Discussion Papers
Department of Economics
University of Copenhagen

No. 17-26

The Dynamics of Bertrand Price Competition with Cost-Reducing Investments

Mogens Fosgerau, Emerson Melo, André de Palma & Matthew Shum

Øster Farimagsgade 5, Building 26, DK-1353 Copenhagen K., Denmark
Tel.: +45 35 32 30 01 – Fax: +45 35 32 30 00
http://www.econ.ku.dk

ISSN: 1601-2461 (E)

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

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

November 27, 2017

Abstract

This paper establishes a general equivalence between discrete choice and rational inattention models. Matejka and McKay (2015, *AER*) showed that when information costs are modelled using the Shannon entropy, the resulting choice probabilities in the rational inattention model take the multinomial logit form. We show that when information costs are modelled using a class of generalized entropies, then the choice probabilities in *any* rational inattention model are observationally equivalent to some additive random utility discrete choice model and vice versa. Thus any additive random utility model can be given an interpretation in terms of boundedly rational behavior. We provide examples of this equivalence utilizing the nested logit model, an empirically relevant random utility model allowing for flexible substitution possibilities between choices.

JEL codes: D03, C25, D81, E03
Keywords: Rational Inattention; discrete choice; random utility; convex analysis; generalized entropy

1 Motivation

In many situations where agents must 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

[*] First draft: December 22, 2016. We thank Bob Becker, Marcus Berliant, Mark Dean, Federico Echenique, Juan Carlos Escanciano, Filip Matejka, Paulo Natenzon, Henrique de Oliveira, Antonio Rangel, Ryan Webb, and Michael Woodford for useful comments. Lucie Letrouit, Julien Monardo, and Alejandro Robinson Cortes provided research assistance.

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

[‡] Indiana University; emelo@iu.edu

[§] ENS Cachan, Université Paris-Saclay, CREST; andre.depalma@ens-cachan.fr.

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

available options. This idea underlies the theory of rational inattention, 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 rational inattention models. Matějka and McKay (2015) showed that when information costs are modelled using the Shannon entropy, the resulting choice probabilities in the rational inattention model take the familiar multinomial logit form. This is a very appealing result, providing both a microfoundation as well as alternative interpretation for the multinomial logit model.

However, the multinomial logit choice probabilities implied by the rational inattention model based on the Shannon entropy have the "independence of irrelevant alternatives", or IIA property, which is that the ratio of the probabilities of two alternatives does not depend on the utility of a third (irrelevant) alternative. In many empirical contexts, this property is violated: it is rather the norm that some goods are closer substitutes than others. We illustrate this in a motivating example.

Example 1 *Consider a rationally inattentive consumer facing a choice between an apple, a mango, and a piece of cheesecake, say. A priori, the consumer does not know his/her payoff associated with each good but considers the valuation vector* $\mathbf{V}$ *to be random with a known distribution. Conditional on the value of* $\mathbf{V}$*, the rationally inattentive consumer chooses among the three goods with probabilities* $\mathbf{p}(\mathbf{V}) = (p_1(\mathbf{V}), p_2(\mathbf{V}), p_3(\mathbf{V}))$*. Consider two potential realizations of* $\mathbf{V}$*:* $\mathbf{v}^1 = (1, 1, 1)$ *and* $\mathbf{v}^2 = (0.9, 1, 1)$ *and assume that we observe corresponding choice probabilities* $\mathbf{p}(\mathbf{v}^1) = (0.27, 0.27, 0.46)$ *and* $\mathbf{p}(\mathbf{v}^2) = (0.20, 0.33, 0.47)$*. These probabilities reflect that a decrease in the payoff of apple causes the consumer to substitute mostly towards mango and only a little towards cheesecake, as one might reasonably expect. However, such an outcome cannot be predicted by the rational inattention model with Shannon information cost, since the IIA property is violated by* $0.27/0.46 \neq 0.33/0.47$. ■

The root of the problem is that the use of the Shannon entropy as a measure of the cost of processing information embodies an important and strong assumption of *symmetry*: the Shannon entropy is symmetric, or invariant to permutations 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). 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. Our generalized entropy allows patterns such as those in the example above to be accommodated in a rational inattention model, which contributes to making rational inattention models empirically relevant. In fact, we show that a rational inattention model can yield the same choice probability system as *any* additive random utility model, depending on the choice of information cost; this includes specifications such as nested logit, multinomial probit, and so on, that are often employed in empirical work. At the same time, this equivalence permits the interpretation of any additive random utility discrete choice model as arising from boundedly rational behavior.

We introduce a class of *Generalized Entropy Rational Inattention* (GERI) models. In a GERI model, the Shannon entropy is replaced by an information cost defined as the convex conjugate to the surplus function of an additive random utility model ("ARUM"). This generalizes the Shannon entropy since the Shannon entropy arises when the ARUM is the multinomial logit model. As we will show, a GERI model exists corresponding to any ARUM, implying that rationally inattentive behavior can generate choice probabilities which can be context dependent and violate the IIA property, as in Example 1 above.

Related literature. Besides the papers already mentioned above, the main equivalence result in this paper is related to several strands of literature. First, this paper contributes to the growing literature on rational inattention with more general cost functions. Caplin et al. (2017) provides a behavioral characterization of Shannon entropy. In a dynamic setting, Hébert and Woodford (2016) also consider generalizations of the information cost. Morris and Yang (2016) uses ideas from global games to develop a rational inattention framework where the cost of processing information satisfies natural properties such as convexity and continuity. Second, the results in this paper are related to the literature on perturbed utility models. In this context, Anderson et al. (1988) uses the Shannon entropy to derive the multinomial logit model. This observation is generalized by Hofbauer and Sandholm (2002), who show 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. We contribute to that literature by providing an explicit characterization of the perturbation term corresponding to general ARUM.

Finally, Joo (2017) and Caplin, Leahy and Matějka (2016) consider empirical applications in the rational inattention paradigm, using the Shannon/multinomial logit framework. The results in this paper may enable researchers to apply rational inattention models far beyond the multinomial logit setting, as they imply that choice behavior emerging from *any* ARUM model may be explained by rationally inattentive behavior.

Notation: Throughout this paper, for vectors $\mathbf{a}$ and $\mathbf{b}$, we use the notation $\mathbf{a} \cdot \mathbf{b}$ to denote the vector scalar product $\sum_i a_i b_i$. Δ denotes the unit simplex in $\mathbb{R}^N$.

Layout. Section 2 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 about them. Section 3 introduces the rational inattention model. We show how generalized entropy can be used to define the information cost in the rational inattention model, leading to the class of GERI 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 GERI models. Section 4 discusses the specific case of the empirically-relevant nested logit model, and shows how rationally inattentive behavior can generate choice probabilities which can exhibit non-IIA behavior, as in Example 1. Three examples illustrate both differences and similarities of GERI models relative to the Shannon-entropic model. Section 5 concludes. All proofs are in the Appendix.

