Discrete Choice and Rational Inattention: a General Equivalence Result [*]

Mogens Fosgerau[†] Emerson Melo[‡] André de Palma[§] Matthew Shum[¶]

February 7, 2020

Abstract

This paper establishes a general equivalence between discrete choice and rational inattention models. Matejka and McKay (2015) showed that when information costs are modelled using the Shannon entropy, the choice probabilities in the rational inattention (RI) model take the multinomial logit form. We show that, for one given prior over states, RI choice probabilities may take the form of any additive random utility discrete choice model (ARUM) when the information cost is a Bregman information, belonging to a class defined in this paper. The prior information of the rationally inattentive agent is summarized in a constant vector of utilities in the corresponding ARUM. We illustrate our results utilizing the nested logit, an empirically relevant discrete choice model.

JEL codes: D03, C25, D81, E03
Keywords: rational inattention; discrete choice models; random utility; convex analysis; generalized entropy; Bregman Information

[*] First draft: December 22, 2016. We thank Bob Becker, Marcus Berliant, Mark Dean, Federico Echenique, Juan Carlos Escanciano, Paul Heidhues, Joonhwi Joo, Filip Matejka, Paulo Natenzon, Henrique de Oliveira, Antonio Rangel, Ryan Webb, Michael Woodford, and participants at the 2018 AEA meetings (Philadelphia) for useful comments. Alejandro Robinson Cortes, Lucie Letrouit, and Julien Monardo provided research assistance. We are grateful to the editor and three referees for their very valuable comments and suggestions. Mogens Fosgerau has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 740369).

[†] University of Copenhagen; mogens.fosgerau@econ.ku.dk

[‡] Indiana University; emelo@iu.edu

[§] ENS, University Paris-Saclay, CREST; andre.depalma@ens-paris-saclay.fr.

[¶] California Institute of Technology; mshum@caltech.edu

1 Introduction

In many situations where agents make decisions under uncertainty, information acquisition is costly (involving pecuniary, time, or psychological costs); therefore, agents may rationally choose to remain imperfectly informed about the available options. This idea underlies the theory of Rational Inattention (RI), which has become an important paradigm for modeling boundedly rational behavior in many areas of economics (Sims, 2003, 2010). In this paper, our main contribution is to establish a general equivalence between additive random utility discrete choice and RI models. Matějka and McKay (2015) showed that when information costs are modelled using the Shannon mutual information between actions and states, the resulting choice probabilities in the RI model take the familiar multinomial logit form, leading to the "RI-logit" model, as we will refer to it below. This is a very appealing result, providing a microfoundation as well as alternative interpretation for the multinomial logit model.

However, the RI-logit model has the "independence of irrelevant alternatives" (IIA) property, which states that, in a given state, the ratio of the choice probabilities of two alternatives does not depend on the utility of a third (irrelevant) alternative.[1] In many empirical contexts, the IIA property implies restrictive and unrealistic substitution patterns among the choice options, as illustrated in the following example.

Example 1 *Consider a rationally inattentive consumer facing a choice between pineapple (good 1), mango (good 2), and cheesecake (good 3). A priori, the consumer does not know the value associated with each good but has fixed prior beliefs about the possible realizations of the valuation vector* $\mathbf{V}$*. Assume that* $\mathbf{V}$ *has four equally likely possible states:*

$$\left(\mathbf{v}^1, \mathbf{v}^2, \mathbf{v}^3, \mathbf{v}^4\right) = \begin{pmatrix} 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0.1 \end{pmatrix},$$

[1] In the context of the RI model, we interpret IIA as a comparison across states, holding the DM's prior fixed. For further details about IIA we refer the reader to Maddala (1986, 3.2) and Anderson et al. (1992).

and assume that we observe corresponding choice probabilities $p\left(\mathbf{v}^1\right) = (0.41, 0.41, 0.18)$ *and* $p\left(\mathbf{v}^2\right) = (0.46, 0.37, 0.17)$ *for the first two states.*[2] *These probabilities reflect a situation where an increase (from* 0 *to* 0.1*) in the value of good 1, the pineapple, causes consumers to substitute disproportionately from the mango rather than from the cheesecake. However, such outcomes violate the IIA property, as the choice probability for mango decreases by 10 % while the choice probability for cheesecake decreases only by 4 %; hence, they cannot arise from the RI-logit model.* ■

The root of the problem in the previous example is that the Shannon mutual information embodies an important and strong assumption of *symmetry*: the Shannon entropy is invariant to permutation in its arguments; therefore reordering the choice options does not affect the information cost. This makes the cost of processing information *context independent* (Hobson, 1969) and hence it cannot take into account that some choice options are more similar than other options.[3]

In this paper we introduce a new class of generalized entropies that allows us to define cost functions that embody information related to the identity of alternatives. Formally, our generalized entropy is defined as the negative convex conjugate of the surplus function of an Additive Random Utility Model (ARUM), and allows patterns such as those in the example above to be accommodated in a RI model. The new generalized entropies are not required to be symmetric. This contributes to making RI models empirically relevant. In fact, we show that, depending on the choice of information cost, a RI model can yield the same choice probability system as any ARUM; this includes specifications such as nested logit, multinomial probit, and so on, that are often employed in empirical work.

Based on our definition of generalized entropy, we introduce a class of *generalized RI models*. In particular, we define a general class of information cost functions where the Shannon entropy is replaced by our generalized entropy. This

[2] The choice probabilities here are generated, not by a RI-logit model, but by a RI-nested logit model, introduced in this paper.

[3] This example considers how choices vary across different states of the world. This is distinct from the exercises in Matějka and McKay (2015), who consider how choices vary as priors or choice sets change. For empirical analysis, the change in choices across states is often most relevant: for instance, in demand analysis, researchers wish to uncover how consumer choices depend on changes in product prices, which can be considered as a change from one state to another.

generalizes the Shannon mutual information since this cost function arises when the ARUM is the multinomial logit model. Because of this connection with the ARUM class we label our general RI model as "RI-ARUM". As we will show, an RI-ARUM model exists corresponding to any ARUM, implying that rationally inattentive behavior can lead to choice probabilities that violate the IIA property, as in the above example.

Related literature. Besides the papers already mentioned above, the main equivalence result in this paper is related to several strands of literature. This paper contributes to the growing literature on rational inattention with more general cost functions. Hébert and Woodford (2017) provide a foundation for the rational inattention model based on a dynamic information accumulation process. In particular, they introduce a class of "neighborhood-cost" functions, which allows them to reflect varying similarity of states to one another. Morris and Yang (2016) use ideas from global games to develop a rational inattention framework, in which it is more difficult for players to distinguish between nearby states. Our results complement these, but instead of allowing cost functions to reflect that some states are more similar than others, we introduce cost functions that may reflect that some choice *options* are more similar than others which, as illustrated in the example above, is relevant for the empirical implications of the rational inattention model.

Caplin et al. (2017) and Frankel and Kamenica (2018) study properties that may be required of information cost functions. We propose to use Bregman information (Banerjee et al., 2005) based on generalized entropy.

Our results also relate to the literature on perturbed utility models. Anderson et al. (1988) derived the representative consumer model underlying the logit model, showing that the direct utility has an entropy form. This observation was generalized by Hofbauer and Sandholm (2002), who showed that the choice probabilities generated by any ARUM can be derived from a deterministic model based on payoff perturbations that depend nonlinearly on the vector of choice probabilities. Fosgerau and McFadden (2012) provide a foundation for applications of consumer theory to perturbed utility problems with nonlinear budget constraints. Fudenberg et al. (2015) provide an axiomatic characterization of a class of perturbed random utility models. Allen and Rehbeck (2019) consider identification. Fosgerau et al. (2019) construct a class of inverse demand models that are useful for estimating demand for differentiated products using Berry's (1994) method. We contribute to that literature in two ways. First, we provide an explicit characterization of the perturbation term corresponding to general ARUM. Second, our equivalence result allows us to interpret the class of perturbed random utility models in terms of RI arguments.

Finally, the rational inattention framework has also inspired some recent empirical work; see Caplin et al. (2016), Joo (2019), Brown and Jeon (2019) and Porcher (2019). These papers primarily utilize the Shannon/multinomial logit framework. The results in this paper may enable researchers to apply rational inattention models far more general than the multinomial logit model, as they show that choice behavior emerging from *any* ARUM model may emerge from rationally inattentive behavior.

Layout. Section 2 introduces the rational inattention model. Section 3 introduces the ARUM framework, and uses convex analysis to generate some insights into the fundamental structure of these models. Using this structure, we introduce a class of generalized entropies and present a few key results. Section 4 shows how generalized entropy can be used to define the information cost in the rational inattention model, leading to the class of RI-ARUM models. Then we present the key result from this paper, which establishes the equivalence between choice probabilities emerging from the discrete choice model, and those emerging from RI-ARUM models. Section 5 discusses the specific case of the nested logit model, which has proven useful in many empirical models. We show how rationally inattentive behavior can generate choice probabilities with substitution patterns that violate IIA, as in the above example. Two examples demonstrate some properties of the RI-nested logit. Section 6 concludes. All proofs are in the Appendix.

Notation: Throughout this paper, for vectors $\mathbf{a}$ and $\mathbf{b}$, $\mathbf{a} \cdot \mathbf{b}$ denotes the vector scalar product $\sum_i a_i b_i$, such that, e.g., $\mathbf{1} \cdot \mathbf{q} = \sum_i q_i$. A univariate function applied to a vector is understood as coordinate-wise application of the function, e.g., $e^{\mathbf{q}} = (e^{q_1}, ..., e^{q_N})$. Consequently, $a + \mathbf{q} = (a + q_1, ..., a + q_J)$ for scalar a. The gradient with respect to a vector $\mathbf{v}$ is $\nabla_{\mathbf{v}}$; e.g., for $\mathbf{v} = (v_1, ..., v_N)$, $\nabla_{\mathbf{v}} W(\mathbf{v}) = \left(\frac{\partial W(\mathbf{v})}{\partial v_1}, ..., \frac{\partial W(\mathbf{v})}{\partial v_N}\right)$. The unit simplex in $\mathbb{R}^N$ is Δ.

2 Rational inattention

We introduce the rational inattention model. The decision maker is presented with a group of N options, from which he must choose one. Each option has an associated payoff $\mathbf{v} = (v_1, ..., v_N)$, but the vector of payoffs is *unobserved* by the decision-maker (DM). Instead, the DM considers the payoff vector $\mathbf{V}$ to be random, taking values in a set $\mathcal{V} \subset \mathbb{R}^N$; for simplicity, we take $\mathcal{V}$ to be finite. The DM possesses some prior knowledge about the available options, given by a probability measure μ, where $\mu(\mathbf{v}) = \mathbb{P}(\mathbf{V} = \mathbf{v}) > 0$ for all $\mathbf{v} \in \mathcal{V}$.

