j \in A} e^{\frac{u(j,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}^*_j, \boldsymbol{\omega}_{1..n-1})} \right) \\
\leq \lambda_n & \sum_{\boldsymbol{\omega}_{1..n-1}} g(\boldsymbol{\omega}_{1..n-1}) \log \sum_{i \in A} \Pi(\mathbf{p}_i, \boldsymbol{\omega}_{1..n-1}) \\
= \lambda_n & \sum_{\boldsymbol{\omega}} g(\boldsymbol{\omega}) \log \sum_{i \in A} e^{\frac{0}{\lambda_n}} \Pi(\mathbf{p}_i, \boldsymbol{\omega}_{1..n-1}),
\end{aligned}$$

where the first inequality follows with Jensen's inequality and the second inequality follows with (16). The expression in the last line is the utility obtained from using $\mathbf{p}$ in an RI problem where all utilities are zero in all states. This utility can at best be zero so that $W(\mathbf{p}) - W(\mathbf{p}^*) \leq 0$, which establishes optimality of $\mathbf{p}^*$.

In order to show *necessity of* (16), we show how to improve a $\mathbf{p}$ that violates (16). Let $\Delta(i, \boldsymbol{\omega})$ denote

$$\Delta(i, \boldsymbol{\omega}) = \frac{e^{u(i,\boldsymbol{\omega})/\lambda_n}}{\sum_j e^{u(j,\boldsymbol{\omega})/\lambda_n} \Pi(\mathbf{p}_j, \boldsymbol{\omega}_{1..n-1})} - 1$$

and let $\Delta(i)$ denote the violation of (16) of i, $\Delta(i) = \sum_{\boldsymbol{\omega}} g(\boldsymbol{\omega}) \Delta(i, \boldsymbol{\omega})$. Let $\hat{\imath}$ be the alternative for which (16) is violated most, $\hat{\imath} = \arg_i \max \Delta(i)$ and let $\hat{\omega}$ be the state that maximizes the absolute value of $\Delta(i, \boldsymbol{\omega})$ for $\hat{\imath}$,

$\hat{\boldsymbol{\omega}} = \arg_{\boldsymbol{\omega}} \max |\Delta(\hat{\imath}, \boldsymbol{\omega})|$. Set

$$\varepsilon = \frac{1}{\Delta(\hat{\imath}, \hat{\boldsymbol{\omega}})^2 + \Delta(\hat{\imath})\,\Delta(\hat{\imath}, \hat{\boldsymbol{\omega}})}. \tag{30}$$

Denote by $\hat{\mathbf{p}}$ the choice probabilities giving probability 1 to $\hat{\imath}$. Define $\mathbf{p}^{\varepsilon} = (1-\varepsilon)\,\mathbf{p} + \varepsilon\hat{\mathbf{p}}$. By (28), we get

$$\begin{aligned} & W(\mathbf{p}^{\varepsilon}) - W(\mathbf{p}) \\ & \geq \lambda_n \sum_{\boldsymbol{\omega}} g(\boldsymbol{\omega}) \log \frac{\sum_i e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}} (1-\varepsilon)\,\Pi(\mathbf{p}_i, \boldsymbol{\omega}_{1..n-1}) + \sum_i e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}} \varepsilon \Pi(\hat{\mathbf{p}}_i, \boldsymbol{\omega}_{1..n-1})}{\sum_j e^{\frac{u(j,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}_j, \boldsymbol{\omega}_{1..n-1})} \end{aligned}$$

This reduces to $W(\mathbf{p}^{\varepsilon}) - W(\mathbf{p}) = \lambda_n \sum_{\boldsymbol{\omega}} g(\boldsymbol{\omega}) \log(1 + \varepsilon\Delta(\hat{\imath}, \boldsymbol{\omega}))$. Using the approximation $\log(1+\varepsilon\Delta(\hat{\imath}, \boldsymbol{\omega})) = \varepsilon\Delta(\hat{\imath}, \boldsymbol{\omega}) + R_2(\varepsilon\Delta(\hat{\imath}, \boldsymbol{\omega}))$, where $|R_2(\varepsilon\Delta(\hat{\imath}, \boldsymbol{\omega}))| \leq \varepsilon^2 \Delta(\hat{\imath}, \boldsymbol{\omega})^2 / (1 - \varepsilon\Delta(\hat{\imath}, \boldsymbol{\omega}))$, we obtain

$$W(\mathbf{p}^{\varepsilon}) - W(\mathbf{p}) \geq \lambda_n \sum_{\boldsymbol{\omega}} g(\boldsymbol{\omega}) \left(\varepsilon\Delta(\hat{\imath}, \boldsymbol{\omega}) - \varepsilon^2 \Delta(\hat{\imath}, \boldsymbol{\omega})^2 / (1 - \varepsilon\Delta(\hat{\imath}, \boldsymbol{\omega})) \right).$$

