

Approximating Choice Data by Discrete Choice Models

August 1, 2026

Abstract

Random-coefficient discrete choice models, including mixed-logit, are commonly used in empirical choice analysis for their ability to capture rich substitution patterns, with applications in demand prediction, revenue management, and merger analysis. We ask when these models are sufficiently flexible to approximate any nonparametric random utility model arbitrarily well across choice sets. We show that the necessary and sufficient condition is the affine independence of the characteristic vectors. When the condition fails, we identify strict preference rankings that are difficult to approximate and propose algorithms to quantify the resulting approximation error. Applying the framework to a dataset on DVD sales, we show that widely used random-coefficient models substantially restrict substitution patterns and may lead to suboptimal assortment decisions. We offer practical guidance for diagnosing and mitigating these problems.

Keywords: Discrete choice analysis, stochastic choice theory, misspecification.

1 Introduction

Random-coefficient discrete choice models are widely used in economics, marketing, and operations research to analyze individual and aggregate choice behavior. They are commonly used to approximate rich preferences and to capture flexible substitution patterns. The models are key inputs for counterfactual analyses in applications such as demand prediction, revenue management, and merger analysis. However, the exact degree of flexibility and the limitations of random-coefficient models have not been fully understood.¹ When a random-coefficient model fails to adequately capture the underlying choice behavior, the resulting misspecification can generate prediction errors that lead to misleading policy conclusions and suboptimal managerial decisions.

¹An important early work in this direction is McFadden and Train (2000). See the section on related literature for details.

In this paper, we provide a sharp characterization of the approximation power of random-coefficient additive random utility models (ARUMs). We obtain a necessary and sufficient condition under which random-coefficient models can approximate the choice behavior generated by any nonparametric random utility model arbitrarily well. Some widely used specifications—for example, linear mixed-logit models with a limited set of observed characteristics—violate this condition and therefore cannot flexibly capture key features of choice behavior, including substitution patterns. We show how to identify strict preference rankings that may be difficult for a given model to approximate closely. We then propose algorithms to quantify the magnitude of the resulting approximation errors. Together, these results offer researchers practical tools to assess whether their chosen discrete choice specification is rich enough for the application at hand.

We consider the standard discrete-choice setup (see, e.g., Train (2009)). Let J denote the set of all alternatives. For each alternative $j \in J$, we use $x_j \in \mathbf{R}^K$ to denote the vector of characteristics of alternative j , where K is the number of characteristics. Let $\rho(D, j)$ denote the probability of choosing j from a choice set $D \subset J$. In this paper, we focus on the class of *random-coefficient additive random utility models*. Such models—including mixed-logit specifications—have become workhorse tools for empirical work in industrial organization, marketing, transportation, and operations. They are routinely used to capture rich substitution patterns and to accommodate both observed and unobserved heterogeneity in preferences.

To introduce these models, we first define an *additive random utility model (ARUM)* as one in which the choice probabilities take the form $\rho(D, j) = \mu(\{\varepsilon | \beta \cdot x_j + \eta_j + \varepsilon_j > \beta \cdot x_l + \eta_l + \varepsilon_l, \forall l \in D \setminus \{j\}\})$, where $\beta \in \mathbf{R}^K$ is a coefficient vector capturing an agent's preferences, $\eta = (\eta_j)_{j \in J}$ is a vector of *fixed effects* capturing unobserved characteristics of alternatives, and $\varepsilon = (\varepsilon_j)_{j \in J}$ is a vector of random utility shocks with a probability measure μ . The class of ARUMs is general and includes probit and logit models as special cases. A *random-coefficient ARUM* extends this framework by allowing taste coefficients to vary across agents. Formally, the choice probabilities are obtained by integrating ARUM choice probabilities against a distribution over β . In one standard interpretation, the distribution m captures the heterogeneity of preferences among the population of agents. When μ is an iid type-I extreme-value distribution, ρ reduces to a *mixed-logit model*.

Given the popularity of the random-coefficient ARUMs, it is important to understand their flexibility and limitations. For this purpose, we obtain a necessary and sufficient condition under which the random-coefficient ARUMs are rich enough to approximate any choice probabilities generated by nonparametric random utility models arbitrarily well across choice sets. For the approximation target, we choose the random utility models, which are defined as probability measures over strict preference rankings over alternatives. We choose this class of models as the

approximation target because it is the most general and agnostic framework that only assumes individual rationality. We study approximation *across choice sets* because many questions of interest, such as substitution patterns, rely on analyzing behavior across choice sets. Choice data across multiple choice sets are commonly encountered in applications such as conjoint analysis, revenue management, and merger analysis. Counterfactual analyses also require predicting behavior in new or modified choice sets. In such cases, the random-coefficient structure imposes restrictions across choice sets.²

We briefly summarize our main results. Our main theorem (Theorem 1) states that the necessary and sufficient condition for approximating arbitrarily well is the *affine-independence* of the set $\{x_j \in \mathbf{R}^K | j \in J\}$.³ To interpret the condition, consider a typical setup, where a researcher first fixes one probability measure μ over the shock ε . Then the researcher estimates the distribution m over coefficients β and the fixed effects η after observing a dataset. If the affine-independence condition is satisfied, then the researcher should be able to approximate any given dataset by using some random-coefficient ARUMs arbitrarily well across choice sets. On the other hand, if the affine-independence condition is violated, there exists a dataset generated by a random utility model that cannot be approximated arbitrarily well by any sequence of random-coefficient ARUMs, no matter which random-coefficient distribution m and fixed effects η are used. The affine-independence condition is straightforward to test: the condition is generically equivalent to the simpler condition: $K \geq |J| - 1$, where $|J|$ is the number of alternatives and K is the number of characteristics observed for each alternative.

Many empirical papers use the mixed-logit models that are linear in the original characteristics and do not contain additional terms such as polynomials. We call such models *linear mixed-logit models*. In these papers, we often observe a deviation from the condition $K \geq |J| - 1$, resulting in the violation of the affine-independence condition. This means that the linear mixed-logit models may not be rich enough to approximate the true substitution pattern arbitrarily well across subsets of J , no matter how one chooses the distribution m and the fixed effects η .

We show that the source of this approximation failure can be traced to strict preference rankings that are difficult for the model to approximate. These rankings can be identified efficiently using linear programming methods.

To further quantify the flexibility of any given random-coefficient ARUM specification, we develop two algorithms to compute the approximation error to strict preference rankings that are particularly difficult to approximate. First, we implement an EM (Expectation–Maximization) algorithm, following Dempster, Laird, and Rubin (1977). Second, we adapt the greedy algorithm

²In applications with fixed effects, these fixed effects are often estimated from observed choice sets and then held fixed or otherwise specified for counterfactual choice sets.

³A set $Y \equiv \{y_1, \dots, y_n\}$ is *affinely independent* if for any $y_i \in Y$, there exist no real numbers $\{\mu_j\}_{j \neq i}$ such that $y_i = \sum_{j \neq i} \mu_j y_j$ and $\sum_{j \neq i} \mu_j = 1$. A set $Y \equiv \{y_1, \dots, y_n\}$ is *affinely independent* if and only if the set $\{y_2 - y_1, \dots, y_n - y_1\}$ is linearly independent.

proposed in Barron, Cohen, Dahmen, and DeVore (2008). With a slight modification, we also compute the maximal substitution patterns that are attainable within a given class of random-coefficient ARUMs.

We apply our theorem and algorithms to a DVD sales dataset from Rusmevichientong, Shen, and Shmoys (2010). We focus on a subsample with four alternatives (products 1–4 hereafter) and two characteristics. In this setting, the affine-independence condition fails because $K = 2 < 3 = |J| - 1$. We show that half of all strict preference rankings cannot be well approximated by linear mixed-logit specifications, and the resulting approximation errors can be substantial. Moreover, as shown in Section 5.2, the linear mixed-logit model imposes strong restrictions on substitution patterns. In our setting, *no* linear mixed-logit specification can generate substantial substitution between products 1 and 3 or between products 2 and 4.

In this application, we show that the above limitation of the linear mixed-logit model can distort assortment decisions. Consider a market with three consumer segments: 50% rank $3 \succ 4 \succ 1 \succ 2$, 25% rank $1 \succ 3 \succ 2 \succ 4$, and 25% rank $2 \succ 4 \succ 3 \succ 1$, where product 4 is the outside option. Suppose product 3 has the highest margin, followed by products 1 and 2, and shelf-space constraints require the retailer to remove either product 1 or product 2. A linear mixed-logit model estimated from historical sales cannot predict sizable reallocation from product 1 to product 3 when product 1 is removed, but may predict that removing product 2 redirects demand toward products 1 and 3 with little market exit. The retailer may therefore drop product 2. In reality, however, the 25% segment with preferences $2 \succ 4 \succ 3 \succ 1$ exits the market when product 2 is removed, causing an avoidable revenue loss. Thus, restrictive substitution patterns in linear mixed-logit models can lead firms to mis-rank assortments and make suboptimal decisions.

The structure of the paper is as follows. Section 2 introduces the discrete-choice setup and the definitions of various models used in our analysis. Section 3 presents our main theorem and related results and discusses their implications for representability and approximation across choice sets. Section 4 develops our distance measure and the EM and greedy algorithms for quantifying approximation errors. Section 5 applies our framework to a DVD sales dataset and documents the resulting approximation errors and implications for substitution patterns and assortment planning. All proofs are collected in the appendix.

Related Literature

The work most closely related to our paper is Dagsvik (1994) and, especially, McFadden and Train (2000), who show that any given nonparametric continuous random utility model can be approximated arbitrarily well by a mixed-logit model.⁴ The key differences concern both the

⁴Our result is consistent with theirs: heuristically, the result of McFadden and Train (2000) corresponds to the case in which the researcher allows arbitrarily high-order polynomials (i.e., $K \rightarrow \infty$), which satisfies our condition

setup and the sharpness of the results. First, we provide a *necessary and sufficient* characterization of when random-coefficient ARUMs can approximate arbitrary random-utility behavior, whereas McFadden and Train (2000) only establish a sufficient condition for approximation. We obtain this sharper characterization by focusing on approximation for a fixed, finite set of alternatives. Second, we analyze approximation of choice probabilities *across choice sets*. This cross-menu perspective with fixed alternatives is central for applications in assortment optimization and revenue management (Farias, Jagabathula, and Shah, 2009; Jagabathula and Venkataraman, 2022). Third, our affine-independence condition is necessary and sufficient, applies beyond mixed-logit, and is tailored to environments with fixed finite alternatives. By contrast, McFadden and Train (2000) provide asymptotic approximation guarantees, focusing primarily on settings with continuous covariates, and they offer no comparable guidance on how to construct random-coefficient ARUMs to achieve approximation in the finite-alternatives, cross-menu setting.

Another paper closely related to ours is Norets and Takahashi (2013). They study whether ARUMs can represent any stochastic choice on a single multinomial choice problem. In particular, under a mild regularity condition, they show that for a given menu D , by appropriately choosing the vector of mean utilities (the deterministic utility indices of the alternatives), an ARUM without random coefficients can generate any desired choice probability vector $\{\rho(D, j)\}_{j \in D}$ on the interior of $|D| - 1$ -simplex. Our scope is different in two important respects. First, we study approximation across multiple menus for a fixed set of alternatives. This cross-menu perspective requires the model to approximate the data well simultaneously across different choice sets. This issue does not arise in the single-menu setting. Second, we allow for random coefficients and ask when a given class of random-coefficient ARUMs can approximate any random utility model across menus. As we show (see Lemma 4), menu-by-menu constructions à la Norets–Takahashi cannot in general be made mutually consistent across menus, so their single-menu surjectivity result does not imply our cross-menu approximation result, and vice versa. Other related papers also focus on a fixed choice set. Berry (1994) provides an earlier and classical inversion result useful for representing any stochastic choice on a given choice set. Athey and Imbens (2007) investigate how a rich specification of the unobserved components is needed to represent any stochastic choice function.

A recent paper by Lu and Saito (2021) is also related to our work, but it is conceptually closer to McFadden and Train (2000) than to our paper. Lu and Saito (2021) differ from our analysis in three key respects. First, they work in a continuous characteristics space, as in McFadden and Train (2000), rather than on a finite set of alternatives with fixed characteristics. Second, they restrict the approximation target to pure-characteristics models (continuous random utility),

$K \geq |J| - 1$.

whereas we allow for arbitrary finite-domain random utility models (probability measures over strict rankings). Third, they focus on parametric mixed-logit specifications whose systematic utility is a polynomial of degree at most a given bound, while we consider a general class of random-coefficient ARUMs. They provide a necessary and sufficient condition under which a given pure-characteristics model can be approximated by a mixed-logit model with a specified polynomial degree across choice sets. Thus, although both papers address related approximation questions, they operate in different domains and yield conceptually distinct characterizations. A more detailed comparison between the two frameworks is provided in Section B of the online appendix.⁵

Our paper builds on the decision theory literature, where logit models and random utility models have been analyzed extensively since Luce (1959) and Block and Marschak (1960). For more recent generalizations of logit models, see Cerreia-Vioglio, Maccheroni, Marinacci, and Rustichini (2023), Cerreia-Vioglio, Lindberg, Maccheroni, Marinacci, and Rustichini (2021), and Horan (2021). Recent studies by Apesteguia and Ballester (2018) and Frick, Iijima, and Strzalecki (2019) further highlight the distinctions in choice behavior between random utility models and logit models. Furthermore, several recent studies examine the substitution property in discrete choice analysis. In Horan and Adam (2023), the substitution patterns captured by random utility models are discussed, while Allen and Rehbeck (2020) analyze aggregate complementarity in latent utility models applied to discrete choice. Additionally, Chambers, Masatlioglu, Natenzon, and Raymond (2025) introduce a new model of discrete choice, demonstrating that their model allows for flexible substitution patterns and variations in market shares across different choice sets.

Our analysis also shares some of its spirit with the growing literature that identifies and estimates flexible discrete choice/demand models under minimal assumptions. See, for example, Berry and Haile (2014), Compiani (2022), and Tebaldi, Torgovitsky, and Yang (2023). Despite the similarity in the spirit, our problem is different from the standard econometric problems. We are not concerned with statistical estimation or identification problems, i.e., recovering model parameters in either a sample or a population setting. In contrast, our primary goal is to explore the limitations, in terms of the approximation power, of common modeling strategies within the discrete choice literature. Our work focuses on a specification or approximation question rather than identification, estimation, or inference.

Applications of discrete choice models span a wide range of fields, including economics, transportation, management, and marketing. Berbeglia, Garassino, and Vulcano (2022) and Feng, Shanthikumar, and Xue (2022) review applications in the context of management science

⁵In the section, we discuss McFadden and Train (2000), Norets and Takahashi (2013), and Lu and Saito (2021) in more detail.

and operations research.⁶

2 Model

We first introduce the setup.

Set of alternatives: The set of all alternatives is denoted by J . J is assumed to be finite and $|J| > 2$.

Choice sets: Let $\mathcal{D} \subset 2^J \setminus \{\emptyset\}$ be the set of choice sets. $\mathcal{D}$ can be a proper subset of $2^J \setminus \{\emptyset\}$. Unless otherwise noted, we assume throughout the paper that $\mathcal{D}$ contains all binary and ternary choice sets: $\{j, l\} \in \mathcal{D}$ and $\{j, l, r\} \in \mathcal{D}$ for any distinct $j, l, r \in J$. The set $\mathcal{D}$ may contain both observed choice sets as well as hypothetical choice sets the researcher is interested in. For example, even when the researcher observes consumers' choices only over $\{\text{train}, \text{bus}, \text{car}\}$, he may also be interested in choices over $\{\text{train}, \text{bus}\}$, $\{\text{train}, \text{car}\}$, and $\{\text{bus}, \text{car}\}$ to learn about the consumers' substitution pattern.

Explanatory variables: An alternative $j \in J$ is described by a real vector $x_j \in \mathbf{R}^K$ of explanatory variables, where K is the number of explanatory variables. For instance, if an alternative j is a consumption good, the alternative may be described by its price p_j and its quality index q_j ; in that case $x_j = (p_j, q_j)$. Moreover, the researcher can include functions of original characteristics in x_j such as higher order polynomials as well as splines or wavelets (Chen, 2007). For example, with the original characteristics (p_j, q_j) of alternative j , the researcher may include higher order polynomials such as $p_j^2, q_j^2, p_j q_j$ in the characteristic vector x_j and can make the number K of characteristics larger. If the researcher includes all second-order terms, then $x_j = (p_j, q_j, p_j^2, q_j^2, p_j q_j)$ and $K = 5$.

Stochastic choice function: A function $\rho : \mathcal{D} \times J \rightarrow [0, 1]$ is called a *stochastic choice function* if $\sum_{j \in D} \rho(D, j) = 1$ and $\rho(D, j) = 0$ for any D and $j \notin D$. The set of stochastic choice functions is denoted by $\mathcal{P}$. For each $(D, j) \in \mathcal{D} \times J$, the number $\rho(D, j)$ is interpreted as the probability that an alternative j is chosen from a choice set D . In market-level analysis, $\rho(D, j)$ can also be interpreted as the market share of product j in a market in which the set of available products is D .

Rankings: Let Π be the set of bijections between J and $\{1, \dots, |J|\}$, where $|J|$ is the number of elements of J . For any element $\pi \in \Pi$, if $\pi(j) = i$, then we interpret alternative j to be the $|J| + 1 - i$ -th best element of J with respect to π . If $\pi(j) > \pi(l)$, then j is better than l with respect to π . An element π of Π is called a *ranking* over J . A ranking describes an agent's strict preference relation. There are $|J|!$ elements in Π .

⁶Applications in management science include but are not limited to Kamakura, Kim, and Lee (1996); Chintagunta, Dubé, and Goh (2005); Fiebig, Keane, Louviere, and Wasi (2010); Musalem, Olivares, Bradlow, Terwiesch, and Corsten (2010); Farias, Jagabathula, and Shah (2013); Alptekinoğlu and Semple (2016); Jagabathula and Venkataraman (2022); Akchen and Mitrofanov (2025).

In the main body of the paper, we consider all possible rankings for simplicity. In Section A.2 of the online appendix, we show how our analysis can be extended to accommodate cases where only a subset of rankings is considered, particularly when certain rankings are deemed unreasonable.

Fixed effects: We define fixed effects as a function $\eta : \mathcal{D} \times J \rightarrow \mathbf{R}$. We interpret $\eta_{(D,j)}$ as the value of the fixed effect of alternative j in set D . Thus, $\eta_{(D,j)} = 0$ if $j \notin D$ for all $D \in \mathcal{D}$. We denote the set of fixed effects by $\mathcal{F}$. We say that a fixed effect η is *menu-independent*, if $\eta_{(D,j)} = \eta_{(E,j)}$ for all $D, E \in \mathcal{D}$ and $j \in D \cap E$. For simplicity, we assume fixed effects are *menu-independent* for the main results. In that case, we simply write $\eta_{(D,j)}$ as η_j and the set $\mathcal{F}$ of fixed effects can be identified as $\mathbf{R}^{|J|}$. With *menu-dependent* fixed effects, our main results continue to hold under an additional consistency condition. See Section 3.2 for details.

We now introduce the definition of random utility models. We denote the set of probability measures over Π by $\Delta(\Pi)$. Since Π is finite, $\Delta(\Pi) = \{(\nu_1, \dots, \nu_{|\Pi|}) \in \mathbf{R}_+^{|\Pi|} \mid \sum_{i=1}^{|\Pi|} \nu_i = 1\}$, where $\mathbf{R}_+$ is the set of nonnegative real numbers.

Definition 1. A stochastic choice function ρ is called a *random utility model* if there exists a probability measure $\nu \in \Delta(\Pi)$ such that for all $(D, j) \in \mathcal{D} \times J$, if $j \in D$, then $\rho(D, j) = \nu(\{\pi \in \Pi \mid \pi(j) > \pi(l) \text{ for any } l \in D \setminus \{j\}\})$. The set of random utility models is denoted by $\mathcal{P}_r$.⁷

In both theoretical and empirical literature, researchers approximate the random utility models by using random-coefficient models. First, we introduce one definition.

Definition 2. A Borel probability measure μ on the Borel σ -algebra of $\mathbf{R}^{|J|}$ is said to be a *standard probability measure* if μ is absolutely continuous with respect to the Lebesgue measure and the support is convex.⁸ Let $\mathcal{M}$ be the set of all standard probability measures.

Although this definition is specific to our paper, it is general enough to nest many additive utility-shock distributions common in the literature, such as normal, logistic, and exponential shock distributions.

In the following, we denote the inner product of two vectors x and y by $x \cdot y$.

Definition 3. Let $\eta \in \mathcal{F}$ be a vector of fixed effects. A stochastic choice function ρ is called a *random-coefficient additive random utility model (random-coefficient ARUM) with fixed effects* η if there exist a standard probability measure μ and a Borel probability measure m such that

⁷While the function above is often called a random ranking function, a random utility model is often defined differently by using the existence of a probability measure μ over utilities such that for all $(D, j) \in \mathcal{D} \times J$, if $j \in D$, then $\rho(D, j) = \mu(\{u \in \mathbf{R}^{|J|} \mid u(j) \geq u(i) \text{ for all } i \in D \setminus \{j\}\})$. Block and Marschak (1960)'s Theorem 3.1 proves that the two definitions are equivalent.

⁸Remember that the support $\text{supp } \mu$ is defined as $\{\varepsilon \in \mathbf{R}^{|J|} \mid \mu(N_\varepsilon) > 0 \text{ for any open neighborhood } N_\varepsilon \text{ of } \varepsilon\}$.

for all $(D, j) \in \mathcal{D} \times J$, if $j \in D$, then

$$\rho(D, j) = \int \mu\left(\{\varepsilon | \beta \cdot x_j + \eta_j + \varepsilon_j > \beta \cdot x_l + \eta_l + \varepsilon_l, \forall l \in D \setminus \{j\}\}\right) dm(\beta). \quad (1)$$

The random vector $(\varepsilon_j)_{j \in J} \in \mathbf{R}^{|J|}$ follows the distribution μ . When the support of m has only one point, the stochastic choice function ρ is called an *additive random utility model (ARUM) with fixed effects* η : for all $(D, j) \in \mathcal{D} \times J$, if $j \in D$, then

$$\rho(D, j) = \mu\left(\{\varepsilon | \beta \cdot x_j + \eta_j + \varepsilon_j > \beta \cdot x_l + \eta_l + \varepsilon_l, \forall l \in D \setminus \{j\}\}\right). \quad (2)$$

The set of random-coefficient ARUMs is denoted by $\mathcal{P}_{ra}(\eta|\mu)$ and the set of ARUMs by $\mathcal{P}_a(\eta|\mu)$. When the context makes clear which standard probability measure μ we consider, we do not specify the standard probability measure μ .

The term $\beta \cdot x_j$ is the *systematic (deterministic)* part of the utility of alternative j captured by the observed characteristics x_j . The vector β captures preferences of an agent and the distribution m over coefficients β describes the heterogeneity of preferences among the population of agents. The constant η_j is called a *fixed effect* that captures the utility of alternative j from the unobserved characteristics; ε_j is the preference shock to the utility of alternative j .

Almost all probability measures used in practice are standard. For a mixed-logit model, μ is an iid extreme-value type-I distribution; for a probit model, μ is the multivariate standard normal distribution; for an exponential model, μ is an iid multivariate exponential distribution with scale parameter 1. In most empirical applications of these models, the mixing distribution m is a parametric distribution like a multivariate normal distribution. In our case, the mixing distributions of the random coefficients do not come from a particular parametric family.

For later exposition, we define the *mixed-logit models* formally as follows:

Definition 4. Let $\eta \in \mathcal{F}$ be a vector of fixed effects. A stochastic choice function ρ is called a *mixed-logit model with fixed effects* η if there exists a Borel probability measure m such that for all $(D, j) \in \mathcal{D} \times J$, if $j \in D$, then

$$\rho(D, j) = \int \frac{\exp(\beta \cdot x_j + \eta_j)}{\sum_{l \in D} \exp(\beta \cdot x_l + \eta_l)} dm(\beta). \quad (3)$$

The set of mixed-logit models with fixed effects η is denoted by $\mathcal{P}_{ml}(\eta)$. When m puts the unit mass on a particular β in (3), then ρ is called a *logit model*.

We give two remarks on the models. First, in the definitions above, we write fixed effects as η_j assuming menu-independence for simplicity (i.e., $\mathcal{F} = \mathbf{R}^{|J|}$). However, our results hold more generally with menu-dependent fixed effects. See Section 3.2 for details. Second, in the models above, following McFadden and Train (2000) as well as many other papers in discrete

choice analysis, we write the systematic part of utility by $\beta \cdot x_j$. Although we have a linear structure in the explanatory variables x_j , it is important to remember that x_j can be a function of original characteristics such as higher order polynomials of original characteristics as used in the proof of McFadden and Train (2000). Since any continuous function can be approximated by polynomials, our models are general enough to allow for flexible systematic parts of utility functions.

Finally, we review essential mathematical concepts. A *polytope* is a convex hull of finitely many points. The closure of a set C is denoted by $\text{cl } C$ with respect to the standard finite dimensional Euclidean topology. The *affine hull* of a set C is the smallest affine set that contains C , and it is denoted by $\text{aff } C$. The convex hull of a set C is denoted by $\text{co } C$. The *relative interior* of a convex set C is the interior of C in the relative topology with respect to $\text{aff } C$. The relative interior of C is denoted by $\text{rint } C$. We recall a basic concept in geometry: a set $\{x_j \in \mathbf{R}^K | j \in J\}$ is *affinely independent* if no x_j can be written as an affine combination of the other elements $\{x_l\}_{l \neq j}$. Formally, for any $j \in J$, there exist no real numbers $\{\alpha_l\}_{l \in J \setminus \{j\}}$ such that $x_j = \sum_{l \in J \setminus \{j\}} \alpha_l x_l$ and $\sum_{l \in J \setminus \{j\}} \alpha_l = 1$.