The DM's choice is an action $i \in \{1, ..., N\}$ and we write $p_i(\mathbf{v})$ as shorthand for the conditional probability that the action is i in state $\mathbf{V} = \mathbf{v}$. The payoff resulting from action i is V_i . The vector of choice probabilities conditional on $\mathbf{V} = \mathbf{v}$ is then $\mathbf{p}(\mathbf{v}) = (p_1(\mathbf{v}), \ldots, p_N(\mathbf{v}))$, and $\mathbf{p}(\cdot) = \{\mathbf{p}(\mathbf{v})\}_{\mathbf{v} \in \mathcal{V}}$ is the collection of conditional probabilities. Given the conditional probabilities $\mathbf{p}(\cdot)$ and the prior μ, it is convenient to have notation for the unconditional choice probabilities and we let $p_i^0 = \mathbb{E}p_i(\mathbf{V}) = \sum_{\mathbf{v} \in \mathcal{V}} p_i(\mathbf{v}) \mu(\mathbf{v})$ and $\mathbf{p}^0 = (p_1^0, \ldots, p_N^0)$.

The problem of the rationally inattentive DM is to choose the conditional distribution $\mathbf{p}(\cdot)$, balancing the expected payoff against the cost of information. The DM's strategy is a solution to the following problem:

$$\max_{\mathbf{p}(\cdot)} \{\mathbb{E}(\mathbf{V} \cdot \mathbf{p}(\mathbf{V})) - \text{Information Cost}\} . \tag{1}$$

2.1 The RI-logit model: the Matějka and McKay (2015) result

The key element in program (1) is the choice of the information cost function. In specifying information costs, researchers have used concepts from information theory (Cover and Thomas, 2006). Specifically, much of the existing literature (Sims, 2003; Matějka and McKay, 2015) has utilized the mutual Shannon information between payoffs and actions to measure the information costs[4]; that is, letting $\Omega(\mathbf{q}) = -\mathbf{q} \cdot \log \mathbf{q}$ denote the Shannon entropy, the information cost is specified

[4] More formally, Matějka and McKay (2015) study the problem where agents first choose an information structure (mapping from state of the world to information signals) and then, based on signals, choose optimal actions.

as

$$\begin{aligned}\kappa(\mathbf{p}(\cdot),\mu) &\equiv \Omega(\mathbb{E}(\mathbf{p}(\mathbf{V}))) - \mathbb{E}(\Omega(\mathbf{p}(\mathbf{V}))) \\ &= -\sum_{i=1}^{N} p_i^0 \log p_i^0 + \sum_{\mathbf{v}\in\mathcal{V}} \left(\sum_{i=1}^{N} p_i(\mathbf{v}) \log p_i(\mathbf{v}) \right) \mu(\mathbf{v}).\end{aligned} \tag{2}$$

Plugging this into (1), the rationally inattentive DM chooses the system of conditional probabilities $\mathbf{p}(\cdot)$ to optimize[5]

$$\max_{\mathbf{p}(\cdot)} \{\mathbb{E}(\mathbf{V}\cdot\mathbf{p}(\mathbf{V})) - \kappa(\mathbf{p}(\cdot),\mu)\} \tag{3}$$
$$= \max_{\mathbf{p}(\cdot)} \left\{ \sum_{\mathbf{v}\in\mathcal{V}} \left\{ \mathbf{p}(\mathbf{v}) \cdot [\mathbf{v} + \log \mathbf{p}^0] - \mathbf{p}(\mathbf{v}) \cdot \log \mathbf{p}(\mathbf{v}) \right\} \mu(\mathbf{v}) \right\}$$

subject to

$$p_i(\mathbf{v}) \geq 0 \text{ for all } i, \quad \sum_{i=1}^{N} p_i(\mathbf{v}) = 1. \tag{4}$$

Solving this, the DM finds conditional choice probabilities

$$p_i(\mathbf{v}) = \frac{p_i^0 e^{v_i}}{\sum_{j=1}^{N} p_j^0 e^{v_j}} \quad \text{for } i = 1, \ldots, N, \tag{5}$$

that satisfy $p_i^0 = \mathbb{E}p_i(\mathbf{V})$. We may rewrite (5) as

$$p_i(\mathbf{v}) = \frac{e^{v_i + \log p_i^0}}{\sum_{j=1}^{N} e^{v_j + \log p_j^0}} = \frac{e^{\tilde{v}_i}}{\sum_{j=1}^{N} e^{\tilde{v}_j}}, \tag{6}$$

where $\tilde{v}_i = v_i + \log p_i^0$. This may be recognized as a multinomial logit model in which the payoff vector $\mathbf{v}$ is shifted by $\log \mathbf{p}^0$. Remarkably, the influence of the prior information μ is completely captured by this shift vector $\log \mathbf{p}^0$.

Below, we show that this equivalence between the rational inattention model and the logit discrete choice model can be extended to the entire class of additive

[5] Our presentation of the rational inattention paradigm here follows Sims (2003, 2010), in which agents are modelled as choosing directly their conditional choice probabilities $\{\mathbf{p}(\mathbf{v})\}_{\mathbf{v}\in\mathcal{V}}$, taking the prior distribution $\mu(\mathbf{v})$ as given.

random utility discrete choice models, by suitably generalizing the information cost function $\Omega()$. We will also find that a location shift vector that completely captures the influence of the prior information. Before turning to these results, we briefly review some properties of the ARUM class and establish some new results that are useful for working with generalized entropies.

3 Random utility models and generalized entropy

Consider a DM making a utility maximizing discrete choice among a set of $i = 1, \dots, N$ options. The utility of option i is

$$u_i = v_i + \epsilon_i, \tag{7}$$

where $\mathbf{v} = (v_1, \dots, v_N)$ is deterministic and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_N)$ is a vector of random utility shocks. This is the classic ARUM framework pioneered by McFadden (1978). Our presentation of the ARUM framework here will emphasize convex-analytic properties which will be important in drawing connections with the rational inattention model in what follows.

Assumption 1 *The random vector $\boldsymbol{\epsilon}$ follows a joint distribution with finite means that is absolutely continuous, independent of $\mathbf{v}$, and fully supported on $\mathbb{R}^N$.*

Assumption 1 leaves the distribution of the ϵ's unspecified, thus allowing for a wide range of choice probability systems far beyond the often used logit model. Importantly, it accommodates arbitrary correlation in the ϵ_i's across options, which is reasonable and realistic in applications.

The DM then has choice probabilities

$$q_i(\mathbf{v}) \equiv \mathbb{P}\left(v_i + \epsilon_i = \max_j [v_j + \epsilon_j]\right), i = 1, ..., N.$$

An important concept in this paper is the *surplus function* of the discrete choice model (so named by McFadden, 1981), defined as

$$W(\mathbf{v}) = \mathbb{E}_{\boldsymbol{\epsilon}}(\max_j [v_j + \epsilon_j]). \tag{8}$$

Under Assumption 1, $W(\mathbf{v})$ is convex and differentiable and the choice probabilities coincide with the derivatives of $W(\mathbf{v})$:[6]

$$\frac{\partial W(\mathbf{v})}{\partial v_i} = q_i(\mathbf{v}) \text{ for } i = 1, \ldots, N$$

or, using vector notation, $\mathbf{q}(\mathbf{v}) = \nabla W(\mathbf{v})$. This is the Williams-Daly-Zachary theorem, famous in the discrete choice literature (McFadden, 1978, 1981).

From the differentiability of W and the Williams-Daly-Zachary theorem it follows that the choice probabilities emerging from any random utility discrete choice model can be expressed in closed-form as[7]:

$$q_i(\mathbf{v}) = \frac{T_i\left(e^{\mathbf{v}}\right)}{\sum_{j=1}^{N} T_j\left(e^{\mathbf{v}}\right)} \quad \text{for } i = 1, \ldots, N, \tag{9}$$

where the vector-valued function $\mathbf{T}(\cdot) = (T_1(\cdot), ..., T_N(\cdot)) : \mathbb{R}_+^N \mapsto \mathbb{R}_+^N$ is defined as the gradient of the exponentiated surplus, i.e.

$$\mathbf{T}(e^{\mathbf{v}}) = \nabla_{\mathbf{v}} \left(e^{W(\mathbf{v})}\right). \tag{10}$$

For the specific case of multinomial logit, the ϵ_i's are i.i.d. across options i with a type 1 extreme value distribution, the surplus function is $W(\mathbf{v}) = \log\left(\sum_{i=1}^{N} e^{v_i}\right)$, implying that $T_i(e^{\mathbf{v}}) = e^{v_i}$. Thus, Eq. (9) becomes the familiar multinomial logit choice formula: $q_i(\mathbf{v}) = e^{v_i} / \sum_j e^{v_j}$.

Based on (9), we may refer to $\mathbf{T}$ as the scaled demand mapping. We will use the inverse of the scaled demand mapping to construct an information cost. The inverse must allow the zero demands that arise in the rational inattention model. Existence of such an inverse is established by the following proposition.

Proposition 2 (Invertibility) *Let Assumption 1 hold. Then the function* $\mathbf{T}(\cdot)$ *has a continuous extension to* $\mathbf{T} : \mathbb{R}_{+0}^N \to \mathbb{R}_{+0}^N$ *that is surjective, injective and hence*

[6] The convexity of $W(\cdot)$ follows from the convexity of the max function. Differentiability follows from the absolute continuity of $\boldsymbol{\epsilon}$. See Shi et al. (2018), Chiong and Shum (2019), and Melo et al. (2019) for semiparametric econometric approaches based on these convex-analytic properties of discrete-choice models.

[7] By direct differentiation of $e^{W(\mathbf{v})}$, and applying the Williams-Daly-Zachary theorem, we have $q_i(\mathbf{v}) = T_i(e^{\mathbf{v}})/e^{W(\mathbf{v})}$ for all i. Imposing $\sum_i q_i(\mathbf{v}) = 1$ we have $\sum_i T_i(e^{\mathbf{v}}) = e^{W(\mathbf{v})}$.

globally invertible. Moreover, the function $\mathbf{S}(\cdot)$ *defined as* $\mathbf{S}(\cdot) = \mathbf{T}^{-1}(\cdot)$ *satisfies* $S_i(\mathbf{q}) = 0$ *iff* $q_i = 0$ *for* $i = 1, \ldots, N$.

In what follows we refer to $\mathbf{S}$ as the *inverse scaled demand* mapping. For any discrete choice model, there is a close relationship between the corresponding inverse scaled demand ($\mathbf{S}$) and surplus (W) functions. They are related in terms of *convex conjugate duality*. Since the social surplus function W for any ARUM is convex, we know that there exists a convex conjugate function W^* satisfying the problem[8]

$$W(\mathbf{v}) = \max_{\mathbf{q} \in \Delta} \{\mathbf{q} \cdot \mathbf{v} - W^*(\mathbf{q})\} \tag{11}$$

where the maximum on the right-hand side is attained at $\mathbf{q}(\mathbf{v}) = \nabla W(\mathbf{v})$.