This simplifies to $\frac{(W(\mathbf{p}^{\varepsilon}) - W(\mathbf{p}))}{\varepsilon\lambda_n} > \Delta(\hat{\imath}) - \varepsilon\Delta(\hat{\imath}, \hat{\boldsymbol{\omega}})^2 / (1 - \varepsilon\Delta(\hat{\imath}, \hat{\omega}))$. Using (30) yields

$$\begin{aligned} & \frac{W(\mathbf{p}^{\varepsilon}) - W(\mathbf{p})}{\varepsilon\lambda_n} \\ \geq\ & \Delta(\hat{\imath}) - \frac{1}{\Delta(\hat{\imath}, \hat{\boldsymbol{\omega}})^2 + \Delta(\hat{\imath})\,\Delta(\hat{\imath}, \hat{\boldsymbol{\omega}})} \frac{\Delta(\hat{\imath}, \hat{\boldsymbol{\omega}})^2}{1 - \frac{1}{\Delta(\hat{\imath}, \hat{\boldsymbol{\omega}})^2 + \Delta(\hat{\imath})\Delta(\hat{\imath}, \hat{\boldsymbol{\omega}})} \Delta(\hat{\imath}, \hat{\boldsymbol{\omega}})} \\ =\ & \Delta(\hat{\imath}, \hat{\boldsymbol{\omega}})\,\Delta(\hat{\imath}) \frac{\Delta(\hat{\imath}) - 1}{\Delta(\hat{\imath}, \hat{\boldsymbol{\omega}})^2 + \Delta(\hat{\imath}, \hat{\boldsymbol{\omega}})\,(\Delta(\hat{\imath}) - 1)} \\ >\ & 0. \end{aligned}$$

Proof of Proposition 2

Note that $p^t(i \mid \boldsymbol{\omega}_{1..n-1}) = 0$ implies $p^{t+1}(i \mid \boldsymbol{\omega}_{1..n-1}) = 0$. Thus, we have

$$
\begin{aligned}
& W\left(\mathbf{p}^{t+1}\right)-W\left(\mathbf{p}^{t}\right) \\
= & \lambda_n \sum_{\boldsymbol{\omega}} g\left(\boldsymbol{\omega}\right) \log\left(\frac{\sum_{i:p^t(i|\boldsymbol{\omega}_{1..n-1})>0} e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}_i^{t+1},\boldsymbol{\omega}_{1..n-1})}{\sum_{j\in A} e^{\frac{u(j,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}_i^{t},\boldsymbol{\omega}_{1..n-1})}\right) \\
= & \lambda_n \sum_{\boldsymbol{\omega}} g(\boldsymbol{\omega}) \log\left(\sum_{i:p^t(i|\boldsymbol{\omega}_{1..n-1})>0} \frac{e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}_i^{t},\boldsymbol{\omega}_{1..n-1})}{\sum_{j\in A} e^{\frac{u(j,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}_j^{t},\boldsymbol{\omega}_{1..n-1})} \frac{\Pi(\mathbf{p}_i^{t+1},\boldsymbol{\omega}_{1..n-1})}{\Pi(\mathbf{p}_i^{t},\boldsymbol{\omega}_{1..n-1})}\right) \\
\geq & \lambda_n \sum_{\boldsymbol{\omega}} g\left(\boldsymbol{\omega}\right) \sum_{i:p^t(i|\boldsymbol{\omega}_{1..n-1})>0} \frac{e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}_i^{t},\boldsymbol{\omega}_{1..n-1})}{\sum_{j\in A} e^{\frac{u(j,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}_j^{t},\boldsymbol{\omega}_{1..n-1})} \log\left(\frac{\Pi(\mathbf{p}_i^{t+1},\boldsymbol{\omega}_{1..n-1})}{\Pi(\mathbf{p}_i^{t},\boldsymbol{\omega}_{1..n-1})}\right) \\
= & \lambda_n \sum_{\boldsymbol{\omega}_{1..n-1}} g\left(\boldsymbol{\omega}_{1..n-1}\right) \sum_{i:p^t(i|\boldsymbol{\omega}_{1..n-1})>0} \log\left(\frac{\Pi(\mathbf{p}_i^{t+1},\boldsymbol{\omega}_{1..n-1})}{\Pi(\mathbf{p}_i^{t},\boldsymbol{\omega}_{1..n-1})}\right) \\
& \qquad \times \sum_{\omega_n} \frac{g\left(\omega_n \mid \boldsymbol{\omega}_{1..n-1}\right) e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}_i^{t},\boldsymbol{\omega}_{1..n-1})}{\sum_{j\in A} e^{\frac{u(j,\boldsymbol{\omega})}{\lambda_n}} \Pi(\mathbf{p}_j^{t},\boldsymbol{\omega}_{1..n-1})},
\end{aligned}
$$

which follows from Jensen's inequality. Collecting the term that equals $p^{t+1}\left(i \mid \boldsymbol{\omega}_{1..n-1}\right)$ in the last row gives

$$
\begin{aligned}
& W\left(\mathbf{p}^{t+1}\right)-W\left(\mathbf{p}^{t}\right) \\
= & \lambda_n \sum_{\boldsymbol{\omega}_{1..n-1}} g(\boldsymbol{\omega}_{1..n-1}) \sum_{i:p^t(i|\boldsymbol{\omega}_{1..n-1})>0} \left(\sum_{k=0}^{n-1} \frac{\lambda_{k+1}-\lambda_k}{\lambda_n} \log\frac{p^{t+1}\left(i \mid \boldsymbol{\omega}_{1..k}\right)}{p^{t}\left(i \mid \boldsymbol{\omega}_{1..k}\right)}\right) p^{t+1}(i \mid \boldsymbol{\omega}_{1..n-1}).
\end{aligned}
$$