2 Random utility models and generalized entropy

Consider a decision-maker (DM) making a discrete choice among a set of $i = 1, \ldots, N$ options. The utility of option i is

$$u_i = v_i + \epsilon_i, \tag{1}$$

where $\mathbf{v} = (v_1, \ldots, v_N)$ is deterministic and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)$ is a vector of random utility shocks. This is the classic ARUM framework pioneered by McFadden (1978).

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 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 utility maximizing DM 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

$$W(\mathbf{v}) = \mathbb{E}_{\boldsymbol{\epsilon}}(\max_j [v_j + \epsilon_j]). \tag{2}$$

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

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

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).

We define a vector-valued function $\mathbf{H}(\cdot) = (H_1(\cdot), ..., H_N(\cdot)) : \mathbb{R}_+^N \mapsto \mathbb{R}_+^N$ as the gradient of the exponentiated surplus, i.e.

$$\mathbf{H}(e^{\mathbf{v}}) = \nabla_{\mathbf{v}} \left(e^{W(\mathbf{v})}\right). \tag{3}$$

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 in terms of the function $\mathbf{H}$ as:[2]

$$q_i(\mathbf{v}) = \frac{H_i(e^{\mathbf{v}})}{\sum_{j=1}^N H_j(e^{\mathbf{v}})}, \quad \text{for } i = 1, \dots, N. \tag{4}$$

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 $H_i(e^{\mathbf{v}}) = e^{v_i}$. Thus, Eq. (4) becomes the familiar multinomial logit

[1] The convexity of $W(\cdot)$ follows from the convexity of the max function. Differentiability follows from the absolute continuity of $\boldsymbol{\epsilon}$.

[2] By direct differentiation of (3), and applying the Williams-Daly-Zachary theorem, we have $q_i(\mathbf{v}) = H_i(e^{W(\mathbf{v})})/e^{W(\mathbf{v})}$ for all i. Imposing $\sum_i q_i(\mathbf{v}) = 1$ we have $\sum_i H_i(e^{W(\mathbf{v})}) = e^{W(\mathbf{v})}$.

choice formula: $q_i(\mathbf{v}) = e^{v_i} / \sum_j e^{v_j}$.

The function $\mathbf{H}(\cdot)$ thus defined has domain $\mathbb{R}^N_+$ and also range $\mathbb{R}^N_+$. As we will see below, rationally inattentive behavior can lead to zero choice probabilities, so it is important to extend the domain and range of this function to $\mathbb{R}^N_{+0}$, to allow for zero probabilities. The following proposition allows us to do that and assures that the extended $\mathbf{H}$ has an inverse.

Proposition 2 (Invertibility) *The function* $\mathbf{H}(\cdot)$ *has to a continuous extension to* $\mathbf{H} : \mathbb{R}^N_{+0} \to \mathbb{R}^N_{+0}$ *that is surjective, injective and hence globally invertible.*

Having established that function $\mathbf{H}$ is a bijection from $\mathbb{R}^N_{+0}$ to $\mathbb{R}^N_{+0}$, we can define a function $\mathbf{S}(\cdot)$ as the inverse of $\mathbf{H}(\cdot)$,

$$\mathbf{S}(\cdot) = \mathbf{H}^{-1}(\cdot). \tag{5}$$

The proof of Proposition 2 leads to the following Corollary.

Corollary 3 $S_i(q) = 0 \Leftrightarrow q_i = 0.$

In what follows, we will refer to $\mathbf{S}$ as a *generator* function, and $\mathbf{H}$ as an *inverse generator* function. The intuition behind the interpretation of $\mathbf{S}$ as a generator comes from the fact that is a close relationship between the function $\mathbf{S}(\cdot)$ and the surplus function $W(\cdot)$ of the corresponding discrete choice model: as the next proposition establishes, the surplus function $W(\cdot)$ and the generator function $\mathbf{S}(\cdot)$ are related in terms of *convex conjugate duality*.[3]

Proposition 4 (Convexity properties and generalized entropy functions) *Let assumption 1 hold. Then:*

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

$$W(\mathbf{v}) = \log\left(\sum_{i=1}^{N} H_i(e^{\mathbf{v}})\right). \tag{6}$$

[3] 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. Roughly speaking, the gradients (or sub-gradients, in case of non-differentiability) of $g(\mathbf{x})$ and $g^*(\mathbf{y})$ are inverse mappings to each other. For more details see (Rockafellar, 1970, ch. 12)

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

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

where $\mathbf{S}(\cdot)$ is a generator function defined in (5). We call the negative convex conjugate $-W^*(\cdot)$ a ***generalized entropy****.*

(iii) *The surplus function $W(\mathbf{v})$ is the convex conjugate of $W^*(\mathbf{q})$, that is*

$$W(\mathbf{v}) = \max_{\mathbf{q} \in \Delta} \{\mathbf{q} \cdot \mathbf{v} - W^*(\mathbf{q})\} \tag{7}$$

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

Parts (i) and (ii) of this proposition establish a specific structure of the surplus function W and its convex conjugate W^*; this is new in the literature on random utility models, and may be of independent interest. In particular, this result contributes 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)).

We will use this structure to define a class of *generalized entropies*. To see how this works, consider first the multinomial logit model. In this case, $\mathbf{H}$ is the identity, implying that its inverse, the corresponding generator $\mathbf{S}(\mathbf{q}) = \mathbf{q}$ is also just the identity. Then by Proposition 4(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.

Generalizing from this, $-W^*$, the negative convex conjugate of the surplus W of any ARUM may be viewed as a *generalized entropy*.[4] Proposition 4(ii) shows how the generalized entropy may be expressed in terms of the function $\mathbf{S}$ as $-W^*(\mathbf{q}) = -\mathbf{q} \cdot \log \mathbf{S}(\mathbf{q})$. In contrast to the Shannon entropy, the generalized entropy will typically not be invariant to permutations of its components: exchanging q_i and q_j, for instance, will change $\mathbf{S}(\mathbf{q})$ and thereby $W^*(\mathbf{q})$. This non-symmetry of the generalized entropy, it will turn out, is crucial for generating choice behavior that violates the IIA property, as in Example 1 above.

[4] We show in Appendix B that generalized entropies share some desirable properties with the Shannon entropy.

Proposition 4(iii) provides an alternative representation of the surplus function of a random utility model, in addition to Eq. (2). It illustrates a close connection between $-W^*(\mathbf{q})$ and the joint distribution of $\boldsymbol{\epsilon}$, the random utility shocks, which aids interpretation of the generalized entropy. Specifically, Eq. (2) 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 (7), we obtain an alternative expression for the generalized entropy, as a choice probability-weighted sum of expectations of the utility shocks $\boldsymbol{\epsilon}$:[5]

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

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

We conclude this section by listing some properties of the generator $\mathbf{S}(\cdot)$, which will be important in what follows.

Proposition 5 (Properties of the generator functions) *Let assumption 1 hold. Then the vector valued-function $\mathbf{S}(\cdot)$ defined by (5) satisfies the following conditions:*

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

(ii) $\mathbf{q} \cdot \log \mathbf{S}(\mathbf{q})$ *is convex.*

(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.

