Random Utility with Aggregation

Yuexin Liao, Kota Saito, Alec Sandroni*
Caltech

September 28, 2025

Abstract

We characterize when discrete-choice datasets that involve aggregation, such as category-level items or an outside option, are consistent with a random utility model (RUM). The underlying alternatives that an aggregated category represents may differ across individuals and remain unobserved by the analyst. We characterize the observable implications of RUMs with unknown composition of aggregated categories and show that they are surprisingly weak, implying only limited monotonicity of choice frequencies and standard RUM consistency on unaggregated menus. These restrictions are insufficient to justify the aggregated random utility model (ARUM) commonly assumed in empirical work. We identify two sufficient conditions that restore the implication of ARUM: non-overlapping preferences and menu-independent aggregation. Simulations show that violations of these conditions generate estimation bias, highlighting the practical importance of how aggregated alternatives are defined.

1 Introduction

In empirical studies of discrete choice, it is common practice to aggregate alternatives. For instance, in car sales data, all variants of the Toyota RAV4 despite differing in features and prices, are often grouped under a single category, "RAV4." A particularly important form of aggregation involves *the outside option*. In many empirical IO studies (e.g., Nevo (2001); Shum (2016)), the outside option is modeled as a single alternative that aggregates all unlisted choices. For example, in Nevo (2001), which studies consumer demand for 25

*Liao: yliao@caltech.edu. Saito: saito@caltech.edu, Sandroni: asandron@caltech.edu. We appreciate participants at a presentation at Waseda University and Caltech Lunch Seminar. Alec Sandroni acknowledges financial support from the Summer Undergraduate Research Fellowships (SURF) program in 2022, 2023, and 2024 at Caltech. Yuexin Liao also acknowledges financial support from SURF program in 2025 at Caltech.

cereal brands, the market share of the outside good is defined as the residual: the difference between one and the total market share of the inside goods, implicitly capturing all other breakfast options, such as pancakes or omelettes.[1]

The standard modeling approach assumes a random utility model (RUM) defined directly over the aggregated categories—what we call the *aggregated random utility model* (ARUM). That is, researchers assume that each category corresponds to a single alternative, and that preferences are defined over these categories as if they were atomic options.

However, this modeling simplification obscures an important complication: while the analyst only observes aggregate frequencies, the consumer may evaluate the underlying alternatives composing an aggregated category very differently. In the car example, different versions of the RAV4 such as the Gasoline LE and the Plug-in Hybrid XSE may vary significantly in both desirability and prices. Likewise, the outside option may encompass alternatives that differ greatly in price and quality. Furthermore, the exact composition of an aggregated category may be *unobservable* to the analyst. As a result, observed choice frequencies over aggregated categories may not represent consumers' underlying preferences.

Our paper is the first to formally analyze the implications of unknown composition in aggregated categories. We characterize the observable restrictions of RUM in this setting and show precisely how they are weaker than those implied by ARUM. Through a simulation exercise, we further demonstrate that this weakness can lead to significant estimation biases. Building on these findings, we provide theoretical results that offer guidance on how to define aggregated categories in ways that mitigate such biases.

1.1 Preview of theoretical results

To illustrate our contribution, consider a similar setup that was studied in Nevo (2001), which involves 25 cereal brands along with an outside good that contains all other breakfast options. Importantly, the composition of the outside good may differ across individuals in the population. For instance, some consumers may not consider omelette a viable option due to egg shortages resulting from bird flu, while others may rule out pancakes because essential ingredients such as syrup are unavailable in stores.

We formalize this using two key concepts. An *aggregation function* X specifies the possible underlying goods that an aggregated category may represent—for example, the outside option may include omelettes, pancakes, or other breakfast items. A *composition distribution* λ then assigns, for each menu, probabilities to these possible realizations of the outside option. Thus, when the outside option is present, λ determines whether a consumer

[1] See Appendix A: Data, page 337.

faces only omelettes, only pancakes, both, or some other combination of breakfast goods. Importantly, analysts observe only the reduced choice frequencies over categories, not the aggregation function or the composition distribution.

We say that the observed dataset ρ is consistent with a RUM if there exists (i) composition distribution λ with aggregation function X and (ii) a probability distribution $\mu_{\mathscr{X}}$ over rankings (i.e., linear orders) $\succ$ on the set $\mathscr{X}$ of the underlying alternatives (in this case, the 25 cereal brands along with other breakfast options)[2], such that for all subsets D of cereals and $x \in D$, the market share $\rho(D \cup \text{out}, x)$ of cereal x in a market where the set D of cereals is available equals

$$\sum_{S \subseteq X(out)} \lambda_{D\cup\text{out}}(S)\, \mu_{\mathscr{X}}(\succ \mid x \succ y \text{ for all } y \in (D \setminus \{x\}) \cup S), \tag{1}$$

where $\lambda_{D\cup\text{out}}(S)$ is the probability that the outside option "out" is composed of the set $S \subseteq X(out)$ in the context of choice set $D \cup \text{out}$. This formulation is a natural extension of the standard RUM: Equation (1) shows that the observed choice probability is a weighted average of standard RUM probabilities, with weights λ reflecting the distribution of outside-option compositions.

Our first question is: *What are the observable implications for ρ that are consistent with some pair $(\mu_{\mathscr{X}}, \lambda)$ as described above?* We show that these implications are characterized exactly by two conditions. The first is *Limited Monotonicity*—that is, the choice frequencies on menus containing only non-aggregated categories decrease when aggregated categories are added to the choice set. The second is a condition that we call *Partial RU-rationality*, that is on the subset of choice sets that do not contain any aggregated categories, choices must be rationalizable by a standard RUM over categories. See Theorems 3.1 and 3.12 in Section 3 for the formal statements.

Crucially, these two conditions are *insufficient* to guarantee the existence of an *aggregated random utility model* (ARUM), that is, a distribution over preference rankings on *aggregated* categories. This is because ARUMs require much more structure. The simplest implication of ARUM is full monotonicity — that is, choice frequencies decrease when alternatives are added to any choice set. The RU-rational choice functions may violate full monotonicity when the choice set contains aggregated categories. This gap is especially concerning given how commonly ARUMs are used in empirical work.

In addition to the characterization in Theorem 3.1, Theorem 3.6 provides a complementary characterization of RU-rationality focusing on the behavior of each consumer in

[2]We write $\mu_{\mathscr{X}}$ to clarify that the measure is defined on rankings over the underline alternatives $\mathscr{X}$, not the aggregated categories.

the population. It shows that RU-rationality allows behavior ruled out by ARUM: an agent may follow a well-defined ranking in some cases but default to the outside option in others. This can be interpreted as menu effects, where complex or unfamiliar menus lead the agent to choose the outside option. The result underscores that, without further restrictions, RU-rationality does not imply ARUM.

Given these results, we next ask: *under what additional conditions can RU-rational stochastic choice functions also be rationalized by an ARUM?* This question is practically important, as it guides how aggregated categories should be defined in empirical work.

We identify two sufficient conditions. The first is the *non-overlapping condition* on $\mu_{\mathscr{X}}$: all of the alternatives that underlie an aggregated category must occupy similar positions in consumers' rankings. In the cereal example, this means no consumer places some cereals between other breakfast options such as omelettes and pancakes. When overlap occurs—e.g., some consumers prefer omelettes to certain cereals but not others—the condition fails (Proposition 4.2). The second is menu independence of the composition distribution λ: the distribution of underlying components must not vary with the choice set. In empirical IO, where each menu corresponds to a market, this requires that λ remain constant across markets. For instance, if eggs are scarce in one market (making omelettes unavailable) but plentiful in another (so omelettes remain available), the condition is violated (Proposition 4.6).

1.2 Simulation results

Our results imply that when either condition—non-overlapping preferences or menu independence—fails, the observed choice frequencies diverge from those predicted by an ARUM. Estimating an ARUM (such as logit model over aggregated categories) can therefore generate substantial bias. Assessing this bias empirically is crucial for understanding the practical relevance of our theory.

We examine this prediction through simulations. Starting from a logit model defined on the underlying alternatives $\mathscr{X}$, we generate the true dataset $\rho_{\mathscr{X}}$ assuming true utility values. We then fix values of the composition distribution λ and an aggregation function X and construct the reduced dataset ρ over aggregated categories. Finally, we re-estimate utilities from ρ as if it followed a logit model over aggregated categories. When λ is menu-independent or preferences are non-overlapping, ρ is consistent with an ARUM and bias must be low, otherwise the bias should become large as ρ cannot be represented by an ARUM.

As our measure of bias, we examined the difference in estimated utilities between two non-aggregated alternatives, and we also computed the geometric distance from ρ to the set of ARUMs. Both measures increase as λ becomes more menu-dependent or preferences more overlapping. In some cases, the bias was substantial enough to reverse the ranking of alternatives—for example, even when $u(x) > u(y)$ in the true model, the estimates produced $\hat{u}(y) > \hat{u}(x)$. These findings align with our theoretical predictions and underscore the empirical importance of potential biases. Further details are provided in Section 5.

1.3 Related Literature

There is no paper studying stochastic choice focusing on the unknown composition of aggregated categories. There are, however, several papers that are related in terms of mathematical setup or motivation.

The first line of research related to our paper is the recent work on random attention models, such as Cattaneo, Ma, Masatlioglu, and Suleymanov (2020), Aguiar, Boccardi, Kashaev, and Kim (2023), Brady and Rehbeck (2016), Manzini and Mariotti (2014),and Horan (2019). See Strzalecki (2022) for a survey of the literature. These papers study models in which agents do not consider the full available choice set but instead randomly consider a subset. In particular, Horan (2019) studies the case where the choice frequency of the outside or "default" option is not observable under the random attention model.

The resemblance is only superficial. The models of random attention assume that each decision maker randomly *attends* to a subset of the alternatives. Our setting, by contrast, is driven by *aggregation-induced unobservability*: the analyst sees aggregated categories whose internal composition differ across consumers and menus. This distinction has sharp mathematical consequences. In our setup, when an aggregated category is present, *at least one* of its underlying components is available to the consumer. The models of random attention impose no such requirement. This constraint yields testable implications for RUM in our paper.

Azrieli and Rehbeck (2022) studies a mathematically related model but with a different focus. They characterize marginal stochastic choice—frequencies not conditioned on particular menus—assuming the analyst observes the distribution of menus but not within-menu choices. By contrast, in our framework the distribution over available alternatives is unobserved and captured by λ, while we observe reduced stochastic choice over aggregated categories. Although Azrieli and Rehbeck (2022) also allows for endogenous menu availability, their analysis centers on models such as Gul and Pesendorfer (2001) and Kreps (1979), and addresses rationalizability rather than estimation bias.

Apesteguia and Ballester (2016) also study aggregation of stochastic choice data, but with a different motivation. They examine whether aggregate choice behavior, obtained by combining individual choices, remains in the same class as the underlying individual behaviors. By contrast, our model aggregates alternatives and allows for aggregation across agents facing different choice problems.

Our companion paper Kono, Saito, and Sandroni (2025) also studies the RUM with an outside option, but in a different setup. The paper does not analyze the case of multiple aggregated categories. Moreover, from the perspective of the observed dataset, the key distinction between the two papers is that the present study takes the reduced dataset ρ as the observed object, whereas the companion paper treats the full dataset over the underlying alternatives (with missing information) as given. In addition, the companion paper assumes the outside option always corresponds to the full set of underlying alternatives—that is, λ is menu-independent and degenerate. Even in this idealized case, much of the implications of the RUM are lost. In the present paper, we extend that framework by allowing λ to be non-degenerate or menu-dependent, and we characterize the observable implications of the RUM under this more general specification.

A few papers in the empirical IO literature such as Brownstone and Li (2017), Stafford (2018), and Huang and Rojas (2013) have examined potential biases arising from aggregation and misspecification of the outside option. More recently, Zhang (2023) analyzed biases resulting from misspecifying market size, and hence the market share of the outside good. However, all of these papers are purely empirical in nature and do not address the issue of unknown composition of aggregated categories or outside options.

Although less directly related, classical statistical work has developed a comprehensive framework for making inferences from data that are partially observed or categorized. For example, see Heitjan (1989) for a review of grouped data; Blumenthal (1968) and Nordheim (1984) for categorized data; and Heitjan and Rubin (1990, 1991) for coarse data. In this framework, subsets of the sample space are observed rather than exact realizations.[3] A key distinction from our setting is that, in our model, uncertainty arises from the actual choice set an agent faces, rather than from the data collection or coarsening process.

2 Setup

2.1 Alternatives

We introduce two notations that we use throughout the paper: For any finite set Z, let $\mathscr{L}(Z)$ be the set of rankings (i.e., linear orders) on Z and let $\Delta(Z)$ be the set of probability

[3]For example, if the sample space is the real line (e.g., blood pressure), we may only observe a random variable taking interval values (e.g., an interval of blood pressures containing the true value).

distributions over Z.

Let $\mathscr{A} = \{a_0, ..., a_N\}$ be the (finite) set categories. Let $\mathscr{X}$ be the set of underlying alternatives. We assume that $\mathscr{X}$ is finite.[4] We consider a partition $\{\mathscr{A}_N, \mathscr{A}_A\}$ of $\mathscr{A}$.[5] We interpret $\mathscr{A}_N$ as the set of *nonaggregated* categories; $\mathscr{A}_A$ the set of *aggregated* categories.

Definition 2.1. *A function $X : \mathscr{A} \to 2^{\mathscr{X}} \setminus \{\emptyset\}$ is called an aggregation function if (i) $X(a_i) \cap X(a_j) = \emptyset$ for any distinct i, j; (ii) $X(a)$ is singleton if $a \in \mathscr{A}_N$; (iii) $X(a)$ is non-singleton if $a \in \mathscr{A}_A$.*

We call X an aggregation function because the function (or more precisely, its inverse) specifies how each underlying alternative is assigned to a category. Condition (i) requires that every underlying alternative be assigned to exactly one category. Conditions (ii) and (iii) justify the interpretation that $\mathscr{A}_N$ represents the set of non-aggregated categories, while $\mathscr{A}_A$ represents the set of aggregated categories.

In many parts of the paper, we consider the setup in which there is only one aggregated category a_0 called *outside option*. In the setup, we assume $\mathscr{A}_N = \{a_0\}$.

Example 2.1. (**Cereal Choice**) Consider a similar setup to Nevo (2001), who estimates demand over the 25 cereal brands. As mentioned in the introduction, a key feature of the analysis is the inclusion of an outside good, representing all breakfast options not explicitly listed among the 25 cereal brands in the dataset. In the setup, all available alternatives are

$$\mathscr{A} = \{\text{Kellogg Corn Flakes, Kellogg Crispix}, \cdots, \text{Quaker Life, Outside goods}\}.$$

Since the cereal brands are finely specified in $\mathscr{A}$, all cereals are non-aggregated; thus we have $\mathscr{A}_N = \{\text{Kellogg Corn Flakes}, \cdots, \text{Quaker Life}\}$; and $\mathscr{A}_A = \{\text{Outside goods}\}$. For example, $X(\text{Kellogg Corn Flakes}) = \{\text{Kellogg Corn Flakes}\}$. If the outside goods contain only two alternatives by our assumption, then $X(\text{Outside goods}) = \{\text{Omelette, Pancakes}\}$.[6]

2.2 Datasets and RU-rationalities

Let $\mathscr{D} \subseteq 2^{\mathscr{A}} \setminus \{\emptyset\}$ be the set of observable choice sets. The dataset in this paper is a stochastic choice function ρ on $\mathscr{D}$. That is, the dataset is a function $\rho : \mathscr{D} \times \mathscr{A} \to \mathbb{R}_+$ such that for

[4]To ensure the existence of all aggregation functions relevant to the RUM, we assume that $\mathscr{X}$ contains at least $|\mathscr{A}_A|^2$ distinct elements.

[5]That is, $\mathscr{A}_N \cup \mathscr{A}_A = \mathscr{A}$ and $\mathscr{A}_N \cap \mathscr{A}_A = \emptyset$.

[6]In our example, we assume omelettes are unavailable when eggs are in short supply. An alternative setup is that eggs remain available but at higher cost, so omelettes are still feasible though less attractive. We can represent this case as "Omelette with high price," reflecting the increased cost while keeping it as an available alternative.

any $A \in \mathscr{D}$, $\sum_{a \in A} \rho(A,a) = 1$, $\rho(A,a) \geq 0$ for all $a \in A$ and $\rho(A,a) = 0$ for all $a \notin A$. Note that the stochastic choice function is defined over menus of categories in $\mathscr{A}$, not over the underlying alternatives $\mathscr{X}$. Thus, we sometimes call ρ *reduced* dataset.

Remark 2.2. *For simplicity, we assume that ρ is defined on all subsets of $\mathscr{A}$. However this assumption is not necessary for our main results. After we present our theorems, we provide a remark in which we specify a weaker domain assumption. (See Remark 3.4 for Theorem 3.1, Remark 3.11 for Theorem 3.6, and Remark 3.13 for Theorem 3.12.)*

As mentioned in the introduction, in discrete choice analysis, the standard method in the presence of aggregated categories is to assume *aggregated random utility models* (ARUMs), which is a probability distribution over rankings over $\mathscr{A}$.

Definition 2.3. *A stochastic choice function ρ is aggregated random utility (ARU)-rational if there exists $\mu_{\mathscr{A}} \in \Delta(\mathscr{L}(\mathscr{A}))$ such that for any $a \in A \in \mathscr{D}$,*[7]

$$\rho(A,a) = \mu_{\mathscr{A}}(\succ \in \mathscr{L}(\mathscr{A}) \;|a \succ b \text{ for all } b \in A \setminus a). \tag{2}$$

In our framework, preferences are defined on the set $\mathscr{X}$ of underlying primitive alternatives rather than on their aggregates. The components of an aggregate category may differ substantially—for example, across versions of the RAV4 or within a heterogeneous outside option—and these differences matter to agents. Accordingly, the relevant random utility model should be formulated over $\mathscr{X}$.

To connect these two levels of analysis, we introduce a formal mapping called a *composition distribution* which links the aggregated categories to the underlying alternatives in their composition. Given an aggregation function, this mapping λ captures the distribution of available underlying alternatives faced by agents in the population, which may not be observable to the analyst.

Definition 2.4. *Given an aggregation function X, a set of functions $(\lambda_A)_{A \subseteq \mathscr{A}}$ is called a composition distribution if for any $A \in \mathscr{D}$,*

- $\lambda_A((S_a)_{a \in A}) \geq 0$ *for any* $(S_a)_{a \in A} \in \prod_{a \in A}(2^{X(a)} \setminus \{\emptyset\})$,
- $\sum_{(S_a)_{a \in A} \in \prod_{a \in A}(2^{X(a)} \setminus \{\emptyset\})} \lambda_A((S_a)_{a \in A}) = 1.$

The value of the composition distribution $\lambda_A((S_a)_{a \in A})$ represents the frequency with which each aggregated category a is composed of the subset $S_a \subseteq X(a)$ within the choice

[7] We write $\mu_{\mathscr{A}}$ to clarify that the measure is defined over rankings over $\mathscr{A}$, not on the set $\mathscr{X}$ of underlying alternatives.

set $A \subseteq \mathscr{A}$. We allow λ to depend on the choice set A, since the availability of underlying alternatives may vary across markets—for example, it may depend on which stores consumers can access. In Example 2.1, eggs may be unavailable in one market (e.g., due to a bird flu outbreak), making omelettes infeasible, while pancake ingredients remain widely available. In this case, λ_A would assign higher probability to S_{out} containing pancakes than to sets including omelettes.

We now define RU-rationality in the presence of aggregated categories whose compositions are unknown to the analyst. To motivate the definition, return to the cereal example and consider a distribution $\mu_{\mathscr{X}}$ over the underlying alternatives (the cereals and other breakfast options). Let D be the set of cereals available in a store and let $x \in D$ denote one such cereal. For a subpopulation of consumers whose outside option consists only of omelettes and pancakes, the market share of x is $\mu_{\mathscr{X}}(\succ| \, x \succ y \text{ for all } y \in D \cup \{\text{omelette}, \text{pancakes}\})$. Analogous expressions arise for other possible compositions of the outside option. Aggregating across these compositions using λ_A yields equation (1). The following extends this construction to the case with multiple aggregated categories:

Definition 2.5. *A stochastic choice function ρ is RU-rational if there exist $\mu_{\mathscr{X}} \in \Delta(\mathscr{L}(\mathscr{X}))$, an aggregation function X, and a composition distribution λ such that for any $a \in A \in \mathscr{D}$,*

$$\rho(A, a) = \sum_{(S_b)_{b \in A} \in \prod_{b \in A} (2^{X(b)} \setminus \{\emptyset\})} \lambda_A((S_b)_{b \in A}) \mu_{\mathscr{X}}(\succ \in \mathscr{L}(\mathscr{X}) | \exists x \in S_a, \ \forall y \in \bigcup_{b \in A \setminus a} S_b, \ x \succ y). \tag{3}$$

We say that the pair $(\mu_{\mathscr{X}}, \lambda)$ rationalizes *ρ.*[8]

Note that the term $\mu_{\mathscr{X}}(\cdots)$ on the right-hand side of equation (3) represents the total market share of category a from menu A within the subpopulation of consumers for whom each aggregated alternative b corresponds to the subset $S_b \subseteq X(b)$. The term $\lambda_A((S_a)_{a \in A})$ denotes the proportion of such consumers in the overall population. Thus, the right-hand side of equation (3) is the *weighted average* of these market shares, with weights given by λ.

Remark 2.6. *Definition 2.5 assumes independence between the distribution $\mu_{\mathscr{X}}$ over preference rankings and the composition distribution λ over underlying alternatives. While one could consider a more general framework allowing dependence between these components, such a model may lack testable implications, since the joint distribution could vary with the menu as λ does. In particular, if the agent's preferences may depend on composition, such*

[8]Strictly speaking, we should write $(\mu_{\mathscr{X}}, X, \lambda)$ rationalizes ρ. However, since the role of X is implicit in λ, we will simply write $(\mu_{\mathscr{X}}, \lambda)$.

dependence would allow the distribution over rankings to vary with the menu, rendering the model unfalsifiable. If, however, we only allow the composition distribution λ to depend on the preferences, the characterization remains the same. See Remark 3.3 for more details.