3 Main Result

Theorem 1. (i) *Let $\mu \in \mathcal{M}$ be any standard probability measure. If the set $\{x_j \in \mathbf{R}^K | j \in J\}$ is affinely independent, then any random utility model can be approximated arbitrarily well by a sequence of random-coefficient ARUMs with the given μ ; moreover, the approximation can be done without fixed effects (i.e., $\eta = 0$). Formally,*

$$\forall \mu \in \mathcal{M}, \forall \rho \in \mathcal{P}_r, \exists \{\rho_n\}_{n=1}^\infty \subset \mathcal{P}_{ra}(0|\mu), \forall D \in \mathcal{D}, \forall j \in D, \lim_{n \rightarrow \infty} \rho_n(D, j) = \rho(D, j).$$

(ii) *If the set $\{x_j \in \mathbf{R}^K | j \in J\}$ is not affinely independent, then there exists a random utility model ρ such that for every standard probability measure μ , ρ cannot be approximated arbitrarily well by any sequence of random-coefficient ARUMs with fixed effects and with the distribution μ . Formally,*

$$\exists \rho \in \mathcal{P}_r, \forall \mu \in \mathcal{M}, \rho \notin \text{cl} \bigcup_{\eta \in \mathcal{F}} \mathcal{P}_{ra}(\eta|\mu).$$

We provide the proof in the appendix. To interpret the main theorem, consider a typical practice where the shock distribution μ over ε is fixed a priori by a researcher; then he or she estimates the distribution m over coefficients β as well as the fixed effects η after seeing a dataset ρ generated from a random utility model.

Statement (i) shows that if the set $\{x_j \in \mathbf{R}^K | j \in J\}$ is affinely independent, then the

researcher should be able to approximate the given dataset ρ arbitrarily well across choice sets $D \in \mathcal{D}$ by choosing an appropriate distribution m over coefficients β . This direction builds upon the classical result of McFadden and Train (2000) about mixed-logit. Note that the result allows a general class of utility-shock distributions, not confined to the type-I extreme-value distribution or a particular choice of mixing distributions.

Statement (ii) shows that if the affine-independence condition fails, then there exists a random utility model that cannot be approximated arbitrarily well by any sequence of random-coefficient ARUMs, no matter how the researcher changes the distribution m over coefficients β as well as the fixed effects η , given any arbitrarily chosen standard probability measure μ over ε . This direction is novel and has not been studied in McFadden and Train (2000). Again, this result holds for the broad class of random-coefficient models. For example, the approximation is impossible using any mixed-logit model or any random-coefficient probit model. In Propositions 1 and 2 in Section 3.1, we give examples of the random utility models that cannot be approximated arbitrarily well.

The affine-independence condition can be simplified further to a generically equivalent condition. To see this, remember the following basic facts: (i) if $|J| > K + 1$, then $\{x_j \in \mathbf{R}^K | j \in J\}$ is not affinely independent; (ii) if $|J| \leq K + 1$, then the set is generically affinely independent.⁹ Given these observations, Theorem 1 implies the following corollary.¹⁰

Corollary 1. *Let K be the number of explanatory variables and $|J|$ be the number of alternatives.*

- (i) *If $K \geq |J| - 1$, then the statements in Theorem 1 (i) hold generically.*
- (ii) *If $K < |J| - 1$, then the statements in Theorem 1 (ii) hold.*

We now mention a few remarks on the results. First, as mentioned earlier, to increase the number K of explanatory variables, researchers may include additional terms such as higher order polynomials (McFadden and Train (2000)) as well as splines or wavelets (Chen, 2007). In their proof, McFadden and Train (2000) use higher order polynomials of arbitrarily high degrees to approximate continuous random utility models. In particular, in their construction, x_j is a vector of monomials of any degree of original characteristics $(y_{j1}, \dots, y_{jn})$:

$$x_j = (\underbrace{y_{j1}, \dots, y_{jn}}_{\text{1st order terms}}, \underbrace{y_{j1}^2, \dots, y_{jn}^2}_{\text{2nd order terms including interactions}}, y_{j1}^3, \dots, y_{jn}^3, \dots) \in \mathbf{R}^K,$$

where $K \rightarrow \infty$. Their result is thus consistent with the sufficiency part of our result: we proved

⁹This is a standard concept of genericity in the literature of discrete geometry. Even if the set is not affinely independent, as long as $|J| \leq K + 1$, for any $\varepsilon > 0$, there exists an affinely independent set X' , obtained from X by moving each point by a distance of at most ε (see Section 3 of Matousek (2013)).

¹⁰One caveat of the result is that even though the generic condition holds, the original condition of the affine independence may not hold when explanatory variables include zeros and ones. In that case, one should check the affine-independence of $\{x_j \in \mathbf{R}^K | j \in J\}$, rather than the generic condition.

that $K \geq |J| - 1$ is sufficient in our setup.¹¹

Second, in the theorem, we consider all possible random utility models (i.e., probability distributions over all rankings) as the prediction target mainly for simplicity. As mentioned after Definition 1, in some cases, the researchers may want to restrict the set of rankings by excluding those deemed unreasonable. In Section A.2 of the online appendix, we provide a necessary and sufficient condition for the approximation of the restricted random utility model.

Third, many empirical papers use the mixed-logit models that are linear in original characteristics and do not contain additional terms such as polynomials. The condition that $K \geq |J| - 1$ is frequently violated in various contexts, which in turn results in a violation of the affine-independence condition. Remember that $|J|$ is the number of alternatives and K is the number of characteristics. There are many choice situations in which $|J|$ is very large such as choices of groceries, hospitals, cars, schools, and restaurants. In such a dataset, the condition is likely to be violated. This means that the class of the models may not be rich enough to approximate the true substitution pattern across choice sets, no matter how one chooses parameters and fixed effects.¹² Theorem 1 considers the case of approximation on a rich collection of choice sets. Less stringent conditions are required for approximation if one considers the case of a single choice set. See Section 3.3 below.

3.1 Unrepresentable Rankings and Approximation Failure

Theorem 1 shows that when the affine-independence condition fails, there exist random utility models that cannot be approximated arbitrarily well by random-coefficient ARUMs. A key step toward this result is to identify deterministic rankings that ARUMs without fixed effects may fail to approximate well. In what follows, we provide a method for identifying such rankings. The following definition is central.

Definition 5. A ranking $\pi \in \Pi$ is *representable in choice sets* $\mathcal{D}$ if there exists a real vector β such that, for all $D \in \mathcal{D}$ and $j \in D$,

$$\pi(j) > \pi(l), \forall l \in D \setminus \{j\} \text{ if and only if } \beta \cdot x_j > \beta \cdot x_l, \forall l \in D \setminus \{j\}. \quad (4)$$

If π is not representable in $\mathcal{D}$, we say that π is *unrepresentable in* $\mathcal{D}$.

To determine the representability of a specific ranking π , one can employ linear programming techniques.¹³ In Section 5, utilizing an empirical dataset, we identify such unrepresentable

¹¹It should be noted that our results are not suggesting the imprudent use of a large number of characteristics in practice. In empirical applications for predictions, one must account for the statistical trade-off between bias and variance when comparing models with different numbers of characteristics. Some model selection procedures (e.g., AIC, BIC, or cross-validation) or penalization methods may be required.

¹²By substitution patterns, in general, we mean how choice probabilities change in different choice sets. In Section 5.2, we provide a more specific definition of substitution patterns.

¹³The ranking $\pi(1) > \pi(2) > \dots > \pi(J)$ is representable if and only if the linear programming problem

rankings, as detailed in Table 1 of that section. The identification of these rankings enables researchers to evaluate which substitution patterns are challenging to capture within their models.

Notice that the requirement (4) depends both on the specification of the characteristic vectors $\{x_j\}_{j \in J}$ as well as the set $\mathcal{D}$ of choice sets. As mentioned earlier, we usually consider the general choice sets $\mathcal{D}$ with binary and ternary choice sets, while in some places, we assume a simpler case in which $\mathcal{D} = \{J\}$. In the rest of this section, we consider the general $\mathcal{D}$. When there is no risk of confusion, we will simply say that a ranking $\pi \in \Pi$ is representable without specifying choice sets $\mathcal{D}$.

The *reverse* of a ranking π plays an important role:

Definition 6. For any ranking $\pi \in \Pi$, define $\pi^- \in \Pi$ such that $\pi(j) > \pi(l)$ if and only if $\pi^-(l) > \pi^-(j)$ for any $j, l \in J$. The ranking π^- is called the *reverse* ranking of π .

Note that if a ranking π is representable, then π^- is also representable.

The proposition below demonstrates that the failure of the affine-independence condition leads to the existence of unrepresentable rankings. To show the proposition, for each ranking $\pi \in \Pi$, define

$$\rho^\pi(D, j) = 1\{\pi(j) > \pi(l) \text{ for all } l \in D \setminus \{j\}\}. \quad (5)$$

The function ρ^π gives probability one to the best alternative j in a choice set D according to the strict ranking π .

Proposition 1. *The set $\{x_j \in \mathbf{R}^K | j \in J\}$ is not affinely independent if and only if there exists a ranking $\pi \in \Pi$ that is not representable in $\mathcal{D}$. For every unrepresentable ranking π and any standard probability measure μ , there exists $\sigma \in \{\pi, \pi^-\}$ and a neighborhood U of ρ^σ such that any random utility model in U cannot be approximated arbitrarily well by any sequence of random-coefficient ARUMs without fixed effects and with the same shock distribution μ .*

Remark 1. *A careful reader may wonder why Proposition 1 considers both an unrepresentable ranking π and its reverse ranking π^- . The reason is that, for a suitably chosen shock distribution, an ARUM without fixed effects can approximate ρ^π arbitrarily well even when π itself is not representable by the characteristic vectors. To see this, suppose that the shock distribution μ satisfies, for every pair $j, l \in J$, $\varepsilon_j > \varepsilon_l$ a.s. if and only if $\pi(j) > \pi(l)$. Fix any coefficient vector $\beta \in \mathbf{R}^K$, and consider the sequence of ARUMs with coefficient vector β/n and shock vector $(\varepsilon_j)_{j \in J}$. As $n \rightarrow \infty$, the systematic utility term $(\beta/n) \cdot x_j$ vanishes for each alternative j . Hence the induced choice probabilities converge to those generated solely by the shock ranking. Therefore, $\rho_n \rightarrow \rho^\pi$ even if π is not representable by the characteristic vectors. This observation*

$\max_{\beta \in \mathbf{R}^K, t \in \mathbf{R}} t$ subject to $\beta \cdot (x_j - x_{j+1}) \geq t$ for each $j = 1, \dots, |J| - 1$ is unbounded. If the ranking is unrepresentable, the problem has the optimal value 0.

explains why the conclusion must be formulated as a joint statement involving ρ^π and ρ^{π^-} , rather than as a statement about ρ^π alone. Such a construction is ruled out under either of the following additional assumptions: (a) the one-dimensional marginal distributions of the shocks are identical across alternatives; or (b) the zero vector belongs to the interior of $\text{supp } \mu$, as in Proposition 3(ii)(a). Under either assumption, if π is not representable, then ρ^π cannot be approximated arbitrarily well.

Remember that we have assumed no fixed effects (i.e., $\eta = 0$) in this proposition. In the next proposition, we consider the case with fixed effects. The next proposition shows that when the affine-independence condition fails, approximating a mixture of ρ^π and ρ^{π^-} is impossible even using fixed effects when π is not representable.

Proposition 2. *Suppose that $\{x_j \in \mathbf{R}^K | j \in J\}$ is not affinely independent. Then there exists an unrepresentable ranking π . For any unrepresentable ranking $\pi \in \Pi$, any standard probability measure μ , and any $\alpha \in (0, 1)$, there exists a neighborhood U of $\alpha\rho^\pi + (1 - \alpha)\rho^{\pi^-}$ such that any random utility model that belongs to U cannot be approximated arbitrarily well by any sequence of random-coefficient ARUMs with fixed effects and with the same shock distribution μ .*

3.2 Menu-dependent Fixed Effects

In the sections above, we assume menu-independent fixed effects for simplicity. All results hold with menu-dependent fixed effects with the following condition. For all $(D, j) \in \mathcal{D} \times J$ such that $j \in D$, remember that $\eta_{(D,j)} \in \mathbf{R}$ denotes a fixed effect of alternative j in a choice set D . We allow $\eta_{(D,j)}$ to be different across D but impose a local restriction: there are three alternatives $j, l, r \in J$ such that for all $\eta \in \mathcal{F}$, $\eta_{(D,j)}$, $\eta_{(D,l)}$, and $\eta_{(D,r)}$ do not depend on D if $D \subseteq \{j, l, r\}$. Formally, there exist $j, l, r \in J$ such that for all $\eta \in \mathcal{F}$,

$$\eta_{(D,s)} = \eta_{(E,s)} \text{ for all } D, E \subseteq \{j, l, r\} \text{ and all } s \in D \cap E. \quad (6)$$

Note that the restriction only requires that there exist three alternatives for which (6) holds; it does *not* require (6) to hold for all triples of alternatives. This restriction is a minimal restriction that allows us to study substitution patterns within the triple $\{j, l, r\}$ separately from the menu effects by comparing $\{j, l, r\}$ to the binary sets $\{j, l\}$, $\{j, r\}$, and $\{l, r\}$. Theorem 1, Proposition 1, and Proposition 2 remain valid with menu-dependent fixed effects as long as restriction (6) holds, replacing η_j with $\eta_{(D,j)}$ in the representations (1) and (2) in Definition 3.¹⁴

In Section A.1 of the online appendix, we also analyze the case in which η is fully menu-dependent without any restriction. Building on the surjectivity result of Norets and Takahashi (2013), we show that fully menu-dependent fixed effects are rich enough to approximate any

¹⁴The proofs in the appendix only assume (6), not menu-independence.

choice behavior, including non-RUM behavior, even without random coefficients. Intuitively, the fixed effects can absorb all variation in choice probabilities.

In practice, however, when the model is used for counterfactual prediction, the fixed effects for unobserved choice sets must be chosen a priori by the researcher—either by reusing fixed effects from some observed choice sets or by setting them to constants. In this case, our theorem implies that a random-coefficient ARUM with such a priori fixed effects cannot approximate arbitrary random utility models across choice sets unless our conditions are satisfied. For example, suppose we observe data only for the full menu J , but wish to make counterfactual predictions for smaller choice sets—such as all binary and ternary choice sets—to learn about substitution patterns. Our theorem implies that if the fixed effects for the counterfactual choice sets are chosen a priori—or set equal to the fixed effects from some observed choice set—then a random-coefficient ARUM cannot approximate arbitrary random utility models unless the affine-independence condition is satisfied. Thus, our main result remains directly relevant to counterfactual analysis.

3.3 Additional Result for Single Choice Set Case

In the following, we provide a supplemental result for the case in which $\mathcal{D} = \{J\}$. Such a case corresponds to a situation in which the researcher is interested only in the observed choice probabilities (i.e., market shares) on a single set J (but not on its subsets).

Proposition 3. *Assume that $\mathcal{D} = \{J\}$.*

- (i) *Let $\mu \in \mathcal{M}$ be any standard probability measure. If the set $\{x_j \in \mathbf{R}^K | j \in J\}$ is convex-independent (i.e., if $x_j \notin \text{co}\{x_l | l \in J \setminus \{j\}\}$ for any $j \in J$), then any random utility model can be approximated arbitrarily well by a sequence of random-coefficient ARUMs with the same shock distribution μ ; moreover, the approximation can be done without fixed effects (i.e., $\eta = 0$). That is,*

$$\forall \mu \in \mathcal{M}, \forall \rho \in \mathcal{P}_r, \exists \{\rho_n\}_{n=1}^\infty \subset \mathcal{P}_{ra}(0|\mu), \forall j \in J, \lim_{n \rightarrow \infty} \rho_n(J, j) = \rho(J, j).$$

- (ii) (a) *If the set $\{x_j \in \mathbf{R}^K | j \in J\}$ is not convex-independent, then there exists a random utility model ρ such that for every standard probability measure μ with $\mathbf{0} \in \text{int}(\text{supp } \mu)$, ρ cannot be approximated arbitrarily well by any sequence of random-coefficient ARUMs without fixed effects (i.e., with $\eta = 0$) and with the same shock distribution μ . Formally, $\exists \rho \in \mathcal{P}_r, \forall \mu \in \mathcal{M}$ s.t. $\mathbf{0} \in \text{int}(\text{supp } \mu)$, $\rho \notin \text{cl } \mathcal{P}_{ra}(0|\mu)$.*
- (b) *However, if fixed effects are used, any random utility model can be approximated arbitrarily well by a sequence of ARUMs with any shock distribution $\mu \in \mathcal{M}$. Formally, $\forall \rho \in \mathcal{P}_r, \forall \mu \in \mathcal{M}, \rho \in \text{cl } \bigcup_{\eta \in \mathcal{F}} \mathcal{P}_{ra}(\eta|\mu)$.*

Note that the convex-independence condition is weaker than the affine-independence condition.

tion. This makes sense because the convex-independence condition guarantees the approximation only on the single choice set (i.e., $\{J\}$), while the affine-independence condition guarantees the approximation across all subsets $D \in \mathcal{D}$ of J (including J itself).

The implications of Theorem 1 and Proposition 3 are similar. In both cases, there exists a random utility model that cannot be approximated arbitrarily well *without* using fixed effects. However, there are two important differences. First, Proposition 3(ii)(a) restricts attention to standard probability measures with $\mathbf{0} \in \text{int}(\text{supp } \mu)$, whereas Theorem 1 allows arbitrary standard probability measures.¹⁵ This restriction is necessary in the single-choice-set case because the shock distribution can otherwise be chosen to fit the target choice probabilities directly: for any given $\rho \in \mathcal{P}_r$, one can choose $\mu \in \mathcal{M}$ so that an ARUM without fixed effects matches ρ , by constructions of the kind described in Remark 1.

Second, as stated in Proposition 3 (ii)(b), if fixed effects are used, any random utility model can be approximated arbitrarily well on one particular choice set J . This result directly follows from Norets and Takahashi (2013). This is in contrast to Theorem 1 (ii), which claims that there exists a random utility model that cannot be approximated arbitrarily well even using fixed effects across choice sets $\mathcal{D}$.¹⁶

Unlike affine-independence, the convex-independence does not restrict the number of elements in a convex-independent set.¹⁷ So there exists no counterpart of Corollary 1.

4 Measuring Approximation Errors

Propositions 1 and 2 in Section 3.1 show that the approximation errors to some random utility models may not be negligible when the affine-independence condition fails. In this section, we provide a way to quantify the approximation errors. We first define the distance function as follows: for any $\hat{\rho}, \rho \in \mathcal{P}_r$, define $d(\hat{\rho}, \rho) \equiv \sqrt{(\sum_{D \in \mathcal{D}} \sum_{j \in D} (\rho(D, j) - \hat{\rho}(D, j))^2) / |\mathcal{D}|}$.

In our analysis, $\hat{\rho}$ is a given random utility model and ρ is a random-coefficient ARUM by which we approximate $\hat{\rho}$. We divide the norm by $\sqrt{|\mathcal{D}|}$ to make the distance independent from the number of choice sets.¹⁸ Notice that the maximal distance is at most $\sqrt{2}$.

¹⁵This assumption can be weakened to require only that $a\mathbf{1} \in \text{int}(\text{supp } \mu)$ for some $a \in \mathbf{R}$ because only the differences $\varepsilon_j - \varepsilon_l$ matter. For simplicity, we state the result under the normalization $a = 0$. Another possible sufficient condition is that the marginal distribution of ε_j is the same for every $j \in J$.

¹⁶This difference originates from the fact that we require approximation on $\mathcal{D}$ in Theorem 1 which contains binary and ternary choice sets, while in Proposition 3, we require approximation only on J .

¹⁷For example, in three-dimensional space (x, y, z) , consider a circumference of radius one whose origin is $(0, 0, 1)$ on a hyperplane of $z = 1$. The number of points on the circumference is a continuum. However, the set of points on the circumference is convex-independent.

¹⁸ $d(\hat{\rho}, \rho)$ can be written as $\|\rho - \hat{\rho}\| / \sqrt{|\mathcal{D}|}$, where $\|\cdot\|$ is the Euclidean norm (i.e., l_2 norm). One can consider an alternative distance function based on l_1 norm as follows: $d_1(\hat{\rho}, \rho) \equiv (\sum_{(j,D) \in J \times \mathcal{D}} |\hat{\rho}(D, j) - \rho(D, j)|) / |\mathcal{D}|$. Since $\sqrt{\sum_{(j,D) \in J \times \mathcal{D}} (\hat{\rho}(D, j) - \rho(D, j))^2} \leq \sum_{(j,D) \in J \times \mathcal{D}} |\hat{\rho}(D, j) - \rho(D, j)|$, we can show that $d(\hat{\rho}, \rho) / \sqrt{|\mathcal{D}|} \leq d_1(\hat{\rho}, \rho)$ for any ρ and $\hat{\rho}$. So our approximation error divided by $\sqrt{|\mathcal{D}|}$ will provide a lower bound of an alternative approximation error measured by d_1 . We use our distance function d rather than d_1 because the inner-product structure of d is

Given an approximation target $\hat{\rho} \in \mathcal{P}_r$ and a standard probability measure μ , when researchers use random-coefficient ARUMs with fixed effects η , the approximation error is defined as:

$$\inf_{\rho \in \mathcal{P}_{ra}(\eta|\mu)} d(\hat{\rho}, \rho). \quad (7)$$

We call (7) the *approximation error to $\hat{\rho}$ by random-coefficient ARUMs with fixed effects η* .

Given $\hat{\rho}$, we propose two algorithms to solve (7) and measure the approximation errors. The first is the standard EM (Expectation-Maximization) algorithm (Dempster, Laird, and Rubin, 1977) to estimate the best possible finite-mixture logit model. However, the EM algorithm may not converge to the global optimum in our setting. To alleviate this concern, we propose a greedy algorithm inspired by Barron, Cohen, Dahmen, and DeVore (2008). This algorithm solves a sequence of optimization problems to converge to the global optimal solution. We provide explanations of these algorithms in Section G of the online appendix. In Section 5 below, we find that the approximation error can be large for some unrepresentable rankings in a real dataset.

5 Application to Data

To quantify approximation errors with and without fixed effects, we use the DVD sales dataset introduced by Rusmevichientong, Shen, and Shmoys (2010).¹⁹ This dataset has been widely used in research on consumer choice modeling and assortment optimization by, for example, Rusmevichientong, Shen, and Shmoys (2010); Farias, Jagabathula, and Shah (2013) and Abdallah and Vulcano (2021). We obtained the dataset from Table 2 of Farias, Jagabathula, and Shah (2009) and replicate it in Appendix J.2. The dataset contains 15 DVD products. As noted by Rusmevichientong, Shen, and Shmoys (2010), the most predictive attributes for choices are the total number of helpful review votes and the price per disc. Motivated by this fact and to keep our computation manageable, we focus on these two characteristics and 4 DVDs to illustrate our theoretical results.²⁰ Hence we have $|J| = 4$ and $K = 2$. In Appendix Section J.1, we report the approximation errors without fixed effects for a larger number of alternatives with $|J| = 10$ and $K = 3$. The findings are similar.

Our empirical analysis concentrates primarily on mixed-logit models given their widespread use. We consider two specifications of mixed-logit models. The first is the linear mixed-logit models that are linear in the original characteristics. The second is the mixed-logit model

useful for constructing an algorithm.

¹⁹In Section K of the online appendix, we measure approximation errors using an additional fishing-site choice dataset from Thomson and Crooke (1991), illustrating that our methods apply more generally and are not specific to this particular dataset.

²⁰We select the four DVDs from the first four rows of Table 2 in Farias, Jagabathula, and Shah (2009), which correspond to the four most expensive alternatives.

defined with quadratic polynomials.²¹ We call the model *quadratic mixed-logit model*. If we use the linear mixed-logit models, then $K = 2$ and the condition in Corollary 1 is violated (i.e., $K = 2 \not\geq 3 = |J| - 1$) and the set of characteristics is not affinely independent. Thus, by Theorem 1, the linear mixed-logit models with fixed effects are not flexible enough to approximate some random utility models. On the other hand, with quadratic polynomials, the generic condition for representability in Corollary 1 is satisfied, since $K = 5 \geq 3 = |J| - 1$ and the characteristics are affinely independent. Thus, by Theorem 1, the quadratic mixed-logit models are flexible enough to approximate any random utility model. These theoretical implications are numerically verified and quantified below, where we report approximation errors for the linear and quadratic logit models with and without fixed effects.

5.1 Approximation Errors with and without Fixed Effects

We assume that $\mathcal{D} = 2^J \setminus \{\emptyset\}$. We say that a ranking π is *linearly representable* if π is representable in $\mathcal{D}$ with the original characteristics and π is *linearly unrepresentable* if π is not linearly representable. With $|J| = 4$, there are 24 possible rankings and we find that 12 rankings are not linearly representable. Table 1 below reports the approximation errors for rankings that are not linearly representable, calculated using both the Greedy and EM algorithms. (In the table, we label the four DVDs by the indices 1, 2, 3, and 4.) Approximation errors for deterministic rankings are of particular interest, because the worst-case approximation error over all stochastic choice rules is attained at such a ranking.²²

As implied by Theorem 1, quadratic mixed-logit models yield zero approximation errors for all rankings and linear mixed-logit models yield zero errors for linearly representable rankings. The approximation errors of linear mixed-logit models for unrepresentable rankings π are almost always larger than 0.4. Thus, even the best-fitting linear mixed-logit model remains far from the corresponding deterministic choice rule ρ^π in the normalized Euclidean metric. Some errors are substantially larger. For example, the approximation error for the ranking $\pi(2) > \pi(4) > \pi(3) > \pi(1)$ is more than 0.685.

Table 2 reports the approximation errors *with fixed effects* to a half-half mixture, $\frac{1}{2}\rho^\pi + \frac{1}{2}\rho^{\pi^-}$, for each unrepresentable ranking π . The random utility model $\frac{1}{2}\rho^\pi + \frac{1}{2}\rho^{\pi^-}$ can be interpreted as the aggregate behavior of a population in which half of consumers have ranking π and the other half have ranking π^- .

