Discrete Choice and Rational Inattention: a General Equivalence Result

Fosgerau, Mogens and Melo, Emerson and Andre, De Palma and Shum, Matt

Technical University of Denmark, Indiana University, ENS Cachan, Universite Paris-Saclay, CREST;, Caltech

February 2017

Discrete Choice and Rational Inattention: a General Equivalence Result*

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

April 1, 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 function, the resulting choice probabilities in the rational inattention model take the multinomial logit form. By exploiting convex-analytic properties of the discrete choice model, we show that when information costs are modelled using a class of generalized entropy functions, 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. This includes empirically relevant specifications such as the probit and nested logit models.

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 psychic costs); therefore, agents may rationally choose to remain imperfectly informed about the 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

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

†Technical University of Denmark; mfos@dtu.dk

‡Indiana University; emelo@iu.edu

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

¶California Institute of Technology; mshum@caltech.edu

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. Matejka and McKay (2015) showed that when information costs are modelled using the Shannon entropy function, the resulting choice probabilities in the rational inattention model take the multinomial logit (MNL) form. In order for the rational inattention model to generate non-MNL choice probabilities, we need to generalize the information cost function beyond the Shannon entropy function assumed in much of the existing literature. We do this by exploiting convex-analytic properties of the additive random utility model to demonstrate a duality between discrete choice and rational inattention models.[1]

Specifically, we introduce a class of *Generalized Entropy Rational Inattention* (GERI) models.[2] In GERI models, the information cost functions are constructed from a class of "generalized entropy" functions; these generalized entropy functions are, essentially, "dual" to the class of random utility discrete choice models; precisely, the generalized entropy functions are the convex conjugate functions to the surplus functions in any arbitrary general additive random utility model. Hence, GERI models naturally yield choice probabilities that can equivalently be generated from general additive random utility models; the resulting choice probabilities can take forms far beyond the multinomial logit, including specifications such as nested logit, multinomial probit, and so on, which are often employed in empirical work.

Importantly, these generalized entropy functions allow for random utility models in which the random shocks are dependent across options; this corresponds to information cost functions that exhibit information spillovers across options with shared features, which may be reasonable in many decision environments. In contrast, the multinomial logit model assumes independent shocks; correspondingly, the Shannon entropy function precludes information spillovers.

The paper is organized as follows. Section 2 presents insights into the fundamental convex-analytic structure of the additive random utility discrete choice model. Using this structure, we formulate a class of generalized entropy functions and present key results about them. Section 3 introduces the rational inattention model. We show how the generalized entropy functions can be used to define the information cost function in the rational inattention model. 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 the rational inattention model based on the generalized entropy functions. Section 4 discusses an example while Section 5 concludes.

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 $\mathbb{R}^N$.

[1] Throughout this paper, we will use the terms "additive random utility model" and "discrete choice model" interchangeably.

[2] This complements work by Hébert and Woodford (2016), who also consider generalizations of the information cost function.

2 Random utility models and generalized entropy functions

Consider a decision-maker (DM) making discrete choices among a set of $i = 1, \ldots, N$ options, where, for each option i, the utility is given by

$$u_i = \tilde{v}_i + \epsilon_i, \tag{1}$$

where $\tilde{\mathbf{v}} = (\tilde{v}_1, \ldots, \tilde{v}_N)$ is deterministic and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)$ is a vector of random utility shocks. This is the classic additive random utility framework pioneered by McFadden (1978).

Assumption 1 *The random vector $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)$ follows a joint distribution with finite means that is absolutely continuous, independent of $\tilde{\mathbf{v}}$, and fully supported on $\mathbb{R}^N$.*

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

$$W(\tilde{\mathbf{v}}) = \mathbb{E}_{\boldsymbol{\epsilon}}(\max_i [\tilde{v}_i + \epsilon_i]). \tag{2}$$

Under Assumption 1, $W(\tilde{\mathbf{v}})$ is convex and differentiable[3] and the choice probabilities coincide with the derivatives of $W(\tilde{\mathbf{v}})$:

$$\frac{\partial W(\tilde{\mathbf{v}})}{\partial \tilde{v}_i} = q_i(\tilde{\mathbf{v}}) \equiv \mathbb{P}\left(\tilde{v}_i + \epsilon_i \geq \tilde{v}_j + \epsilon_j, \, \forall j \neq i\right) \text{ for } i = 1, \ldots, N$$

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

As a running example, we consider the familiar logit model. When the ϵ_i's are distributed i.i.d. across options i according to the type 1 extreme value distribution, then the resulting choice probabilities take the familiar multinomial logit form: $q_i(\tilde{\mathbf{v}}) = e^{\tilde{v}_i} / \sum_j e^{\tilde{v}_j}$. Assumption 1 above leaves the distribution of the ϵ's unspecified, thus allowing for choice probabilities beyond the multinomial logit case. Importantly, it accommodates arbitrary correlation in the ϵ_i's across choices, which is reasonable and realistic in applications.