Finally, we rearrange to get the desired inequality,

$$
\begin{aligned}
& W\left(\mathbf{p}^{t+1}\right)-W\left(\mathbf{p}^{t}\right) \\
\geq & \sum_{i:p^t(i|\boldsymbol{\omega}_{1..n-1})>0} \sum_{k=0}^{n-1} (\lambda_{k+1}-\lambda_k) \sum_{\boldsymbol{\omega}_{1..n-1}} p^{t+1}\left(i,\boldsymbol{\omega}_{1..n-1}\right) \log\left(\frac{p^{t+1}\left(i \mid \boldsymbol{\omega}_{1..k}\right)}{p^{t}\left(i \mid \boldsymbol{\omega}_{1..k}\right)}\right) \\
= & \sum_{k=0}^{n-1} (\lambda_{k+1}-\lambda_k) \sum_{\boldsymbol{\omega}_{1..k}} g(\boldsymbol{\omega}_{1..k}) \sum_{i:p^t(i|\boldsymbol{\omega}_{1..n-1})>0} p^{t+1}(i \mid \boldsymbol{\omega}_{1..k}) \log\left(\frac{p^{t+1}\left(i \mid \boldsymbol{\omega}_{1..k}\right)}{p^{t}\left(i \mid \boldsymbol{\omega}_{1..k}\right)}\right),
\end{aligned}
$$

which by definition of the Kullback-Leibler divergence equals the deserved expression.

Proof of Theorem 3

Because of the compactness of the domain, the set of accumulation points of the sequence $\{\mathbf{p}^t\}$ is nonempty. Denote this set by $\tilde{\mathbf{P}}$. Moreover, $\delta(\mathbf{p}^t)$ given by $\delta(\mathbf{p}^t) = W\left(\mathbf{p}^{t+1}\right) - W\left(\mathbf{p}^t\right)$ is a continuous extended real valued function in $\mathbf{p}^t$. Thus, for every accumulation point $\tilde{\mathbf{p}}$ of $\{\mathbf{p}^t\}$ and subsequence $\{\mathbf{p}^{t_r}\}$ that converges to $\tilde{\mathbf{p}}$, we have $\delta(\mathbf{p}^{t_r}) \to \delta(\tilde{\mathbf{p}})$. By Proposition 2, we have $\delta(\mathbf{p}^t) \geq 0$, and with $W\left(\mathbf{p}^t\right) \leq W^* < \infty$ (because $u\left(i, \boldsymbol{\omega}\right)$ is finite), we get $\delta(\mathbf{p}^t) \to 0$ for $t \to \infty$. We infer $\delta(\tilde{\mathbf{p}}) = 0$. This with Proposition 2 implies that $\tilde{\mathbf{p}}^{+1} = \tilde{\mathbf{p}}$, i.e., $\tilde{\mathbf{p}}$ satisfies (15). Thus, once $\delta(\mathbf{p}^t) = 0$, Step 2 of the algorithm found a point, say $\mathbf{p}^*$, that satisfies (15). This $\mathbf{p}^*$ constitutes a solution for the problem of Proposition 1 when specified on the restricted domain where all $p\left(i \mid \boldsymbol{\omega}_{1..n-1}\right)$ that take values zero in $\mathbf{p}^*$ are required to be zero, i.e., $\mathbf{p}^*$ in conjunction with (10) gives conditional choice probabilities $\left\{p^*\left(i \mid \boldsymbol{\omega}\right)\right\}_{i \in A, \boldsymbol{\omega} \in \boldsymbol{\Omega}}$ that solve

$$\max_{\{p(i|\boldsymbol{\omega})\}_{i \in A, \boldsymbol{\omega} \in \boldsymbol{\Omega}}} \sum_{i \in A} \sum_{\boldsymbol{\omega} \in \boldsymbol{\Omega}} u\left(i, \boldsymbol{\omega}\right) p\left(i \mid \boldsymbol{\omega}\right) g\left(\boldsymbol{\omega}\right) - \sum_{k=1}^{n} \lambda_k \cdot I_p\left(\Omega_k, A \mid \boldsymbol{\Omega}_{1..k-1}\right) \quad (31)$$

s. t.