Proposition 2 implies that such random utility models cannot be approximated arbitrarily well by any sequence of linear mixed-logit models with fixed effects. In both algorithms, the approximation errors to $\frac{1}{2}\rho^\pi + \frac{1}{2}\rho^{\pi^-}$ are around 0.2 or above whenever π is not representable. Thus, even after allowing fixed effects, the best-fitting linear mixed-logit model remains bounded

²¹That is, in addition to the original characteristics, we include their squared terms and interactions.

²²A proof of this statement is provided in Section F of the online appendix.

Table 1: Approximation errors to ρ^π

Ranking π	Linear mixed-logit		Quadratic mixed-logit	
	Greedy (1)	EM (2)	Greedy (3)	EM (4)
Linearly Unrepresentable Rankings				
$\pi(1) > \pi(3) > \pi(4) > \pi(2)$	0.649	0.649	0.000	0.000
$\pi(1) > \pi(3) > \pi(2) > \pi(4)$	0.587	0.603	0.000	0.000
$\pi(1) > \pi(4) > \pi(3) > \pi(2)$	0.437	0.445	0.000	0.000
$\pi(2) > \pi(4) > \pi(3) > \pi(1)$	0.685	0.691	0.000	0.000
$\pi(2) > \pi(4) > \pi(1) > \pi(3)$	0.674	0.685	0.000	0.000
$\pi(2) > \pi(3) > \pi(4) > \pi(1)$	0.436	0.445	0.000	0.000
$\pi(2) > \pi(1) > \pi(4) > \pi(3)$	0.422	0.431	0.000	0.000
$\pi(3) > \pi(1) > \pi(4) > \pi(2)$	0.596	0.642	0.000	0.000
$\pi(3) > \pi(1) > \pi(2) > \pi(4)$	0.493	0.522	0.000	0.000
$\pi(3) > \pi(4) > \pi(1) > \pi(2)$	0.388	0.404	0.000	0.000
$\pi(4) > \pi(2) > \pi(3) > \pi(1)$	0.511	0.566	0.000	0.000
$\pi(4) > \pi(2) > \pi(1) > \pi(3)$	0.455	0.490	0.000	0.000
Linearly Representable Rankings	0.000	0.000	0.000	0.000

Note: The numbers in the table show the approximation errors without fixed effects for each ρ^π , where each ranking π is defined in the leftmost column. For each ranking, columns (1) and (2) show the approximation errors of the linear mixed-logit models computed by the greedy algorithm and the EM algorithm, respectively. We choose 18 mixtures for the EM algorithm, as implied by Propositions 6 and 7 in Section I because $(\dim \mathcal{P}_r) + 1 = 1 + \sum_{D \in \mathcal{D}} (|D| - 1) = 18$. Columns (3) and (4) show the approximation errors of the quadratic mixed-logit models calculated by each algorithm. All numbers are rounded to three decimal places. For the greedy algorithm we set the number of iterations to 1000. For the EM algorithm we set the number of random initial points to 10.

away from $\frac{1}{2}\rho^\pi + \frac{1}{2}\rho^{\pi^-}$ in the normalized Euclidean metric. By contrast, the approximation errors for $\frac{1}{2}\rho^\pi + \frac{1}{2}\rho^{\pi^-}$ are essentially zero when π is representable, as predicted by Theorem 1.

5.2 Maximal Substitution

This section quantifies the flexibility of the linear mixed-logit models by measuring the *maximal substitution patterns* that can be generated by these models. Specifically, for two alternatives j and l , we define the maximal substitution pattern for mixed-logit models as

$$\sup_{\rho \in \mathcal{P}_{ml}(0)} \left(\rho(J \setminus \{j\}, l) - \rho(J, l) \right). \quad (8)$$

The quantity $\rho(J \setminus \{j\}, l) - \rho(J, l)$ describes how many consumers would substitute to alternative l if alternative j becomes unavailable. The supremum of such quantities captures the maximal substitution pattern that can be generated by mixed-logit models without fixed effects.²³ Since

²³Conlon and Mortimer (2021) analyze similar concepts called *diversion ratios*, which measure the fraction of consumers who switch their choices from alternative j to l after the price of alternative j increases marginally. One quantity highlighted by Conlon and Mortimer (2021) (p. 701) is a functional of diversion ratios that characterizes substitution patterns when a product becomes unavailable. Our measure corresponds to the limit of the numerator of that quantity. Some recent papers study substitution and complementarity properties in the discrete choice analysis. See Horan and Adam (2023) and Allen and Rehbeck (2020).

Table 2: Approximation errors to random utility models $\frac{1}{2}\rho^\pi + \frac{1}{2}\rho^{\pi^-}$

Ranking π	Linear mixed-logit		Quadratic mixed-logit	
	Greedy (1)	EM (2)	Greedy (3)	EM (4)
Linearly unrepresentable rankings				
$\pi(1) > \pi(3) > \pi(4) > \pi(2)$	0.265	0.300	0.000	0.000
$\pi(1) > \pi(3) > \pi(2) > \pi(4)$	0.225	0.287	0.000	0.000
$\pi(1) > \pi(4) > \pi(3) > \pi(2)$	0.191	0.243	0.000	0.000
$\pi(2) > \pi(4) > \pi(1) > \pi(3)$	0.191	0.243	0.000	0.000
$\pi(2) > \pi(1) > \pi(4) > \pi(3)$	0.191	0.206	0.000	0.000
$\pi(3) > \pi(1) > \pi(2) > \pi(4)$	0.191	0.206	0.000	0.000
Linearly representable rankings	0.000	0.000	0.000	0.000

Note: The numbers in the table show the approximation errors to $\frac{1}{2}\rho^\pi + \frac{1}{2}\rho^{\pi^-}$, where π is defined in the leftmost column. All numbers are rounded to three decimal places. For the greedy algorithm we set the number of iterations to 1000. For the EM algorithm we set the number of random initial points to 10. We optimize over fixed effects from -10 to 10 with a step size of 1. Therefore, the reported errors with fixed effects should be interpreted as grid-search approximation errors; the true infimum over unrestricted fixed effects can only be weakly smaller.

ρ is a random utility model, we always have $\rho(J \setminus \{j\}, l) \geq \rho(J, l)$. Thus the quantity in (8) takes values in $[0, 1]$. We use the greedy algorithm to solve (8), as detailed in Section G.3 of the online appendix.²⁴

Table 3: Maximal substitution of the linear mixed-logit models

j (drop) \ l (choose)	1	2	3	4
1	-	0.999	0.138	0.999
2	1.000	-	1.000	0.111
3	0.250	1.000	-	1.000
4	1.000	0.303	1.000	-

Note: The numbers in the table show the value of (8) for each $j, l \in \{1, 2, 3, 4\}$ s.t. $j \neq l$. All numbers are rounded to three decimal places.

Table 3 reports the maximal substitutions for each pair of alternatives. It is evident that the linear mixed-logit model has difficulty capturing the substitution behavior between products 1 and 3 and between products 2 and 4. In particular, the maximal substitution from product 2 to product 4 is only 0.111. This finding aligns with the result presented in Table 1 as we observe substantial approximation errors for the two specific rankings: $\pi(2) > \pi(4) > \pi(3) > \pi(1)$ and $\pi(2) > \pi(4) > \pi(1) > \pi(3)$. These are the only rankings in which 4 replaces 2 when 2 becomes unavailable.

²⁴The quantity in (8) is always 1 when we can choose fixed effects freely. This is because we can always choose fixed effects large enough to approximate degenerate preferences. We do not use the EM algorithm as it cannot be easily transformed to solve the problem in (8). We can replace the sup in (8) with inf, for which we calculate the *minimal substitution pattern*. The minimal substitution pattern is 0, since the dataset satisfies the convex independence.

The fact that linear mixed-logit choice models fail to approximate substitution patterns may have important implications for the classical *assortment planning problem* in operations management. See K  k, Fisher, and Vaidyanathan (2008) for a survey on the topic. For example, consider a simple example where there are three alternatives (labeled 1, 2 and 3) and a no-purchase option (labeled 4). Suppose the population consists of three groups; 50% of customers have preferences $\pi(3) > \pi(4) > \pi(1) > \pi(2)$, 25% have preferences $\pi(1) > \pi(3) > \pi(2) > \pi(4)$; and the remaining 25% have preferences $\pi(2) > \pi(4) > \pi(3) > \pi(1)$. Assume that product 3 yields the highest profit, followed by products 1 and 2. The store initially offers products 1, 2, and 3, but a shelf-space constraint forces it to remove one product. Since product 3 is both the most popular and the most profitable, the choice is between removing product 1 or 2.

Suppose the store manager fits a linear mixed-logit model to historical sales data and uses this model to forecast the impact of dropping one product. If product 1 is removed, the model may predict that only a small fraction of customers (at most 13.8%, as in Table 3) switch from 1 to 3, with most reallocating from 1 to 2 or 4. If product 2 is removed, the model may predict that only a small fraction of customers (at most 11.1%, as in Table 3) switch from 2 to 4, with most reallocating from 2 to 1 or 3. On this basis, the manager might decide to remove product 2, hoping to shift demand toward the higher-margin products 1 and 3. In reality, however, one quarter of customers, those with preference $\pi(2) > \pi(4) > \pi(3) > \pi(1)$, will choose the no-purchase option 4. As a result, the store sacrifices revenue due to the poor substitution predictions of the linear mixed-logit model.

Appendix: Proofs

In Proposition 6 of Section I of the online appendix, we prove that $\mathcal{P}_{ra}(\eta|\mu) = \text{co } \mathcal{P}_a(\eta|\mu)$ for any fixed effects vector η and standard probability measure μ , where co denotes the convex hull. We use this result throughout the proof.

For all $(D, j) \in \mathcal{D} \times J$ such that $j \in D$, if $\pi(j) > \pi(l)$ for all $l \in D \setminus \{j\}$, then we write $j = \max_D \pi$. We use this notation throughout the proof.

A Proofs of Theorem 1 (i), Proposition 1 and 3

To prove Proposition 1, Proposition 3, and Theorem 1-(i), we will show several lemmas. We first consider models without fixed effects (i.e., $\eta = 0$). The following fact is elementary but fundamental:

Observation: *The set $\mathcal{P}_r$ of random utility models is a polytope, that is, $\mathcal{P}_r = \text{co } \{\rho^\pi | \pi \in \Pi\}$. For each $\pi \in \Pi$, ρ^π is a vertex of the polytope.*

The observation holds because for any random utility model $\rho \in \mathcal{P}_r$, we have $\rho = \sum_{\pi \in \Pi} \nu(\pi) \rho^\pi$, where ν is the probability measure on Π rationalizing the random utility model. See section

H.2 in the online appendix for details. The hexagons in Figure A.2 in the section illustrate the polytope.

Lemma 1 gives a necessary and sufficient condition under which any random utility models can be approximated arbitrarily well by a sequence of random-coefficient ARUMs without fixed effects.

Lemma 1. *Fix any $\mu \in \mathcal{M}$. $\mathcal{P}_r \subset \text{cl } \mathcal{P}_{ra}(0|\mu)$ if and only if $\rho^\pi \in \text{cl } \mathcal{P}_a(0|\mu)$ for any $\pi \in \Pi$.*

The next lemma makes it easier for us to check the conditions of Lemma 1.

Lemma 2. *For any ranking $\pi \in \Pi$, the following statements hold: (1) If π is representable in $\mathcal{D}$, then for any $\mu \in \mathcal{M}$, $\{\rho^\pi, \rho^{\pi^-}\} \subset \text{cl } \mathcal{P}_a(0|\mu)$; (2) If π is not representable in $\mathcal{D}$, then for any $\mu \in \mathcal{M}$, $\{\rho^\pi, \rho^{\pi^-}\} \not\subset \text{cl } \mathcal{P}_{ra}(0|\mu)$.*

As mentioned earlier, checking the representability of a particular ranking is easy. However, checking the representability of *all* rankings may be computationally demanding. This is because the number of rankings equals $|J|!$ and can be large. To overcome this problem, we obtain a simpler necessary and sufficient condition for all rankings to be representable:

Lemma 3. *(1) Every ranking is representable in $\mathcal{D}$ if and only if the set $\{x_j \in \mathbf{R}^K | j \in J\}$ is affinely independent; (2) Every ranking is representable in $\{J\}$ if and only if the set $\{x_j \in \mathbf{R}^K | j \in J\}$ is convex-independent.*

See Section H.1 in the online appendix for the illustration of Lemma 3.

A.1 Proof of Theorem 1 (i)

By Lemma 3, the affine independence of $\{x_j \in \mathbf{R}^K | j \in J\}$ implies that all rankings are representable and thus can be approximated by a sequence of ARUMs by Lemma 2-(1). Applying Lemma 1 proves Theorem 1-(i).

A.2 Proof of Proposition 1 and 3

We first prove Proposition 1. By Lemma 3-(1), the non-affine independence of $\{x_j | j \in J\}$ implies that some rankings are not representable. By Lemma 2-(2), for the given μ at least one element ρ^σ with $\sigma \in \{\pi, \pi^-\}$ is not in $\text{cl } \mathcal{P}_{ra}(0 | \mu)$. Since $\text{cl } \mathcal{P}_{ra}(0 | \mu)$ is closed, there exists an open neighborhood U of ρ^σ such that $U \cap \text{cl } \mathcal{P}_{ra}(0 | \mu) = \emptyset$. Hence, Proposition 1 holds.

Lemma 1, 2, and 3 (2) imply statement (i) of Proposition 3. The proof of statement (ii) is in Section C of the online appendix.

A.3 Proof of Lemma 1

We first prove the if direction. Suppose that $\rho^\pi \in \text{cl } \mathcal{P}_a(0|\mu)$ for all $\pi \in \Pi$. Then $\text{co } \{\rho^\pi | \pi \in \Pi\} \subset \text{co cl } \mathcal{P}_a(0|\mu)$. By the fact that $\mathcal{P}_a(0|\mu)$ is bounded, it follows from Theorem 17.2 of

Rockafellar (2015) that $\text{cl co } \mathcal{P}_a(0|\mu) = \text{co cl } \mathcal{P}_a(0|\mu)$. Thus, $\mathcal{P}_r \subset \text{cl co } \mathcal{P}_a(0|\mu) \subset \text{cl } \mathcal{P}_{ra}(0|\mu)$, where we used the fact that $\mathcal{P}_{ra}(0|\mu) = \text{co } \mathcal{P}_a(0|\mu)$ for any $\mu \in \mathcal{M}$.²⁵

We now prove the only-if direction. By assumption, $\mathcal{P}_r \subset \text{cl co } \mathcal{P}_a(0|\mu) = \text{co cl } \mathcal{P}_a(0|\mu)$, where the equality holds by the previous argument. Since $\mathcal{P}_r \subset \text{co cl } \mathcal{P}_a(0|\mu)$, for any $\pi \in \Pi$, there exist positive numbers $\{\lambda_i\}_{i=1}^m$ and $\rho^i \in \text{cl } \mathcal{P}_a(0|\mu)$ for all $i \in \{1, \dots, m\}$ such that $\sum_{i=1}^m \lambda_i = 1$ and $\sum_{i=1}^m \lambda_i \rho^i = \rho^\pi$. Note that $\rho^i \in \text{cl } \mathcal{P}_a(0|\mu) \subset \mathcal{P}_r$ for each i because $\mathcal{P}_r$ is closed. Since ρ^π is a vertex of $\mathcal{P}_r$, thus ρ^π is an exposed point.²⁶ Hence, $\sum_{i=1}^m \lambda_i \rho^i = \rho^\pi$ implies $\rho^i = \rho^\pi$ for all i . This means that $\rho^\pi \in \text{cl } \mathcal{P}_a(0|\mu)$ for all $\pi \in \Pi$.

A.4 Proof of Lemma 2

We prove Lemma 2-(1) in the next step.

Step 1: For any $\pi \in \Pi$ and any $\mu \in \mathcal{M}$, if a ranking π is representable, then there exists a sequence $\{\rho_n\}$ of $\mathcal{P}_a(0|\mu)$ such that $\rho_n \rightarrow \rho^\pi$.

Proof. Assume that a ranking π is representable. Fix $D \in \mathcal{D}$ and let $j = \max_D \pi$. Then $\beta \cdot x_j - \max_{l \in D \setminus \{j\}} \beta \cdot x_l > 0$. Let ρ_n be the sequence of ARUMs with coefficient $n\beta$. For any positive integer n and any $(D, j) \in \mathcal{D} \times J$ such that $j \in D$, note that $\rho_n(D, j) \geq \int 1\{n(\beta \cdot x_j - \max_{l \in D \setminus \{j\}} \beta \cdot x_l) \geq \max_{l \in D \setminus \{j\}} \varepsilon_l - \varepsilon_j\} d\mu$, where $1\{\cdot\}$ is the indicator function.²⁷ By the dominated convergence theorem, $\lim_{n \rightarrow \infty} \rho_n(D, j) \geq \int \lim_{n \rightarrow \infty} 1\{n(\beta \cdot x_j - \max_{l \in D \setminus \{j\}} \beta \cdot x_l) \geq \max_{l \in D \setminus \{j\}} \varepsilon_l - \varepsilon_j\} d\mu = 1$. Since choice frequencies must be nonnegative and sum up to one, $\lim_{n \rightarrow \infty} \rho_n(D, j) \geq 1$ implies $\rho_n \rightarrow \rho^\pi$. $\square$

We prove the contrapositive of Lemma 2-(2) in Step 2 and 3.

Step 2: If there exists a $\mu \in \mathcal{M}$ and sequences $\{\rho_n\}$ and $\{\rho'_n\}$ of $\mathcal{P}_a(0|\mu)$ such that $\rho_n \rightarrow \rho^\pi$ and $\rho'_n \rightarrow \rho^{\pi^-}$, then π and π^- are representable.

Proof. Suppose $\rho_n \in \mathcal{P}_a(0|\mu)$ and $\rho'_n \in \mathcal{P}_a(0|\mu)$ satisfy $\rho_n \rightarrow \rho^\pi$ and $\rho'_n \rightarrow \rho^{\pi^-}$. Let β_n and β'_n be the corresponding coefficient vectors. Fix a binary menu $\{j, l\}$ with $\pi(j) > \pi(l)$. Define $Z_{lj} = \varepsilon_l - \varepsilon_j$, $\gamma_n = \beta_n \cdot (x_j - x_l)$, and $\gamma'_n = \beta'_n \cdot (x_l - x_j)$. Since $\rho_n \rightarrow \rho^\pi$ and $\rho'_n \rightarrow \rho^{\pi^-}$, we have $\mu(\{\varepsilon | Z_{lj} < \gamma_n\}) \rightarrow 1$ and $\mu(\{\varepsilon | Z_{lj} > -\gamma'_n\}) \rightarrow 1$. If $\gamma_n + \gamma'_n \leq 0$ along an infinite subsequence, then along that subsequence $\{\varepsilon | Z_{lj} < \gamma_n\} \cap \{\varepsilon | Z_{lj} > -\gamma'_n\} = \emptyset$, because $\gamma_n \leq -\gamma'_n$. Therefore, $\mu(\{\varepsilon | Z_{lj} < \gamma_n\}) + \mu(\{\varepsilon | Z_{lj} > -\gamma'_n\}) \leq 1$, contradicting the fact that both probabilities converge to one. Hence, for this pair (j, l) , $\gamma_n + \gamma'_n = (\beta_n - \beta'_n) \cdot (x_j - x_l) > 0$ for all sufficiently large n .

Since there are finitely many pairs with $\pi(j) > \pi(l)$, there is a single N such that for all $n \geq N$,

²⁵See Proposition 6 of Section I of the online appendix.

²⁶A point of a convex set is an exposed point if there is a supporting hyperplane which contains no other points of the set (Rockafellar (2015), Page 162)

²⁷Note that $\rho_n(D, j) \equiv \mu(\{\varepsilon | n\beta \cdot x_j + \varepsilon_j \geq \max_{l \in D \setminus \{j\}} \{n\beta \cdot x_l + \varepsilon_l\}\}) \geq \mu(\{\varepsilon | n\beta \cdot x_j + \varepsilon_j \geq \max_{l \in D \setminus \{j\}} n\beta \cdot x_l + \max_{l \in D \setminus \{j\}} \varepsilon_l\}) = \mu(\{\varepsilon | n(\beta \cdot x_j - \max_{l \in D \setminus \{j\}} \beta \cdot x_l) \geq \max_{l \in D \setminus \{j\}} \varepsilon_l - \varepsilon_j\}) = \int 1\{n(\beta \cdot x_j - \max_{l \in D \setminus \{j\}} \beta \cdot x_l) \geq \max_{l \in D \setminus \{j\}} \varepsilon_l - \varepsilon_j\} d\mu$.

$(\beta_n - \beta'_n) \cdot (x_j - x_l) > 0$ for every pair j, l with $\pi(j) > \pi(l)$. Thus, $\beta_n - \beta'_n$ represents π and $\beta'_n - \beta_n$ represents π^- for all sufficiently large n . $\square$

In the following, we again use the fact that $\mathcal{P}_{ra}(0|\mu) = \text{co } \mathcal{P}_a(0|\mu)$ for any $\mu \in \mathcal{M}$.

Step 3: If there exists $\mu \in \mathcal{M}$ and a sequence $\{\rho_n\}$ of $\text{co } \mathcal{P}_a(0|\mu)$ such that $\rho_n \rightarrow \rho^\pi$, then there exists a sequence $\{\rho'_n\}$ of $\mathcal{P}_a(0|\mu)$ such that $\rho'_n \rightarrow \rho^\pi$.

Proof. Fix $\pi \in \Pi$. Suppose that there exists a sequence ρ_n of $\text{co } \mathcal{P}_a(0|\mu)$ such that $\rho_n \rightarrow \rho^\pi$ as $n \rightarrow \infty$. Let $M = \dim \text{co } \mathcal{P}_a(0|\mu)$. Then for each ρ_n , by Caratheodory's theorem, there exist $\{\rho_n^i\}_{i=1}^{M+1} \subset \mathcal{P}_a(0|\mu)$ and nonnegative numbers $\{\alpha_n^i\}_{i=1}^{M+1}$ such that $\rho_n = \sum_{i=1}^{M+1} \alpha_n^i \rho_n^i$ and $\sum_{i=1}^{M+1} \alpha_n^i = 1$. Denote $(\alpha_n^i)_{i=1}^{M+1}$ by α_n . Then α_n belongs to a compact set (i.e., M -dimensional simplex). There exists a convergent subsequence $\{\alpha_{n'}\}$. Thus $\rho'_{n'} \equiv \sum_{i=1}^{M+1} \alpha_{n'}^i \rho_{n'}^i$ is a subsequence of $\{\rho_n\}$. For each i , let α_*^i be the limit of $\{\alpha_{n'}^i\}$. Since $\sum_{i=1}^{M+1} \alpha_{n'}^i = 1$ for all n' , we have $\sum_{i=1}^{M+1} \alpha_*^i = 1$, so that there must exist i^* such that $\alpha_*^{i^*} \neq 0$.

In the following, we will show that $\rho_{n'}^{i^*} \rightarrow \rho^\pi$ as $n' \rightarrow \infty$. To show the claim, we prove that if $\rho_{n'}^{i^*} \not\rightarrow \rho^\pi$, then $\alpha_{n'}^{i^*} \rightarrow 0$, which is a contradiction. Assume that $\rho_{n'}^{i^*} \not\rightarrow \rho^\pi$. Then there exist $D \in \mathcal{D}$, $j \in D$, and $\varepsilon > 0$ such that for any integer N there exists $n' > N$ such that $|\rho_{n'}^{i^*}(D, j) - \rho^\pi(D, j)| > \varepsilon$. This implies that for any N there exists $n' > N$ such that $\left| \sum_{i=1}^{M+1} \alpha_{n'}^i \rho_{n'}^i(D, j) - \rho^\pi(D, j) \right| = \sum_{i=1}^{M+1} \alpha_{n'}^i |\rho_{n'}^i(D, j) - \rho^\pi(D, j)| \geq \alpha_{n'}^{i^*} \varepsilon$, where the first equality holds because if $j = \max_D \pi$ then $\rho_{n'}^i(D, j) - \rho^\pi(D, j) \leq 0$ for all i ; if $j \neq \max_D \pi$, then $\rho_{n'}^i(D, j) - \rho^\pi(D, j) \geq 0$ for all i . Since $\sum_{i=1}^{M+1} \alpha_{n'}^i \rho_{n'}^i(D, j) \rightarrow \rho^\pi(D, j)$, it must hold that $\alpha_{n'}^{i^*} \rightarrow 0$. $\square$

Step 2 and 3 show that for any $\mu \in \mathcal{M}$, if there exist sequences $\{\rho_n\}$ and $\{\rho'_n\}$ of $\text{co } \mathcal{P}_a(0|\mu)$ such that $\rho_n \rightarrow \rho^\pi$ and $\rho'_n \rightarrow \rho^{\pi^-}$, then π is representable. The contraposition of this statement is Lemma 2-(2).

A.5 Proof of Lemma 3

We first prove statement (1). For simplicity of notation, let $J = \{1, 2, \dots, |J|\}$. For any ranking $\pi \in \Pi$, by relabeling J if necessary, we assume that $\pi(i) > \pi(i+1)$ for all $i \leq |J|-1$ without loss of generality. We label the following condition as Condition (*): if $\lambda_1 x_1 + \sum_{i=2}^{|J|-1} (\lambda_i - \lambda_{i-1}) x_i - \lambda_{|J|-1} x_{|J|} = 0$ and $\lambda_i \geq 0$ for all $i \in \{1, \dots, |J|-1\}$, then $\lambda_i = 0$ for all $i \in \{1, \dots, |J|-1\}$.

Step 1: For each $\pi \in \Pi$, Condition (*) holds if and only if π is representable.

