Universitat Pompeu Fabra Barcelona
Department of Economics and Business

Economics Working Paper Series

Working Paper No. 1936

Attention-Entropy Random Utility: Endogenous: Attention and Context Effects in Discrete Choice

Mohammad Ghaderi

January 2026

Attention-Entropy Random Utility: Endogenous Attention and Context Effects in Discrete Choice

Mohammad Ghaderi,[a,b,c]

[a]Department of Economics and Business, Pompeu Fabra University, [b]Barcelona School of Economics, [c]Barcelona School of Management, mohammad.ghaderi@upf.edu

Abstract. This paper introduces the attention-entropy random utility (AERU) model, a behavioral model of discrete choice in which a decision-maker endogenously allocates attention across subsets of attributes in order to increase subjective confidence by reducing ex post choice uncertainty, and subsequently chooses an option based solely on the attended information. By endogenizing attention, the decision problem is reformulated from "which alternative to choose" to "which informational cues to process," with the observed choice emerging as the outcome of this attentional allocation. The AERU framework nests random utility model (RUM)-like behavior under transparent conditions, yet it is not restricted by Luce's independence of irrelevant alternatives (IIA), order-independence, or regularity. This flexibility enables AERU to capture key context effects in a disciplined manner and to generate sharp, testable predictions regarding the conditions for each context effect. From an empirical standpoint, AERU preserves the parsimony of the multinomial logit, requiring only a single additional attention parameter. Employing a scalable estimation procedure based on block coordinate ascent combined with a quasi-Newton method, I provide results from computational experiments demonstrating that AERU can produce better in-sample and out-of-sample predictions. Overall, AERU provides a flexible, parsimonious, and interpretable model of boundedly rational choice with a clear behavioral foundation and implications for context effects.

Funding: This work was supported by Ramón y Cajal grant, funded by the Spanish Ministry of Science and Innovation [RYC2021-034981-I], and the Severo Ochoa Programme for Centres of Excellence in R&D [CEX2024-001476-S].

Key words: discrete choice, endogenous attention, entropy, subjective confidence, random utility, context effects, regularity

1. Introduction

Making a choice inevitably involves trade-offs and conflicts. Alternatives differ across multiple attributes, and decision-makers (DMs) are often unsure how to trade off one attribute against another. Thus, preferences are constructed, at least partly, in context (Lichtenstein and Slovic 2006). Understanding choice behavior, therefore, requires understanding not only DMs' underlying tastes, but also how those tastes are expressed in the specific context of each choice.

I introduce the attention-entropy random utility (AERU) model, a behavioral framework in which context-dependent choice behavior emerges from an endogenous attention-allocation mechanism. AERU's choice behavior is driven by standard utility maximization combined with an *ex post* uncertainty-minimization through which the DM enjoys maintaining a high level of subjective confidence in their choice. These two behavioral principles capture, respectively, trade-offs and conflict resolution in choice. [1] AERU builds on the idea that the DM does not process the entire set of available attribute information symmetrically. Instead, they allocate attention across subsets of attributes in a way that favors information structures that reduce ex post uncertainty in the induced choice probabilities, in the sense of having a low entropy. In other words, choice is shaped by a tendency toward decisiveness, in which attention is endogenously skewed toward fragments of available information that make one option stand out. The model distinguishes between the information set available in the choice environment and the information subset that is effectively used for choice, while leaving the underlying taste parameters over attributes intact.

AERU implements this idea by specifying (i) a random utility representation at the attribute level, and (ii) a menu-dependent attention distribution over subsets of attributes, where the weight on each subset is decreasing in the entropy of the resulting choice probabilities. Overall choice probabilities are the mixture of within-subset multinomial logits under this attention distribution.

This structure has several implications. First, it gives a behavioral foundation for context effects. By endogenizing attention, context effects such as compromise and attraction emerge through systematic shifts in attention rather than ad hoc utility distortions. Second, it provides a unified structure connecting classic random utility models (RUMs), random attention or consideration models, and rational inattention. AERU retains the random utility core and applies an information-theoretic discipline to the distribution of attention

[1] There are two types of conflicts influencing the choice: (i) the within-alternative conflict, that is, the conflict among the set of attributes of a single alternative, and (ii) within-menu conflict, that is, the conflict due to choosing from competing options in a menu. The utility maximization aspect captures the first conflict type, whereas ex post choice-uncertainty minimization, or equivalently, subjective confidence maximization, addresses the second.

across attributes. Third, AERU is internally disciplined, it nests multinomial logit (MNL) as a limiting case when attention is uniform, and relaxes IIA, regularity, and order-independence in a transparent and testable way. Fourth, because the attention mechanism is defined in the preference space (over induced choice distributions) rather than directly on attribute contrasts, AERU is sensitive to which attributes actually matter for context effects, consistent with evidence from the attribute non-attendance and selective processing literatures, and hence is relatively robust to misspecification.

AERU is intentionally parsimonious. Compared to standard MNL, it introduces a single additional parameter that governs the strength of the entropy-based attention filter. This one parameter spans a continuum from "full" attention to all attributes, as in the RUM/MNL benchmark, to highly selective attention that amplifies context effects. Despite its behavioral richness, the resulting choice probabilities retain a simple form as mixtures of logits over attribute subsets. This yields a tractable, menu-sensitive model that is substantially more flexible than MNL, yet far more structured than fully nonparametric choice models. Moreover, I will show how the same framework can be extended to incorporate additional behavioral primitives, such as attribute-specific processing costs or complexity aversion, by altering the attention prior over attribute subsets without changing the core choice structure.

The contributions of the paper are threefold. First, I derive AERU from an explicit decision problem in which the DM chooses an attention strategy over attribute subsets subject to an entropy-based criterion on ex post choice uncertainty. This delivers the AERU representation as the choice rule of a conflict-resolving, attention-allocating agent. Second, I analyze the theoretical properties of AERU, characterizing when and how it departs from classical random utility. I show that it can capture key context effects, such as similarity and compromise, as well as particular patterns of attraction. I then identify conditions under which specific effects arise or are excluded, and establish links between these effects and the entropy profile of menus and the strength of the attention filter. Moreover, I characterize the role of the attention parameter and establish limit and convergence-rate results that clarify the geometry of AERU choice probabilities relative to the MNL benchmark and extreme attention regimes. Third, I develop a scalable maximum-likelihood estimation procedure and evaluate AERU in a computational experiment using finite choice data and based on in-sample and out-of-sample fit.

The remainder of the paper is organized as follows. Section 2 situates AERU within the literature on discrete choice, inattention, and context-dependent preferences. Section 3 introduces the primitives. Section 4 formally defines the AERU model. Section 5 presents the behavioral decision problem from which AERU arises. Section 6 studies context effects, e.g., similarity, compromise, and attraction-type patterns, and links AERU predictions to classical axioms such as regularity, IIA, and order-independence. Section 7 analyzes the attention filter. Section 8 describes the estimation procedure, and Section 9 reports the computational experiment. Finally, I conclude and discuss future directions in Section 10.

2. Related Literature

Understanding human choices is a central theme in mainstream economics (Samuelson 1938, Richter 1966, McFadden 2001), decision theory (Fishburn 1970), operations research (Ben-Akiva and Lerman 1985, Train 2009, Farias et al. 2013, Berbeglia et al. 2022, Ghaderi et al. 2025), marketing (Louviere et al. 2000, Toubia et al. 2003), and psychology (Lichtenstein and Slovic 2006, Bettman et al. 1998), with applications in areas such as pricing, revenue management, healthcare, product development, and transportation (Talluri and Van Ryzin 2004, Hensher 1994, Ben-Akiva and Bierlaire 1999, de Bekker-Grob et al. 2012).

Conventional random utility models, such as logit-based models, posit that the DM evaluates all available information relevant to their choice, which, in conjunction with their individual taste parameters, determines their choice. Central to these models are the assumptions of *regularity* and *Luce IIA* (Block and Marschak 1959, Luce 1959). Regularity suggests that adding more options to a menu should not increase the popularity of any existing option. Luce's IIA asserts that the relative popularity of two options remains unchanged regardless of the availability of other options. A large body of empirical and experimental research over the past few years has documented systematic departures from these implications. As an early example, in their experiment, Tversky and Shafir (1992) observes that adding a higher-quality music player reduces the share of a previously popular low-price music player and increases "no choice," violating regularity. When both options are present, conflict between price and quality appears to trigger deferral. Similar context effects have been observed outside the lab (Simonson and Tversky 1992, Kivetz et al. 2004, Wu and Cosguner 2020, Webb et al. 2021).

Motivated by such findings and by the aim of developing realistic choice models, an active interdisciplinary research area has emerged in recent years, focusing on bounded rationality in choice. In the domain of discrete choice, the boundedly rational behavior often arises when the menu as a whole serves as a unique stimulus, conveying information beyond its constituent elements. This phenomenon is referred to as the *context effect*. In this literature, AERU sits at the intersection of random choice (Marschak 1959, Luce 1959, McFadden 1973) and context-dependent choice behavior (Simonson and Tversky 1992, Tversky and Simonson 1993), with an endogenous attention mechanism at its core.

Building on standard random utility primitives, AERU contributes to the context effect literature (Simonson and Tversky 1992, Tversky and Simonson 1993, Kivetz et al. 2004, Rooderkerk et al. 2011, Simonson 2014) by offering a structurally transparent account in which context-dependence arises from an endogenous attention allocation that favors attribute subsets yielding low ex post choice uncertainty. The attention mechanism is driven by the DM's desire to maintain a high level of subjective confidence in their choice by resolving within-menu conflict when choosing from competing options. Thus, conflict resolution remains a central concept in AERU.

In an attempt to understand how people resolve conflict in binary choice, Slovic (1975) conducted an experiment in which subjects were asked to choose between pairs of alternatives that they had previously equated in value, thereby involving a high level of preferential conflict. They report that subjects tend to resolve conflicts by choosing the option that is superior on the more important dimension. Consistent with reason-based choice (Shafir et al. 1993, Dietrich and List 2016), people seek justifiable grounds for their decision. In other words, people resolve such conflicts by seeking *reasons* for choosing one option over another. In AERU, this appears as attention shifting toward attribute subsets that better discriminate between options in the preference space, thereby producing more decisive (lower-entropy) judgments. Thus, AERU relates to reason-based choice where the *reasons* to attend to or ignore particular subsets of attributes depend on how they change ex post uncertainty in induced choice probabilities over the options in the menu.

In choice among multiple options, when conflict among competing alternatives is difficult to resolve through direct comparison, decision-makers leverage menu composition to *construct* reasons in favor of one option. For instance, adding new options to the menu can make certain trade-offs between attributes more justifiable. This gives rise to context-dependent choice behavior. AERU captures such patterns through its entropy-based attention rule, in which menu composition shapes the reasons for attending specific attribute subsets based on the choice uncertainty they yield. The same attribute subset may enhance or diminish induced choice uncertainty, depending on which other options are available in the menu.

Context dependence often leads to choice behavior that departs from standard predictions of random utility models. A prominent account is salience theory (Bordalo et al. 2013, 2020, 2022), which explains context effects through context-sensitive reweighting of attribute differences based on contrast or prominence. Compared to salience models, AERU holds tastes fixed and generates context dependence by endogenously allocating attention across *collections* of attributes, thereby reshaping the induced choice probability distributions. The two perspectives are conceptually related. Salience emphasizes taste reweighting in the attribute space, while AERU emphasizes attention allocation in the preference space.

AERU is also related to rational inattention (Sims 2003). Similar to the inattention models, AERU uses information-theoretic discipline on attention (Caplin 2016, Caplin et al. 2022), but applies it to the attribute subsets of options within a menu rather than to signals about states. This shift allows AERU to generate context effects, whereas rational-inattention logit models inherit IIA (Matějka and McKay 2015). Conceptually, AERU aligns with the endogenous attention view (Gabaix 2019), operationalizing attribute sparsity (Gabaix 2014) via attention allocation over information subsets, that is also consistent with eye-tracking evidence on attribute non-attendance (Orquin and Loose 2013, DellaVigna 2009). [2]

Relatedly, Gul et al. (2014) models random choice as behavioral optimization using attribute-based primitives and weaker independence axioms. Their representation is not aimed at context-dependent choice behavior and preserves regularity, whereas AERU's menu-dependent attention permits disciplined violations of regularity.

[2] In cognitive terms, rational inattention formalizes top-down (deliberative, goal-directed) information acquisition. In contrast, AERU can be read as a decisiveness-seeking, stimulus-driven, and reflexive attention allocation, hence belonging to the bottom-up category of attention mechanisms. For more details on this categorization, see Loewenstein and Wojtowicz (2025).

On the other hand, AERU differs from consideration-set models and random attention over options (Manzini and Mariotti 2014, Caplin et al. 2019, Cattaneo et al. 2020, Abaluck and Adams-Prassl 2021, Gallego and Li 2024). Those frameworks relate attention to awareness or product salience in the menu, and are well suited when many simple options compete, and limited consideration is the bottleneck. AERU is complementary. It models which information cues or product features are processed when options are complex, reallocating attention across attributes and thereby operating in the preference space rather than pruning the option set.

Several works in marketing and operations have examined limited information acquisition and attribute-level search. Branco et al. (2016) explores costly consumer information search and derives the seller's optimal information disclosure in situations where consumers can choose to search only a subset of attributes based on their ex ante valuation of the product. Ke et al. (2016) develops a framework for continuous information search, where expected utility depends on the attributes consumers decide to search for, with a search cost that is alternative-specific but fixed across attributes for the same alternative. Ning et al. (2025) explores the case of time-varying search cost. Other works have developed empirical models of choice under limited information and applied them to new product introduction (Joo 2023), pricing (Boyacı and Akçay 2018), and the evolution of product assortment (Natan 2025). These papers endogenize attribute inspection effort or stopping rules in the presence of search costs. AERU shares the focus on the attribute-level information processing mechanism but differs in its menu-dependent attention discipline, yielding closed-form mixture-of-logits probabilities rather than dynamic stopping rules.

