Approximating Choice Data by Discrete Choice Models

Haoge Chang, Yusuke Narita, and Kota Saito

September 11, 2022

Abstract

We obtain a necessary and sufficient condition under which random-coefficient discrete choice models such as the mixed logit models can approximate the choice behavior generated by nonparametric random utility models. The condition turns out to be very simple and tractable. For the case under which the condition is *not* satisfied and, hence, any random-coefficient discrete choice model cannot approximate some choice data generated by a random utility model, we provide algorithms to measure the approximation errors. After applying our theoretical results and the algorithms to real data, we find that the approximation errors can be large in practice. [1]

Keywords: Discrete choice, stochastic choice, mixed logit, random coefficients, finite mixture.

[1]Chang: Yale University, email: haoge.chang@yale.edu. Narita: Yale University, email: yusuke.narita@yale.edu. Saito: Caltech, email: saito@caltech.edu. A part of this paper was first presented at the University of Tokyo on July 29, 2017. This paper subsumes parts of "Axiomatizations of the Mixed Logit Model" by Saito (The paper is available at `http://www.hss.caltech.edu/content/axiomatizations-mixed-logit-model`). We would like to thank Hiroki Saruya and Haruki Kono for their help as RAs. We appreciate the valuable discussions we had with Brendan Beare, Victor Aguirregabiria, Doignon Jean-Paul , John Rust, Giovanni Compiani, Steven Berry, Yi Xin, Alfred Galichon, Phil Haile, Jay Lu, Ariel Pakes, Whitney K. Newey, and Matt Shum. Jay Lu also read an earlier version of the manuscript and offered helpful comments. We appreciate the insightful comments made by Victor Aguirregabiria at the ASSA meetings in January 2022. Saito acknowledges the financial support of the NSF through grants SES-1919263 and SES-1558757.

1 Introduction

Random-coefficient discrete choice models are workhorse models in many empirical economics applications. These models have been used to approximate preferences and capture substitution patterns. However, the exact degree of flexibility and the limitations of the random-coefficient models have not yet been fully understood. In this paper, we obtain a necessary and sufficient condition under which random coefficient models can approximate the choice behavior generated by any random utility model. The condition turns out to be very simple and tractable.

We consider the following class of models. Let $X \subset \mathbf{R}^k$ be the set of all alternatives, where k is the number of explanatory variables. In a *utility-shock model*, the choice probability of an alternative x in a choice set $D \subset X$ is given by $\rho(D,x) = \mu(\{\varepsilon | u(x) + \varepsilon(x) > u(y) + \varepsilon(y) \forall y \in D \setminus \{x\}\})$, where ε is a random utility shock that follows the distribution μ.[2] The utility-shock model is general and includes the probit, logit, and nested-logit models as special cases. The *random-coefficient* version of the utility-shock model is defined as follows. The choice probability is given by

$$\rho(D,x) = \int \mu(\{\varepsilon | u(x) + \eta(x) + \varepsilon(x) > u(y) + \eta(y) + \varepsilon(y) \forall y \in D \setminus \{x\}\}) dm(u), \quad (1)$$

where m is a distribution over u and η is a vector of *fixed effects*. In the standard interpretation, m captures the heterogeneity of preferences among consumers. The *fixed effects* η capture unobserved characteristics of alternatives.[3] When μ is a multivariate iid extreme-value type-I distribution, then ρ reduces to a *mixed logit model*, which is one of the most widely used random-coefficient models (see Train (2009)). Usually, researchers make a parametric assumption on $u(\cdot)$ in (1), such as $u(x)$ is a polynomial. In many papers, researchers assume u is linear (i.e., $u(x) = \beta \cdot x$). If u is a polynomial of degree d m-a.s., then the model is called the *degree-d random-coefficient utility-shock model*.[4]

Given the popularity of the model, it is important to understand its exact extent of flexibility and limitations. For this purpose, we obtain a necessary and sufficient

[2]The model is often called additive random utility model. In this paper, we assume that μ has a density function f and the support of μ (i.e., $\{\varepsilon | f(\varepsilon) > 0\}$) is convex and open and includes zero.

[3]Notice that the roles of u and η are different. The distribution m is only on u but not on η.

[4]In empirical analysis, researchers usually make a restrictive assumption on the distribution m. Such typical implementations of mixed logit models impose strong restrictions on the data-generating process (Ackerberg and Rysman, 2005).

condition under which the degree-d random-coefficient utility-shock models are rich enough to approximate any choice probabilities generated by random utility models across choice sets. For the data-generating process and the approximation target, we choose nonparametric random utility models, which are defined as probability measures over rankings on alternatives. This is because the class of the models is one of the most flexible classes in the literature. We study approximation *across choice sets* because many questions of interest (such as substitution patterns) are possible to answer only if we analyze behaviors across choice sets.