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.

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

From the differentiability of W and the Williams-Daly-Zachary theorem it follows that the choice probabilities emerging from any random utility discrete choice

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

model can be expressed in closed-form in terms of the $\mathbf{H}$ function as:[4]

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

For the multinomial logit case, the surplus function is $W(\tilde{\mathbf{v}}) = \log\left(\sum_{i=1}^{N} e^{\tilde{v}_i}\right)$, implying that $H_i(e^{\tilde{\mathbf{v}}}) = e^{\tilde{v}_i}$. Thus, for this case Eq. (4) becomes the multinomial logit choice formula.

The function $\mathbf{H}$ is globally invertible (see Lemma 8 in the Appendix), and we introduce a function $\mathbf{S}$ defined as the inverse of $\mathbf{H}$,

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

For reasons that will be apparent below, we refer to $\mathbf{S}$ as a *generator* function. There is a close relationship between the function $\mathbf{S}(\cdot)$ and the surplus function $W(\tilde{\mathbf{v}})$ 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* (Rockafellar, 1970, ch. 12).[5]

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

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

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

(ii) *The convex conjugate function for the surplus function $W(\tilde{\mathbf{v}})$ 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 function $-W^*(\cdot)$ a **generalized entropy function**.*

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

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

[4] By direct differentiation of (3), and applying the Williams-Daly-Zachary theorem, we have $q_i(\tilde{\mathbf{v}}) = H_i(e^{W(\tilde{\mathbf{v}})})/e^{W(\tilde{\mathbf{v}})}$ for all i. Imposing the summability restriction $\sum_i q_i(\tilde{\mathbf{v}}) = 1$ we have $\sum_i H_i(e^{W(\tilde{\mathbf{v}})}) = e^{W(\tilde{\mathbf{v}})}$ leading to Eq. (4).

[5] 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.

and the RHS is optimized at the choice probabilities $\mathbf{q}(\tilde{\mathbf{v}}) = \nabla W(\tilde{\mathbf{v}})$.

Parts (i) and (ii) 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. We use this structure to define the class of *generalized entropy* functions. To see how this works, consider again the multinomial logit model, for which $\mathbf{H}$ is the identity, implying that the corresponding generator function $\mathbf{S}(\mathbf{q}) = \mathbf{q}$ is also just the identity. Then by Proposition 1(ii), the negative convex conjugate 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 function.

Generalizing from this, Proposition 1(ii) shows how the conjugate function for any discrete choice model can be generated by the function $\mathbf{S}$. Therefore we refer to the negative conjugate function $-W^*(\mathbf{q}) = -\mathbf{q} \cdot \log \mathbf{S}(\mathbf{q}) = -\sum_i q_i \log S_i(\mathbf{q})$ for any general discrete choice model as a *generalized entropy* function. Comparing the generalized and Shannon entropies, the former allows for cross-effects, in the sense that the choice probability for option j, q_j, enters the entropic term for option i, $S_i(\mathbf{q})$. As we will see below, these cross-effects allow for information spillovers when we use these generalized entropy functions to construct rational inattention models.

Proposition 1(iii) provides an alternative representation of the surplus function from 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 function. Specifically, Eq. (2) implies that the surplus function can be written as

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

Combining this with (7), we obtain an alternative expression for the generalized entropy function, as a choice probability-weighted sum of expectations of the utility shocks $\boldsymbol{\epsilon}$:[6]

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

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

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

Proposition 2 (Properties of the generator functions) *Let assumption 1 hold. Then the vector valued-function* $\mathbf{S}(\cdot)$ *defined by (5) satisfies the following conditions:*

[6] See Chiong, Galichon, and Shum (2016). Additionally, we conjecture that $\log S_i(\mathbf{q}) = -\mathbb{E}[\epsilon_i | u_i \geq u_j, j \neq i]$ for $i = 1, \ldots, N$. For the multinomial logit case, corresponding to $\mathbf{S}(\mathbf{q}) = \mathbf{q}$, McFadden (1978) showed that $\gamma - \log q_i = \mathbb{E}[\epsilon_i | u_i \geq u_j, j \neq i]$, for γ being Euler's constant.

(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*.*

The possibility of zero choice probabilities will play a role in what follows. We impose an additional regularity assumption on the generator functions $\mathbf{S}$.