Finally, sequential sampling and decision field theory views decision-making as a dynamic process in which the DM sequentially samples information (Roe et al. 2001) while shifting their attention back and forth between stimuli (Noguchi and Stewart 2018). AERU is static but consistent with the idea that attention shifts across attributes. From this perspective, the entropy-based attention mechanism can be interpreted as favoring stimuli that would reach a stopping boundary more decisively. For a concise discussion and eye-tracking evidence, see Krajbich (2019), Noguchi and Stewart (2014).

3. Primitives

Let A be a finite set, where each element $a \in A$ represents a choice option or alternative, including possibly a no-choice option. [3] Denote by $\mathcal{A}$ the collection of all non-empty subsets of A, where each $S \in \mathcal{A}$ represents a choice set or menu. When facing a menu $S \in \mathcal{A}$, the DM chooses an option $a \in S$ with a probability $\rho(a, S)$. [4] The function $\rho : A \times \mathcal{A} \to [0, 1]$, where $\rho(a, S) = 0$ for all $a \notin S$, and $\sum_n \rho(a_n, S) = 1$, is called a *stochastic choice function* (SCF). [5] A *choice model* is a parameterization of an SCF. Different parameterizations lead to different choice models.

A general class of choice models includes the popular random utility model, where parametrization of the SCF ρ is via a random vector $\boldsymbol{U}$ over alternatives such that:

$$\rho(a_n, S) = \mathbb{P}\left\{U_n = \max_{a_i \in S} U_i\right\}, \tag{1}$$

where $\mathbb{P}$ is a probability measure and U_n, the n-th component of $\boldsymbol{U}$, is the random utility for $a_n \in A$. Defining $v_n = \mathbb{E}[U_n]$, the nominal utility, and $\varepsilon_n = U_n - v_n$, and assuming *positivity*, that is $\rho > 0$, the model is expressed as:

$$\rho(a_n, S) = \mathbb{P}\left\{\varepsilon_n \geq \max_{i: a_i \in S}(v_i - v_n + \varepsilon_i)\right\}, \tag{2}$$

a formulation known as the *discrete choice model* (Ben-Akiva and Lerman 1985). Varying the joint distribution of the $\boldsymbol{\varepsilon} = (\varepsilon_n)_n$ yields different choice models (Train 2009). One of the most popular choice models is the MNL (McFadden 1973, 2001), which is obtained by assuming independent and identically distributed

[3] Throughout the paper, I use the terms *option* and *alternative* interchangeably.

[4] The stochasticity of such choice functions admits multiple interpretations. At the population level, it can represent taste heterogeneity among individuals. At the individual level, choices may appear *observationally stochastic* due to limited information from the analyst's perspective, inferential uncertainty, measurement errors, or *intrinsically stochastic* behavior arising from cognitive factors such as inattention, inertia, or intentional randomization as modeled in perturbed utility frameworks (Fudenberg et al. 2015). While interpretations vary across domains, the formalism and operationalization of stochastic choice functions remain consistent. Thus, they are natural modeling tools for observed choice behavior and are amenable to empirical validation.

[5] When values of ρ are restricted to $\{0, 1\}$, it collapses to a deterministic choice behavior. Moreover, if choices are consistent with the utility maximization principle, this deterministic choice function can be represented by a strict linear ordering on A.

Type-I Extreme Value ε_n terms. MNL yields a closed-form expression for the choice probabilities as follows: [6]

$$\rho_{MNL}(a_n, S) = \frac{e^{v_n}}{\sum_{a_i \in S} e^{v_i}}. \tag{3}$$

Following the MNL parametrization, $\rho_{MNL}(a_n, S)/\rho_{MNL}(a_m, S) = e^{v_n - v_m}$, is independent of S, suggesting that the choice probability ratio of two alternatives is independent of the presence or absence of any other alternative in the menu. This property, known as the IIA, is recognized as a restrictive assumption and often inconsistent with empirical observations. Moreover, the model implies that an option's choice probability cannot increase when more options are added to the menu, a property known as regularity or monotonicity. While MNL can be extended to accommodate non-IIA choices, for instance, as in the nested logit model (Ben-Akiva 1973, Galichon 2022), regularity remains a central assumption to the entire class of random utility models (Falmagne 1978).

However, research in behavioral economics and marketing has documented many situations in which regularity does not hold. In most such cases, the composition of the menu systematically influences choice. For instance, adding an asymmetrically dominated option, also known as a decoy, can increase the choice probability of a target superior option in the menu (Huber et al. 1982), adding an extreme option can make the compromise option more attractive (Simonson 1989, Kivetz et al. 2004), or introducing a new option can hurt similar options more than dissimilar options, hence creating preference for products that stand out (Tversky 1972). These observations, respectively known as the attraction, compromise, and similarity effects, highlight instances of *context-dependent preferences* arising from various underlying behavioral patterns, such as the trade-off contrast and extremeness aversion (Simonson and Tversky 1992) and salience (Bordalo et al. 2013). Context effects suggest that menus function as independent information signals beyond the information provided by their constituent elements and thus influence choices in ways that standard random utility models cannot explain.

[6] Note that under the random utility model (1), choice probabilities are invariant to any affine transformation of U, and thus the scale parameter in the variance of ε can be scaled to one.

4. Model

AERU builds on the idea that the information contained in the menu influences the DM's attention by shifting it to some subset of attributes in a systematic way. In other words, the DM does not rely on the same set of attributes in evaluating alternatives across different menus. I define an endogenous attention mechanism in the preference space that captures this structured context-dependent attention filter.

Two factors influence the attention mechanism: (i) the DM's taste parameters, and (ii) the ex-post uncertainty in choice. The first component is fixed and stable across menus, and captures substitution patterns among attributes. The second component concerns the extent to which a specific subset of attributes provides the DM with greater subjective confidence in their choice and builds on the observation that resolving a higher choice conflict entails a higher cognitive cost, requires more effort, and therefore imposes disutility.

Let $\mathcal{M} = \{1, 2, \cdots, M\}$ be a finite set of attributes or variables describing or evaluating the choice options, and $\mathcal{J}$ the collection of all non-empty subsets of $\mathcal{M}$. Moreover, let $\boldsymbol{\upsilon}_n = (\upsilon_{1n}, \upsilon_{2n}, \cdots, \upsilon_{Mn})$ be the vector of nominal utilities of a_n over the attributes. An attention filter is a mapping $f : \mathcal{A} \to \mathcal{J}$ that specifies the set of attributes in the attention set for a given set of options in a menu. Conditional on the attention set, that is, assuming that the DM considers attributes $J \in \mathcal{J}$ for evaluating alternatives in S, the choice probabilities follow the MNL form: [7]

$$\mathbb{P}(a_n|J,S) = \frac{e^{U_{Jn}}}{\sum_{s \in S} e^{U_{Js}}} \cdot \tag{4}$$

where $U_{Jn} = \sum_{j \in J} \upsilon_{jn}$.

The attention filter is inherently stochastic, and the DM might select a different subset J for a menu in repeated choice. Thus, J is a random variable. The probability of attending to a specific subset of attributes J is endogenously determined based on the degree of ex-post uncertainty it induces in choice. Endogeneity arises because taste parameters and the attention filter influence choice probabilities, whereas attention itself depends on those probabilities. I use the Shannon entropy measure to quantify uncertainty in the induced choice probability vector over the menu as follows.

$$H(J,S) = -\sum_{k \in S} \mathbb{P}(a_k|J,S) \times \log \mathbb{P}(a_k|J,S) \cdot \tag{5}$$

[7] With slight abuse of the notation, I use $s \in S$ instead of $s : a_s \in S$.

Given the taste parameters or the nominal utilities $(\boldsymbol{v}_n)_n$, a high entropy $H(J,S)$ suggests that attending the attribute subset $J \in \mathcal{J}$ when choosing from the menu S yields high ex post uncertainty, or, equivalently, low subjective confidence, in choice. Therefore, I define the attention filter, proportional to this uncertainty, as follows.

$$\mathbb{P}(J|S) = \frac{e^{-\alpha H(J,S)}}{\sum_{K \in \mathcal{J}} e^{-\alpha H(K,S)}}, \tag{6}$$

where $\alpha \in \mathbb{R}_{\geq 0}$ is a parameter of the attention filter. Thus, the DM probabilistically samples information from subsets of attributes, with sampling probabilities proportional to the negative ex-post uncertainty. Therefore, the stochastic choice function under the attention-entropy random utility model is given by:

$$\rho_{AERU}(a_n, S) = \sum_{J \in \mathcal{J}} \mathbb{P}(J|S) \times \mathbb{P}(a_n|J,S)\cdot \tag{7}$$

The AERU parameters are $\{\{v_{jn}\}, \alpha\}$, that is, taste parameters and the attention parameter. [8] Although it has only one additional parameter relative to MNL, I demonstrate that AERU is considerably more flexible and can capture non-IIA choice data and various forms of context effects.

5. The Behavioral Foundation: Choice as an Attention Allocation Problem

I will first focus on the subjective-confidence maximization component of choice behavior through endogenous attention and derive the posterior attention distribution. Then I will show that the attention-entropy random utility model emerges naturally as the output of a behavioral model in which, for every menu S, the DM solves a constrained optimization problem. Consider the following optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{w}} \mathcal{F}_{\{\alpha,\boldsymbol{u}\}}(\boldsymbol{w}) = \alpha \mathbb{E}_{J\sim \boldsymbol{w}}\Big(H(J,S)\Big) + KL\Big(\boldsymbol{w}||\boldsymbol{u}\Big) \\ &s.t. \;\; \boldsymbol{w} \in \Delta(\mathcal{J}) \end{aligned} \tag{8}$$

where H is the ex-post entropy in choice probabilities after fixing attention on a subset of attributes $J \in \mathcal{J}$, $\boldsymbol{u}$ is the DM's attention prior on $\mathcal{J}$, and $KL(.||.)$ is the Kullback-Leibler divergence between probability vectors. The model interpretation is that the DM is willing to pay a convex penalty, captured by the KL divergence, to lower the ex-post uncertainty in choice, where α is the price of uncertainty. Then, the DM's attention posterior is obtained as follows.

[8] Note that, except for additivity, the model does not assume any functional form for the utility function and v_{jn} can be any function of the j-th attribute level.

THEOREM 1. *For every menu S and $\alpha \geq 0$, problem (8) has a unique solution*

$$w_J^*(\alpha, \boldsymbol{u}) = \frac{u_J e^{-\alpha H(J,S)}}{\sum_K u_K e^{-\alpha H(K,S)}} \tag{9}$$

with optimal value $\log \sum_J u_J e^{-\alpha H(J,S)}$.

Proofs are provided in the Appendix.

COROLLARY 1. *Given $w_J^*(\alpha, \boldsymbol{u})$ in (9), the induced choice probability is $\rho(a_n, S) = \sum_J w_J^*(\alpha, \boldsymbol{u}) \mathbb{P}(a_n | J, S)$.*

The proof follows immediately from the law of total probability.

The following result shows how the AERU structure is directly derived from a joint randomization policy over the attribute subsets and choice options.

THEOREM 2. *For a menu S, let $\pi(J, n)$ be a joint randomized policy over the attribute subsets and choice options, and consider the following DM's problem.*

$$\min_{\pi \in \Delta(\mathcal{J} \times S)} \mathcal{G}_{\{\alpha, \boldsymbol{u}\}}(\pi) = \alpha \mathbb{E}_{J \sim \boldsymbol{\pi}(J)} \Big(H(J,S) \Big) + KL\Big(\pi(J) || \boldsymbol{u}\Big) + \mathbb{E}_{J \sim \boldsymbol{\pi}(J)} \Big(KL\big(\pi(\cdot|J) || P_J\big) \Big) \tag{10}$$

where $\pi(J) = \sum_{n \in S} \pi(J, n)$, $\pi(\cdot|J) = \Big(\pi(J,n)/\pi(J)\Big)_{n \in S}$ is the vector $\pi(n|J)$ for $n \in S$, and $P_J = \Big(\mathbb{P}(a_n|J,S)\Big)_{n \in S}$ is the MNL choice probabilities vector for a fixed set of attributes J. Then, $\mathcal{G}_{\{\alpha, \boldsymbol{u}\}}(\pi)$ is strictly convex in π, and has the unique minimizer

$$\pi^*(J, n) = w_J^*(\alpha, \boldsymbol{u}) P_J(n) \tag{11}$$

where

$$w_J^*(\alpha, \boldsymbol{u}) = \frac{u_J e^{-\alpha H(J,S)}}{\sum_K u_K e^{-\alpha H(K,S)}} \cdot$$

Moreover, the induced choice function is as follows.

$$\rho(a_n, S) = \pi^*(n) = \sum_J \pi^*(J, n) = \sum_J w_J^*(\alpha, \boldsymbol{u}) \mathbb{P}(a_n | J, S) \cdot$$

The proof is similar to that of Theorem 1 except for one additional term in the Lagrangian.

COROLLARY 2. *If the prior $\boldsymbol{u}$ is uniform, then the induced choice model is AERU.*

The proof follows immediately from Theorem 1 and by taking $u_J = 1/|\mathcal{J}|$. The induced choice function ρ is obtained following the law of total probability, $\rho(a_n, S) = \sum_J w^*_J(\alpha, \mathbf{1}/|\mathcal{J}|)\mathbb{P}(a_n|J, S)$.

The third term in (10) guarantees that the within-attribute-subset choice probabilities are RUM. The first two terms capture the subjective confidence-maximizing behavior by minimizing ex post choice uncertainty (the first term) and avoiding large deviations from the attention prior (the second term). Then, the posterior attention takes the form given in (9), in which AERU arises from a uniform attention prior.

5.1. Generalized AERU

Taking a uniform prior in (8) yields AERU. Nevertheless, AERU can be conveniently generalized by considering different priors. Here I provide two brief examples.