Remark 2.7. *ARU-rationality implies RU-rationality. To see why suppose ρ is rationalized by $\mu_{\mathscr{A}}$ over categories $\mathscr{A}$. Choose an aggregation function X so that $X(a)$ is a singleton for all $a \in \mathscr{A}$. Label the sole element of $X(a)$ by x_a. Then for each $\succ$ in the support of $\mu_{\mathscr{A}}$ define $\succ'$ by $x_a \succ' x_b$ whenever $a \succ b$ and define all other comparisons arbitrary. By setting $\mu_{\mathscr{X}}(\succ') = \mu_{\mathscr{A}}(\succ)$ for all $\succ$ in the support of $\mu_{\mathscr{A}}$, it is apparent that $(\mu_{\mathscr{X}}, \lambda)$ rationalizes ρ.*

We define one more concept of rationality.

Definition 2.8. *A stochastic choice function ρ is partially RU-rational if there exists $\mu \in \Delta(\mathscr{L}(\mathscr{A}_N))$ such that for any $a \in A \subseteq \mathscr{A}_N$ such that $A \in \mathscr{D}$,*

$$\rho(A, a) = \mu(\succ \in \mathscr{L}(\mathscr{A}_N) | a \succ b \text{ for all } b \in A \setminus a). \tag{4}$$

Remark 2.9. *RU-rationality implies the partial RU-rationality. When the menu contains only non-aggregated categories, each category maps to a single primitive alternative, so the RU-representation collapses to a standard RUM on the set of non-aggregated categories, which corresponds to the partial RU rationality; λ terms disappear because there is no compositional uncertainty on such menus.*

Remark 2.10. *Summarizing the relationship between the three rationality concepts, we have*

$$\text{ARU } (2) \implies \text{ RU } (3) \implies \text{Partial RU } (4).$$

As mentioned earlier, ARU-rationality is the standard assumption in much of the empirical literature. RU-rationality notion introduced in this paper is a natural refinement that accounts for the unknown composition of aggregated categories.

In the next section, we characterize RU-rationality and show precisely how it is weaker than ARU-rationality. The characterization is derived by bridging the gap between RU-rationality and partial RU-rationality.

3 Characterization

In this section, we consider *the outside option setup* in which there exists only one aggregated category $a_0 \in \mathscr{A}$. Formally, we set $\mathscr{A}_A = \{a_0\}$ and $\mathscr{A}_N = \mathscr{A} \setminus \{a_0\}$. We call a_0 *the outside option.*

This case is useful for two reasons. First, it reflects the common modeling practice of including an outside option—an aggregated category that captures unobserved or unlisted choices—frequently used in empirical applications. Although empirical models often involve more than one aggregated category, the outside option is particularly prominent. Second, this simplified setting serves as a foundation for analyzing the more general case with multiple aggregated categories in Section 3.2.

Theorem 3.1. *Consider the outside option setup (i.e., $\mathscr{A}_A = \{a_0\}$). A stochastic choice function ρ is RU-rational if and only if ρ satisfies the following two conditions:*

(i) (Limited Monotonicity): if $b \in D \subseteq \mathscr{A}_N$ then $\rho(D, b) \geq \rho(D \cup \{a_0\}, b)$,

(ii) (Partial RU-rationality) ρ is partially RU rational.

The necessity of (i) and (ii) is immediate. The sufficiency part of the proof is constructive and nontrivial: given a stochatsic chioce function ρ that satisfies the two conditions (i) and (ii), we need to construct an aggregation function, a composition distribution and a distribution over linear orders on $\mathscr{X}$ that satisfies (3). The proof is in the appendix.

The theorem makes precise the weakness of RU-rationality relative to ARU-rationality: RU-rationality is substantially weaker. To see this, note that Limited Monotonicity requires only that adding the outside option does not increase the choice frequencies of *non-aggregated categories*. In particular, it does *not* require the monotonicity condition with respect to menus containing *aggregated categories*. To see this, consider a choice set D such that $a_0 \in D$ and $b \notin D$. The theorem shows that under RU-rationality, it is possible to observe

$$\rho(D, a_0) < \rho(D \cup \{b\}, a_0),$$

which violates the standard monotonicity condition. This may arise because, in the expanded set $D \cup b$, the outside option a_0 may become more desirable than it was in D. For example, in the context of cereal choice, the presence of a fancy, expensive cereal brand may signal that the market corresponds to a high-income area leading to a change in the interpretation or attractiveness of the outside good: the outside goods may contain more appealing options such as smoked salmon or quiche. By contrast, the meanings of non-aggregated categories remain the same when a_0 is added to a choice set with only non-aggregated categories; thus the limited monotonicity condition holds.

ARU-rationality, by contrast, imposes much stronger restrictions including full monotonicity and beyond. When ρ is observed on all subsets of $\mathscr{A}$ (i.e., $\mathscr{D} = 2^{\mathscr{A}} \setminus \{\emptyset\}$), ARU-rationality is equivalent to the non-negativity of all Block-Marschak (BM) polynomials.

Among these, monotonicity is one of the simplest and most intuitive implications. However, as Theorem 3.1 shows, RU-rationality only implies monotonicity on limited choice sets.

This result highlights the substantial gap between RU-rationality and ARU-rationality. In Section 4, we examine the specific conditions on the underlying distributions μ and λ under which RU-rational stochastic choice functions are also compatible with an ARUM.

Two more technical remarks are in order:

Remark 3.2. *In our discussion, we focus on Limited Monotonicity rather than on partial RU-rationality. The reason is straightforward: partial RU-rationality is simply RU-rationality restricted to menus consisting only of non-aggregated categories, and under the full domain assumption (i.e.,* $\mathscr{D} = 2^{\mathscr{A}} \setminus \{\emptyset\}$*) it can be characterized by the non-negativity of Block–Marschak (BM) polynomials defined over sets composed solely of non-aggregated categories.*

Remark 3.3. *In our notion of RU rationalizability, we assume that the composition of categories and the agent's preferences are independent. A natural extension is to allow the composition distribution to depend on the agent's preferences. That is, a composition distribution* $\lambda_{\succ,A}$ *is a distribution over* $\prod_{a\in A}(2^{X(a)} \setminus \{\emptyset\})$ *that may jointly depend on the available categories* A *and the agent's preference* $\succ \in \mathscr{L}(\mathscr{X})$*. In this case, we say that* ρ *is General RU rationalizable if there exists* $(\mu_{\mathscr{X}}, \lambda_{\succ,A})$ *such that,*

$$\rho(A, a) = \sum_{(S_b)_{b\in A} \in \prod\limits_{b\in A}(2^{X(b)} \setminus \{\emptyset\})} \sum_{\succ \in \mathscr{L}(\mathscr{X})} 1(\exists x \in S_a,\ \forall y \in \bigcup_{b\in A\setminus\{a\}} S_b,\ x \succ y))\lambda_{\succ,A}((S_b)_{b\in A})\mu_{\mathscr{X}}(\succ). \tag{5}$$

Since this model includes the independent case, RU rationality implies General RU rationality. Furthermore, it can be shown that General RU rationalizability implies both Limited Monotonicity and Partial RU. Thus RU rationalizability and General RU rationalizability are equivalent.

Remark 3.4. *For the theorem, in order to make sense of Limited Monotonicity, we need the following domain assumption on which choice probabilities are observable: if* ρ *is observable on* $(D \cup \{a_0\}, \cdot)$*, then* ρ *is observable on* $(D, \cdot)$ *(i.e., if* $D \cup \{a_0\} \in \mathscr{D}$*, then* $D \in \mathscr{D}$*). In this case, Limited Monotonicity should be defined as: for all* $D \subseteq \mathscr{A}_N$ *such that* $D \cup \{a_0\} \in \mathscr{D}$*,* ρ *satisfies* $\rho(D, a) \geq \rho(D \cup \{a_0\}, a)$ *for all* $a \in D$*. In the next section (Section 3.1), we omit the assumption and provide an alternative characterization of RU rationality.*

3.1 Vertex representation

In this section, we characterize admissible consumer behaviors under RU-rationality and show that many such behaviors are not ARU-rational. No assumption on the domain is required. However, to simplify exposition, we impose that every observable choice set contains the outside option a_0 (i.e., $a_0 \in D$ for all $D \in \mathscr{D}$), consistent with the common view that a status quo alternative is always available. (Even without this assumption, the main message of the section remains unchanged; see Remark 3.11 for details.)

We first introduce a notation: for each $\succ \in \mathscr{L}(\mathscr{A})$ and $\mathscr{E} \subseteq \mathscr{D}$, define a function denoted by $c_{\mathscr{E}}^{\succ}$ over $\mathscr{D}$ as follows: for all $D \in \mathscr{D}$,

$$c_{\mathscr{E}}^{\succ}(D) = \begin{cases} \max_D \succ & \text{if } D \in \mathscr{E}, \\ a_0 & \text{if } D \notin \mathscr{E}. \end{cases}$$

The function $c_{\mathscr{E}}^{\succ}$ is a choice function that maximizes $\succ$ when $D \in \mathscr{E}$ and otherwise chooses a_0. Now define a degenerate stochastic choice function $\rho_{\mathscr{E}}^{\succ}$ that corresponds to $c_{\mathscr{E}}^{\succ}$. That is, for all $D \in \mathscr{D}$ and $a \in D$,

$$\rho_{\mathscr{E}}^{\succ}(D, a) = 1(c_{\mathscr{E}}^{\succ}(D) = a). \tag{6}$$

Remark 3.5. *$\rho_{\mathscr{E}}^{\succ}$ is RU-rational for all $\succ \in \mathscr{L}(\mathscr{A})$ and $\mathscr{E} \subseteq \mathscr{D}$.*[9]

We can interpret $\rho_{\mathscr{E}}^{\succ}$ as follows. The set $\mathscr{E}$ represents the collection of choice sets in which the agent makes decisions according to the linear order $\succ$. Outside this collection (i.e., on $\mathscr{E}^c$), the agent always selects the outside option a_0. This behavior is natural: $\mathscr{E}$ can be viewed as the domain on which the consumer behaves rationally according to $\succ$, while outside $\mathscr{E}$, *menu effects* lead the agent to default to a_0. For example, in marketing experiments consumers may follow a stable ranking when choosing among familiar products, but when faced with an unfamiliar or unusually large assortment, they often revert to the default option a_0.

We now show the main results of this section:

Theorem 3.6. *Consider the outside option setup (i.e., $\mathscr{A}_A = \{a_0\}$). A stochastic choice*

[9] To see this, fix $\succ \in \mathscr{L}(\mathscr{A})$ and $\mathscr{E} \subseteq \mathscr{D}$. Choose any two elements of $X(a_0)$ denoted by $\overline{x}$ and $\underline{x}$. Consider the following ranking $\succ' \in \mathscr{L}(\mathscr{X})$ such that (i) $\succ$ and $\succ'$ coincide on $\mathscr{A}_N$ (i.e., $a \succ b$ if and only if $x \succ' y$ for all $a, b \in \mathscr{A}_N$, where $X(a) = \{x\}$ and $X(b) = \{y\}$); (ii) for any $y \in \mathscr{A}_N$, $\overline{x} \succ' y \succ' \underline{x}$. For all $D \in \mathscr{D}$, define $\lambda_D(\{\overline{x}\}) = 1$ for any $D \notin \mathscr{E}$; and $\lambda_D(\{\underline{x}\}) = 1$ for any $D \in \mathscr{E}$. It is easy to see that $(\delta_{\succ'}, \lambda)$ RU- rationalizes $\rho_{\mathscr{E}}^{\succ}$. Note that this construction only relies on $|X(a_0)| \geq 2$ which is implied by the outside option approach, thus Remark 3.5 does not depend on $X(a_0)$.

function ρ is RU-rational if and only if

$$\rho \in co.\{\rho^{\succ}_{\mathscr{E}} \mid \succ \in \mathscr{L}(\mathscr{A}) \text{ and } \mathscr{E} \subseteq \mathscr{D}\}.$$

We refer to this set as the RU polytope. Moreover, each $\rho^{\succ}_{\mathscr{E}}$ is a vertex of the RU polytope.

Mathematically, the theorem provides a vertex characterization of RU-rational stochastic choice functions, whereas Theorem 3.1 offers a hyperplane characterization. The vertex characterization shows that RU-rationality permits a type of behavior excluded under ARU-rationality: the agent may sometimes choose the outside option (the default option) a_0, possibly due to menu effects, while in other cases adhering to a rational ranking.

To explicitly compare the RU polytope and the set of stochastic choice fictions that are ARU-rational, we first define the latter set as follows denoted by $\mathcal{P}_{\mathscr{A}}$:

Definition 3.7.

$$\mathcal{P}_{\mathscr{A}} = \left\{ \rho : \rho(A,a) = \sum_{\succ \in \mathscr{L}(\mathscr{A})} \mu_{\mathscr{A}}(\succ)\rho^{\succ}_{\mathscr{D}}(A,a) \quad \text{for all } A \in \mathscr{D},\ a \in A,\ \text{with } \mu_{\mathscr{A}} \in \Delta(\mathcal{L}(\mathscr{A})) \right\},$$

where $\Delta(\mathcal{L}(\mathscr{A}))$ is the set of probability distributions over $\mathcal{L}(\mathscr{A})$ and $\rho^{\succ}_{\mathscr{D}}$ is defined by (6).

Remark 3.8. *The set $\mathcal{P}_{\mathscr{A}}$ is a polytope with vertices $\rho^{\succ}_{\mathscr{D}}$, or*

$$\mathcal{P}_{\mathscr{A}} = co.\{\rho^{\succ}_{\mathscr{D}} \mid \succ \in \mathscr{L}(\mathscr{A})\}.$$

We refer to this set as the ARU polytope.

Corollary 3.9. *The RU polytope contains the ARU polytope $\mathcal{P}_{\mathscr{A}}$ as a subpolytope.*[10]

Theorem 3.6 and Corollary 3.9 together show that the ARU polytope is strictly smaller than the RU polytope: it is a subpolytope whose vertices form only a small subset of the much larger collection of RU vertices. Specifically, the vertices of the RU polytope are given by $\rho^{\succ}_{\mathscr{E}}$ for all $\succ \in \mathscr{L}(\mathscr{A})$ and $\mathscr{E} \subseteq \mathscr{D}$, while the vertices of the ARU polytope are given by $\rho^{\succ}_{\mathscr{D}}$ for all $\succ \in \mathscr{L}(\mathscr{A})$. Thus, ARU vertices vary only with the ranking $\succ$, whereas RU vertices vary with both $\succ$ and the choice of $\mathscr{E}$. See Figure 4 in Section 5.3 for an illustration. The remark quantifies this gap:

Remark 3.10. *The number of vertices (i.e., the number of $\rho^{\succ}_{\mathscr{E}}$) of the RU polytope is double-exponentially larger than the number of vertices of the ARU polytope (i.e., the number of*

[10] A subpolytope is a polytope obtained by taking a subset of the vertices of the original polytope and forming their convex hull.

$\rho^{\succ}_{2^{\mathscr{A}}}$*), provided that* $\mathscr{D}$ *is sufficiently rich. For example, if* $\mathscr{D}$ *contains all sets that include* a_0*, then the ratio of the number of vertices in the RU polytope to the number in the ARU polytope is at least*

$$\frac{2^{(2^n - \binom{n}{2} - 1)}}{n+1},$$

where $|\mathscr{A}_N| = n$*. Since the dominant term in the numerator is double exponential, the ratio grows extremely rapidly. In fact, when* $n \geq 6$*, the ratio already exceeds* $2^{2^{n-1}}$*.*

The next remark shows that as the number of non-aggregated categories n increases, the number of RU vertices grows at a double-exponential rate, whereas the ARU polytope grows much more slowly. This stark contrast highlights that RU-rationality imposes far weaker restrictions than ARU-rationality.

Remark 3.11. *As mentioned, the main result of this section (Theorem 3.6) does not require the domain assumption that* $a_0 \in D$ *for all* $D \in \mathscr{D}$*. The theorem continues to hold with a slight modification of the definition of* $c^{\succ}_{\mathscr{E}}$*. The adjustment is needed because a set* D *may fail to contain* a_0 *even when* $D \notin \mathscr{E}$*. The generalized definition is:*

$$c^{\succ}_{\mathscr{E}}(D) = \begin{cases} \max_D \succ & \text{if } a_0 \notin D \text{ or if } D \in \mathscr{E}, \\ a_0 & \text{otherwise.} \end{cases}$$

This is the same as before, except that $c^{\succ}_{\mathscr{E}}$ *is now defined for menus that do not contain* a_0*. With this modification, Theorem 3.6 holds without change.*

3.2 Multiple aggregated categories

In this section, we consider the case in which there exist multiple aggregated categories as well non-aggregated categories. The main results from the single aggregated category case continue to hold in essentially the same form, as we show below:

Theorem 3.12. *A stochastic choice function* ρ *is RU-rational if and only if* ρ *satisfies the following two conditions:*

(i) (Limited Monotonicity): if $b \in D \subseteq \mathscr{A}_N$ *and* $E \subseteq \mathscr{A}_A$ *then* $\rho(D, b) \geq \rho(D \cup E, b)$*,*

(ii) (Partial RU-rationality) ρ *is partially RU rational.*

The only substantive difference from Theorem 3.1 is that the condition (i) becomes stronger. In the earlier theorem, the condition (i) required monotonicity with respect to the addition of a single aggregated category—namely, the outside option a_0. In the current theorem, where multiple aggregated categories are present, the condition (i) instead requires monotonicity with respect to the addition of arbitrary subsets E of the aggregated categories. This is a natural extension of Theorem 3.1: Theorem 3.1 corresponds to the case in which $\mathscr{A}_A$ is singleton $\{a_0\}$. Therefore, all of the implications of Theorem 3.1, as explained in the previous section hold with Theorem 3.12.

The proof is constructive and more involved as we need to define a composition distribution, or the distribution function defined on the product space $\prod_{a\in\mathscr{A}_A}(2^{X(a)}\setminus\{\emptyset\})$, not just on $2^{X(a_0)}\setminus\{\emptyset\}$. We provide the formal proof is in the appendix.

Remark 3.13. *In line with the single aggregated category case, the only domain assumption required for Theorem 3.12 to hold is that if $D\cup E\in\mathscr{D}$ for $D\subseteq\mathscr{A}_N$ and $E\subseteq\mathscr{A}_A$, then $D\in\mathscr{D}$. This assumption is only needed in order to define Limited Monotonicity and plays no conceptual role beyond that.*

4 Conditions Restoring ARUM

This section investigates the conditions under which RU-rationality implies ARU-rationality. These results are especially relevant given the widespread use of ARUMs in empirical IO and the common assumption that the data-generating process satisfies RU-rationality. They provide concrete guidance on how to define aggregated categories in a way that justifies the use of ARUMs.

In particular, we provide sufficient restrictions on the underlying preference distribution μ_X and the composition distribution λ that ensure RU-rationality and ARU-rationality coincide on the observed choice domain $\mathscr{A}$. The two conditions we introduce are independent: one applies to μ_X (Section 4.1) and the other to λ (Section 4.2).

4.1 Condition on $\mu_{\mathscr{X}}$: non-overlappingness

In this section we provide a restriction on the distribution of preferences that recovers the implications of ARUM. The restriction says for each aggregated category a, the underlying alternatives in $X(a)$ are similar in utility.

Definition 4.1. *Given an aggregation function X, a distribution $\mu_{\mathscr{X}}\in\Delta(\mathscr{L}(\mathscr{X}))$ is non-overlapping if for any $\succ\in supp(\mu_{\mathscr{X}})$, $a\in\mathscr{A}_A$, and $y,z\in X(a)$,*

$$y\succ x\succ z \implies x\in X(a).$$

The definition says that any preference in the support of $\mu_{\mathscr{X}}$ has the property that each $X(a)$ does not overlap with other aggregated categories in terms of preference ranking.[11]

Proposition 4.2. *A stochastic choice function ρ is RU rational with a non-overlapping $\mu_{\mathscr{X}}$ if and only if ρ is ARU rational.*

Proposition 4.2 shows that, without any assumptions on λ, the use of ARUM is valid when the alternatives grouped within each aggregate occupy similar positions in agents' preference rankings. Consequently, when constructing an aggregated data set, alternatives with closely related characteristics should not be assigned to different aggregates. A similar observation has been made empirically by Stafford (2018) using a dataset of lobster fishing. In particular, she demonstrated that classifying crab-fishing sites as part of the outside option leads to biased coefficient estimates. Because lobster and crab fishing are close substitutes, an excellent lobster site is preferred to a poor crab site and vice-versa, their rankings overlap, thereby violating the non-overlapping condition specified in Proposition 4.2.

4.2 Condition on λ: menu-independence

The second condition that we discuss is that the distribution of aggregated categories is the same across choice sets, or markets.

Definition 4.3. *Given an aggregation function X, a composition distribution λ is said to be menu independent if there exists $\lambda \in \Delta\Big(\prod_{a\in\mathscr{A}}(2^{X(a)} \setminus \{\emptyset\})\Big)$ such that for any $B \subseteq \mathscr{A}$ and $(S_a)_{a\in B} \in \prod_{a\in B}(2^{X(a)} \setminus \{\emptyset\})$,*

$$\lambda_B((S_a)_{a\in B}) = \lambda\Big(\prod_{a\in B}(S_a) \times \prod_{a\in\mathscr{A}\setminus B}(2^{X(a)} \setminus \{\emptyset\})\Big).^{12} \tag{7}$$

Remark 4.4.