[5] See Chiong et al. (2016). Additionally, we conjecture that, up to constant, $\log S_i(\mathbf{q})$ is equal to $-\mathbb{E}[\epsilon_i | u_i \geq u_j, j \neq i]$ for $i = 1, \ldots, N$. In fact, for the multinomial logit case, corresponding to $\mathbf{S}(\mathbf{q}) = \mathbf{q}$, McFadden (1978) showed that $\gamma - \log q_i = \mathbb{E}[\epsilon_i | u_i \geq u_j, j \neq i]$, where γ is Euler's constant.

3 Rational inattention

We now introduce the rational inattention model. The decision maker is again 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 in contrast to the ARUM, the vector of payoffs is unobserved by the 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 $\mu(\mathbf{v}) = \mathbb{P}(\mathbf{V} = \mathbf{v})$.

The DM's choice is represented as a random action $\mathbf{A}$ that is a canonical unit vector in $\mathbb{R}^N$. The payoff resulting from the action is $\mathbf{V} \cdot \mathbf{A}$, namely the value of the entry in $\mathbf{V}$ indicated by the action $\mathbf{A}$. The problem of the rationally inattentive DM is to choose the conditional distribution $\mathbb{P}(\mathbf{A}|\mathbf{V})$, balancing the expected payoff against the cost of information.

Denote an action by i and write $p_i(\mathbf{v})$ as shorthand for $\mathbb{P}\left(\mathbf{A} = i|\mathbf{V} = \mathbf{v}\right)$. Denote also the vector of choice probabilities conditional on $\mathbf{V} = \mathbf{v}$ by $\mathbf{p}(\mathbf{v}) = (p_1(\mathbf{v}), \ldots, p_N(\mathbf{v}))$, and let $\mathbf{p}(\cdot) = \{\mathbf{p}(\mathbf{v})\}_{\mathbf{v}\in\mathcal{V}}$ denote the collection of conditional probabilities. The DM's strategy is a solution to the following variational problem:

$$\max_{\mathbf{p}(\cdot)} \left\{\mathbb{E}\left(\mathbf{V} \cdot \mathbf{A}\right) - \text{information cost}\right\}. \tag{8}$$

In this model, and in line with Sims (2003), agents directly choose their choice probabilities $\mathbf{p}(\cdot)$.[6] In Matějka and McKay (2015, Lemma 1), the connection between state, the payoff vector, and action is derived from a more fundamental 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. Due to the symmetry of the Shannon entropy, there is no loss of generality from identifying signals with choices, which provides a rationale for the model where agents choose choice probabilities directly. In our case, the generalized entropy is not symmetric in general, implying that this result no longer holds. As we have discussed, however, the asymmetry is an important virtue of our approach as it allows us to generate plausible choice patterns in contexts where the IIA property of the multinomial logit model does not hold.

In Section 3.1 we review results from the existing literature connecting the rational inattention model with the multinomial logit. Then in Section 3.2 we intro-

[6] It is worth noting that this approach has been applied in many contexts. See for instance, Sims (2010) and references therein.

duce generalized entropy to the problem, leading to a class of *Generalized Entropy Rational Inattention* (GERI) models. This connects the rational inattention model to general ARUM, and Section 3.3 contains the main result of this paper, which establishes an observational equivalence between discrete choice and rational inattention models.

3.1 Shannon entropy and multinomial logit

The key element in the program above is the cost of information. Much of the previous literature has utilized the mutual Shannon information between payoffs $\mathbf{V}$ and the actions $\mathbf{A}$ to measure the information costs. Denote the Shannon entropy by $\Omega(\mathbf{q}) = -\mathbf{q} \cdot \log \mathbf{q}$. Denote also the unconditional choice probabilities by $p_i^0 = \mathbb{E} p_i(\mathbf{V})$ and $\mathbf{p}^0 = (p_1^0, \ldots, p_N^0)$. Then the mutual Shannon information between $\mathbf{V}$ and $\mathbf{A}$ is

$$\begin{aligned} \kappa(\mathbf{p}(\cdot), \mu) &= \Omega(\mathbb{E}(\mathbf{p}(\mathbf{V}))) - \mathbb{E}(\Omega(\mathbf{p}(\mathbf{V}))) && (9) \\ &= -\sum_{i=1}^{N} p_i^0 \log p_i^0 + \sum_{\mathbf{v} \in \mathcal{V}} \left(\sum_{i=1}^{N} p_i(\mathbf{v}) \log p_i(\mathbf{v}) \right) \mu(\mathbf{v}). && (10) \end{aligned}$$

Accordingly, we can specify the information cost as $\lambda \kappa(\mathbf{p}, \mu)$ where $\lambda > 0$ is the unit cost of information. As the distribution of payoffs is unspecified, we may take $\lambda = 1$ at no loss of generality. The choice strategy of the rationally inattentive DM is the distribution of the action $\mathbf{A}$ conditional on the payoff $\mathbf{V}$ that maximizes the expected payoff less the cost of information, which is the solution to the optimization problem

$$\max_{\mathbf{p}(\cdot)} \{\mathbb{E}(\mathbf{V} \cdot \mathbf{A}) - \kappa(\mathbf{p}(\cdot), \mu)\} \tag{11}$$

subject to

$$p_i(\mathbf{v}) \geq 0 \text{ for all } i, \quad \sum_{i=1}^{N} p_i(\mathbf{v}) = 1. \tag{12}$$

Solving this, the DM finds conditional choice probabilities

$$p_i(\mathbf{v}) = \frac{p_i^0 e^{v_i}}{\sum_{j=1}^{N} p_j^0 e^{v_j}} \quad \text{for } i = 1, \ldots, N, \tag{13}$$

that satisfy $p_i^0 = \mathbb{E} p_i(\mathbf{V})$.

Under the convention that $\log 0 = -\infty$ and $\exp(-\infty) = 0$, we may rewrite (13) as

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

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 $\log \mathbf{p}^0$. For options that are not in the consideration set, the shifted payoff is $v_i = -\infty$. From the perspective of the multinomial logit model these options have zero probability of maximizing the random utility (1) and they have effectively been eliminated from the model.

3.2 The Generalized Entropy Rational Inattention (GERI) model

In this paper we generalize the preceding equivalence result between rational inattention and multinomial logit. We begin by generalizing the rational inattention framework described above, using generalized entropy in place of the Shannon entropy. Specifically, we let $\mathbf{S}$ be the generator 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. We define accordingly a general information cost by

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

A *Generalized Entropy Rational Inattention* (GERI) model describes 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

$$\max_{\mathbf{p}(\cdot)} \{\mathbb{E}(\mathbf{V} \cdot \mathbf{A}) - \kappa_{\mathbf{S}}(\mathbf{p}(\cdot), \mu)\}. \qquad (15)$$

The following proposition characterizes the solution to the GERI model.