5.1.1. Complexity-aversion: Complexity can be defined by the cardinality of attribute subsets that the DM attends to make a choice. Minimal complexity occurs when the choice is based on a single attribute. Let the attention prior be

$$u_J(\gamma) = A(\gamma)\exp(-\gamma|J|) \tag{12}$$

where $A(\gamma)$ is the normalizing constant. Taking this prior gives

$$w^*(\alpha, \gamma) = \frac{e^{-\alpha H(J,S) - \gamma|J|}}{\sum_{K \in \mathcal{J}} e^{-\alpha H(K,S) - \gamma|K|}}, \tag{13}$$

implying complexity aversion since the DM is less likely to attend a higher cardinality attribute subset. At the same time, uncertainty aversion, avoiding attribute subsets that induce a higher ex-post entropy, remains part of the attention filter.

5.1.2. Attribute-specific information processing cost: Collecting and processing information from different attributes might entail different levels of difficulty. For instance, while information on product price is often readily accessible and has a clear numerical representation, information on product durability or quality is more subjective and harder to evaluate, thereby entailing a higher cognitive cost in the choice process. This cost may influence the degree of attention the DM allocates to an attribute. To capture the attribute-specific information processing cost in AERU, let the attention prior be

$$u_J(\boldsymbol{\lambda}) = A(\boldsymbol{\lambda})e^{-\sum_{j \in J} \lambda_j}, \tag{14}$$

with $\lambda_j > 0$ being the cost of obtaining or processing information from attribute j, and $A(\boldsymbol{\lambda})$ is the normalizing constant. Taking this prior gives

$$w^*(\alpha, \boldsymbol{\lambda}) = \frac{e^{-\alpha H(J,S) - \sum_{j \in J} \lambda_j}}{\sum_{K \in \mathcal{J}} e^{-\alpha H(K,S) - \sum_{k \in K} \lambda_k}}, \tag{15}$$

implying that the DM is less likely to attend attribute subsets that entail high-cost (large λ). A related setting in which some attributes of a product may be easier to evaluate than others is discussed in Bar-Isaac et al. (2012). For further discussions on inattention and information cost, see Matějka and McKay (2015), Huettner et al. (2019), Brown and Jeon (2024).

6. Context Effect

In this section, I analyze context effects through varying menu S while holding the attention filter parameter α fixed. I start with an initial menu S and expand the menu by adding a new alternative $d \in A \setminus S$, hence $S' = S \cup \{d\}$. Throughout this section, I fix the attention parameter and defer the analysis of the attention filter to the next section. Hence, I simplify the notation by denoting $\rho_n(S) = \rho(a_n, S)$, $w_J(S) = \mathbb{P}(J|S)$, $H_J(S) = H(J,S)$, and $P_J^S(n) = \mathbb{P}(a_n|J,S)$. Moreover, let $q_J = \mathbb{P}(d|J,S') = \frac{e^{U_{Jd}}}{\sum_{s \in S'} e^{U_{Js}}}$ be the within-subset MNL choice probability of the new option d, and $\Delta_J = H_J(S') - H_J(S)$ be the entropy change in subset J after adding option d.

PROPOSITION 1. *Expand a menu S by adding a new option d, $S' = S \cup \{d\}$. For any option $a_n \in S$*

$$\rho_n(S') - \rho_n(S) = \sum_J \underbrace{w_J(S)\Big(P_J^{S'}(n) - P_J^S(n)\Big)}_{\textit{within-subset MNL}} + \sum_J \underbrace{\Big(w_J(S') - w_J(S)\Big)P_J^{S'}(n)}_{\textit{attention shift}}. \tag{16}$$

Proposition 1, derived directly from the definition of $\rho = \sum w_J P_J$ and by adding and subtracting $\sum w_J P_J'$, states that the change in the choice probability of an existing option in a menu, after adding a new option, is the sum of two terms. i) The first term is the weighted average of within-subset MNL choice probability changes. Because MNL assumes regularity, this term is always negative. ii) The second term is the attention shift and captures the context effect. It can be positive or negative and is the only channel that can overturn regularity. Lemma 1 below describes the behavior of the attention shift.

LEMMA 1. *Let $S' = S \cup \{d\}$. Then for a fixed attention parameter α,*

$$w_J(S') = w_J(S)\frac{e^{-\alpha\Delta_J}}{\mathbb{E}_{w_J(S)}\left[e^{-\alpha\Delta_J}\right]}, \tag{17}$$

$$\Delta_J = -q_J \log q_J - (1-q_J)\log(1-q_J) - q_J H_J(S), \tag{18}$$

$$\textit{and } P_J^{S'}(n) = (1-q_J)P_J^S(n)\cdot \tag{19}$$

Therefore, under AERU, adding a new option redistributes attention based on changes in relative entropy, Δ_J. On the other hand, two factors shape Δ_J: the baseline entropy $H_J(S)$, and the within-attribute shares of the new option, q_J. Holding q_J fixed, Δ_J decreases linearly with H_J since $\frac{\partial}{\partial H_J(S)}\Delta_J = -q_J \le 0$. Holding H_J fixed, Δ_J is concave in q_J since $\frac{\partial}{\partial q_J}\Delta_J = \log\frac{1-q_J}{q_J} - H_J$ and $\frac{\partial^2}{\partial q_J^2}\Delta_J = -\frac{1}{q_J} - \frac{1}{1-q_J} \le 0$. Hence, across the attribute subsets, the attention reweighting is governed by the combined profile (H_J, q_J). Attribute subsets where the new option share q_J is small incur small Δ_J and gain attention, even if their baseline entropy H_J is low. Attribute subsets where the new option is more competitive, hence larger q_J up to a threshold $1/(1+e^{H_J})$, can suffer a larger Δ_J and lose attention, even if baseline entropy is high.

Next, I describe how the AERU choice probabilities, ρ, change when a new option is added to the menu.

LEMMA 2. *Let $S' = S \cup \{d\}$. Then for a fixed attention parameter α and any option $a_n \neq d$,*

$$\rho_n(S') = \frac{\mathbb{E}_{w_J(S)}\left[(1-q_J)P_J^S(n)e^{-\alpha\Delta_J}\right]}{\mathbb{E}_{w_J(S)}\left[e^{-\alpha\Delta_J}\right]}\cdot \tag{20}$$

Using Lemma 2 to characterize how ρ changes under menu expansion, we are now ready to derive our main result on changes in choice probabilities for the analysis of context effects.

THEOREM 3 **(AERU Context-Effect Covariance Identity).** *The AERU choice probability change for any option $a_n \in S$, $\Delta\rho_n = \rho_n(S\cup\{d\}) - \rho_n(S)$, when adding a new option $\{d\}$, is obtained as follows,*

$$\Delta\rho_n = \mathrm{Cov}_{w_J(S)}\left(P_J^S(n), K_J\right) - \mathbb{E}_{w_J(S)}\left[P_J^S(n)q_J K_J\right] \tag{21}$$

where $K_J = \frac{e^{-\alpha\Delta_J}}{\mathbb{E}_{w_J(S)}\left[e^{-\alpha\Delta_J}\right]}\cdot$

Notice that in the absence of attention filter, $K_J = 1$ for every J, the covariance term vanishes, and thus $\Delta\rho_n = -\mathbb{E}_{w_J(S)}\left[P_J^S(n)q_J\right] \leq 0$, satisfying regularity.

To analyze the AERU context effect, first note that under the Luce independence assumption and the MNL model, the following holds.

PROPOSITION 2. *Under IIA, for any $a_n \in S$ and $d \in A \setminus S$,*

$$\Delta\rho_n^{IIA} = \rho(a_n, S \cup \{d\}) - \rho(a_n, S) = -Q\rho(a_n, S) \tag{22}$$

where $Q = \rho(d, S \cup \{d\})$ is the choice share attained by the new option.

Therefore, under IIA, $\Delta\rho_n/\rho_n = -Q$ is the same for all $a_n \in S$, meaning that the new option draws choice shares symmetrically from all options in S. Below, I show how AERU can capture context effects in which options in S are asymmetrically affected by the new option.

Following Theorem (3) and Proposition (2), I define

$$D_n = \Delta\rho_n - \Delta\rho_n^{IIA}$$

in order to capture the difference in choice probability changes when adding a new option, relative to the baseline IIA without a context effect. If $D_n \equiv 0$ for all $a_n \in S$, then IIA holds and there's no context effect. A $D_n < 0$ implies that a_n loses more choice share compared to what no-context-effect predicts. This happens, for instance, in the case of the similarity effect, a classic example of context effect in which an option is hurt more by similar options than by dissimilar ones, implying preference for options that stand out (Tversky 1972). When $D_n > 0$, it implies that the loss in a_n choice share is smaller than what the no-context-effect predicts. This can occur in the case of the compromise effect, in which adding an extreme option increases the *relative* popularity of an existing compromise option (Tversky and Simonson 1993). It is important to notice that, despite a common misunderstanding, context effect may exist without violating regularity. In fact, if $0 < D_n < Q\rho_n$, then there is a positive context effect without violating regularity. The regularity axiom is violated only in the case of extreme context effects when $D_n > Q\rho_n$. For a discussion, see Frederick et al. (2014) and Ghaderi et al. (2025). Below, in Lemma 3 and the following section, I show that under AERU, D_n can take both negative and positive values, including violations of regularity, depending on the within-subset choice shares of the new option, q_J, and the values of the two covariance terms.

LEMMA 3. *For any $a_n \in S$,*

$$D_n = \mathrm{Cov}_{w_J(S)}\Big(P^S_J(n), K_J\Big) - \mathrm{Cov}_{w_J(S)}\Big(P^S_J(n), q_J K_J\Big)\cdot \tag{23}$$

6.1. What drives the context-effect?

Lemma 3 shows that the sign of D_n depends on how $P_J(n)$ covaries with attention reweighting factor $K_J = \frac{e^{-\alpha\Delta_J}}{\mathbb{E}_{w_J(S)}\left[e^{-\alpha\Delta_J}\right]}$ versus with the $q_J K_J$. On the other hand, K_J is increasing in $H_J(S)$ because $\frac{\partial}{\partial H_J(S)}\Delta_J = -q_J \leq 0$ (see the discussion of Lemma 1). Thus, the sign of D_n depends on the following two terms.

- The first covariance $\mathrm{Cov}(P^S_J(n), K_J)$ is positive when a_n tends to be strong (high $P^S_J(n)$) on higher-entropy subsets since K_J is increasing in $H_J(S)$.
- The second covariance term, $\mathrm{Cov}(P^S_J(n), q_J K_J)$, is positive when the new option captures more share precisely in subsets where a_n is strong.

Therefore, if the target option $a_n \in S$ is strong on high-entropy subsets and the new option is weak there, that is, $\mathrm{Cov}(P^S_J(n), H_J(S)) > 0$ and $\mathrm{Cov}(P^S_J(n), q_J) < 0$, then $\mathrm{Cov}(P^S_J(n), K_J) > 0$ while $\mathrm{Cov}(P^S_J(n), q_J K_J) < 0$, hence $D_n > 0$ and a_n benefits from adding the new option to the menu. This is the *compromise effect* pattern. On the other hand, if a_n is strong on low-entropy subsets and the new option too is strong there, then $\mathrm{Cov}(P^S_J(n), K_J) < 0$ and $\mathrm{Cov}(P^S_J(n), q_J K_J) > 0$, and $D_n < 0$. This is the *similarity effect* pattern. Finally, regularity can be violated only when $\mathrm{Cov}(P^S_J(n), H_J(S)) \gg 0$ and $\mathrm{Cov}(P^S_J(n), q_J) \ll 0$ so that

$$\mathrm{Cov}_{w_J(S)}\Big(P^S_J(n), (1-q_J)K_J\Big) > Q\rho_n(S)\cdot$$

Next, I present three numerical examples to illustrate the three important context effects in the literature: similarity (Tversky 1972), compromise (Simonson 1989, Kivetz et al. 2004), and attraction (Huber et al. 1982). Following the established measures in the literature (Tversky and Simonson 1993, Kivetz et al. 2004, Rooderkerk et al. 2011), I focus on changes in the popularity of a focal option relative to a competing option when menu composition changes in a particular way that generates a context effect.

EXAMPLE 1 (**SIMILARITY**). The similarity effect occurs when adding a similar option to the menu makes the dissimilar option more popular. That is, $\rho(A,\{A,B,A'\}) < \rho(A,\{A,B,B'\})$ where A' and B' are options *similar* to A and B, respectively, as depicted in Figure 1.

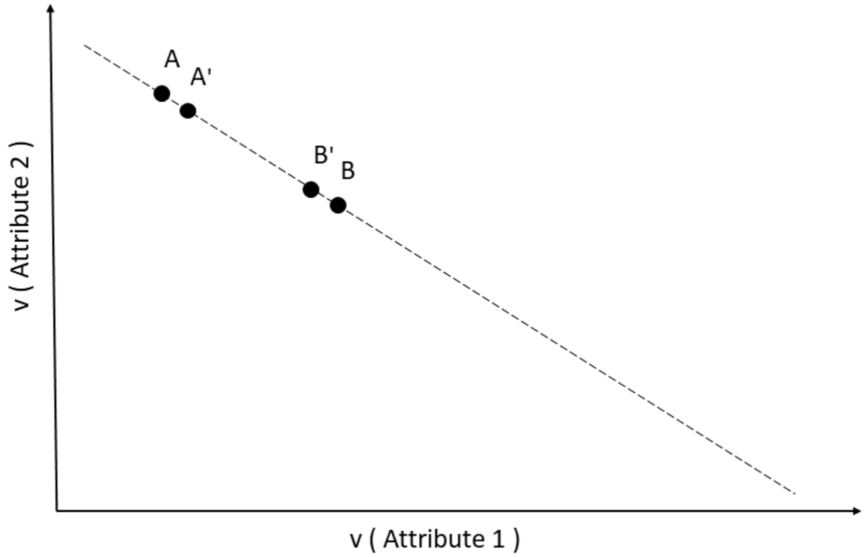

Figure 1 Similarity effect: The share of option A relative to option B is greater in the presence of B' than A'.