Assumption 2 *Let* $\mathbf{q}$ *be a probability vector. Then* $q_i = 0$ *if and only if* $S_i(\mathbf{q}) = 0$*.*

This assumption is satisfied for the generator functions $\mathbf{S}$ corresponding to many familiar additive random utility models, including the multinomial logit and the nested logit models.[7]

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 additive random utility model, 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 that 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 $\mathbb{P}(\mathbf{A} = i|\mathbf{V} = \mathbf{v})$. 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:

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

[7] In fact, the necessity part of Assumption 2 arises immediately from the results in this section. As $\tilde{v}_i \to -\infty$, $q_i(\tilde{\mathbf{v}}) \to 0$, which by (4) implies that $H_i\left(e^{\tilde{\mathbf{v}}}\right) \to 0$. Then, since $\log \mathbf{S}(\mathbf{q}(\tilde{\mathbf{v}})) = \tilde{\mathbf{v}} - \log \sum_j H_j(e^{\tilde{\mathbf{v}}})$, we have $\log S_1(q) \to -\infty$ (by homogeneity of $\mathbf{H}$, we may suppose that $\log \sum_j H_j(e^{\tilde{\mathbf{v}}})$ is a constant).

The previous literature has used the Shannon entropy to specify the information cost, which connects the rational inattention model to the multinomial logit model. We review these results in the next Section 3.1. Then in Section 3.2 we introduce generalized entropy to the problem. This connects the rational inattention model to general additive random utility 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 $\mathbf{A}$ may be written as

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

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

$$\max_{\mathbf{p}(\cdot)} \left\{ \sum_{\mathbf{v}\in\mathcal{V}} \left(\sum_{i=1}^{N} v_i p_i(\mathbf{v}) \right) \mu(\mathbf{v}) - \kappa(\mathbf{p}(\cdot), \mu) \right\} \tag{11}$$

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 $p_i^0 = \mathbb{E}p_i(\mathbf{V})$. It is an important feature of the rational inattention model that some p_i^0 may be zero, in which case the corresponding $p_i(\mathbf{v})$ are also zero. Then the rational inattention model implies the formation of a *consideration set*, comprising those options that have strictly positive probability of being chosen (cf. Caplin, Dean, and Leahy (2016)).

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 $\tilde{\mathbf{v}}$ is $\mathbf{v}$ shifted by $\log \mathbf{p}^0$. For options that are not in the consideration set, the shifted payoff is $\tilde{v}_i = -\infty$. 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. To achieve that, we replace the Shannon entropy by the generalized entropy introduced in Section 2 above. Since each generalized entropy implies a corresponding discrete choice model (Proposition 2), it turns out that each RI model with an information cost derived from a generalized entropy will generate choice probabilities consistent with a corresponding discrete choice model (Proposition 4 below); this implies that *any* additive random utility discrete choice model can be microfounded by a corresponding rational inattention model, thus generalizing substantially the results in the previous section.

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 entropy generator corresponding to some additive random utility model satisfying Assumptions 1 and 2, 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

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

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

$$\max_{\mathbf{p}(\cdot)} \sum_{\mathbf{v} \in \mathcal{V}} \left(\sum_{i=1}^N v_i p_i(\mathbf{v}) \right) \mu(\mathbf{v}) - \kappa_{\mathbf{S}}(\mathbf{p}(\cdot), \mu). \tag{15}$$

The following proposition characterizes the optimal solution to the GERI model.

Proposition 3 *The solution to the GERI model:*

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

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

(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. Part (iii) illustrates the close connection between convex analysis and the GERI problem. To see this, note that the GERI information cost function 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 1(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 3(iii).

It is worth remarking that some of the optimal unconditional choice probabilities may be zero. For these options, the corresponding conditional choice probabilities will also be zero for all $\mathbf{v}$.[8] The rational inattention model then also describes the formation of consideration sets, i.e. the set of options that are chosen with positive probability.[9]

[8] To see this, consider the solution to the GERI problem given in Eq. (17) and define $\tilde{\mathbf{v}} = \mathbf{v} + \log \mathbf{S}(\mathbf{p}^0)$. Let $p_i^0 = 0$. Then by assumption 2 it follows that $\log S_i(\mathbf{p}^0) = -\infty$, or equivalently, $\tilde{v}_i \longrightarrow -\infty$ and hence $p_i(\mathbf{v}) = 0$ for all $\mathbf{v} \in \mathcal{V}$.

[9] Because of the possibility of zero choice probabilities for some options, GERI models can also generate failures of the "regularity" property (adding an option to a choice set cannot increase the choice probability for any of the original choices). See section B in the Appendix for an example.

While Proposition 3 does not characterize explicitly the optimal 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 4 *For some option a, and for all $\mathbf{v} \in \mathcal{V}$, let $v_a \leq v_i$ for all $i \neq a$, and assume that the inequality is strict with positive probability. Then $p_a^0 = 0$ (that is, option a is not in the optimal consideration set).*

For the special case of Shannon entropy (when $\mathbf{S}$ is the identity function), the result can be strengthened even further. Corollary 7 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 optimal consideration set.