Proposition 6 *Let* $\mathbf{p}(\cdot), \mathbf{p}^0$ *be the solution to the GERI model. Then*

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

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

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

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

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

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

Part (i) of the proposition shows that the solution of the GERI model involves a fixed point problem; in what follows, we assume that a solution exists.[7] Part (iii) illustrates the close connection between convex analysis and the GERI problem. To see this, note that the GERI information cost may be written as

$$\kappa_{\mathbf{S}}(\mathbf{p}(\cdot),\mu)=-W^*(\mathbf{p}^0)+\mathbb{E}W^*(\mathbf{p}(\mathbf{V})). \tag{18}$$

Hence, given $\mathbf{p}^0$, the conditional choice probabilities $\mathbf{p}(\mathbf{v})$ can be generated, for each $\mathbf{v}\in\mathcal{V}$, by the problem

$$\max_{\mathbf{p}(\mathbf{v})\in\Delta}\left\{\mathbf{p}(\mathbf{v})\cdot(\mathbf{v}+\log\mathbf{S}(\mathbf{p}^0))-W^*(\mathbf{p}(\mathbf{v}))\right\}, \tag{19}$$

the optimized value of which, by Proposition 4(iii), is

$$W(\mathbf{v}+\log\mathbf{S}(\mathbf{p}^0)),\quad \text{for each } \mathbf{v}\in\mathcal{V} \tag{20}$$

corresponding to Proposition 6(iii).

3.2.1 Zero choice probabilities and the consideration set

It is an important feature of the rational inattention model that some p_i^0 may be zero, in which case the corresponding $p_i\left(\mathbf{v}\right)$ are also zero.[8] Then the rational inattention model implies the formation of a *consideration set*, comprising those options that

[7] It is worth noticing that in general there is not guarantee that a solution is unique. Matějka and McKay (2015, p. 284) show that for the logit case, the uniqueness of a solution may fail in environments where the values of $\mathbf{v}$ comove across states in a very rigid way. Thus without additional restrictions on $\mathcal{V}$, the uniqueness of a solution to GERI is not assured .

[8] To see this, consider the solution to the GERI problem given in Eq. (17) and define $\tilde{\mathbf{v}}=\mathbf{v}+\log\mathbf{S}(\mathbf{p}^0)$ for some $\mathbf{v}\in\mathcal{V}$. Let $p_i^0=0$. Then Corollary 3 implies that $\log S_i(\mathbf{p}^0)=-\infty$, or equivalently, $\tilde{v}_i=-\infty$ and hence $p_i(\mathbf{v})=0$.

have strictly positive probability of being chosen (c.f. Matějka and McKay, 2015; Caplin, Dean and Leahy, 2016).

Definition 7 *The consideration set of a GERI model is the set of options that have positive unconditional choice probabilities*

$$\mathcal{C} = \left\{ i | p_i^0 > 0 \right\}.$$

The restriction of a vector $\mathbf{v}$ *to* $\mathcal{C}$ *is denoted* $\mathbf{v}_{|\mathcal{C}}$*.*

Because of the possibility of zero choice probability for some options, GERI models can also generate failures of the "regularity" property, i.e. that adding an option to a choice set leads to an increase in the choice probability for one of the original choice options.[9] Section 4.3 provides an example.

While Proposition 6 does not explicitly characterize the consideration set emerging from a GERI model, the following corollary describes one important feature that it has, namely that it excludes options that offer the lowest utility in all states of the world.

Corollary 8 *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).*[10]

3.3 Equivalence between discrete choice and rational inattention

We now establish the central result of this paper, namely the equivalence between additive random utility discrete choice models and rational inattention models. In particular, we show that the choice probabilities generated by a GERI model lead to the same choice probabilities as a corresponding ARUM and vice versa. From the expressions for the choice probabilities in a GERI model in (17) and in an ARUM in (4) 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 GERI model.

Two technical issues arise, that we must deal with. First, we need to fix the prior $(\mu, \mathcal{V})$, which is part of the GERI model but not of the random utility model.

[9] Matějka and McKay (2015, pp. 293ff)

[10] For the special case of the Shannon entropy (when $\mathbf{S}$ is the identity), the result can be strengthened even further. Corollary 10 in the Appendix shows that in that case, an option that is dominated by another option in all states of the world will not be in the consideration set.

Second, we need to take care of the fact that the GERI model allows some options to have zero unconditional choice probabilities p_i^0, while choice probabilities are necessarily positive in the standard ARUM. We then have the following proposition.

Proposition 9 *For every ARUM with choice probabilities* $\mathbf{q}\left(\mathbf{v}\right)$ *and generator* $\mathbf{S}$ *and given a prior* $\left(\mu, \mathcal{V}\right)$*, there is a location shift vector* $\mathbf{c}$ *such that the GERI model with prior* $\left(\mathbf{v} \to \mu\left(\mathbf{v}-\mathbf{c}\right), \mathcal{V}+\mathbf{c}\right)$ *and generator* $\mathbf{S}$ *has choice probabilities* $\mathbf{p}$ *that satisfy* $\mathbf{p}\left(\mathbf{v}-\mathbf{c}\right) = \mathbf{q}\left(\mathbf{v}\right)$ *for all* $\mathbf{v} \in \mathcal{V}$*.*

Conversely, for every GERI model with prior $\left(\mu, \mathcal{V}\right)$*, generator* $\mathbf{S}$ *and choice probabilities* $\mathbf{p}\left(\mathbf{v}\right)$ *there is an ARUM defined on the consideration set of the GERI model that yields the same choice probabilities for all* $\mathbf{v} \in \mathcal{V}$*.*

In what follows, we apply this proposition to study a GERI model in which the choice probabilities are equivalent to those from a nested logit discrete choice model, which is a frequently-used model in empirical applications.

4 Example: The nested logit GERI model

From an applied point of view, an important implication of Proposition 9 is that it allows us to formulate rational inattention models that have complex substitution patterns, going beyond the multinomial logit case. In this example, we consider a GERI model with an 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 an 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 IIA criticism.[11]

We partition the set of options $i \in \{1, \ldots, N\}$ 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

$$q_i(\mathbf{v}) = \frac{e^{v_i/\zeta_{g_i}}}{\sum_{j \in g_i} e^{v_j/\zeta_{g_i}}} \cdot \frac{e^{\zeta_{g_i} \log\left(\sum_{j \in g_i} e^{v_j/\zeta_{g_i}}\right)}}{\sum_{\text{all nests } g} e^{\zeta_g \log\left(\sum_{j \in g} e^{v_j/\zeta_g}\right)}}. \tag{21}$$

[11] See, for instance, Maddala (1986, Chap. 2), and Anderson et al. (1996).