The next proposition establishes a specific structure of the surplus function W and its convex conjugate W^*.[9]

Proposition 3 (Generalized entropy functions) *Consider an ARUM discrete choice model satisfying Assumption 1, with surplus function* $W(\cdot)$. *Then*

(i) *The surplus function* $W(\mathbf{v})$ *is equal to*

$$W(\mathbf{v}) = \log\left(\sum_{i=1}^{N} T_i(e^{\mathbf{v}})\right) \tag{12}$$

for the vector-valued function $\mathbf{T}$ *as defined in Eq. (10).*

(ii) *The convex conjugate of the surplus function* $W(\mathbf{v})$ *is*

$$W^*(\mathbf{q}) = \begin{cases} \mathbf{q} \cdot \log \mathbf{S}(\mathbf{q}) & \mathbf{q} \in \Delta \\ +\infty & \textit{otherwise}, \end{cases} \tag{13}$$

[8] For details see (Rockafellar, 1970, ch. 12). Briefly, for a convex function $g(\mathbf{x})$, its convex conjugate function is defined as $g^*(\mathbf{y}) = \max_{\mathbf{x}} \{\mathbf{x} \cdot \mathbf{y} - g(\mathbf{x})\}$, which is also convex. Fenchel's theorem then establishes that $g(\mathbf{x}) = \max_{\mathbf{y}} \mathbf{x} \cdot \mathbf{y} - g^*(\mathbf{y})$. When $\mathbf{x}$ and $\mathbf{y}$ are scalar and $g(x)$ is differentiable, then $g(\mathbf{x})$ and $g^*(\mathbf{y})$ are inverse mappings to each other. Vohra (2011) applies these ideas to the mechanism design setting.

[9] To the best of our knowledge, this result is new in the literature on random utility models, and may be of independent interest. In particular, this result is related to the literature on perturbed random utility models, which has been focused on characterizing choice probabilities as the solution of a deterministic optimization problem (Hofbauer and Sandholm (2002); Fosgerau and McFadden (2012); Fudenberg et al. (2015)).

where $\mathbf{S}(\cdot)$ *is the inverse mapping for the* $\mathbf{T}$ *function in Eq. (10). We call the negative convex conjugate* $-W^*(\cdot)$ *a* ***generalized entropy****.*

Remark 4 (RI-logit revisited.) *To see how this works in a special case, let us consider the multinomial logit model. In this case,* $\mathbf{T}$ *is the identity, implying that its inverse,* $\mathbf{S}(\mathbf{q}) = \mathbf{q}$*, is also just the identity. Then by Proposition 3(ii), the negative convex conjugate of the surplus function is* $-W^*(\mathbf{q}) = -\mathbf{q} \cdot \log \mathbf{q} = -\sum_i q_i \log q_i$*, which is just the Shannon (1948) entropy.*

Moreover, we see that Eq. (13) implies that the RI-logit optimization problem (3), written as

$$\max_{\mathbf{p}(\cdot)} \sum_{\mathbf{v} \in \mathcal{V}} \left\{ \mathbf{p}(\mathbf{v}) \cdot [\mathbf{v} + \log \mathbf{p}^0] - W^*(\mathbf{p}(\mathbf{v})) \right\} \mu(\mathbf{v}),$$

has the multinomial logit choice probabilities in Eq. (6) as solution. ■

Proposition 3(i) generalizes the "logsum" formula for the multinomial logit model to the entire class ARUM. Similarly, generalizing from the logit case, $-W^*$, the negative convex conjugate of the surplus W of any ARUM may be viewed as a *generalized entropy*. In particular, Proposition 3(ii) shows how the generalized entropy may be expressed in terms of the inverse scaled demand **S** as $-W^*(\mathbf{q}) = -\mathbf{q} \cdot \log \mathbf{S}(\mathbf{q})$.

To aid further interpretation of the generalized entropy function, note that Eq. (8) implies that the surplus function can be written as

$$W(\mathbf{v}) = \sum_{i=1}^{N} q_i(\mathbf{v})(v_i + \mathbb{E}(\epsilon_i | u_i \geq u_j, j \neq i)).$$

Combining this with (11), we obtain an alternative expression for the generalized entropy, as a choice probability-weighted sum of expectations of the utility shocks $\boldsymbol{\epsilon}$:[10]

$$-W^*(\mathbf{q}) = \sum_i q_i \mathbb{E}[\epsilon_i | u_i \geq u_j, j \neq i].$$

[10] See Chiong et al. (2016).

In this way, different distributions for the utility shocks $\boldsymbol{\epsilon}$ in the random utility model will imply different generalized entropies.

We may also interpret $-\log \mathbf{S}(\mathbf{q})$ as follows. Given $\mathbf{q}$ in the interior of the simplex, there exists $\mathbf{v}$ such that $(\mathbf{q}, \mathbf{v})$ satisfy (9), see Norets and Takahashi (2013). Then, using (12),[11]

$$-\log S_j(\mathbf{q}) = W(\mathbf{v}) - v_j,$$

which means that $-\log S_j(\mathbf{q})$ is the expected utility gain from the discrete choice relative to the deterministic utility component of option j. This coincides with the $\boldsymbol{\psi}_j(\cdot)$ mapping introduced in Lemma 1 of Arcidiacono and Miller (2011), which is a key component for the estimation procedures developed in that paper.

Proposition 8 in the Appendix contains important mathematical properties of the class of inverse scaled demand functions $\mathbf{S}$, which are used in proving the key propositions in the remainder of the paper.

4 Generalizing the RI-logit model: RI-ARUM models with Bregman Information Cost

In this section we generalize the equivalence result between rational inattention and multinomial logit. We begin by generalizing the rational inattention framework described in Section 2, using generalized entropy in place of the Shannon entropy. Specifically, we let $\mathbf{S}$ be the inverse scaled demand corresponding to some ARUM satisfying Assumption 1, and define $\Omega_{\mathbf{S}}(\mathbf{p}) = -\mathbf{p} \cdot \log \mathbf{S}(\mathbf{p})$ as the corresponding generalized entropy.

For a strictly convex function f, the Bregman (1967) divergence associated with f is a function of probability vectors $(\mathbf{p}, \mathbf{q})$ to the real line defined by $D_f(\mathbf{p}||\mathbf{q}) \equiv f(\mathbf{p}) - f(\mathbf{q}) - \nabla f(\mathbf{q}) \cdot (\mathbf{p} - \mathbf{q})$. It measures the vertical distance from $f(\mathbf{p})$ to a tangent hyperplane to f at the point $\mathbf{q}$. By convexity of f, this distance is positive and increasing away from $\mathbf{q}$. When $\mathbf{p} = \mathbf{p}(\mathbf{V})$ is random, we can consider the expected Bregman divergence $\mathbb{E}D_f(\mathbf{p}(\mathbf{V})||\mathbf{q})$, which measures the expected divergence of $\mathbf{p}(\mathbf{V})$ from $\mathbf{q}$. Banerjee et al. (2005) show that $\mathbb{E}D_f(\mathbf{p}(\mathbf{V})||\mathbf{q})$ is

[11] We have $\log \mathbf{S}(\mathbf{q}) = \log \mathbf{S}(\mathbf{T}(e^{\mathbf{v}})/(\sum_j T_j(e^{\mathbf{v}}))) = \log(e^{\mathbf{v}}/(\sum_j T_j(e^{\mathbf{v}}))) = \mathbf{v} - W(\mathbf{v})$. where we have used the fact that $\mathbf{S}$ is homogeneous of degree 1.

minimized at $\mathbf{q} = \mathbb{E}\mathbf{p}\left(\mathbf{V}\right) = \mathbf{p}^0$ for any choice of f. Banerjee et al. (2005) use this observation to define the Bregman information of $\mathbf{p}\left(\mathbf{V}\right)$ as $\mathbb{E}D_f\left(\mathbf{p}\left(\mathbf{V}\right)||\mathbf{p}^0\right)$, noting that this is the expected distortion, as measured by the Bregman divergence, when replacing a random $\mathbf{p}\left(\mathbf{V}\right)$ by the optimal constant vector $\mathbf{p}^0$.

We define accordingly an information cost as the Bregman information associated with the negative of the generalized entropy $\Omega_{\mathbf{S}}$, i.e. $\kappa_{\mathbf{S}}\left(\mathbf{p}\left(\cdot\right),\mu\right) = \mathbb{E}D_{-\Omega_{\mathbf{S}}}\left(\mathbf{p}\left(\mathbf{V}\right)||\mathbf{p}^0\right)$.[12] Proposition 5 below establishes that

$$\begin{aligned} \kappa_{\mathbf{S}}\left(\mathbf{p}\left(\cdot\right),\mu\right) &= \Omega_{\mathbf{S}}\left(\mathbf{p}^0\right) - \mathbb{E}\Omega_{\mathbf{S}}\left(\mathbf{p}(\mathbf{V})\right) \qquad (14) \\ &= -\mathbf{p}^0 \cdot \log \mathbf{S}\left(\mathbf{p}^0\right) + \sum_{\mathbf{v}\in\mathcal{V}} \left[\mathbf{p}\left(\mathbf{v}\right) \cdot \log \mathbf{S}\left(\mathbf{p}\left(\mathbf{v}\right)\right)\right]\mu\left(\mathbf{v}\right). \end{aligned}$$

In particular, the information cost $\kappa(\mathbf{p}(\cdot),\mu)$ is equal to the Bregman information, associated with the (negative) Shannon entropy, which is well known as the mutual (Shannon) information. The interpretation of our information cost $\kappa_{\mathbf{S}}$ is analogous to the information cost for the RI-logit model, in Eq. (2) above. It measures consumers' *action adjustment costs* associated with shifting behavior from the state-independent unconditional choice probabilities $\mathbf{p}_0$ to the state-dependent conditional choice probabilities $\mathbf{p}(\mathbf{v})$.[13] Taking information cost $\kappa_{\mathbf{S}}$ as Bregman information means that the information cost inherits properties of the Bregman divergence, as stated in the next proposition.

Proposition 5 *For any generalized entropy function $\Omega_{\mathbf{S}}$, the information cost $\kappa_{\mathbf{S}}\left(\mathbf{p}\left(\cdot\right),\mu\right)$ in (14) is the expectation of the Bregman divergence associated with $\Omega_{\mathbf{S}}$ of $\mathbf{p}\left(\cdot\right)$ and $\mathbf{p}^0$. Hence it is convex in $\mathbf{p}\left(\cdot\right)$ when holding $\mathbf{p}^0$ constant and $\kappa_{\mathbf{S}}\left(\mathbf{p}\left(\cdot\right),\mu\right) = 0$ if action and state are independent.*

The information cost $\kappa_{\mathbf{S}}$ takes context into account by construction; that is, going back to Example 1, exchanging the labels of good 1 (pineapple) and good 3 (cheesecake) affects information costs and therefore choices by design. Allowing the information cost function to depend on context in this way entails some loss of

[12] Strict concavity of $\Omega_{\mathbf{S}}$ is established in Proposition 8 in the Appendix.

[13] Using Bayes' rule, such a shift in choice probabilities corresponds to a change in beliefs from the prior μ to a posterior $\mu(\mathbf{v}|i) \propto p_i(\mathbf{v})\mu(\mathbf{v})$, which Caplin and Dean (2015) and Chambers et al. (2018) refer to as "revealed posterior" distributions.

generality, as it need not satisfy Blackwell's information ordering. In particular, $\mathbf{S}$ is not invariant with respect to permutation in its arguments. Example 5.2 below illustrates this for the case of $\kappa_{\mathbf{S}}$ corresponding to a nested logit model.