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 additive random utility model and vice versa.

Combining the choice probabilities $p_i(\mathbf{v})$ in (17) from the GERI model and the choice probabilities $q_i(\tilde{\mathbf{v}})$ in (4) from the additive random utility model, we find that if payoffs are related by

$$\tilde{v}_i = v_i + \log S_i(\mathbf{p}^0) \quad \text{for } i = 1, \ldots, N, \tag{21}$$

then the two models yield the same choice probabilities

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

Given a GERI model with payoffs $\mathbf{v} \in \mathcal{V}$ and unconditional choice probabilities $\mathbf{p}^0$, we may then use (21) to construct deterministic utility components $\tilde{\mathbf{v}}$ for the additive random utility model. If the GERI model has some zero unconditional choice probabilities p_i^0, then Assumption 2 ensures that $p_i(\mathbf{v}) = 0$ if and only if $q_i(\tilde{\mathbf{v}}) = 0$. The additive random utility model that corresponds to the GERI model is then an extended additive random utility model in which some deterministic utility components are minus infinity.

Conversely, given an additive random utility model with flexible generator $\mathbf{S}$ and a prior distribution $\tilde{\mu}$ of the deterministic utility components $\tilde{\mathbf{v}} \in \tilde{\mathcal{V}}$, define $\mathbf{p}^0 = \mathbb{E}\mathbf{q}(\tilde{\mathbf{v}})$ and note that all $p_i^0 > 0$. Then define $\mathbf{v}$ using (21) and define similarly μ and $\mathcal{V}$ using the same location shift $\log \mathbf{S}(\mathbf{p}^0)$. By the same argument as before, the GERI model with payoffs $\mathbf{v} \in \mathcal{V}$, prior μ and flexible generator $\mathbf{S}$ for

the generalized entropy has the same choice probabilities as the additive random utility model.

Hence, we have shown the following proposition.

Proposition 5 *For every additive random utility discrete choice model and every prior distribution on $\tilde{\mathcal{V}}$, there is an equivalent GERI model with a prior distribution on $\mathcal{V}$, where $\mathcal{V}$ is equal to $\tilde{\mathcal{V}}$ up to a location shift.*

Conversely, every GERI model is equivalent to an additive random utility discrete choice model in which the utility components for options chosen with zero probability are set to minus infinity.

In Section 4, we will apply this proposition to study a GERI model in which the choice probabilities are equivalent to those from a nested logit discrete choice model.

3.4 Additional properties of generalized entropy cost functions

We have shown that the generalized rational inattention model is always equivalent to an additive random utility model and conversely that the generalized rational inattention model may provide a boundedly rational foundation for any additive random utility model. The key to this result is the generalization of the information cost function $\kappa_{\mathbf{S}}(\mathbf{p}(\cdot), \mu)$ using generalized entropy as defined in Eq. (14). It is then natural to ask whether $\kappa_{\mathbf{S}}(\mathbf{p}(\cdot), \mu)$ has the 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, Denti, Mihm, Ozbek (2015), 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 $\mathbf{p}$. This is useful as it ensures a unique solution to the problem of the rationally inattentive DM. 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 function $\kappa_{\mathbf{S}}(\mathbf{p}(\cdot), \mu)$ 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 the generalized entropy functions also satisfies these properties.

Proposition 6 *The information cost defined in Eq. (14) satisfies the independence and convexity conditions.*

4 Example: The nested logit GERI model

From an applied point of view, an important implication of Proposition 5 is that it allows us to formulate rational inattention models that have complex substitution patterns, beyond the multinomial logit case. In this example, we consider a GERI model with an information cost function derived from a nested logit discrete choice model. The nested logit choice probabilities are consistent with a discrete choice model in which the utility shocks $\boldsymbol{\epsilon}$ are jointly distributed in the class of generalized extreme value distributions. 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 "red bus/blue bus" criticism.[10]

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 $\tilde{\mathbf{v}}$, the nested logit choice probabilities are given by:

$$q_i(\tilde{\mathbf{v}}) = \frac{e^{\tilde{v}_i/\zeta_{g_i}}}{\sum_{j \in g_i} e^{\tilde{v}_j/\zeta_{g_i}}} \cdot \frac{e^{\zeta_{g_i} \log\left(\sum_{j \in g_i} e^{\tilde{v}_j/\zeta_{g_i}}\right)}}{\sum_{\text{all nests } g} e^{\zeta_g \log\left(\sum_{j \in g} e^{\tilde{v}_j/\zeta_g}\right)}}. \tag{22}$$

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

$$S_i(\mathbf{q}) = q_i^{\zeta_{g_i}} \left(\sum_{j \in g_i} q_j \right)^{1-\zeta_{g_i}} \tag{23}$$