$$p\left(i \mid \boldsymbol{\omega}\right) \geq 0 \quad \text{for all } i \in A \text{ and } \boldsymbol{\omega} \in \boldsymbol{\Omega}, \quad (32)$$

$$p\left(i \mid \boldsymbol{\omega}\right) = 0 \quad \text{for all } i \in A \text{ and } \boldsymbol{\omega} \in \boldsymbol{\Omega} \text{ such that } p^*\left(i \mid \boldsymbol{\omega}\right) = 0, \quad (33)$$

$$\sum_{i \in A} p\left(i \mid \boldsymbol{\omega}\right) = 1 \quad \text{for all } \boldsymbol{\omega} \in \boldsymbol{\Omega}. \quad (34)$$

In Step 3 of the algorithm, either (16) is confirmed and by Theorem 2 we have found an optimal $\mathbf{p}$; or the incremental change of $\mathbf{p}^*$ strictly increases the objective, as is show in the "necessity of (16)"-part in the proof of Theorem 2. Then, the outcome of Step 2 will give a $\mathbf{p}^{**}$ for which $W(\mathbf{p}^{**}) > W(\mathbf{p}^*)$. Thus, the zeros in $\mathbf{p}^{**}$ must be different from the zeros in $\mathbf{p}^*$ as otherwise $\mathbf{p}^*$ would not be a solution to (31)-(34). Since the possible combinations of zeros in $\mathbf{p}$ are finite, the algorithm terminates with optimal zeros in $\mathbf{p}$, satisfying (15) on the other entries, thus solving (31)-(34) for optimal zeros, thus solving the original problem.

C. Derivation of the Closed Solution of Example 1

Note that $p(\omega_c \mid \omega_a, \omega_b, i) = 1$ for all $\boldsymbol{\omega} \in \bar{\boldsymbol{\Omega}}$. As a result, conditions (14) simplify to

$$\frac{p\left(\omega_b \mid \omega_a, i\right) p\left(\omega_a \mid i\right)^{\frac{\lambda_a}{\lambda_b}}}{p\left(\omega_b \mid \omega_a, j\right) p\left(\omega_a \mid j\right)^{\frac{\lambda_a}{\lambda_b}}} = \frac{e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_b}}}{e^{\frac{u(j,\boldsymbol{\omega})}{\lambda_b}}}. \tag{35}$$

We now use (35) and consistency, $\sum_{i\in\{a,c,b\}} p\left(\omega_b \mid \omega_a, i\right) \cdot p\left(\omega_a \mid i\right) \cdot p\left(i\right) = g(\omega_a, \omega_b)$, to infer the closed-form solution for the choice probabilities for if $p(a), p(b), p(c) > 0$ (The results for $p(c) = 0$ are obtained through the algorithm in §6.2.). First, we express (35) for the states $\boldsymbol{\omega} = (0, v_b, 0)$ and $\boldsymbol{\omega} = (0, 0, v_c)$, respectively:

$$\frac{p\left(v_b \mid 0, a\right) p\left(0 \mid a\right)^{\frac{\lambda_a}{\lambda_b}}}{p\left(v_b \mid 0, b\right) p\left(0 \mid b\right)^{\frac{\lambda_a}{\lambda_b}}} = \frac{e^{\frac{0}{\lambda_a}}}{e^{\frac{v_b}{\lambda_b}}} = \frac{1}{e^{v_b/\lambda_b}} \tag{36}$$

$$\frac{p\left(v_b \mid 0, b\right) p\left(0 \mid b\right)^{\frac{\lambda_a}{\lambda_b}}}{p\left(v_b \mid 0, c\right) p\left(0 \mid c\right)^{\frac{\lambda_a}{\lambda_b}}} = \frac{e^{\frac{v_b}{\lambda_a}}}{e^{\frac{0}{\lambda_b}}} = e^{v_b/\lambda_b} \tag{37}$$

$$\frac{p\left(0 \mid 0, a\right) p\left(0 \mid a\right)^{\frac{\lambda_a}{\lambda_b}}}{p\left(0 \mid 0, b\right) p\left(0 \mid b\right)^{\frac{\lambda_a}{\lambda_b}}} = \frac{e^{\frac{0}{\lambda_a}}}{e^{\frac{0}{\lambda_b}}} = 1 \tag{38}$$

$$\frac{p\left(0 \mid 0, b\right) p\left(0 \mid b\right)^{\frac{\lambda_a}{\lambda_b}}}{p\left(0 \mid 0, c\right) p\left(0 \mid c\right)^{\frac{\lambda_a}{\lambda_b}}} = \frac{e^{\frac{0}{\lambda_a}}}{e^{\frac{v_c}{\lambda_b}}} = \frac{1}{e^{v_c/\lambda_b}} \tag{39}$$