- *To explain the meaning of our condition, consider a stronger condition as follows: there exists $\lambda^a \in \Delta(2^{X(a)} \setminus \{\emptyset\})$ for each $a \in \mathscr{A}$ such that for any $B \subseteq \mathscr{A}$ and $(S_a)_{a\in B} \in \prod_{a\in B}(2^{X(a)} \setminus \{\emptyset\})$,*

$$\lambda_B((S_a)_{a\in B}) = \prod_{a\in B} \lambda^a(S_a). \tag{8}$$

[11] It is possible to weaken the definition as follows but we keep this condition for simplicity: A distribution $\mu_{\mathscr{X}} \in \Delta(\mathscr{L}(\mathscr{X}))$ is almost non-overlapping if for any $\succ \in supp(\mu_{\mathscr{X}})$, there exists $b, c \in \mathscr{A}$ such that (i) $y \succ x$ for all $x \in X(b) \cup X(c)$ and $y \in \mathscr{X} \setminus (X(b) \cup X(c))$ and (ii) for all $a \in \mathscr{A}_A \setminus \{b, c\}$, and $y, z \in X(a)$, $y \succ x \succ z \implies x \in X(a)$.

[12] The right hand side can be written simply as follows $\lambda\Big((S_a)_{a\in B}, (2^{X(a)} \setminus \{\emptyset\})_{\mathscr{A}\setminus B}\Big)$

The condition (7) says that there exists a joint distribution that works across choice sets.[13] *(7) is weaker than condition (8) in the sense that it allows correlation across aggregated categories. The two conditions coincide when there is a single aggregated category. In that case, both reduce to menu-independent aggregation:* $\lambda_D = \lambda_E$ *for all* $D, E \subseteq \mathscr{A}$.

- *By the finiteness and the additivity, for any* $B \subseteq \mathscr{A}$ *and* $(S_a)_{a\in B} \in \prod_{a\in B}(2^{X(a)} \setminus \{\emptyset\})$, *the menu independence implies*

$$\lambda_B((S_a)_{a\in B}) = \sum_{(S_c)_{a\in B^c} \in \prod_{c\in B^c}(2^{X(c)}\setminus\{\emptyset\})} \lambda((S_a)_{a\in B}, (S_c)_{c\in B^c}).$$

When λ is independent, representation (3) in Definition 2.5 can be simplified as in the following lemma.

Lemma 4.5. *Given* $\mu_{\mathscr{X}} \in \Delta(\mathscr{L}(\mathscr{X}))$ *and a menu independent composition distribution* λ, *we have for all* $A \subseteq \mathscr{A}$ *and* $a \in A$

$$\rho(A, a) = \sum_{(S_b)_{b\in\mathscr{A}}} \lambda((S_b)_{b\in\mathscr{A}})\rho_{\mathscr{X}}(\cup_{b\in A}S_b, S_a),$$

where $\rho_{\mathscr{X}}(\cup_{b\in A}S_b, S_a) = \mu_{\mathscr{X}}(\succ\in \mathscr{L}(\mathscr{X}) | \exists x \in S_a,\ \forall y \in \cup_{b\in A\setminus a}S_b,\ x \succ y)$

Lemma 4.5 means that the aggregated stochastic choice function ρ is a convex combination of $\rho_{\mathscr{X}}$ for each composition of the aggregated categories. If $\rho_{\mathscr{X}}$ is RUM, then ρ is also a RUM. This intuition is formalized in the following proposition.

Proposition 4.6. *A stochastic choice function* ρ *is RU rational with menu independent* λ *if and only if* ρ *is ARU rational.*

Proposition 4.6 shows that if the composition distribution λ is menu independent, then each aggregated category can be treated as a single good. Whether this assumption holds depends on the setting. In some markets, λ may indeed be menu independent. A natural example is automobiles: when alternatives are aggregated by brand, the distribution of models within a brand is relatively stable, conditional on that brand being available, and thus invariant to which other brands are offered. In other markets, however, λ is menu

[13]For example, consider $a, b \in \mathscr{A}_A$ and let $X(a) = \{x_1, y_1\}$ and $X(b) = \{x_2, y_2\}$. Definition 2.5 allows the situation that a means alternative x_1 (y_1) if and only if b means alternative x_2 (y_2, respectively); and a means x_1 and y_1 equally likely. In this example, the probability that a means x_1 and b means y_2 is zero, which is not allowed in (8).

dependent. As discussed, the outside option often varies across markets (e.g., with income, store access, or local amenities), so its composition—and hence λ_A—changes with the menu A.

5 Simulation

The theoretical results from the previous section show that, in general, the implications of RU-rationality for the observed dataset ρ are too weak to guarantee ARU-rationality except under two special conditions: when the composition distribution λ is menu independent, or when the underlying preference structure is non-overlapping.

These results motivate the following conjecture: *As the composition distribution λ departs from menu independence, or as preferences depart from the non-overlapping condition, the likelihood that ρ can be rationalized by an ARUM diminishes. Consequently, estimation based on an ARUM, such as the logit model, is expected to yield increasingly biased utility estimates.* This section empirically investigates this conjecture through simulation analysis and also examines *how large such biases can be*, an important practical question.

Specifically, we generate true unobservable choice data $\rho_{\mathscr{X}}$ defined over the set $\mathscr{X}$ of underlying alternatives using a logit model with fixed utilities. Given $\rho_{\mathscr{X}}$ and a specified composition distribution λ as well as an aggregation function X, we construct the observed dataset ρ on the set $\mathscr{A}$ of categories. We then estimate utility parameters from ρ using a logit specification and measure the resulting estimation bias. In addition, we compute the distance from ρ to the set of ARUMs (i.e., the ARU polytope) to evaluate the extent of rationality loss.

Our analysis is organized around the following conjectures:

(A) As λ becomes more menu dependent or preferences more overlapping, ρ is increasingly unlikely to be rationalized by an ARUM, let alone by a logit model defined over $\mathscr{A}$. This leads to larger estimation biases (see Section 5.2).

(B) The reason for Conjecture (A) is that the distance from ρ to the ARU polytope grows with greater menu dependence or preference overlap (see Section 5.3).

We first confirm Conjecture (A). In particular, we demonstrate that *the resulting biases are often substantial enough not only to distort estimated utility levels but also to overturn inferred preference orderings.* We then demonstrate that Conjecture (B) explains why this occurs.

To establish these conjectures, we fix certain parameter values that are not the main focus of the analysis. Section B of the online appendix reports simulations using alternative parameter values.

5.1 Setup

Consider three alternatives x, y, a_0, where x, y are non-aggregated categories[14]; a_0 is an aggregated category (the outside option). The outside option can mean good z or w. That is, $X(a_0) = \{z, w\}$. Given fixed utility values $u(x), u(y), u(z)$, and $u(w)$, we generate the true dataset $\rho_{\mathscr{X}}$ over the set $\mathscr{X}$ of underlying alternatives by assuming the logit model: for all $D \subseteq \{x, y, z, w\}$ and $i \in D$,

$$\rho_{\mathscr{X}}(D, i) = \frac{\exp(u(i))}{\sum_{j \in D} \exp(u(j))}.$$

We examine three distinct markets, each represented by one of the following choice sets: $\{x, a_0\}$, $\{y, a_0\}$, and $\{x, y, a_0\}$. Importantly, the composition of the aggregated category a_0 varies across these markets; that is, $\lambda_{\{x,a_0\}}$, $\lambda_{\{y,a_0\}}$, and $\lambda_{\{x,y,a_0\}}$ over $\{\{w\}, \{z\}, \{w, z\}\}$ are allowed to differ. With $\rho_{\mathscr{X}}$ and λ, we construct the observable reduced dataset ρ over the set $\mathscr{A}$ of aggregated categories. That is for any $D \subseteq \{x, y\}$ and $i \in D$, we define $\rho(D \cup a_0, i)$ as

$$\lambda_{D \cup a_0}(\{w\})\rho_{\mathscr{X}}(D \cup \{w\}, i) + \lambda_{D \cup a_0}(\{z\})\rho_{\mathscr{X}}(D \cup \{z\}, i) + \lambda_{D \cup a_0}(\{z, w\})\rho_{\mathscr{X}}(D \cup \{z, w\}, i).$$

Given the observed aggregated dataset ρ, we estimate the utility parameters $\hat{u}(x)$ and $\hat{u}(y)$ under the assumption that ρ follows a multinomial logit model defined over $\mathscr{A}$ and that the utility of the outside option is zero, i.e., $\exp(u(a_0)) = 1$. Note that in this setup

- The composition distribution λ becomes more *menu dependent* as $\lambda_{\{x,a_0\}}$, $\lambda_{\{y,a_0\}}$, and $\lambda_{\{x,y,a_0\}}$ become more distinct.
- Fixing the true utilities of $u(x)$ and $u(y)$—say, $u(x) > u(y)$—the preference structure becomes *more overlapping* as $\max\{u(z), u(w)\}$ exceeds $u(x)$ and $\min\{u(z), u(w)\}$ falls below $u(y)$.

[14]For simplcity, we abuse the notation x and y to denote both the non-aggregated categories and their respective underlying alternatives.

5.2 Estimation Bias from ARUM Misspecification

To confirm Conjecture (A), we quantify the estimation bias using the following measure:

$$(\hat{u}(x) - \hat{u}(y)) - (u(x) - u(y)), \tag{9}$$

where $\hat{u}(x)$ and $\hat{u}(y)$ denote the maximum-likelihood estimates of $u(x)$ and $u(y)$, respectively. This measure captures the distortion in relative utilities between alternatives x and y. Remember that in our setup the values of $u(x)$ and $u(y)$ are fixed say $u(x) = 2$ and $u(y) = 1$. We are interested in the values of the measure as well as the sign of the measure. When the measure is less than -1 it means that in the estimate we have $\hat{u}(y) > \hat{u}(x)$, even though in the true data we have $u(x) > u(y)$. This means that bias affects not only the estimated utility levels but also the inferred preference orderings.

The values of the measure are also meaningful. To see this, consider the case in which $\hat{u}(x) - \hat{u}(y) > u(x) - u(y)$. Then, by taking the exponential of the measure, we have

$$\frac{\exp(\hat{u}(x))}{\exp(\hat{u}(y))} \Big/ \frac{\exp(u(x))}{\exp(u(y))},$$

which represents the ratio of the estimated odds ratio to the true odds ratio of choosing x over y. A deviation in this ratio reflects how the estimated utilities—and thus predicted behavior—diverges from the true underlying model.

5.2.1 Effect of composition distribution λ on bias

In the first simulation, we test Conjecture (A) focusing on the role of λ: *the bias defined in (9) becomes larger as λ becomes more menu dependent (i.e., as λ changes across choice sets $\{x, a_0\}, \{y, a_0\}$, $\{x, y, a_0\}$).* As the preference structure is not the focus of the analysis, we fixed the values $u(x) = 2$, $u(y) = 1$, $u(z) = 3, u(w) = 0$.

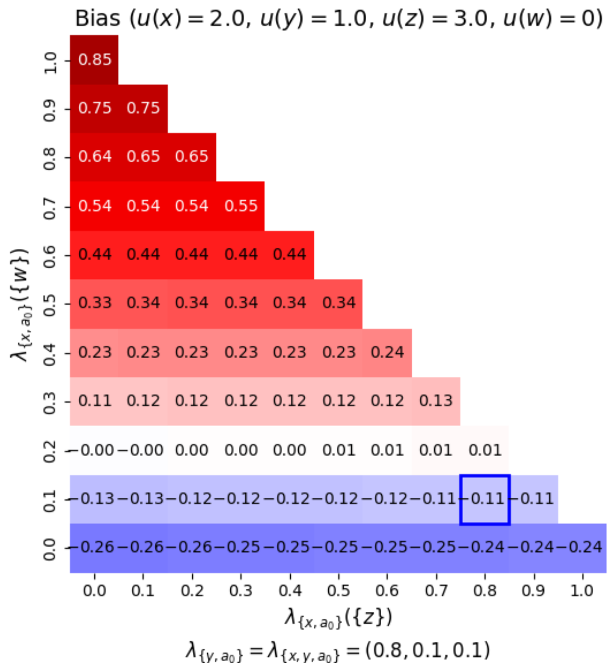


Figure 1: Heatmap of biases across values of $\lambda_{\{x,a_0\}}$, with $\lambda_{\{y,a_0\}}$ and $\lambda_{\{x,y,a_0\}}$ fixed at $(0.8, 0.1, 0.1)$ and utilities fixed at $u(x) = 2$, $u(y) = 1$, $u(z) = 3$, $u(w) = 0$. The horizontal axis reports $\lambda_{\{x,a_0\}}(\{z\})$, the vertical axis reports $\lambda_{\{x,a_0\}}(\{w\})$, and color intensity represents the magnitude of the bias defined in (9). The blue-outlined cell marks the independent case, where $\lambda_{\{x,a_0\}} = \lambda_{\{y,a_0\}} = \lambda_{\{x,y,a_0\}}$; cells farther from the blue cell correspond to more menu-dependent cases, where the bias is larger.

First, we construct a heatmap to visualize the sensitivity of bias to deviations from the independence case. In Figure 1, we vary the values of $\lambda_{\{x,a_0\}}$ while equating and holding $\lambda_{\{x,y,a_0\}}$ and $\lambda_{\{y,a_0\}}$ fixed at $(0.8,\ 0.1,\ 0.1)$, where the first, second, and third entries represent the probabilities assigned to $\{z\}$, $\{w\}$, and $\{z,w\}$, respectively. Note that the horizontal axis measures $\lambda_{\{x,a_0\}}(\{z\})$, while the vertical axis measures $\lambda_{\{x,a_0\}}(\{w\})$. (Note that the numbers $\lambda_{\{x,a_0\}}(\{z\})$ and $\lambda_{\{x,a_0\}}(\{w\})$ determine $\lambda_{\{x,a_0\}}$ uniquely as $\lambda_{\{x,a_0\}}(\{z,w\}) = 1 - \lambda_{\{x,a_0\}}(\{w\}) - \lambda_{\{x,a_0\}}(\{z\})$.) Each grid point in the heatmap represents a specific assignment of $\lambda_{\{x,a_0\}}$; the bias is computed at each grid point. The blue-outlined cell corresponds to the independent case, in which $\lambda_{\{x,a_0\}}(\{z\}) = 0.8$ and $\lambda_{\{x,a_0\}}(\{w\}) = 0.1$, implying that $\lambda_{\{x,a_0\}} = \lambda_{\{y,a_0\}} = \lambda_{\{x,y,a_0\}}$.

The heatmap supports Conjecture (A): bias grows as λ becomes more menu dependent—that is, as the cell moves further away from the independent case. The mechanism is straightforward: positive bias arises when x is overvalued relative to a_0 in the menu $\{x, a_0\}$, while negative bias arises when x is undervalued relative to a_0 in the menu $\{x, a_0\}$. When a_0 consists of $\{w\}$, it is relatively unattractive, whereas when it consists of $\{z\}$ or $\{z, w\}$, it is more attractive, since in our setup $u(z)$ is high and $u(w)$ is low. Hence, the large positive bias in the figure occurs when a_0 is composed of w more often in the menu $\{x, a_0\}$ (e.g., when $\lambda_{\{x,a_0\}}(\{w\}) = 1$). Conversely, the negative bias arises when a_0 is composed of $\{z\}$ or $\{z, w\}$ (e.g., when $\lambda_{\{x,a_0\}}(\{w\}) = 0$).

5.2.2 Maximum and minimum biases and the independent case

To construct Figure 1, we assumed $\lambda_{\{y,a_0\}} = \lambda_{\{x,y,a_0\}}$. In this section, we relax this assumption and vary the values of $\lambda_{\{y,a_0\}}$ and $\lambda_{\{x,y,a_0\}}$ and report the resulting maximum and minimum biases in Figure 2. Each point in the figure corresponds to four possible bias measures, evaluated at nonnegative values of $\lambda_{\{x,y,a_0\}}(\{w\})$ and $\lambda_{\{x,y,a_0\}}(\{z\})$ satisfying $\lambda_{\{x,y,a_0\}}(\{w\}) + \lambda_{\{x,y,a_0\}}(\{z\}) \leq 1$:

- (1) Maximum positive bias, obtained by varying all values of $\lambda_{\{x,a_0\}}$ and $\lambda_{\{y,a_0\}}$ (red)
- (2) Maximum negative bias, obtained by varying all values of $\lambda_{\{x,a_0\}}$ and $\lambda_{\{y,a_0\}}$ (purple)
- (3) Minimum absolute bias, obtained by varying all values of $\lambda_{\{x,a_0\}}$ and $\lambda_{\{y,a_0\}}$ (blue)
- (4) Independent case, where $\lambda_{\{x,y,a_0\}} = \lambda_{\{x,a_0\}} = \lambda_{\{y,a_0\}}$ (green)

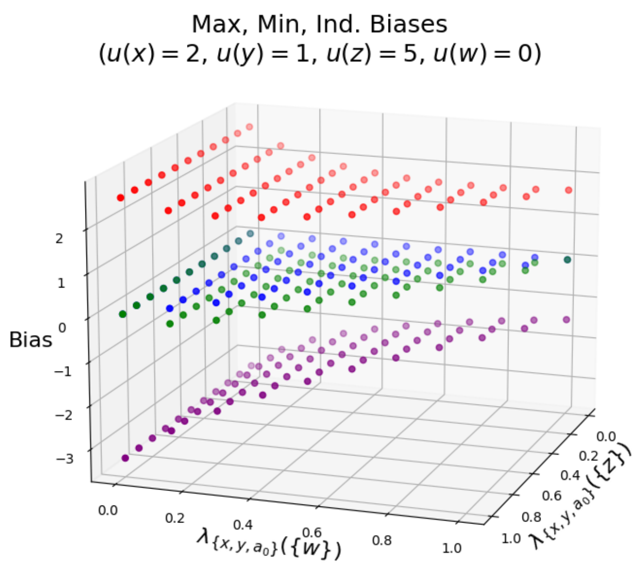


Figure 2: Maximum positive bias (red), maximum negative bias (purple), minimum absolute bias (blue), and bias under independent λ (green), plotted across values of $\lambda_{\{x,y,a_0\}}$. The horizontal axis reports $\lambda_{\{x,y,a_0\}}(\{w\})$ and the vertical axis reports $\lambda_{\{x,y,a_0\}}(\{z\})$. The maximum and minimum biases are obtained by optimizing over all admissible values of $\lambda_{\{x,a_0\}}$ and $\lambda_{\{y,a_0\}}$.

The figure illustrates that, in the independent case, the biases are significantly smaller than the maximum biases and are often close to the minimum biases.[15] Notably, both the maximum positive and negative biases can be substantial in magnitude. Specifically, the bias can exceed 2 or fall below -3. As mentioned, when the bias $(\hat{u}(x) - \hat{u}(y)) - (u(x) - u(y))$ is less than -1, it implies that the estimated utilities satisfy $\hat{u}(x) < \hat{u}(y)$, despite the true ordering being $u(x) > u(y)$. Moreover, the difference is large enough that y appears substantially better than x in the estimates. This demonstrates that *bias distorts not only the estimated utility levels but also the inferred preference orderings, potentially leading to incorrect conclusions and decisions.*

Conversely, when the bias exceeds 2, we obtain $\hat{u}(x) - \hat{u}(y) > 3$, indicating a substantial overstatement of the utility difference. In terms of odds ratios, this implies

$$\frac{\exp(\hat{u}(x))}{\exp(\hat{u}(y))} > e^2 \cdot \frac{\exp(u(x))}{\exp(u(y))}.$$

In other words, the estimated odds ratio overstates the true odds ratio by a factor greater than $e^2 \approx 7$. Such distortions in relative odds can result in highly inaccurate predictions of choice probabilities.

5.2.3 Effect of preference structure on bias

In the last simulation on biases, we vary the preference structure to evaluate how overlapping preferences affect bias. We assume that the true unobservable dataset $\rho_{\mathscr{X}}$ is generated by logit models with $u(x) = 2$, $u(y) = 1$, and various values of $u(z)$ and $u(w)$ on a uniform grid over $[-5, 5]$. The composition distributions λ are fixed to be non-independent, since λ is not the focus of this analysis. Specifically, we set $\lambda_{\{x,y,a_0\}} = (0.8, 0.1, 0.1)$, $\lambda_{\{x,a_0\}} = (0.8, 0.1, 0.1)$, and $\lambda_{\{y,a_0\}} = (0.1, 0.8, 0.1)$, where the three coordinates represent the probabilities of $\{z\}$, $\{w\}$, and $\{z, w\}$, respectively.

For each pair $(u(z), u(w))$, we compute the bias defined in (9). Since $u(x) = 2$ and $u(y) = 1$, violations of the non-overlapping condition are more likely when $u(z)$ and $u(w)$ are far apart, with $\max\{u(z), u(w)\} > 2$ and $\min\{u(z), u(w)\} < 1$.

Figure 3 displays the resulting heatmap, with $u(z)$ and $u(w)$ on the axes. Red indicates positive bias at the top left corner and blue indicates negative bias at the bottom right corner, with color intensity reflecting the magnitude. The figure confirms our Conjecture (A): bias is largest when the preference structure is highly overlapping. In particular, as $u(z)$ and $u(w)$ move farther apart, cells lie farther from the diagonal, and the magnitude of

[15] In fact, when $\lambda_{\{x,y,a_0\}}(\{w\}) = 1$ or $\lambda_{\{x,y,a_0\}}(\{w\}) = 0$, the blue points and the green points coincide.

bias increases. Conversely, when both $u(z)$ and $u(w)$ are either very high or very low, the bias decreases.

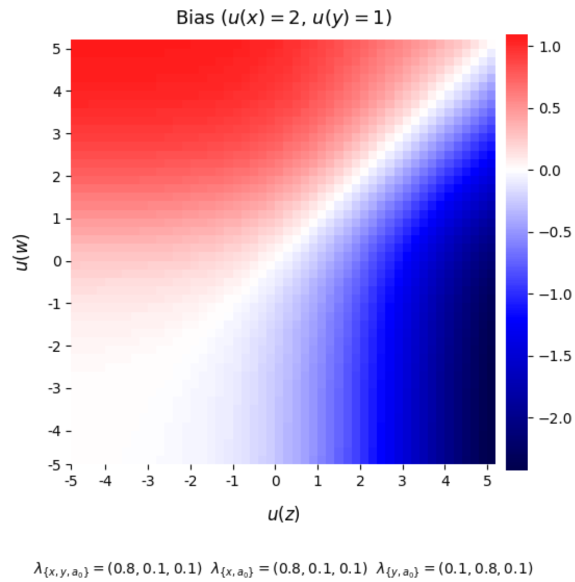


Figure 3: Heatmap of biases across values of $u(z)$ and $u(w)$, with $u(x) = 2$, $u(y) = 1$, and composition distributions fixed at $\lambda_{\{x,y,a_0\}} = (0.8, 0.1, 0.1)$, $\lambda_{\{x,a_0\}} = (0.8, 0.1, 0.1)$, and $\lambda_{\{y,a_0\}} = (0.1, 0.8, 0.1)$. The horizontal axis reports $u(z)$, the vertical axis reports $u(w)$, and color intensity indicates the magnitude of the bias defined in (9) (red for positive bias, blue for negative bias). Biases are largest when preferences overlap strongly, i.e., when $u(z)$ and $u(w)$ straddle $u(x) = 2$ and $u(y) = 1$.