Using this, and applying Proposition 5, the nested logit choice probabilities (22) are also equivalent to those from a GERI model with valuations

$$v_i = \tilde{v}_i - \zeta_{g_i} \log p_i^0 - (1 - \zeta_{g_i}) \log \left(\sum_{j \in g_i} p_j^0 \right), \quad i \in \{1, \ldots, n\}. \tag{24}$$

The $\mathbf{S}$ function for the nested logit model in Eq. (23) has several interesting features, relative to the Shannon entropy. First, Eq. (23) allows us to write the generalized entropy $\Omega_{\mathbf{S}}(\mathbf{p})$ as

$$\Omega_{\mathbf{S}}(\mathbf{p}) = -\sum_{i=1}^{N} \zeta_{g_i} p_i \log p_i - \sum_{i=1}^{N} (1 - \zeta_{g_i}) p_i \log \left(\sum_{j \in g_i} p_j \right). \tag{25}$$

The first term in Eq (25) captures the Shannon entropy within nests, whereas the second term captures the information between nests. According to this, we may interpret Eq. (25) as an "augmented" version of Shannon entropy.

Second, when the nesting parameter $\zeta_{g_j} = 1$, then $\mathbf{S}$ is the identity function

[10] See, for instance, Maddala (1983, Chap. 2), and Anderson, de Palma, Thisse (1992).

($S_j(\mathbf{p}) = p_j$ 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 which 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.

From the discrete choice perspective, nested logit choice probabilities allow for correlation in the utility shocks (ϵ's) corresponding to the different choice options. Analogously, in an information cost function constructed from the nested logit $\mathbf{S}$ function in Eq. (23), there will be a common cost component across all options belonging to the same nest, corresponding to the term $\left(\sum_{j \in g_j} p_j\right)^{1-\zeta_{g_j}}$ which is common to all $S_j(\mathbf{p})$ for $j \in g_j$. From an information processing perspective, this suggests that there are spillovers in gathering information for options in the same nest. Information spillovers across choices arise in many decision environments. For example, a supermarket shopper gains information about common features of the vegetables, such as the average freshness and quality, while looking at any of them. In animal foraging, animals who have information about presence of predators in one grazing site also use that information to update about predator presence at other nearby sites.

For the Shannon entropy, in contrast, these common terms do not exist, so that there are no spillovers across options in information processing. From a behavioral point of view, then, more correlated utility shocks makes each option's signal harder to distinguish – there is more redundant information – implying that multinomial logit choice probabilities, which would ignore this correlation, manifest a type of correlation neglect.

To illustrate this point, we compute a GERI model using the nested-logit cost function. (This requires solving the fixed point equation (16).) In this example, 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, we might expect that any non-uniform choice probabilities should reflect underlying asymmetries in the information cost function.

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

As we expect, 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 function ($S_i(\mathbf{p}) = p_i$) in which information costs are separable across all five choices.[11] Unlike the multinomial logit case, we see that choice probabilities are higher for the options

[11] 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.

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

1,2 and 3, which constitute the larger nest, and smaller for options 4,5 which constitute the smaller nest. (However, within nest, the choice probabilities are identical.) The non-uniform choice probabilities for the nested logit model reflect the cost spillovers across options in the structure of the nested logit information cost function.

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, however, is higher, being over 35%.

This simple example demonstrates the substantive importance of allowing for information cost functions beyond the Shannon entropy, which leads to multinomial logit choice probabilities. Obviously, it makes a difference for a DM to be processing information using the nested logit cost function, as compared to the Shannon cost function, as the highest valuation option is chosen with higher probability on average using the nested logit cost function.

5 Summary

The central result in this paper is the observational equivalence between a 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.[12] Then, in order to apply an additive random utility discrete choice model, it is no

[12] 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.

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 that decision makers to be aware of all options in a large set of options.

The underlying idea is that, by exploiting convex analytic properties of the discrete choice model, we show 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 functions are built from the convex conjugate functions of some additive random utility discrete choice model.

A few remarks are in order. First, the equivalence result in this paper is at the individual level, hence it also holds for additive random utility models with random parameters, including the mixed logit or random coefficient logit models which have been popular in applied work.[13] 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, Marinacci, and Rustichini (2006) propose to represent ambiguity averse preferences, as well as to the revealed perturbed utility paradigm proposed by Fudenberg, Iijima, and Strzalecki (2015) to model stochastic choice behavior. Gul, Natenzon, and Pesendorfer (2014) shows 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 cost functions outside the class of generalized entropy functions introduced in this paper.[14] Obviously, choice probabilities from these non-GERI models would not be equivalent to those which can be generated from random utility discrete-choice models; it will be interesting to examine the empirical distinctions that non-GERI choice probabilities would have.