For the state $\boldsymbol{\omega} = (v_a, 0, 0)$, using $p(0 \mid v_a, a) = p(0 \mid v_a, b) = 1$, equations (35) even simplify to

$$p(v_a \mid a) = p(v_a \mid b)e^{v_a/\lambda_a} \tag{40}$$

$$p(v_a \mid b) = p(v_a|c) \tag{41}$$

We also know $p(v_a \mid b) + p(0 \mid b) = 1$ and $p(v_a \mid c) + p(0 \mid c) = 1$, which with (41) imply $p(0 \mid b) = p(0 \mid c)$. Hence, we can also re-write (37) and (39) as

$$p(v_b \mid 0, b) = p(v_b \mid 0, c)e^{v_b/\lambda_b} \tag{42}$$

$$p(0 \mid 0, c) = p(0 \mid 0, b)e^{v_c/\lambda_b} \tag{43}$$

From (42) and (43), and using $p(v_b \mid 0, b) + p(0 \mid 0, b) = 1$ and $p(v_b \mid 0, c) + p(0 \mid 0, c) = 1$, we obtain

$$p(v_b \mid 0,b) = p(0 \mid 0,c) = \frac{e^{(v_b+v_c)/\lambda_b} - e^{v_b/\lambda_b}}{e^{(v_b+v_c)/\lambda_b} - 1} \tag{44}$$

$$p(v_b \mid 0,c) = p(0 \mid 0,b) = \frac{e^{v_c/\lambda_b} - 1}{e^{(v_b+v_c)/\lambda_b} - 1} \tag{45}$$

Similarly, from (36) and (38), using $p(v_b \mid 0,a) + p(0 \mid 0,a) = 1$ and $p(v_c \mid 0,a) + p(0 \mid 0,a) = 1$, we solve for

$$p(v_b \mid 0,a) = p(0 \mid 0,a) = \frac{e^{v_c/\lambda_b} - 1}{e^{v_b/\lambda_b} + e^{v_c/\lambda_b} - 2} \tag{46}$$

Noting that $p(0 \mid v_a, i) = 1$ and $p(v_b \mid v_a, i) = 0$ for all $i \in \{a,b,c\}$, we have found the posterior probabilities of the form $p(\cdot \mid \cdot, i)$, $i \in \{a,b,c\}$. Next, we proceed with the derivations of the posteriors of the form $p(\cdot \mid i)$ for all $i \in \{a,b,c\}$. We substitute $p(0 \mid 0,a)$ and $p(0 \mid 0,b)$ in (38), and obtain

$$p(0 \mid a) = p(0 \mid b) \left(\frac{e^{v_b/\lambda_b} + e^{v_c/\lambda_b} - 2}{e^{(v_b+v_c)/\lambda_b} - 1} \right)^{\frac{\lambda_b}{\lambda_a}} \tag{47}$$

For notational convenience, let us define

$$\xi := \left(\frac{e^{v_b/\lambda_b} + e^{v_c/\lambda_b} - 2}{e^{(v_b+v_c)/\lambda_b} - 1} \right)^{\frac{\lambda_b}{\lambda_a}} \tag{48}$$

Recognizing that $p(v_a \mid i) + p(0 \mid i) = 1$ for all $i \in \{a,b,c\}$, and substituting into (40) and (47) yields

$$p(v_a \mid a) = \frac{e^{v_a/\lambda_a}(1-\xi)}{e^{v_a/\lambda_a} - \xi} \tag{49}$$

$$p(0 \mid a) = \frac{\xi(e^{v_a/\lambda_a} - 1)}{e^{v_a/\lambda_a} - \xi} \tag{50}$$

$$p(v_a \mid b) = p(v_a \mid c) = \frac{1-\xi}{e^{v_a/\lambda_a} - \xi} \tag{51}$$

$$p(0 \mid b) = p(0 \mid c) = \frac{e^{v_a/\lambda_a} - 1}{e^{v_a/\lambda_a} - \xi} \tag{52}$$

As a result, we have completed the derivation of all posterior probabilities in our problem. These posteriors need to be consistent with the prior beliefs of the consumer. Specifically, for alternative a, we have

$$g(v_a, 0, 0) = p(v_a \mid a)p(a) + p(v_a \mid b)p(b) + p(v_a \mid c)p(c) \tag{53}$$

Substituting $p(v_a \mid a)$, $p(v_a \mid b)$ and $p(v_a \mid c)$, and using $p(a)+p(b)+p(c)=1$, we obtain the closed-form solution for $p(a)$ as

$$p(a) = \left(e^{v_a/\lambda_a}-1\right)^{-1}\left(g(v_a,0,0)\frac{e^{v_a/\lambda_a}-\xi}{1-\xi}-1\right) \tag{54}$$

where ξ is given in (48). In a similar fashion, and through more algebra, we can also solve for $p(b)$ and $p(c)$,

$$\begin{aligned} p(c) &= g(0,0,v_c)\frac{\left(e^{a/\lambda_a}-\xi\right)\left(e^{(v_b+v_c)/\lambda_b}-1\right)}{\left(e^{v_a/\lambda_a}-1\right)\left(e^{v_b/\lambda_b}-1\right)\left(e^{v_c/\lambda_b}-1\right)}-\frac{1}{\left(e^{v_c/\lambda_b}-1\right)} \\ &\quad -\left(\frac{\xi\left(e^{(v_b+v_c)/\lambda_b}-1\right)}{\left(e^{v_b/\lambda_b}+e^{v_c/\lambda_b}-2\right)\left(e^{v_c/\lambda_b}-1\right)}-\frac{1}{\left(e^{v_c/\lambda_b}-1\right)}\right)p(a) \qquad (55) \\ p(b) &= 1-p(a)-p(c). \qquad (56) \end{aligned}$$