The generator $\mathbf{S}$ corresponding to a nested logit model is

$$S_i(\mathbf{q}) = q_i^{\zeta_{g_i}} \left(\sum_{j \in g_i} q_j \right)^{1-\zeta_{g_i}} . \tag{22}$$

Applying Proposition 9, the nested logit choice probabilities (21) are the same as those from a GERI model with valuations

$$v_i - \zeta_{g_i} \log p_i^0 - (1 - \zeta_{g_i}) \log \left(\sum_{j \in g_i} p_j^0 \right), \quad i \in \{1, \ldots, n\} . \tag{23}$$

The generator $\mathbf{S}$ for the nested logit model in Eq. (22) has several interesting features, relative to the Shannon entropy. First, Eq. (22) allows us to write the generalized entropy $\Omega_{\mathbf{S}}(\mathbf{p})$ as

$$\Omega_{\mathbf{S}}(\mathbf{p}) = -\sum_{i=1}^{N} \zeta_{g_i} p_i \log p_i - \sum_{i=1}^{N} (1 - \zeta_{g_i}) p_i \log \left(\sum_{j \in g_i} p_j \right) . \tag{24}$$

The first term in Eq (24) captures the Shannon entropy within nests, whereas the second term captures the information between nests. According to this, we may interpret Eq. (24) as an augmented version of Shannon entropy. It is also apparent from (24) 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.

Using this model, we consider three examples, emphasizing both differences and similarities of the GERI model vis-a-vis the Shannon entropy rational inattention model.

4.1 Example 1: mango-apple-cheesecake continued

We return to the apple-mango-cheesecake example from earlier. For these three products, we consider a model with two nests, in which goods 1 and 2 (apple and mango) are placed in a "fruits" nest g_1, and good 3 (cheesecake) is by itself in a second nest g_2. For the nesting parameters, we choose $\zeta_{g_1} = 0.2$ and $\zeta_{g_2} = 0.8$.

Recall that there are two possible outcomes for the valuations: $\mathbf{v}_1 = (1, 1, 1)$ or $\mathbf{v}_2 = (0.9, 1, 1)$. Outcome $\mathbf{v}_1$ and $\mathbf{v}_2$ are equally likely a priori. Starting with a GERI model, we note that the valuations for apple are never higher than for mango and cheese cake and strictly lower with positive probability, and hence by Corollary 8, the DM will never choose apple ($p_1^0 = 0$). Since the valuations for mango and cheesecake are identical across the two states of the world, so are the GERI choice probabilities; they are $(0.0, 0.66, 0.34)$ in both states.[12] The corresponding location shift vector is $\mathbf{c^1} = \log \mathbf{S}(\mathbf{p}^0) = (-\infty, -0.41, -1.09)$. This can be rationalized by the same nested logit discrete choice model with $\mathbf{v}_1 + \mathbf{c^1} = \mathbf{v}_2 + \mathbf{c^1} = (-\infty, 0.59, -0.09)$.

Conversely, we can start with the nested logit model, with valuations $\mathbf{v}_1 = (1, 1, 1)$ and $\mathbf{v}_2 = (0.9, 1, 1)$. The choice probabilities are $(0.27, 0.27, 0.46)$ under $\mathbf{v}_1$ and $(0.20, 0.32, 0.48)$ under $\mathbf{v}_2$. Under equal priors on $\mathbf{v}_1$ and $\mathbf{v}_2$, the unconditional choice probabilities are $\mathbf{p}^0 = (0.23, 0.30, 0.47)$, leading to a location shift vector $\mathbf{c^2} = \log \mathbf{S}(\mathbf{p}^0) = (-0.80, -0.75, -0.75)$. This can be generated from a GERI model with valuations $\mathbf{v}_1 - \mathbf{c^2} = (1.80, 1.75, 1.75)$ and $\mathbf{v}_2 - \mathbf{c^2} = (1.70, 1.75, 1.75)$.

We contrast these results with the RI model using the Shannon entropy, as in Matějka and McKay (2015). Starting with the GERI model with equiprobable valuations $\mathbf{v}_1 = (1, 1, 1)$ and $\mathbf{v}_2 = (0.9, 1, 1)$, the GERI choice probabilities are $(0, 0.5, 0.5)$, corresponding to a location shift vector $\mathbf{c^3} = (\log 0 = -\infty, \log(0.5), \log(0.5))$. This can be rationalized by a multinomial logit model with valuations $\mathbf{v}_1 + \mathbf{c^3} = \mathbf{v}_2 + \mathbf{c^3} = (-\infty, 0.31, 0.31)$.

Conversely, starting with a random utility model with equiprobable valuations $\mathbf{v}_1 = (1, 1, 1)$ and $\mathbf{v}_2 = (0.9, 1, 1)$, the choice probabilities are $(0.33, 0.33, 0.33)$ under $\mathbf{v}_1$ but $(0.31, 0.34, 0.34)$ under $\mathbf{v}_2$: note that the symmetry properties of the Shannon entropy imply that the decrease in attractiveness of the apple under $\tilde{v}_2$ leads to equal substitution towards mango and cheesecake. These two sets of choice probabilities can be rationalized in a Shannon entropy RI model with valuation vector equal to, respectively, $(2.13, 2.08, 2.08)$ and $(2.03, 2.08, 2.08)$.

[12] Computing these choice probabilities requires solving the fixed point equation (16).

4.2 Example 2: Efficiency comparison of nested logit vs. multinomial logit

In our second example, we expand the number of goods and do further comparison of the choice probabilities from a GERI model with a nested logit information cost, vs. the RI model with the Shannon entropy. There are five options, in which the valuations $\mathbf{v} = (v_1, v_2, \ldots, v_5)'$ are drawn i.i.d. uniformly from the unit interval. We assume that options (1,2,3) are in one nest, and options (4,5) are in a second nest. With this specification, all five options are *a priori* identical, and have equal probability of being the option with the highest valuation. Hence, any asymmetry in the choice probabilities reflects the underlying asymmetry in the information cost.

In Table 1, we report the average choice probability for each option according to two specifications of the nested logit cost function. In the upper panel, we set $\zeta_1 = \zeta_2 = 1$, corresponding to the multinomial logit model. In the lower panel, we set $\zeta_1 = \zeta_2 = 0.5$.

Choice probs:	Option 1	Option 2	Option 3	Option 4	Option 5
	Multinomial logit: $\zeta_1 = 1,\ \zeta_2 = 1$				
Avg:	0.200	0.200	0.200	0.200	0.200
Median:	0.194	0.194	0.194	0.194	0.194
Std:	0.060	0.060	0.060	0.060	0.060
Overall efficiency:	Pr(Choosing the best option) = 0.283				
	Nested logit: $\zeta_1 = 0.5,\ \zeta_2 = 0.5$				
Avg:	0.221	0.221	0.221	0.169	0.169
Median:	0.200	0.200	0.200	0.157	0.157
Std:	0.116	0.116	0.116	0.081	0.081
Overall efficiency:	Pr(Choosing the best option) = 0.355				

Table 1: Choice Probabilities in GERI model: Nested Logit vs. Multinomial Logit

As expected, we see that the average choice probabilities are identically equal to 0.2 across all five options in the multinomial logit case. As we remarked before, this reflects the feature of the Shannon-based information cost ($S_i(\mathbf{p}) = p_i$) in which information costs are separable across all five options.[13] In the nested logit case, in contrast, we see that choice probabilities are higher for the options 1,2 and

[13] In the nested logit case, we obtained the unconditional distribution by iterating over the fixed point relation $\mathbf{p}^0 = \mathbb{E}\mathbf{p}(\mathbf{V})$, starting from the multinomial logit distribution.