[13] See, for instance, Berry, Levinsohn, and Pakes (1995), McFadden and Train (2000), Fox, Kim, Ryan, Bajari (2012).

[14] 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

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

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

A. Caplin, M. Dean, and J. Leahy (2016). Rational Inattention, Optimal consideration sets and stochastic choice. Working paper.

A. Caplin, J. Leahy, and F. Matejka (2016). Rational Inattention and Inference from market Share Data. Working paper.

K. Chiong, A. Galichon, and M. Shum (2016). Duality in Dynamic Discrete Choice Models. *Quantitative Economics*, 7 (1), pp. 83-115.

H. de Oliveira, T. Denti, M. Mihm, K. Ozbek (2015). Rationally Inattentive Preferences and Hidden Information Costs. Working paper.

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

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

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

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

R. D. Luce (1959). *Individual Choice Behavior*. John Wiley, 1959.

F. Maccheroni, M. Marinacci, and A. Rustichini (2006). Ambiguity Aversion, Robustness, and the Variational Representation of Preferences, *Econometrica*, 74(6): 1447–1498.

G. Maddala (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press, 1983.

F. Matêjka and A. McKay (2015). Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model, *American Economic Review*, 105(1): 272–98.

D. McFadden (1978). Modelling the choice of residential location. In *Spatial Interaction Theory and Residential Location* (A. Karlquist et. al., eds.), North-Holland, Amsterdam.

D. McFadden (1981). Econometric Models of Probabilistic Choice. In: C.Manski and D. McFadden (Eds), Structural Analysis of Discrete Data with Economic Applications, Cambridge, MA: MIT Press, 198–272.

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

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

M. Ruzhansky and M. Sugimoto (2015) On global inversion of homogeneous maps. *Bulletin of Mathematical Sciences* 5: 13-18.

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

C. Sims (2003). Implications of Rational inattention. *Journal of Monetary Economics*, 50(3), pp. 665-690.

C. Sims (2010). Rational inattention and monetary economics. *Handbook of Monetary Economics*, Volume 3, pp. 155-181.

L. Thurstone (1927). A law of comparative judgment. *Psychological review*, 34(4): 273-278.

R. Webb (2016). The Dynamics of Stochastic Choice. Working paper.

A Proofs and additional results

Notation. Vectors are denoted simply 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 $\tilde{\mathbf{v}}$ is $\nabla_{\tilde{\mathbf{v}}}$; e.g., for $\tilde{\mathbf{v}} = (v_1, ..., v_N)$, $\nabla_{\tilde{\mathbf{v}}} W(\tilde{\mathbf{v}}) = \left(\frac{\partial W(\tilde{\mathbf{v}})}{\partial \tilde{v}_1}, ..., \frac{\partial W(\tilde{\mathbf{v}})}{\partial \tilde{v}_N}\right)$. The Jacobian is denoted J with, for example,

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

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

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

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

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

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

Note that for any constant k we have $W(\tilde{\mathbf{v}} + k \cdot \mathbf{1}) = k + W(\tilde{\mathbf{v}})$, so that we normalize $\mathbf{1} \cdot \tilde{\mathbf{v}} = 0$. Maximize then the Lagrangian $\mathbf{q} \cdot \tilde{\mathbf{v}} - W(\tilde{\mathbf{v}}) - \lambda(\mathbf{1} \cdot \tilde{\mathbf{v}})$ with

first-order conditions $0 = q_j - \frac{\partial W(\tilde{\mathbf{v}})}{\partial \tilde{v}_j} - \lambda$, which lead to $\lambda = 0$. Then

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

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 $\mathbf{q} = \frac{1}{\alpha}\mathbf{H}(e^{\tilde{\mathbf{v}}})$, where $\alpha = \sum_{j=1}^{N} H_j(e^{\tilde{\mathbf{v}}})$. Then, by the definition of $\mathbf{S}$ it follows that $\mathbf{S}(\mathbf{q}) = \frac{e^{\tilde{\mathbf{v}}}}{\alpha}$. Replacing the latter expression in Eq. (7) we get

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

■

Proof of Proposition 2. 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

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

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

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

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 1. 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

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

as required.

Part (ii) follows from Proposition 1(ii). ■

Proof of proposition 3. The Lagrangian for the DM's problem is

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

where $\gamma(\mathbf{V})$ and $\xi_j(\mathbf{V})$ are Lagrange multipliers corresponding to condition (12).