The intuition is as follows. In this setup, a_0 is likely to be composed of z in both $\{x, y, a_0\}$ and $\{x, a_0\}$, since $\lambda_{\{x,y,a_0\}}(\{z\}) = \lambda_{\{x,a_0\}}(\{z\})$ is high, whereas a_0 is likely to be composed of w in $\{y, a_0\}$, given that $\lambda_{\{y,a_0\}}(\{w\})$ is high. Consequently, when $u(w) > u(z)$, the relative attractiveness of x in $\{x, a_0\}$ and $\{x, y, a_0\}$ is high, leading to the overestimation of $\hat{u}(x)$; and the relative attractiveness of y in $\{y, a_0\}$ is low, leading to an underestimation of $\hat{u}(y)$ and thus a positive bias in the top-left corner. Conversely, when $u(z) > u(w)$, the relative attractiveness of x in $\{x, y, a_0\}$ and $\{x, a_0\}$ is low, leading to an underestimation of $\hat{u}(x)$; and the relative attractiveness of y in $\{y, a_0\}$ is high, leading to an overestimation of $\hat{u}(y)$. This results in a negative bias in the bottom-right corner.

5.3 Distance to the ARU Polytope

The previous subsection confirmed Conjecture (A), namely that estimation biases increase as the composition distribution λ becomes more menu dependent and the preference structure becomes more overlapping. The purpose of this subsection is to explain why it holds. Our explanation is provided by Conjecture (B), which states that *the distance from the observed dataset ρ to the ARU polytope grows as the composition distribution λ becomes more menu dependent or as the underlying preferences become more overlapping.*

To quantify the deviation of the observable dataset ρ from ARU-rationality, we calculate the minimum distance between ρ and the set $\mathcal{P}_{\mathscr{A}}$ of ARUMs (i.e., the ARU polytope). (See Definition 3.7 in Section 3.1 for the definition.) Recall that a dataset ρ belongs to $\mathcal{P}_{\mathscr{A}}$ if and only if it is ARU rational. Hence, the minimal distance from ρ to $\mathcal{P}_{\mathscr{A}}$ provides a natural metric for assessing the extent of violation of the ARU rationality. The minimum distance is defined as follows:

$$\min_{\rho' \in \mathcal{P}_{\mathscr{A}}} \|\rho - \rho'\|. \tag{10}$$

Figure 4 shows a point ρ that is RU rational but not ARU rational. The line from ρ to the red triangle shows the minimal Euclidean distance from ρ to the ARU polytope $\mathcal{P}_{\mathscr{A}}$ which is given by the orthogonal projection.

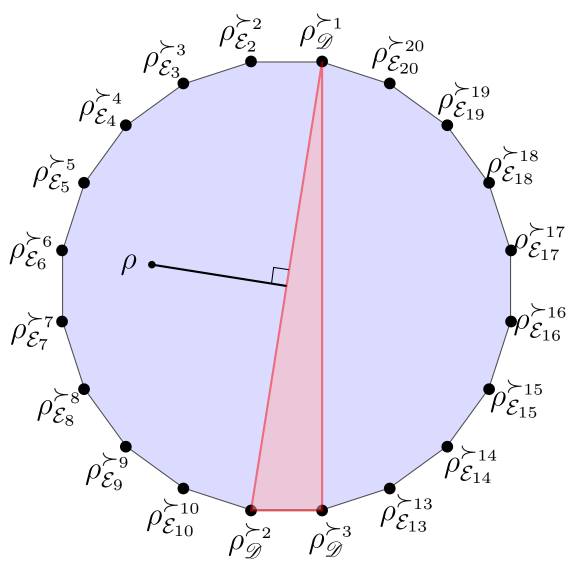


Figure 4: Random utility polytope (blue) and aggregated random utility polytope $\mathcal{P}_{\mathscr{A}}$ (red). The orthogonal projection illustrates the distance from ρ to $\mathcal{P}_{\mathscr{A}}$

To see how the minimum distance can be calculated in our setup, remember that we have three alternatives x, y, a_0. Thus the number of linear orders is 6; the number of coordinates of ρ (i.e., (D, i) such that $i \in D \subseteq \{x, y, a_0\}$) is 12. Then, the distance (10) can be calculated as follows: given $\rho \in \mathbf{R}^{12}$,

$$\min_{\mu \in \Delta^5} \left\| \rho - \sum_{i=1}^{6} \mu_i \rho^{\succ_i}_{\mathscr{D}} \right\|_2^2, \tag{11}$$

where the norm is the Euclidean norm and $\rho_{\mathscr{D}}^{\succ_i}$ is defined by equation (6). Here, Δ^5 is the five dimensional simplex and an element of the simplex satisfies $\mu \in \mathbb{R}^6$, $\mu_i \geq 0$, and $\sum_i \mu_i = 1$.

5.3.1 Effect of composition distribution λ on distance

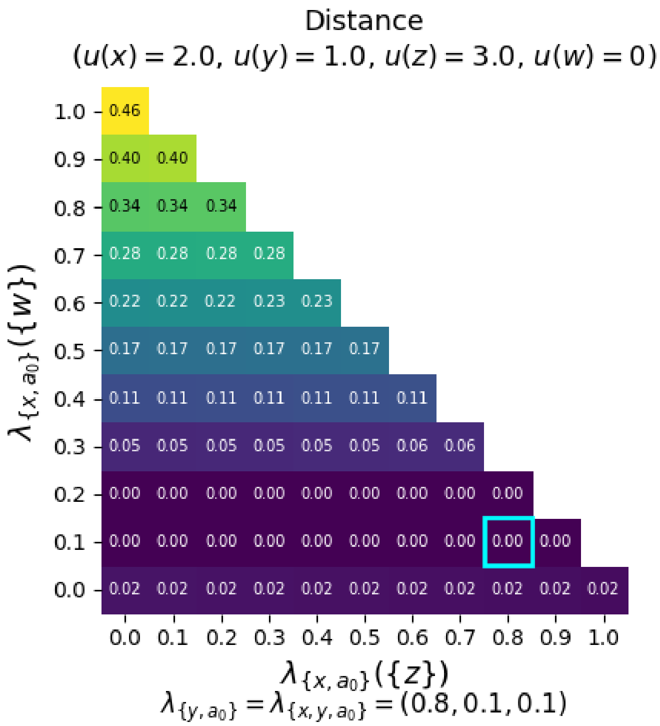


Figure 5: Heatmap of the distance defined in (11) across values of $\lambda_{\{x,a_0\}}$, with $\lambda_{\{y,a_0\}}$ and $\lambda_{\{x,y,a_0\}}$ fixed at $(0.8, 0.1, 0.1)$ and utilities fixed at $u(x) = 2$, $u(y) = 1$, $u(z) = 3$, and $u(w) = 0$. The horizontal axis reports $\lambda_{\{x,a_0\}}(\{z\})$, the vertical axis reports $\lambda_{\{x,a_0\}}(\{w\})$, and color intensity indicates the magnitude of the distance. The blue-outlined cell marks the independent case ($\lambda_{\{x,a_0\}} = \lambda_{\{y,a_0\}} = \lambda_{\{x,y,a_0\}}$); cells farther from the blue cell correspond to more menu-dependent cases, where the distance is larger.

In this simulation, we examine how the distance defined in (11) varies with λ while holding utility values fixed. We use the same utility and λ specifications as in Subsection 5.2.1. Specifically, we vary $\lambda_{\{x,a_0\}}$ while fixing $\lambda_{\{x,y,a_0\}}$ and $\lambda_{\{y,a_0\}}$.

We compute the distance at each grid point. Figure 5 reports the resulting heatmap. Consistent with Proposition 4.6, the distance is zero in the independent case (blue-outlined cell). As λ moves farther from the menu independence, the distance increases, confirming Conjecture (B).

To understand the figures, recall that we consider three alternatives x, y, and a_0. Thus, it can be shown that ρ is ARUM if and only if it satisfies the standard monotonicity condition for all menus. Since ρ is RU-rational, it satisfies Limited Monotonicity, that is monotonicity on $\{x, y\}$ with respect to adding a_0. Also, since $\lambda_{\{y,a_0\}} = \lambda_{\{x,y,a_0\}}$, monotonicity is satisfied on menu $\{y, a_0\}$ with respect to adding x. Thus the only possible ways to violate monotonicity are (i) $\rho(\{x, y, a_0\}, a_0) > \rho(\{x, a_0\}, a_0)$ and (ii) $\rho(\{x, y, a_0\}, a_0) > \rho(\{y, a_0\}, a_0)$. The first occurs when a_0 is less attractive on $\{x, a_0\}$ than on $\{x, y, a_0\}$ and the second occurs when a_0 is more attractive on $\{x, a_0\}$ than on $\{x, y, a_0\}$. Since $u(z)$ is high and $u(w)$ is low, the first case occurs when a_0 is more likely to be $\{w\}$ (e.g., $\lambda_{\{x,a_0\}}(\{w\}) = 1$); the second case occurs when a_0 is more likely to be $\{z\}$ or $\{z, w\}$ (e.g., $\lambda_{\{x,a_0\}}(\{w\}) = 0$). These are two cases where the distance is large in Figure 5.

5.3.2 Effect of preference structure on distance

Finally, we examine how the distance varies with the preference structure while keeping λ fixed. The simulation setup follows Subsection 5.2.3: we fix $u(x) = 2$ and $u(y) = 1$, and vary $u(z)$ and $u(w)$ over a uniform grid.

We produced a heatmap with axes corresponding to $u(z)$ and $u(w)$ and color intensity represents the magnitude of the distance. Figure 6 shows that the distance is largest in the lower-right corner, where the preference structure exhibits strong overlap, and smallest along the diagonal, where preferences are closer to non-overlapping, which is consistent with Proposition 4.2 and Conjecture (B).

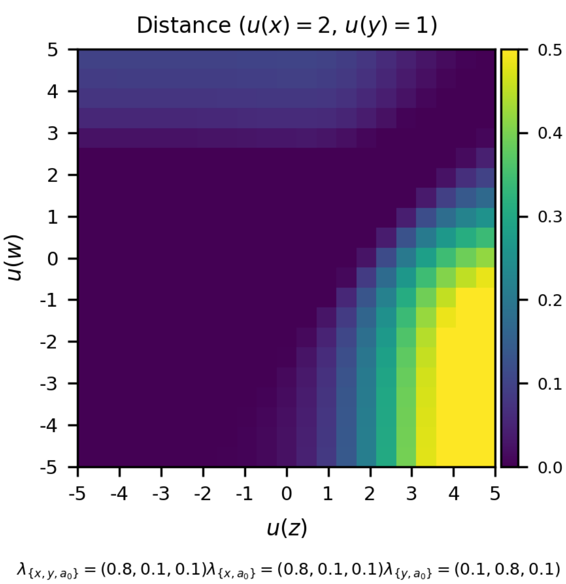


Figure 6: Heatmap of the distance defined in (11) across values of $u(z)$ and $u(w)$, with $u(x) = 2$, $u(y) = 1$, and composition distributions fixed at $\lambda_{\{x,y,a_0\}} = (0.8, 0.1, 0.1)$, $\lambda_{\{x,a_0\}} = (0.8, 0.1, 0.1)$, and $\lambda_{\{y,a_0\}} = (0.1, 0.8, 0.1)$. The horizontal axis reports $u(z)$, the vertical axis reports $u(w)$, and color intensity indicates the magnitude of the distance. The distance is largest in the lower-right corner and smaller along the diagonal.

The intuition behind the figure is as follows. As in the previous example, ARUM is characterized by monotonicity, so we can interpret the figure by identifying the parameters that generate the largest violations of monotonicity. Because $\lambda_{\{x,y,a_0\}} = \lambda_{\{x,a_0\}}$, monotonicity violations only occur when comparing $\{x, y, a_0\}$ and $\{y, a_0\}$.

When $u(z)$ is high and $u(w)$ is low, a_0 is relatively more attractive in $\{x, y, a_0\}$ than in $\{y, a_0\}$. This is because a_0 is more often composed of z in $\{x, y, a_0\}$, whereas it is more often composed of w in $\{y, a_0\}$. This implies $\rho(\{x, y, a_0\}, a_0) > \rho(\{y, a_0\}, a_0)$, producing a large violation of monotonicity in the bottom-right corner. By contrast, when $u(w)$ is high and $u(z)$ is low, a_0 becomes relatively less attractive in $\{x, y, a_0\}$ than in $\{y, a_0\}$. As a result, y is more attractive in $\{y, a_0\}$, so that $\rho(\{x, y, a_0\}, y) > \rho(\{y, a_0\}, y)$—again a violation of monotonicity. However, because y is much less attractive than x (i.e., $u(x) > u(y)$), the size of this violation is considerably smaller than in the previous case. This explains why the distances in the top-left corner are smaller than those in the bottom-right.

6 Summary

We study discrete choice with aggregated categories whose compositions are unobserved and show that RU-rationality imposes only limited implications for the observed dataset ρ: (i) Limited Monotonicty (adding aggregated categories to a menu of non-aggregated categories decreases choice probabilites) (ii) Partial RU-rationality (standard RUM consistency on menus that do not contain aggregated categories). We also characterize the admissible behaviors of consumers under RU-rationality and show that the characterization includes various irrational behaviors under ARU-rationality.

These results imply that RU-rationality alone is too weak to justify ARUMs (distributions over rankings on aggregated categories) commonly assumed in empirical work. Given this, we identify two conditions under which RU-rationality implies ARU-rationality: (i) *menu-independence of* λ and (ii) a *non-overlapping preference structure*. The following figure summarizes our theoretical results:

$$\text{ARU} \quad \underset{\substack{+\text{ menu independent } \lambda \text{ or} \\ \text{non-overlapping } \mu_{\mathscr{X}}}}{\overset{\text{Remark 2.7}}{\rightleftarrows}} \quad \text{RU} \quad \underset{+\text{ Limited Monotonicity}}{\overset{\text{By definition}}{\rightleftarrows}} \quad \text{Partial RU.}$$

To assess empirical relevance, we conduct simulations. We first generate "true" logit data $\rho_{\mathscr{X}}$ on the underlying alternatives using the true utilities. Then, fixing a composition distribution λ, we construct the observable reduced data ρ on categories. From ρ, we estimate utilities and evaluate (i) the bias $(\hat{u}(x) - \hat{u}(y)) - (u(x) - u(y))$ and (ii) the distance from

ρ to the ARU polytope $\mathcal{P}_{\mathscr{A}}$. The results are stark: greater menu dependence and greater preference overlap both increase bias and distance. Bias magnitudes can be large enough to flip the inferred ranking of x and y, showing how misspecified aggregation distorts both utility levels and orderings.

References

Aguiar, V. H., M. J. Boccardi, N. Kashaev, and J. Kim (2023): "Random utility and limited consideration," *Quantitative Economics*, 14(1), 71–116.

Apesteguia, J., and M. A. Ballester (2016): "Stochastic Representative Agent," Discussion paper.

Azrieli, Y., and J. Rehbeck (2022): "Marginal stochastic choice," Papers 2208.08492, arXiv.org.

Blumenthal, S. (1968): "Multinomial Sampling with Partially Categorized Data," *Journal of the American Statistical Association*, 63(322), 542–551.

Brady, R. L., and J. Rehbeck (2016): "Menu-Dependent Stochastic Feasibility," *Econometrica*, 84(3), 1203–1223.

Brownstone, D., and P. Li (2017): "A Model for Broad Choice Data," .

Cattaneo, M. D., X. Ma, Y. Masatlioglu, and E. Suleymanov (2020): "A Random Attention Model," *Journal of Political Economy*, 128(7), 2796–2836.

Gul, F., and W. Pesendorfer (2001): "Temptation and Self-Control," *Econometrica*, 69(6), 1403–1435.

Heitjan, D. F. (1989): "Inference from Grouped Continuous Data: A Review," *Statistical Science*, 4(2), 164 – 179.

Heitjan, D. F., and D. B. Rubin (1990): "Inference from Coarse Data Via Multiple Imputation with Application to Age Heaping," *Journal of the American Statistical Association*, 85(410), 304–314.

——— (1991): "Ignorability and Coarse Data," *The Annals of Statistics*, 19(4), 2244 – 2253.

Horan, S. (2019): "Random consideration and choice: A case study of "default" options," *Mathematical Social Sciences*, 102, 73–84.

Huang, D., and C. Rojas (2013): "The Outside Good Bias in Logit Models of Demand with Aggregate Data," *Economics Bulletin*, 33(1), 198–206.

Kono, H., K. Saito, and A. Sandroni (2025): "Random utility model with unobservable alternatives," *working paper*.

Kreps, D. M. (1979): "A Representation Theorem for Preference for Flexibility," *Econometrica*, 47(3), 565–577.

Manzini, P., and M. Mariotti (2014): "Stochastic Choice and Consideration Sets," *Econometrica*, 82(3), 1153–1176.

Nevo, A. (2001): "Measuring market power in the ready-to-eat cereal industry," *Econometrica*, 69(2), 307–342.

Nordheim, E. V. (1984): "Inference from Nonrandomly Missing Categorical Data: An Example from a Genetic Study on Turner's Syndrome," *Journal of the American Statistical Association*, 79(388), 772–780.

Shum, M. (2016): *Econometric Models for Industrial Organization*. World Scientific, Singapore, 2 edn.

Stafford, T. M. (2018): "Accounting for outside options in discrete choice models: An application to commercial fishing effort," *Journal of Environmental Economics and Management*, 88(C), 159–179.

Strzalecki, T. (2022): *Stochastic Choice Theory*. working paper.

Zhang, L. (2023): "Identification and estimation of market size in discrete choice demand models," *Available at SSRN 4545848*.

A Proofs

For ease of notation, we will assume throughout this section that each category $a \in \mathscr{A}_N$ is composed of itself, that is $X(a)$ contains only a. Thus by abuse of notation, we may identify the non-aggregated categories $\mathscr{A}_N$ and the corresponding set of underlying alternatives $\cup_{a \in \mathscr{A}_N} X(a)$.

A.1 Proof of Theorem 3.1

The necessity is trivial. We prove the sufficiency. Assume the two conditions and fix an aggregation function X such that $|X(a_0)| = |\mathscr{A}_N| + 1$.

By Partial-RU, there exists $\tilde{\mu} \in \Delta(\mathscr{L}(\mathscr{A}_N))$ that rationalizes ρ on $2^{\mathscr{A}_N} \setminus \{\emptyset\}$. For each $y \in \mathscr{A}_N$, choose an element of $X(a_0)$ denoted by $x_0(y)$. We will also fix a unobservable alternative $\underline{x}$ that plays a special role in the following. Note that such alternatives exist because $|X(a_0)| = |\mathscr{A}_N| + 1$.

For any ranking $\tilde{\succ}$ on $\mathscr{A}_N$ that belongs to the support of $\tilde{\mu}$, we extend $\tilde{\succ}$ and define a ranking $\succ$ on $\mathscr{X}$ as follows. First, for each $y \in \mathscr{A}_N$, place $x_0(y) \in X(a_0)$ just above y and $\underline{x}$ at the bottom for all $\tilde{\succ} \in supp(\tilde{\mu})$. For example, if the ranking $y_1 \tilde{\succ} y_2 \tilde{\succ} y_3$ is in the support then we extend $\tilde{\succ}$ to $\succ$ on $\mathscr{X}$ as follows

$$x_0(y_1) \succ y_1 \succ x_0(y_2) \succ y_2 \succ x_0(y_3) \succ y_3 \succ \underline{x}.$$

Finally, we define $\mu_{\mathscr{X}} \in \Delta(\mathscr{L}(\mathscr{X}))$ as follows: for each extended ranking $\succ$ defined above, we let $\mu_{\mathscr{X}}(\succ) = \tilde{\mu}(\tilde{\succ})$.

Let $\rho_{\mathscr{X}}$ be a random utility model that is rationalized by $\mu_{\mathscr{X}}$. Now fix $D \subseteq \mathscr{A}_N$. In the following, we will construct $\lambda \in \prod_{b \in D \cup a_0}(2^{X(b)} \setminus \{\emptyset\})$ to establish the following equality for all $b \in D \cup a_0$:

$$\sum_{(S_b)_{b \in D \cup a_0} \subseteq \prod_{b \in D \cup a_0}(2^{X(b)} \setminus \{\emptyset\})} \lambda((S_b)_{b \in D \cup a_0}) \rho_{\mathscr{X}}(\cup_{b \in D \cup a_0} S_b, S_b) = \rho(D \cup a_0, b).$$

Since $D \subseteq \tilde{X}$, for all $b \in D$, there exists y_b such that $X(b) = \{y_b\}$.