Proof. Since $\mathcal{D}$ contains all binary choice sets, π is representable if and only if there exists $\beta \in \mathbb{R}^K$ such that $\beta \cdot x_1 > \beta \cdot x_2 > \dots > \beta \cdot x_{|J|}$. Equivalently, $\beta \cdot (x_i - x_{i+1}) > 0$ for all $i = 1, \dots, |J|-1$. By the theorem of alternatives, this system of strict inequalities has a solution if and only if there is no nonzero vector $(\lambda_i)_{i=1}^{|J|-1}$ with $\lambda_i \geq 0$ for all i such that $\sum_{i=1}^{|J|-1} \lambda_i (x_i - x_{i+1}) = 0$. See Section D in the online appendix for the detail. Expanding the last display gives

exactly Condition (*). $\square$

Step 2: The set $\{x_j | j \in J\}$ is affinely independent if and only if Condition (*) holds for any $\pi \in \Pi$.

Proof. First suppose that $\{x_j : j \in J\}$ is affinely independent. If the equality in Condition (*) holds, define $\mu_1 = \lambda_1$, $\mu_i = \lambda_i - \lambda_{i-1}$ ($i = 2, \dots, |J| - 1$), $\mu_{|J|} = -\lambda_{|J|-1}$. Then $\sum_{i=1}^{|J|} \mu_i x_i = 0$ and $\sum_{i=1}^{|J|} \mu_i = 0$. Affine independence implies $\mu_i = 0$ for all i , and hence $\lambda_i = 0$ for all i . Thus Condition (*) holds. Conversely, suppose Condition (*) holds for every ranking. Take any $(\mu_i)_{i=1}^{|J|}$ such that $\sum_{i=1}^{|J|} \mu_i x_i = 0$ and $\sum_{i=1}^{|J|} \mu_i = 0$. Relabel alternatives so that $\mu_1 \geq \mu_2 \geq \dots \geq \mu_{|J|}$, and define $\lambda_i = \sum_{k=1}^i \mu_k$ for all $i \in \{1, \dots, |J| - 1\}$. Then $\lambda_i \geq 0$ for all i , and $\sum_{i=1}^{|J|-1} \lambda_i (x_i - x_{i+1}) = 0$. By Condition (*), $\lambda_i = 0$ for all i , and therefore $\mu_i = 0$ for all i . This proves affine independence. Combining Steps 1 and 2 proves Lemma 3-(1). $\square$

The proof of statement (2) is straightforward and is relegated to Section D of the online appendix.

B Proof of Theorem 1-(ii) and Proposition 2

As mentioned in Section 3.2, the main results hold with menu-dependent fixed effects with a mild restriction. We prove Theorem 1-(ii) and Proposition 2 with general $\mathcal{F}$ defined by (6).

To prove Theorem 1-(ii) and Proposition 2, we first prove the following two statements: (a) for any ranking π , any strict convex combination between ρ^π and ρ^{π^-} cannot be approximated arbitrarily well by any sequence of ARUMs with fixed effects; and (b) moreover, if π is not representable, then any strict convex combination between ρ^π and ρ^{π^-} cannot be approximated arbitrarily well by any sequence of random-coefficient ARUMs even with fixed effects. To prove statement (a), we prove a stronger result:

Lemma 4. *Fix three alternatives j, l, r such that for any $\eta \in \mathcal{F}$, $\eta_{(D,s)} = \eta_{(E,s)}$ for all $D, E \subseteq \{j, l, r\}$ and $s \in D \cap E$. Fix two rankings π and π' such that $\pi(j) > \pi(l) > \pi(r)$ and $\pi'(r) > \pi'(l) > \pi'(j)$.²⁸ Let $\mu \in \mathcal{M}$ be a standard probability measure. For any $\alpha \in (0, 1)$, a random utility model $\rho^\alpha \equiv \alpha \rho^\pi + (1 - \alpha) \rho^{\pi'}$ cannot be approximated by any sequence of ARUMs with fixed effects and the given μ . Formally, for all $\alpha \in (0, 1)$ and $\mu \in \mathcal{M}$, $\rho^\alpha \notin \text{cl} \bigcup_{\eta \in \mathcal{F}} \mathcal{P}_a(\eta | \mu)$.*

This lemma obviously implies statement (a) by letting $\pi' = \pi^-$. This lemma shows the limitation of the representative power of ARUMs even with fixed effects.

To show statement (b), we introduce the following concept:

²⁸Such three alternatives exist by assumption made in Section 3.2. The lemma could be stated without explicitly referring to them by letting $\pi' = \pi^-$. However, we present the result in the current form because the lemma itself is relevant to the literature, especially to Norets and Takahashi (2013). See Section A.1 in the online appendix for the details.

Definition 7. The two rankings π and π' are *adjacent* if there exist a real vector t and a real number a such that (i) $\rho^\pi \cdot t = \rho^{\pi'} \cdot t = a$, and (ii) for any $\hat{\pi} \in \Pi$, if $\hat{\pi} \notin \{\pi, \pi'\}$, then $\rho^{\hat{\pi}} \cdot t > a$.²⁹

For example, in Figure A.2 in the online appendix, ρ^{π_1} and ρ^{π_6} as well as ρ^{π_i} and $\rho^{\pi_{i+1}}$ for each $i \leq 5$ are adjacent and no other pairs are adjacent.³⁰ Since π and π^- are reversed with each other, ρ^π and ρ^{π^-} seem very different. It turns out, however, that they are adjacent:³¹

Lemma 5. *For any ranking $\pi \in \Pi$, ρ^π and ρ^{π^-} are adjacent.*

Lemma 5 follows from Proposition 4.2 of Doignon and Saito (2023), who obtain the characterization of adjacency of vertices for the case $\mathcal{D} = 2^J \setminus \{\emptyset\}$. In the online appendix, we show that Lemma 5 holds even for the case in which $\mathcal{D} \neq 2^J \setminus \{\emptyset\}$ as long as $\mathcal{D}$ contains all binary and ternary sets.³²

By using Lemma 4 and 5, we prove the following lemma, which implies statement (b).

Lemma 6. *Let μ be any element of $\mathcal{M}$. For any $\alpha \in (0, 1)$ and any ranking π that is not representable, there exists a neighborhood U of $\alpha\rho^\pi + (1 - \alpha)\rho^{\pi^-}$ such that any random utility model that belongs to U cannot be approximated arbitrarily well by any sequence of random-coefficient ARUMs with fixed effects and the given μ . Formally, $\forall \mu \in \mathcal{M}, \forall \alpha \in (0, 1), \forall \pi$: unrepresentable, $\exists U$ s.t. $\alpha\rho^\pi + (1 - \alpha)\rho^{\pi^-} \in U, \forall \rho \in \mathcal{P}_r \cap U, \rho \notin \text{cl} \bigcup_{\eta \in \mathcal{F}} \mathcal{P}_{ra}(\eta|\mu)$.*

Proposition 2 follows directly from Lemma 3-(1) and Lemma 6. Theorem 1-(ii) trivially follows from Proposition 2.

B.1 Proof of Lemma 4

Fix any $\mu \in \mathcal{M}$ and $\alpha \in (0, 1)$. Fix the three alternatives $j, l, r \in J$ and π and π' as in the statement of the lemma. Suppose by contradiction that $\rho^\alpha \in \text{cl} \bigcup_{\eta \in \mathcal{F}} \mathcal{P}_a(\eta|\mu)$. This implies that there exists a sequence $\{\rho_n\}_{n=1}^\infty \subset \bigcup_{\eta \in \mathcal{F}} \mathcal{P}_a(\eta|\mu)$ converging to ρ^α .

Let β_n and η_n be the sequence of coefficients and fixed effects vector associated with ρ_n . Define $\gamma_{n,jl} \equiv \beta_n \cdot (x_j - x_l) + \eta_{n,(\{j,l\},j)} - \eta_{n,(\{j,l\},l)}$. (Remember that we allow menu dependent fixed effects in the proof.) Define $\gamma_{n,jr}, \gamma_{n,lr}, \gamma_{n,lj}, \gamma_{n,rl}$ and $\gamma_{n,rj}$ similarly. First consider the sequence $\{\gamma_{n,jl}\}_{n=1}^\infty$.

Step 1: The sequence $\{\gamma_{n,jl}\}_{n=1}^\infty$ is bounded.

Proof. We prove this by contradiction. First, note that $\varepsilon_l - \varepsilon_j$ is a tight random variable: for each $\delta > 0$, there exists a positive number N_δ such that $\mu(\{\varepsilon_l - \varepsilon_j \in (-N_\delta, N_\delta)^c\}) < \delta$.

Fix a positive real number σ such that $\sigma < 1 - \alpha$. Let $\delta = \min\{\frac{\alpha}{2}, 1 - \alpha - \sigma\}$. Now, if $\{\gamma_{n,jl}\}_{n=1}^\infty$ is unbounded, then either there exists a subsequence $\{\gamma_{n_k,jl}\}_{k=1}^\infty$ such that $\gamma_{n_k,jl} > N_\delta$

²⁹For any $\rho \in \mathcal{P}$, $\rho \cdot t = \sum_{(D,j) \in D \times J} \rho(D,j)t(D,j)$.

³⁰The figure is intended only as an illustration and does not represent the set of random utility models exactly.

³¹Our discussions with Jean-Paul Doignon and Haruki Kono were very helpful for obtaining this result.

³²We are grateful to Haruki Kono for pointing out this fact.

for all k , or there exists a subsequence $\{\gamma_{n_k, jl}\}_{k=1}^{\infty}$ such that $\gamma_{n_k, jl} < -N_\delta$ for all k . In the first case, we have $\rho_{n_k}(\{j, l\}, j) = \mu(\{\varepsilon | \gamma_{n_k, jl} > \varepsilon_l - \varepsilon_j\}) \geq \mu(\{\varepsilon | N_\delta > \varepsilon_l - \varepsilon_j\}) \geq 1 - \delta \geq \alpha + \sigma > \alpha$; in the second case, $\rho_{n_k}(\{j, l\}, j) = \mu(\{\varepsilon | \gamma_{n_k, jl} > \varepsilon_l - \varepsilon_j\}) \leq \mu(\{\varepsilon | -N_\delta > \varepsilon_l - \varepsilon_j\}) \leq \delta \leq \frac{\alpha}{2} < \alpha$. Either case contradicts convergence to $\alpha = \rho^\alpha(\{j, l\}, j)$. Thus $\{\gamma_{n, jl}\}_{n=1}^{\infty}$ must be bounded. ■

In the same way, we can show that $\gamma_{n, lr}$, $\gamma_{n, lj}$, $\gamma_{n, rl}$, and $\gamma_{n, rj}$ are all bounded sequences. Given Step 1, we can select convergent subsequences $\{(\gamma_{n_k, jl}, \gamma_{n_k, jr}, \gamma_{n_k, lr}, \gamma_{n_k, lj}, \gamma_{n_k, rj}, \gamma_{n_k, rl})\}_{k \in \mathbf{N}}$. We denote the limits as $(\gamma_{jl}^*, \gamma_{jr}^*, \gamma_{lr}^*, \gamma_{lj}^*, \gamma_{rj}^*, \gamma_{rl}^*)$. We consider the corresponding stochastic choice functions ρ_{n_k} . Note that, by definition, $\lim_{n_k} \rho_{n_k} = \rho^\alpha$.

For any $s, t \in \{j, l, r\}$ and any positive integer n , define a measurable set $E_{n, st} = \{\varepsilon | \gamma_{n, st} > \varepsilon_t - \varepsilon_s\}$. In words, $E_{n, st}$ is the event where s is preferred over t . We further define $E_{st} = \{\varepsilon | \gamma_{st}^* \geq \varepsilon_t - \varepsilon_s\}$, and $E'_{st} = \{\varepsilon | \gamma_{st}^* > \varepsilon_t - \varepsilon_s\}$. Since $\mu \in \mathcal{M}$ is absolutely continuous with respect to the Lebesgue measure, $\mu\{\varepsilon | \gamma_{st}^* = \varepsilon_t - \varepsilon_s\} = 0$. Thus $\mu(E_{st}) = \mu(E'_{st})$. Note also by the definition of $\liminf E_{n, st}$, we have $E'_{st} \subset \liminf E_{n, st} \subset E_{st}$.

Step 2: (i) $E_{jl} = E_{jr}$ and $E_{rl} = E_{rj}$ up to a measure zero set; (ii) $\mu(E_{jl} \cap E_{jr}) = \alpha$ and $\mu(E_{rl} \cap E_{rj}) = 1 - \alpha$.

Proof. First note that for each n , we have $\rho_n(\{j, l, r\}, j) = \mu(E_{n, jl} \cap E_{n, jr})$ because of the assumption on the fixed effects: $\eta_{(\{j, l, r\}, t)} = \eta_{(\{j, l\}, t)} = \eta_{(\{j, r\}, t)}$ for all $t \in \{j, l, r\}$. Thus, by Fatou's lemma, we have $\alpha = \rho^\alpha(\{j, l, r\}, j) = \limsup \rho_n(\{j, l, r\}, j) = \limsup \mu(E_{n, jl} \cap E_{n, jr}) \leq \mu(\limsup (E_{n, jl} \cap E_{n, jr})) \leq \mu(E_{jl} \cap E_{jr})$, and $\alpha = \rho^\alpha(\{j, l\}, j) = \liminf \mu(E_{n, jl}) \geq \mu(\liminf E_{n, jl}) \geq \mu(E'_{jl}) = \mu(E_{jl})$. By an identical argument applied to $\{j, r\}$, we have $\alpha \geq \mu(E_{jr})$. Thus we have $\mu(E_{jl} \cap E_{jr}) \geq \alpha \geq \max\{\mu(E_{jl}), \mu(E_{jr})\} \geq \mu(E_{jl} \cap E_{jr})$. It follows that $\mu(E_{jl} \cap E_{jr}) = \alpha = \mu(E_{jl}) = \mu(E_{jr})$; thus $E_{jl} = E_{jr}$ up to a measure zero set. By applying the same argument to $\{r, l\}$ and $\{r, j\}$, we obtain $\mu(E_{rl} \cap E_{rj}) = 1 - \alpha$ and $E_{rl} = E_{rj}$ up to a measure zero set. ■

For each $\varepsilon \in \mathbf{R}^J$, define a projected column vector $\varepsilon_{jlr} = (\varepsilon_j, \varepsilon_l, \varepsilon_r)$. Further define $A = \{\varepsilon | U\varepsilon_{jlr} \geq c\}$ and $B = \{\varepsilon | U\varepsilon_{jlr} \leq c\}$, where

$$U = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}, \quad c = \begin{bmatrix} \gamma_{lj}^* \\ \gamma_{rl}^* \end{bmatrix}. \quad (9)$$

In words, A is the event that j is preferred over l and l is preferred over r , and B is the event with the reverse ordering. Note that both sets are closed and thus measurable.

Step 3: $\mu(A) = \alpha$, $\mu(B) = 1 - \alpha$, and $\mu(A \cup B) = 1$.

Proof. By Step 2, $\mu(E_{jl} \cap E_{jr}) = \alpha$ and $\mu(E_{rl} \cap E_{rj}) = 1 - \alpha$. Remember E_{st} is the event that s is chosen over t in the binary set $\{s, t\}$. Notice $E_{jl} \cap E_{jr}$ and $E_{rl} \cap E_{rj}$ have measure zero intersections, so the two events partition the probability space (ignoring measure zero events).

Notice that $\alpha = \rho^\alpha(\{r, l\}, l) = \mu(E_{lr})$. Since the event E_{lr} is incompatible with the event $E_{rl} \cap E_{rj}$ up to a measure zero set, E_{lr} must completely lie within the event $E_{jl} \cap E_{jr}$ (up to a measure zero set). Moreover, since $\mu(E_{lr}) = \alpha = \mu(E_{jl} \cap E_{jr})$, the event E_{lr} coincides with the event $E_{jl} \cap E_{jr}$ (ignoring measure zero events). Finally, notice the event A is the intersection of $E_{jl} \cap E_{jr}$ and E_{lr} . Thus, $\mu(A) = \alpha$. In a similar way, we can show that $\mu(B) = 1 - \alpha$. Thus we have $\mu(A \cup B) = 1$. $\blacksquare$

Define $A' \equiv \{\varepsilon \mid U\varepsilon_{jlr} > c\}$ and $B' \equiv \{\varepsilon \mid U\varepsilon_{jlr} < c\}$.

Step 4: There exist $\varepsilon^a \in A' \cap \text{supp } \mu$ and $\varepsilon^b \in B' \cap \text{supp } \mu$ such that $\varepsilon^\lambda \equiv \lambda\varepsilon^a + (1-\lambda)\varepsilon^b \notin A \cup B$ for some $\lambda \in [0, 1]$.

Proof. Since μ is a standard probability measure, $\text{supp } \mu$ contains an open set in $\mathbf{R}^J$. Moreover, since U has full row rank, there exist $\varepsilon^a \in A' \cap \text{supp } \mu$ and $\varepsilon^b \in B' \cap \text{supp } \mu$ such that three points $U\varepsilon_{jlr}^a$, $U\varepsilon_{jlr}^b$ and c in $\mathbf{R}^2$ are not collinear, where $\varepsilon_{jlr}^a = (\varepsilon_j^a, \varepsilon_l^a, \varepsilon_r^a)$ and $\varepsilon_{jlr}^b = (\varepsilon_j^b, \varepsilon_l^b, \varepsilon_r^b)$ are projections of ε^a and ε^b , respectively.³³

For all $\lambda \in [0, 1]$, define $\varepsilon^\lambda \equiv \lambda\varepsilon^a + (1-\lambda)\varepsilon^b$ and let $\varepsilon_{jlr}^\lambda$ denote the corresponding projection. It remains to show that there exists $\lambda^* \in (0, 1)$ such that $\varepsilon^{\lambda^*} \notin A \cup B$. For each $\lambda \in [0, 1]$, define $y^\lambda \equiv U\varepsilon_{jlr}^\lambda - c = \lambda(U\varepsilon_{jlr}^a - c) + (1-\lambda)(U\varepsilon_{jlr}^b - c) \in \mathbf{R}^2$. Write $y^\lambda = (y_1^\lambda, y_2^\lambda)$. Since $\varepsilon^a \in A'$ and $\varepsilon^b \in B'$, we have $y_1^1 > 0$, $y_2^1 > 0$, and $y_1^0 < 0$, $y_2^0 < 0$. Hence, for each coordinate $i = 1, 2$, the affine function $\lambda \mapsto y_i^\lambda$ is negative at $\lambda = 0$, positive at $\lambda = 1$, and y_i^λ is strictly increasing and continuous with respect to λ for all i . Therefore, there exists a unique $\lambda_i \in (0, 1)$ such that $y_i^{\lambda_i} = 0$. We claim that $\lambda_1 \neq \lambda_2$. Indeed, if $\lambda_1 = \lambda_2$, then $y^{\lambda_1} = 0$, or equivalently, $U\varepsilon_{jlr}^{\lambda_1} = c$. This would imply that c lies on the line segment joining $U\varepsilon_{jlr}^a$ and $U\varepsilon_{jlr}^b$, contradicting that three points $U\varepsilon_{jlr}^a$, $U\varepsilon_{jlr}^b$ and c in $\mathbf{R}^2$ are not collinear. Thus $\lambda_1 \neq \lambda_2$. Without loss of generality, suppose $\lambda_1 < \lambda_2$. Since $y_i^1 > 0 > y_i^0$, the affine function $\lambda \mapsto y_i^\lambda$ is strictly increasing for all i . Choose any $\lambda^* \in (\lambda_1, \lambda_2)$. Since y_i^λ is strictly increasing, $y_1^{\lambda_1} = 0$, and $y_2^{\lambda_2} = 0$, we have $y_1^{\lambda^*} > 0$ and $y_2^{\lambda^*} < 0$. Thus y^{λ^*} has mixed signs. Hence $U\varepsilon_{jlr}^{\lambda^*} - c$ is neither weakly positive componentwise nor weakly negative componentwise. By the definitions of A and B , this implies $\varepsilon^{\lambda^*} \notin A \cup B$. This proves the desired claim. $\blacksquare$

By Step 4, it follows from the convexity of $\text{supp } \mu$ that $\varepsilon^\lambda \in \text{supp } \mu \cap (A \cup B)^c$ for some $\lambda \in [0, 1]$. Moreover, $(A \cup B)^c$ is open. It follows from the definition of the support that there exists $r > 0$ such that the ball centered at ε^λ with radius r , $B_r(\varepsilon^\lambda)$, satisfies $\mu(B_r(\varepsilon^\lambda)) > 0$ and $B_r(\varepsilon^\lambda) \subset (A \cup B)^c$. This contradicts $\mu(A \cup B) = 1$.

³³That $\text{supp } \mu$ contains an open set follows from the definition of standard probability measures. By Definition 2, μ is absolutely continuous with respect to Lebesgue measure and has convex support. A convex set with positive Lebesgue measure must have nonempty interior (Theorem 6.2 of Rockafellar (2015)), and therefore contains an open ball.

B.2 Proof of Lemma 6

Let $\mu \in \mathcal{M}$. Let π be a ranking that is not representable. Fix $\alpha \in (0, 1)$. Suppose by way of contradiction that ρ_α^π is approximated arbitrarily well by a sequence of random-coefficient ARUMs with fixed effects; formally, $\rho_\alpha^\pi \in \text{cl } \bigcup_{\eta \in \mathcal{F}} \mathcal{P}_{ra}(\eta|\mu)$.

We first show the following claim:

Claim: *There exist two sequences of ARUMs with the same fixed effects that converge to ρ^π and ρ^{π^-} , respectively.*

Proof. Let ρ_n be the elements of $\bigcup_{\eta \in \mathcal{F}} \mathcal{P}_{ra}(\eta|\mu)$ converging to ρ_α^π . Let η_n be the sequence of the fixed effects associated with ρ_n . By Proposition 6 of the online appendix, for each n , we have $\mathcal{P}_{ra}(\eta_n|\mu) = \text{co}\mathcal{P}_a(\eta_n|\mu)$; thus $\rho_n \in \text{co}\mathcal{P}_a(\eta_n|\mu)$ for all n . Therefore, we can write ρ_n as $\sum_{i=1}^M \sigma_{ni} \rho_{ni}$, where $\rho_{ni} \in \mathcal{P}_a(\eta_n|\mu)$ and $M = \dim \mathcal{P}_r + 1$.

Since the sets of weights and random utility models are compact, we can extract converging subsequences such that for all i , $\sigma_{n_k i} \rightarrow \sigma_i^*$ and $\rho_{n_k i} \rightarrow \rho_i^*$ as $n_k \rightarrow \infty$. Thus, $\sum_i \sigma_{n_k i} \rho_{n_k i} \rightarrow \sum_i \sigma_i^* \rho_i^* = \rho_\alpha^\pi$. Moreover, we have $\rho_{n_k i} \in \mathcal{P}_a(\eta_{n_k}|\mu)$ for each i ; and thus $\rho_i^* \in \text{cl } \bigcup_{\eta \in \mathcal{F}} \mathcal{P}_a(\eta|\mu)$ for each i .

In the following, we will show that there exist some i, j such that $\rho_i^* = \rho^\pi$ and $\rho_j^* = \rho^{\pi^-}$. This completes the proof since it implies $\rho_{n_k i} \rightarrow \rho^\pi$ and $\rho_{n_k j} \rightarrow \rho^{\pi^-}$ as $n_k \rightarrow \infty$; and moreover both sequences can be chosen to share the same sequence of fixed effects η_{n_k} because $\rho_{n_k i}, \rho_{n_k j} \in \mathcal{P}_a(\eta_{n_k}|\mu)$ for each n_k .

Suppose, by contradiction and without loss of generality, that $\rho_i^* \neq \rho^\pi$ for any i .³⁴

By Lemma 5, ρ^π and ρ^{π^-} are adjacent. Let (t, a) be as in Definition 7 with the pair (ρ^π, ρ^{π^-}) of adjacent rankings. We will consider two cases.

Case 1: $\rho_i^* \neq \rho^{\pi^-}$ for any i . Thus, we have $\rho_i^* \notin \{\rho^\pi, \rho^{\pi^-}\}$ for all i . Since $\rho_i^* \in \text{cl } \bigcup_{\eta \in \mathcal{F}} \mathcal{P}_a(\eta|\mu)$ for all i , Lemma 4 implies $\rho_i^* \notin \{\rho_\alpha^\pi | \alpha \in (0, 1)\}$ for all i . It follows that $\rho_i^* \notin \{\rho_\alpha^\pi | \alpha \in [0, 1]\}$ for all i . Since π and π^- are adjacent vertices, by Definition 7, we have $\rho_i^* \cdot t > a$ for all i . Thus, $\left(\sum_i \sigma_i^* \rho_i^*\right) \cdot t = \sum_i \sigma_i^* \rho_i^* \cdot t > a$. On the other hand by Definition 7, $\left(\sum_i \sigma_i^* \rho_i^*\right) \cdot t = \rho_\alpha^\pi \cdot t = a$. This is a contradiction.

Case 2: $\rho_i^* = \rho^{\pi^-}$ for some i . Define $I = \{i \in \{1, \dots, M+1\} | \rho_i^* = \rho^{\pi^-}\}$. First notice that there exists $i \in \{1, \dots, M\} \setminus I$ such that $\sigma_i^* > 0$. (If such i does not exist, then $\rho^{\pi^-} = \sum_i \sigma_i^* \rho_i^* = \rho_\alpha^\pi$, which contradicts $\alpha \notin \{0, 1\}$.) Then, $a = \rho_\alpha^\pi \cdot t = \sum_i \sigma_i^* \rho_i^* \cdot t = \sum_{i \in I} \sigma_i^* \rho^{\pi^-} \cdot t + \sum_{i \notin I} \sigma_i^* \rho_i^* \cdot t > a$, where the strict inequality holds because $\rho_i^* \cdot t > a$ for all $i \notin I$ by Definition 7. This is a contradiction. $\blacksquare$

Then by the claim above, there exist a sequence $\{\eta_n\}$ of fixed effects and two sequences $\{\rho_n\}$ and $\{\rho'_n\}$ in $\bigcup_{\eta \in \mathcal{F}} \mathcal{P}_a(\eta|\mu)$ such that (i) $\rho_n \rightarrow \rho^\pi$; (ii) $\rho'_n \rightarrow \rho^{\pi^-}$; and (iii) $\rho_n, \rho'_n \in \mathcal{P}_a(\eta_n|\mu)$ for

³⁴The proof for the other case is exactly the same after changing ρ^{π^-} to ρ^π and ρ^π to ρ^{π^-} .

each n . (We write n instead of n_k for simplicity.) Let β_n be the coefficient vector associated with ρ_n and β'_n the coefficient vector associated with ρ'_n ; by (iii), both share the fixed effects η_n .