When the condition is *not* satisfied, it is important to ask how large the approximation errors can be. To answer this question, we provide algorithms that calculate the approximation errors. Moreover, we apply the algorithms to a real dataset. In the following paragraphs, we will explain the necessary and sufficient condition. Then, we will describe its empirical application and the algorithms to calculate the minimal approximation errors.

The necessary and sufficient condition is the affine-independence of the set $\{p_d(x)|x \in X\}$, where $p_d(x)$ is the vector consisting of monomials of at most degree d of characteristics x.[5] One surprising fact about the condition is that it does not depend on the distribution μ over the utility shock ε. This implies that if the condition is violated, there exists some choice probabilities generated by a random utility model that cannot be approximated by any degree-d random-coefficient utility-shock model, no matter which distribution μ of ε we use (and no matter which fixed effects we use).

Although the affine-independence condition is easy to test, the condition is generically equivalent to a further simpler condition: $|X| \leq \binom{d+k}{k}$, where $|X|$ is the number of alternatives, k is the number of characteristics observed for each alternative, and d is the degree of polynomial utility functions u. The number $\binom{d+k}{k}$ is the number of ways of choosing k elements out of $d+k$ elements, which is increasing both in d and k. The condition requires that the degree d of polynomial utility functions and the number k of coefficients be large enough to satisfy the inequality. The condition is easy to check. For example, when $d = 1$, as most papers assume, the condition reduces to $|X| \leq k + 1$.

Remember that the affine-independence condition is for the approximation across

[5]For example, if $k = 2$ then $p_2(x) = (x(1), x(2), x(1)^2, x(1)x(2), x(2)^2)$ and $p_1(x) = (x(1), x(2))$, where $x = (x(1), x(2))$. A set Y is *affinely independent* if for any $y \in Y$, there exists no real number $\{\mu_x\}_{x \in X}$ such that $y = \sum_{x \in Y \setminus \{y\}} \mu_x x$ and $\sum_{x \in Y \setminus \{y\}} \mu_x = 1$.

choice sets. In some cases, researchers may be interested only in fitting a model to the observed choice probabilities (i.e., market shares) on the fixed choice set X. In that case, the necessary and sufficient condition reduces to be the convex-independence of the set $\{p_d(x)|x \in X\}$ (i.e., $p_d(x) \notin \text{co.}\{p_d(y)|y \in X \setminus x\}$ for any $x \in X$), which is weaker than the affine-independence condition.[6]

In many empirical papers, researchers use mixed logit models (i.e., models in which μ in (1) is a multivariate iid extreme-value type-I distribution); moreover the utility function u in (1) is linear (i.e., $d = 1$). In all such papers, we find that the convex-independence condition is satisfied. On the other hand, the condition that $|X| \leq k + 1$ is not satisfied in many papers. This means that the linear mixed-logit models are rich enough to approximate observed choice probabilities from a single choice set X; however, the models may not be rich enough to approximate the true substitution pattern across subsets of X, no matter how one chooses the parameters and fixed effects.

In the case in which the affine-independence condition is not satisfied, we propose two algorithms to measure the approximation errors. One algorithm is a variant of the greedy algorithm proposed in Barron et al. (2008). The algorithm has the useful property that it is guaranteed to find the best approximation using the random-coefficient model in the l_2 metric. The other algorithm is the EM (Expectation-Maximization) algorithm drawn from Dempster et al. (1977). We provide some theoretical results that facilitate the use of the EM algorithm by providing an upper bound for the number of mixtures to use.

We apply our theorem and the two algorithms to a dataset of fishing-site choices from Thomson and Crooke (1991). In the data set, $k = 2$ and $|X| = 4$; the affine-independence condition with $d = 1$ is thus violated. We measure the approximation errors by estimating the best possible linear mixed logit model using the greedy algorithm and the EM algorithm. Regardless of the method used, we find that the approximation errors are large and often larger than 10 percent on average. The results suggest that using the linear mixed logit model may make it difficult to capture a reasonable substitution pattern.

The rest of the paper is organized as follows. In the next subsection, we discuss the related literature. In section 2, we introduce the models. In section 3, we provide the main results. In section 4, we provide theoretical results for measuring

[6]Unlike the affine-independence condition, the convex-independence condition does not have such a simple generic condition. See footnote 17.

approximation errors. In section 5, we provide an empirical illustration.