Thus, for any $(S_b)_{b \in D \cup a_0} \subseteq \prod_{b \in D \cup a_0}(2^{X(b)} \setminus \{\emptyset\})$, S_b is determined as $\{y_b\}$ for all $b \in D$. Thus, for simplicity, we will focus on the marginal distribution of λ on $2^{X(a_0)} \setminus \{\emptyset\}$ and obtain the following equality:

$$\sum_{S \subseteq X(a_0)} \lambda(S) \rho_{\mathscr{X}}(\{y_b | b \in D\} \cup S, y_b) = \rho(D \cup a_0, b) \quad \text{for all } b \in D \tag{12}$$

and

$$\sum_{S \subseteq X(a_0)} \lambda(S) \rho_{\mathscr{X}}(\{y_b | b \in D\} \cup S, S) = \rho(D \cup a_0, a_0). \tag{13}$$

Note that (13) is implied by (14) by the fact that the probability must sum to one. The purpose of our proof is, thus, to show the existence of λ on $2^{X(a_0)} \setminus \{\emptyset\}$ that satisfies the following equality for all $y \in D \subseteq \mathscr{A}_N$,[16]

$$\sum_{S \subseteq X(a_0)} \lambda(S) \rho_{\mathscr{X}}(D \cup S, y) = \rho(D \cup a_0, y). \tag{14}$$

Now we will define λ recursively. We first label elements of D by y_i so that $\frac{\rho(D \cup a_0, y_0)}{\rho(D, y_0)}$ is

[16]Remember that this λ is a composition distribution for the choice set $D \cup \{a_0\}$. For another choice set $E \cup \{a_0\}$, one can construct a desirable composition distribution in the same way.

maximum among $\frac{\rho(D\cup a_0,y_i)}{\rho(D,y_i)}$. Define

$$\lambda_0(\underline{x}) = \frac{\rho(D \cup a_0, y_0)}{\rho(D, y_0)}, \quad \lambda_0(X(a_0)) = 1 - \lambda_0(\underline{x})$$

and all other values are zero. Then λ_0 is a probability measure because $\lambda_0(\underline{x}) \leq 1$ by Limited Monotonicity. Moreover, the chance that y_0 is chosen is equal to $\rho(D \cup a_0, y_0)$ as desired:

$$\begin{aligned}
\sum_{S \subseteq X(a_0)} \lambda_0(S)\rho_{\mathscr{X}}(D \cup S, y_0) &= \lambda_0(\underline{x})\rho_{\mathscr{X}}(D \cup \{\underline{x}\}, y_0) + \lambda_0(X(a_0))\rho_{\mathscr{X}}(D \cup X(a_0), y_0) \\
&= \lambda_0(\underline{x})\rho_{\mathscr{X}}(D \cup \{\underline{x}\}, y_0) \quad (\because x(y_0) \succ y_0 \text{ for all } \succ \in \mathit{supp}(\mu_{\mathscr{X}})) \\
&= \lambda_0(\underline{x})\rho_{\mathscr{X}}(D, y_0) \qquad\qquad (\because y_0 \succ \underline{x} \text{ for all } \succ \in \mathit{supp}(\mu_{\mathscr{X}})) \\
&= \lambda_0(\underline{x})\rho(D, y_0) \\
&= \rho(D \cup a_0, y_0),
\end{aligned}$$

where the fourth equality holds because $\rho_{\mathscr{X}}(D, y_0) = \rho(D, y_0)$ because all elements in D are non-aggregated categories.

Given λ_0, we will define λ_1 that satisfies (14) for y_0 and y_1. First remember that since $\frac{\rho(D\cup a_0,y_0)}{\rho(D,y_0)}$ is maximum among $\frac{\rho(D\cup a_0,y_i)}{\rho(D,y_i)}$, we have

$$\lambda_0(\underline{x}) \equiv \frac{\rho(D \cup a_0, y_0)}{\rho(D, y_0)} \geq \frac{\rho(D \cup a_0, y_i)}{\rho(D, y_i)} = \frac{\rho(D \cup a_0, y_i)}{\rho_{\mathscr{X}}(D, y_i)} = \frac{\rho(D \cup a_0, y_i)}{\rho_{\mathscr{X}}(D \cup \{\underline{x}\}, y_i)},$$

where the second to the last equality holds because $\rho_{\mathscr{X}}(D, y_i) = \rho(D, y_i)$; and the last equality holds because $\underline{x}$ is the worst alternative for any $\succ \in \mathit{supp}(\mu_{\mathscr{X}})$. Thus $\lambda_0(\underline{x})\rho_{\mathscr{X}}(D \cup \{\underline{x}\}, y_i) \geq \rho(D \cup a_0, y_i)$. By using this inequality, we have

$$\begin{aligned}
\sum_{S \subseteq X(a_0)} \lambda_0(S)\rho_{\mathscr{X}}(D \cup S, y_i) &= \lambda_0(\underline{x})\rho_{\mathscr{X}}(D \cup \{\underline{x}\}, y_i) + \lambda_0(X(a_0))\rho_{\mathscr{X}}(D \cup X(a_0), y_i) \\
&= \lambda_0(\underline{x})\rho_{\mathscr{X}}(D \cup \{\underline{x}\}, y_i) \\
&\geq \rho(D \cup a_0, y_i).
\end{aligned}$$

This inequality suggests that we should decrease the value of $\lambda_0(\underline{x})$ to have the desired equality for $i > 0$. Given the observation, we define λ_1 given λ_0 as follows. The idea is to move probability from $\lambda_0(\underline{x})$ to $\lambda_0(\underline{x}, x_0(y_1))$ so that the chance that y_1 is chosen decreases to $\rho(D \cup a_0, y_1)$ as desired. More precisely, we define λ_1 as follows:

$$\lambda_1(\underline{x}) = \frac{\rho(D \cup a_0, y_1)}{\rho(D, y_1)}, \quad \lambda_1(\underline{x}, x_0(y_1)) = \lambda_0(\underline{x}) - \lambda_1(\underline{x}), \quad \lambda_1(X(a_0)) = \lambda_0(X(a_0)).$$

To see that λ_1 is a probability measure, notice that $\lambda_1(\underline{x}, x_0(y_1)) \geq 0$ because $\frac{\rho(D \cup a_0, y_0)}{\rho(D, y_0)} \geq \frac{\rho(D \cup a_0, y_1)}{\rho(D, y_1)}$. Moreover, we have the desired equalities for y_1 and y_0 as follows:

$$\begin{aligned}
&\sum_{S \subseteq X(a_0)} \lambda_1(S) \rho_{\mathscr{X}}(D \cup S, y_1) \\
&= \lambda_1(\underline{x}) \rho_{\mathscr{X}}(D \cup \{\underline{x}\}, y_1) + \lambda_1(\underline{x}, x_0(y_1)) \rho_{\mathscr{X}}(D \cup \{\underline{x}, x_0(y_1)\}, y_1) + \lambda_1(X(a_0)) \rho_{\mathscr{X}}(D \cup X(a_0), y_1) \\
&= \lambda_1(\underline{x}) \rho_{\mathscr{X}}(D \cup \{\underline{x}\}, y_1) \quad (\because x_0(y_1) \succ y_1 \text{ for all } \succ \in \mathit{supp}(\mu)) \\
&= \lambda_1(\underline{x}) \rho_{\mathscr{X}}(D, y_1) \quad (\because y_1 \succ \underline{x} \text{ for all } \succ \in \mathit{supp}(\mu)) \\
&= \lambda_1(\underline{x}) \rho(D, y_1) \\
&= \rho(D \cup a_0, y_1)
\end{aligned}$$

and

$$\begin{aligned}
&\sum_{S \subseteq X(a_0)} \lambda_1(S) \rho_{\mathscr{X}}(D \cup S, y_0) \\
&= \lambda_1(\underline{x}) \rho_{\mathscr{X}}(D \cup \{\underline{x}\}, y_0) + \lambda_1(\underline{x}, x_0(y_1)) \rho_{\mathscr{X}}(D \cup \{\underline{x}, x_0(y_1)\}, y_0) + \lambda_1(X(a_0)) \rho_{\mathscr{X}}(D \cup X(a_0), y_0) \\
&= \big(\lambda_1(\underline{x}) + \lambda_1(\underline{x}, x_0(y_1))\big) \rho_{\mathscr{X}}(D \cup \{\underline{x}\}, y_0) + \lambda_1(X(a_0)) \rho_{\mathscr{X}}(D \cup X(a_0), y_0) \qquad (15) \\
&= \lambda_0(\underline{x}) \rho_{\mathscr{X}}(D \cup \{\underline{x}\}, y_0) + \lambda_0(X(a_0)) \rho_{\mathscr{X}}(D \cup X(a_0), y_0) \\
&= \lambda_0(\underline{x}) \rho_{\mathscr{X}}(D \cup \{\underline{x}\}, y_0) \\
&= \rho(D \cup a_0, y_0), \quad (\because \text{Definition of } \lambda_0)
\end{aligned}$$

where the second equality holds because for all $\succ \in \mathit{supp}(\mu_{\mathscr{X}})$, $x_0(y_1)$ is the immediate predecessor of y_1 so adding $x_0(y_1)$ changes the only the probability of choice frequency of y_0, thus $\rho_{\mathscr{X}}(D \cup \{\underline{x}\}, y_0) = \rho_{\mathscr{X}}(D \cup \{\underline{x}, x_0(y_1)\}, y_0)$.

In general, at nth step, given λ_{n-1} that satisfy (14) for all $y \in \{y_1, \ldots, y_{n-1}\}$, we will define λ_n that satisfy (14) for all $y \in \{y_1, \ldots, y_{n-1}, y_n\}$. First define

$$c_n = \frac{\rho(D \cup a_0, y_n)}{\rho(D, y_n) \lambda_0(\underline{x})}.$$

Note $0 \leq c_n \leq 1$ because $\frac{\rho(D \cup a_0, y_0)}{\rho(D, y_0)}$ is maximum among $\frac{\rho(D \cup a_0, y_i)}{\rho(D, y_i)}$. Let $\lambda_n(X(a_0)) = \lambda_0(X(a_0))$

and for each S in the support of λ_{n-1} except $X(a_0)$

$$\lambda_n(S) = c_n \lambda_{n-1}(S), \quad \lambda_n(S \cup x_0(y_n)) = \lambda_{n-1}(S) - \lambda_n(S).$$

Note first that λ_n is a probability measure given that λ_{n-1} is a probability measure. Note also this modification of λ_n from λ_{n-1} does not change the chance that y_i is chosen for any $i \leq n-1$; this only changes the chance that y_n is chosen because adding $x_0(y_n)$ changes only the choice frequency of y_n as in (15). Moreover, we obtain the desired equalities for the choice frequency of y_n as follows:

$$\begin{aligned}
&\sum_{S \subseteq X(a_0)} \lambda_n(S) \rho_{\mathscr{X}}(D \cup S, y_n) \\
&= \lambda_{n-1}(X(a_0)) \rho_{\mathscr{X}}(D \cup X(a_0), y_n) + \\
&\sum_{S \in supp(\lambda_{n-1}) \setminus \{X(a_0)\}} c_n \lambda_{n-1}(S) \rho_{\mathscr{X}}(D \cup S, y_n) + (1 - c_n) \lambda_{n-1}(S) \rho_{\mathscr{X}}(D \cup S \cup x_0(y_n), y_n) \\
&= \sum_{S \in supp(\lambda_{n-1}) \setminus \{X(a_0)\}} c_n \lambda_{n-1}(S) \rho_{\mathscr{X}}(D \cup S, y_n) \quad (\because x_0(y_n) \succ y_n \text{ for all } \succ \in supp(\mu_{\mathscr{X}})) \\
&= \sum_{S \in supp(\lambda_{n-1}) \setminus \{X(a_0)\}} c_n \lambda_{n-1}(S) \rho_{\mathscr{X}}(D, y_n) \\
&= c_n \sum_{S \in supp(\lambda_{n-1}) \setminus \{X(a_0)\}} \lambda_{n-1}(S) \rho(D, y_n) \quad (\because \rho(D, y_n) = \rho_{\mathscr{X}}(D, y_n)) \\
&= c_n \lambda_0(\underline{x}) \rho(D, y_n) \quad (\because \sum_{S \in supp(\lambda_{n-1}) \setminus \{X(a_0)\}} \lambda_{n-1}(S) = 1 - \lambda_{n-1}(X(a_0)) = 1 - \lambda_0(X(a_0)) = \lambda_0(\{\underline{x}\})) \\
&= \rho(D \cup a_0, y_n),
\end{aligned}$$

where the last equality holds by the definition of c_n; the fourth to the last equality holds because y_n is the maximum element among D with respect to $\succ$ if and only if y_n is the maximum element among $D \cup S$ with respect to $\succ$ since $S \subseteq \{x_0(y_i) | i < n\} \cup \{\underline{x}\}$ and $x_0(y_i)$ is the immediate predecessor of y_i for all $\succ \in supp(\mu_{\mathscr{X}})$.

A.2 Proof of Theorem 3.12

The necessity is trivial. We show the sufficiency. Fix an aggregation function X such that $|X(a)| = |\mathscr{A}_N| + 2$ for all $a \in \mathscr{A}_A$. For each $a \in \mathscr{A}_A$ and $y \in \mathscr{A}_N$, choose an element of $X(a)$, denoted by $x_a(y)$. We also fix two alternatives $\overline{x}_a, \underline{x}_a \in X(a)$. Note that there exist such alternatives under the assumption that $|X(a)| = |\mathscr{A}_N| + 2$ for all $a \in \mathscr{A}_A$. We again abuse notation and assume that each category $a \in \mathscr{A}_N$ is composed of itself, that is $X(a) = \{a\}$.

By Partial RU- rationality, there exists a distribution $\tilde{\mu}$ over linear orders $\tilde{\succ}$ on $\mathscr{A}_N$ that rationalizes ρ on subsets of $\mathscr{A}_N$. Based on this distribution, we will construct a RUM ρ on the entire set $\mathscr{X}$ of the underlying alternatives as follows. For each ranking on $\mathscr{A}_N$ in the support of $\tilde{\mu}$, we construct ranking on $\mathscr{X}$: we place $x_a(y)$ right above y, $\overline{x}_a$ and $\underline{x}_a$ at the bottom with $\overline{x}_a$s above $\underline{x}_a$s. The order among aggregated categories is arbitrarily fixed. For example, let $\{a_0, a_1\} = \mathscr{A}_A$ and fix an order by index. Also let $\mathscr{A}_N = \{y_0, y_1\}$ and $y_0 \tilde{\succ} y_1$. Then, we define $\succ$ on $\mathscr{X}$ as follows:

$$x_0(y_0) \succ x_1(y_0) \succ y_0 \succ x_0(y_1) \succ x_1(y_1) \succ y_1 \succ \overline{x}_0 \succ \overline{x}_1 \succ \underline{x}_0 \succ \underline{x}_1.$$

Fix $D \subseteq \mathscr{A}_N$ and $E \subseteq \mathscr{A}_A$. In the following, we will construct $\lambda \in \prod_{b \in D \cup E}(2^{X(b)} \setminus \{\emptyset\})$ to establish the following equality for all $b \in D \cup E$:

$$\sum_{(S_b)_{b \in D \cup E} \subseteq \prod_{b \in D \cup E}(2^{X(b)} \setminus \{\emptyset\})} \lambda((S_b)_{b \in D \cup E}) \rho_{\mathscr{X}}(\cup_{b \in D \cup E} S_b, S_b) = \rho(D \cup E, b).$$

With using the same simplified notation and the reasoning in the proof of Theorem 3.1, the purpose of the proof is to show the existence of a probability distribution λ on $\prod_{b \in E}(2^{X(b)} \setminus \{\emptyset\})$ that satisfies

$$\sum_{(S_b)_{b \in E} \in \prod_{b \in E}(2^{X(b)} \setminus \{\emptyset\})} \lambda((S_b)_{b \in E}) \rho_{\mathscr{X}}(D \cup (\cup_{b \in E} S_b), y) = \rho(D \cup E, y) \text{ for all } y \in D, \tag{16}$$

and

$$\sum_{(S_b)_{b \in E} \in \prod_{b \in E}(2^{X(b)} \setminus \{\emptyset\})} \lambda((S_b)_{b \in E}) \rho_{\mathscr{X}}(D \cup (\cup_{b \in E} S_b), S_b) = \rho(D \cup E, b) \text{ for all } b \in E. \tag{17}$$

The proof consists of two steps. At the first step, for each $a \in E$ we will construct a λ^a in the same way as the single outside option case (i.e., Theorem 3.1) except replace $\underline{x}$ with $\overline{x}_a$ and add $\underline{x}_a$ to all the sets. Given λ^a, in the second step, by taking a convex combination of λ^a, we will obtain the desired λ that satisfy (16) and (17) for the given D and E.

Step 1: We first label elements of D by y_i so that $\frac{\rho(D \cup E, y_0)}{\rho(D, y_0)}$ is maximum among $\frac{\rho(D \cup E, y_i)}{\rho(D, y_i)}$. We will construct a probability distribution λ^a on $\prod_{b \in E}(2^{X(b)} \setminus \{\emptyset\})$ recursively. Initially, we define λ_0^a by

$$\lambda_0^a(\overline{x}_a, (\underline{x})_{E \setminus a}) = \frac{\rho(D \cup E, y_0)}{\rho(D, y_0)}, \quad \lambda_0^a((X(b))_{b \in E}) = 1 - \lambda_0^a(\overline{x}_a, (\underline{x})_{E \setminus a}) \tag{18}$$

and all other lambdas are zero. By Limited Monotonicity, we have $\lambda_0^a(\overline{x}_a, (\underline{x})_{E\setminus a}) \leq 1$, thus λ_0^a is a probability measure. Moreover, the probability that y_0 is chosen equals $\rho(D \cup E, y_0)$ as desired:

$$
\begin{aligned}
&\sum_{(S_b)_{b\in E}\in\prod_{b\in E}(2^{X(b)}\setminus\{\emptyset\})} \lambda_0^a((S_b)_{b\in E})\rho_{\mathscr{X}}(D\cup(\cup_{b\in E}S_b), y_0)\\
&= \lambda_0^a(\overline{x}_a, (\underline{x})_{E\setminus a})\rho_{\mathscr{X}}(D\cup\{\overline{x}_a, (\underline{x})_{E\setminus a}\}, y_0) + \lambda_0^a((X(b))_{b\in E})\rho_{\mathscr{X}}(D\cup(\cup_{b\in E}X(b)), y_0)\\
&= \lambda_0^a(\overline{x}_a, (\underline{x})_{E\setminus a})\rho_{\mathscr{X}}(D\cup\{\overline{x}_a, (\underline{x})_{E\setminus a}\}, y_0) \qquad (\because x_a(y_0)\succ y_0 \text{ for all } \succ\in supp(\mu_{\mathscr{X}}))\\
&= \lambda_0^a(\overline{x}_a, (\underline{x}_a)_{E\setminus a})\rho(D, y_0) \qquad (\because y_0\succ \overline{x}_a, \underline{x}_b \text{ for all } b\in E\setminus a \text{ and } \succ\in supp(\mu_{\mathscr{X}}))\\
&= \rho(D\cup E, y_0).
\end{aligned}
$$

Next we define λ_1^a based on λ_0^a to satisfy the equality for $\rho(D \cup E, y_1)$. In general at the nth step to obtain the desired equality for $\rho(D \cup E, y_n)$, let

$$
c_n = \frac{\rho(D\cup E, y_n)}{\rho(D, y_n)\lambda_0^a(\overline{x}_a, (\underline{x})_{E\setminus a})}.
$$

Note $0 \leq c_n \leq 1$ by our choice of y_0. We define λ_n^a by $\lambda_n^a((X(b))_{b\in E}) = \lambda_0^a((X(b))_{b\in E})$ and for all $(S_b)_{b\in E} \in supp(\lambda_{n-1}^a) \setminus (X(b))_{b\in E}$

$$
\lambda_n^a((S_b)_{b\in E}) = c_n\lambda_{n-1}^a((S_b)_{b\in E}), \quad \lambda_n^a(S_a\cup x_a(y_n)\ ,\ (S_b)_{b\in E\setminus a}) = (1-c_n)\lambda_{n-1}^a((S_b)_{b\in E}).
$$

Note first that λ_n^a is a probability measure given that λ_{n-1}^a is a probability measure. Note also this modification of λ_n^a from λ_{n-1}^a does not change the chance that y_i is chosen for any $i \leq n-1$; this only changes the chance that y_n is chosen because adding $x_a(y_n)$ changes only the choice frequency of y_n. Moreover, we obtain the desired equalities for the choice frequency of y_n as follows:

$$
\begin{aligned}
&\sum_{(S_b)_{b\in E}\in\prod_{b\in E}(2^{X(b)}\setminus\{\emptyset\})} \lambda^a_n((S_b)_{b\in E})\rho_{\mathscr{X}}(D\cup(\cup_{b\in E}S_b),y_n)\\
&=\lambda^a_0((X(b))_{b\in E})\rho_{\mathscr{X}}(D\cup(\cup_{b\in E}X(b)),y_n)\\
&+\sum_{(S_b)_{b\in E}\in supp(\lambda^a_{n-1})\setminus(X(b))_{b\in E}}\Big[c_n\lambda^a_{n-1}((S_b)_{b\in E})\rho_{\mathscr{X}}(D\cup(\cup_{b\in E}S_b),y_n)\\
&\qquad\qquad+(1-c_n)\lambda^a_{n-1}((S_b)_{b\in E})\rho_{\mathscr{X}}(D\cup(\cup_{b\in E}S_b)\cup x_a(y_n),y_n)\Big]\\
&=\sum_{(S_b)_{b\in E}\in supp(\lambda^a_{n-1})\setminus(X(b))_{b\in E}}c_n\lambda^a_{n-1}((S_b)_{b\in E})\rho_{\mathscr{X}}(D\cup(\cup_{b\in E}S_b),y_n)\\
&=\sum_{(S_b)_{b\in E}\in supp(\lambda^a_{n-1})\setminus(X(b))_{b\in E}}c_n\lambda^a_{n-1}((S_b)_{b\in E})\rho(D,y_n)\\
&=c_n\lambda^a_0(\overline{x}_a,(\underline{x})_{E\setminus a})\rho(D,y_n)\\
&=\rho(D\cup E,y_n),
\end{aligned}
$$

where the second equality holds because $x_a(y_n)\succ y_n$ for all $\succ\in supp(\mu_{\mathscr{X}})$; the second to the last equality holds because $\sum_{(S_b)_{b\in E}\in supp(\lambda^a_{n-1})\setminus(X(b))_{b\in E}}\lambda^a_{n-1}((S_b)_{b\in E})=1-\lambda^a_{n-1}((X(b))_{b\in E})=1-\lambda^a_0((X(b))_{b\in E})=\lambda^a_0(\overline{x}_a,(\underline{x})_{E\setminus a})$; and finally the third to the last equality holds because y_n is a maximum element among D with respect to $\succ$ if and only if y_n is a maximum element among $D\cup(\cup_{b\in E}S_b)$ with respect to $\succ$ (Note hat this equivalence holds because $S_b\subseteq\{x_b(y_i)|i<n\}\cup\{\underline{x}_b,\overline{x}_b|b\in\mathscr{A}_A\}$ and $x_b(y_i)$ is the immediate predecessor of y_i for all $\succ\in supp(\mu_{\mathscr{X}})$). It follows that λ^a satisfies (16) as desired: for all $y\in D$

$$\sum_{(S_b)_{b\in E}\in\prod_{b\in E}(2^{X(b)}\setminus\{\emptyset\})}\lambda^a((S_b)_{b\in E})\rho_{\mathscr{X}}(D\cup(\cup_{b\in E}S_b),y)=\rho(D\cup E,y). \tag{19}$$