For each pair j, l with $\pi(j) > \pi(l)$, define $Z_{lj} := \varepsilon_l - \varepsilon_j$, $\gamma_{n,jl} := \beta_n \cdot (x_j - x_l) + \eta_{n,\{j,l\},j} - \eta_{n,\{j,l\},l}$, and $\gamma'_{n,lj} := \beta'_n \cdot (x_l - x_j) + \eta_{n,\{j,l\},l} - \eta_{n,\{j,l\},j}$. The convergence to ρ^π and ρ^{π^-} gives $\mu(\{\varepsilon|Z_{lj} < \gamma_{n,jl}\}) \rightarrow 1$ and $\mu(\{\varepsilon|Z_{lj} > -\gamma'_{n,lj}\}) \rightarrow 1$. If $\gamma_{n,jl} + \gamma'_{n,lj} \leq 0$ for infinitely many n , then along that infinite subsequence the two sets $\{\varepsilon|Z_{lj} < \gamma_{n,jl}\}$ and $\{\varepsilon|Z_{lj} > -\gamma'_{n,lj}\}$ are disjoint, so their probabilities have sum at most one. This contradicts the fact that both probabilities converge to one. Therefore, for this pair j, l , $\gamma_{n,jl} + \gamma'_{n,lj} > 0$ for all sufficiently large n . Since there are finitely many pairs, there exists a single N such that for all $n \geq N$ and all j, l with $\pi(j) > \pi(l)$, $\gamma_{n,jl} + \gamma'_{n,lj} > 0$. The fixed effects cancel, so $\gamma_{n,jl} + \gamma'_{n,lj} = (\beta_n - \beta'_n) \cdot (x_j - x_l) > 0$. Thus $\beta_n - \beta'_n$ represents π for all sufficiently large n , contradicting the unrepresentability of π .

References

- ABDALLAH, T., AND G. VULCANO (2021): “Demand Estimation under the Multinomial Logit Model from Sales Transaction Data,” *Manufacturing & Service Operations Management*, 23(5), 1196–1216.
- AKCHEN, Y.-C., AND D. MITROFANOV (2025): “Consider or Choose? The Role and Power of Consideration Sets,” *Management Science*.
- ALLEN, R., AND J. REHBECK (2020): “Hicksian Complementarity and Perturbed Utility Models,” *Economic Theory Bulletin*, 8, 245–261.
- ALPTEKINOĞLU, A., AND J. H. SEMPLE (2016): “The Exponential Choice Model: A New Alternative for Assortment and Price Optimization,” *Operations Research*, 64(1), 79–93.
- APESTEGUIA, J., AND M. BALLESTER (2018): “Monotone Stochastic Choice Models: The Case of Risk and Time Preferences,” *Journal of Political Economy*, 126(1), 74–106.
- ATHEY, S., AND G. W. IMBENS (2007): “Discrete Choice Models with Multiple Unobserved Choice Characteristics,” *International Economic Review*, 48(4), 1159–1192.
- BARRON, A. R., A. COHEN, W. DAHMEN, AND R. A. DEVORE (2008): “Approximation and Learning by Greedy Algorithms,” *Annals of Statistics*, 36(1), 64–94.
- BERBEGLIA, G., A. GARASSINO, AND G. VULCANO (2022): “A Comparative Empirical Study of Discrete Choice Models in Retail Operations,” *Management Science*, 68(6), 4005–4023.
- BERRY, S. T. (1994): “Estimating Discrete-choice Models of Product Differentiation,” *RAND Journal of Economics*, pp. 242–262.
- BERRY, S. T., AND P. A. HAILE (2014): “Identification in Differentiated Products Markets Using Market Level Data,” *Econometrica*, 82(5), 1749–1797.
- BLOCK, H. D., AND J. MARSCHAK (1960): “Random Orderings and Stochastic Theories of Responses,” *Contributions to Probability and Statistics*, 2, 97–132.

- CAMERON, A. C., AND P. K. TRIVEDI (2005): *Microeconometrics: Methods and Applications*. Cambridge University Press.
- CERREIA-VIOGLIO, S., P. O. LINDBERG, F. MACCHERONI, M. MARINACCI, AND A. RUSTICHINI (2021): “A Canon of Probabilistic Rationality,” *Journal of Economic Theory*, 196, 105289.
- CERREIA-VIOGLIO, S., F. MACCHERONI, M. MARINACCI, AND A. RUSTICHINI (2023): “Multinomial Logit Processes and Preference Discovery: Inside and Outside the Black Box,” *Review of Economic Studies*, 90(3), 1155–1194.
- CHAMBERS, C. P., Y. MASATLIOGLU, P. NATENZON, AND C. RAYMOND (2025): “Weighted Linear Discrete Choice,” *American Economic Review*.
- CHEN, X. (2007): “Large Sample Sieve Estimation of Semi-nonparametric Models,” *Handbook of Econometrics*, 6, 5549–5632.
- CHINTAGUNTA, P., J.-P. DUBÉ, AND K. Y. GOH (2005): “Beyond the Endogeneity Bias: The Effect of Unmeasured Brand Characteristics on Household-level Brand Choice Models,” *Management Science*, 51(5), 832–849.
- COMPIANI, G. (2022): “Market Counterfactuals and the Specification of Multiproduct Demand: A Nonparametric Approach,” *Quantitative Economics*, 13(2), 545–591.
- CONLON, C., AND J. H. MORTIMER (2021): “Empirical Properties of Diversion Ratios,” *RAND Journal of Economics*, 52(4), 693–726.
- CROISSANT, Y. (2020): “Estimation of Random Utility Models in R: The mlogit Package,” *Journal of Statistical Software*, 95(11), 1–41.
- DAGSVIK, J. K. (1994): “Discrete and Continuous Choice, Max-Stable Processes, and Independence from Irrelevant Attributes,” *Econometrica*, pp. 1179–1205.
- DEMPSTER, A. P., N. M. LAIRD, AND D. B. RUBIN (1977): “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1), 1–22.
- DOIGNON, J.-P., AND K. SAITO (2023): “Adjacencies on Random Ordering Polytopes and Flow Polytopes,” *Journal of Mathematical Psychology*, 114, 102768.
- FARIAS, V. F., S. JAGABATHULA, AND D. SHAH (2009): “A Non-parametric Approach to Modeling Choice with Limited Data,” *arXiv preprint arXiv:0910.0063*.
- (2013): “A Nonparametric Approach to Modeling Choice with Limited Data,” *Management Science*, 59(2), 305–322.
- FENG, Q., J. G. SHANTHIKUMAR, AND M. XUE (2022): “Consumer Choice Models and Estimation: A Review and Extension,” *Production and Operations Management*, 31(2), 847–867.
- FIEBIG, D. G., M. P. KEANE, J. LOUVIERE, AND N. WASI (2010): “The Generalized Multinomial Logit Model: Accounting for Scale and Coefficient Heterogeneity,” *Marketing Science*, 29(3), 393–421.
- FRICK, M., R. IJIMA, AND T. STRZALECKI (2019): “Dynamic Random Utility,” *Econometrica*, 87(6), 1941–2002.
- GHALANOS, A., AND S. THEUSSL (2015): *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*. R package version 1.16.

- HERRIGES, J. A., AND C. L. KLING (1999): “Nonlinear Income Effects in Random Utility Models,” *Review of Economics and Statistics*, 81(1), 62–72.
- HORAN, S. (2021): “Stochastic Semi-orders,” *Journal of Economic Theory*, 192, 105171.
- HORAN, S., AND A. ADAM (2023): “Slutsky Patterns of Substitution in Discrete Choice,” Working Paper.
- JAGABATHULA, S., AND A. VENKATARAMAN (2022): “Nonparametric Estimation of Choice Models,” in *The Elements of Joint Learning and Optimization in Operations Management*, pp. 177–209. Springer.
- KAMAKURA, W. A., B.-D. KIM, AND J. LEE (1996): “Modeling Preference and Structural Heterogeneity in Consumer Choice,” *Marketing Science*, 15(2), 152–172.
- KÖK, A. G., M. L. FISHER, AND R. VAIDYANATHAN (2008): “Assortment Planning: Review of Literature and Industry Practice,” *Retail Supply Chain Management: Quantitative Models and Empirical Studies*, pp. 99–153.
- LU, J., AND K. SAITO (2021): “Mixed Logit and Pure Characteristic Models,” Working Paper.
- LUCE, R. D. (1959): *Individual Choice Behavior*. New York: Wiley.
- MATOUSEK, J. (2013): *Lectures on Discrete Geometry*, vol. 212. Springer Science & Business Media.
- McFADDEN, D., AND K. TRAIN (2000): “Mixed MNL Models for Discrete Response,” *Journal of Applied Econometrics*, pp. 447–470.
- MUSALEM, A., M. OLIVARES, E. T. BRADLOW, C. TERWIESCH, AND D. CORSTEN (2010): “Structural Estimation of the Effect of Out-of-stocks,” *Management Science*, 56(7), 1180–1197.
- NORETS, A., AND S. TAKAHASHI (2013): “On the Surjectivity of the Mapping between Utilities and Choice Probabilities,” *Quantitative Economics*, 4(1), 149–155.
- ROCKAFELLAR, R. T. (2015): *Convex Analysis*, Princeton Landmarks in Mathematics and Physics. Princeton University Press, Princeton, NJ.
- RUSMEVICHIENTONG, P., Z.-J. M. SHEN, AND D. B. SHMOYS (2010): “Dynamic Assortment Optimization with a Multinomial Logit Choice Model and Capacity Constraint,” *Operations Research*, 58(6), 1666–1680.
- STOER, J., AND C. WITZGALL (2012): *Convexity and Optimization in Finite Dimensions I*, vol. 163. Springer Science & Business Media.
- TEBALDI, P., A. TORGOVITSKY, AND H. YANG (2023): “Nonparametric Estimates of Demand in the California Health Insurance Exchange,” *Econometrica*, 91(1), 107–146.
- THOMSON, C. J., AND S. J. CROOKE (1991): “Results of the Southern California Sportfish Economic Survey,” *NOAA Technical Memorandum NMFS*.
- TRAIN, K. E. (2009): *Discrete Choice Methods with Simulation*. Cambridge University Press.
- YE, Y. (1987): “Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming,” Ph.D. thesis, Department of ESS, Stanford University.

For Online Publication

Online Appendix for “Approximating Choice Data by Discrete Choice Models”

A Generalization

A.1 Menu-dependent Fixed Effects

In Section 3.2, we stated that our key results continue to hold with menu-dependent fixed effects under a mild locality restriction: for some triple $\{j, l, r\} \subseteq J$, the fixed effects $\eta_{(D,i)}$ are constant across all menus $D \subseteq \{j, l, r\}$ and for each $i \in D$. Here, we drop this restriction and allow $\eta_{(D,i)}$ to vary arbitrarily across menus D and items $i \in D$.

By a result in Norets and Takahashi (2013), fully menu-dependent fixed effects are rich enough to approximate any random utility model over the choice sets in D , even without random coefficients. To see this, recall that our goal is to choose $m(\beta)$ and $(\eta_{(D,i)})_{i \in D}$ for each menu $D \in \mathcal{D}$ so that

$$\int \mu\left(\{\varepsilon : \beta \cdot x_j + \eta_{(D,j)} + \varepsilon_j > \beta \cdot x_\ell + \eta_{(D,\ell)} + \varepsilon_\ell, \forall \ell \in D \setminus \{j\}\}\right) dm(\beta)$$

can approximate $\rho(D, j)$ arbitrarily well for every menu $D \in \mathcal{D}$ and $j \in D$. Observe that we may absorb the systematic term $\beta \cdot x_j$ into the fixed effect by defining $\tilde{\eta}_{(D,j)} = \eta_{(D,j)} + \beta \cdot x_j$. Let the random coefficient distribution place all its mass at $\beta = 0$. The model then reduces to

$$\mu\left(\{\varepsilon : \tilde{\eta}_{(D,j)} + \varepsilon_j > \tilde{\eta}_{(D,\ell)} + \varepsilon_\ell, \forall \ell \in D \setminus \{j\}\}\right).^{35} \quad (10)$$

Mathematically, given the $|D|$ target probabilities $\{\rho(D, j)\}_{j \in D}$, we seek $|D|$ location parameters $\{\tilde{\eta}_{(D,j)}\}_{j \in D}$ solving (10). This is precisely a surjectivity problem for the multinomial choice mapping from location shifts to choice probabilities. Norets and Takahashi (2013) shows that under a weak regularity condition that ε is continuously distributed with respect to the Lebesgue measure, the mapping contains the interiors of all probability simplices associated with $\mathcal{D}$ and hence can approximate any random utility model.

As the results above show, allowing menu-dependent fixed effects can substantially increase the model’s expressive power. With fully flexible menu-dependent fixed effects, the model can even approximate choice behavior that is inconsistent with any RUM, thereby achieving perfect

³⁵Fixed effects are menu dependent but not random: m is not a distribution over η . If one were to allow *random* fixed effects (i.e., draw η from a distribution), then by matching that distribution to a distribution over rankings one can trivially replicate any RUM. Our focus here is the nonrandom η case.

in-sample fit. In counterfactual applications, however, researchers typically use the estimated random-coefficient distribution together with fixed effects that are either set to a priori constants or inherited from observed choice sets. In this case, our main theorem implies that, unless the affine-independence condition is satisfied, there exist random utility models on the relevant counterfactual choice sets that cannot be approximated arbitrarily well. For example, suppose that we observe data only for the full menu J , but wish to make counterfactual predictions for all binary and ternary choice sets. Then affine independence is required for perfect approximation. We note, however, that this requirement depends on the collection of choice sets for which counterfactual predictions are of interest. For instance, if we observe choice frequencies for all binary and ternary choice sets and only wish to conduct counterfactual predictions for the full menu, then convex independence is sufficient.

A.2 Restricting Possible Rankings

As highlighted in the main text, there are situations where a researcher deems certain rankings as unreasonable, leading them to constrain the possible set of rankings to $\hat{\Pi} \subset \Pi$. To accommodate this case, define $\mathcal{P}_r(\hat{\Pi}) \equiv \text{co} \{ \rho^\pi | \pi \in \hat{\Pi} \}$. We call an element of $\mathcal{P}_r(\hat{\Pi})$ a *random utility model on $\hat{\Pi}$* .

In our main results (Theorem 1 and Proposition 1), we considered the case where $\hat{\Pi} = \Pi$ for simplicity. In this section, we provide a necessary and sufficient condition for the approximation of the random utility models on $\hat{\Pi}$.

Since the set $\mathcal{P}_r(\hat{\Pi})$ of random utility models on $\hat{\Pi}$ is also a polytope, we can generalize Lemma 1 for $\mathcal{P}_r(\hat{\Pi})$ by simply changing Π to $\hat{\Pi}$. That is, we have the following result:

Lemma 7. *Let $\hat{\Pi} \subset \Pi$. Let $\mu \in \mathcal{M}$ be any standard probability measure. Then $\mathcal{P}_r(\hat{\Pi}) \subset \text{cl } \mathcal{P}_{ra}(0|\mu)$ if and only if $\rho^\pi \in \text{cl } \mathcal{P}_a(0|\mu)$ for any $\pi \in \hat{\Pi}$.*

The proof is essentially the same as the proof of Lemma 1. This result, together with Lemma 2 and Proposition 2, implies the following result.

Corollary 2. *Let $\hat{\Pi} \subset \Pi$.*

- (i) *Let $\mu \in \mathcal{M}$ be any standard probability measure. If every ranking $\pi \in \hat{\Pi}$ is representable in $\mathcal{D}$, then any random utility model on $\hat{\Pi}$ can be approximated arbitrarily well by a sequence of random-coefficient ARUMs with the given distribution μ . Moreover, the approximation can be done without fixed effects (i.e., $\eta = 0$).*
- (ii) *If some ranking $\pi \in \hat{\Pi}$ is not representable and $\pi^- \in \hat{\Pi}$, then there exists a random utility model ρ on $\hat{\Pi}$ such that for any standard probability measure $\mu \in \mathcal{M}$, ρ cannot be approximated arbitrarily well by any sequence of random-coefficient ARUMs with fixed effects and with the given distribution $\mu \in \mathcal{M}$.*

Proof. Statement (i) follows from Lemma 2-(1) and Lemma 7. Statement (ii) follows from Proposition 2. Indeed, take $\alpha \in (0, 1)$ and set $\rho^* = \alpha\rho^\pi + (1 - \alpha)\rho^{\pi^-}$. Since $\pi, \pi^- \in \hat{\Pi}$, we have $\rho^* \in \mathcal{P}_r(\hat{\Pi})$. By Proposition 2, for every $\mu \in \mathcal{M}$, $\rho^* \notin \text{cl } \bigcup_{\eta \in F} \mathcal{P}_{ra}(\eta \mid \mu)$. $\square$

Corollary 2 offers a testable condition for determining if the random-coefficient ARUMs can adequately approximate any random utility model on $\hat{\Pi} \subset \Pi$. If the researcher wishes to omit certain rankings from the analysis, this corollary may be more useful than Theorem 1 and Proposition 1.

B More Comments on the Related Literature

B.1 McFadden and Train (2000)

In the main body of the paper, we mentioned McFadden and Train (2000) in several sections; here we provide a summary.

McFadden and Train (2000) show that a mixed-logit model can approximate an arbitrary (nonparametric) continuous random-utility model arbitrarily well. We also explain that this flexibility is achieved by using higher-order polynomials of arbitrarily high degrees, which corresponds to letting $K \rightarrow \infty$.

The key differences concern both the setup and the sharpness of the results. We provide a *necessary and sufficient* characterization of when random-coefficient ARUMs can approximate arbitrary random-utility behavior, whereas McFadden and Train (2000) establish a *sufficient* condition for approximation. We obtain this sharper characterization by focusing on approximation for a *fixed, finite set of alternatives*. In addition, we analyze approximation of choice probabilities *across choice sets*. This cross-menu perspective with fixed alternatives is central for applications in assortment optimization and revenue management (Farias, Jagabathula, and Shah, 2009; Jagabathula and Venkataraman, 2022). Our affine-independence condition is necessary and sufficient, applies beyond mixed-logit, and is tailored to environments with fixed finite alternatives. By contrast, McFadden and Train (2000) provide asymptotic approximation guarantees, focusing primarily on settings with continuous covariates, and they offer no comparable guidance on how to construct random-coefficient ARUMs to achieve approximation in the finite-alternatives, cross-menu setting.

B.2 Norets and Takahashi (2013)

As explained in the main body of the paper, Norets and Takahashi (2013) show that for a *single* multinomial choice problem with menu D , under a mild regularity condition, one can always find an ARUM that matches the given dataset. In particular, given any interior probability

vector $\{\rho(D, j)\}_{j \in D}$, there exists a vector of fixed effect parameters $\{\eta_{(D, j)}\}_{j \in D}$ such that the ARUM with the fixed effect vector but without random coefficients generates those probabilities on that menu.

The essential differences are the choice-set configuration under which the approximation problem is studied; and the approximation class they use. Norets and Takahashi (2013) study the approximation power of ARUMs *on a single choice set*. On the other hand, we study the approximation power of random-coefficient ARUMs with fixed effects *across multiple choice sets simultaneously* given a fixed set of alternatives.

In Lemma 4, we examine the ARUM without random coefficients across *multiple* menus. While Norets–Takahashi’s surjectivity ensures that, menu by menu, any interior vector $\{\rho(D, j)\}_{j \in D}$ can be matched by some $\{\eta_{(D, j)}\}_{j \in D}$, Lemma 4 shows that such per-menu constructions *cannot, in general, be made mutually consistent across menus*.

B.3 Lu and Saito (2021)

Lu and Saito (2021) differ from our analysis in three key respects. First, they work in a continuous characteristics space, as in McFadden and Train (2000). Second, they restrict the approximation target to *pure-characteristics* models (i.e., continuous random-utility models). Third, they focus on parametric mixed-logit specifications whose systematic utility is a polynomial of degree at most a given bound.

In particular, Lu and Saito (2021) ask, within the class of pure-characteristics models, which subclasses can be approximated by a degree- d mixed-logit? Their analysis is conducted under continuity restrictions that are natural in a continuous characteristics space but are not well defined on our finite alternative domain.

By contrast, we ask for necessary and sufficient conditions under which any RUM can be approximated by a popular parametric class—random-coefficient ARUMs—when only a finite collection of menus is observed. Consequently, there is no exact finite-domain analogue of the question studied in Lu and Saito (2021).

In what follows, we explain the similarities and differences in more detail, using a formal mathematical comparison. Both Lu and Saito (2021) and our paper establish a baseline approximation inclusion that holds for every d :

$$\text{cl } \mathcal{P}_{mi}(d) \supseteq \mathcal{P}_r(d), \quad (11)$$

where $\mathcal{P}_{mi}(d)$ is the class of mixed-logit models whose systematic utility is a polynomial of degree at most d in characteristics (random coefficients allowed), and $\mathcal{P}_r(d)$ the subclass of

target models whose systematic utility likewise has degree at most d .³⁶ Intuitively, the result states that a degree- d mixed-logit can approximate any degree- d target model (in closure). This statement is the bridge connecting the two frameworks.

Continuous domain (Lu–Saito). In the continuous setting of pure-characteristic models, Lu–Saito show that intersecting with the continuous class yields an equality:

$$\mathcal{P}_{cr} \cap \text{cl } \mathcal{P}_{ml}(d) = \mathcal{P}_r(d), \quad (12)$$

where $\mathcal{P}_{cr}$ is the set of pure-characteristics models (i.e., continuous random-utility models). The result means that within the continuous pure-characteristics domain, the closure of degree- d mixed-logits exactly recovers the degree- d target class. Moreover, their “degree matching” results clarify that approximation quality hinges on the polynomial degree of the systematic utility.

Discrete domain (our paper). On a finite alternative set, continuity imposes no restriction. The affine independence of the set of characteristics plays a key role. Formally, let $x_j \in \mathbf{R}^K$ be characteristics and let $x_j^{(d)}$ be the vector of monomials of x_j of degrees at most d . Define

$$X(d) = \{x_j^{(d)} | j \in J\}. \quad (13)$$

Our central discrete-domain statement is:

$$\mathcal{P}_r(d) = \mathcal{P}_r \iff X(d) \text{ is affinely independent}, \quad (14)$$

where $\mathcal{P}_r$ is the set of RUMs on the finite domain. Combining (14) with the common inclusion (11) yields

$$\text{cl } \mathcal{P}_{ml}(d) = \mathcal{P}_r \iff X(d) \text{ is affinely independent},$$

which is the discrete analogue of (12) and a special case of our main theorem. In words: degree- d mixed-logits (indeed, random-coefficient ARUMs with degree- d systematic utility) approximate *every* finite-domain RUM in closure if and only if $X(d)$ is affinely independent.

Similarity and difference. The similarity is the shared approximation set-inclusion (11): in both papers, degree- d mixed-logits are as expressive as degree- d targets, up to closure. The difference lies in the domain-specific condition needed to sharpen inclusion to equality. In the continuous case, equality follows after intersecting with $\mathcal{P}_{cr}$ (continuity). In the discrete case,

³⁶For the definition $\mathcal{P}_r(d)$, see Definition 1 in Lu and Saito (2021).

equality holds when $X(d)$ defined in (13) is affinely independent.³⁷

C Proof of Proposition 3 (ii)

To prove statement (ii)(a), suppose that $x_j : j \in J$ is not convex-independent. Then there exists an alternative $j^* \in J$ such that $x_{j^*} \in \text{co}\{x_l | l \in J \setminus \{j^*\}\}$. Thus there are nonnegative weights $(\lambda_l)_{l \neq j^*}$ satisfying $\sum_{l \neq j^*} \lambda_l = 1$ and $x_{j^*} = \sum_{l \neq j^*} \lambda_l x_l$. Let $L := \{l \neq j^* : \lambda_l > 0\}$. Define $\rho^* \in \mathcal{P}_r$ to be the deterministic random utility model defined by $\rho^*(J, j^*) = 1$ and $\rho^*(J, l) = 0$ for all $l \neq j^*$. We show that $\rho^* \notin \text{cl } \mathcal{P}_{ra}(0 | \mu)$ for every $\mu \in \mathcal{M}$ such that $\mathbf{0} \in \text{int}(\text{supp } \mu)$.

Fix any $\mu \in \mathcal{M}$ such that $\mathbf{0} \in \text{int}(\text{supp } \mu)$. There exists $r > 0$ such that $B_r(\mathbf{0}) \subseteq \text{supp } \mu$, where $B_r(\mathbf{0})$ denotes the open Euclidean ball of radius r centered at the zero vector. For each $l \in L$, define $O_l := \{\varepsilon \in B_r(\mathbf{0}) : \varepsilon_l > \varepsilon_{j^*}\}$. The set O_l is nonempty and open. Since $O_l \subseteq \text{supp } \mu$, the definition of support implies $\mu(O_l) > 0$. Because L is finite, we can define $\delta := \min_{l \in L} \mu(O_l) > 0$.

Now fix any coefficient vector $\beta \in \mathbf{R}^K$. By the convex-combination representation of x_{j^*} , $\beta \cdot x_{j^*} = \sum_{l \in L} \lambda_l \beta \cdot x_l \leq \max_{l \in L} \beta \cdot x_l$. Hence there exists $l(\beta) \in L$ such that $\beta \cdot x_{l(\beta)} \geq \beta \cdot x_{j^*}$. On the event $O_{l(\beta)}$, we also have $\varepsilon_{l(\beta)} > \varepsilon_{j^*}$. Therefore, on $O_{l(\beta)}$, $\beta \cdot x_{l(\beta)} + \varepsilon_{l(\beta)} > \beta \cdot x_{j^*} + \varepsilon_{j^*}$. Thus alternative j^* cannot be chosen from J on the event $O_{l(\beta)}$. It follows that, for the ARUM with coefficient vector β and no fixed effects, the probability that j^* is chosen from J is bounded by $1 - \mu(O_{l(\beta)}) \leq 1 - \delta$. This upper bound is uniform in β .