Using the generalized cost function $\kappa_{\mathbf{S}}$ just introduced, we now define a new RI model describing a DM who chooses the collection of conditional probabilities $\mathbf{p}(\cdot) = \{\mathbf{p}(\mathbf{v})\}_{\mathbf{v}\in\mathcal{V}}$ to maximize his expected payoff less the general information cost

$$\max_{\mathbf{p}(\cdot)} \{\mathbb{E}(\mathbf{V}\cdot\mathbf{p}(\mathbf{V})) - \kappa_{\mathbf{S}}(\mathbf{p}(\cdot), \mu)\} \tag{15}$$
$$= \max_{\mathbf{p}(\cdot)} \left\{ \sum_{\mathbf{v}\in\mathcal{V}} \{\mathbf{p}(\mathbf{v})\cdot[\mathbf{v} + \log\mathbf{S}(\mathbf{p}^0)] - \mathbf{p}(\mathbf{v})\cdot\log\mathbf{S}(\mathbf{p}(\mathbf{v}))\}\,\mu(\mathbf{v}) \right\}.$$

We refer this model as RI-ARUM in order to make explicit the fact that the cost function (14) is defined in terms of the generalized entropy $\Omega_{\mathbf{S}}(\mathbf{q})$ which is derived from an ARUM. The maximization problem in Eq. (15) is an extension of the maximization problem in Eq. (11) that is representative for the ARUM. In fact, holding $\mathbf{p}^0$ *fixed*, the RI-ARUM objective function (15) above coincides, pointwise in $\mathbf{v}$, with the problem (11), where the generalized entropy function is $W^*(\mathbf{p}) = -\mathbf{p}\cdot\log\mathbf{S}(\mathbf{p})$ as stated in Proposition 3. This connection underlies the finding, elaborated in the following proposition, that the optimal conditional choice probabilities for any RI-ARUM model have the logit-like closed form from Eq. (9) above.

Proposition 6 *Let T be the scaled demand of an ARUM and let $S = T^{-1}$ be the inverse scaled demand. Let $\mathbf{p}(\cdot)$ be the solution to the corresponding RI-ARUM model and $\mathbf{p}^0 = \mathbb{E}\mathbf{p}(\mathbf{V})$. Then*

(i) *The unconditional probabilities satisfy the fixed point equation*

$$\mathbf{p}^0 = \mathbb{E}\left(\frac{\mathbf{T}\left(e^{\mathbf{V}+\log\mathbf{S}(\mathbf{p}^0)}\right)}{\sum_{j=1}^N T_j\left(e^{\mathbf{V}+\log\mathbf{S}(\mathbf{p}^0)}\right)}\right). \tag{16}$$

(ii) *The conditional probabilities are given in terms of the unconditional probabil-*

ities by

$$\mathbf{p}\left(\mathbf{v}\right)=\frac{\mathbf{T}\left(e^{\mathbf{v}+\log\mathbf{S}\left(\mathbf{p}^{0}\right)}\right)}{\sum_{j=1}^{N}T_{j}\left(e^{\mathbf{v}+\log\mathbf{S}\left(\mathbf{p}^{0}\right)}\right)}. \tag{17}$$

(iii) *The optimized value of (15) is*

$$\mathbb{E}\log\sum_{j=1}^{N}T_{j}\left(e^{\mathbf{V}+\log\mathbf{S}\left(\mathbf{p}^{0}\right)}\right)=\mathbb{E}W\left(\mathbf{V}+\log\mathbf{S}\left(\mathbf{p}^{0}\right)\right). \tag{18}$$

The unconditional and conditional choice probabilities in (16) and (17) generalize the corresponding expressions for the RI-logit model in a straightforward way. We note in particular that the influence of the prior information is captured completely by the vector $\log\mathbf{S}(\mathbf{p}^{0})$. The implications of prior for the behavior of an RI-ARUM agent can be summarized by a vector in $\mathbb{R}^{N}$ where N is the number of choice options. This is true regardless of the form of the prior beliefs.

Part (i) of the proposition shows that the solution of the RI-ARUM model involves a fixed point problem; in what follows, we assume that a solution exists. In general, the uniqueness of a solution to RI-ARUM is not guaranteed. By Cover and Thomas (2006, Thm 2.7.4), the mutual (Shannon) information $\kappa(\mathbf{p}(\cdot),\mu)$ is convex as a function of the conditional probability $\mathbf{p}\left(\cdot\right)$, and strictly convex when the conditional probabilities differ from the unconditional probability $\mathbf{p}^{0}$. In this case, the equations in Proposition 6 uniquely identify the solution to the RI-ARUM model. By extension, as we show in Proposition 10 in the Appendix, this conclusion also applies to the nested logit model. Whether it applies to all RI-ARUM remains an open question.

Proposition 6(i) also implies that an important feature of the RI-ARUM model is that some p_{i}^{0} may be zero, in which case the corresponding $p_{i}\left(\mathbf{v}\right)$ are also zero. To see this, consider Eq. (17). When $p_{i}^{0}=0$, then Proposition 2(ii) implies that $\log S_{i}(\mathbf{p}^{0})=-\infty$ and hence $p_{i}(\mathbf{v})=0$. Following the literature, we refer to the set of options chosen with positive probability as the *consideration set* (e.g. Caplin et al. (2018)).

Proposition 6(iii) indicates an alternative way for calculating the value attained by optimal rationally inattentive behavior. In particular, Eq. (18) shows that the optimal value of program (15) can be computed as the expected surplus function of the appropriately shifted ARUM. This generalizes the corresponding Lemma 2 in Matějka and McKay (2015).

While Proposition 6 does not explicitly characterize the consideration set emerging from a RI-ARUM problem, Corollary 9 in the Appendix describes one important feature that it has, namely that it excludes options that offer the lowest utility in all states of the world.

4.1 Equivalence between discrete choice (ARUM) and rational inattention

We now establish the central result of this paper, namely the equivalence between additive random utility discrete choice models and RI models. In particular, we show that the choice probabilities generated by an RI-ARUM lead to the same choice probabilities as a corresponding ARUM and vice versa. In particular, comparing the expressions for the choice probabilities in the RI-ARUM model in (17) to those in an ARUM in (9), it is clear that such a result is available: the expressions for the choice probabilities are identical except for the location shift of the deterministic utility components $\mathbf{v}$ by the vector $\log \mathbf{S}\left(\mathbf{p}^0\right)$ in the RI-ARUM model.

Proposition 7 *For every RI-ARUM with prior $(\mu, \mathcal{V})$, inverse scaled demand $\mathbf{S}$ and choice probabilities $\mathbf{p}(\mathbf{v})$ there is an ARUM defined on the consideration set of the RI-ARUM that yields the same choice probabilities for all $\mathbf{v} \in \mathcal{V}$.*

Conversely, for every ARUM with choice probabilities $\mathbf{q}(\mathbf{v})$ and inverse scaled demand $\mathbf{S}$ and given a prior $(\mu, \mathcal{V})$ such that the corresponding information cost $\kappa_{\mathbf{S}}(\mathbf{p}(\cdot), \mu)$ is strictly convex, there is a location shift vector $\mathbf{c}$ such that the RI-ARUM with prior $(\mathbf{v} \rightarrow \mu(\mathbf{v}-\mathbf{c}), \mathcal{V}+\mathbf{c})$ and inverse scaled demand $\mathbf{S}$ has choice probabilities $\mathbf{p}$ that satisfy $\mathbf{p}(\mathbf{v}-\mathbf{c}) = \mathbf{q}(\mathbf{v})$ for all $\mathbf{v} \in \mathcal{V}$.

This proposition implies a new interpretation of ARUM models as describing boundedly rational behavior, which suggests that in order to apply an ARUM, one need not assume that decision makers are completely aware of the valuations of all the available options. This is important, as it is clearly unrealistic to expect decision makers to be aware of all options when the number is large.

At the same time, despite the formal equivalence in the choice probabilities in ARUM and RI-ARUM models under the conditions of Proposition 7, there are several important differences between them. The RI-ARUM model also allows some options to have zero unconditional choice probabilities p_i^0. Since choice probabilities are necessarily positive in the ARUM under Assumption 1, the equivalence is defined only for the options which are in the consideration set of the RI-ARUM model, as is clear from Proposition 7.

Moreover, the equivalence requires fixing the prior over states $(\mu, \mathcal{V})$; a prior is part of the RI-ARUM model but needs to be added to the random utility model. When we consider two choice scenarios with different priors, the subsequent choices in the RI-ARUM and ARUM models can deviate considerably. In particular, the RI-ARUM class allows options in the choice set to be *complements*, in the sense that increasing the payoff of an option in some state may lead to an increase in the choice probability of some other options. Such complementarities are explicitly ruled out by ARUM discrete choice models.[14]

Additionally, we have necessary and sufficient conditions for a system of choice probabilities to be consistent with an ARUM (Fosgerau et al., 2013). By Proposition 7, the same conditions are then necessary for a system of choice probabilities that derives from an RI-ARUM and some fixed prior.

Starting from a given ARUM model, proceeding to the corresponding RI-ARUM model requires deriving the convex conjugate function corresponding to the social surplus function of the given ARUM. Explicit closed forms for the convex conjugate functions are available, as far as we are aware, only for the multinomial and nested logit models. However, in general, the mass transport approach in Chiong et al. (2016) can be used to simulate the convex conjugate function for *any* ARUM, via computationally straightforward linear programming algorithms.

[14] Indeed, in empirical papers utilizing discrete-choice demand models, complementarities between choices can be typically accommodated only by modelling consumers as choosing “bundles” of options in the choice set; see, eg. Gentzkow (2007) and Fox and Lazzati (2017). Such an approach may become intractable as the dimensionality of the choice set increases. In contrast, complementarities can arise in the RI model both from correlation in the priors (as pointed out by Matějka and McKay (2015)), and also from the form of the generalized entropy information cost functions considered in this paper.

In what follows, we illustrate these features for a specific example; namely, we study a RI-ARUM model in which the choice probabilities are equivalent to those from a *nested logit* discrete choice model, a frequently-used model in empirical applications.