Let $\boldsymbol{v}_A = (0,9)$, $\boldsymbol{v}_B = (3,6)$, and the AERU attention parameter $\alpha = 2$. For the two similar options, let their utility values be $\boldsymbol{v}_{A'} = (0.1, 8.9)$ and $\boldsymbol{v}_{B'} = (2.9, 6.1)$. In the baseline menu $S_{base} = \{A,B\}$, $\rho^{IIA}(A,S_{base}) = \rho^{AERU}(A,S_{base}) = 1/2$. When adding either A' or B', the IIA choice share of each option becomes $1/3$. However, consistent with the prediction of the similarity effect, AERU yields different results. When adding the option similar to A, $\rho^{AERU}(A, S_{base} \cup A') = 0.21$ whereas $\rho^{AERU}(B, S_{base} \cup A') = 0.59$, meaning that adding A' hurts A while increases share of B. Hence, the popularity of A relative to B drops to $\frac{0.21}{0.21+0.59} = 0.26$ compared to the IIA prediction of 0.50. Conversely, When adding the option similar to B, $\rho^{AERU}(A, S_{base} \cup B') = 0.59$ whereas $\rho^{AERU}(B, S_{base} \cup B') = 0.21$, meaning that adding B' hurst B while making A more popular. Hence, the popularity of A relative to B increases to $\frac{0.59}{0.59+0.21} = 0.74$ compared to IIA prediction of 0.50. Hence, the share of A relative to B is 25.9% in set $\{A,B,A'\}$ versus 74.1% in set $\{A,B,B'\}$, a substantial similarity effect of 48.2%. These results are summarized in Table 1.

In both expansions, regularity is violated. When adding either A' or B' to the base menu $\{A,B\}$, the share of the dissimilar option increases. Specifically, in this example, while IIA predicts that the new option

Table 1 Summary of the results in Example 1 for the similarity effect. The new option draws disproportionately from the similar option, while increasing the share of the competing option.

Menu	$\rho^{IIA}(\cdot,S)$	ρ^{AERU}		$\frac{\rho(A,S)}{\rho(A,S)+\rho(B,S)}$
		A	B	
$S_{base} = \{A,B\}$	1/2	0.50	0.50	0.50
$S_{base} \cup A'$	1/3	0.208	0.594	0.259
$S_{base} \cup B'$	1/3	0.594	0.208	0.741

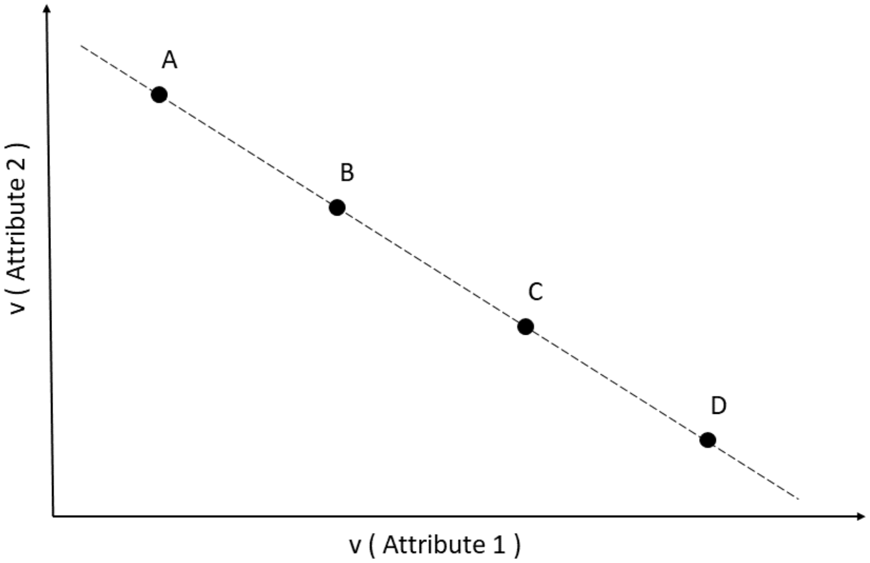

Figure 2 Compromise effect: The share of option B relative to option C is greater in the presence of A (when B is a compromise option) than in the presence of D (when B is an extreme option).

draws a share of $1/6$ from each options A and B, AERU predicts that the new option draws a share of 0.198 from the similar option and transfers an additional share of 0.094 from the similar option to the competing dissimilar option.

EXAMPLE 2 (**COMPROMISE**). The compromise effect occurs when the share of a focal option relative to a competing option increases as it becomes an intermediate option and decreases as it becomes an extreme option in the menu. That is, in Figure 2, the popularity of option B relative to option C is greater in menu $\{A,B,C\}$ than in menu $\{B,C,D\}$.

Let $\boldsymbol{v}_B = (0.50, 2)$, $\boldsymbol{v}_C = (0.55, 1)$, and the AERU attention parameter $\alpha = 10$. In the base menu $S_{base} = \{B, C\}$, $\rho^{IIA}(B, S_{base}) = 0.72$ and $\rho^{AERU}(B, S_{base}) = 0.69$. Now consider two different expansions of the base menu: one is obtained by adding a left-extreme option $\boldsymbol{v}_A = (0.45, 5)$ to S_{base}, and therefore making option B a compromise and C an extreme. The other menu is obtained by adding a right-extreme option $\boldsymbol{v}_D = (4, 0)$, making option B an extreme and C a compromise. Compromise effect predicts that $\frac{\rho(B,S)}{\rho(B,S)+\rho(C,S)}$ is larger in $S_{base} \cup A$ than in $S_{base} \cup D$. Under AERU, these relative shares are 72.6% and 51.1% in the $S_{base} \cup A$ and $S_{base} \cup D$ menus, respectively. Therefore, adding the extreme option A increased the share of B relative to C from 69.1% to 72.6%, whereas adding the extreme option D decreased it to 51.1%, indicating a compromise effect of 21.5%.

6.1.1. How to achieve a compromise effect? The compromise effect arises from two forces. First, adding an extreme option reduces the focal option's within-attribute-subset shares in the attributes where the new option is strong. In other words, adding an extreme option that is strong on some attributes shrinks the within-attribute share of both the focal and the competing option in those attributes. Second, because the extreme option lowers entropy within those attribute subsets, AERU shifts attention towards it. A relative gain for the focal option arises when this tilt shifts attention away from attributes where it is weak relative to the competing option and toward those where it is strong. Thus, if the focal option is stronger than the competing option in the attribute subset where the new extreme option is strong, this attention shift boosts its relative share.

Consequently, to achieve a compromise effect, the way the new option's extremeness is defined is crucial. For the extreme option to boost the relative share of a focal option, it must shift attention away from the attribute subsets in which the focal option is weak by increasing entropy; hence, to achieve this, it must be comparably attractive (similar utility values). Next, and more importantly, to shift the attention to the attributes where the focal option is strong, it must exhibit extreme attractiveness up to a certain point. If the new option becomes too extreme on any attribute, it dominates both the focal and competing options in all attribute subsets, thereby preventing high entropy in attribute subsets where the focal option is weak. To illustrate this, Figure 3 shows the compromise effect across different extremeness levels in $\boldsymbol{v}_D$ for the first

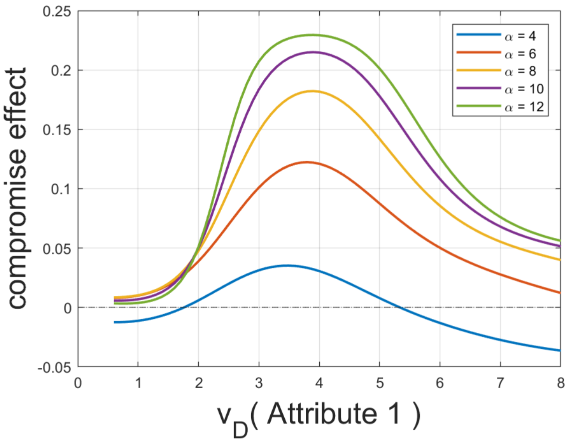

Figure 3 **(Color online) Compromise effect versus extreme option utility for different α parameter values.**

attribute, and across different levels of attention filter intensity. As discussed, the extreme option must be strong enough to shift attention toward its own favored attribute, and where the focal option is strong, but not so dominant as to annihilate the focal option everywhere.

6.2. Attraction

The attraction effect occurs when adding an asymmetrically dominated option, a decoy, increases the choice probability of the dominating option relative to a competing option. The attraction effect has been among the widely debated and controversial context effects regarding its robustness and replicability (Frederick et al. 2014). For a discussion, see Huber et al. (2014), Simonson (2014).

The asymmetric dominance in attraction effect requires the decoy to be dominated by the target option, but not by the competing option. For instance, in Figure 4, A_d is a decoy to A and B_d a decoy to B. The attraction effect predicts that the share of option A relative to B will increase by adding its decoy A_d to the menu $S_{base} = \{A, B\}$. If the decoy captures no choice share at all, this leads to a violation of regularity. This is how the attraction effect has been originally introduced (Huber et al. 1982), and I refer to it as the *absolute* attraction effect. Alternatively, in the psychology literature, the attraction effect has been measured as the difference in the relative popularity of A to B between two menu expansions, $S_{base} \cup A_d$ versus $S_{base} \cup B_d$. From this view, $\rho_A/(\rho_A + \rho_B)$ is expected to be larger in the first compared to the second menu expansion, and $\rho_B/(\rho_A + \rho_B)$ is expected to be larger in the second compared to the first menu expansion (Trueblood et al. 2013). I refer to this as the *relative* attraction effect. I show that AERU can

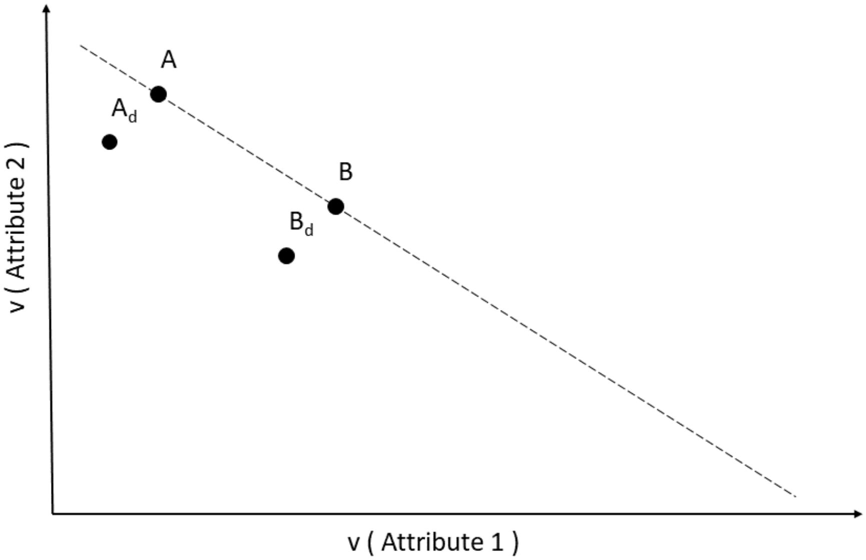

Figure 4 Absolute attraction effect: The share of an option, A or B, increases by adding its decoy, A_d or B_d, respectively, to the menu. Relative attraction effect: The share of option A relative to B is greater in the presence of its own decoy A_d than in the presence of the competing option's decoy, B_d.

capture the absolute attraction effect when the focal option satisfies certain conditions, but does not generate the relative attraction effect.

PROPOSITION 3. *Let $S_{base} = \{A, B\}$ and A_d be a decoy to A, that is $\upsilon_{jd} \leq \upsilon_{jA}$ for all j with strict inequality for at least one attribute. Suppose there exists a nonempty collection $\bar{\mathcal{J}}$ of attribute subsets such that:*

(i) $P_J^{S_{base}}(A) > P_J^{S_{base}}(B)$ for all $J \in \bar{\mathcal{J}}$,

(ii) for all J, $q_J \leq 1/(1+e^{H_J(S_{base})})$ so that Δ_J is increasing in q_J,

(iii) for all $J \in \bar{\mathcal{J}}$, the decoy's within-subset share q_J is sufficiently small;

(iv) for all $J \notin \bar{\mathcal{J}}$, q_J is uniformly larger than on $\bar{\mathcal{J}}$, that is, $\min_{J\notin\bar{\mathcal{J}}} q_J > \max_{J\in\bar{\mathcal{J}}} q_J$.

Then there exists $\alpha^* > 0$ such that for all $\alpha > \alpha^*$,

$$\Delta\rho_A \;=\; \rho\big(A,\, S_{base} \cup \{A_d\}\big) - \rho\big(A,\, S_{base}\big) \;\geq\; 0.$$

Proposition 3 states that for the attraction effect to occur, an option A must be the option that is stronger than B on low-entropy attributes, and the decoy A_d must capture sufficiently small choice probabilities on these attribute subsets, whereas sufficiently large shares on the rest of the attribute subsets. Therefore, in this menu, the attraction effect can be produced for A, but adding B_d to the menu $\{A, B\}$ need not produce an attraction effect.

Therefore, attraction is asymmetric and target–specific. It is *possible* but cannot be manufactured for everyone in the same menu. The sufficient conditions in Proposition 3 describe the *target–specific* nature of the attraction effect. They require a partitioning of attribute–subsets into a set $\bar{\mathcal{J}}$ on which the target A is stronger (than B) and the decoy's within–attribute shares are uniformly smaller than within its complement subsets where those shares are uniformly larger. For the *other* option B to satisfy the same conditions with its own decoy B_d, one would need the reverse partition to receive uniformly smaller decoy shares for B_d and larger shares elsewhere. These two requirements are generally incompatible for a given base menu $S = \{A, B\}$ and a fixed attention parameter α because they induce opposite attention shifts. Hence, absolute attraction can be produced for some target option(s) in S but, in general, not for all options. This observation is crucial and can partly explain the mixed results in the literature, where attraction is sometimes observed and at other times adding a decoy does not produce the effect (Frederick et al. 2014, Huber et al. 2014, Simonson 2014).