D. Inference from Market Shares – Model and Steady State Specification

Let Ω denote the finite type space. The common prior distribution over distributions of types is denoted by $G \in \Delta(\Delta(\Omega))$. The type distributions initially deemed possible is given as $\Gamma_0 = supp(G) \subseteq \Delta(\Omega)$ and includes the true distribution $g^\star \in int(supp(G))$, which thus is assigned a positive probability $G(g^\star) > 0$. The common prior G induces an expected distribution of types at time 0, $\mu_0(\boldsymbol{\omega}) = \int_{g\in\Gamma_0} g(\boldsymbol{\omega})\, dG$.

Given μ_0, the agents learn $\boldsymbol{\omega}$ and choose in an rationally-inattentive fashion, yielding

$$p_1(i \mid \boldsymbol{\omega}) = \frac{e^{\frac{u(i,\boldsymbol{\omega})}{\lambda_n}} p_1(i)^{\frac{\lambda_1}{\lambda_n}} \prod_{k=1}^n p_1(i \mid \boldsymbol{\omega}_{1..k})^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}}{\sum_{j\in A} e^{\frac{u(j,\boldsymbol{\omega})}{\lambda_n}} p_1(j)^{\frac{\lambda_1}{\lambda_n}} \prod_{k=1}^n p_1(j \mid \boldsymbol{\omega}_{1..k})^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}}$$

where $p_1(j \mid \boldsymbol{\omega}_{1..k}) = \sum_{\boldsymbol{\omega}'_{k+1..n}} \mu_0\left(\boldsymbol{\omega}'_{k+1..n} \mid \boldsymbol{\omega}_{1..k}\right) p_1\left(j \mid \boldsymbol{\omega}'_{1..k}\boldsymbol{\omega}_{k+1..n}\right)$, i.e., they learn based on the belief μ_0. The realized observable partial type dependent market shares are created from conditional choice and true type distribution, $M_1(i \mid \boldsymbol{\omega}_{1..n-1}) = \sum_{\omega_n} g^\star(\omega_n \mid \boldsymbol{\omega}_{1..n-1})\, p_1(i \mid \boldsymbol{\omega})$. Now, the set of possible type distributions Γ_0 is refined to those compatible with the observed $M_1(i \mid \boldsymbol{\omega}_{1..n-1})$,

$$\Gamma_1=\left\{g\in\Gamma_0|M_1(i|\boldsymbol{\omega}_{1\cdot\cdot n-1})=\sum_{\omega_n} p_1(i|\boldsymbol{\omega})\, g(\omega_n|\boldsymbol{\omega}_{1\cdot\cdot n-1}) \text{ for all } \boldsymbol{\omega}_{1\cdot\cdot n-1}\in\boldsymbol{\Omega}_{1\cdot\cdot n-1}\right\}$$

In general, given the set of possible distributions Γ_t, the prior μ_t is determined as

$$\mu_t(\boldsymbol{\omega}) = \frac{1}{G(\Gamma_t)} \int_{g\in\Gamma_t} g(\boldsymbol{\omega})\, dG \tag{57}$$

Again, (10) pins down the conditional choice probabilities, which induce partial type dependent market share

$$M_t(i \mid \boldsymbol{\omega}_{1\cdot\cdot n-1}) = \sum_{\omega_n} g^\star(\omega_n \mid \boldsymbol{\omega}_{1\cdot\cdot n-1})\, p_t(i \mid \boldsymbol{\omega}) \tag{58}$$

This is only compatible with some type distributions $g \in \Gamma_t$ giving the new set of beliefs Γ_{t+1},

$$\Gamma_{t+1}=\left\{g\in\Gamma_t \mid M_t(i,\boldsymbol{\omega}_{1\cdot\cdot n-1}) = \sum_{\omega_n} g(\boldsymbol{\omega})\, p_t(i \mid \boldsymbol{\omega}) \text{ for all } \boldsymbol{\omega}_{1\cdot\cdot n-1}\in\boldsymbol{\Omega}_{1\cdot\cdot n-1}\right\}$$

Steady-state is reached if $\Gamma_{t+1} = \Gamma_t$. In a finite number of steps, this procedure reaches the steady-state. The argument is exactly the same as for Caplin et al. (2016b, Lemma 1), and is omitted.