Step 2: In the following, given $(\lambda^a)_{a\in\mathscr{A}_A}$, we will construct λ that satisfies (17) as well as (16). First observe that for all $a\in\mathscr{A}_A$,

$$\sum_{(S_b)_{b\in E}\in\prod_{b\in E}(2^{X(b)}\setminus\{\emptyset\})}\lambda^a((S_b)_{b\in E})\rho_{\mathscr{X}}(D\cup(\cup_{b\in E}S_b),S_b)=\begin{cases}1-\sum_{y\in D}\rho(D\cup E,y) & \text{if } b=a,\\ 0 & \text{if } b\neq a.\end{cases} \tag{20}$$

We denote the left hand side of the above equation as $\rho^a(D\cup E,b)$. Define

$$\lambda=\sum_{a\in E}\lambda^a\left(\frac{\rho(D\cup E,a)}{1-\sum_{y\in D}\rho(D\cup E,y)}\right).$$

Since $\sum_{a\in E}\rho(D\cup E,a)=1-\sum_{y\in D}\rho(D\cup E,y)$ and each λ^a is a probability distribution on $\prod_{a\in E}2^{X(a)}\setminus\{\emptyset\}$, so is λ. Moreover, we have the desired equalities for choice frequency of each aggregated category $b\in E$ as follows:

$$
\begin{aligned}
&\sum_{(S_b)_{b\in E}\in\prod_{b\in E}(2^{X(b)}\setminus\{\emptyset\})}\lambda((S_b)_{b\in E})\rho_{\mathscr{X}}(D\cup(\cup_{b\in E}S_b),S_b)\\
&=\sum_{(S_b)_{b\in E}\in\prod_{b\in E}(2^{X(b)}\setminus\{\emptyset\})}\left(\sum_{a\in E}\lambda^a((S_b)_{b\in E})\frac{\rho(D\cup E,a)}{1-\sum_{y\in D}\rho(D\cup E,y)}\right)\rho_{\mathscr{X}}(D\cup(\cup_{b\in E}S_b),S_b)\\
&=\sum_{a\in E}\left(\frac{\rho(D\cup E,a)}{1-\sum_{y\in D}\rho(D\cup E,y)}\right)\sum_{(S_b)_{b\in E}\in\prod_{b\in E}(2^{X(b)}\setminus\{\emptyset\})}\lambda^a((S_b)_{b\in E})\rho_{\mathscr{X}}(D\cup(\cup_{b\in E}S_b),S_b)\\
&=\sum_{a\in E}\left(\frac{\rho(D\cup E,a)}{1-\sum_{y\in D}\rho(D\cup E,y)}\right)\rho^a(D\cup E,b)\quad(\because \text{Definition of } \rho^a(D\cup E,b))\\
&=\left(\frac{\rho(D\cup E,b)}{1-\sum_{y\in D}\rho(D\cup E,y)}\right)\rho^b(D\cup E,b)+\left(\sum_{a\in E\setminus b}\frac{\rho(D\cup E,a)}{1-\sum_{y\in D}\rho(D\cup E,y)}\right)\rho^a(D\cup E,b)\\
&=\rho(D\cup E,b).\qquad(\because (20))
\end{aligned}
$$

Thus, λ satisfies (17). Finally, to show that λ satisfies (16), choose $z\in D$. Then, we have

$$
\begin{aligned}
&\sum_{(S_b)_{b\in E}\in\prod_{b\in E}(2^{X(b)}\setminus\{\emptyset\})}\lambda((S_b)_{b\in E})\rho_{\mathscr{X}}(D\cup(\cup_{b\in E}S_b),z)\\
&=\sum_{(S_b)_{b\in E}\in\prod_{b\in E}(2^{X(b)}\setminus\{\emptyset\})}\left(\sum_{a\in E}\lambda^a((S_b)_{b\in E})\frac{\rho(D\cup E,a)}{1-\sum_{y\in D}\rho(D\cup E,y)}\right)\rho_{\mathscr{X}}(D\cup(\cup_{b\in E}S_b),z)\\
&=\sum_{a\in E}\left(\frac{\rho(D\cup E,a)}{1-\sum_{y\in D}\rho(D\cup E,y)}\right)\sum_{(S_b)_{b\in E}\in\prod_{b\in E}(2^{X(b)}\setminus\{\emptyset\})}\lambda^a((S_b)_{b\in E})\rho_{\mathscr{X}}(D\cup(\cup_{b\in E}S_b),z)\\
&=\sum_{a\in E}\left(\frac{\rho(D\cup E,a)}{1-\sum_{y\in D}\rho(D\cup E,y)}\right)\rho(D\cup E,z)\quad(\because (19))\\
&=\rho(D\cup E,z),
\end{aligned}
$$

where the last equality holds because $\sum_{a\in E}\rho(D\cup E,a)=1-\sum_{y\in D}\rho(D\cup E,y)$.

A.3 Proof of Theorem 3.6 and Remark 3.10

A.3.1 Proof of Theorem 3.6

First we prove that the set of RU rational choice functions coincides with co.$\{\rho^{\succ}_{\mathscr{E}} \mid \succ \in \mathscr{L}(\mathscr{A})$ and $\mathscr{E} \subseteq \mathscr{D}\}$.

Fix $\succ \in \mathscr{L}(\mathscr{A}_N \cup X(a_0))$. For each $A \in \mathscr{D}$, fix $(S_a)_{a\in A} \in \prod_{a\in A}(2^{X(a)} \setminus \{\emptyset\})$. For each collection $((S_a)_{a\in A})_{A\in\mathscr{D}}$, we define $\rho^{\succ}_{((S_a)_{a\in A})_{A\in\mathscr{D}}}$ as follows. To make the notation simple, we write $((S_a)_{a\in A})_{A\in\mathscr{D}}$ as $(S_A)_{A\in\mathscr{D}}$. For each $b \in B \in \mathscr{D}$,

$$\rho^{\succ}_{(S_A)_{A\in\mathscr{D}}}(B, b) = 1\Big(\exists x \in S_b \forall y \in \cup_{a\in B\setminus b} S_a [x \succ y]\Big).$$

Step 1: For each $\succ \in \mathscr{L}(\mathscr{A}_N \cup X(a_0))$ and each collection $(S_A)_{A\in\mathscr{D}}$, there exists $\succ' \in \mathscr{L}(\mathscr{A})$ and $\mathscr{E} \subseteq \mathscr{D}$ such that $\rho^{\succ}_{(S_A)_{A\in\mathscr{D}}} = \rho^{\succ'}_{\mathscr{E}}$.

Proof. Fix each $\succ \in \mathscr{L}(\mathscr{A}_N \cup X(a_0))$ and each collection $(S_A)_{A\in\mathscr{D}}$, define

$$\mathcal{F} = \left\{ D \in \mathscr{D} \;\middle|\; \forall y \in \bigcup_{b\in D\setminus a_0} X(b) \; [x^* \succ y] \right\},$$

where $x^* = \max_{S_{a_0}} \succ$. That is, $\mathcal{F}$ consists of all choice sets D such that the $\succ$-best element lies in S_{a_0} and is preferred to every element contained in $\bigcup_{b\in D\setminus a_0} X(b)$. Define $\mathscr{E} = \mathscr{D} \setminus \mathcal{F}$.

Define $\succ' \in \mathscr{L}(\mathscr{A}_N)$ such that $a \succ' b$ if and only if $x \succ y$ for any $a, b \in \mathscr{A}_N$, where $\{x\} = X(a)$ and $\{y\} = X(b)$. For any $D \in \mathscr{D}$, if $D \in \mathscr{E}^c \equiv \mathcal{F}$, the maximum elements of $\succ$ over D belongs to $X(a_0)$, thus $\rho^{\succ}_{(S_A)_{A\in\mathscr{D}}}(D, \cdot) = \rho^{\succ'}_{\mathscr{E}}(D, \cdot)$. Moreover, if $D \in \mathscr{E}$, the maximum elements of $\succ$ and $\succ'$ over D coincide; thus $\rho^{\succ}_{(S_A)_{A\in\mathscr{D}}}(D, \cdot) = \rho^{\succ'}_{\mathscr{E}}(D, \cdot)$. □

Given Step 1, it suffices to show that any RU-rational ρ can be expressed as a convex combination of terms of the form $\rho^{\succ}_{(S_A)_{A\in\mathscr{D}}}$. Assume that ρ is rationalized by $(\mu_{\mathscr{X}}, \lambda)$. We first establish this in the degenerate case where the rationalizing measure $\mu_{\mathscr{X}}$ is a point mass in Step 2.[17] In particular, we show that ρ can be represented as a convex combination of $\rho^{\succ}_{(S_A)_{A\in\mathscr{D}}}$, with weights $\alpha((S_A)_{A\in\mathscr{D}})$ defined by

$$\alpha\big((S_A)_{A\in\mathscr{D}}\big) \;\equiv\; \prod_{A\in\mathscr{D}} \lambda_A(S_A).$$

Since each λ_A is a probability distribution over S_A, the product measure α is also a probability distribution over the entire collection $(S_A)_{A\in\mathscr{D}}$.

[17]The general case is proved in Step 3

Step 2: Suppose that ρ is RU-rationalized by $(\delta_{\succ}, \lambda)$. For each (B, b) such that $b \in B \in \mathscr{D}$,

$$\rho(B, b) = \sum_{(S_A)_{A\in\mathscr{D}} \in \prod_{A\in\mathscr{D}} \text{ supp } \lambda_A} \alpha((S_A)_{A\in\mathscr{D}})\rho^{\succ}_{(S_A)_{A\in\mathscr{D}}}(B, b).$$

Proof. For each (B, b) such that $b \in B \in \mathscr{D}$, the left hand side equals

$$\rho(B, b) = \sum_{(S_b)_{b\in B} \in \prod_{b\in B}(2^{X(b)}\setminus\{\emptyset\})} \lambda_B((S_b)_{b\in B})1(\exists x \in S_b \forall y \in \cup_{a\in B\setminus b} S_a[x \succ y]).$$

As for the right hand side

$$\begin{aligned}
&\sum_{(S_A)_{A\in\mathscr{D}} \in \prod_{A\in\mathscr{D}} \text{ supp } \lambda_A} \alpha((S_A)_{A\in\mathscr{D}})\rho^{\succ}_{(S_A)_{A\in\mathscr{D}}}(B, b) \\
&= \sum_{(S_A)_{A\in\mathscr{D}} \in \prod_{A\in\mathscr{D}} \text{ supp } \lambda_A} \Big(\prod_{A\in\mathscr{D}} \lambda_A(S_A)\Big)\rho^{\succ}_{(S_A)_{A\in\mathscr{D}}}(B, b) \\
&= \sum_{S_B \in \text{ supp } \lambda_B} \lambda_B(S_B)\rho^{\succ}_{(S_A)_{A\in\mathscr{D}}}(B, b) \times \sum_{(S_A)_{A\in\mathscr{D}\setminus B} \in \prod_{A\in\mathscr{D}\setminus B} \text{ supp } \lambda_A} \Big(\prod_{A\in\mathscr{D}\setminus B} \lambda_A(S_A)\Big) \\
&\quad (\because \rho^{\succ}_{(S_A)_{A\in\mathscr{D}}}(B, b) \text{ depends only on } S_B) \\
&= \sum_{S_B \in \text{ supp } \lambda_B} \lambda_B(S_B)\rho^{\succ}_{(S_A)_{A\in\mathscr{D}}}(B, b) \quad (\because \text{ the second term is } 1) \\
&= \sum_{(S_b)_{b\in B} \in \prod_{b\in B}(2^{X(b)}\setminus\{\emptyset\})} \lambda_B(S_B)1(\exists x \in S_b \forall y \in \cup_{a\in B\setminus b} S_a[x \succ y]).
\end{aligned}$$

□

Given Step 2, the general case follows. To see this suppose that ρ is RU-rationalized by $(\mu_{\mathscr{X}}, \lambda)$. Observe that for any (A, a) such that $a \in A \in \mathscr{D}$,

$$\rho(A, a) = \sum_{\succ \in \mathscr{L}(\mathscr{A}_N \cup X(a_0))} \mu_{\mathscr{X}}(\succ) \underbrace{\sum_{(S_a)_{a\in A} \in \prod_{a\in A}(2^{X(a)}\setminus\{\emptyset\})} \lambda_A((S_a)_{a\in A})1(\exists x \in S_a \forall y \in \cup_{b\in A\setminus a} S_b[x \succ y])}_{(*)}. \tag{21}$$

Step 2 shows that each term $(*)$ in equation (21) can be written as a convex combination of elements of the form $\rho^{\succ}_{(S_A)_{A\in\mathscr{D}}}$. Hence, the entire right-hand side of (21) is itself a convex combination of such terms with $\mu_{\mathscr{X}}$.

We have just shown that the set of RU-rational choice functions is a subset of co.$\{\rho^{\succ}_{\mathscr{E}} | \succ \in \mathscr{L}(\mathscr{A})$ and $\mathscr{E} \subseteq 2^{\mathscr{A}}\}$.[18]

[18] This holds regardless of the size of $X(a_0)$.

Now we show the other inclusion. By Theorem 3.1, it can be shown that the set of RU-rational choice functions is convex since . Since $\rho^{\succ}_{\mathscr{E}}$ is RU-rational for all $\succ \in \mathscr{L}(\mathscr{A})$ and $\mathscr{E} \in \mathscr{D}$ (see Remark 3.5), thus its convex combination is also RU rational; thus the set $\text{co.}\{\rho^{\succ}_{\mathscr{E}} | \succ \in \mathscr{L}(\mathscr{A}) \text{ and } \mathscr{E} \subseteq 2^{\mathscr{A}}\}$ is contained by the set of RU-rational choice functions. Therefore, the RU rational set coincides with $\text{co.}\{\rho^{\succ}_{\mathscr{E}} | \succ \in \mathscr{L}(\mathscr{A}) \text{ and } \mathscr{E} \subseteq 2^{\mathscr{A}}\}$.

To finish the proof of Theorem 3.6, we must show that each $\rho^{\succ}_{\mathscr{E}}$ is a vertex. Notice that $\rho^{\succ}_{\mathscr{E}}$ cannot be represented as a convex combination of the other RU rational ρs as the values of the vector $\rho^{\succ}_{\mathscr{E}}$ are all 0 and 1 and all RU rational vectors are nonnegative. This means that all $\rho^{\succ}_{\mathscr{E}}$ are vertices of RU rational polytope. The fact that the RU rational set is $\text{co.}\{\rho^{\succ}_{\mathscr{E}} | \succ \in \mathscr{L}(\mathscr{A}) \text{ and } \mathscr{E} \subseteq 2^{\mathscr{A}}\}$ imply that there are no other vertices. This concludes the proof of statement (ii).

A.3.2 Proof of Remark 3.10

We obtain the lower bound of the number of vertices. Fix $\mathscr{E}$ that contains all subsets of $\mathscr{A}$ that contains a_0. Note that there are $2^{(2^{|\mathscr{A}_N|} - \binom{|\mathscr{A}_N|}{2} - 1)}$ distinct such collection $\mathscr{E}$. Let $\mathcal{C}$ be the collection of such $\mathscr{E}$s. Let $\mathscr{L}'$ be the set of linear orders on $\mathscr{A}$ such that a_0 is the worst alternative. There are $|\mathscr{A}_N|!$ distinct such linear orders.

It can be shown that for any $\mathscr{E}, \mathscr{E}' \in \mathcal{C}$ and $\succ, \succ' \in \mathscr{L}'$ such that $(\succ, \mathscr{E}) \neq (\succ', \mathscr{E}')$, we have $\rho^{\succ}_{\mathscr{E}} \neq \rho^{\succ'}_{\mathscr{E}'}$. Thus, the number of distinct vertices is at least

$$|\mathscr{A}_N|! \times 2^{(2^{|\mathscr{A}_N|} - \binom{|\mathscr{A}_N|}{2} - 1)}.$$

Finally we show $\rho^{\succ}_{\mathscr{E}} \neq \rho^{\succ'}_{\mathscr{E}'}$ if $(\succ, \mathscr{E}) \neq (\succ', \mathscr{E}')$. To see this suppose by way of contradiction that $\rho^{\succ}_{\mathscr{E}} = \rho^{\succ'}_{\mathscr{E}'}$. Since $\succ, \succ' \in \mathscr{L}'$, the worst elements of $\succ$ and $\succ'$ are a_0. Moreover, both $\mathscr{E}$ and $\mathscr{E}'$ contain the sets of the form $\{a, b, a_0\}$ for all $\{a, b\} \subseteq \mathscr{A}_N$. Thus, $\rho^{\succ}_{\mathscr{E}} = \rho^{\succ'}_{\mathscr{E}'}$ implies that $\succ$ and $\succ'$ coincide binary comparison of any $a, b \in \mathscr{A}_N$. It follows that $\succ = \succ'$, which in turn implies $\mathscr{E} = \mathscr{E}'$ given $\rho^{\succ}_{\mathscr{E}} = \rho^{\succ'}_{\mathscr{E}'}$ and the fact that $\mathscr{E}$ contains all sets containing a_0.[19]

A.4 Proofs for Section 4

To prove Proposition 4.2, we first provide a lemma:

Lemma A.1. *Let ρ be a stochastic choice function.*

(a) Suppose that $(\mu_{\mathscr{X}}, \lambda)$ rationalizes ρ and $\mu_{\mathscr{X}}$ is non-overlapping. Then for any composition distribution λ', $(\mu_{\mathscr{X}}, \lambda')$ rationalizes ρ.

[19]There are many redundant vertices. For example, for any $\mathscr{E}$, fix two linear orders $\succ$ and $\succ'$ such that $a_1 \succ \cdots \succ a_{|\mathscr{A}_N|} \succ a_0$ and $a_1 \succ' \cdots \succ' a_{|\mathscr{A}_N|-1} \succ' a_0 \succ' a_{|\mathscr{A}_N|}$. Then we have $\rho^{\succ}_{\mathscr{E}} = \rho^{\succ'}_{\mathscr{E}}$.

(b) *Suppose that there is no non-overlapping $\mu_{\mathscr{X}}$ such that $(\mu_{\mathscr{X}}, \lambda)$ rationalizes ρ for some λ. Then, for any composition distribution λ', there does not exist non-overlapping $\mu'_{\mathscr{X}}$ such that $(\mu'_{\mathscr{X}}, \lambda')$ rationalizes ρ.*

Proof. Statement (a) follows from the fact that if $\mu_{\mathscr{X}}$ is non-overlapping then for all $(S_a)_{a\in A} \in \prod_{a\in A}(2^{X(a)}\backslash\{\emptyset\})$ and $(S'_a)_{a\in A} \in \prod_{a\in A}(2^{X(a)}\backslash\{\emptyset\})$ we have that $\rho_{\mathscr{X}}(\cup_{a\in A}S_a, S_b) = \rho_{\mathscr{X}}(\cup_{a\in A}S'_a, S'_b)$ for all for all $b \in A$ where $\rho_{\mathscr{X}}$ is the underlying stochastic choice function on $\mathscr{X}$ rationalized by $\mu \in \Delta(\mathscr{L}(\mathscr{A}_N \cup X(a_0)))$. Fix $(S^*_a)_{a\in A}$. Then for all $b \in A$, we have

$$\begin{aligned}
&\sum_{(S_a)_{a\in A}\in\prod_{a\in A}(2^{X(a)}\backslash\{\emptyset\})} \lambda_A((S_a)_{a\in A})\rho_{\mathscr{X}}(\cup_{a\in A}S_a, S_b)\\
&= \rho_{\mathscr{X}}(\cup_{a\in A}S^*_a, S^*_b) \sum_{(S_a)_{a\in A}\in\prod_{a\in A}(2^{X(a)}\backslash\{\emptyset\})} \lambda_A((S_a)_{a\in A})\\
&= \rho_{\mathscr{X}}(\cup_{a\in A}S^*_a, S^*_b)\\
&= \mu_{\mathscr{X}}(\succ |\exists x \in S^*_b,\ \forall y \in \cup_{a\in A\backslash b}S_a,\ x \succ y).
\end{aligned}$$

Thus every λ produces the same choice function. Thus, by the definition of rationalization (Definition 2.5), we established statement (a).

To prove statement (b), assume there is no non-overlapping $\mu_{\mathscr{X}}$ that RU rationalizes ρ. By way of contradiction assume that there exist λ' and non-overlapping $\mu'_{\mathscr{X}}$ such that $(\mu'_{\mathscr{X}}, \lambda')$ that rationalizes ρ. Given this, statement (a) implies a contradiction. $\square$

A.4.1 Proof of Proposition 4.2

Only if direction: It suffices to show this for the case where $\mu_{\mathscr{X}}$ is deterministic. Suppose $\mu_{\mathscr{X}}$ is non-overlapping and $\mu_{\mathscr{X}}(\succ) = 1$ for some $\succ \in \mathscr{L}(\mathscr{X})$ and fix λ. Label the elements of $\mathscr{A}$ as $a_1, a_2, ...a_{|\mathscr{A}|}$ in such a way that if $i < j$ then $x \succ y$ for all $x \in X(a_i)$ and $y \in X(a_j)$. This labeling is possible by the almost non-overlapping condition.
Let $\succ_{\mathscr{A}}$ be defined such that

$$a_1 \succ_{\mathscr{A}} a_2 \succ_{\mathscr{A}} \cdots \succ_{\mathscr{A}} a_{|\mathscr{A}|}.$$

It is easy to see that if $D \subseteq \mathscr{A}$ contains any element of $\mathscr{A}$ then ρ puts probability 1 on the a_i with the largest index. Thus, $\mu_{\mathscr{A}}$ rationalizes ρ.