Related Literature

The work most closely related to our paper is McFadden and Train (2000), whose result shows that any given (nonparametric) continuous random utility model can be approximated by a mixed logit model.[7] Nevertheless, there are three important differences to note. In particular, our result holds for a much more general class of random-coefficient utility-shock models, including but not confined to the mixed logit models. Second, our result is not only sufficient but also necessary. This is crucial given our purpose of clarifying the exact extent of flexibility and limitations of the random-coefficient utility-shock models. Moreover, through our condition, our results provide a tight bound on how many parameters we need for an arbitrarily good approximation. Third, the setup of McFadden and Train (2000) and our setup differ in that McFadden and Train (2000) focus on the case where X is continuous, while we assume that X is finite. Hence, neither result implies the other. A recent paper by Lu and Saito (2021a) also studies the extent to which the approximation of continuous random utility is possible by using mixed-logit models.

Another paper related to ours is Norets and Takahashi (2013). They also study utility-shock models but without considering random coefficients. They study whether utility-shock models can represent any stochastic choice on a fixed choice set. The differences between our paper and their paper come from the fact that they do not study choices across subsets nor do they allow random coefficients. Athey and Imbens (2007) also investigate how a rich specification of the unobserved components (i.e., the fixed effects) is needed to represent any stochastic choice. Their setup is also different from ours in that they focus on logit models.

Our analysis shares some of its spirit with the growing literature that identifies and estimates flexible discrete choice models under minimal assumptions. See, for example, Berry and Haile (2014), Compiani (2022), and Tebaldi et al. (2019). Our paper is also related in motivation to recent studies that apply machine learning to specify flexible utility functions in discrete choice models to improve approximation performance (Bajari et al., 2015; Ruiz et al., 2020; Gillen et al., 2019).

In the decision theory literature, we know of no research that directly relates to our papers. However, logit models and random utility models have been analyzed

[7] Our result is consistent with their result: heuristically speaking, the result by McFadden and Train (2000) corresponds to the case when $d = \infty$, which satisfies our condition.

for a long time ever since Luce (1959) and Block and Marschak (1960), although the utility-shock model has not been studied directly. In the following, we will mention recent work on logit models and random utility models.

Recent papers in decision theory have considered generalizations of logit models, which include mixed logit models (Gul, Natenzon, and Pesendorfer (2014), Saito (2018)) and nested-logit models (Kovach and Tserenjigmid (2020)). Echenique and Saito (2019), Cerreia-Vioglio, Maccheroni, Marinacci and Rustichini (2022; 2018), and Horan (2018) consider the Luce axiom without positivity. Fudenberg and Strzalecki (2015) consider dynamic extensions of logit models. Chambers, Cuhadaroglu and Masatlioglu (2020) consider a variation of the logit models in a social setting, while Chambers, Masatlioglu and Raymond (2021a) add an additional parameter reflecting salience or other economic frictions. Some papers, such as Apesteguia and Ballester (2018) and Frick, Iijima, and Strzalecki (2019), point out the differences in choice behavior between random utility models and logit models.

Gul and Pesendorfer (2006), which axiomatizes the random expected utility theory, has inspired many works that study the random utility model and its generalization. These include Ahn and Sarver (2013), Apesteguia, Ballester and Lu (2017), Lin (2019), and Chambers, and Masatlioglu and Turansick (2021b), as well as Lu (2016, 2021) on extensions to choice under uncertainty, and Lu and Saito (2018), Duraj (2018), Frick, Iijima and Strzalecki (2019), and Lu and Saito (2021b) on dynamic extensions.

2 Model

2.1 Setup

Set of alternatives: The set of all alternatives is denoted by X. X is assumed to be finite. An alternative x is described by a real vector of explanatory variables of the alternative. For example, if an alternative is a consumption good, the alternative is described by its price and its various other characteristics. Hence, we let X be a finite subset of $\mathbf{R}^k$, where k is the number of the explanatory variables. For each $x \in X$ and $l \in \{1, \ldots, k\}$, we write $x(l)$ to denote the l-th element of x.

Choice sets: Let $\mathcal{D} \subset 2^X \setminus \{\emptyset\}$ be the set of choice sets. Notice that $\mathcal{D}$ can be a proper subset of $2^X \setminus \emptyset$. Unless otherwise noted, throughout the paper we assume that $\{x, y\} \in \mathcal{D}$ and $\{x, y, z\} \in \mathcal{D}$ for any $x, y, z \in X$. In some parts of the paper,

however, we drop this assumption and assume that $\mathcal{D} = \{X\}$ when we consider the case in which an econometrician's purpose is fitting a model to the observed choice probabilities from the single choice set.

The set $\mathcal{D}$ may contain both observed choice sets as well as hypothetical choice sets the econometrician is interested in. For example, even when the econometrician 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 the consumers' substitution pattern.