5 The RI-nested logit model

From an applied point of view, an important implication of Proposition 7 is that it allows us to formulate rational inattention models that have complex substitution patterns, going beyond the multinomial logit case. In this section, we consider an RI-ARUM model with information cost derived from a nested logit model. The nested logit choice probabilities are consistent with a discrete choice model in which the utility shocks $\boldsymbol{\epsilon}$ have a certain generalized extreme value joint distribution. Among applied researchers, the nested logit model is often preferred over the multinomial logit model because it allows some products to be closer substitutes than others, thus avoiding the restrictions implied by the IIA property.[15]

We partition the set of options $i \in \{1, \dots, N\}$ into mutually exclusive nests, and let g_i denote the nest containing option i. Let $\zeta_{g_i} \in (0, 1]$ be nest-specific parameters. For a valuation vector $\mathbf{v}$, the nested logit choice probabilities are given by

$$q_i(\mathbf{v}) = \frac{e^{v_i/\zeta_{g_i}}}{\sum_{j\in g_i} e^{v_j/\zeta_{g_i}}} \cdot \frac{e^{\zeta_{g_i} \log\left(\sum_{j\in g_i} e^{v_j/\zeta_{g_i}}\right)}}{\sum_{\text{all nests } g} e^{\zeta_g \log\left(\sum_{j\in g} e^{v_j/\zeta_g}\right)}}. \tag{19}$$

The inverse scaled demand $\mathbf{S}$ corresponding to a nested logit model is

$$S_i(\mathbf{q}) = q_i^{\zeta_{g_i}} \left(\sum_{j\in g_i} q_j\right)^{1-\zeta_{g_i}}. \tag{20}$$

Applying Proposition 7, the nested logit choice probabilities (19) are the same as

[15] In order to be consistent with the definition of IIA, when applied to RI models we assume that the DM's prior is fixed. This assumption allows us to keep the choice set constant so that we can focus on changes in the utilities associated to alternatives. For further details about IIA see Maddala (1986, Chap. 2), and Anderson et al. (1992).

those from a RI-ARUM model with valuations

$$v_i - \zeta_{g_i} \log p_i^0 - (1 - \zeta_{g_i}) \log \left(\sum_{j \in g_i} p_j^0 \right), \quad i \in \{1, \ldots, n\}. \tag{21}$$

The inverse scaled demand $\mathbf{S}$ for the nested logit model in Eq. (20) has several interesting features, relative to the Shannon entropy. First, Eq. (20) allows us to write the generalized entropy $\Omega_{\mathbf{S}}(\mathbf{p})$ as

$$\Omega_{\mathbf{S}}(\mathbf{p}) = -\sum_{i=1}^{N} \zeta_{g_i} p_i \log p_i - \sum_{i=1}^{N} (1 - \zeta_{g_i}) p_i \log \left(\sum_{j \in g_i} p_j \right). \tag{22}$$

The first term in Eq (22) captures the Shannon entropy within nests, whereas the second term captures the information between nests. According to this, we may interpret Eq. (22) as an augmented version of the Shannon entropy. It is also apparent from (22) that $\Omega_{\mathbf{S}}(\mathbf{p})$ is not invariant to reordering of the choice probabilities, due to the second term.

Second, when the nesting parameters $\zeta_{g_j} = 1$, then $\mathbf{S}$ is the identity ($S_j(\mathbf{p}) = p_j$ for all j), corresponding to the Shannon entropy. When $\zeta_{g_j} < 1$, then $S_j(\mathbf{p}) \geq p_j$; here, $\mathbf{S}(\mathbf{p})$ behaves as a probability weighting function that tends to overweight options j belonging to larger nests. At the extreme $\zeta_{g_j} \to 0$, all options within the same nest effectively collapse into one aggregate option and become perfect substitutes.

We denote this model as RI-nested logit (hereafter RI-NL). Using this model, we consider two examples, emphasizing both differences and similarities of the RI-NL vis-a-vis the RI-logit model.

5.1 Example 1: mango-pineapple-cheesecake continued

We return to the earlier pineapple-mango-cheesecake example from Section 1. For these three products, we consider a model with two nests, in which the tropical fruits pineapple (good 1) and mango (good 2) are placed in one nest g_1, while cheesecake (good 3) is placed by itself in a second nest g_2. For the nesting parameters, we choose $\zeta_{g_1} = 0.5$. The value of ζ_{g_2} is irrelevant since nest g_2 comprises just one alternative. Recall that there are four equally likely possible states:

$$\left(\mathbf{v}^1,\mathbf{v}^2,\mathbf{v}^3,\mathbf{v}^4\right)=\begin{pmatrix} 0 & 0.1 & 0 & 0 \\ 0 & 0 & 0.1 & 0 \\ 0 & 0 & 0 & 0.1 \end{pmatrix}. \tag{23}$$

Solving this RI-NL model leads to $\log \mathbf{S}(\mathbf{p}^0)$ which is a constant vector plus $(0,0,-1.18)^\top$. Hence the nested logit model with payoffs shifted by this vector produces the same choice probabilities as the RI-NL. The RI-NL conditional choice probabilities,

$$\begin{pmatrix} 0.41 & 0.46 & 0.37 & 0.40 \\ 0.41 & 0.37 & 0.46 & 0.40 \\ 0.18 & 0.17 & 0.17 & 0.19 \end{pmatrix},$$

do not satisfy IIA and are hence not compatible with the RI-logit model.[16]

Conversely, we can start with the nested logit model with the payoffs given in (23) above and the same nest parameters as before. The conditional choice probability vectors for this model are

$$\left(\mathbf{p}^1,\mathbf{p}^2,\mathbf{p}^3,\mathbf{p}^4\right)=\begin{pmatrix} 0.29 & 0.33 & 0.27 & 0.28 \\ 0.29 & 0.27 & 0.33 & 0.28 \\ 0.41 & 0.40 & 0.40 & 0.44 \end{pmatrix},$$

and the unconditional choice probability vector is $\mathbf{p}^0=(0.29,0.29,0.41)^\top$ under the uniform prior. The corresponding location shift vector $\log \mathbf{S}\left(\mathbf{p}^0\right)$ is $(0,0,-0.001)^\top$ up to a constant and shifting the payoffs by this amount produces RI-NL conditional choice probabilities that are equal to the nested logit choice probabilities.

For comparison, we also compute the case where the nesting parameter has been set to $\zeta_{g_1}=0.4$, which makes the alternatives pineapple and mango closer substitutes than before. The RI-NL unconditional choice probability vector be-

[16] But even with the logit specification, the IIA property can break down if the DM is able to consume more than one product, ie. *bundles* of goods (c.f. Gentzkow, 2007).

comes $\mathbf{p}^0 = (0.45, 0.45, 0.10)^\top$ and the RI-NL conditional choice probabilities become

$$\begin{pmatrix} 0.45 & 0.51 & 0.39 & 0.44 \\ 0.45 & 0.39 & 0.51 & 0.44 \\ 0.10 & 0.10 & 0.10 & 0.11 \end{pmatrix}.$$

The changes across states deviate more from IIA than when $\zeta_{g_1} = 0.5$. It appears that, as goods 1 and 2 become more substitutable, the DM is able to make better choices in states where goods 1 or 2 are optimal. This comparative statics suggests that increasing the substitutability between a set of goods corresponds to shifts in the information structure towards signals which allow the DM to better distinguish between states in which these goods are optimal.

5.2 Example 2: swapping alternatives can lead to increased information cost

As we have stated through the paper, our cost functions embody information related to the identity of alternatives. In order to see how this feature works in practice, consider a nested logit with four choice options, nests formed by options 1-2 and 3-4, and with two states $\mathbf{v}^1 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$ and $\mathbf{v}^2 = \left(\frac{1}{8}, \frac{3}{8}, \frac{1}{8}, \frac{3}{8}\right)$. Define $\log S_i(\boldsymbol{q}) = \zeta \log q_i + (1-\zeta) \log \left(\sum_{j \in g_i} q_j\right)$, where g_i is the nest that contains option i. A garbling that swaps alternatives 2 and 3 will move probability mass across nests in state 2 but not in state 1. Then the probability distribution across nests is independent of the state without garbling but not with garbling and therefore this garbling increases the information cost κ_S.

6 Conclusions

The central result in this paper is the equivalence between an additive random utility discrete choice model and a corresponding RI-ARUM. Thus any additive random utility discrete choice model can be cast as a model of rationally inattentive behavior, and vice-versa for any RI-ARUM; we can go back and forth between the two paradigms.[17] Then, in order to apply an ARUM, it is no longer necessary

[17] In a similar vein, Webb (2019) demonstrates an equivalence between random utility models and bounded-accumulation or drift-diffusion models of choice and reaction times used in the neuroeco-

to assume that decision makers are completely aware of the valuations of all the available options. This is important, as it is clearly unrealistic to expect decision makers to be aware of all options when the number is large.

Our equivalence result is at the individual level, hence it also holds for ARUM with random parameters, including the mixed logit or random coefficient logit models which have been popular in applied work.[18]

Our equivalence result generalizes to perturbed random utility models (e.g. Hofbauer and Sandholm (2002) and Fudenberg et al. (2015)) where the information cost is the Bregman information associated with a (negative) generalized entropy (Fosgerau et al., 2019). We are also exploring connections between our results and those in the decision theory literature. Gul et al. (2014), for instance, show an equivalence between random utility and an "attribute rule" model of stochastic choice, and we conjecture that our results may be useful in showing similar results for other decision-theoretic models.

Finally, there are rational inattention models outside the RI-ARUM framework; that is, rational inattention models with information costs outside the class of generalized entropies introduced in this paper.[19] Obviously, choice probabilities from these non-RI-ARUM models would not be equivalent to those which can be generated from ARUM models; it will be interesting to examine the empirical distinctions that non-RI-ARUM choice probabilities would have.

The properties in Proposition 8 are satisfied by any ISD corresponding to an ARUM but do not characterize ARUM. In fact, the properties may be used to define a class of generalized entropies that is strictly larger than the class consisting of those generalized entropies corresponding to ARUM (see Fosgerau et al., 2019). We have not found direct conditions that characterize those generalized entropies that correspond to ARUM. We have chosen in this paper to work with the generalized entropies that derive from ARUM in order to emphasize the main point of the paper: the connection between RI with our information cost and ARUM. However, the conclusions of Proposition 6 extend without change to the case when $\mathbf{S}$

nomics and psychology literature.

[18] See, for instance, Berry et al. (1995), McFadden and Train (2000), Fox et al. (2012).

[19] As an example, the function $g(\mathbf{p}) = -\sum_{i=1}^{N} \log(p_i)$ is not a generalized entropy function; thus a rational inattention model using this as an information cost function would lie outside the RI-ARUM framework.

is an inverse scaled demand that satisfies the conclusions of Proposition 8 but not necessarily corresponds to an ARUM. For applications it is then possible to work with such generalized entropies without needing to check that they correspond to an ARUM.

References

Allen, R. and Rehbeck, J. (2019) Identification With Additively Separable Heterogeneity *Econometrica* **87**(3), 1021–1054.

Anderson, S. P., De Palma, A. and Thisse, J.-F. (1988) A Representative Consumer Theory of the Logit Model *International Economic Review* **29**(3), 461–466.

Anderson, S. P., De Palma, A. and Thisse, J. F. (1992) *Discrete choice theory of product differentiation* MIT Press Cambridge, MA.

Arcidiacono, P. and Miller, R. A. (2011) Conditional Choice Probability Estimation of Dynamic Discrete Choice Models With Unobserved Heterogeneity *Econometrica* **79**(6), 1823–1867.

Banerjee, A., Merugu, S., Dhillon, I. S. and Ghosh, J. (2005) Clustering with Bregman Divergences *Journal of Machine Learning Research* **6**, 1705–1749.

Berry, S. (1994) Estimating discrete-choice models of market equilibrium *The RAND Journal of Economics* **25**(2), 242–262.

Berry, S. T., Levinsohn, J. and Pakes, A. (1995) Automobile Prices in Market Equilibrium *Econometrica* **63**(4), 841–890.