Next, we argue that $\Gamma_{t+1} = \Gamma_t$ implies $M_t(i,\boldsymbol{\omega}_{1\cdot\cdot n-1}) = p_t(i,\boldsymbol{\omega}_{1\cdot\cdot n-1})$. This also means $g^\star(\boldsymbol{\omega}_{1\cdot\cdot n-1}) = g_t(\boldsymbol{\omega}_{1\cdot\cdot n-1})$ and $M_t(i \mid \boldsymbol{\omega}_{1\cdot\cdot k}) = p_t(i \mid \boldsymbol{\omega}_{1\cdot\cdot k})$ for all $k = 1,\ldots,n-1$. From this follows (19). Note that $\Gamma_{t+1} = \Gamma_t$ means $M_t(i,\boldsymbol{\omega}_{1\cdot\cdot n-1}) = \sum_{\omega_n} g(\boldsymbol{\omega})\, p_t(i \mid \boldsymbol{\omega})$ for all $g \in \Gamma_t$ and $\boldsymbol{\omega}_{1\cdot\cdot n-1} \in \Omega_{1\cdot\cdot n-1}$. Plugging this into

$$\begin{aligned} p_t(i,\boldsymbol{\omega}_{1\cdot\cdot n-1}) &= \sum_{\omega_n} \mu_t(\omega_n,\boldsymbol{\omega}_{1\cdot\cdot n-1})\, p_t(i \mid \boldsymbol{\omega}) \\ &\overset{(57)}{=} \sum_{\omega_n} \left(\frac{1}{G(\Gamma_t)} \sum_{g\in\Gamma_t} g(\boldsymbol{\omega})\, G(g) \right) p_t(i \mid \boldsymbol{\omega}) \\ &= \frac{1}{G(\Gamma_t)} \sum_{g\in\Gamma_t} G(g) \sum_{\omega_n} g(\boldsymbol{\omega})\, p_t(i \mid \boldsymbol{\omega}) \end{aligned}$$

gives the desired result:

$$p_t(i,\boldsymbol{\omega}_{1\cdot\cdot n-1}) = \frac{1}{G(\Gamma_t)} \sum_{g\in\Gamma_t} G(g)\, M_t(i,\boldsymbol{\omega}_{1\cdot\cdot n-1}) = M_t(i,\boldsymbol{\omega}_{1\cdot\cdot n-1}) \tag{59}$$

Finally, we also establish that reducing the scope to those alternatives that are chosen in steady-state allows us to consider them chosen as if rationally-inattentive consumer's had chosen them knowing the true prior: Suppose we reached a steady-state with non-eliminated type distributions $\bar{\Gamma}$ and where market share is nonzero for alternatives $\bar{A} = \{i \mid M(i) > 0\}$. With

$$\begin{aligned}
& p_t\left(i \mid \boldsymbol{\omega}_{1..n-1}\right) \\
\overset{(59)}{=}\ & M_t\left(i \mid \boldsymbol{\omega}_{1..n-1}\right) \\
\overset{(58)}{=}\ & \sum_{\omega_n} g^\star\left(\omega_n \mid \boldsymbol{\omega}_{1..n-1}\right) p_t\left(i \mid \boldsymbol{\omega}\right) \\
\overset{(10)}{=}\ & \sum_{\omega_n} g^\star\left(\omega_n \mid \boldsymbol{\omega}_{1..n-1}\right) \frac{e^{\frac{u\left(i,\boldsymbol{\omega}_{1..n-1}\omega_n\right)}{\lambda_n}} \prod_{k=0}^{n-1} p_t\left(i \mid \boldsymbol{\omega}_{1..k}\right)^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}}{\sum_{j\in A} e^{\frac{u\left(j,\boldsymbol{\omega}_{1..n-1}\omega_n\right)}{\lambda_n}} \prod_{k=0}^{n-1} p_t\left(j \mid \boldsymbol{\omega}_{1..k}\right)^{\frac{\lambda_{k+1}-\lambda_k}{\lambda_n}}},
\end{aligned}$$

it follows that $p_t\left(i \mid \boldsymbol{\omega}_{1..n-1}\right)$ satisfy the necessary and sufficient conditions for prior $g^\star$.

References

Ackerberg DA (2003) Advertising, learning, and consumer choice in experience good markets: an empirical examination. *International Economic Review* 44(3):1007–1040.

Alptekinoğlu A, Semple JH (2015) The exponomial choice model: A new alternative for assortment and price optimization. *Operations Research* 64(1):79–93.

Anderson SP, de Palma A, Thisse JF (1992) *Discrete Choice Theory of Product Differentiation* (Cambridge: MIT Press).

Blanchet JH, Gallego G, Goyal V (2016) A markov chain approximation to choice modeling. *Operations Research* 64(4):886–905.

Boyacı T, Akçay Y (2018) Pricing when customers have limited attention. *Management Science* 64(7):2995–3014.

Branco F, Sun M, Villas-Boas JM (2012) Optimal search for product information. *Management Science* 58(11):2037–2056.

Caplin A, Dean M (2013) The behavioral implications of rational inattention with Shannon entropy. Technical report, NBER.

Caplin A, Dean M (2015) Revealed preference, rational inattention, and costly information acquisition. *American Economic Review* 105(7):2183–2203.

Caplin A, Dean M, Leahy J (2016a) Rational inattention, optimal consideration sets and stochastic choice. Technical report, Working paper.

Caplin A, Dean M, Leahy J (2017) Rationally inattentive behavior: Characterizing and generalizing shannon entropy. Technical Report No. 23652, NBER Working Paper.

Caplin A, Leahy J, Matějka F (2016b) Rational inattention and inference from market share data. Technical report, Working Paper.