If direction: There exists $\mu_{\mathscr{A}} \in \Delta(\mathscr{L}(\mathscr{A}))$ that rationalizes ρ. For each $\succ \in \operatorname{supp} \mu_{\mathscr{A}}$ consider a ranking $\succ' \in \mathcal{L}(\mathscr{X})$ that extends $\succ$ by

$$x \succ' y \text{ if } x \in X(a) \text{ and } y \in X(b) \text{ for some } a \ \succ \ b$$

and fill the order among $X(a)$ arbitrarily for all a; then define $\mu_{\mathscr{X}}(\succ') = \mu_{\mathscr{A}}(\succ)$. Also define λ arbitrarily. Since $\mu_{\mathscr{X}}$ is non-overlapping, it follows from statement (b) of Lemma A.1 that $(\mu_{\mathscr{X}}, \lambda)$ rationalizes ρ.

A.4.2 Proof of Lemma 4.5

Fix $A \subseteq \mathscr{A}$ and $a \in A$. Then,

$$\begin{aligned}
\rho(A,a) &= \sum_{(S_b)_{b\in A}\in\prod_{b\in A}\left(2^{X(b)}\setminus\{\emptyset\}\right)} \lambda_A\left((S_b)_{b\in A}\right)\rho_{\mathscr{X}}(\cup_{b\in A}S_b, S_a)\\
&= \sum_{(S_b)_{b\in A}\in\prod_{b\in A}\left(2^{X(b)}\setminus\{\emptyset\}\right)} \lambda\Big(\prod_{b\in A}(S_b)\times\prod_{b\in\mathscr{A}\setminus A}(2^{X(b)}\setminus\{\emptyset\})\Big)\rho_{\mathscr{X}}(\cup_{b\in A}S_b, S_a)\\
&= \sum_{(S_b)_{b\in A}\in\prod_{b\in A}\left(2^{X(b)}\setminus\{\emptyset\}\right)}\ \sum_{(S_c)_{a\in B^c}\in\prod_{c\in B^c}\left(2^{X(c)}\setminus\{\emptyset\}\right)} \lambda((S_a)_{a\in B},(S_c)_{c\in B^c})\rho_{\mathscr{X}}(\cup_{b\in A}S_b, S_a)\\
&= \sum_{(S_b)_{b\in\mathscr{A}}\in\prod_{b\in\mathscr{A}}(2^{X(b)}\setminus\{\emptyset\})} \lambda\left((S_b)_{b\in\mathscr{A}}\right)\rho_{\mathscr{X}}(\cup_{b\in B}S_b, S_a).
\end{aligned}$$

A.4.3 Proof of Proposition 4.6

If direction: If ρ is rationalizable by a random utility model $\mu_{\mathscr{A}}$ over $\mathscr{A}$, then as we showed in Proposition 4.2 that there exists $(\mu_{\mathscr{X}}, \lambda)$ that rationalizes ρ and $\mu_{\mathscr{X}}$ is non-overlapping. Moreover, as proved in Lemma A.1 (a), for non-overlapping representation any λ works; so we can choose an independent λ as desired.

Only if direction: Suppose that ρ is RU rationalizable with $(\mu_{\mathscr{X}}, \lambda)$, where λ is independent. Let $\rho_{\mathscr{X}}$ be the random utility function associated with $\mu_{\mathscr{X}}$. Given $(S_a)_{a\in\mathscr{A}} \subseteq \prod_{a\in\mathscr{A}} X(a)$, for each (B,b) such that $b \in B \subseteq \mathscr{A}$ define $\rho_{(S_a)_{a\in\mathscr{A}}}(B,b) = \rho_{\mathscr{X}}(\cup_{a\in B}S_a, S_b)$.[20] Given $(S_a)_{a\in\mathscr{A}} \subseteq \prod_{a\in\mathscr{A}} 2^{X(a)}\setminus\{\emptyset\}$, for each $\succ\in\mathcal{L}(\mathscr{X})$ define $\succ_{(S_a)_{a\in\mathscr{A}}}$ by $a \succ_{(S_a)_{a\in\mathscr{A}}} b$ if and only if there exists $x \in S_a$ such that $x \succ y$ for all $y \in S_b$. Note that $\succ_{(S_a)_{a\in\mathscr{A}}}$ is a linear order on $\mathscr{A}$. Now we define a probability measure $\mu_{(S_a)_{a\in\mathscr{A}}}$ over linear orders on $\mathscr{A}$ as follows: for any linear order $\succ'$ on $\mathscr{A}$, define $\mu_{(S_a)_{a\in\mathscr{A}}}(\succ') = \mu_{\mathscr{X}}(\succ\in\mathscr{L}(\mathscr{X})|\succ_{(S_a)_{a\in\mathscr{A}}}=\succ')$.

Note that, by definition, $\mu_{(S_a)_{a\in\mathscr{A}}}$ RU rationalizes $\rho_{(S_a)_{a\in\mathscr{A}}}$, that is for any (A,a) such that $a \in A \subseteq \mathscr{A}$, $\rho_{(S_a)_{a\in\mathscr{A}}}(A,a) = \mu_{(S_a)_{a\in\mathscr{A}}}(\succ'\in\mathscr{L}(\mathscr{A})|a\succ' b,$ for all $b\in A\setminus a)$. Now let

[20] Note that $\rho_{(S_a)_{a\in\mathscr{A}}}$ is a data set where a always means S_a.

$\mu_{\mathscr{A}} = \sum_{(S_a)_{a\in\mathscr{A}}} \lambda((S_a)_{a\in\mathscr{A}})\mu_{(S_a)_{a\in\mathscr{A}}}$. Then

$$\begin{aligned}
\rho(B,a) &= \sum_{(S_a)_{a\in\mathscr{A}}} \lambda((S_a)_{a\in\mathscr{A}})\rho_{\mathscr{X}}(\cup_{a\in B}S_a, S_a) && (\because \text{Lemma 4.5}) \\
&= \sum_{(S_a)_{a\in\mathscr{A}}} \lambda((S_a)_{a\in\mathscr{A}})\mu_{(S_a)_{a\in\mathscr{A}}}(\succ'\in \mathscr{L}(\mathscr{A})|a \succ' b, \text{ for all } b \in A\setminus a) \\
&= \mu_{\mathscr{A}}(\succ'\in \mathscr{L}(\mathscr{A})|a \succ' b, \text{ for all } b \in A\setminus a).
\end{aligned}$$

Thus $\mu_{\mathscr{A}}$ rationalizes ρ.

B Online Appendix: Supplementary Simulation Results–Bias and Distance

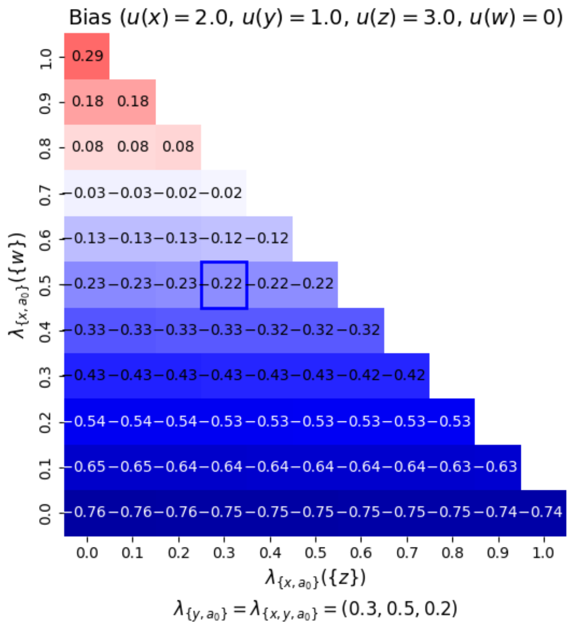
B.1 Estimation Bias

B.1.1 Effect of composition distributions λ

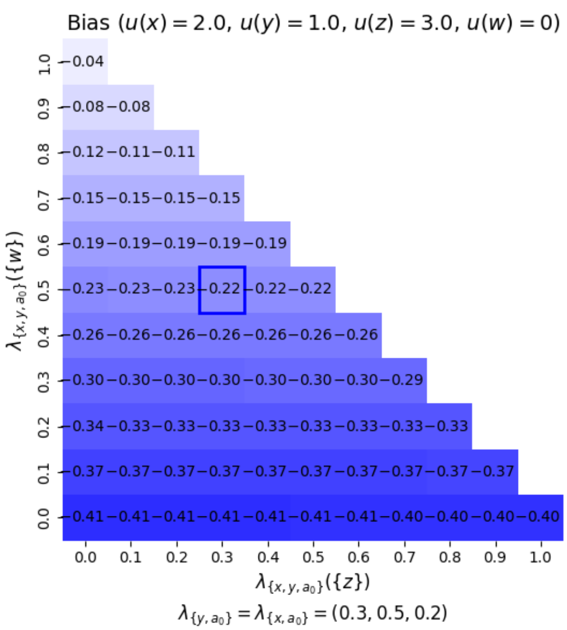
In this section, we replicate the simulation in Subsection 5.2.1 using different parameter values. Overall, the results are qualitatively similar: biases are smaller in the menu-independent cases, and as we move away from independence, the biases tend to increase.

There are, however, some exceptions. Consider the two heatmaps of biases across $\lambda_{\{x,y,a_0\}}$. In these heatmaps, the pattern differs slightly. The blue cell (the independent case) does not necessarily correspond to the smallest bias, and cells farther from independence do not always correspond to larger biases. This finding can be explained by the fact that these two heatmaps are generated by varying the values of $\lambda_{\{x,y,a_0\}}$, which in turn changes the meaning of a_0 in the choice set $\{x, y, a_0\}$. In this choice set, both x and y are present, so changes in the interpretation of a_0 affect the relative desirability of x and y symmetrically. As a result, such changes do not substantially affect the bias, which is defined as the difference in utilities between x and y.

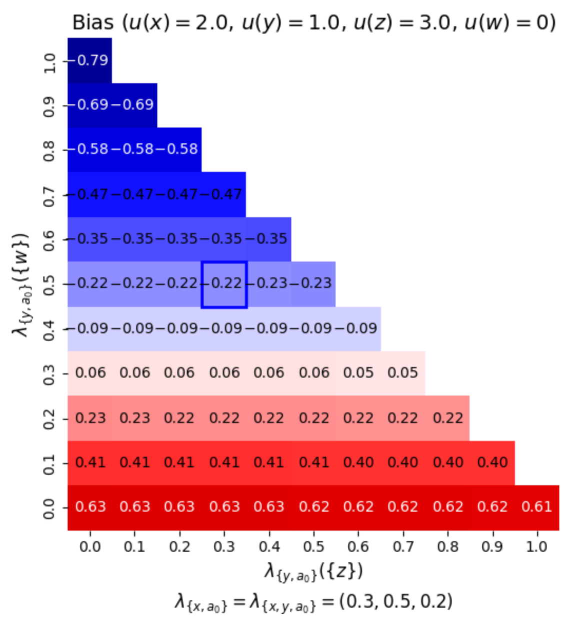
By contrast, when we vary the values of $\lambda_{\{x,a_0\}}$ or $\lambda_{\{y,a_0\}}$, the effect is asymmetric: changes in $\lambda_{\{x,a_0\}}$ affect only x, while changes in $\lambda_{\{y,a_0\}}$ affect only y. These changes directly influence the relative evaluation of x and y, thereby producing larger estimation biases, as expected.



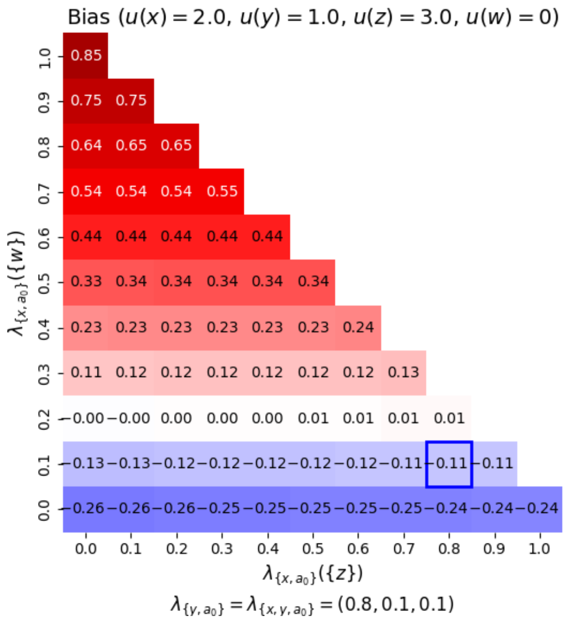
Heatmap of Biases across $\lambda_{\{x,a_0\}}$.



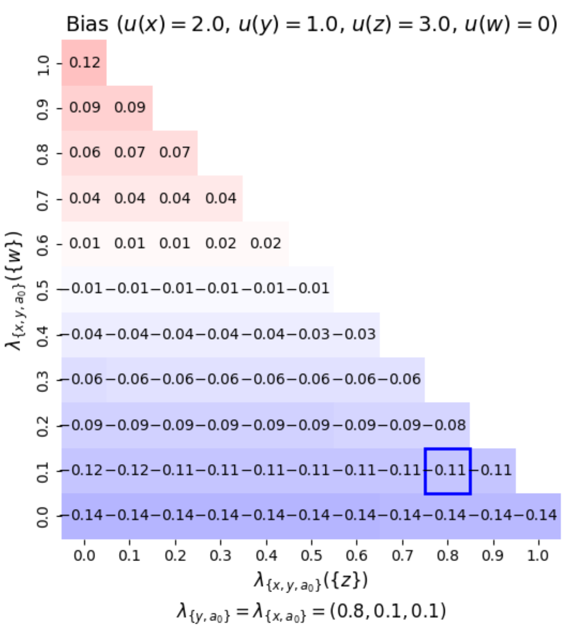
Heatmap of Biases across $\lambda_{\{x,y,a_0\}}$.



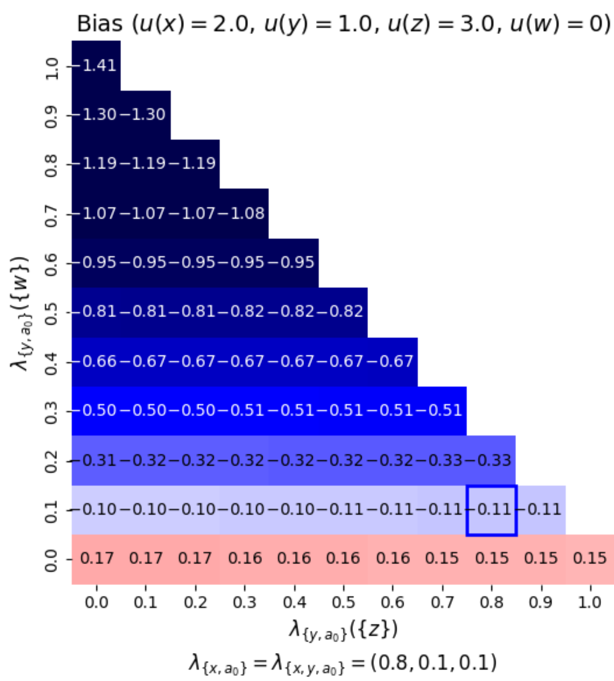
Heatmap of Biases across $\lambda_{\{y,a_0\}}$.



Heatmap of Biases across $\lambda_{\{x,a_0\}}$.



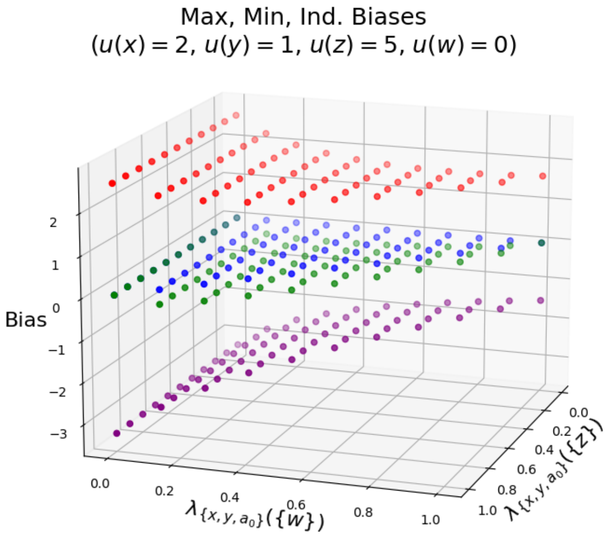
Heatmap of Biases across $\lambda_{\{x,y,a_0\}}$.

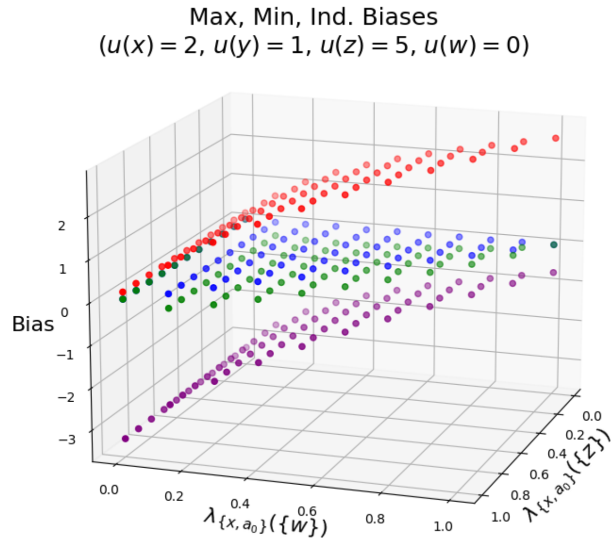


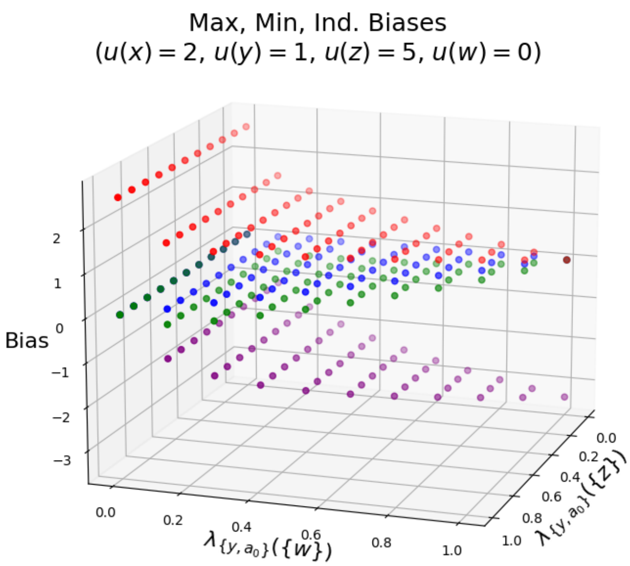
Heatmap of Biases across $\lambda_{\{y,a_0\}}$.

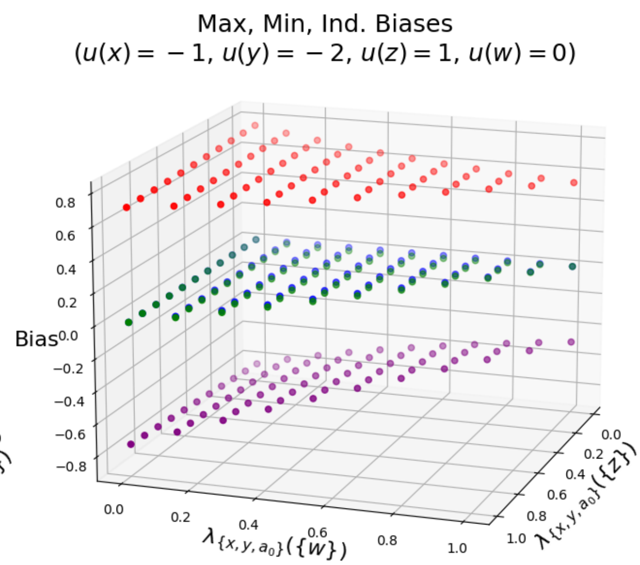
B.1.2 Maximum/Minimum, and independent case

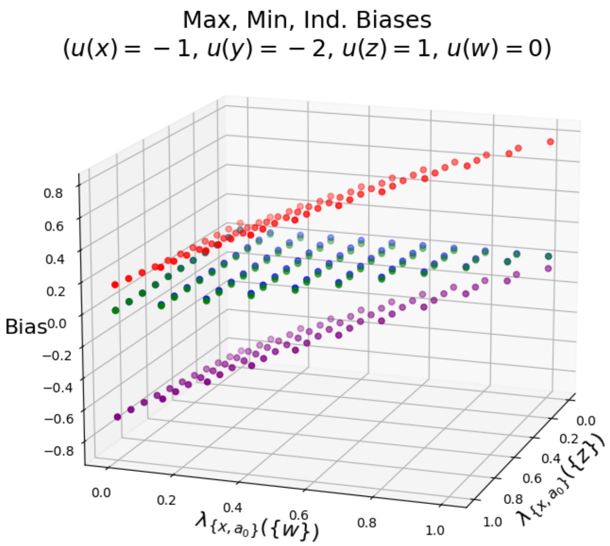
In this section, we replicate the simulation in Subsection 5.2.2 under different parameter values. The setup is the same as in Figure 2, except that we use alternative utility levels and different axes. The overall patterns are consistent with those reported in the main body of the paper.