In conclusion, AERU accommodates context effects in a disciplined way. From Proposition 1 and Theorem 3, the change in choice shares decomposes into an MNL within–attribute dilution and an endogenous *entropy–weighted* attention reallocation; only the latter can overturn regularity. In summary: (i) *Similarity* arises when the new option captures more choice share precisely in attributes where the focal option is strong, shifting attention against it; (ii) *Compromise* arises when an extreme option lowers entropy more in attributes that favor the compromise option, shifting attention toward those attributes; (iii) *Attraction* (absolute) can occur only for some target options under transparent conditions (Proposition 3), but, in general, not all options in the same menu are qualified to become a target option for generating attraction effect. Moreover, the alternative *relative* attraction effect is ruled out, at least under strict decoys and attention over *all* attribute subsets. Nevertheless, this can potentially occur under the generalized AERU described in section 5.1. This requires further analysis.

7. The Attention Filter

In this section, I analyze the attention filter by varying the attention parameter while holding the menu S fixed. The aim is to link the endogenous attention mechanism to observable choice probabilities. The results serve three purposes. i) Theoretically, they provide necessary conditions for AERU-representability of an SCF ρ. Specifically, they provide easy-to-test conditions for identifying a stochastic choice function that does not belong to the AERU family. ii) Empirically, the results provide tools for partial identification of the attention parameter. Specifically, I identify bounds on the attention parameter directly from observable variables, such as choice shares and menu size. iii) Finally, they impose cross-environment restrictions relating AERU attention filters, useful when attention intensity varies with different intensities of attention filter, for instance due to different time pressure or engagement.

Holding the menu fixed for the analysis in this section, I simplify the notation by denoting $\rho_n(\alpha) = \rho(a_n, S)$, $w_J(\alpha) = \mathbb{P}(J|S)$, $P_J(n) = \mathbb{P}(a_n|J,S)$, and $H_J = H(J,S)$. Hence, the next observation follows immediately.

REMARK 1. $\rho_n = \mathbb{E}_{w_J(\alpha)}\big[P_J(n)\big] \in \mathrm{Conv}\Big(P_J(n)\Big)_J$. Thus, $\min_J P_J(n) \leq \rho_n(\alpha) \leq \max_J P_J(n), \forall\alpha$.

PROPOSITION 4. *For a fixed $(S, \{v_{jn}\})$, the map $\alpha \to \rho_n(\alpha)$ is real analytic on $\mathbb{R}$, hence infinitely differentiable with converging Taylor series for any α. Moreover*

$$\rho_n(0) = \frac{1}{|\mathcal{J}|}\sum_{J\in\mathcal{J}} P_J(n) \tag{24}$$

and

$$\rho_n(\infty) = \frac{1}{|\mathcal{J}^*|}\sum_{J\in\mathcal{J}^*} P_J(n) \tag{25}$$

where $\mathcal{J}^* = \{J : H_J = \min_K H_K\}$.

Remark 1 and Proposition 4 describe the geometry of ρ. It lives in the convex hull of the within-subset MNL choice probabilities, and the AERU choice probabilities for extreme attention parameters are the arithmatic means over all or the minimum entropy attribute subsets. The following Lemma is the main result of this section. It characterizes the behavior of ρ with respect to the attention filter and is used to derive subsequent bounds and convergence results.

LEMMA 4 **(AERU Attention Covariance Identity).** *For a fixed menu S and any alternative a_n,*

$$\frac{d\rho_n}{d\alpha} = -\mathrm{Cov}_{w_J(\alpha)}\Big(P_J(n), H_J\Big)\cdot \tag{26}$$

The attention covariance identity states that if an alternative is strong on attributes with low induced entropy and weak on attributes with high induced entropy, it gains choice share as the attention filter becomes stronger. In other words, whether strengthening the attention filter benefits or harms an option depends on its entropy profile with respect to the induced choice probabilities.

LEMMA 5 **(Global Bound).** *For a fixed menu S and any alternative $a_n \in S$ and any $\alpha \in \mathbb{R}_{\geq 0}$*

$$\left|\rho_n(\alpha) - \frac{1}{|\mathcal{J}|}\sum_{J\in\mathcal{J}} P_J(n)\right| \leq \frac{\log|S|}{4}\alpha\cdot \tag{27}$$

Thus, ρ lies in the convex hull of P_J within a Lipschitz distance of the centroid $\rho_n(0)$. Consequently, for the same menu S, observed at two environments with attention parameters α_1, α_2, the AERU attention parameters difference is bounded below by

$$|\alpha_2 - \alpha_1| \geq \frac{4}{\log|S|}\left|\rho_n(\alpha_2) - \rho_n(\alpha_1)\right|\cdot \tag{28}$$

The global bound in Lemma 5 is tight when $\rho_n(\alpha) = 0.5$. The following curvature-sensitive bound, although not global, provides tighter bounds for all values of ρ.

LEMMA 6 **(Curvature-Sensitive Bound).** *For a fixed menu S and any alternative $a_n \in S$ and any $\alpha_1, \alpha_2 \in \mathbb{R}_{\geq 0}$, $\alpha_2 \geq \alpha_2$*

$$\arcsin\sqrt{\rho_n(\alpha_2)} - \arcsin\sqrt{\rho_n(\alpha_1)} \leq \frac{\log|S|}{4}(\alpha_2 - \alpha_1)\cdot \tag{29}$$

Note that when $\rho_n(\alpha) = 0.5$, $\rho_n(1-\rho_n) = 1/4$, and the Bhatia-Davis inequality gives the same upper bound obtained for the global bound. The curvature-sensitive bound becomes increasingly tighter as the choice becomes more deterministic, that is, when ρ_n gets closer to 0 or 1.

Lemma 5 provides a global bound for ρ_{AERU} in relation to the base choice model $\rho(0)$ from Proposition 4. The following lemma provides the upper bound and an exponential convergence rate in relation to the other extreme choice model $\rho(\infty)$.

LEMMA 7 **(Exponential Convergence Rate).** *For a fixed menu S, let $\mathcal{J}^* = \{J : H_J = \min_K H_K\}$. Define the entropy gap $\Delta = \min_{J \notin \mathcal{J}^*}(H_J - \min_K H_K)$. Then, for any p-norm $||.||_p$,*

$$\left\| \left| \rho_n(\alpha) - \frac{1}{|\mathcal{J}^*|} \sum_{J \in \mathcal{J}^*} P_J(n) \right| \right\|_p \leq 2^{1/p} C e^{-\alpha\Delta} \tag{30}$$

for some C depending only on $|\mathcal{J}^*|$. Specifically, for L^∞ norm,

$$\max_{n \in S} \left| \rho_n(\alpha) - \frac{1}{|\mathcal{J}^*|} \sum_{J \in \mathcal{J}^*} P_J(n) \right| \leq C e^{-\alpha\Delta}.$$

8. Estimation

In this section, I describe the AERU maximum-likelihood estimation using block coordinate ascent and a quasi-Newton method. Let $\{(a_t, S_t)\}_{t=1}^T$ denote a finite collection of observed choices where each element (a_t, S_t) reads as option $a_t \in S_t$ was chosen when the menu S_t was presented. The corresponding likelihood function is defined as:

$$\mathcal{L}(\alpha, \{\upsilon_{jn}\}) = \prod_{t=1}^{T} \sum_{J \in \mathcal{J}} \Big(\frac{e^{-\alpha H(J, S_t)}}{\sum_{K \in \mathcal{J}} e^{-\alpha H(K, S_t)}} \Big) \Big(\frac{e^{\sum_{j \in J} \upsilon_{jt}}}{\sum_{s \in S} e^{\sum_{j \in J} \upsilon_{js}}} \Big) \tag{31}$$

where

$$H(J, S_t) = -\sum_{k \in S} \Big(\frac{e^{\sum_{j \in J} \upsilon_{jk}}}{\sum_{s \in S} e^{\sum_{j \in J} \upsilon_{js}}} \Big) \Big(\sum_{j \in J} \upsilon_{jk} - \log \sum_{s \in S} e^{\sum_{j \in J} \upsilon_{js}} \Big).$$

This likelihood function is nonconvex in the unknown parameters $\{\{\upsilon_{jn}\}, \alpha\}$.

Notice that no specification assumptions were made on the shape of utility functions. Nevertheless, if the utility functions are parametrized, then the likelihood function can be easily modified by replacing the taste parameters υ_{jn} with the utility function parameters. For instance, for a linear utility function $\upsilon_n = \boldsymbol{\theta}^T \boldsymbol{z}_n$, where $\boldsymbol{\theta}$ is the vector of utility parameters and $\boldsymbol{z}_n$ is the vector of attribute values for option a_n, $\upsilon_{jn} = \theta_j z_{jn}$ and therefore the taste parameters $\{\upsilon_{jn}\}$ will be replaced by the utility function parameters $\{\theta_j\}$ in the likelihood function. In the numerical experiment in the next section, a linear utility function is used, and the model parameters are estimated by solving the following constrained optimization problem.

$$\max_{\alpha \geq 0, \boldsymbol{\theta}} \mathcal{L}(\alpha, \boldsymbol{\theta}), \tag{32}$$

where $\mathcal{L}$ is obtained by replacing υ_{jn} elements with $\theta_j z_{jn}$ in the likelihood function (31).

To solve the optimization problem (32), an iterative procedure combining the block coordinate descent method (Wright 2015) with a quasi-Newton method is employed. At each iteration, I first maximize the likelihood function with respect to α, holding $\boldsymbol{\theta}$ fixed. I then maximize the likelihood function with respect to $\boldsymbol{\theta}$, holding α fixed at its value from the previous step. Maximizing with respect to α is straightforward since this is a univariate optimization problem with a sign constraint. I solve the optimization problem in the second step, updating the $\boldsymbol{\theta}$ parameters, using the Newton method with the BFGS (Broyden-Fletcher–Goldfarb–Shanno) approximation of the Hessian matrix. Therefore, this solution method requires only the first-order partial derivatives of the likelihood function with respect to α and $\boldsymbol{\theta}$, which can be computed easily since the stochastic choice function 7 has a closed form. It is easy to show that the likelihood function improves in each iteration since

$$\mathcal{L}(\alpha^{itr+1}, \boldsymbol{\theta}^{itr+1}) \underset{\text{by maximizing .w.r.t } \boldsymbol{\theta}}{\geq} \mathcal{L}(\alpha^{itr+1}, \boldsymbol{\theta}^{itr}) \underset{\text{by maximizing .w.r.t } \alpha}{\geq} \mathcal{L}(\alpha^{itr}, \boldsymbol{\theta}^{itr})$$

and therefore the solution from each iteration weakly dominates the previous iteration solution. [9] I repeat this process until improvement in the likelihood function falls below a prespecified threshold, set to 10^{-3} in the simulation analysis in section 9.

9. Simulation

To evaluate AERU performance, this section presents a computational experiment using synthetic choice data generated under various configurations. Following the setup in Ghaderi et al. (2025), I employed a random lexicographic model (Tversky 1972, Kohli and Jedidi 2007) since i) it is a flexible non-compensatory choice model capable of generating non-IIA and context-dependent choices, and ii) it closely resembles a realistic choice process.

[9] The block coordinate descent is not guaranteed to find the global maximum in nonconvex cases. In the case of our analyses, it consistently achieved a better likelihood value compared to using the BFGS quasi-Newton method alone to solve the optimization problem (32), but it also nearly doubled the solution time.

9.1. Setup

To generate the choice data, I first draw attribute weights from a Dirichlet distribution, $(\alpha_1, \ldots, \alpha_M) \sim \text{Dir}(\mathbf{1}_M)$, where $\mathbf{1}_M$ denotes an M-dimensional vector of ones. For each choice task S, attribute priorities are then generated according to the following rule:

$$\text{Priority}_m = \log(\alpha_m) - \log(-\log(q_m)), \tag{33}$$

where $q_m \sim \text{Uniform}(0, 1)$. The resulting ordered priority values determine the sequence in which alternatives are screened for that particular choice task S. A new priority sequence is generated independently for each choice task by drawing a new uniform random number q_m, for each attribute m, and then by applying Eq. (33).

In each iteration of the simulation study, the menu length is set to 3 or 6, then 25 menus of that length are randomly generated. Each menu option is randomly generated from $M = 4$ or 6 attributes, each with five levels. Menus are constructed to ensure that no alternatives are dominated within a menu. For each menu, I randomly generated 40 choice tasks according to the random lexicographic model described above. Therefore, a total of 1000 choice instances are generated in each replication. I then randomly partitioned the menus into the training and test sets. The training set included only the choice data from menus assigned to the training condition. The training set comprises 15 or 20 menus out of the 25. For each setting, I repeated the process 50 times, hence a total of $2 \times 2 \times 2 \times 25 \times 40 \times 50 = 40,000$ synthetic $(menu,\ choice)$ pairs.

9.2. Results

I report results on choice probability estimation using the mean absolute error (MAE) for in-sample and out-of-sample menus, with MNL as a benchmark. The results show that AERU improvement of out-of-sample MAE ranges from 20.1% to 30.7% across the simulation settings, with the lowest improvements in the settings with many attributes (6), and the highest improvements in the setting with few attributes (4) and a large training set (20 menus). The average AERU improvement of out-of-sample MAE is 24.6% (and $22.7\% - 26.3\%$ for 95% bootstrap confidence interval with $10,000$ replications).