Bregman, L. (1967) The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming *USSR Computational Mathematics and Mathematical Physics* **7**(3), 200–217.

Brown, Z. Y. and Jeon, J. (2019) Endogenous Information Acquisition and Insurance Choice.

Caplin, A. and Dean, M. (2015) Revealed Preference, Rational Inattention, and Costly Information Acquisition *The American Economic Review* **105**(7), 2183–2203.

Caplin, A., Dean, M. and Leahy, J. (2017) Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy *Technical report* National Bureau of Economic Research Cambridge, MA.

Caplin, A., Dean, M. and Leahy, J. (2018) Rational Inattention, Optimal Consideration Sets, and Stochastic Choice *The Review of Economic Studies* .

Caplin, A., Leahy, J. and Matějka, F. (2016) Rational Inattention and Inference from Market Share Data.

Chambers, C. P., Liu, C. and Rehbeck, J. (2018) Costly Information Acquisition.

Chiong, K. X., Galichon, A. and Shum, M. (2016) Duality in dynamic discrete-choice models *Quantitative Economics* **7**(1), 83–115.

Chiong, K. X. and Shum, M. (2019) Random Projection Estimation of Discrete-Choice Models with Large Choice Sets *Management Science* **65**(1), 256–271.

Cover, T. M. and Thomas, J. A. (2006) *Elements of information theory* 2nd edn Wiley-Interscience.

Fosgerau, M., McFadden, D. and Bierlaire, M. (2013) Choice probability generating functions *Journal of Choice Modelling* **8**.

Fosgerau, M. and McFadden, D. L. (2012) A theory of the perturbed consumer with general budgets *NBER Working Paper* pp. 1–27.

Fosgerau, M., Monardo, J. and de Palma, A. (2019) The inverse product differentiation logit model.

Fox, J. T., Kim, K. I., Ryan, S. P. and Bajari, P. (2012) The random coefficients logit model is identified *Journal of Econometrics* **166**(2), 204–212.

Fox, J. T. and Lazzati, N. (2017) A note on identification of discrete choice models for bundles and binary games *Quantitative Economics* **8**(3), 1021–1036.

Frankel, A. and Kamenica, E. (2018) Quantifying information and uncertainty.

Fudenberg, D., Iijima, R. and Strzalecki, T. (2015) Stochastic Choice and Revealed Perturbed Utility *Econometrica* **83**(6), 2371–2409.

Gentzkow, M. (2007) Valuing New Goods in a Model with Complementarity: Online Newspapers *The American Economic Review* **97**(3), 713–744.

Gul, F., Natenzon, P. and Pesendorfer, W. (2014) Random Choice as Behavioral Optimization *Econometrica* **82**(5), 1873–1912.

Hébert, B. and Woodford, M. (2017) Rational Inattention with Sequential Information Sampling *NBER Working Paper* (w23787).

Hobson, A. (1969) A new theorem of information theory *Journal of Statistical Physics* **1**(3), 383–391.

Hofbauer, J. and Sandholm, W. H. (2002) On the global convergence of stochastic fictitious play *Econometrica* **70**(6), 2265–2294.

Joo, J. (2019) Rational Inattention as an Empirical Framework – With an Application to the Welfare Effects of New Product Introduction and Endogenous Promotion.

Maddala, G. S. (1986) *Limited-dependent and qualitative variables in econometrics*. Cambridge University Press Cambridge.

Matějka, F. and 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. (1978) Modelling the choice of residential location *in* A. Karlquist, F. Snickars and J. W. Weibull (eds), *Spatial Interaction Theory and Planning Models* Vol. 673 North Holland Amsterdam pp. 75–96.

McFadden, D. (1981) Econometric Models of Probabilistic Choice *in* C. Manski and D. McFadden (eds), *Structural Analysis of Discrete Data with Econometric Applications* MIT Press Cambridge, MA, USA pp. 198–272.

McFadden, D. and Train, K. (2000) Mixed MNL Models for Discrete Response *Journal of Applied Econometrics* **15**(November 1998), 447–470.

Melo, E., Pogorelskiy, K. and Shum, M. (2019) TESTING THE QUANTAL RESPONSE HYPOTHESIS *International Economic Review* **60**(1), 53–74.

Morris, S. and Yang, M. (2016) Coordination and Continuous Choice *SSRN Electronic Journal* .

Norets, A. and Takahashi, S. (2013) On the surjectivity of the mapping between utilities and choice probabilities *Quantitative Economics* **4**(1), 149–155.

Porcher, C. (2019) Migration with Costly Information.

Rockafellar, R. T. (1970) *Convex Analysis* Princeton University Press Princeton, N.J.

Shannon, C. E. (1948) A Mathematical Theory of Communication *Bell System Technical Journal* **27**(3), 379–423.

Shi, X., Shum, M. and Song, W. (2018) Estimating Semi-Parametric Panel Multinomial Choice Models Using Cyclic Monotonicity *Econometrica* **86**(2), 737–761.

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

Sims, C. A. (2010) Rational inattention and monetary economics *Handbook of Monetary Economics* Vol. 3 Elsevier chapter 4, pp. 155–181.

Vohra, R. (2011) *Mechanism Design: A Linear Programming Approach* Cambridge University Press Cambridge.

Webb, R. (2019) The (Neural) Dynamics of Stochastic Choice *Management Science* **65**(1), 230–255.

A Proofs of results in main text

Proof of Proposition 2. Note first that we may write

$$\mathbf{T}\left(e^{\mathbf{v}}\right) = e^{W(\mathbf{v})}\mathbf{q}\left(\mathbf{v}\right).$$

The probabilities in $\mathbf{q}$ are never zero since the random utility shocks have full support. Define for convenience $X = \left\{\mathbf{v} \in \mathbb{R}^N | v_1 = 0\right\}$. The results in Norets and Takahashi (2013) apply to the mapping $\mathbf{q}$: Hence $\mathbf{q}$ is a bijection between X and the interior of the unit simplex Δ.

To obtain injectivity of $\mathbf{T}$ on $\mathbb{R}^N_+$, suppose that $\mathbf{T}\left(e^{\mathbf{v}}\right) = \mathbf{T}\left(e^{\mathbf{v}'}\right)$ and aim to show that $\mathbf{v} = \mathbf{v}'$. Since $T_i\left(e^{\mathbf{v}}\right) = e^{W(\mathbf{v})}q_i\left(\mathbf{v}\right)$ and $\sum_{i=1}^{N} q_i = 1$, we may sum $\sum_{i=1}^{N} T_i\left(e^{\mathbf{v}}\right) = \sum_{i=1}^{N} T_i\left(e^{\mathbf{v}'}\right)$ to find that $W\left(\mathbf{v}\right) = W\left(\mathbf{v}'\right)$ and hence that $\mathbf{q}\left(\mathbf{v}\right) = \mathbf{q}\left(\mathbf{v}'\right)$. Then by the Norets and Takahashi (2013) result, $\mathbf{v} = \mathbf{v}' + (c, ..., c)$ which leads to $W\left(\ \mathbf{v}\right) = W\left(\mathbf{v}'\right) + c = W\left(\mathbf{v}\right) + c$, and hence $c = 0$.

Consider next surjectivity and let $\mathbf{x} \in \mathbb{R}^N_+$ be an arbitrary point. We aim to solve the equation $\mathbf{T}\left(\mathbf{y}\right) = \mathbf{x}$. By Norets and Takahashi, there exists $\mathbf{v} \in X$ such that $\mathbf{q}\left(\mathbf{v}\right) = \mathbf{x} / \sum_{i=1}^{N} x_i$. Let $c = -W\left(\mathbf{v}\right) + \log \sum_{i=1}^{N} x_i$. Then

$$\mathbf{T}\left(e^{\mathbf{v}+\mathbf{c}}\right) = e^{W(\mathbf{v}+\mathbf{c})}\mathbf{q}\left(\mathbf{v}\right) = \mathbf{q}\left(\mathbf{v}\right)\sum_{i=1}^{N} x_i = \mathbf{x},$$

which establishes that $\mathbf{T}$ is a surjection from $\mathbb{R}^N_+$ to $\mathbb{R}^N_+$.

The next point is to extend $\mathbf{T}$ to $\mathbb{R}_{+0}^{N}$. For $\mathbf{y}$ on the boundary of $\mathbb{R}_{+0}^{N}$, let $z = \{i \in \{1, ..., N\} \,| y_i > 0\}$ index the non-zero components of $\mathbf{y}$. If $z = \emptyset$, then we let $\mathbf{T}(\mathbf{y}) = (0, ..., 0)$. For $z \neq \emptyset$, consider the discrete choice model (7) with choice restricted to z. Let $\tilde{p}_i, i \in z(\mathbf{y})$ be the choice probabilities from this restricted model and let $\tilde{p}_i = 0$ for $i \notin z$. Similarly let $\tilde{W}$ be the expected maximum utility for the restricted model. Define then $\mathbf{T}(\mathbf{y}) = e^{\tilde{W}}(\tilde{p}_1, ..., \tilde{p}_N)$.

The argument that $\mathbf{T}$ is a bijection from $\mathbb{R}_{+}^{N}$ to $\mathbb{R}_{+}^{N}$ may be repeated for each combination of zeros reflected in the set z. Hence the extended function is a bijection from $\mathbb{R}_{+0}^{N}$ to $\mathbb{R}_{+0}^{N}$.

It remains to show that $\mathbf{T}$ is continuous. We will do this by establishing that the values of $\mathbf{T}$ on the boundary of $\mathbb{R}_{+0}^{N}$ are limits of values from sequences in the interior. A limit point of a continuous function is unique, hence for each boundary point we need just consider one sequence converging to that point.

Consider first a sequence $\{\mathbf{y}^n\}_{n=1}^{\infty}$ with $\lim_{n\to\infty} \mathbf{y}^n = (0, ...0)$. As the limit is unique if it exists, consider $\mathbf{y}^n = \mathbf{y}/n$ for some $\mathbf{y} \in \mathbb{R}_{+}^{N}$. Note that $W(\log \mathbf{y}^n) = W(\log \mathbf{y}) - \log n \to -\infty$. Then since $q_i(\mathbf{y}^n)$ are bounded between 0 and 1, $\mathbf{T}(\mathbf{y}^n) \to (0, .., 0)$ as required.

Consider then $\mathbf{y} \in \mathbb{R}_{+}^{N}$, let $z \subset \{1, ..., N\}$ be non-empty and define $y_i^n = y_i$ for $i \in z$ and $y_i^n = y_i/n$ for $i \notin z$. Let F be the cumulative distribution function of the vector of random utility shocks and let F_i be its partial derivatives. Then choice probabilities may be written as

$$q_i(\mathbf{v}) = \int_{-\infty}^{\infty} F_i(u + v_i - v_1, ..., u + v_i - v_N)\, du. \tag{24}$$

As above, let $\tilde{q}$ be the choice probabilities when choice is restricted to z. At no loss of generality, let $z = \left\{1, ..., \tilde{N}\right\}$, where $0 < \tilde{N} < N$. For $i \in z$, use the dominated convergence theorem together with (24) to see that

$$\begin{aligned} \lim_{n\to\infty} q_i(\log \mathbf{y}^n) &= \int_{-\infty}^{\infty} \lim_{n\to\infty} F_i(u + \log y_i^n - \log y_1^n, ..., u + \log y_i^n - \log y_N^n)\, du \\ &= \int_{-\infty}^{\infty} F_i\left(u + \log y_i - \log y_1, ..., u + \log y_i - \log y_{\tilde{N}}, \infty ..., \infty\right) du \\ &= \tilde{q}_i. \end{aligned}$$

These probabilities sum to 1. Hence $\lim_{n\to\infty} q_i(\log \mathbf{y}^n) = 0$ for $i \notin z$.