Now consider any random-coefficient ARUM without fixed effects and with mixing distribution m over coefficient vectors. Its probability of choosing j^* from J is

$$\int \mu(\{\varepsilon | \beta \cdot x_{j^*} + \varepsilon_{j^*} > \beta \cdot x_l + \varepsilon_l, \forall l \neq j^*\}) dm(\beta).$$

By the uniform bound above, this integral is at most $1 - \delta$. Hence every element $\rho \in \mathcal{P}_{ra}(0 | \mu)$ satisfies $\rho(J, j^*) \leq 1 - \delta$. Since every element of $\text{cl } \mathcal{P}_{ra}(0 | \mu)$ also satisfies $\rho(J, j^*) \leq 1 - \delta$, $\rho^* \notin \text{cl } \mathcal{P}_{ra}(0 | \mu)$.

Statement (ii) (b) can be proved as follows. Consider any stochastic choice function ρ on $\{J\}$. Then there exists a sequence of stochastic choice functions $\{\rho_n\}$ such that $\rho_n \rightarrow \rho$ and $\rho_n(J, j) > 0$ for any $j \in J$. Fix $\mu \in \mathcal{M}$. By Corollary 1 of Norets and Takahashi (2013), ρ_n can be represented as ARUMs for each n .

³⁷To see this, note that under affine independence of $X(d)$, we have $\mathcal{P}_r \subset \mathcal{P}_r(d) \subset \text{cl } \mathcal{P}_{ml}(d) \subset \mathcal{P}_r$, which implies $\mathcal{P}_r = \mathcal{P}_r(d) = \text{cl } \mathcal{P}_{ml}(d)$. Thus, continuity in Lu-Saito and affine independence of $X(d)$ in our paper play analogous roles: each pinpoints when the degree- d mixed-logit family attains the full approximation frontier relevant to its domain.

D Proof of Lemma 3

D.1 Details of Proof of Step 1 of Lemma 3-(1)

We use the following lemma:

Lemma 8. *Let A be an $r \times n$ real matrix, B be an $l \times n$ real matrix, and E be a real $m \times n$ matrix. Exactly one of the following alternatives is true: (1) There is $u \in \mathbf{R}^n$ such that $Au = 0$, $Bu \geq 0$, $Eu \gg 0$; (2) There is $\theta \in \mathbf{R}^r$, $\eta \in \mathbf{R}^l$, and $\lambda \in \mathbf{R}^m$ such that $\theta A + \eta B + \lambda E = 0$, $\lambda > 0$ and $\eta \geq 0$, where $\gg 0$ means all entries are positive, > 0 means all entries are nonnegative and positive for some entry, and $\geq$ means all entries are nonnegative.*

See Theorem 1.6.1 of Stoer and Witzgall (2012) for the proof.

For simplicity of notation, let $J = \{1, 2, \dots, |J|\}$. For any ranking $\pi \in \Pi$, by relabeling J if necessary, we assume that $\pi(i) > \pi(i+1)$ for all $i \leq |J| - 1$ without loss of generality. We label the following condition as Condition (*): if $\lambda_1 x_1 + \sum_{i=2}^{|J|-1} (\lambda_i - \lambda_{i-1}) x_i - \lambda_{|J|-1} x_{|J|} = 0$ and $\lambda_i \geq 0$ for all $i \in \{1, \dots, |J| - 1\}$, then $\lambda_i = 0$ for all $i \in \{1, \dots, |J| - 1\}$.

Step 1: For each $\pi \in \Pi$, Condition (*) holds if and only if π is representable.

Proof. Since $\mathcal{D}$ contains all binary sets, $\pi \in \Pi$ is representable if and only if there exists β such that for any $j, l \in J$, $\pi(j) > \pi(l) \Leftrightarrow \beta \cdot x_j > \beta \cdot x_l$. Fix $\pi \in \Pi$.

$$\begin{aligned}
& \exists \beta [x_1 \cdot \beta > x_2 \cdot \beta > \dots > x_{|J|-1} \cdot \beta > x_{|J|} \cdot \beta] \\
& \iff \exists \beta [(x_1 - x_2) \cdot \beta > 0, \dots, (x_{|J|-1} - x_{|J|}) \cdot \beta > 0] \iff \exists \beta [E\beta \gg 0] \\
& \iff \nexists \lambda \in \mathbf{R}^{|J|-1} [\lambda > 0, \lambda E = 0] \iff \nexists \lambda \in \mathbf{R}^{|J|-1} [\lambda > 0, \sum_{i=1}^{|J|-1} \lambda_i (x_i - x_{i+1}) = 0] \\
& \iff \nexists \lambda \in \mathbf{R}^{|J|-1} [\lambda > 0, \lambda_1 x_1 + \sum_{i=2}^{|J|-1} (\lambda_i - \lambda_{i-1}) x_i - \lambda_{|J|-1} x_{|J|} = 0] \\
& \iff \text{Condition}(*),
\end{aligned}$$

where $\lambda \equiv (\lambda_1, \dots, \lambda_{|J|-1})$ and the third equivalence is obtained by using Lemma 8 with $A, B = 0$ and $E \equiv (x_1 - x_2; \dots; x_{|J|-1} - x_{|J|}) \in \mathbf{R}^{(|J|-1) \times K}$. $\square$

D.2 Proof of Lemma 3-(2)

For any $j \in J$, $x_j \notin \text{co}(\{x_l | l \in J \setminus \{j\}\}) \Leftrightarrow x_j$ is an extreme point of $\text{co}(\{x_l | l \in J\}) \Leftrightarrow x_j$ is an exposed point of $\text{co}(\{x_l | l \in J\}) \Leftrightarrow \exists \beta \forall l \in J \setminus \{j\} [\beta \cdot x_j > \beta \cdot x_l] \Leftrightarrow$ all rankings π on $\{J\}$, whose best alternative is j , are representable. The first and third equivalences are by the definitions of extreme points and exposed points, respectively, while the second equivalence is by the fact that $\text{co}(\{x_l | l \in J\})$ is a polytope.

E Proof of Lemma 5

For each positive integer n , define $J_n = \{1, \dots, n\}$. We prove the lemma by an induction on n . Let Π_n be the set of all rankings on J_n .

Let π_n be a ranking over J_n such that $\pi_n(i) > \pi_n(i+1)$ for any $1 \leq i \leq n-1$. We will prove that ρ^{π_n} and $\rho^{\pi_n^-}$ are adjacent. Since the labeling of the alternatives is arbitrary, this proves the lemma.

Induction Base: Let us consider the case of $n = 3$. (Remember that all binary and ternary choice sets are in $\mathcal{D}_n$. The cases for $n = 1$ and $n = 2$ are trivial.) WLOG we consider $\pi(1) > \pi(2) > \pi(3)$ and its reverse. For $b > a > 0$, let $t_3(\{1, 2\}, 1) = a$, $t_3(\{2, 3\}, 2) = -b$, $t_3(\{1, 3\}, 1) = b - a$, and $t_3(\{1, 2, 3\}, 2) = a + b$. For all other $(D, j) \in \mathcal{D} \times J$, $t_3(D, j) = 0$. A direct calculation shows that $\rho^{\pi_3} \cdot t_3 = \rho^{\pi_3^-} \cdot t_3 = 0$, and $\rho^{\sigma_3} \cdot t_3 > 0$ for any $\sigma_3 \in \Pi_3 \setminus \{\pi_3, \pi_3^-\}$.

Assume that $n \geq 4$. For each i such that $3 \leq i \leq n-1$, we define a set of sets $\mathcal{D}_i \subset 2^{J_i} \setminus \emptyset$ such that (i) $\mathcal{D}_i \subset \mathcal{D}_{i+1}$ and $\mathcal{D}_n = \mathcal{D}$; (ii) for each i , $\{j, l\} \in \mathcal{D}_i$ and $\{j, l, r\} \in \mathcal{D}_i$ for any $j, l, r \in J_i$.

Induction Step: WLOG, let π_{n-1} be the ranking over J_{n-1} such that $\pi_{n-1}(i) > \pi_{n-1}(i+1)$ for any $i \leq n-2$. By the induction hypothesis there exists $t_{n-1} \in \mathbf{R}^{|\mathcal{D}_{n-1}| \times |J_{n-1}|}$ such that $\rho^{\pi_{n-1}} \cdot t_{n-1} = \rho^{\pi_{n-1}^-} \cdot t_{n-1} = 0$, and $\rho^{\sigma_{n-1}} \cdot t_{n-1} > 0$ for $\sigma_{n-1} \in \Pi_{n-1} \setminus \{\pi_{n-1}, \pi_{n-1}^-\}$. Choose a positive number ε_{n-1} such that $0 < \varepsilon_{n-1} < \min_{\sigma_{n-1} \in \Pi_{n-1} \setminus \{\pi_{n-1}, \pi_{n-1}^-\}} \rho^{\sigma_{n-1}} \cdot t_{n-1}$. We define $t_n \in \mathbf{R}^{|\mathcal{D}_n| \times |J_n|}$ as follows: for each $(D, j) \in \mathcal{D}_n \times J_n$,

$$t_n(D, j) = \begin{cases} t_{n-1}(D, j) & \text{if } (D, j) \in (\mathcal{D}_{n-1} \times J_{n-1}) \setminus (\{1, 2\}, 1), \\ t_{n-1}(D, j) + \varepsilon_{n-1} & \text{if } (D, j) = (\{1, 2\}, 1), \\ -\varepsilon_{n-1}/(n-1) & \text{if } (D, j) = (\{i, n\}, i) \text{ for some } i \in \{1, \dots, n-1\}, \\ 2\varepsilon_{n-1} & \text{if } (D, j) = (\{n-2, n-1, n\}, n-1), \\ 0 & \text{otherwise.} \end{cases}$$

It is clear that $\rho^{\pi_n} \cdot t_n = \rho^{\pi_n^-} \cdot t_n = 0$. Fix $\sigma_n \in \Pi_n \setminus \{\pi_n, \pi_n^-\}$. Let $j \in \{1, \dots, n\}$ be such that n is the j th best alternative under σ_n , so there are exactly $j-1$ alternatives strictly preferred to n in σ_n . Let $\sigma_{n-1} \in \Pi_{n-1}$ be the restriction of σ_n to J_{n-1} .

First notice that by the definition of t_n and $\rho^{\sigma_{n-1}} = \rho^{\sigma_n}$ on $\{1, 2\}$, $\rho^{\sigma_n} \cdot t_n = \rho^{\sigma_{n-1}} \cdot t_{n-1} + \varepsilon_{n-1} \rho^{\sigma_{n-1}}(\{1, 2\}, 1) - \frac{\varepsilon_{n-1}}{n-1}(j-1) + 2\varepsilon_{n-1} \rho^{\sigma_n}(\{n-2, n-1, n\}, n-1)$, where the second term of the right hand side follows since in σ_n , there are $j-1$ elements that are better than n .

Case 1: $\sigma_{n-1} = \pi_{n-1}$. Note that $\rho^{\sigma_{n-1}}(\{1, 2\}, 1) = \rho^{\pi_{n-1}}(\{1, 2\}, 1) = 1$. Note also that $\rho^{\sigma_n}(\{n-2, n-1, n\}, n-1) = 0$. Thus, $\rho^{\sigma_n} \cdot t_n = 0 + \varepsilon_{n-1} - \frac{\varepsilon_{n-1}}{n-1}(j-1) + 0 > 0$, where the last inequality holds because $j < n$. Note that if $j = n$, then $\sigma_{n-1} = \pi_{n-1}$ implies that $\sigma_n = \pi_n$.

Case 2: $\sigma_{n-1} = \pi_{n-1}^-$. Note that $\rho^{\sigma_{n-1}} \cdot t_{n-1} = \rho^{\pi_{n-1}^-} \cdot t_{n-1} = 0$ and $\rho^{\sigma_{n-1}}(\{1, 2\}, 1) =$

$\rho^{\pi_{n-1}^-}(\{1, 2\}, 1) = 0$. Note also that $\rho^{\sigma_n}(\{n-2, n-1, n\}, n-1) = 1$ because $n-1$ is the best element in σ_n (except the case in which n is the best element in σ_n and $\sigma_{n-1} = \pi_{n-1}^-$, then the ranking in σ_n coincides with π_n^-). Thus, $\rho^{\sigma_n} \cdot t_n = 0 + 0 - \frac{\varepsilon_{n-1}}{n-1}(j-1) + 2\varepsilon_{n-1} > 0$.

Case 3: $\sigma_{n-1} \notin \{\pi_{n-1}, \pi_{n-1}^-\}$. Thus, $\rho^{\sigma_n} \cdot t_n > \varepsilon_{n-1} - \frac{\varepsilon_{n-1}}{n-1}(j-1) \geq 0$, where the first inequality holds by $\rho^{\sigma_{n-1}} \cdot t_{n-1} > \varepsilon_{n-1}$ and the second inequality holds by $j \leq n$.

F Worst-case Approximation Error is Achieved at a Deterministic Preference

We consider the worst-case approximation error

$$\sup_{\hat{\rho} \in \mathcal{P}_r} \inf_{\rho \in \mathcal{P}_{ra}(\eta|\mu)} d(\hat{\rho}, \rho), \quad (15)$$

where $d(\cdot, \cdot)$ is the distance function defined in Section 4. We show that the worst-case approximation error is attained at one of the deterministic preferences.

For each $\pi \in \Pi$, let ρ^π denote the stochastic choice function associated with the deterministic preference ranking π .

Proposition 4. *For any $\mu \in \mathcal{M}$ and $\eta \in \mathcal{F}$,*

$$\sup_{\hat{\rho} \in \mathcal{P}_r} \inf_{\rho \in \mathcal{P}_{ra}(\eta|\mu)} d(\hat{\rho}, \rho) = \sup_{\hat{\rho} \in \{\rho^\pi\}_{\pi \in \Pi}} \inf_{\rho \in \mathcal{P}_{ra}(\eta|\mu)} d(\hat{\rho}, \rho).$$

Proof. Given an $\varepsilon > 0$, pick $\hat{\rho}_\varepsilon \in \mathcal{P}_r$ such that

$$\inf_{\rho \in \mathcal{P}_{ra}(\eta|\mu)} d(\hat{\rho}_\varepsilon, \rho) \geq \sup_{\hat{\rho} \in \mathcal{P}_r} \inf_{\rho \in \mathcal{P}_{ra}(\eta|\mu)} d(\hat{\rho}, \rho) - \varepsilon. \quad (16)$$

Since $\mathcal{P}_r$ is the convex hull of $\{\rho^\pi\}_{\pi \in \Pi}$, we can write the vertex representation of $\hat{\rho}_\varepsilon$ as

$$\hat{\rho}_\varepsilon = \sum_{\pi \in \Pi} \lambda_\varepsilon^\pi \rho^\pi,$$

where the weights satisfy $\sum_{\pi \in \Pi} \lambda_\varepsilon^\pi = 1$ and $\lambda_\varepsilon^\pi \geq 0$ for all $\pi \in \Pi$.

For a sufficiently small $\delta > 0$ and for each $\pi \in \Pi$, let $\rho_\delta^\pi \in \mathcal{P}_{ra}(\eta|\mu)$ be such that

$$d(\rho^\pi, \rho_\delta^\pi) \leq \inf_{\rho \in \mathcal{P}_{ra}(\eta|\mu)} d(\rho^\pi, \rho) + \delta. \quad (17)$$

Combining this with (16) and (17) yields

$$\begin{aligned}
\sup_{\hat{\rho} \in \mathcal{P}_r} \inf_{\rho \in \mathcal{P}_{ra}(\eta|\mu)} d(\hat{\rho}, \rho) - \varepsilon &\leq \inf_{\rho \in \mathcal{P}_{ra}(\eta|\mu)} d(\hat{\rho}_\varepsilon, \rho) & (\because (16)) \\
&\leq d\left(\hat{\rho}_\varepsilon, \sum_{\pi \in \Pi} \lambda_\varepsilon^\pi \rho_\delta^\pi\right) \\
&= d\left(\sum_{\pi \in \Pi} \lambda_\varepsilon^\pi \rho^\pi, \sum_{\pi \in \Pi} \lambda_\varepsilon^\pi \rho_\delta^\pi\right) \\
&\leq \sum_{\pi \in \Pi} \lambda_\varepsilon^\pi d(\rho^\pi, \rho_\delta^\pi) & (\because \text{by convexity of } d) \\
&\leq \sum_{\pi \in \Pi} \lambda_\varepsilon^\pi \left(\inf_{\rho \in \mathcal{P}_{ra}(\eta|\mu)} d(\rho^\pi, \rho) + \delta \right) & (\because (17)) \\
&\leq \sup_{\hat{\rho} \in \{\rho^\pi\}_{\pi \in \Pi}} \inf_{\rho \in \mathcal{P}_{ra}(\eta|\mu)} d(\hat{\rho}, \rho) + \delta.
\end{aligned}$$

Since ε and δ are arbitrary, letting $\varepsilon \rightarrow 0$ and $\delta \rightarrow 0$ yields the desired result. $\square$

G Algorithms to Compute Approximation Errors

G.1 EM Algorithm

To compute approximation errors in Table 1, we fit a finite-mixture logit model to each deterministic ranking by the method of maximum likelihood. The data input is the stochastic choice function $\hat{\rho}(D, j)$ and characteristics of each alternative j . We choose the number of mixtures, $M = 18$, according to the theoretical upper bound suggested by Proposition 6 in Section I. Given the number of mixtures, the model has two sets of parameters: (1) mixture weights $\{\lambda_i\}_{i=1}^M$ and (2) coefficients for each mixture $\{\beta_i\}_{i=1}^M$. The log-likelihood function of a finite mixture model with M mixtures is

$$\mathcal{L}(\{\lambda_i\}_{i=1}^M, \{\beta_i\}_{i=1}^M) \equiv \sum_{D \in \mathcal{D}} \sum_{j \in D} \hat{\rho}(D, j) \log \sum_{i=1}^M \lambda_i \frac{\exp(\beta_i \cdot x_j)}{\sum_{l \in D} \exp(\beta_l \cdot x_l)}.$$

We estimate the parameters by the EM algorithm (Dempster, Laird, and Rubin, 1977; Train, 2009). We implement the algorithm according to Chapter 14 in Train (2009). We terminate the algorithm when the change of the implied l_2 distance between the estimated choice probability and the target choice probability becomes smaller than $\frac{1}{10^6}$ between two successive iterations.

Our use of the maximum likelihood method with the EM algorithm is motivated by the following observation: if the affine-independence condition is satisfied and the target choice probability is an interior random utility model $\hat{\rho} \in \text{rint } \mathcal{P}_r$, then the model that maximizes the likelihood will yield a perfect fit to the target probability. The maximum likelihood method

therefore minimizes the approximation error metric in (7).

To see this, notice that under the affine-independence condition, Proposition 6 and 7 in Section I of the online appendix imply that any interior random utility model can be represented by a finite mixture of logit models with $M = \sum_{D \in \mathcal{D}} (|D| - 1) + 1$ mixtures. That is, if $M = \sum_{D \in \mathcal{D}} (|D| - 1) + 1$, there exists a set of parameters $\{(\beta_i^*, \lambda_i^*)\}_{i=1}^M$ such that $\sum_{i=1}^M \lambda_i^* \frac{\exp(\beta_i^* \cdot x_j)}{\sum_{l \in D} \exp(\beta_i^* \cdot x_l)} = \hat{\rho}(D, j)$ for any $D \in \mathcal{D}$, $j \in D$. The choice probability generated by this set of parameters yields a perfect fit of the target probability. Hence this set of parameters maximizes the likelihood.³⁸

G.2 Greedy Algorithm

It is known that the EM algorithm may not converge to the global optimum. To alleviate this concern, we propose a second greedy algorithm inspired by Barron, Cohen, Dahmen, and DeVore (2008). This algorithm solves a sequence of optimization problems and converges to the global optimal solution provided each inner problem is solved correctly. The structure of the random-coefficient RUMs is important for the proof.

The algorithm takes a stochastic choice function $\hat{\rho}$ and a fixed effects vector η as input and returns a solution to (7). The algorithm is iterative: each step seeks to optimize based on the results of previous steps:

- **Step 1:** Given $\hat{\rho}$, choose ρ^1 such that $\rho^1 = \arg \inf_{\rho \in \text{cl } \mathcal{P}_a(\eta|\mu)} \|\hat{\rho} - \rho\|^2$.
- **Step n, $n \geq 2$:**
 - Consider a set of grids $\alpha_n = \{\frac{2}{i+1}\}_{i=1}^n$.
 - Find $(\alpha_n^*, \rho_n^*) = \arg \inf_{(\alpha, \rho) \in \alpha_n \times \text{cl } \mathcal{P}_a(\eta|\mu)} \|\hat{\rho} - (1 - \alpha)\rho^{n-1} - \alpha\rho\|^2$.
 - Define $\rho^n = (1 - \alpha_n^*)\rho^{n-1} + \alpha_n^*\rho_n^*$ and let $\rho^{out} = \rho^n$.
- Stop if a terminating criterion is reached.
- Return ρ^{out} at the final step.

The next proposition shows that the algorithm will converge for our problems in Section 4, provided each inner problem is solved correctly. Define $d(\rho, \hat{\rho})$ as in (7).

³⁸Recall that for any other stochastic choice function ρ , the likelihood is $\sum_{D \in \mathcal{D}} \sum_{j \in D} \hat{\rho}(D, j) \log(\rho(D, j))$. Note that $\hat{\rho}$ maximizes the likelihood since

$$\begin{aligned} & \sum_{D \in \mathcal{D}} \sum_{j \in D} \hat{\rho}(D, j) (\log(\hat{\rho}(D, j)) - \log(\rho(D, j))) = \sum_{D \in \mathcal{D}} \sum_{j \in D} \hat{\rho}(D, j) \log \frac{\hat{\rho}(D, j)}{\rho(D, j)} \\ &= - \sum_{D \in \mathcal{D}} \sum_{j \in D} \hat{\rho}(D, j) \log \frac{\rho(D, j)}{\hat{\rho}(D, j)} \geq - \sum_{D \in \mathcal{D}} \sum_{j \in D} \hat{\rho}(D, j) \left(\frac{\rho(D, j)}{\hat{\rho}(D, j)} - 1 \right) \\ &= - \sum_{D \in \mathcal{D}} \sum_{j \in D} \rho(D, j) + \sum_{D \in \mathcal{D}} \sum_{j \in D} \hat{\rho}(D, j) = -|\mathcal{D}| + |\mathcal{D}| = 0, \end{aligned}$$

where we use the fact $-\log(x) \geq -(x - 1)$ when $x \geq 0$.

Proposition 5. Let $\hat{\rho} \in \mathcal{P}$ be any stochastic choice function, $\eta \in \mathbf{R}^{|\mathcal{J}|}$, and $\mu \in \mathcal{M}$. Define $d^* = \inf_{\rho \in \mathcal{P}_{ra}(\eta|\mu)} d(\rho, \hat{\rho})$. Let n denote the number of steps and ρ^n denote the output after the completion of the n -th step of the algorithm. Then,

$$d(\rho^n, \hat{\rho}) - d^* \leq \sqrt{\frac{8}{n+1}}. \quad (18)$$

For our implementation, the terminating criterion is when the number of steps taken reaches 1000. With 1000 steps, (18) implies the margin of error is within 0.09.³⁹ When we approximate $\hat{\rho}$ without fixed effects, we let $\eta = 0$. When we approximate $\hat{\rho}$ with fixed effects, we couple the algorithm with a grid of fixed effects to search for the minimum.

Proof. Since the set $\text{clco}\mathcal{P}_a(\eta|\mu)$ is compact and convex, $\rho^* = \arg \inf_{\rho \in \text{clco}\mathcal{P}_a(\eta|\mu)} d(\rho, \hat{\rho}) = \arg \inf_{\rho \in \text{clco}\mathcal{P}_a(\eta|\mu)} \|\rho - \hat{\rho}\|_2^2$ exists and it can be written as a convex combination of elements of $\text{cl}\mathcal{P}_a(\eta|\mu)$. By Caratheodory's theorem, it can be written as $\rho^* = \sum_{i=1}^M \lambda_i \rho_i$, where $\lambda_i \geq 0$, $\sum_i \lambda_i = 1$ and $\rho_i \in \text{cl}\mathcal{P}_a(\eta|\mu)$, $M = \dim \mathcal{P}_r + 1$. For each step n , define $E_n = \|\hat{\rho} - \rho^n\|^2 - \|\hat{\rho} - \rho^*\|^2$, where $\|\cdot\|$ denotes the Euclidean l_2 norm. For each step n , let α_n^* and ρ_n^* be the minimizers over the grids $\{\alpha_n\}$ and $\text{cl}\mathcal{P}_a(\eta|\mu)$, respectively. Define $C = \sum_i \lambda_i \|\rho_i - \rho^*\|^2$ and $T = \max\{2E_1, 4C\}$. E_1 and C can be upper-bounded by the squared diameter of the random utility polytope. For example, $C \leq \sum_i \lambda_i \|\rho_i - \rho^*\|^2 \leq \sup_{\rho, \hat{\rho} \in \mathcal{P}_r} \|\rho - \hat{\rho}\|_2^2 = 2 \times \text{the number of choice sets}$. The extremum is achieved by selecting ρ to be a degenerate ranking and $\hat{\rho}$ its reverse ranking. Similarly E_1 can be upper bounded by $2\|\hat{\rho}\|^2 + 2\|\rho^n\|^2 \leq 4|\mathcal{D}|$. Hence we can choose $T = 8|\mathcal{D}|$. Notice this implies $T = 8 \times (\text{the number of choice sets})$.