Similarly, for in-sample, AERU consistently provides better results, with MAE improvement ranging from 24.0% (in the settings with many attributes (6) and large menu length (6)) to 32.1% (in the setting with

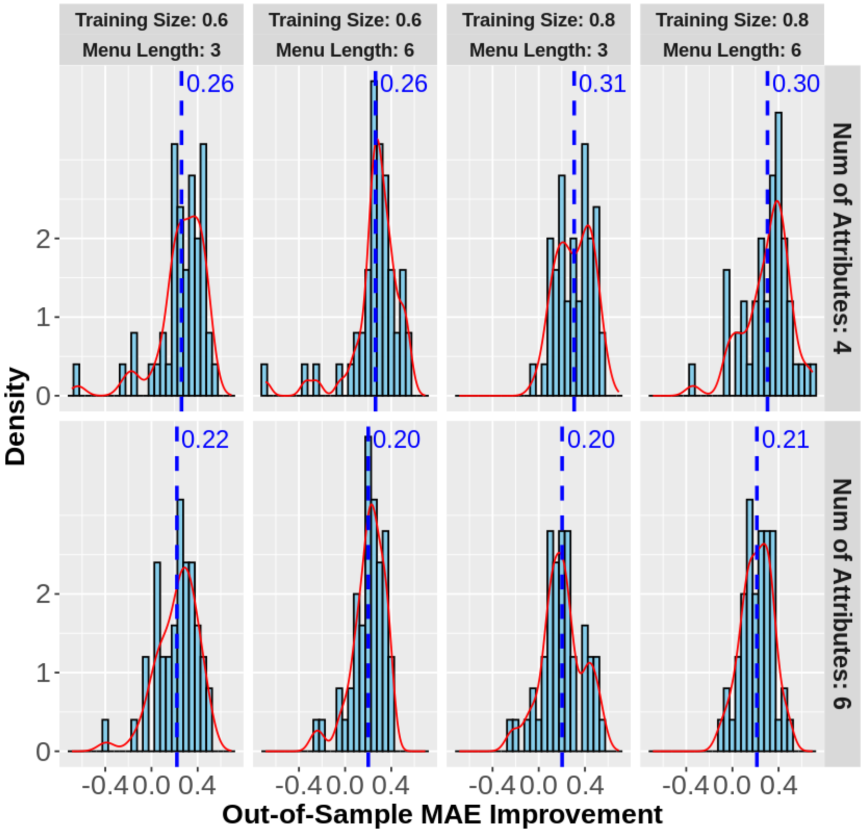

Figure 5 Distribution of out-of-sample MAE improvement by Number of Attributes (4 or 6 attributes), Training Size (60% or 80% of total 25 menus), and Menu Length (3 or 6 options in the menu). Dashed lines show the mean values, which are also displayed in the graph.

few attributes (4), large training set $(20$ menus), and large menu length (6)), with an average improvement of 27.7% (and $26.1\% - 29.1\%$ for 95% bootstrap confidence interval with $10,000$ replications). Figure 5 shows the distribution of out-of-sample MAE improvement for different simulation settings.

Moreover, Figure 6 shows a positive relationship between the estimated attention filter parameter (α) and the out-of-sample MAE improvement (Pearson correlation coefficient 0.107, p-value 0.034), suggesting that improvements over the benchmark MNL come from instances where the attention filter needed to be activated more strongly.

The results, summarized in Table 2, show that AERU consistently provides better in-sample and out-of-sample predictions compared to the benchmark MNL model.

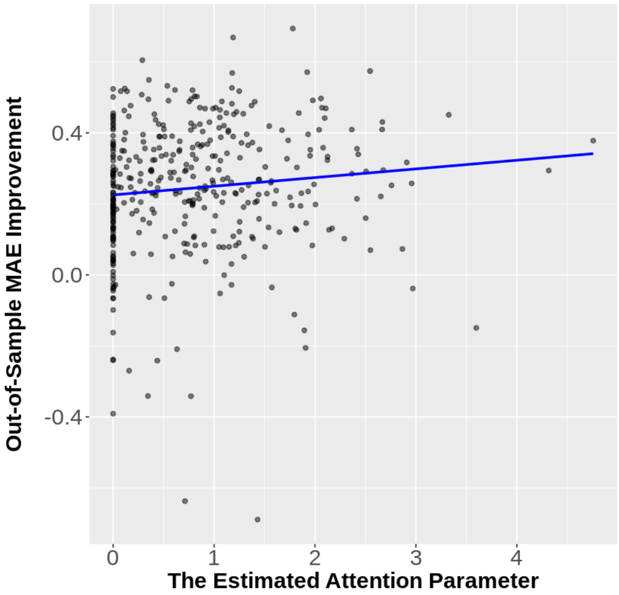

Figure 6 **Correlation between the out-of-sample MAE improvement and the attention parameter (α).**

Table 2 **Mean Model Performance by Number of Attributes, Training Size, and Menu Length. Lower MAE indicates a better performance.**

Num. of Attributes	Training Size	Menu Length	Out-of-Sample MAE		In-Sample MAE		Log Likelihood	
			AERU	MNL	AERU	MNL	AERU	MNL
Few	Small	Small	0.096	0.132	0.085	0.119	-451	-496
		Large	0.070	0.098	0.064	0.095	-694	-803
	Large	Small	0.094	0.138	0.087	0.129	-619	-689
		Large	0.067	0.101	0.066	0.096	-937	-1079
Many	Small	Small	0.094	0.123	0.081	0.114	-475	-512
		Large	0.067	0.085	0.060	0.080	-743	-802
	Large	Small	0.099	0.126	0.086	0.117	-668	-715
		Large	0.065	0.084	0.061	0.081	-991	-1075

In summary, these results complement the previous theoretical results and show that, while preserving the parsimony of MNL, AERU consistently fits and predicts better.

10. Concluding Remarks

This paper develops a behavioral model of discrete choice combining random utility maximization with a novel subjective-confidence maximization. The resulting choice model *endogenizes attention* and captures various context effects. Subjective-confidence maximization can be viewed as a drive for decisiveness, in which the decision maker allocates attention across *attribute subsets* and favors those subsets that yield low ex post uncertainty in the induced choice probabilities. Choice emerges as a mixture of within-subset logits, weighted by a menu-sensitive attention filter. The model preserves interpretable taste parameters, nests MNL as a limit case, and relaxes regularity, IIA, and order-independence in a transparent and testable way.

On the theoretical side, I demonstrate how AERU departs from the classical random utility model, why this departure is important for explaining boundedly rational choice behavior, and how it achieves this in a disciplined way. The changes in AERU choice probabilities under menu expansion decompose into (i) dilution within attribute subsets (regularity) and (ii) an entropy-based attention shift (the only source of regularity violations). This yields sharp conditions for context effects. I show how AERU can capture the three important context-effects, compromise, similarity, and an *absolute* attraction effect only for options that satisfy certain conditions in the menu, alongside a negative prediction for the *relative* decoy effect under strict dominance. These results can also be used to reconcile previous mixed empirical findings (Frederick et al. 2014, Simonson 2014, Huber et al. 2014).

Empirically, AERU is parsimonious. Relative to MNL, it adds a single attention parameter while substantially enlarging the set of behaviors the model can capture. Moreover, as the computational experiments show, it consistently improves both in-sample (MAE improvement from 24% to 32%) and out-of-sample (MAE improvement from 20% to 31%) fit while maintaining interpretability.

Future directions. This paper focuses on modeling, theory, and operationalization of AERU, while a full algorithmic treatment and computational analysis is deferred. Exact evaluation of AERU requires summing over all 2^M attribute subsets. This motivates sparsity and screening. Two practical strategies can be employed to control the computational cost: *(i) Cardinality truncation* by restricting attention to attrubute subset with small cardinality, $|J| \leq K$ (a special case of complexity aversion). This reduces the worst-case count to $\sum_{k=1}^{K} \binom{K}{k}$ and has a clear behavioral interpretation. *(ii) Sparse mixing* by approximating ρ as a convex combination of a few low-entropy attribute subsets of any cardinality. To populate the candidate set in the latter approximation strategy, heuristic search methods can be used to minimize approximation error by adding one attribute subset at a time, for instance via column generation or Frank-Wolfe method. Both cardinality truncation and sparse mixing approximations preserve the model's behavioral content while keeping computation tractable when the number of attributes is very large. A systematic study of the accuracy–speed trade-offs and finite-sample guarantees for these approximation strategies is a promising avenue for future research.

AERU offers falsifiable, managerially relevant predictions. Because the attention shift is disciplined by entropy in preference space, the model predicts *when* assortment changes should increase a focal option's share (and when they should not), provides conditions for context effects, and delivers bounds on their magnitude as functions of attention intensity and the menu's entropy profile. There are at least three promising avenues:

1. **Experimental tests of the mechanism.** Tests that manipulate menu composition to alter entropy profile can be used to evaluate the AERU predicted choice share shifts for attraction and compromise effects, alongside its null prediction for relative attraction under strict dominance. The results have immediate implications for assortment and information design.
2. **Consumer search behavior.** AERU links naturally to reason-based choice, and thus can be used to integrate behavioral insights into standard search models. [10] From the AERU perspective, search

[10] Consider this example from Tversky and Shafir (1992). They reported that the DM is more likely to pay a cost to receive a new lottery, that is, incur a search cost to discover a new option, when choosing between (A) Winning \$15 with 65% chance and (B) Winning 35 with 30% chance, compared to when choosing between (A) Winning 15 with 65% chance and (C) Winning 14 with 65% chance. In the latter case, the DM has a good reason to select option (A), whereas in the former case, this reason is absent, and thus the DM continues the search. Notice the considerable difference between this search rule and the standard view, where the search terminates as soon as the expected value of continuing the search exceeds its cost.

should terminate when choice conflict is sufficiently resolved, rather than when search cost exceeds expected improvement in decision making. Embedding AERU into consumer search, by incorporating insights from the rapidly growing literature at the intersection of consumer search and consideration set formation (Zwick et al. 2003, Onzo and Ansari 2025, Kosilova and Alptekinoğlu 2025), can yield novel results, particularly for assortment policies in markets with complex products.

3. **Richer attention primitives.** The same attention architecture accommodates attribute-specific processing costs, complexity aversion, or heterogeneity in the attention parameter. These extensions retain the AERU behavioral interpretation while broadening its applicability.

Finally, a natural next step is to embed AERU into *assortment optimization* problems, where a firm chooses a menu S to maximize its expected revenue. Generally, this has two parts: *estimation* of a choice model that captures substitution, and *optimization* of the assortment using price or profit information and the choice-share estimates for each feasible option. AERU is well-suited here because it treats assortments not merely as availability sets but as attention-shaping instruments influencing the choice behavior. By changing the menu, the firm shifts attention across attribute subsets and can, under transparent conditions, improve the choice share of high-margin items. Thus, in an AERU-based assortment optimization framework, the estimation component already encodes the levers that the optimization problem would exploit. Methodologically, because AERU's menu-dependent attention can violate regularity, classical structural properties exploited in MNL-type models, such as revenue-ordered optimality, need not apply, creating new algorithmic challenges. Developing this theory, scalable solution algorithms, and revenue implications is an appealing direction for future work.

Overall, AERU provides a behaviorally grounded, interpretable, and empirically tractable model of bounded rationality in choice. It augments random utility maximization with a novel subjective-confidence maximization, implemented via an endogenous, entropy-based allocation of attention across attribute subsets. The model preserves the parsimony of standard RUM (Louviere et al. 2000, McFadden 2001), while capturing key context effects and offering flexibility comparable to that of the nonparametric choice models (Farias et al. 2013, Ghaderi et al. 2025, Susan et al. 2025). In practice, it is substantially more flexible than MNL yet far more structured than fully nonparametric approaches.

References

Abaluck J, Adams-Prassl A (2021) What do consumers consider before they choose? identification from asymmetric demand responses. *The Quarterly Journal of Economics* 136(3):1611–1663.

Bar-Isaac H, Caruana G, Cuñat V (2012) Information gathering externalities for a multi-attribute good. *The Journal of Industrial Economics* 60(1):162–185.

Ben-Akiva M, Bierlaire M (1999) Discrete choice methods and their applications to short term travel decisions. *Handbook of Transportation Science*, 5–33 (Springer).

Ben-Akiva M, Lerman SR (1985) *Discrete Choice Analysis: Theory and Application to Travel Demand* (Cambridge, MA: MIT Press).

Ben-Akiva ME (1973) *Structure of passenger travel demand models.* Ph.D. thesis, Massachusetts Institute of Technology.

Berbeglia G, Garassino A, Vulcano G (2022) A comparative empirical study of discrete choice models in retail operations. *Management Science* 68(6):4005–4023.

Bettman J, Luce M, Payne J (1998) Constructive consumer choice processes. *Journal of Consumer Research* 25(3):187–217.

Block HD, Marschak J (1959) Random orderings and stochastic theories of response. Discussion Paper 66, Cowles Foundation for Research in Economics, Yale University, New Haven, CT.

Bordalo P, Gennaioli N, Shleifer A (2013) Salience and consumer choice. *Journal of Political Economy* 121(5):803–843.

Bordalo P, Gennaioli N, Shleifer A (2022) Salience. *Annual Review of Economics* 14(1):521–544.

Bordalo P, Gennaioli N, Shleifer A, et al. (2020) Memory, attention, and choice. *The Quarterly Journal of Economics* 135(3):1399–1442.

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

Branco F, Sun M, Villas-Boas JM (2016) Too much information? information provision and search costs. *Marketing Science* 35(4):605–618.

Brown ZY, Jeon J (2024) Endogenous information and simplifying insurance choice. *Econometrica* 92(3):881–911.

Caplin A (2016) Measuring and modeling attention. *Annual Review of Economics* 8(1):379–403.

Caplin A, Dean M, Leahy J (2019) Rational inattention, optimal consideration sets, and stochastic choice. *The Review of Economic Studies* 86(3):1061–1094.

Caplin A, Dean M, Leahy J (2022) Rationally inattentive behavior: Characterizing and generalizing shannon entropy. *Journal of Political Economy* 130(6):1676–1715.

Cattaneo MD, Ma X, Masatlioglu Y, Suleymanov E (2020) A random attention model. *Journal of Political Economy* 128(7):2796–2836.

de Bekker-Grob EW, Ryan M, Gerard K (2012) Discrete choice experiments in health economics: a review of the literature. *Health Economics* 21(2):145–172.

DellaVigna S (2009) Psychology and economics: Evidence from the field. *Journal of Economic literature* 47(2):315–372.

Dietrich F, List C (2016) Reason-based choice and context-dependence: An explanatory framework. *Economics & Philosophy* 32(2):175–229.

Falmagne JC (1978) A representation theorem for finite random scale systems. *Journal of Mathematical Psychology* 18(1):52–72.

Farias VF, Jagabathula S, Shah D (2013) A nonparametric approach to modeling choice with limited data. *Management Science* 59(2):305–322.

Fishburn PC (1970) *Utility Theory for Decision Making*. Publications in Operations Research, Vol. 18 (New York: John Wiley & Sons).