By dominated convergence,

$$\begin{aligned}
\lim_{n\to\infty} W(\mathbf{y}^n) &= \lim_{n\to\infty}\left(\int_0^\infty (1 - F(u - \log \mathbf{y}^n))\, du - \int_{-\infty}^0 F(u - \log \mathbf{y}^n)\, du\right) \\
&= \int_0^\infty \left(1 - \lim_{n\to\infty} F(u - \log \mathbf{y}^n)\right) du - \int_{-\infty}^0 \lim_{n\to\infty} F(u - \log \mathbf{y}^n)\, du \\
&= \int_0^\infty \left(1 - F\left(u - \log y_1, ..., u - \log y_{\tilde{N}}, \infty, ..., \infty\right)\right) du \\
&\quad - \int_{-\infty}^0 F\left(u - \log y_1, ..., u - \log y_{\tilde{N}}, \infty, ..., \infty\right) du \\
&= \tilde{W}.
\end{aligned}$$

Combining these results, find that $\mathbf{T}(\lim_{n\to\infty} \mathbf{y}^n) = \lim_{n\to\infty} \mathbf{T}(\mathbf{y}^n)$ as required.

Finally, defining $\mathbf{S}(\cdot) = \mathbf{T}^{-1}(\cdot)$ the conclusion follows at once. ■

Proof of proposition 3. We first evaluate $W^*(\mathbf{q})$. If $\mathbf{1} \cdot \mathbf{q} \neq 1$, then

$$\mathbf{q} \cdot (\mathbf{v} + \gamma) - W(\mathbf{v} + \gamma) = \mathbf{q} \cdot \mathbf{v} - W(\mathbf{v}) + (\mathbf{1} \cdot \mathbf{q} - 1)\gamma,$$

which can be made arbitrarily large by changing γ and hence $W^*(\mathbf{q}) = \infty$. Next consider $\mathbf{q}$ with some $q_j < 0$. $W(\mathbf{v})$ decreases towards a lower bound as $v_j \to -\infty$. Then $\mathbf{q} \cdot \mathbf{v} - W(\mathbf{v})$ increases towards $+\infty$ and hence W^* is $+\infty$ outside the unit simplex Δ.

For $\mathbf{q} \in \Delta$, we solve the maximization problem

$$W^*(\mathbf{q}) = \sup_{\mathbf{v}}\{\mathbf{q} \cdot \mathbf{v} - W(\mathbf{v})\}. \tag{25}$$

Note that for any constant k we have $W(\mathbf{v} + k \cdot \mathbf{1}) = k + W(\mathbf{v})$, so that we may normalize $\mathbf{1} \cdot \mathbf{v} = 0$. Maximize then the Lagrangian $\mathbf{q} \cdot \mathbf{v} - W(\mathbf{v}) - \lambda(\mathbf{1} \cdot \mathbf{v})$ with first-order conditions $0 = q_j - \frac{\partial W(\mathbf{v})}{\partial v_j} - \lambda$, which lead to $\lambda = 0$. Then

$$\begin{aligned}
\mathbf{q} &= \nabla_{\mathbf{v}} W(\mathbf{v}) \Leftrightarrow \\
\mathbf{q} e^{W(\mathbf{v})} &= \nabla_{\mathbf{v}} \left(e^{W(\mathbf{v})} \right) = \mathbf{T}(e^{\mathbf{v}}) \Leftrightarrow \\
\mathbf{S}(\mathbf{q}) e^{W(\mathbf{v})} &= e^{\mathbf{v}} \Leftrightarrow \\
\log \mathbf{S}(\mathbf{q}) + W(\mathbf{v}) &= \mathbf{v} \Rightarrow \\
\mathbf{q} \cdot \log \mathbf{S}(\mathbf{q}) + W(\mathbf{v}) &= \mathbf{q} \cdot \mathbf{v}.
\end{aligned}$$

Inserting this into (25) leads to the desired result.

For part (i), let $\mathbf{q}$ be a solution to problem (11). Then, by the homogeneity of $\mathbf{T}$ we have $\mathbf{q} = \frac{1}{\alpha}\mathbf{T}(e^{\mathbf{v}})$, where $\alpha = \sum_{j=1}^{N} T_j(e^{\mathbf{v}})$. Then, by the definition of $\mathbf{S}$ it follows that $\mathbf{S}(\mathbf{q}) = \frac{e^{\mathbf{v}}}{\alpha}$. Replacing the latter expression in Eq. (11) we get

$$\begin{aligned}
W(\mathbf{v}) &= \mathbf{qv} - \mathbf{q} \log(e^{\mathbf{v}}/\alpha), \\
&= \mathbf{qv} - \mathbf{q}(\log e^{\mathbf{v}} + \log \alpha), \\
&= \log\left(\sum_{j=1}^{N} T_j(e^{\mathbf{v}}) \right).
\end{aligned}$$

■

Proof of Proposition 5. Let $\Omega(\mathbf{p}) = -\mathbf{p} \cdot \log \mathbf{S}(\mathbf{p})$ be a generalized entropy. Then, using Proposition 8, the associated Bregman divergence becomes

$$\begin{aligned}
D(\mathbf{p}||\mathbf{q}) &= -\mathbf{q} \cdot \log \mathbf{S}(\mathbf{q}) + \mathbf{p} \cdot \log \mathbf{S}(\mathbf{p}) - (\log \mathbf{S}(\mathbf{q}) + \mathbf{1}) \cdot (\mathbf{p} - \mathbf{q}) \\
&= \mathbf{p} \cdot (\log \mathbf{S}(\mathbf{p}) - \log \mathbf{S}(\mathbf{q})),
\end{aligned}$$

where we have used that $\mathbf{1} \cdot (\mathbf{p} - \mathbf{q}) = 0$.

Convexity of $D(\mathbf{p}||\mathbf{q})$ in $\mathbf{p}$ follows from Proposition 8 or from the fact that it is a Bregman divergence. Clearly, $D(\mathbf{q}||\mathbf{q}) = 0$.

Our information cost is by definition an expected Bregman divergence. We therefore immediately obtain that it is convex in $\mathbf{p}(\cdot)$ holding $\mathbf{p}^0$ constant and that $\kappa_S(\mathbf{p}(\cdot), \mu) = 0$ if action and state are independent, since in that case $\mathbf{p}(\mathbf{V}) = \mathbf{p}^0$. ■

Proof of proposition 6. The Lagrangian for the DM's problem is

$$\Lambda = \mathbb{E}\left(\mathbf{V}\cdot\mathbf{A}\right) - \kappa_{\mathbf{S}}(\mathbf{p},\mu) + \mathbb{E}\left(\gamma\left(\mathbf{V}\right)\left(1-\sum_j p_j\left(\mathbf{V}\right)\right)\right) + \mathbb{E}\left(\sum_j \xi_j\left(\mathbf{V}\right)p_j\left(\mathbf{V}\right)\right),$$

where $\gamma\left(\mathbf{V}\right)$ and $\xi_j\left(\mathbf{V}\right)$ are Lagrange multipliers corresponding to condition (4).

Before we derive the first-order conditions for $p_j\left(\mathbf{v}\right)$ it is useful to note that we may regard the terms $\log\mathbf{S}\left(\mathbf{p}^0\right)$ and $\log\mathbf{S}\left(\mathbf{p}\left(\mathbf{v}\right)\right)$ in the information cost $\kappa_{\mathbf{S}}(\mathbf{p},\mu)$ as constant, since their derivatives cancel out by Proposition 8(iii). Define $\tilde{v}_j = v_j + \xi_j\left(\mathbf{v}\right) + \log S_j\left(\mathbf{p}^0\right)$ and $\tilde{\mathbf{v}} = \left(\tilde{v}_1,...,\tilde{v}_N\right)$. Then the first-order condition for $p_j\left(\mathbf{v}\right)$ is easily found to be

$$\log\mathbf{S}_j\left(\mathbf{p}\left(\mathbf{v}\right)\right) = \tilde{v}_j - \gamma\left(\mathbf{v}\right). \tag{26}$$

This fixes $\mathbf{p}\left(\mathbf{v}\right)$ as a function of $\mathbf{p}^0$ since then

$$\mathbf{p}\left(\mathbf{v}\right) = \mathbf{T}\left(e^{\tilde{\mathbf{v}}}\right)\exp\left(-\gamma\left(\mathbf{v}\right)\right). \tag{27}$$

If some $p_j\left(\mathbf{v}\right) = 0$, then we must have $\tilde{v}_j = -\infty$, which implies that $S_j\left(\mathbf{p}^0\right) = 0$ and the value of $\xi_j\left(\mathbf{v}\right)$ is irrelevant. If $p_j\left(\mathbf{v}\right) > 0$, then $\xi_j\left(\mathbf{v}\right) = 0$. We may then simplify by setting $\xi_j\left(\mathbf{v}\right) = 0$ for all $j, \mathbf{v}$ at no loss of generality, which means that $\tilde{v}_j = v_j + \log S_j\left(\mathbf{p}^0\right)$.

Using that probabilities sum to 1 leads to

$$\exp\left(\gamma\left(\mathbf{v}\right)\right) = \sum_j T_j\left(e^{\tilde{\mathbf{v}}}\right)$$

and hence (i) follows. Item (ii) then follows immediately.

Now substitute (17) back into the objective, using $p_j\left(\mathbf{v}\right)\xi_j\left(\mathbf{v}\right) = 0$, to find that it reduces to

$$\Lambda = \mathbb{E}\gamma\left(\mathbf{V}\right) = \mathbb{E}\log\sum_j T_j\left(e^{\mathbf{V}+\log\mathbf{S}\left(\mathbf{p}^0\right)}\right) \tag{28}$$

We may then use (28) to determine $\mathbf{p}^0$. Now apply Eq. (12) to establish part

(iii) of the proposition. ■

Proof of Proposition 7. Consider a RI-ARUM model with prior $(\mu, \mathcal{V})$, scaled demand $\mathbf{T}$ and choice probabilities $\mathbf{p}(\mathbf{v})$ and let $\mathcal{C}$ be its consideration set. For $i \notin \mathcal{C}$ we have $p_i(\mathbf{v}) = 0$ and $\log S_i(\mathbf{p^0}) = -\infty$ by Proposition 2. Let $\mathbf{c} = \log \mathbf{S}(\mathbf{p^0})$. Then for $i \in \mathcal{C}$ we have

$$\begin{aligned} p_i(\mathbf{v}) &= \frac{T_i(e^{\mathbf{v}+\log \mathbf{S}(\mathbf{p^0})})}{\sum_{j=1}^N T_j(e^{\mathbf{v}+\log \mathbf{S}(\mathbf{p^0})})} \\ &= P\left(v_i + c_i + \epsilon_i = \max_j \{v_j + c_j + \epsilon_j\}\right) \\ &= P\left(v_i + c_i + \epsilon_i = \max_{j \in \mathcal{C}} \{v_j + c_j + \epsilon_j\}\right), \end{aligned}$$

which is an ARUM on $\mathcal{C}$.