3, which constitute the larger nest, and smaller for options 4,5 which constitute the smaller nest. The choice probabilities are identical within nests.

Moreover, the performance of the two models is surprisingly different. Under the multinomial logit specification, the overall efficiency – defined as the average probability of choosing the option with the highest valuation – is 28%. The overall efficiency for the nested logit is substantially higher, being over 35%. Then it makes a substantial difference for a DM to be processing information using the nested logit information cost rather than the Shannon information cost function.

4.3 Example 3: Consideration sets and failure of regularity

Finally, we consider a fully solved out example illustrating the possibility of zero unconditional choice probabilities and failure of regularity, which can occur in the rational inattention framework but not in the ARUM, and represents an important point of difference between the two models. Matějka and McKay (2015, pp. 293ff) have demonstrated that failures of regularity can occur in the RI model under Shannon entropy. We show that such failures also occur in a GERI model, in particular for the nested logit information cost.

Consider a setting with four choice options. Table 2 lists the valuation vectors for these four options in the three equiprobable states of the world. We consider both the Shannon and GERI-nested logit models. For the nested logit specification, we assume that nest 1 consists of options (1,2) with nesting parameter $\zeta_1 = 0.7$, and nest 2 consists of options (3,4) with parameter $\zeta_2 = 0.8$.

State:	$\mathbf{v}^1$	$\mathbf{v}^2$	$\mathbf{v}^3$
Option 1	2	3	3
Option 2	1	2	2
Option 3	3	1	3
Option 4	2	4	2

Table 2: Valuation vectors in Example 2

For each model, we compute the unconditional probabilities, solving the fixed-point equation (16) first for the choice set $\{1, 2, 3\}$, and then for the expanded choice set $\{1, 2, 3, 4\}$.

Table 3 shows the unconditional choice probabilities. The Shannon and GERI-nested logit specifications yield similar results. With the smaller set of options, only options 1,2 are chosen with positive probabilities. When option 4 is added,

Model:	Shannon	Shannon	GERI-nested logit	GERI-nested logit
Choice set:	$\{1,2,3\}$	$\{1,2,3,4\}$	$\{1,2,3\}$	$\{1,2,3,4\}$
p_1^0	0.71	0.00	0.71	0.00
p_2^0	0.00	0.00	0.00	0.00
p_3^0	0.29	0.51	0.29	0.57
p_4^0	—	0.49	—	0.43
Optimized surplus: $\mathbb{E}W(\mathbf{V}+\log\mathbf{S}(\mathbf{p}^0))$	2.705	2.865	4.222	6.032

Table 3: Unconditional probabilities for Example 3

however, option 1 drops out of the consideration set, and only options 3,4 are chosen with positive probability. This is a failure of the regularity property, as the addition of option 4 *increases* the choice probability for choice 1.

The underlying mechanism is that the addition of option 4 allows the DM to form an effective "hedge" in conjunction with option 3. Option 3 yields a low payoff in state $\mathbf{v}^2$, but then option 4 yields a high payoff; option 3 yields high payoffs in the other states.

Finally, note that with the expanded choice set, option 2 is chosen with zero probability, even though it is not inferior in all states. This shows that the characterization of consideration sets in Corollary 8 is not exhaustive.

5 Summary

The central result in this paper is the observational equivalence between an additive random utility discrete choice model and a corresponding Generalized Entropy Rational Inattention (GERI) model. Thus the choice probabilities of any additive random utility discrete choice model can be viewed as emerging from rationally inattentive behavior, and vice-versa; we can go back and forth between the two paradigms.[14] Then, in order to apply an ARUM, it is no longer necessary 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 in a large set of options.

[14] In a similar vein, Webb (2016) demonstrates an equivalence between random utility models and bounded-accumulation or drift-diffusion models of choice and reaction times used in the neuroeconomics and psychology literature.

The underlying idea is that, by exploiting convex analytic properties of the discrete choice model, we establish a "duality" between the discrete choice and GERI models in the sense of convex conjugacy. Precisely, the surplus function of a discrete choice model has a convex conjugate that is a generalized entropy. Succinctly, then, GERI models are rational inattention problems in which the information cost is built from the convex conjugate of some ARUM.

A few remarks are in order. First, the equivalence result in this paper 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.[15] Any mixed discrete choice model such as these is observationally equivalent to a mixed GERI model.

In addition, there is also a connection between the results here and papers in the decision theory literature. The GERI optimization problem (15) bears resemblance to the variational preferences that Maccheroni et al. (2006) propose to represent ambiguity averse preferences, as well as to the revealed perturbed utility paradigm proposed by Fudenberg et al. (2015) to model stochastic choice behavior. Gul et al. (2014) show an equivalence between random utility and an "attribute rule" model of stochastic choice. The main point in this paper is to establish a duality between rational inattention models and random utility discrete choice models, which results in observational equivalence of their choice probabilities. A similar duality might arise between random utility discrete choice models and these other models from decision theory.

Finally, there are rational inattention models outside the GERI framework; that is, rational inattention models with information costs outside the class of generalized entropies introduced in this paper.[16] Obviously, choice probabilities from these non-GERI models would not be equivalent to those which can be generated from additive random utility discrete-choice models; it will be interesting to examine the empirical distinctions that non-GERI choice probabilities would have.

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

[16] 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 GERI framework.

References

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. (1996) *Discrete choice theory of product differentiation* MIT Press Cambridge, MA.

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

Caplin, A., Dean, M. and Leahy, J. (2016) Rational Inattention, Optimal Consideration Sets and Stochastic Choice *Working Paper* .

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., Leahy, J. and Matějka, F. (2016) Rational Inattention and Inference from Market Share Data *Working paper*.

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

de Oliveira, H., Denti, T., Mihm, M. and Ozbek, K. (2017) Rationally inattentive preferences and hidden information costs *Theoretical Economics* **12**(2), 621–654.

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

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.

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

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

Hébert, B. and Woodford, M. (2016) Rational Inattention with Sequential Information Sampling *Working paper*.

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. (2017) Buying a Larger Package with Quantity Surcharge: Information Friction or Preference Heterogeneity? *Working Paper*

Maccheroni, F., Marinacci, M. and Rustichini, A. (2006) Ambiguity aversion, robustness, and the variational representation of preferences *Econometrica* **74**(6), 1447–1498.

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.

Morris, S. and Yang, M. (2016) Coordination and Continuous Choice *Working Paper* .

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

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.

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 chapter 4, pp. 155–181.

Webb, R. (2016) The Dynamics of Stochastic Choice *Management Science*, forthcoming.