Frederick S, Lee L, Baskin E (2014) The limits of attraction. *Journal of Marketing Research* 51(4):487–507.

Fudenberg D, Iijima R, Strzalecki T (2015) Stochastic choice and revealed perturbed utility. *Econometrica* 83(6):2371–2409.

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

Gabaix X (2019) Behavioral inattention. *Handbook of behavioral economics: Applications and foundations 1*, volume 2, 261–343 (Elsevier).

Galichon A (2022) On the representation of the nested logit model. *Econometric Theory* 38(2):370–380.

Gallego G, Li A (2024) A random consideration set model for demand estimation, assortment optimization, and pricing. *Operations Research* 72(6):2358–2374.

Ghaderi M, Jedidi K, Kadziński M, Donkers B (2025) Random preference model. *BSE Working Paper* .

Gul F, Natenzon P, Pesendorfer W (2014) Random choice as behavioral optimization. *Econometrica* 82(5):1873–1912.

Hensher DA (1994) Stated preference analysis of travel choices: the state of practice. *Transportation* 21:107–133.

Huber J, Payne JW, Puto C (1982) Adding asymmetrically dominated alternatives: Violations of regularity and the similarity hypothesis. *Journal of Consumer Research* 9(1):90–98.

Huber J, Payne JW, Puto CP (2014) Let's be honest about the attraction effect. *Journal of Marketing Research* 51(4):520–525.

Huettner F, Boyacı T, Akçay Y (2019) Consumer choice under limited attention when alternatives have different information costs. *Operations Research* 67(3):671–699.

Joo J (2023) Rational inattention as an empirical framework for discrete choice and consumer-welfare evaluation. *Journal of Marketing Research* 60(2):278–298.

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

Kivetz R, Netzer O, Srinivasan V (2004) Alternative models for capturing the compromise effect. *Journal of Marketing Research* 41(3):237–257.

Kohli R, Jedidi K (2007) Representation and inference of lexicographic preference models and their variants. *Marketing Science* 26(3):380–399.

Kosilova N, Alptekinoğlu A (2025) Discrete choice via sequential search. *Management Science* .

Krajbich I (2019) Accounting for attention in sequential sampling models of decision making. *Current opinion in psychology* 29:6–11.

Lichtenstein S, Slovic P, eds. (2006) *The Construction of Preference* (New York, NY: Cambridge University Press).

Loewenstein G, Wojtowicz Z (2025) The economics of attention. *Journal of Economic Literature* 63(3):1038–1089.

Louviere JJ, Hensher DA, Swait JD (2000) *Stated Choice Methods: Analysis and Applications* (Cambridge, UK: Cambridge University Press).

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

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

Marschak J (1959) Binary choice constraints on random utility indicators .

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

McFadden D (1973) *Conditional logit analysis of qualitative choice behavior*, 105–142 (Academic Press).

McFadden D (2001) Economic choices. *American Economic Review* 91(3):351–378.

Natan OR (2025) Choice frictions in large assortments. *Marketing Science* 44(3):593–625.

Ning ZE, Villas-Boas JM, Yao Y (2025) Search fatigue, choice deferral, and closure. *Marketing Science* .

Noguchi T, Stewart N (2014) In the attraction, compromise, and similarity effects, alternatives are repeatedly compared in pairs on single dimensions. *Cognition* 132(1):44–56.

Noguchi T, Stewart N (2018) Multialternative decision by sampling: A model of decision making constrained by process data. *Psychological review* 125(4):512.

Onzo K, Ansari A (2025) Bayesian nonparametric sequential search. *Journal of Marketing Research* 62(2):362–385.

Orquin JL, Loose SM (2013) Attention and choice: A review on eye movements in decision making. *Acta psychologica* 144(1):190–206.

Richter MK (1966) Revealed preference theory. *Econometrica: Journal of the Econometric Society* 635–645.

Roe RM, Busemeyer JR, Townsend JT (2001) Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological review* 108(2):370.

Rooderkerk RP, Van Heerde HJ, Bijmolt TH (2011) Incorporating context effects into a choice model. *Journal of Marketing Research* 48(4):767–780.

Samuelson PA (1938) A note on the pure theory of consumer's behaviour. *Economica* 5(17):61–71.

Shafir E, Simonson I, Tversky A (1993) Reason-based choice. *Cognition* 49(1-2):11–36.

Simonson I (1989) Choice based on reasons: The case of attraction and compromise effects. *Journal of consumer research* 16(2):158–174.

Simonson I (2014) Vices and virtues of misguided replications: The case of asymmetric dominance. *Journal of Marketing Research* 51(4):514–519.

Simonson I, Tversky A (1992) Choice in context: Tradeoff contrast and extremeness aversion. *Journal of Marketing Research* 29(3):281–295.

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

Slovic P (1975) Choice between equally valued alternatives. *Journal of Experimental Psychology: Human perception and performance* 1(3):280.

Susan F, Golrezaei N, Emamjomeh-Zadeh E, Kempe D (2025) Active learning for nonparametric choice models. *Operations Research* .

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

Toubia O, Simester DI, Hauser JR, Dahan E (2003) Fast polyhedral adaptive conjoint estimation. *Marketing Science* 22(3):273–303.

Train KE (2009) *Discrete Choice Methods with Simulation* (New York, NY: Cambridge University Press).

Trueblood JS, Brown SD, Heathcote A, Busemeyer JR (2013) Not just for consumers: Context effects are fundamental to decision making. *Psychological science* 24(6):901–908.

Tversky A (1972) Elimination by aspects: A theory of choice. *Psychological Review* 79(4):281–299.

Tversky A, Shafir E (1992) Choice under conflict: The dynamics of deferred decision. *Psychological science* 3(6):358–361.

Tversky A, Simonson I (1993) Context-dependent preferences. *Management Science* 39:1179–1189.

Webb R, Glimcher PW, Louie K (2021) The normalization of consumer valuations: Context-dependent preferences from neurobiological constraints. *Management Science* 67(1):93–125.

Wright SJ (2015) Coordinate descent algorithms. *Mathematical programming* 151(1):3–34.

Wu C, Cosguner K (2020) Profiting from the decoy effect: A case study of an online diamond retailer. *Marketing Science* 39(5):974–995.

Zwick R, Rapoport A, Lo AKC, Muthukrishnan A (2003) Consumer sequential search: Not enough or too much? *Marketing Science* 22(4):503–519.

Appendix

10.1. Proof of Theorem 1

The $KL(.||\boldsymbol{u})$ is convex and $\mathbb{E}_{\boldsymbol{w}}\Big[H(J,S)\Big]$ is linear in $\boldsymbol{w}$. Therefore $\mathcal{F}_{\{\alpha,\boldsymbol{u}\}}(\boldsymbol{w})$ is strictly convex. Moreover, the simplex $\Delta(\mathcal{J})$ is closed and compact, so a minimizer exists and it is unique. Writing the Lagrangian with multiplier λ,

$$L(\boldsymbol{w},\lambda)=\sum_J w_J\Big(\alpha H(J,S)+\log(\frac{w_J}{u_J})\Big)-\lambda(\sum_J w_J-1)$$

and taking derivative

$$\partial L/\partial w_J=\alpha H(J,S)+\log(\frac{w_J}{u_J})+1+\lambda=0$$

yield $w_J=\log u_J-\alpha H(J,S)-(1+\lambda)$, that is, $w_J=u_J e^{-(1+\lambda)}e^{-\alpha H(J,S)}$. The $\partial L/\partial\lambda=\sum_J w_J-1=0$ gives $e^{-(1+\lambda)}=1/\sum_J u_J e^{-\alpha H(J,S)}$, and therefore

$$w_J^*=\frac{u_J e^{-\alpha H(J,S)}}{\sum_K u_K e^{-\alpha H(K,S)}}.$$

Plugging w_J^* into $\mathcal{F}_{\{\alpha,\boldsymbol{u}\}}(\boldsymbol{w})$ and using the optimal form $\log(w_J^*/u_J)=-\alpha H(J,S)-\log\sum_J u_J e^{-\alpha H(J,S)}$ gives $\mathcal{F}_{\{\alpha,\boldsymbol{u}\}}(w_J^*)=-\log\sum_J u_J e^{-\alpha H(J,S)}$. □

10.2. Proof of Lemma 1

The third line of Lemma follows directly from the construction of the within-subset MNL choice probabilities. The second line follows from the definition of Shannon entropy and from replacing the $P_J^{S'}$ terms in the third line. The first line follows from $w_J(S')\propto e^{-\alpha H(J,S')}$ and replacing $H(J,S')$ with $H(J,S)+\Delta_J$. □

10.3. Proof of Lemma 2

By plugging in the $w_J(S')$ and $P_J^{S'}(n)$ from Lemma 1,

$$\rho_n(S')=\sum_{J\in\mathcal{J}}w_J(S')P_J^{S'}(n)=\sum_{J\in\mathcal{J}}\frac{w_J(s)(1-q_J)P_J^S(n)e^{-\alpha\Delta_J}}{\mathbb{E}_{w_J(S)}\Big[e^{-\alpha\Delta_J}\Big]}.$$

The proof follows from the fact that the sum over the attribute subsets of the numerator is in fact $\mathbb{E}_{w_J(S)}\Big[(1-q_J)P_J^S(n)e^{-\alpha\Delta_J}\Big]$. □

10.4. Proof of Theorem 3

Firs, note that, since $K_J = \frac{e^{-\alpha\Delta_J}}{\mathbb{E}_{w_J(S)}\left[e^{-\alpha\Delta_J}\right]}$, if follows that $\mathbb{E}_{w_J(S)}\big[K_J\big] = 1$. Moreover, $\mathbb{E}_{w_J(S)}\big[P_J^S(n)\big] = \sum_{J\in\mathcal{J}} w_J(S)P_J^S(n) = \rho_n(S)$. Therefore,

$$\begin{aligned}
\mathrm{Cov}_{w_J(S)}\Big(P_J^S(n), K_J\Big) &= \sum_{J\in\mathcal{J}} w_J(S)\Big(P_J^S(n)-\rho_n(S)\Big)\Big(K_J-1\Big)\\
&= \sum_{J\in\mathcal{J}} w_J(S)P_J^S(n)K_J - \sum_{J\in\mathcal{J}} w_J(S)P_J^S(n) - \rho_n(S)\sum_{J\in\mathcal{J}} w_J(S)K_J - \rho_n(S)\sum_{J\in\mathcal{J}} w_J(S)\\
&= \sum_{J\in\mathcal{J}} w_J(S)P_J^S(n)K_J - \rho_n(S)\cdot
\end{aligned}$$

The last equality follows from $\sum_{J\in\mathcal{J}} w_J(S)P_J^S(n) = \rho_n(S)$, $\sum_{J\in\mathcal{J}} w_J(S)K_J = \mathbb{E}_{w_J(S)}\big[K_J\big] = 1$, and $\sum_{J\in\mathcal{J}} w_J(S) = 1$. Therefore,

$$\begin{aligned}
\mathrm{Cov}_{w_J(S)}\Big(P_J^S(n), K_J\Big) - \mathbb{E}_{w_J(S)}\Big[P_J^S(n)q_JK_J\Big] &= \sum_{J\in\mathcal{J}} w_J(S)P_J^S(n)K_J - \rho_n(S) - \sum_{J\in\mathcal{J}} w_J(S)P_J^S(n)q_JK_J\\
&= \underbrace{\sum_{J\in\mathcal{J}} w_J(S)(1-q_J)P_J^S(n)K_J}_{=\rho_n(S\cup\{d\})\text{ (by inserting } K_J \text{ and using Lemma 2)}} - \rho_n(S) = \Delta\rho_n\cdot
\end{aligned}$$

□

10.5. Proof of Proposition 2

Under IIA, $\rho(a_n, S\cup\{d\}) = \frac{e^{v_n}}{\sum_{s\in S} e^{v_s}+e^{v_d}} = (1-\frac{e^{v_d}}{\sum_{s\in S} e^{v_s}+e^{v_d}})\frac{e^{v_n}}{\sum_{s\in S} e^{v_s}} = (1-Q)\rho(a_n,S)$ and therefore $\Delta\rho_n^{IIA} = (1-Q)\rho(a_n,S) - \rho(a_n,S) = -Q\rho(a_n,S)$. □

10.6. Proof of Lemma 3

$$\begin{aligned}
&D_n = \Delta\rho_n - \Delta\rho_n^{IIA} && \text{(by definition)}\\
&= \Delta\rho_n + Q\rho_n(S) && \text{(by Proposition 2)}\\
&= \mathrm{Cov}_{w_J(S)}\Big(P_J^S(n), K_J\Big) - \mathbb{E}_{w_J(S)}\Big[P_J^S(n)q_JK_J\Big] - Q\rho_n(S) && \text{(by Theorem 3)}\\
&= \mathrm{Cov}_{w_J(S)}\Big(P_J^S(n), K_J\Big) - \mathbb{E}_{w_J(S)}\Big[P_J^S(n)q_JK_J\Big] + \mathbb{E}_{w_J(S')}[q_J]\mathbb{E}_{w_J(S)}[P_J(n)] && \text{(by } Q = \mathbb{E}_{w_J(S')}[q_J] \text{ and } \rho_n(S) = \mathbb{E}_{w_J(S)}[P_J(n)]\text{)}
\end{aligned}$$

Finally, by Lemma 1, $\mathbb{E}_{w_J(S')}[q_J] = \mathbb{E}_{w_J(S)}[q_JK_J]$. Using the definition $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ where $X = P_J^S(n)$ and $Y = q_JK_J$, $D_n = \mathrm{Cov}_{w_J(S)}\Big(P_J^S(n), K_J\Big) - \mathrm{Cov}_{w_J(S)}\Big(P_J^S(n), q_JK_J\Big)$. □

10.7. Proof of Proposition 3

By Lemma 1, $w_J(S_{base} \cup \{A_d\}) \propto w_J(S_{base})\, e^{-\alpha \Delta_J}$. By part (ii) of the proposition, Δ_J is increasing in q_J. Hence, by (iii)–(iv),

$$\min_{J \notin \bar{\mathcal{J}}} \Delta_J \; > \; \max_{J \in \bar{\mathcal{J}}} \Delta_J .$$