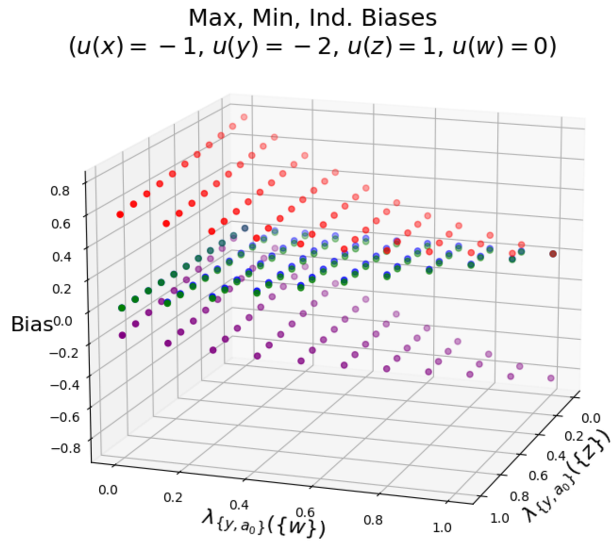

Max, Min, Ind. Biases
$(u(x) = 2, u(y) = 1, u(z) = 5, u(w) = 0)$
Bias
$\lambda_{\{x,y,a_0\}}(\{w\})$
$\lambda_{\{x,y,a_0\}}(\{z\})$


Max, Min, Ind. Biases
$(u(x) = 2, u(y) = 1, u(z) = 5, u(w) = 0)$
Bias
$\lambda_{\{x,a_0\}}(\{w\})$
$\lambda_{\{x,a_0\}}(\{z\})$


Max, Min, Ind. Biases
$(u(x) = 2, u(y) = 1, u(z) = 5, u(w) = 0)$
Bias
$\lambda_{\{y,a_0\}}(\{w\})$
$\lambda_{\{y,a_0\}}(\{z\})$


Max, Min, Ind. Biases
$(u(x) = -1, u(y) = -2, u(z) = 1, u(w) = 0)$
Bias
$\lambda_{\{x,y,a_0\}}(\{w\})$
$\lambda_{\{x,y,a_0\}}(\{z\})$

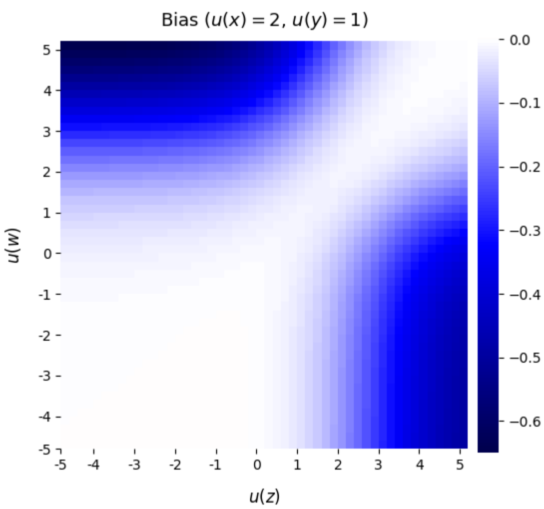

Max, Min, Ind. Biases
$(u(x)=-1, u(y)=-2, u(z)=1, u(w)=0)$
Bias
$\lambda_{\{x,a_0\}}(\{W\})$
$\lambda_{\{x,a_0\}}(\{Z\})$


Max, Min, Ind. Biases
$(u(x)=-1, u(y)=-2, u(z)=1, u(w)=0)$
Bias
$\lambda_{\{y,a_0\}}(\{W\})$
$\lambda_{\{y,a_0\}}(\{Z\})$

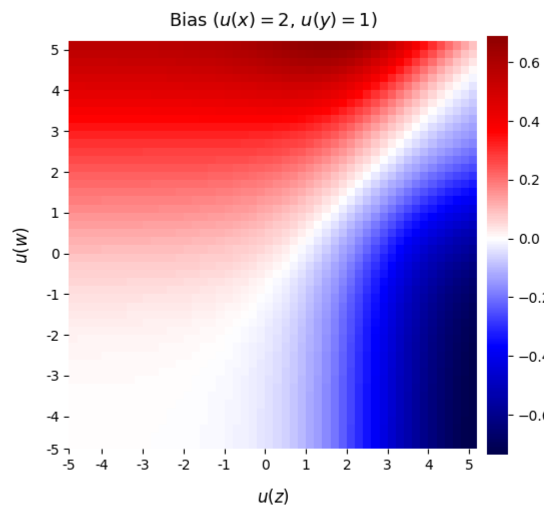
B.2 Effect of preference structure on bias

In this section, we repeat the simulation in Subsection 5.2.3 with different parameter values. The setup is the same as in Figure 3 but with different values of λ.

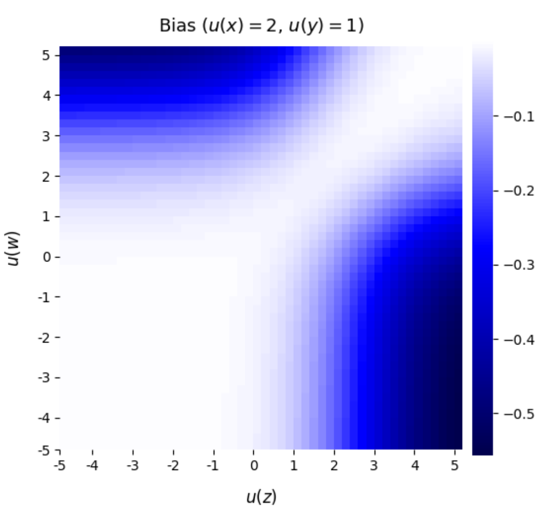
The figures show the biases across utility values of z and w. In general, the bias is largest when overlappingness is most likely (the top left and bottom right corners of each square) and smallest when overlappingness is less likely (the top right and bottom left corners of each square), consistent with Conjecture (B) and Figure 3.



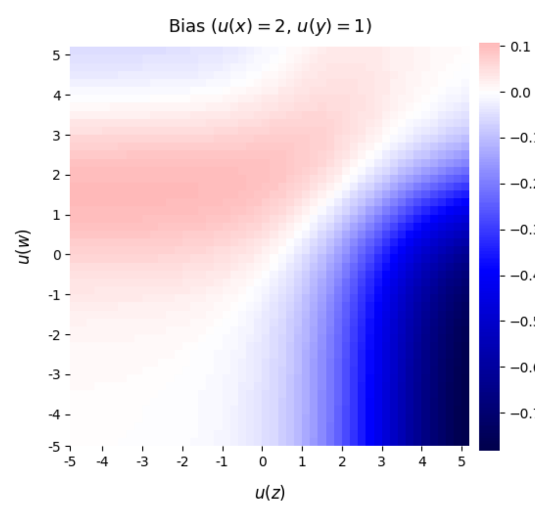
$\lambda_{\{x,y,a_0\}} = (0.1, 0.4, 0.5)$ $\lambda_{\{x,a_0\}} = (0.2, 0.3, 0.5)$ $\lambda_{\{y,a_0\}} = (0.2, 0.3, 0.5)$



$\lambda_{\{x,y,a_0\}} = (0.1, 0.4, 0.5)$ $\lambda_{\{x,a_0\}} = (0.5, 0.3, 0.2)$ $\lambda_{\{y,a_0\}} = (0.1, 0.4, 0.5)$



$\lambda_{\{x,y,a_0\}} = (0.1, 0.4, 0.5)$ $\lambda_{\{x,a_0\}} = (0.1, 0.4, 0.5)$ $\lambda_{\{y,a_0\}} = (0.1, 0.4, 0.5)$



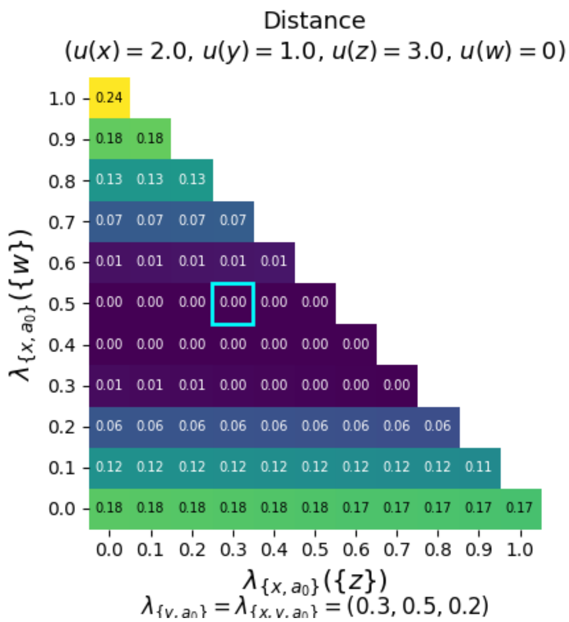
$\lambda_{\{x,y,a_0\}} = (0.8, 0.1, 0.1)$ $\lambda_{\{x,a_0\}} = (0.1, 0.4, 0.5)$ $\lambda_{\{y,a_0\}} = (0.1, 0.4, 0.5)$

B.3 Distance to the ARU Polytope

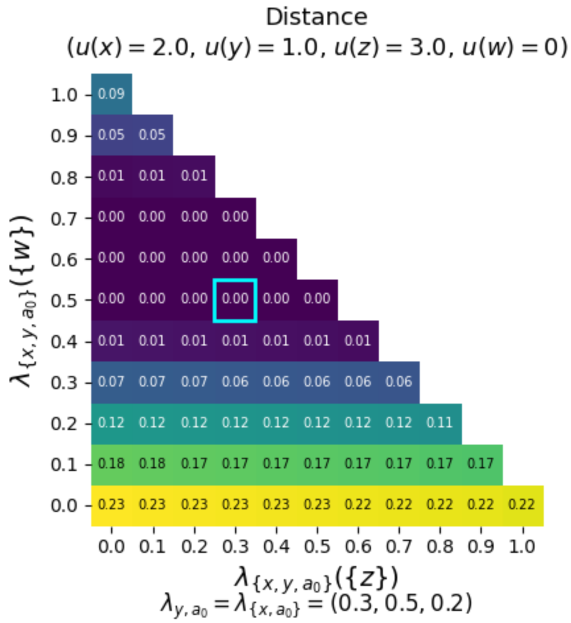
B.3.1 Effect of composition distribution λ on distance

In this section, we repeat the simulation in Subsection 5.3.1 with different parameter values. The setup is the same as in Figure 5 but with different values of λ.

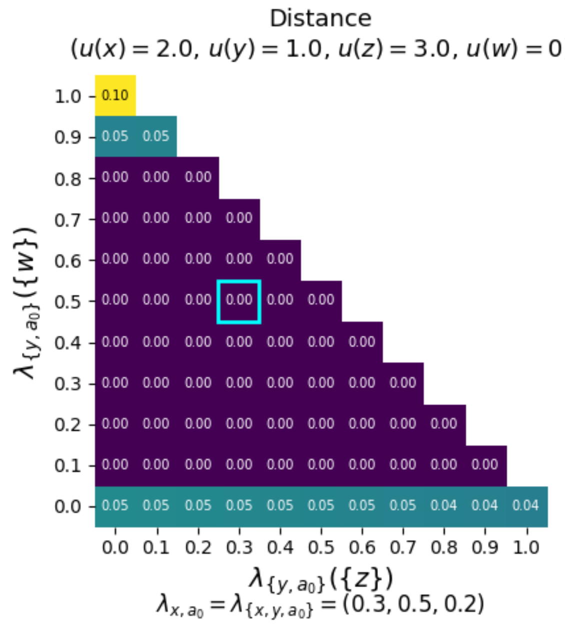
The figures show the distance across values of λ. The resutls are the same as in Subsection 5.3.1: consistent with Proposition 4.6, the distance is zero in the independent case (blue-outlined cell). As λ moves farther from the menu independence, the distance increases, confirming Conjecture (B).



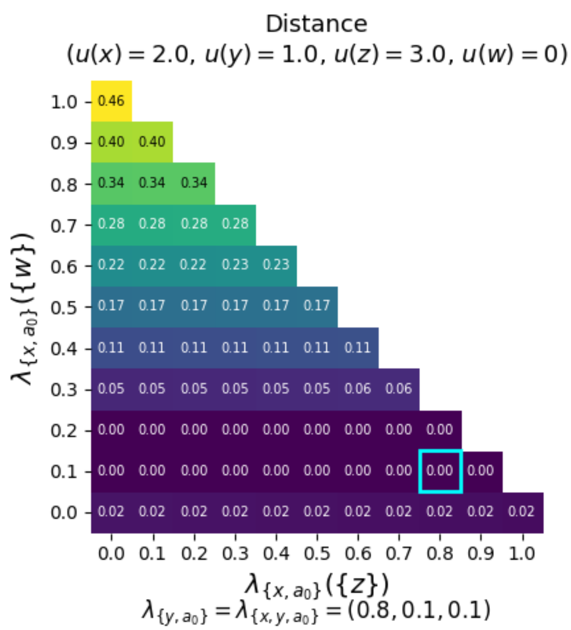
Heatmap of Distance across $\lambda_{\{x,a_0\}}$.



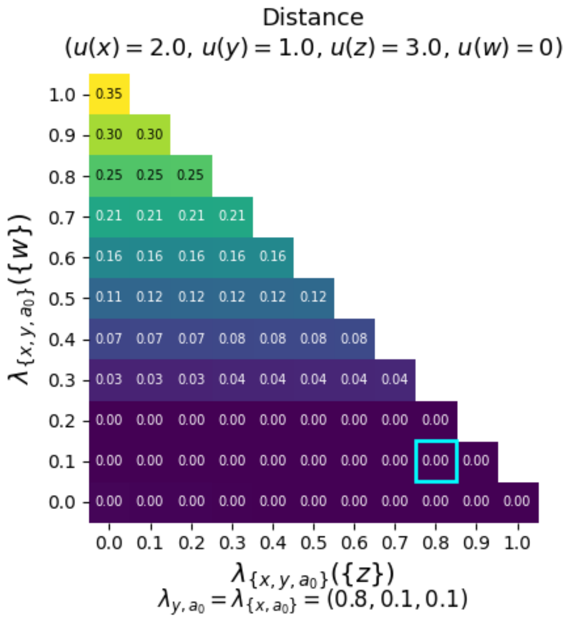
Heatmap of Distance across $\lambda_{\{x,y,a_0\}}$.



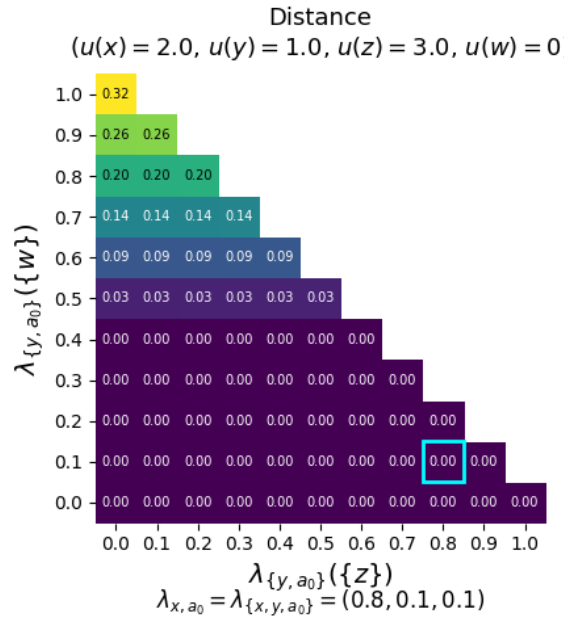
Heatmap of Distance across $\lambda_{\{y,a_0\}}$.



Heatmap of Distance across $\lambda_{\{x,a_0\}}$.



Heatmap of Distance across $\lambda_{\{x,y,a_0\}}$.



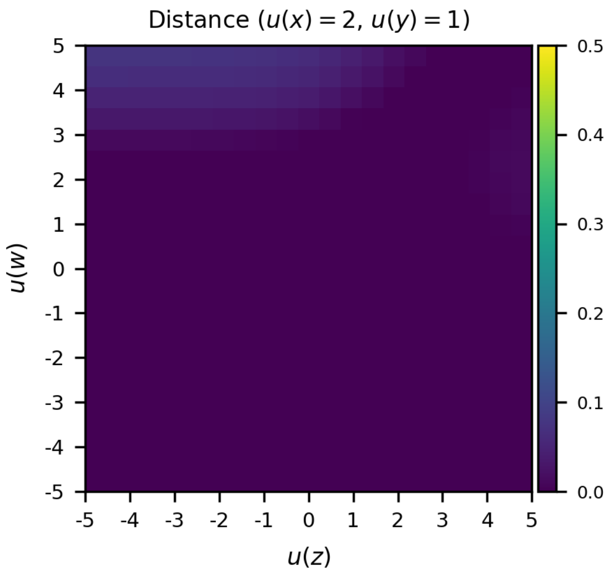
Heatmap of Distance across $\lambda_{\{y,a_0\}}$.

B.3.2 Effect of preference structure on distance

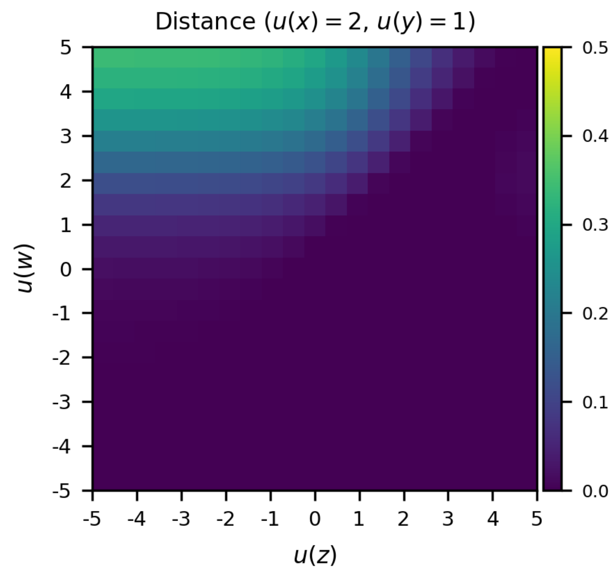
Finally we repeat the simulation in Section 5.3.2. These four figures are the same as Figure 6 with different values of λ — they show the distance from the ARUM polytope across across utility values of z and w.

In general, the distance is largest when overlappingness is most likely (the top left and bottom right corners of each square) and smallest when overlappingness is less likely (the top right and bottom left corners of each square), consistent with Conjecture (B) and the results obtained in the body of the paper.

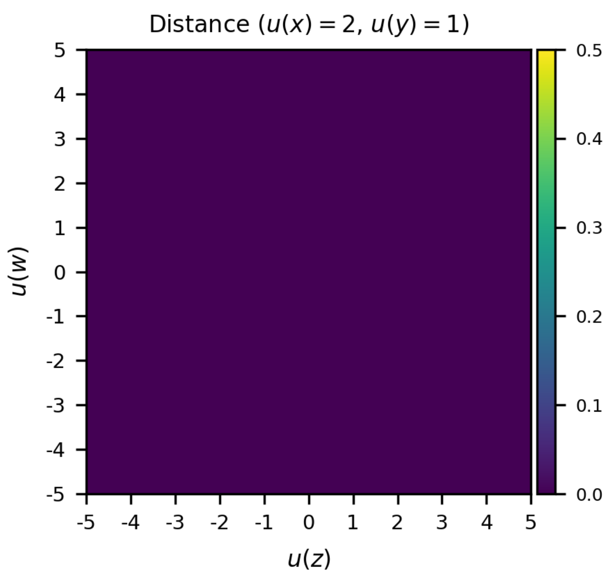
In the first figure ($\lambda_{\{x,y,a_0\}} = \lambda_{\{y,a_0\}} = (0.1, 0.4, 0.5)$ and $\lambda_{\{x,a_0\}} = (0.2, 0.3, 0.5)$ and the third figure ($\lambda_{\{x,y,a_0\}} = \lambda_{\{x,a_0\}} = \lambda_{\{y,a_0\}} = (0.1, 0.4, 0.5)$), the distances are close to zero since λ is either menu-independent as in the third figure or close to menu-independent as in the first figure. In particular, when λ is menu-independent the distance is always zero which is consistent with Proposition 4.6.



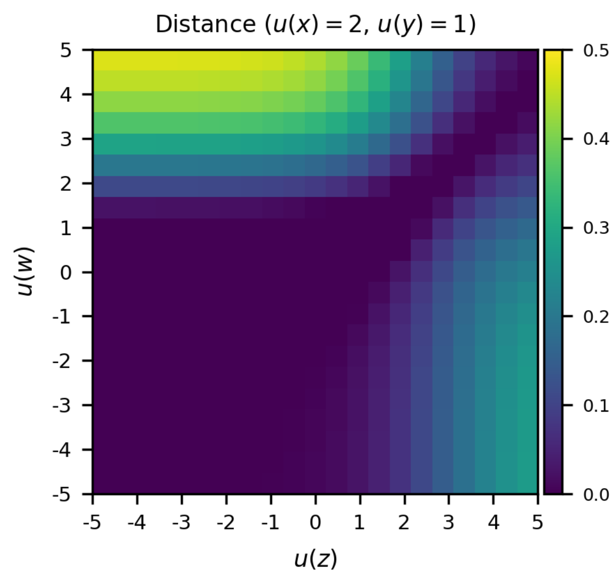
$\lambda_{\{x,y,a_0\}} = (0.1, 0.4, 0.5)\lambda_{\{x,a_0\}} = (0.2, 0.3, 0.5)\lambda_{\{y,a_0\}} = (0.2, 0.3, 0.5)$



$\lambda_{\{x,y,a_0\}} = (0.1, 0.4, 0.5)\lambda_{\{x,a_0\}} = (0.5, 0.3, 0.2)\lambda_{\{y,a_0\}} = (0.1, 0.4, 0.5)$



$\lambda_{\{x,y,a_0\}} = (0.1, 0.4, 0.5)\lambda_{\{x,a_0\}} = (0.1, 0.4, 0.5)\lambda_{\{y,a_0\}} = (0.1, 0.4, 0.5)$



$\lambda_{\{x,y,a_0\}} = (0.8, 0.1, 0.1)\lambda_{\{x,a_0\}} = (0.1, 0.4, 0.5)\lambda_{\{y,a_0\}} = (0.1, 0.4, 0.5)$