A Proofs and additional results

Notation. Vectors are denoted with boldface as $\mathbf{q} = (q_1, ..., q_N)$. 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, if a is a real number then $a + \mathbf{q} = (a + q_1, ..., a + q_J)$. 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 Jacobian is denoted J with, for example,

$$J_{\log \mathbf{S}}(\mathbf{q}) = \begin{pmatrix} \frac{\partial \log S_1(\mathbf{q})}{\partial q_1} & \dots & \frac{\partial \log S_1(\mathbf{q})}{\partial q_N} \\ \dots & \dots & \dots \\ \frac{\partial \log S_N(\mathbf{q})}{\partial q_1} & \dots & \frac{\partial \log S_N(\mathbf{q})}{\partial q_N} \end{pmatrix}.$$

A dot indicates an inner product or products of vectors and matrixes. For a vector $\mathbf{q}$, we use the shorthand $\mathbf{1} \cdot \mathbf{q} = \sum_i q_i$. The unit simplex in $\mathbb{R}^N$ is Δ.

Proof of Proposition 2. Note first that we may write

$$\mathbf{H}(e^{\mathbf{v}}) = e^{W(\mathbf{v})}\mathbf{q}(\mathbf{v}).$$

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 $\mathbf{H}$ on $\mathbb{R}^N_+$, suppose that $\mathbf{H}(e^{\mathbf{v}}) = \mathbf{H}\left(e^{\mathbf{v}'}\right)$ and aim to show that $\mathbf{v} = \mathbf{v}'$. Since $H_i(e^{\mathbf{v}}) = e^{W(\mathbf{v})}q_i(\mathbf{v})$ and $\sum_{i=1}^N q_i = 1$, we may sum $\sum_{i=1}^N H_i(e^{\mathbf{v}}) = \sum_{i=1}^N H_i\left(e^{\mathbf{v}'}\right)$ to find that $W(\mathbf{v}) = W(\mathbf{v}')$ and hence that $\mathbf{q}(\mathbf{v}) = \mathbf{q}(\mathbf{v}')$. Then by the Norets and Takahashi (2013) result, $\mathbf{v} = \mathbf{v}' + (c, ..., c)$ which leads to $W(\mathbf{v}) = W(\mathbf{v}') + c = W(\mathbf{v}) + c$, 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 $\mathbf{H}(\mathbf{y}) = \mathbf{x}$. By Norets and Takahashi, there exists $\mathbf{v} \in X$ such that $\mathbf{q}(\mathbf{v}) = \mathbf{x} / \sum_{i=1}^N x_i$. Let $c = -W(\mathbf{v}) + \ln \sum_{i=1}^N x_i$. Then

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

which establishes that $\mathbf{H}$ is a surjection from $\mathbb{R}_{+}^{N}$ to $\mathbb{R}_{+}^{N}$.

The next point is to extend $\mathbf{H}$ 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 $\mathbf{H}(\mathbf{y}) = (0,...,0)$. For $z \neq \emptyset$, consider the discrete choice model (1) 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 $\mathbf{H}(\mathbf{y}) = e^{\tilde{W}}(\tilde{p}_1,...,\tilde{p}_N)$.

The argument that $\mathbf{H}$ 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 $\mathbf{H}$ is continuous. We will do this by establishing that the values of $\mathbf{H}$ 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 $W(\ln \mathbf{y}^n) = W(\ln \mathbf{y}) - \ln n \to -\infty$. Then since $q_i(\mathbf{y}^n)$ are bounded between 0 and 1, $\mathbf{H}(\mathbf{y}^n) \to (0,..,0)$ 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

$$q_i(\mathbf{v}) = \int_{-\infty}^{\infty} F_i(u + v_i - v_1, ..., u + v_i - v_N)\, du. \tag{25}$$

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 (25) to see that

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

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

By dominated convergence,

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

Combining these results, find that $\mathbf{H}\left(\lim_{n\to\infty}\mathbf{y}^n\right) = \lim_{n\to\infty}\mathbf{H}\left(\mathbf{y}^n\right)$ as required. This completes the proof. ■

Proof of proposition 4. We first evaluate $W^*\left(\mathbf{q}\right)$. If $\mathbf{1}\cdot\mathbf{q}\neq 1$, then

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

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

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

$$W^*(\mathbf{q})=\sup_{\mathbf{v}}\{\mathbf{q}\cdot\mathbf{v}-W(\mathbf{v})\}. \tag{26}$$

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

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

Inserting this into (26) leads to the desired result.

W is convex and closed and hence W is the convex conjugate of W^* (Rockafellar, 1970, Thm. 12.2). This, along with Fenchel's equality (Rockafellar, 1970, Thm. 23.5), proves part (iii). Finally, for part (i), let $\mathbf{q}$ be a solution to problem (7). Then, by the homogeneity of $\mathbf{H}$ we have $\mathbf{q} = \frac{1}{\alpha}\mathbf{H}(e^{\mathbf{v}})$, where $\alpha = \sum_{j=1}^{N} H_j(e^{\mathbf{v}})$. 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. (7) we get

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

■

Proof of Proposition 5. 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 $\mathbf{H}$. 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 $\mathbf{H}$ 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 $\mathbf{q} = \mathbf{H}\left(e^{\hat{\mathbf{t}}}\right)$ and hence $W\left(e^{\hat{\mathbf{t}}}\right) = \log(\mathbf{1} \cdot \mathbf{H}(e^{\hat{\mathbf{t}}})) = \log 1 = 0$ by Proposition 4. Noting that $(\log(\mathbf{S}))^{-1}(\hat{\mathbf{t}}) = \mathbf{H}(e^{\hat{\mathbf{t}}})$ the requirement in part (ii) is equivalent to

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

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

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

as required.

Part (ii) follows from Proposition 4(ii). ■

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 (12).

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 5(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

$$\log\mathbf{S}_j\left(\mathbf{p}\left(\mathbf{v}\right)\right)=\tilde{v}_j-\gamma\left(\mathbf{v}\right). \tag{27}$$

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

$$\mathbf{p}\left(\mathbf{v}\right)=\mathbf{H}\left(e^{\tilde{\mathbf{v}}}\right)\exp\left(-\gamma\left(\mathbf{v}\right)\right). \tag{28}$$

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

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

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

$$\Lambda = \mathbb{E}\gamma\left(\mathbf{V}\right) = \mathbb{E}\log\sum_{j} H_j\left(e^{\mathbf{V}+\log\mathbf{S}\left(\mathbf{p}^0\right)}\right) \tag{29}$$

We may then use (29) to determine $\mathbf{p}^0$. Now apply Eq. (6) to establish part (iii) of the proposition. ■

Proof of corollary 8. 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

$$\begin{aligned} p_j^0 &= \mathbb{E}\left(\frac{H_j\left(e^{\mathbf{V}} \circ \mathbf{S}\left(\mathbf{p}^0\right)\right)}{\sum_k H_k\left(e^{\mathbf{V}} \circ \mathbf{S}\left(\mathbf{p}^0\right)\right)}\right) && (30) \\ &< \mathbb{E}\left(\frac{H_j\left(e^{V_j}\mathbf{S}\left(\mathbf{p}^0\right)\right)}{\sum_k H_k\left(e^{V_j}\mathbf{S}\left(\mathbf{p}^0\right)\right)}\right) && (31) \\ &= \mathbb{E}\left(\frac{e^{V_j}H_j\left(\mathbf{S}\left(\mathbf{p}^0\right)\right)}{e^{V_j}\sum_k H_k\left(\mathbf{S}\left(\mathbf{p}^0\right)\right)}\right) = \mathbb{E}\left(\frac{p_j^0}{\sum_k p_k^0}\right) = p_j^0. && (32) \end{aligned}$$