For each step n , let $\alpha_n = \frac{2}{n+1}$. Then, we have the following:

$$\begin{aligned} E_n &= \|\hat{\rho} - (1 - \alpha_n^*)\rho^{n-1} - \alpha_n^* \rho_n^*\|^2 - \|\hat{\rho} - \rho^*\|^2 \\ &\leq \sum_i \lambda_i \|\hat{\rho} - (1 - \alpha_n)\rho^{n-1} - \alpha_n \rho_i\|^2 - \|\hat{\rho} - \rho^*\|^2 \\ &= \sum_i \lambda_i \{(1 - \alpha_n)^2 \|\hat{\rho} - \rho^{n-1}\|^2 + 2\alpha_n(1 - \alpha_n)((\hat{\rho} - \rho^{n-1}) \cdot (\hat{\rho} - \rho_i)) + \alpha_n^2 \|\hat{\rho} - \rho_i\|^2\} \\ &\quad - \|\hat{\rho} - \rho^*\|^2 \\ &= (1 - \alpha_n)^2 \|\hat{\rho} - \rho^{n-1}\|^2 + 2\alpha_n(1 - \alpha_n)((\hat{\rho} - \rho^{n-1}) \cdot (\hat{\rho} - \sum_i \lambda_i \rho_i)) \\ &\quad + \alpha_n^2 \sum_i \lambda_i \|\hat{\rho} - \rho_i\|^2 - \|\hat{\rho} - \rho^*\|^2 \\ &\leq (1 - \alpha_n)^2 \|\hat{\rho} - \rho^{n-1}\|^2 + \alpha_n(1 - \alpha_n)(\|\hat{\rho} - \rho^{n-1}\|^2 + \|\hat{\rho} - \rho^*\|^2) - \|\hat{\rho} - \rho^*\|^2 \\ &\quad + \alpha_n^2 \sum_i \lambda_i \|\hat{\rho} - \rho_i\|^2 \\ &\leq (1 - \alpha_n)^2 \|\hat{\rho} - \rho^{n-1}\|^2 + \alpha_n(1 - \alpha_n)(\|\hat{\rho} - \rho^{n-1}\|^2 + \|\hat{\rho} - \rho^*\|^2) - \|\hat{\rho} - \rho^*\|^2 \\ &\quad + \alpha_n^2 \sum_i \lambda_i \|\rho_i - \rho^*\|^2 + \alpha_n^2 \|\rho^* - \hat{\rho}\|^2 \\ &= (1 - \alpha_n) \|\hat{\rho} - \rho^{n-1}\|^2 - (1 - \alpha_n) \|\hat{\rho} - \rho^*\|^2 + \alpha_n^2 \sum_i \lambda_i \|\rho_i - \rho^*\|^2 \\ &= (1 - \alpha_n) E_{n-1} + \alpha_n^2 \sum_i \lambda_i \|\rho_i - \rho^*\|^2. \end{aligned}$$

³⁹This is calculated by noting $\sqrt{8/1001} \approx 0.09$.

Thus we have

$$E_n \leq (1 - \alpha_n)E_{n-1} + C\alpha_n^2. \quad (19)$$

In the following, we will show $E_n \leq \frac{T}{n+1}$ for each n . We prove this by induction. The inequality holds with $n = 1$. Fix n . Suppose $E_{n-1} \leq \frac{T}{n}$. By substituting $\alpha_n = \frac{2}{n+1}$ to (19), we have
(i): $E_n \leq \frac{T}{n+1}$.⁴⁰ Let $d^n = d(\hat{\rho}, \rho^n)$ and $d^* = d(\hat{\rho}, \rho^*)$. Since $E_n = |\mathcal{D}|((d^n)^2 - (d^*)^2)$, we have
(ii): $(d^n)^2 - (d^*)^2 \leq \frac{T'}{n+1}$, where $T' = \frac{T}{|\mathcal{D}|} = 8$. Then we have $(d^n - d^*)^2 \leq (d^n - d^*)(d^n + d^*) = (d^n)^2 - (d^*)^2 \leq \frac{8}{n+1}$, where we use the fact that $d^n \geq d^*$ and $d^* \geq 0$. This implies $d^n - d^* \leq \sqrt{\frac{8}{n+1}}$. $\blacksquare$

G.3 Calculating the Maximal Substitution in (8)

To calculate the maximal substitution (8), we consider the problem

$$\inf_{\rho \in \mathcal{P}_{ra}(0|\mu)} \left(\sum_{r \in J \setminus \{j, l\}} \rho(J \setminus \{j\}, r) + \rho(J, l) \right)^2, \quad (20)$$

which can be readily solved by the greedy algorithm. Taking 1 minus the square root of the minimized value in (20) gives the solution to the problem in (8). To see this, notice

$$\sup_{\rho \in \mathcal{P}_{ra}(0|\mu)} (\rho(J \setminus \{j\}, l) - \rho(J, l)) = 1 - \inf_{\rho \in \mathcal{P}_{ra}(0|\mu)} \left(\sum_{r \in J \setminus \{j, l\}} \rho(J \setminus \{j\}, r) + \rho(J, l) \right).$$

Because $\sum_{r \in J \setminus \{j, l\}} \rho(J \setminus \{j\}, r) + \rho(J, l)$ is nonnegative, minimizer(s) of $\sum_{r \in J \setminus \{j, l\}} \rho(J \setminus \{j\}, r) + \rho(J, l)$ is the same as the minimizer(s) of the problem when the criterion is squared.

We modify the greedy algorithm for maximal substitution as follows:

- **Step 1:** Choose ρ^1 as a solution of $\inf_{\rho \in \text{cl } \mathcal{P}_a(0|\mu)} \left(\sum_{r \in J \setminus \{j, l\}} \rho(J \setminus \{j\}, r) + \rho(J, l) \right)^2$.

- **Step n, $n \geq 2$:**

- Consider a set of grids $\alpha_n = \{\frac{2}{i+1}\}_{i=1}^n$.
- Find (α_n^*, ρ_n^*) as a solution of

$$\inf_{(\alpha, \rho) \in \alpha_n \times \text{cl } \mathcal{P}_a(0|\mu)} \left((1-\alpha) \left(\sum_{r \in J \setminus \{j, l\}} \rho^{n-1}(J \setminus \{j\}, r) + \rho^{n-1}(J, l) \right) + \alpha \left(\sum_{r \in J \setminus \{j, l\}} \rho(J \setminus \{j\}, r) + \rho(J, l) \right) \right)^2.$$

- Define $\rho^n = (1 - \alpha_n^*)\rho^{n-1} + \alpha_n^*\rho_n^*$ and let $\rho^{out} = \rho^n$.

- Stop if a terminating criterion is reached.

- Return ρ^{out} at the final step.

$$^{40}E_n \leq \frac{n-1}{n+1} \frac{T}{n} + C \frac{4}{(n+1)^2} = \frac{(n^2-1)T + 4Cn}{(n+1)^2n} \leq \frac{(n^2-1)T + Tn}{(n+1)^2n} \leq \frac{n^2T + Tn}{(n+1)^2n} = \frac{Tn(n+1)}{(n+1)^2n} = \frac{T}{n+1}.$$

H Geometric Intuition for Theorem

H.1 Affine Independence Condition

We first provide the intuition behind Lemma 3. To understand Lemma 3 (1) geometrically, see Figure A.1. In the figure, we assume that there are two original characteristic variables, say (p_j, q_j) for each alternative $j \in J$. In Figure A.1 (a) and (b), we consider the models with the original characteristics (i.e., $K = 2$ and $x_j = (p_j, q_j)$ for each $j \in J$). In Figure A.1 (a), the set $\{x_1, x_2, x_3\}$ is affinely independent. Thus, by Lemma 3 (1) (the “if” part), any ranking is representable. For example, the ranking $\pi(1) > \pi(2) > \pi(3)$ is representable by $\beta \in \mathbf{R}^2$, which defines the parallel hyperplanes (indifference curves) in Figure A.1 (a). On the other hand, in Figure A.1 (b), the set $\{x_1, x_2, x_3, x_4\}$ is not affinely independent. The ranking $\pi(1) > \pi(4) > \pi(3) > \pi(2)$ is not representable. As the figure shows, no matter how one chooses $\beta \in \mathbf{R}^2$ and draws parallel hyperplanes as indifference curves, it does not hold that $\beta \cdot x_1 > \beta \cdot x_4 > \beta \cdot x_3 > \beta \cdot x_2$. The existence of such an unrepresentable ranking is implied by the “only if” part of Lemma 3 (1).

If we use ellipses as indifference curves, however, we can represent the ranking $\pi(1) > \pi(4) > \pi(3) > \pi(2)$ as in Figure A.1 (c).⁴¹ The existence of such curves is again implied by the “if” part of Lemma 3 (1) since ellipses can be defined with the quadratic polynomials $\beta \cdot x_j$ with $x_j = (p_j, q_j, p_j^2, q_j^2, p_j q_j)$. Moreover, the generic condition with quadratic polynomials is satisfied (i.e., $K = 5 \geq 3 = |J| - 1$) in this example. Lemma 3 (2) is more straightforward.⁴² By using Lemmas 1, 2, and 3, we obtain statement (i) of Theorem 1 and Proposition 3.

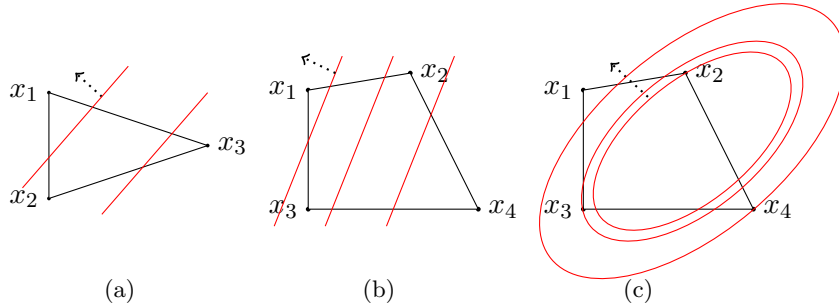


Figure A.1: Illustration of the affine-independence condition.

⁴¹As the radius of the ellipses becomes larger, the represented utility of the alternatives become larger.

⁴²To see this, notice that when $\mathcal{D} = \{J\}$, any $\pi \in \Pi$ is representable in $\{J\}$ if and only if, for any $j \in J$, there exists β such that $\beta \cdot x_j > \beta \cdot x_l$ for all $l \in J \setminus \{j\}$, which means that J is convex-independent.

H.2 The Set of Random Utility Models

Let $\mathcal{P}_r := \text{co}\{\rho^\pi : \pi \in \Pi\}$ denote the polytope of random utility models; its vertices are the deterministic choice rules ρ^π for $\pi \in \Pi$ (left panel of Figure A.2). Each point in $\mathcal{P}_r$ corresponds to a random utility model. The figure is heuristic: each model is a high-dimensional vector indexed by pairs (D, j) with $j \in D \in \mathcal{D}$. It should not be confused with the characteristics space in Figure A.1.

Our question is: under what conditions can all points in $\mathcal{P}_r$ be represented or approximated by random-coefficient ARUMs with fixed effects? The analyst may choose a standard distribution $\mu \in \mathcal{M}$ and a fixed-effect vector $\eta \in \mathcal{F}$. Each pair (η, μ) induces a (convex) set of random-coefficient ARUMs with fixed effects, denoted $\mathcal{P}_{ra}(\eta | \mu)$. Different choices of η and μ generate different subsets of $\mathcal{P}_r$ (right panel of Figure A.2).

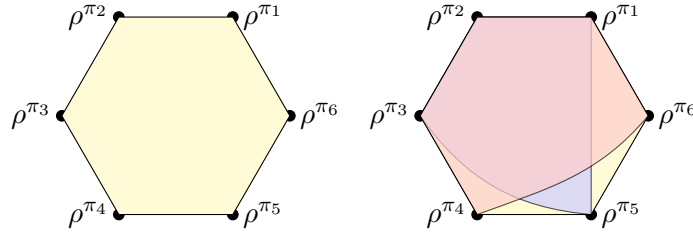


Figure A.2: Illustration of $\mathcal{P}_r$ (yellow), $\mathcal{P}_{ra}(\eta_1 | \mu)$ (red), and $\mathcal{P}_{ra}(\eta_2 | \mu)$ (blue)

We consider the union $\bigcup_{\eta \in \mathcal{F}} \mathcal{P}_{ra}(\eta | \mu)$ given arbitrarily chosen standard probability measure $\mu \in \mathcal{M}$ and ask whether it can approximate the entire polytope $\mathcal{P}_r$. Lemma 6 (Appendix) shows that if an unrepresentable ranking exists, then some random utility models cannot be approximated arbitrarily well—regardless of the choice of fixed effects or probability distributions.

I Additional Theoretical Results

This section provides an additional theoretical result that upper-bounds the number of mixtures required for best possible approximation. The results in this section determine the number of mixtures to be used in the EM algorithm.

The first proposition (Proposition 6) implies that any random-coefficient ARUM can be represented as a finite mixture of ARUMs. The second proposition (Proposition 7) gives us an upper bound on the number of mixtures required by calculating the dimension of the set of random utility models. To prove Proposition 6, we need the following lemma:

Lemma 9. *Let K be a fixed integer. For any bounded Borel set $C \subset \mathbf{R}^K$, let $\Delta(C)$ denote the*

set of Borel probability measures over C .⁴³ Then,

$$\text{co } C = \left\{ \int x dm(x) \mid m \in \Delta(C) \right\},$$

where $\int x dm(x)$ denotes the K -dimensional vector whose l -th element is $\int x(l) dm(x)$ for any $l \in \{1, \dots, K\}$.

The proof is in the next section.⁴⁴ Recall the definition of $\mathcal{P}_{ra}(\eta|\mu)$ and $\mathcal{P}_a(\eta|\mu)$ from Definition 3. By using Lemma 9, we can show the following results that we use in the main body of the paper. Let $\dim \mathcal{P}_r$ denote the dimension of the set of the random utility models.

Proposition 6. *For any (η, μ) ,*

$$\begin{aligned} \mathcal{P}_{ra}(\eta|\mu) &= \text{co } \mathcal{P}_a(\eta|\mu) \\ &= \left\{ \sum_{m=1}^{\dim \mathcal{P}_r + 1} \lambda_m \rho_m \mid \rho_m \in \mathcal{P}_a(\eta|\mu), \lambda_m \geq 0, \forall m = 1, \dots, \dim \mathcal{P}_r + 1, \sum_{m=1}^{\dim \mathcal{P}_r + 1} \lambda_m = 1 \right\}. \end{aligned}$$

Proof. The second equality follows from the Caratheodory's theorem. To show the first equality, fix (η, μ) . The inclusion $\text{co } \mathcal{P}_a(\eta|\mu) \subset \mathcal{P}_{ra}(\eta|\mu)$ holds because every discrete probability measure is a Borel probability measure. In the following, we will show $\mathcal{P}_{ra}(\eta|\mu) \subset \text{co } \mathcal{P}_a(\eta|\mu)$.

Define a mapping $F : \mathbf{R}^K \rightarrow \mathcal{P}_a(\eta|\mu) \subset \mathbf{R}^{|\mathcal{D}| \times J}$ by

$$F(\beta)(D, j) = \mu \left(\varepsilon | \beta \cdot x_j + \eta_{(D,j)} + \varepsilon_j > \beta \cdot x_l + \eta_{(D,l)} + \varepsilon_l, \forall l \in D \setminus \{j\} \right), \quad (21)$$

and 0 if $j \notin D$. The mapping is continuous by the dominated convergence theorem given the fact that μ is absolutely continuous with respect to the Lebesgue measure.

First note that $\mathcal{P}_a(\eta|\mu)$ is a Borel set because $\mathcal{P}_a(\eta|\mu)$ is the image of the continuous mapping F from $\mathbf{R}^K$ to $\mathbf{R}^{|\mathcal{D}| \times J}$ defined in the proof. In particular, $\mathcal{P}_a(\eta|\mu)$ is the countable union of closed images of continuous mapping of compact cubes in $\mathbf{R}^K$.

To show $\mathcal{P}_{ra}(\eta|\mu) \subset \text{co } \mathcal{P}_a(\eta|\mu)$, let $\mathcal{P}_a(\eta|\mu)$ be the set C in Lemma 9, then we have $\text{co } \mathcal{P}_a(\eta|\mu) = \left\{ \int \rho dm(\rho) \mid m \in \Delta(\mathcal{P}_a(\eta|\mu)) \right\}$. It suffices to show $\mathcal{P}_{ra}(\eta|\mu) \subset \left\{ \int \rho dm(\rho) \mid m \in \Delta(\mathcal{P}_a(\eta|\mu)) \right\}$.

Fix a given $m \in \Delta(\mathbf{R}^K)$, we show $\int \rho_\beta dm(\beta) \in \left\{ \int \rho dm(\rho) \mid m \in \Delta(\mathcal{P}_a(\eta|\mu)) \right\}$, where $\rho_\beta \in \mathcal{P}_a(\eta|\mu)$. For any Borel set $C \subset \mathcal{P}_a(\eta|\mu)$, define $\hat{m}(C) = m(F^{-1}(C))$, where $F^{-1}(C)$ is a Borel set because F is continuous. Then, $\hat{m} \in \Delta(\mathcal{P}_a(\eta|\mu))$ and we have $\int \rho_\beta dm(\beta) = \int \rho d\hat{m}(\rho)$, as

⁴³In particular, $m(C) = 1$ for any $m \in \Delta(C)$.

⁴⁴Observe that the result is not true in an infinite dimensional space. To see this, let $\{e_i\}_{i=1}^\infty$ be the base of the infinite-dimensional real space. Define $C = \{e_i\}_{i=1}^\infty$. Define a measure m on C such that $m(e_i) = (1/2)^i$ for each i . Then, $\sum_{i=1}^\infty m(e_i) = 1$, so that m is a probability measure on C . $\int x dm$ cannot be represented as any finite mixture of elements of C . For any $y \in \text{co } C$, there exists i such that $y(e_i) = 0$.

desired. $\square$

We now calculate the number $\dim \mathcal{P}_r$. We note that the set $\mathcal{P}_r$ defined in Definition 1 is associated with a set of choice sets $\mathcal{D}$.

Proposition 7. $\dim \mathcal{P}_r = \sum_{D \in \mathcal{D}} (|D| - 1)$.

The proof is in a later section. Propositions 6 and 7 imply that in order to obtain the best approximating random-coefficient model to the observed choice probabilities, it is sufficient to consider finite mixture models with at most $1 + \sum_{D \in \mathcal{D}} (|D| - 1)$ mixtures. For example, in Section 5, we analyze a choice dataset with $|J| = 4$ with $\mathcal{D} = 2^J \setminus \{\emptyset\}$. These results imply that it is enough to consider the finite mixture models with at most 18 mixtures if one considers the whole choice sets.

I.1 Proof of Lemma 9

By definition, $\text{co } C$ is a subset of $\{\int x dm(x) | m \in \Delta(C)\}$. To show the reverse direction, we first establish a relaxed statement: $\left\{ \int x dm(x) | m \in \Delta(C) \right\} \subset \text{cl co } C$. Suppose by way of contradiction that $\int x dm(x) \notin \text{cl co } C$ for some $m \in \Delta(C)$. By the separating hyperplane theorem (Corollary 11.4.2 of Rockafellar (2015)), there exist a $t \in \mathbf{R}^K \setminus \{0\}$ and $\alpha \in \mathbf{R}$ such that $(\int x dm(x)) \cdot t = \alpha > x \cdot t$ for any $x \in \text{cl co } C$. This is a contradiction because $\alpha = (\int x dm(x)) \cdot t = \int (x \cdot t) dm(x) < \int \alpha dm(x) = \alpha$.

We now prove $\left\{ \int x dm(x) | m \in \Delta(C) \right\} \subset \text{co } C$ by an induction on the dimension of $\text{co } C$.

Induction Base: If $\dim \text{co } C = 1$, there exist $y = \inf\{x | x \in \text{co } C\}$ and $z = \sup\{x | x \in \text{co } C\}$ such that $\text{co } C$ can be represented as the line segment between y and z . We consider the case where the line segment contains neither y nor z . Proofs for the other cases are similar. For any $x \in C$, there exists a weighting function $\alpha(x) = \frac{z-x}{z-y} \in (0, 1)$ such that $x = \alpha(x)y + (1 - \alpha(x))z$. The function α is bounded, nonnegative, and continuous in x . Hence it is measurable and integrable.

Choose any $m \in \Delta(C)$. By continuity from below of measures, there exists l and u such that $y < l < u < z$ and $m(\{x | x \in (l, u)\}) \geq 1 - \varepsilon$ for some $\varepsilon < 1$. Note here we use the fact that $m(C) = 1$ and $m(\{y\}) = m(\{z\}) = 0$ because we are assuming that the line segment does not contain both y and z . Note that $\alpha(x)$ is uniformly bounded away from 0 and 1 for $x \in (l, u)$. Then it follows that $\int \alpha dm = \int \alpha(x) dm(x)$ exists and $0 < \int \alpha(x) dm(x) < 1$. Then, $\int x dm(x) = \int \alpha(x) dm(x) \times y + \int (1 - \alpha(x)) dm(x) \times z \in \text{co } C$, as desired.

Let $l \geq 2$ be an integer.

Induction Step: Suppose that $\left\{ \int x dm(x) | m \in \Delta(C) \right\} \subset \text{co } C$ holds for any C such that $\dim \text{co } C \leq l$. Then it holds for any C such that $\dim \text{co } C = l + 1$.

To prove the step, choose any $m \in \Delta(C)$. We have $\int x dm(x) \in \text{cl co } C$. First consider the case where $\int x dm(x) \in \text{rint cl co } C$. Since $\text{rint cl co } C = \text{rint co } C$ (by Theorem 6.3 of Rockafellar (2015)), we have $\int x dm(x) \in \text{co } C$, as desired.

Next consider the case where $\int x dm(x) \notin \text{rint cl co } C$. Then, $\int x dm(x) \in \partial \text{cl co } C$ and by Theorem 11.6 of Rockafellar (2015), there exists a hyperplane H of $\text{cl co } C$ at $\int x dm(x)$ such that $\text{cl co } C \not\subset H$. There exist $t \in \mathbf{R}^K \setminus \{0\}$ and $\alpha \in \mathbf{R}$ such that $H = \{x | x \cdot t = \alpha\}$ and $\int x dm(x) \cdot t = \alpha > x \cdot t$ for any $x \in \text{cl co } C \cap H^c$. This implies that $m(H) = 1$ and hence $m(H \cap C) = 1$. Since H is a supporting hyperplane and $\text{cl co } C \not\subset H$, we obtain $\dim(H \cap \text{aff } C) \leq l$. Hence, $\dim(\text{co}(H \cap C)) \leq l$ because $\text{co}(H \cap C) \subset \text{co } H \cap \text{co } C \subset H \cap \text{aff } C$. Therefore, the induction hypothesis shows that $\int x dm(x) \in \text{co}(H \cap C) \subset \text{co } C$, as desired.

I.2 Proof of Proposition 7

To prove Proposition 7, we prove two lemmas. Lemma 10 is a technical lemma that facilitates the characterization of the affine hull of random utility polytopes in Lemma 11. The dimension of a set is defined as the dimension of its affine hull, and Proposition 7 follows. Recall that we use $\rho^\pi \in \mathbf{R}^{|\mathcal{D}| \times |J|}$ to denote the stochastic choice function associated with the strict preference ranking $\pi \in \Pi$ where Π is defined in Section 2.

Lemma 10. *For any $t \in \mathbf{R}^{|\mathcal{D}| \times |J|}$, $\rho^\pi \cdot t = \rho^{\pi'} \cdot t$ for all $\pi, \pi' \in \Pi$ if and only if $t(D, j) = t(D, l)$ for all $D \in \mathcal{D}$ and $j, l \in D$.⁴⁵*

Proof. To prove the if part, assume $t(D, j) = t(D, l)$ for all $D \in \mathcal{D}$ and $j, l \in D$. Define $t(D) = t(D, j)$ for any $j \in D$. Then for any $\pi \in \Pi$, $\rho^\pi \cdot t = \sum_{D \in \mathcal{D}} \sum_{j \in D} \rho^\pi(D, j) t(D, j) = \sum_{D \in \mathcal{D}} t(D) \sum_{j \in D} \rho^\pi(D, j) = \sum_{D \in \mathcal{D}} t(D)$, completing the proof of the if part.

The only-if direction is trivially true when the set D contains only one element. We prove the only-if direction for the remaining D , $|D| \geq 2$, by induction. Consider the sets in $\mathcal{D}$ with a size greater or equal to 2. Let m be the smallest cardinality of the sets in $\mathcal{D}$, $m = \min\{|D| | D \in \mathcal{D}, |D| \geq 2\}$.⁴⁶

Induction Base: For any $D \in \mathcal{D}$ such that $|D| = m$ and any $j, l \in D$, $t(D, j) = t(D, l)$.

Proof. To prove the claim, choose any $\pi, \pi' \in \Pi$ such that for $\pi(J \setminus D) > \pi(j) > \pi(l) > \pi(D \setminus \{j, l\})$, and $\pi'(J \setminus D) > \pi'(l) > \pi'(j) > \pi'(D \setminus \{j, l\})$.

Note that the only choice set that has different choice probabilities under π and π' is D . For all the remaining choice sets, either only one of j or l is in the set, or there exists an element that dominates j and l under both π and π' . Notice here that we used the fact that D has the smallest cardinality. Moreover, $\rho^\pi(D, j) = 1$, $\rho^\pi(D, r) = 0$ for any $r \in D \setminus \{j\}$ and $\rho^{\pi'}(D, l) = 1$, $\rho^{\pi'}(D, r) = 0$ for any $r \in D \setminus \{l\}$.

⁴⁵We are identifying each $\rho \in \mathcal{P}$ as an element of $\mathbf{R}^{\mathcal{D} \times J}$.

⁴⁶Note we do not induct from $|D| = 1$ because $\mathcal{D}$ may not contain any set with one element.

Hence, since $t \cdot \rho^\pi = t \cdot \rho^{\pi'}$,

$$\begin{aligned} 0 &= \sum_{(E,r) \in \mathcal{D} \times J} t(E,r) (\rho^\pi(E,r) - \rho^{\pi'}(E,r)) = \sum_{r \in D} t(D,r) (\rho^\pi(D,r) - \rho^{\pi'}(D,r)) \\ &= t(D,j) - t(D,l). \end{aligned}$$

This completes the proof of the induction base. $\square$

Let $k \geq m$.