Before we derive the first-order conditions for $p_j(\mathbf{v})$ it is useful to note that we may regard the terms $\log \mathbf{S}(\mathbf{p}^0)$ and $\log \mathbf{S}(\mathbf{p}(\mathbf{v}))$ in the information cost $\kappa_{\mathbf{S}}(\mathbf{p}, \mu)$ as constant, since their derivatives cancel out by Proposition 2(iii). Define $\tilde{v}_j = v_j + \xi_j(\mathbf{v}) + \log S_j(\mathbf{p}^0)$ and $\tilde{\mathbf{v}} = (\tilde{v}_1, ..., \tilde{v}_N)$. Then the first-order condition for $p_j(\mathbf{v})$ is easily found to be

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

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

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

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

Using that probabilities sum to 1 leads to

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

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

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

$$\Lambda = \mathbb{E}\gamma(\mathbf{V}) = \mathbb{E}\log \sum_j H_j\left(e^{\tilde{\mathbf{v}}}\right) \tag{29}$$

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

Proof of proposition 4. Assume, towards a contradiction, that $p_a^0 > 0$. Then

$$p_a^0 = \mathbb{E}\left(\frac{H_a\left(\left\{e^{V_c}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)}{\sum_b H_b\left(\left\{e^{V_c}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)}\right) \tag{30}$$

$$< \mathbb{E}\left(\frac{H_a\left(\left\{e^{V_a}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)}{\sum_b H_b\left(\left\{e^{V_a}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)}\right) \tag{31}$$

$$= \mathbb{E}\left(\frac{e^{V_a}H_a\left(\left\{S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)}{e^{V_a}\sum_b H_b\left(\left\{S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)}\right) = \mathbb{E}\left(\frac{p_a^0}{\sum_b p_b^0}\right) = p_a^0. \tag{32}$$

The first inequality (31) follows from cyclic monotonicity, which is a property of the gradient of convex functions. (See, for instance, Rockafellar (1970, Thm. 23.5).) Since the surplus function W is convex, its gradient, corresponding to the choice probabilities $\mathbf{p}(\cdot)$ is a cyclic monotone mapping, implying that

$$\left[\mathbf{p}\left(\left\{e^{v_a}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right) - \mathbf{p}\left(\left\{e^{v_c}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)\right]\cdot\left[\left\{e^{v_a}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N - \left\{e^{v_c}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right] \geq 0.$$

All the terms within the second pair of brackets on the LHS are ≤ 0, except for the a-th term, which is equal to zero. In order to satisfy the inequality, then, we must have

$$p_a\left(\left\{e^{v_a}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right) \geq p_i\left(\left\{e^{v_c}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)$$

with the inequality strict with positive probability. Otherwise,

$$\sum_{i\neq a}\left\{p_i\left(\left\{e^{v_a}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right) - p_i\left(\left\{e^{v_c}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)\right\} > 0$$

and

$$\begin{aligned}
&\left[\mathbf{p}\left(\left\{e^{v_a}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right) - \mathbf{p}\left(\left\{e^{v_c}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)\right]\cdot\left[\left\{e^{v_a}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N - \left\{e^{v_c}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right] \\
&= \sum_{c\neq a}\left[p_c\left(\left\{e^{v_a}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right) - p_c\left(\left\{e^{v_c}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)\right]\left[e^{v_a} - e^{v_c}\right]S_c\left(\mathbf{p}^0\right) \\
&\leq \max_{c\neq a}\left[\left(e^{v_a} - e^{v_c}\right)S_c\left(\mathbf{p}^0\right)\right]\sum_{c\neq a}\left[p_c\left(\left\{e^{v_a}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right) - p_c\left(\left\{e^{v_c}S_c\left(\mathbf{p}^0\right)\right\}_{c=1}^N\right)\right] \leq 0
\end{aligned}$$

with the final inequality strict with positive probability. Hence, we conclude that $p_a^0 = 0$. ■

In the case of the Shannon entropy, Corollary 4 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 7 *Let S be the identity. Suppose that option a is dominated by option d in the sense that $\forall \mathbf{v} \in \mathcal{V} : v_a \leq v_d$ with strict inequality for some $\mathbf{v}$. Then $p_a^0 = 0$.*

Proof. Suppose to get a contradiction that $p_a^0 > 0$. From (13), obtain that for all options a

$$1 = \frac{p_a^0}{p_a^0} = \frac{1}{p_a^0}\mathbb{E}p_a(\mathbf{V}) = \mathbb{E}\left(\frac{\exp(V_a)}{\sum_b \exp(V_b)\, p_b^0}\right).$$

Then

$$\mathbb{E}\left(\frac{\exp(V_d)}{\sum_b \exp(V_b)\, p_b^0}\right) > 1,$$

which is a contradiction. ■

Proof of Proposition 6. *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 $\mathbf{p}_1(\mathbf{v}), \mathbf{p}_2(\mathbf{v}), \mathbf{v} \in \mathcal{V}$, where both 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 convexification $\rho\mathbf{p}_1(\mathbf{v}) + (1-\rho)\mathbf{p}_2(\mathbf{v})$. 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 2(ii)). ■