Ching AT, Erdem T, Keane MP (2014) A simple method to estimate the roles of learning, inventories and category consideration in consumer choice. *Journal of Choice Modelling* 13:60–72.

Cover T (1984) An algorithm for maximizing expected log investment return. *IEEE Transactions on Information Theory* 30(2):369–373.

Cover TM, Thomas JA (2006) *Elements of Information Theory* (New York: Wiley), 2nd edition, ISBN 0471241954.

Davis AM, Katok E, Kwasnica AM (2014) Should sellers prefer auctions? a laboratory comparison of auctions and sequential mechanisms. *Management Science* 60(4):990–1008.

De los Santos B, Hortaçsu A, Wildenbeest M (2012) Testing models of consumer search using data on web browsing and purchasing behavior. *The American Economic Review* 102(6):2955–2980.

Dong L, Kouvelis P, Tian Z (2009) Dynamic pricing and inventory control of substitute products. *Manufacturing & Service Operations Management* 11(2):317–339.

Erdem T, Keane MP (1996) Decision-making under uncertainty: Capturing dynamic brand choice processes in turbulent consumer goods markets. *Marketing Science* 15(1):1–20.

Gabaix X (2014) A sparsity-based model of bounded rationality. *The Quarterly Journal of Economics* 129(4):1661–1710.

Hamilton RW, Thompson DV (2007) Is there a substitute for direct experience? Comparing consumers' preferences after direct and indirect product experiences. *Journal of Consumer Research* 34(4):546–555.

Hanson W, Martin K (1996) Optimizing multinomial logit profit functions. *Management Science* 42(7):992–1003.

Ke TT, Shen ZJM, Villas-Boas JM (2016) Search for information on multiple products. *Management Science.* 62(12):3576–3603.

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

Luce RD, Suppes P (1965) *Preference, utility, and subjective probability*, volume 3 of *Handbook of Mathematical Psychology* (New York: Wiley).

Maćkowiak B, Wiederholt M (2009) Optimal sticky prices under rational inattention. *American Economic Review* 99(3):769–803.

Maćkowiak B, Wiederholt M (2015) Business cycle dynamics under rational inattention. *The Review of Economic Studies* 82(4):1502 – 1532.

Manzini P, Mariotti M (2014) Stochastic choice and consideration sets. *Econometrica* 82(3):1153–1176.

Matějka F (2015) Rigid pricing and rationally inattentive consumer. *Journal of Economic Theory* 158:656–678.

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

Matějka F, McKay A (2015) Rational inattention to discrete choices: A new foundation for the multinomial logit model. *American Economic Review* 105(1):272–298.

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

McKelvey RD, Palfrey TR (1995) Quantal response equilibria for normal form games. *Games and economic behavior* 10(1):6–38.

Narayanan S, Manchanda P, Chintagunta PK (2005) Temporal differences in the role of marketing communication in new product categories. *Journal of Marketing Research* 42(3):278–290.

Reis R (2006) Inattentive producers. *The Review of Economic Studies* 73(3):793–821.

Roberts J, Lattin JM (1991) Development and testing of a model of consideration set composition. *Journal of Marketing Research* 429–440.

Hoch SJ, Ha YW (1986) Consumer learning: Advertising and the ambiguity of product experience. *Journal of Consumer Research* 13(2):221–233.

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

Simon HA (1955) A behavioral model of rational choice. *The Quarterly Journal of Economics* 69(1):99–118.

Simon HA (1979) Information processing models of cognition. *Annual review of psychology* 30(1):363–396.

Sims CA (1998) Stickiness. *Carnegie-Rochester Conference Series on Public Policy* 49(1):317–356.

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

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

Srikanth J, Rusmevichientong P (2017) A nonparametric joint assortment and price choice model. *Management Science* 63(9):3128–3145.

Stigler GJ (1961) The economics of information. *Journal of Political Economy* 69(3):213–225.

Stoneman P (1981) Intra-firm diffusion, bayesian learning and profitability. *The Economic Journal* 91(362):375–388.

Talluri K, van Ryzin G (2004) Revenue management under a general discrete choice model of consumer behavior. *Management Science* 50(1):15–33.

Todd PM, Gigerenzer G (2000) Précis of simple heuristics that make us smart. *Behavioral and brain sciences* 23(05):727–741.

Tutino A (2013) Rationally inattentive consumption choices. *Review of Economic Dynamics* 16(3):421 – 439.

van Ryzin G, Mahajan S (1999) On the relationship between inventory costs and variety benefits in retailassortments. *Management Science* 45(11):1496–1509.

Verrecchia RE (1982) Information acquisition in a noisy rational expectations economy. *Econometrica* 50(6):1415–1430.

Weitzman ML (1979) Optimal search for the best alternative. *Econometrica* 47(3):641–654.

Wierenga B (2008) *Handbook of marketing decision models* (New York: Springer).

Zhang D, Adelman D (2009) An approximate dynamic programming approach to network revenue management with customer choice. *Transportation Science* 43(3):381–394.