To prove the converse, let $\mathbf{q}^0 = \mathbb{E}\mathbf{q}(\mathbf{v}),\ \mathbf{c} = -\log \mathbf{S}(\mathbf{q}_0)$ and consider the RI-ARUM with prior $\left(\mathbf{v} \to \mu\left(\mathbf{v}+\log \mathbf{S}\left(\mathbf{q}^0\right)\right), \mathcal{V} - \log \mathbf{S}\left(\mathbf{q}^0\right)\right)$ and scaled demand $\mathbf{T}$. The RI-ARUM conditional choice probabilities satisfy the first-order condition

$$p_i(\mathbf{v} - \log \mathbf{S}(\mathbf{q}_0)) = \frac{T_i(e^{\mathbf{v}-\log \mathbf{S}\left(\mathbf{q}^0\right)+\log \mathbf{S}(\mathbf{p^0})})}{\sum_{j=1}^N T_j(e^{\mathbf{v}-\log \mathbf{S}(\mathbf{q}^0)+\log \mathbf{S}(\mathbf{p^0})})}$$

with $\mathbf{p}^0(\mathbf{v} - \log \mathbf{S}(\mathbf{q}_0)) = \mathbb{E}\mathbf{p}(\mathbf{v} - \log \mathbf{S}(\mathbf{q}_0))$. By strict convexity of the information cost, the first-order condition uniquely identifies the optimal RI-ARUM conditional choice probabilities. Then $\mathbf{p}(\mathbf{v} - \log \mathbf{S}(\mathbf{q}_0)) = \mathbf{q}(\mathbf{v})$ solves the RI-ARUM maximization problem. ■

B Additional results

Proposition 8 (Properties of the inverse scaled demand function) *For any ARUM discrete choice model satisfying Assumption 1, the corresponding inverse scaled demand $\mathbf{S}(\cdot)$ satisfies:*

(i) $\mathbf{S}$ *is continuous and homogenous of degree 1.*

(ii) $\mathbf{q} \cdot \log \mathbf{S}(\mathbf{q})$ *is convex and strictly convex on the interior of* Δ.

(iii) $\mathbf{S}$ *is differentiable with :*

$$\sum_{i=1}^{N} q_i \frac{\partial \log S_i(\mathbf{q})}{\partial q_k} = 1, k \in \{1, \ldots, N\},$$

where $\mathbf{q}$ *is a probability vector with* $0 < q_i < 1$ *for all* i.

Proof of Proposition 8. Continuity of $\mathbf{S}$ follows from continuity of the partial derivatives of W, which is immediate from the definition. Homogeneity of $\mathbf{S}$ is equivalent to homogeneity of $\mathbf{T}$. Using the homogeneity property of W

$$\mathbf{S}^{-1}(\lambda e^{\mathbf{v}}) = \nabla_{\mathbf{v}}(e^{W(\mathbf{v}+\log\lambda)}) = \lambda\nabla_{\mathbf{v}}(e^{W(\mathbf{v})}) = \lambda\mathbf{S}^{-1}(e^{\mathbf{v}}),$$

which shows that $\mathbf{T}$ and hence $\mathbf{S}$ are homogenous of degree 1.

The requirement that $\sum_{i=1}^{N} q_i \frac{\partial \log S_i(\mathbf{q})}{\partial q_k} = 1$ in the relative interior of the unit simplex Δ may be expressed in matrix notation as

$$(q_1, \ldots, q_N) \cdot J_{\log \mathbf{S}}(\mathbf{q}) = (1, \ldots, 1),$$

where

$$J_{\log \mathbf{S}}(\mathbf{q}) = \left\{ \frac{\partial \log S_i\left(\mathbf{q}\right)}{\partial q_j} \right\}_{i,j=1}^{N}$$

is the Jacobian of $\log \mathbf{S}(\mathbf{q})$.

Defining $\hat{\mathbf{t}} \equiv \log \mathbf{S}(\mathbf{q})$, we have $\mathbf{q} = \mathbf{T}\left(e^{\hat{\mathbf{t}}}\right)$ and hence $W\left(e^{\hat{\mathbf{t}}}\right) = \log(\mathbf{1} \cdot \mathbf{T}(e^{\hat{\mathbf{t}}})) = \log 1 = 0$ by Proposition 3. Noting that $(\log(\mathbf{S}))^{-1}(\hat{\mathbf{t}}) = \mathbf{T}(e^{\hat{\mathbf{t}}})$ the requirement in part (ii) is equivalent to

$$(q_1, \ldots, q_N) = (q_1, \ldots, q_N) \cdot J_{\log \mathbf{S}}(\mathbf{q}) \cdot J_{(\log \mathbf{S})^{-1}}(\hat{\mathbf{t}}) = (1, \ldots, 1) \cdot J_{\mathbf{T}(e^{\hat{\mathbf{t}}})}(\hat{\mathbf{t}}).$$

Now, use the Williams-Daly-Zachary theorem to find that

$$(1, \ldots, 1) \cdot J_{\mathbf{T}(e^{\hat{\mathbf{t}}})}(\hat{\mathbf{t}}) = \nabla_{\hat{\mathbf{t}}}\left(e^{W(\hat{\mathbf{t}})}\right) = e^{W(\tilde{\mathbf{v}})}\left(q_1, \ldots q_N\right) = \left(q_1, \ldots q_N\right).$$

as required.

Convexity of $\mathbf{q} \cdot \log \mathbf{S}(\mathbf{q})$ follows from Proposition 3(ii). To show that $\mathbf{q} \cdot \log \mathbf{S}(\mathbf{q})$ is strictly convex for $\mathbf{q} \in int\Delta$, note first that $e^{W(\mathbf{v})}$ has positive definite Hessian on the set $\left\{\mathbf{v} \in \mathbb{R}^N : \mathbf{1} \cdot \mathbf{v} = 1\right\}$ (Hofbauer and Sandholm, 2002). This Hessian is equal to the Jacobian of $\mathbf{T}\left(e^{\mathbf{v}}\right) = e^{W(\mathbf{v})}\mathbf{q}\left(\mathbf{v}\right)$, which is then positive definite. The inverse of $\mathbf{T}\left(e^{\mathbf{v}}\right)$ is $\log S\left(\mathbf{q}\right)$, which then also has a positive definite Jacobian. But the Hessian of $\mathbf{q} \cdot \log \mathbf{S}(\mathbf{q})$ is $\log S\left(\mathbf{q}\right) + \mathbf{1}$. ■

Corollary 9 *For some option j, and for all $\mathbf{v} \in \mathcal{V}$, let $v_j \leq v_i$ for all $i \neq j$, and assume that the inequality is strict with positive probability. Then $p_j^0 = 0$ (that is, option j is not in the consideration set).*

Proof of corollary 9. Let $\circ$ denote the Hadamard product, i.e. $(a_1, ..., a_N) \circ (b_1, ..., b_N) = (a_1 b_1, ..., a_N b_N)$. Assume, towards a contradiction, that $p_j^0 > 0$. It follows from cyclic monotonicity (Rockafellar, 1970, Thm. 23.5) that $p_j\left(\mathbf{v}\right)$ increases as the utility of other options $i, i \notin j$ decrease. Then

$$p_j^0 = \mathbb{E}\left(\frac{T_j\left(e^{\mathbf{V}} \circ \mathbf{S}\left(\mathbf{p}^0\right)\right)}{\sum\limits_k T_k\left(e^{\mathbf{V}} \circ \mathbf{S}\left(\mathbf{p}^0\right)\right)}\right) \tag{29}$$

$$< \mathbb{E}\left(\frac{T_j\left(e^{V_j}\mathbf{S}\left(\mathbf{p}^0\right)\right)}{\sum\limits_k T_k\left(e^{V_j}\mathbf{S}\left(\mathbf{p}^0\right)\right)}\right) \tag{30}$$

$$= \mathbb{E}\left(\frac{e^{V_j}T_j\left(\mathbf{S}\left(\mathbf{p}^0\right)\right)}{e^{V_j}\sum\limits_k T_k\left(\mathbf{S}\left(\mathbf{p}^0\right)\right)}\right) = \mathbb{E}\left(\frac{p_j^0}{\sum\limits_k p_k^0}\right) = p_j^0. \tag{31}$$

This is a contradiction as desired. ■

Proposition 10 (Convexity of the Bregman divergence) *The Bregman information $\kappa_{\mathbf{S}}\left(\mathbf{p}\left(\cdot\right), \mu\right) = \Omega_{\mathbf{S}}\left(\mathbf{p}^0\right) - \mathbb{E}\Omega_{\mathbf{S}}\left(\mathbf{p}(\mathbf{V})\right)$ associated with a nested logit model is convex. It is strictly convex when the conditional probabilities differ from the unconditional probability p^0.*

Proof of Proposition 10. When Ω is the Shannon entropy, i.e. when the associated ARUM is a multinomial logit model, then the Bregman information is the mutual

(Shannon) information. By Cover and Thomas (2006, Thm 2.7.4), the mutual (Shannon) information is convex as a function of the conditional probability $p\left(\cdot\right)$, and strictly convex when the conditional probabilities differ from the unconditional probability p^0.

Consider now a nested logit model and note that the corresponding generalized entropy may be written $\Omega_{\mathbf{S}}\left(\mathbf{p}\right) = \Omega\left(\Gamma\mathbf{p}\right) + c^\top\mathbf{p}$, where Γ is a matrix whose columns are linearly independent probability vectors. For example, the inverse scaled demand (20) of the two-level nested logit model may be written as

$$\begin{aligned}\log S_i(\mathbf{q}) \quad = \quad & \zeta_{g_i}\log\zeta_{g_i}q_i + \left(1-\zeta_{g_i}\right)\log\left(\sum_{j\in g_i}\left(1-\zeta_{g_j}\right)q_j\right) \\ & -\left(1-\zeta_{g_i}\right)\log\left(1-\zeta_{g_i}\right) - \zeta_{g_i}\log\zeta_{g_i}.\end{aligned}$$

Then the NL generalized entropy is the composition of a linear function and a concave function, plus a linear function. The Bregman information may then be written

$$\kappa_{\mathbf{S}}\left(\mathbf{p}\left(\cdot\right),\mu\right) = \Omega\left(\Gamma\mathbf{p}^0\right) - \mathbb{E}\Omega\left(\Gamma\mathbf{p}(\mathbf{V})\right),$$

which is a composition of a linear function and the convex mutual Shannon information, while the linear terms cancel out. Hence it is convex and strictly convex when the conditional probabilities differ from the unconditional probability p^0. ■