This is a contradiction as desired. ■

Proof of Proposition 9. Let $\mathbf{q}^0 = \mathbb{E}\mathbf{q}\left(\mathbf{v}\right)$, $\mathbf{c} = -\log\mathbf{S}\left(\mathbf{q}_0\right)$ and consider the GERI model 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 inverse generator $\mathbf{H}$. The GERI conditional choice probabilities satisfy

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

with $\mathbf{p}^0(\mathbf{v} - \log\mathbf{S}\left(\mathbf{q}_0\right)) = \mathbb{E}\mathbf{p}(\mathbf{v} - \log\mathbf{S}\left(\mathbf{q}_0\right))$. Then $\mathbf{p}(\mathbf{v} - \log\mathbf{S}\left(\mathbf{q}_0\right)) = \mathbf{q}(\mathbf{v})$ solves the GERI maximization problem.

To prove the converse, consider a GERI model with prior $(\mu, \mathcal{V})$, inverse generator $\mathbf{H}$ 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 Corollary 3. Let $\mathbf{c} = \log\mathbf{S}(\mathbf{p^0})$.

Then for $i \in \mathcal{C}$ we have

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

which is an ARUM on $\mathcal{C}$ with random utility shocks $(v+\epsilon)_{|\mathcal{C}}$. ■

In the case of the Shannon entropy, Corollary 8 can be strengthened considerably. In that case, any alternative that is dominated by another alternative in all states of the world will never be chosen, as shown in the following corollary:

Corollary 10 *Let* $\mathbf{S}$ *be the identity. Suppose that option* j *is dominated by option* i *in the sense that* $\forall \mathbf{v} \in \mathcal{V} : v_j \leq v_i$ *with strict inequality for some* $\mathbf{v}$. *Then* $p_j^0 = 0$.

Proof. Suppose to get a contradiction that $p_j^0 > 0$. From (13), obtain that for all options k with $p_k^0 > 0$ we have

$$1 = \frac{p_k^0}{p_k^0} = \frac{1}{p_k^0} \mathbb{E} p_k(\mathbf{V}) = \mathbb{E}\left(\frac{\exp(V_k)}{\sum_{k'} \exp(V_{k'}) p_{k'}^0}\right).$$

Then

$$1 = \mathbb{E}\left(\frac{\exp(V_j)}{\sum_{k'} \exp(V_{k'}) p_{k'}^0}\right) < \mathbb{E}\left(\frac{\exp(V_i)}{\sum_{k'} \exp(V_{k'}) p_{k'}^0}\right) = 1,$$

which is a contradiction. ■

B Additional properties of generalized entropy

We have shown that the generalized rational inattention model is always equivalent to an ARUM and conversely that the generalized rational inattention model may provide a boundedly rational foundation for any ARUM. We have discussed that the symmetry of generalized entropy is not desirable but it is still natural to ask whether an information cost based on generalized entropy has other properties that one would desire for an information cost. In this section we show that $\kappa_{\mathbf{S}}(\mathbf{p}(\cdot), \mu)$

does indeed possess two reasonable and desirable properties of cost functions that have been discussed in the existing literature (cf. de Oliveira et al. (2017), Hébert and Woodford (2016)), thus providing normative support for the GERI framework.

First, when $\mathbf{A}$ and $\mathbf{V}$ are independent, then the action $\mathbf{A}$ carries no information about the payoff $\mathbf{V}$. In that case the information cost should be zero, i.e.

Independence. *If $\mathbf{A}$ and $\mathbf{V}$ are independent, then $\kappa_{\mathbf{S}}(\mathbf{p}(\cdot), \mu) = 0$.*

Second, the mutual Shannon information $\kappa(\mathbf{p}(\cdot), \mu)$ is a convex function of $\mathbf{p}$. We show that the information cost $\kappa_{\mathbf{S}}(\mathbf{p}(\cdot), \mu)$ has a slightly weaker property, namely that it is convex on sets where $\mathbb{E}\mathbf{p}(\mathbf{V})$ is constant.

Convexity. *For a given μ, the information cost $\kappa_{\mathbf{S}}(\mathbf{p}(\cdot), \mu)$ is convex on any set of choice probabilities vectors satisfying $\{\mathbf{p} : \mathcal{V} \mapsto \Delta | \, \mathbb{E}\mathbf{p}(\mathbf{V}) = \hat{\mathbf{p}}\}$.*

The mutual Shannon information $\kappa(\mathbf{p}(\cdot), \mu)$ satisfies these two properties. The next proposition establishes that the information cost defined in (14) using generalized entropy also satisfies these properties.

Proposition 11 *The information cost defined in Eq. (14) satisfies the independence and convexity conditions.*

Proof of Proposition 11. *Independence:* By independence, we have, for all i, $p_i(\mathbf{v}) = k_i$, a constant. Then $p_i^0 = k_i$ and $\kappa_{\mathbf{S}}(\mathbf{p}(\cdot), \mu) = 0$.

Convexity: Consider two sets of choice probabilities $\mathbf{p}_1(\mathbf{v}), \mathbf{p}_2(\mathbf{v}), \mathbf{v} \in \mathcal{V}$, that have the same implied unconditional probabilities $\mathbb{E}\mathbf{p}_1(\mathbf{V}) = \mathbb{E}\mathbf{p}_2(\mathbf{V})$. For $\rho \in [0, 1]$, define $\mathbf{p}_\rho$ as the convex combination $\rho \mathbf{p}_1(\mathbf{v}) + (1-\rho)\mathbf{p}_2(\mathbf{v})$. Then we would like to show that

$$\rho \kappa_{\mathbf{S}}(\mathbf{p_1}(\cdot), \mu) + (1-\rho)\kappa_{\mathbf{S}}(\mathbf{p_2}(\cdot), \mu) \geq \kappa(\mathbf{p_\rho}(\cdot), \mu).$$

But

$$\begin{aligned} &\rho \kappa_{\mathbf{S}}(\mathbf{p_1}(\cdot), \mu) + (1-\rho)\kappa_{\mathbf{S}}(\mathbf{p_2}(\cdot)\,\mu) - \kappa(\mathbf{p_\rho}(\cdot), \mu) \\ = &-\rho \Omega_{\mathbf{S}}(\mathbf{p}_1) - (1-\rho)\Omega_{\mathbf{S}}(\mathbf{p}_2) + \Omega_{\mathbf{S}}(\rho \mathbf{p}_1 + (1-\rho)\mathbf{p}_1), \end{aligned}$$

which is positive by concavity of $\Omega_{\mathbf{S}}(\mathbf{p})$ (Proposition 5(ii)). ■