Lemma 8 $\mathbf{H}$ *is invertible.*

Proof of Lemma 8. We shall make use of Ruzhansky and Sugimoto's 2015 invertibility result applied to $\mathbf{H}$. The Jacobian of $\tilde{\mathbf{v}} \to \mathbf{H}\left(e^{\tilde{\mathbf{v}}}\right)$ is $\left\{e^{W(\tilde{\mathbf{v}})}\frac{\partial W(\tilde{\mathbf{v}})}{\partial v_i}\frac{\partial W(\tilde{\mathbf{v}})}{\partial v_j}\right\} + \left\{e^{W(\tilde{\mathbf{v}})}\frac{\partial^2 W(\tilde{\mathbf{v}})}{\partial v_i \partial v_j}\right\}$. The first matrix is positive definite since all choice probabilities are positive, the second matrix is positive semidefinite due to the convexity of W, hence this matrix is everywhere positive definite and then the Jacobian determinant of $\tilde{\mathbf{v}} \to \mathbf{H}\left(e^{\tilde{\mathbf{v}}}\right)$ never vanishes. This implies in turn that the Jacobian determinant

of the composition $\mathbf{y} \to \log \mathbf{y} \to \mathbf{H}(\mathbf{y})$ never vanishes. It remains to show that $\inf_{\mathbf{y}\in\Delta} \|\mathbf{H}(\mathbf{y})\| > 0$. But $\mathbf{y} \in \Delta$ implies that

$$\begin{aligned}
\|\mathbf{H}(\mathbf{y})\| &= e^{W(\log \mathbf{y})} \|\nabla W(\log \mathbf{y})\| \\
&\geq e^{\mathbb{E}\max_j\{\log y_j + \varepsilon_j\}} J^{-1/2} \\
&\geq e^{\max_j\{\log y_j + \mathbb{E}\varepsilon_j\}} J^{-1/2} \\
&= \max_j \left\{ y_j e^{\mathbb{E}\varepsilon_j} \right\} J^{-1/2} \\
&\geq \left\| \left(y_1 e^{\mathbb{E}\varepsilon_1}, ..., y_J e^{\mathbb{E}\varepsilon_N} \right) \right\| J^{-1} \\
&\geq \left(\sum_{j=1}^{N} e^{-2\mathbb{E}\varepsilon_j} \right)^{-1} J^{-1} > 0,
\end{aligned}$$

where we first used that ∇W is on the unit simplex, second that the max operation is convex, third that the sup-norm bounds the euclidean norm, and fourth that the minimum of $\left\| \left(y_1 e^{\mathbb{E}\varepsilon_1}, ..., y_N e^{\mathbb{E}\varepsilon_N} \right) \right\|$ on the unit simplex is attained at $y_j = e^{-2\mathbb{E}\varepsilon_j} \left(\sum_{k=1}^{N} e^{-2\mathbb{E}\varepsilon_k} \right)^{-1}, j = 1, ..., N.$ ■

B Example: Consideration sets and failure of regularity

Next, 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 discrete choice model, and represent an important point of difference between the two models. Matejka 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 function introduced in Section 4 of the main text.

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 choices (1,2) with nesting parameter $\zeta_1 = 0.7$, and nest 2 consists of choices (3,4) with parameter $\zeta_2 = 0.8$.

For each model, we compute the optimal unconditional probabilities (which as in the previous example, requires solving the fixed-point equation (16)) first for the choice set $\{1, 2, 3\}$, and then for the expanded choice set $\{1, 2, 3, 4\}$. This example illustrates how adding option 4 to the choice set can results in increases in the choice probabilities of choices (1,2,3) thus showing a failure of the regularity property. The optimal unconditional probabilities are shown in Table 3. Qualitatively the results are the same between both the Shannon and GERI-nested logit specifications. With the smaller set of options, we see that only options 1,2 are

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

Table 2: Valuation vectors in Example 2

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: Optimal unconditional probabilities for Example 3

chosen with positive probabilities. When option 4 is added, however, then option 1 drops out of the consideration set, and only options 3,4 are chosen with positive probability. This demonstrates a failure of regularity, as the addition of choice 4 *increases* the prior choice probability for choice 1. (Moreover, note that with the expanded choice set, option 2 is chosen with zero probability, even though it is not inferior in all states of the world, which demonstrates that the characterization of consideration sets in Corollary 4 is not exhaustive.)

Basically, the addition of choice 4 allows agents to form an effective "hedge" in conjunction with choice 3. In the state when choice 3 yields a low payoff (state $\mathbf{v}^2$), choice 4 yields a high payoff; on the contrary, when choice 4 yields a lower payoffs (states $\mathbf{v}^1$ and $\mathbf{v}^3$), choice 3 yields high payoffs.