Induction Step: Suppose that for any $D \in \mathcal{D}$ such that $|D| \leq k$ and any $j, l \in D$, $t(D,j) = t(D,l)$. Then the same claim holds for any $D \in \mathcal{D}$ such that $|D| = k+1$.

Proof. By the induction hypothesis, for any $E \in \mathcal{D}$, if $|E| \leq k$ then $t(E,j) = t(E,l)$ for any $j, l \in E$. Repeat the proof above by considering the choice sets $\mathcal{D} \setminus \{E \in \mathcal{D} \mid |E| \leq k\}$. $\square$

■

Lemma 11. *The affine hull of $\mathcal{P}_r$ is*

$$\mathcal{P}_\pm \equiv \{q \in \mathbf{R}^{|\mathcal{D}| \times |J|} \mid (i) \sum_{j \in D} q(D,j) = 1, \forall D \in \mathcal{D}; (ii) q(D,j) = 0, \forall j \notin D \in \mathcal{D}\}.$$

Proof. The set $\mathcal{P}_\pm$ is affine. So it suffices to show that for any affine set A , if $\mathcal{P}_r \subset A$, then $\mathcal{P}_\pm \subset A$. For any affine set A , by Theorem 1.4 of Rockafellar (2015), it has a representation $A = \{q \in \mathbf{R}^{|\mathcal{D}| \times |J|} \mid Bq = b\}$, where B is a $L \times (|\mathcal{D}| \times |J|)$ matrix, b is a L -dimensional vector, and L is an arbitrary positive integer.

For any $l \in \{1, \dots, L\}$, let $B_l(D,j)$ denote the $(l, (D,j))$ entry of B . Note that each column of B is associated with a $(D,j) \in \mathcal{D} \times J$. So $Bq = b$ means that for any row index $l \in \{1, \dots, L\}$,

$$\sum_{D \in \mathcal{D}} \sum_{j \in J} B_l(D,j) q(D,j) = b_l. \quad (22)$$

By assuming $\mathcal{P}_r \subset A = \{q \in \mathbf{R}^{|\mathcal{D}| \times |J|} \mid Bq = b\}$, we will show that if q satisfies (i) and (ii), then (22) holds for any $l \in \{1, \dots, L\}$.

Step 1: We show that $B_l(D,j) = B_l(D,r)$ for any $l \in \{1, \dots, L\}$, $D \in \mathcal{D}$, and $j, r \in D$. For any $\pi \in \Pi$, $\rho^\pi \in \mathcal{P}_r \subset A = \{q \in \mathbf{R}^{|\mathcal{D}| \times |J|} \mid Bq = b\}$. Hence, (22) holds with $q = \rho^\pi$ for any $\pi \in \Pi$. Thus $\rho^\pi \cdot B_l = \rho^{\pi'} \cdot B_l$ for any $\pi, \pi' \in \Pi$ for any l . By Lemma 10, this implies that $B_l(D,j) = B_l(D,r)$ for any $D \in \mathcal{D}$, and $j, r \in D$.

By Step 1, we can define $B_l(D) = B_l(D,j)$ for any $j \in D$.

Step 2: If q satisfies property (i) and (ii), choose any $\pi \in \Pi$ and $l \in \{1, \dots, L\}$. Since $\rho^\pi \in A$,

then by (22),

$$b_l = \sum_{D \in \mathcal{D}} \sum_{r \in J} B_l(D, r) \rho^\pi(D, r) = \sum_{D \in \mathcal{D}} B_l(D), \quad (23)$$

where the second equality holds by $\rho^\pi(D, r) = 1$ if $r = \max_D \pi$ and $\rho^\pi(D, r) = 0$ otherwise.

Finally, by using these equalities, for each $l \in \{1, \dots, L\}$, we obtain the following equations:

$$\begin{aligned} \sum_{D \in \mathcal{D}} \sum_{r \in J} B_l(D, r) q(D, r) &= \sum_{D \in \mathcal{D}} \sum_{r \in D} B_l(D, r) q(D, r) \quad (\text{by property (ii)}) \\ &= \sum_{D \in \mathcal{D}} \sum_{r \in D} B_l(D) q(D, r) \quad (\text{by Step 1}) \\ &= \sum_{D \in \mathcal{D}} B_l(D) \sum_{r \in D} q(D, r) \\ &= \sum_{D \in \mathcal{D}} B_l(D) = b_l. \quad (\text{by property (i) and (23)}) \end{aligned}$$

This establishes that $\text{aff } \mathcal{P}_r = \{q \in \mathbf{R}^{\mathcal{D} \times J} | \text{(i) and (ii)}\}$.

The equalities in (i) and (ii) are independent. The dimension of $\{q \in \mathbf{R}^{\mathcal{D} \times J} | \text{(ii)}\}$ is $\sum_{D \in \mathcal{D}} |D|$. The number of equalities of (i) is $|\mathcal{D}|$. Hence, the dimension of $\mathcal{P}_r$ is $(\sum_{D \in \mathcal{D}} |D|) - |\mathcal{D}| = \sum_{D \in \mathcal{D}} (|D| - 1)$. $\blacksquare$

Proposition 7 follows from Lemma 10 and Lemma 11.

J Further Empirical Analyses for Section 5

J.1 Approximation Errors for $|J| = 10$ and $K = 3$

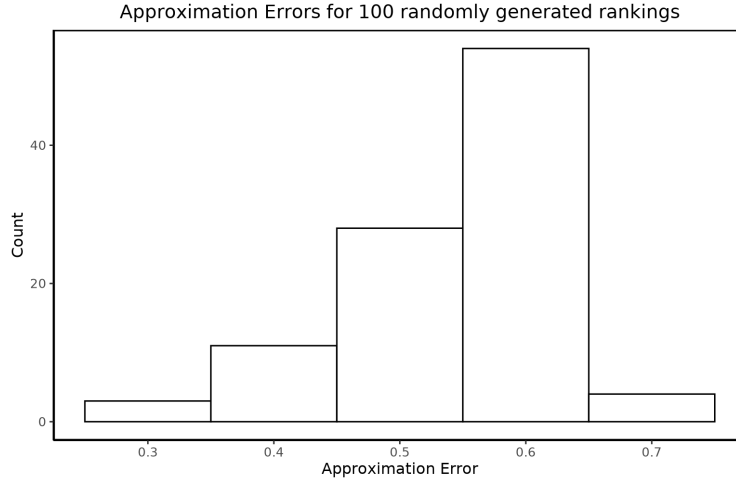


Figure A.3: Approximation Errors of 100 randomly generated rankings with 10 alternatives ($|J| = 10$) and 3 characteristics ($K = 3$).

In Section 5, we examined approximation errors in the setting with four alternatives and two characteristics. To assess how these errors behave with a larger number of alternatives, we now consider 10 alternatives and three characteristics. The 10 alternatives are selected from

the first ten rows of Table 2 in Farias, Jagabathula, and Shah (2009), which correspond to the ten most expensive alternatives. For computational tractability, the analysis is restricted to binary and ternary choice sets. Figure A.3 contains the histogram of approximation errors of linear mixed-logit models without fixed effects for 100 rankings drawn uniformly at random from all possible rankings. The results indicate that all randomly generated rankings exhibit non-negligible approximation errors, ranging from 0.3 to 0.7 in the normalized Euclidean metric.

J.2 Details of the DVD Dataset

We accessed the DVD dataset from Table 2 of Farias, Jagabathula, and Shah (2009), which we replicate as Table A.1. For our analysis in the main text, we selected the products from the first four rows. For our analysis in Section J.1, we selected the products from the first ten rows. For the analysis with four products and two characteristics, we used *Average Price per Disc* and *Total Helpful Votes*. For the analysis with ten products and three characteristics, we additionally included *Price*.

Table A.1: DVD dataset as in Table 2 of Farias, Jagabathula, and Shah (2009)

Product ID	Price	Avg. Price per Disc	Total Helpful Votes
1	115.49	5.7747	462
2	92.03	7.6694	20
3	91.67	13.0955	496
4	79.35	13.2256	8424
5	77.94	6.4949	6924
6	70.12	14.0242	98
7	64.97	16.2423	1116
8	49.95	12.4880	763
9	48.97	6.9962	652
10	46.12	7.6863	227
11	45.53	6.5037	122
12	45.45	11.3637	32541
13	45.41	11.3523	69
14	44.92	11.2292	1113
15	42.94	10.7349	320

K Additional Empirical Application: Fishing-site Choice

In this section, we measure approximation errors with and without fixed effects by using a dataset on fishing-site choices from Thomson and Crooke (1991).⁴⁷ The dataset has been used by Herriges and Kling (1999) and Cameron and Trivedi (2005) (p.464).

In the dataset, 1182 individuals choose among 4 fishing modes, namely, $J = \{\text{beach, boat, charter, pier}\}$, which denote fishing from the beach, a private boat, a charter boat or a pier, respectively. Each alternative $j \in J$ is described by a vector of two characteristics (p_j, q_j) . The first characteristic p_j is the fishing mode j 's price. The other characteristic q_j is the *catch rate*, defined as a per-hour-fished basis for major species by fishing mode j .⁴⁸ Our empirical analysis concentrates primarily on mixed-logit models. We assess the approximation errors using a linear mixed-logit model and a quadratic mixed-logit model, as in Section 5.

K.1 Approximation Errors without Fixed Effects

Table A.2 shows the approximation errors of the linear or quadratic mixed-logit models. We calculated the errors by the greedy algorithm and the EM algorithm. In both algorithms, the approximation errors under quadratic mixed-logit models are zero, as shown in columns (3) and (4) in the table. This finding is consistent with the theorem. For the linear mixed-logit models, the approximation errors for representable rankings π shown in the bottom row of the table are always zero, as the theorem predicts.

On the other hand, the approximation errors for unrepresentable rankings π are almost always larger than 0.4, which means that even the best possible linear mixed-logit model remains far from the corresponding choice probabilities ρ^π . Some errors are much larger. For example, the approximation errors of the two rankings $\pi(1) > \pi(2) > \pi(3) > \pi(4)$ and $\pi(1) > \pi(2) > \pi(4) > \pi(3)$ by the linear models are greater than 0.67. Notice that these two rankings are the only rankings in which alternative 1 is the best and alternative 2 is the second-best. This fact suggests that the substitution from alternative 1 to 2 would be difficult to capture. The next subsection further explores this issue.

K.1.1 Maximal Substitution

Table A.3 shows the values of maximal substitution between the two alternatives j and l . Some numbers in the table are close to one, which implies that the linear mixed-logit models are rich enough to capture flexible substitution from j to l . Some other numbers are smaller.

In particular, the maximal substitution between alternative 1 (i.e., beach) and 2 (i.e., private

⁴⁷The dataset is taken directly from the R package ‘mlogit’ by Croissant (2020).

⁴⁸In the original study, the values of p_j and q_j depend on each individual. For our analysis, we aggregate them by taking the average over individuals.

Table A.2: Approximation errors to ρ^π

Ranking π	Linear mixed-logit		Quadratic mixed-logit	
	Greedy (1)	EM (2)	Greedy (3)	EM (4)
Linearly Unrepresentable Rankings				
$\pi(1) > \pi(2) > \pi(3) > \pi(4)$	0.723	0.753	0.000	0.000
$\pi(1) > \pi(2) > \pi(4) > \pi(3)$	0.670	0.700	0.000	0.000
$\pi(1) > \pi(3) > \pi(2) > \pi(4)$	0.425	0.381	0.000	0.000
$\pi(1) > \pi(4) > \pi(2) > \pi(3)$	0.418	0.547	0.000	0.000
$\pi(2) > \pi(1) > \pi(3) > \pi(4)$	0.458	0.488	0.000	0.000
$\pi(2) > \pi(1) > \pi(4) > \pi(3)$	0.391	0.408	0.000	0.000
$\pi(3) > \pi(2) > \pi(4) > \pi(1)$	0.302	0.318	0.000	0.000
$\pi(3) > \pi(4) > \pi(1) > \pi(2)$	0.401	0.425	0.000	0.000
$\pi(3) > \pi(4) > \pi(2) > \pi(1)$	0.494	0.531	0.000	0.000
$\pi(4) > \pi(2) > \pi(3) > \pi(1)$	0.375	0.381	0.000	0.000
$\pi(4) > \pi(3) > \pi(1) > \pi(2)$	0.521	0.514	0.000	0.000
$\pi(4) > \pi(3) > \pi(2) > \pi(1)$	0.604	0.614	0.000	0.000
Linearly Representable Rankings	0.000	0.000	0.000	0.000

Note: The numbers in the table show the approximation errors for each ρ^π , where each ranking π is defined in the leftmost column. Alternative numbers 1, 2, 3, 4 denote beach, boat, charter, and pier, respectively. For each ranking, columns (1) and (2) show the approximation errors of the linear mixed-logit models computed by the greedy algorithm and the EM algorithm, respectively. Columns (3) and (4) show the approximation errors of the quadratic mixed-logit models calculated by each algorithm. All numbers are rounded to three decimal places. For the greedy algorithm we set the number of iterations to 1000. For the EM algorithm we set the number of random initial points to 10. The greedy algorithm sometimes produces larger approximation errors than the EM algorithm, which is possible with finitely many steps.

boat) as well as the substitution between alternative 3 (i.e., charter) and 4 (i.e., pier) are at most 0.3. In fact, the maximal substitution from 1 to 2 is 0.12. This means that no matter how the parameters of a linear mixed-logit model are chosen, the maximal substitution from alternative 1 to alternative 2 is very limited. This finding aligns with the result presented in Table A.2, where we observe substantial approximation errors for the two specific rankings: $\pi(1) > \pi(2) > \pi(3) > \pi(4)$ and $\pi(1) > \pi(2) > \pi(4) > \pi(3)$. In this way, identifying preferences that are hard to approximate with precision helps researchers in evaluating whether their models successfully capture relevant economic behaviors such as substitution patterns.

Table A.3: Maximal substitution of the linear mixed-logit models

j (drop) \ l (choose)	1	2	3	4
1	-	0.120	0.998	0.998
2	0.317	-	1.000	0.997
3	0.998	1.000	-	0.286
4	0.994	0.998	0.137	-

Note: The numbers in the table show the value of (8) for each $j, l \in \{1, 2, 3, 4\}$ s.t. $j \neq l$. Alternative numbers 1, 2, 3, 4 denote beach, boat, charter, pier, respectively. All numbers are rounded to three decimal places.

K.2 Approximation Errors with Fixed Effects

Table A.4 shows the approximation error to $\frac{1}{2}\rho^\pi + \frac{1}{2}\rho^{\pi^-}$ for each unrepresentable ranking. In both algorithms, the approximation errors to $\frac{1}{2}\rho^\pi + \frac{1}{2}\rho^{\pi^-}$ are always around 0.2 if π is not representable. This means that even the best possible linear mixed-logit model remains bounded away from $\frac{1}{2}\rho^\pi + \frac{1}{2}\rho^{\pi^-}$ in the normalized Euclidean metric. On the other hand, the approximation errors to $\frac{1}{2}\rho^\pi + \frac{1}{2}\rho^{\pi^-}$ are almost zero if π is representable, as the theorem predicts. The approximation errors under quadratic mixed-logit models are also almost zero, as the theorem again predicts.

Table A.4: Approximation errors to random utility models $\frac{1}{2}\rho^\pi + \frac{1}{2}\rho^{\pi^-}$

Ranking π	Linear mixed-logit		Quadratic mixed-logit	
	Greedy (1)	EM (2)	Greedy (3)	EM (4)
Linearly unrepresentable rankings				
$\pi(1) > \pi(2) > \pi(3) > \pi(4)$	0.229	0.255	0.000	0.000
$\pi(1) > \pi(2) > \pi(4) > \pi(3)$	0.229	0.250	0.000	0.000
$\pi(1) > \pi(3) > \pi(2) > \pi(4)$	0.163	0.217	0.000	0.000
$\pi(1) > \pi(4) > \pi(2) > \pi(3)$	0.163	0.173	0.000	0.000
$\pi(2) > \pi(1) > \pi(3) > \pi(4)$	0.192	0.238	0.000	0.000
$\pi(2) > \pi(1) > \pi(4) > \pi(3)$	0.192	0.198	0.000	0.000
Linearly representable rankings	0.000	0.000	0.000	0.000

Note: The numbers in the table show the approximation errors to $\frac{1}{2}\rho^\pi + \frac{1}{2}\rho^{\pi^-}$, where π is defined in the leftmost column. All numbers are rounded to three decimal places. For the greedy algorithm we set the number of iterations to 1000. For the EM algorithm we set the number of random initial points to 10. We optimize over fixed effects from -10 to 10 with a step size of 1.

K.3 In-sample and Out-of-sample Fit

In this section, we evaluate the in-sample and out-of-sample fit of our model. We show that our method performs better or equally well compared to standard methods, not only in terms of in-sample fit but also in terms of out-of-sample fit. We use the same fishing choice dataset used above and predict choice probabilities using aggregated characteristics.

We estimate a random-coefficient logit model with arbitrary mixing distributions. In the dataset, we have four alternatives and we consider only one choice set $\mathcal{D} = \{J\}$. Thus by Proposition 7 and Proposition 6, it suffices to mix four logit models without fixed effects to represent any random utility model. We refer to a 4-mixture mixed-logit model as *our method* hereafter. We also estimate several standard models for comparison. They include 1) a multinomial logit model, 2) nested logit model with two nests (charter and the rest), 3) a nested logit model with two nests (boat and the rest), 4) a random coefficient logit model with a log-normal mixing distribution for each variable, 5) a multinomial logit model with alternative fixed effects,

6) and a random coefficient logit model with log-normal mixing distributions and alternative fixed effects. We detail the definition of each specification in Section K.4.

To evaluate in-sample and out-of-sample fits, we adopt the following strategy. We randomly divide individuals in the sample into a training sample and a test sample of equal sizes. Separately for the training and testing samples, we average individual choices and characteristics to obtain aggregate data on choice probabilities and characteristics. We then estimate the models using the training sample. The models are estimated by maximizing the log-likelihoods. That is, for each model, we solve the problem $\max_{\theta \in \Theta} \sum_{j=1}^{|J|} \hat{\rho}_j \log \rho(j|\theta)$, where j indexes fishing modes, θ is the parameter vector of the model, Θ denotes the set of possible parameter vectors, $\hat{\rho}_j$ is the observed market share for fishing mode j in the training data, and $\rho(j|\theta)$ is the model-predicted choice probability for fishing mode j with characteristic vector x_j . See Section K.4 for a likelihood expression for each model. For the standard models, we maximize the likelihoods with the nonlinear optimization package in R (Ghalanos and Theussl, 2015; Ye, 1987). For our model, we use the EM algorithm.⁴⁹

To evaluate the in-sample fit performance, we compute the predicted choice probabilities in the training sample $\hat{\rho}_{train} \in \mathbf{R}^{|J|}$ and compare them with the observed choice probabilities in the training sample $\rho_{train} \in \mathbf{R}^{|J|}$.⁵⁰ For this comparison, we calculate the l_2 distance between the predicted choice probabilities and the aggregated observed choice probabilities $\|\hat{\rho}_{train} - \rho_{train}\|_2$. Similarly, to evaluate the out-of-sample performance, we compute the predicted choice probabilities using the testing sample $\hat{\rho}_{test} \in \mathbf{R}^{|J|}$ and compare it with the aggregated observed choice probabilities in the testing sample $\rho_{test} \in \mathbf{R}^{|J|}$. We use the l_2 metric $\|\hat{\rho}_{test} - \rho_{test}\|_2$ for this comparison as well.

We repeat this exercise with 50 random splits. The results for in-sample fits are reported in Table A.5. The results for out-of-sample fits are in Table A.6.

As expected, the in-sample fit of our model is perfect. Several standard models, especially those without fixed effects, exhibit imperfect in-sample fit. For example, the random coefficient logit model with the log-normal distributions has the l_2 prediction error 0.038.

Table A.6 shows that the out-of-sample prediction error of our model is positive but small. Standard models without alternative fixed effects have out-of-sample prediction errors substantially larger than our model. The two alternative models with fixed effects have out-of-sample prediction errors comparable to ours. This result suggests that even without using fixed effects, our model performs as well as or better than standard models in this simulation, not only in terms of in-sample fit but also in terms of out-of-sample fit.

⁴⁹We prefer the EM algorithm over the greedy algorithm here because the EM algorithm is faster.

⁵⁰We only consider the single choice set case in this simulation. So the choice probability vector has length $|J|$.

Table A.5: In-Sample Fit

Model	Choice probabilities				Prediction error
	Beach (1)	Boat (2)	Charter (3)	Pier (4)	(5)
Our method	0.114 (0.009)	0.353 (0.015)	0.383 (0.017)	0.151 (0.010)	0.000 (0.000)
Multinomial logit	0.141 (0.007)	0.355 (0.015)	0.378 (0.017)	0.126 (0.007)	0.038 (0.009)
Nested logit (charter and others)	0.114 (0.010)	0.353 (0.015)	0.383 (0.017)	0.150 (0.011)	0.001 (0.002)
Nested logit (boat and others)	0.141 (0.007)	0.355 (0.015)	0.378 (0.017)	0.126 (0.007)	0.038 (0.009)
Mixed-logit with log normal distribution	0.142 (0.008)	0.354 (0.015)	0.378 (0.017)	0.126 (0.007)	0.038 (0.009)
Multinomial logit with fixed effects	0.113 (0.009)	0.353 (0.015)	0.383 (0.017)	0.151 (0.010)	0.000 (0.000)
Mixed-logit with log normal distribution and fixed effects	0.114 (0.009)	0.353 (0.015)	0.383 (0.017)	0.151 (0.010)	0.000 (0.000)

Note: Table A.5 summarizes the in-sample fit of different models. The row “our method” presents choice probabilities predicted by the four-mixture mixed-logit model and the prediction error. The remaining rows present in-sample predicted choice probabilities and prediction errors obtained by standard models. In parentheses are standard deviations obtained by repeating the same analyses 50 times.

K.4 Definitions of Other Models

In each of the standard models used in our empirical section, the choice probability $\rho(J, j) \equiv \rho_j$ of alternative j from J is specified as follows:

- Multinomial logit: $\rho_j = \frac{\exp(\beta \cdot x_j)}{\sum_{j' \in J} \exp(\beta \cdot x_{j'})}$
- Nested logit (charter and others): the choice probability of alternative j that belongs to nest J_g is specified as

$$\rho_j = \frac{\exp(\beta \cdot x_j / \lambda)}{\sum_{j' \in J_g} \exp(\beta \cdot x_{j'} / \lambda)} \times \frac{\left[\sum_{j' \in J_g} \exp(\beta \cdot x_{j'} / \lambda) \right]^\lambda}{\sum_{g' \in G} \left[\sum_{j' \in J_{g'}} \exp(\beta \cdot x_{j'} / \lambda) \right]^\lambda}.$$

The nest is defined by the partition $G = \{\{\text{charter}\}, \{\text{beach}, \text{boat}, \text{pier}\}\}$.

- Nested logit (boat and others): the nested logit model specified above, with the nest defined by $G = \{\{\text{boat}\}, \{\text{beach}, \text{charter}, \text{pier}\}\}$.
- Mixed-logit: $\rho_j = \int \frac{\exp(\beta \cdot x_j)}{\sum_{j' \in J} \exp(\beta \cdot x_{j'})} f(\beta) d\beta$ where f is the density of the distribution of random coefficients. We use independent log-normal distributions for each coefficient. To evaluate the integral, we randomly draw 100 realizations from the random coefficient distribution.

Table A.6: Out-of-Sample Fit

Model	Choice probabilities				Prediction error
	Beach (1)	Boat (2)	Charter (3)	Pier (4)	(5)
Our method	0.114 (0.009)	0.353 (0.015)	0.383 (0.017)	0.151 (0.010)	0.049 (0.017)
Multinomial logit	0.143 (0.011)	0.353 (0.017)	0.377 (0.022)	0.127 (0.011)	0.058 (0.019)
Nested logit (charter and others)	0.115 (0.012)	0.350 (0.022)	0.383 (0.018)	0.152 (0.014)	0.050 (0.020)
Nested logit (boat and others)	0.143 (0.011)	0.353 (0.017)	0.377 (0.022)	0.127 (0.011)	0.058 (0.019)
Mixed-logit with log normal distribution	0.143 (0.010)	0.352 (0.016)	0.377 (0.021)	0.127 (0.010)	0.058 (0.018)
Multinomial logit with fixed effects	0.116 (0.014)	0.350 (0.019)	0.380 (0.022)	0.154 (0.019)	0.048 (0.022)
Mixed-logit with log normal distribution and fixed effects	0.113 (0.008)	0.353 (0.015)	0.383 (0.018)	0.151 (0.010)	0.048 (0.017)

Note: Table A.6 summarizes the out-of-sample fit of different models. The row “our method” presents choice probabilities predicted by the four-mixture mixed-logit model and the prediction error (7). The remaining rows present out-of-sample predicted choice probabilities and prediction errors obtained by standard models. In parentheses are standard deviations obtained by repeating the same analyses 50 times.

- Multinomial logit with fixed effects: the above multinomial logit model with x including dummies for each alternative (except for beach).
- Mixed-logit with fixed effects: the random-coefficient logit model with log normal distributions. We also include fixed effects for each alternative (except for beach). To evaluate the integral, we randomly draw 100 realizations from the random coefficient distribution.

L Notation

Notation	Meaning
J	the set of alternatives
D	a generic choice set
$\mathcal{D}$	the set of generic choice sets
$\mathcal{M}$	the set of standard probability measures
K	the dimension of explanatory variables
$\mathcal{P}$	the set of stochastic choice functions
$\mathcal{P}_r$	the set of random utility models
ρ	a generic stochastic choice function
Π	the set of rankings
$\text{co } C$	the convex hull of a set C
$\text{rint } C$	the relative interior of C
$\text{cl } C$	the closure of C
$\mathcal{P}_{ra}(\eta \mu)$	the set of random-coefficient ARUMs with fixed effects η and a probability measure μ
$\mathcal{P}_a(\eta \mu)$	the set of ARUMs with fixed effects η and a probability measure μ
$\mathcal{P}_{ml}(\eta)$	the set of mixed-logit models with fixed effects η