Therefore, as α grows, the reweighting $K_J \propto e^{-\alpha \Delta_J}$ concentrates on $\bar{\mathcal{J}}$, thus $w_J(S_{base} \cup \{A_d\}) \geq w_J(S_{base})$ for $J \in \bar{\mathcal{J}}$ and $w_J(S_{base} \cup \{A_d\}) \leq w_J(S_{base})$ for $J \notin \bar{\mathcal{J}}$. Using $P_J^{S_{base} \cup \{A_d\}}(A) = (1-q_J)P_J^{S_{base}}(A)$ and $P_J^{S_{base}}(A) > P_J^{S_{base}}(B)$ on $\bar{\mathcal{J}}$, and noting that q_J is uniformly small there, the attention shift toward $\bar{\mathcal{J}}$ will dominate the within-attribute IIA loss as α becomes large enough and therefore $\Delta \rho_A \geq 0$. □

10.8. Proof of Proposition 4

$\rho_n(\alpha) = \sum_{J \in \mathcal{J}} w_J(\alpha) P_J(n)$ where $w_J(\alpha) = \frac{e^{-\alpha H_J}}{\sum_{k \in \mathcal{J}} e^{-\alpha H_K}}$. Thus, for $\alpha = 0$, it immediately follows that $w_J(0) = 1/|\mathcal{J}|$. On the other hand, when $\alpha \to \infty$, let $H_{min} = \min_K H_K$ and $\mathcal{J}^* = \{J : H_J = \min_K H_K\}$. Multiplying and diving $w_J(\alpha)$ by $e^{\alpha H_{min}} > 0$ and noting that $H_J - H_{min} = 0$ for all $J \in \mathcal{J}^*$,

$$\begin{aligned}\rho_n(\alpha) &= \textstyle\sum_{J \in \mathcal{J}} \frac{e^{-\alpha(H_J - H_{min})}}{\sum_{k \in \mathcal{J}} e^{-\alpha(H_K - H_{min})}} P_J(n) \\ &= \textstyle\sum_{J \in \mathcal{J}^*} \frac{1}{|\mathcal{J}^*| + \sum_{k \notin \mathcal{J}} e^{-\alpha(H_K - H_{min})}} P_J(n) + \sum_{J \notin \mathcal{J}^*} \frac{e^{-\alpha(H_J - H_{min})}}{|\mathcal{J}^*| + \sum_{k \notin \mathcal{J}} e^{-\alpha(H_K - H_{min})}} P_J(n) \cdot \end{aligned}$$

When $\alpha \to \infty$, all the $e^{-\alpha(H_J - H_{min})}$ terms vanish if $J \notin \mathcal{J}^*$ since $H_J - H_{min} > 0$. Therefore, it follows that $\lim_{\alpha \to \infty} \rho_n(\alpha) = \sum_{J \in \mathcal{J}^*} \frac{1}{|\mathcal{J}^*|} P_J(n) = \frac{1}{|\mathcal{J}^*|} \sum_{J \in \mathcal{J}^*} P_J(n)$. □

10.9. Proof of Lemma 4

$\frac{d\rho_n}{d\alpha} = \sum_{J \in \mathcal{J}} P_J(n) \frac{d}{d\alpha} w_J(\alpha)$ since $P_J(n)$ is independent from α. Therefore,

$$\begin{aligned}\frac{d\rho_n}{d\alpha} &= \sum_{J \in \mathcal{J}} P_J(n) \frac{d}{d\alpha} \frac{e^{-\alpha H_J}}{\sum_{K \in \mathcal{J}} e^{-\alpha H_K}} \\ &= \sum_{J \in \mathcal{J}} P_J(n) \big(\frac{-H_J e^{-\alpha H_J}}{\sum_{K \in \mathcal{J}} e^{-\alpha H_K}} + \frac{e^{-\alpha H_J} \sum_{K \in \mathcal{J}} H_K e^{-\alpha H_K}}{(\sum_{K \in \mathcal{J}} e^{-\alpha H_K})^2} \big) \\ &= \sum_{J \in \mathcal{J}} -P_J(n) H_J \frac{e^{-\alpha H_J}}{\sum_{k \in \mathcal{J}} e^{-\alpha H_K}} + \sum_{J \in \mathcal{J}} P_J(n) \frac{e^{-\alpha H_J}}{\sum_{k \in \mathcal{J}} e^{-\alpha H_K}} \frac{\sum_{K \in \mathcal{J}} H_K e^{-\alpha H_K}}{\sum_{K \in \mathcal{J}} e^{-\alpha H_K}} \\ &= \sum_{J \in \mathcal{J}} -P_J(n) H_J w_J(\alpha) + \big(\sum_{J \in \mathcal{J}} P_J(n) w_J(\alpha) \big) \big(\sum_{K \in \mathcal{J}} H_K w_K(\alpha) \big) \\ &= -\mathbb{E}_{w_J(\alpha)} \Big[P_J(n) H_J \Big] + \mathbb{E}_{w_J(\alpha)} \Big[P_J(n) \Big] \mathbb{E}_{w_J(\alpha)} \Big[H_J \Big] = -\mathrm{Cov}_{w_J(\alpha)} \Big(P_J(n), H_J \Big). \end{aligned}$$

□

10.10. Proof of Lemma 5

By the AERU identity covariance,

$$\rho_n(\alpha) = \rho_n(0) - \int_0^\alpha \mathrm{Cov}_{w_J(\tau)}\Big(P_J(n), H_J\Big)d\tau$$

Therefore

$$\begin{aligned}
&|\rho_n(\alpha)-\rho_n(0)| = \left|\textstyle\int_0^\alpha \mathrm{Cov}_{w_J(\tau)}\Big(P_J(n),H_J\Big)d\tau\right| \le \textstyle\int_0^\alpha \left|\mathrm{Cov}_{w_J(\tau)}\Big(P_J(n),H_J\Big)\right| d\tau \\
&\textstyle\int_0^\alpha \sqrt{\mathrm{Var}_{w_J(\tau)}\Big[P_J(n)\Big]}\sqrt{\mathrm{Var}_{w_J(\tau)}\Big[H_J\Big]}d\tau \qquad \text{(by Cauchy-Schwarz inequality)} \\
&\le \textstyle\int_0^\alpha \frac{1}{2}\cdot\frac{\log|S|}{2}d\tau \qquad \text{(since } 0 \le P_J(n) \le 1\text{, and } 0 \le H_J \le \log|S|\text{)} \\
&= \tfrac{\log|S|}{4}\alpha. \qquad \square
\end{aligned}$$

10.11. Proof of Lemma 6

Given that $0 \le P_J(n) \le 1$ and $0 \le H_J \le \log|S|$,

$$\begin{aligned}
&\tfrac{d\rho_n}{d\alpha} \le |\tfrac{d\rho_n}{d\alpha}| = \left|\mathrm{Cov}_{w_J(\alpha)}\Big(P_J(n),H_J\Big)\right| && \text{(by Proposition 4)} \\
&\le \sqrt{\mathrm{Var}_{w_J(\tau)}\Big[P_J(n)\Big]}\sqrt{\mathrm{Var}_{w_J(\tau)}\Big[H_J\Big]}d\tau && \text{(by Cauchy-Schwarz inequality)} \\
&\le \sqrt{\Big(1-\mathbb{E}_{w_J(\alpha)}\Big[P_J(n)\Big]\Big)\mathbb{E}_{w_J(\alpha)}\Big[P_J(n)\Big]}\sqrt{\Big(\log|S|-\mathbb{E}_{w_J(\alpha)}\Big[H_J\Big]\Big)\mathbb{E}_{w_J(\alpha)}\Big[H_J\Big]} && \text{(by Bhati-Davis inequality)} \\
&= \sqrt{\rho_n(1-\rho_n)}\sqrt{\mu(\log|S|-\mu)} && \text{(by Proposition 1 and } \mathbb{E}_{w_J(\alpha)}\Big[H_J\Big]=\mu\text{)} \\
&\le \sqrt{\rho_n(1-\rho_n)}\tfrac{\log|S|}{2} && \text{(by } \mu(\log|S|-\mu) \le (\tfrac{\log|S|}{2})^2\text{)}
\end{aligned}$$

Therefore $\frac{d\rho_n}{d\alpha} \le \sqrt{\rho_n(1-\rho_n)}\frac{\log|S|}{2}$, which gives rise to the following ordinary differential equation:

$$\frac{d\rho_n}{\sqrt{\rho_n(1-\rho_n)}} \le \frac{\log|S|}{2}d\alpha\cdot$$

We solve this differential equation by taking $\rho_n \equiv \sin^2\theta$, consequently $d\rho = 2\sin\theta\cos\theta d\theta$. Since $\rho \in [0,1]$, it follows that $\theta \in [0,\pi/2]$, and therefore $\sin\theta \ge 0, \cos\theta \ge 0$. Thus,

$$\int \frac{d\rho_n}{\sqrt{\rho_n(1-\rho_n)}} = \int \frac{2\sin\theta\cos\theta d\theta}{\sqrt{\sin^2\theta(1-\cos^2\theta)}} = \int \frac{2\sin\theta\cos\theta d\theta}{\sin\theta\cos\theta} = 2\theta \le \frac{\log|S|}{2}\alpha + \text{constant}.$$

But $\theta = \arcsin\sqrt{\rho}$. Thus, without loss of generality, for any $\alpha_2 \ge \alpha_1$

$$\arcsin\sqrt{\rho_n(\alpha_2)} - \arcsin\sqrt{\rho_n(\alpha_1)} \le \frac{\log|S|}{4}(\alpha_2-\alpha_1)\cdot \qquad \square$$

10.12. Proof of Lemma 7

Define $\delta_J = H_J - \min_K H_K$. The proof follows from decomposing $w_J(\alpha)$ on $\mathcal{J}/\mathcal{J}^*$ and $\mathcal{J}$ and recognizing two facts: i) $\delta_J = 0$ for $J \in \mathcal{J}^*$ and $\delta_J \geq \Delta$ where Δ is the entropy gap defined in Lemma (7), and ii) $w_J(\alpha) = \frac{e^{-\alpha H_J}}{\sum_{K\in\mathcal{J}} e^{-\alpha H_K}} \overset{\text{divide by } \min_K H_K}{=} \frac{e^{-\alpha\delta_J}}{|\mathcal{J}^*|+\sum_{K\notin\mathcal{J}^*} e^{-\alpha\delta_K}}$.

$$\begin{aligned}
&\rho_n(\alpha) = \sum_J w_J(\alpha)P_J(n) = \sum_{J\notin\mathcal{J}^*} w_J(\alpha)P_J(n) + \sum_{J\in\mathcal{J}^*} w_J(\alpha)P_J(n)\\
&= \sum_{J\notin\mathcal{J}^*} w_J(\alpha)\Big(\frac{\sum_{J\notin\mathcal{J}^*} w_J(\alpha)P_J(n)}{\sum_{J\notin\mathcal{J}^*} w_J(\alpha)}\Big) + \sum_{J\in\mathcal{J}^*} \frac{1-\sum_{J\notin\mathcal{J}^*} w_J(\alpha)}{|\mathcal{J}^*|}P_J(n)\\
&= \nu(\alpha)\tilde{\rho}_n(\alpha) + \big(1-\nu(\alpha)\big)\rho^*(\alpha)
\end{aligned}$$

where $\nu(\alpha) = \sum_{J\notin\mathcal{J}^*} w_J(\alpha)$, $\tilde{\rho}_n(\alpha) = \frac{\sum_{J\notin\mathcal{J}^*} w_J(\alpha)P_J(n)}{\sum_{J\notin\mathcal{J}^*} w_J(\alpha)}$, and $\rho^* = \frac{1}{|\mathcal{J}^*|}\sum_{J\in\mathcal{J}^*} P_J(n)$. Notice that for a $J \in \mathcal{J}^*$, $w_J(\alpha) = \frac{1}{|\mathcal{J}^*|+\sum_{K\notin\mathcal{J}^*} e^{-\alpha\delta_K}} = k$ is a constant and therefore $|\mathcal{J}^*|\,k + \sum_{J\notin\mathcal{J}^*} w_J(\alpha) = 1$, hence $w_J(\alpha) = \frac{1-\sum_{J\notin\mathcal{J}^*} w_J(\alpha)}{|\mathcal{J}^*|}$ for all $J \in \mathcal{J}^*$. Subsequently,

$$||\rho_n(\alpha) - \rho^*(\alpha)||_p = \nu(\alpha)\,||\tilde{\rho}_n(\alpha) - \rho^*(\alpha)||_p \leq 2^{1/p}\nu(\alpha)$$

The last inequality follows from the fact that $\tilde{\rho} and \rho^*$ are probability vectors and $\sup_{u,v} ||u-v||_p = 2^{1/p}$ for any two probability vectors p, q. Finally,

$$\nu(\alpha) = \sum_{J\notin\mathcal{J}^*} w_J(\alpha) = \frac{\sum_{J\notin\mathcal{J}^*} e^{-\alpha\delta_J}}{|\mathcal{J}^*| + \sum_{K\notin\mathcal{J}^*} e^{-\alpha\delta_K}} \leq \frac{\sum_{J\notin\mathcal{J}^*} e^{-\alpha\Delta}}{|\mathcal{J}^*| + \sum_{K\notin\mathcal{J}^*} e^{-\alpha\delta_K}} \leq \frac{\sum_{J\notin\mathcal{J}^*} e^{-\alpha\Delta}}{|\mathcal{J}^*|} \leq \frac{|\mathcal{J}/\mathcal{J}^*|e^{-\alpha\Delta}}{|\mathcal{J}^*|}$$

Therefore, $||\rho_n(\alpha) - \rho^*(\alpha)||_p \leq 2^{1/p}\frac{|\mathcal{J}/\mathcal{J}^*|}{|\mathcal{J}^*|}e^{-\alpha\Delta}$. □