Stochastic choice function: A function $\rho : \mathcal{D} \times X \to [0, 1]$ is called a *stochastic choice function* if $\sum_{x \in D} \rho(D, x) = 1$ and $\rho(D, x) = 0$ for any $x \notin D$. The set of stochastic choice functions is denoted by $\mathcal{P}$. For each $(D, x) \in \mathcal{D} \times X$, the number $\rho(D, x)$ is the probability that an alternative x is chosen from a choice set D. In a context of empirical industrial organization, for example, $\rho(D, x)$ can be interpreted as a market share of product x in a market in which the set of available products is D. We interpret the stochastic choice function ρ as aggregate choice probabilities across individuals.

Rankings: Let Π be the set of bijections between X and $\{1, \ldots, |X|\}$, where $|X|$ is the number of elements of X. For any element $\pi \in \Pi$, if $\pi(x) = i$, then we interpret x to be the $|X| + 1 - i$-th best element of X with respect to π. If $\pi(x) > \pi(y)$, then x is better than y with respect to π. An element π of Π is called a *strict preference ranking* (or simply, a *ranking*) over X. For all $(D, x) \in \mathcal{D} \times X$ such that $x \in D$, if $\pi(x) > \pi(y)$ for all $y \in D \setminus \{x\}$, then we often write $\pi(x) \geq \pi(D)$. There are $|X|!$ elements in Π.

2.2 Models

We denote the set of probability measures over Π by $\Delta(\Pi)$. Since Π is finite, $\Delta(\Pi) = \left\{(\nu_1, \ldots, \nu_{|\Pi|}) \in \mathbf{R}_+^{|\Pi|} \middle| \sum_{i=1}^{|\Pi|} \nu_i = 1\right\}$, where $\mathbf{R}_+$ is the set of nonnegative real numbers.

We now introduce the definition of random utility models:

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 $(D, x) \in \mathcal{D} \times X$, if $x \in D$, then

$$\rho(D, x) = \nu(\pi \in \Pi | \pi(x) \geq \pi(D)).$$

The set of random utility models is denoted by $\mathcal{P}_r$.[8]

Notice that when $\mathcal{D} = \{X\}$, the restriction of random utility is vacuous: any stochastic choice function is a random utility model (i.e., $\mathcal{P}_r = \mathcal{P}$).[9]

To introduce the utility-shock models, we introduce some notations. Given a positive integer d and $x \in X$, the vector $p_d(x)$ consists of *monomials* of at most degree d (i.e., higher order terms such as $x(l)^n$ where $n \leq d$, and interaction terms such as $\prod_{l=1}^{k} x(l)^{n_l}$, where $\sum_{l=1}^{k} n_l \leq d$).[10] Notice that $p_d(x) = x$ when $d = 1$. We consider the models with fixed effects given their popularity in empirical applications. They are used frequently to capture the average preference for unobserved characteristics of alternatives. (See Berry et al. (1995) for an example.)

Definition 2. Let d be a positive integer and $\eta \in \mathbf{R}^{|X|}$ be a real vector of fixed effects. A stochastic choice function ρ is called a *degree-d utility-shock model with fixed effects* η if there exists a probability measure μ such that, for all $(D, x) \in \mathcal{D} \times X$, if $x \in D$, then

$$\rho(D, x) = \mu(\{\varepsilon | \beta \cdot p_d(x) + \eta(x) + \varepsilon(x) > \beta \cdot p_d(y) + \eta(y) + \varepsilon(y) \text{ for all } y \in D \setminus \{x\}\}),$$

where (i) $\beta \cdot p_d(x)$ is a polynomial of x of at most degree d, (ii) μ has a density function f and the support of μ (i.e., $\{\varepsilon | f(\varepsilon) > 0\}$) is convex and open and includes zero.[11] When $d = 1$, then we say the function is *linear* instead of *degree*-1.

The set of degree-d utility-shock models with fixed effects η and distribution μ is denoted by $\mathcal{P}_s(d, \eta | \mu)$.[12]

Let $\mathcal{M}$ be the set of probability distributions over ε that satisfies condition (ii) in Definition 2. Many distributions belong to the set $\mathcal{M}$ and, thus, the models are very

[8]While 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 $(D, x) \in \mathcal{D} \times X$, if $x \in D$, then $\rho(D, x) = \mu(u \in \mathbf{R}^X | u(x) \geq u(D))$. Block and Marschak (1960)'s Theorem 3.1 proves that the two definitions are equivalent.

[9]To see this, observe that $\mathcal{P}_r \subset \mathcal{P}$ by definition. We show the converse. For any $x \in X$, let $\pi_x \in \Pi$ such that $\pi_x(x) > \pi_x(y)$ for all $y \in D$. Then $\rho^{\pi_x}(y) = 1_x(y)$ for any $y \in X$, where $1_x(y) = 1$ if $y = x$ and $1_x(y) = 0$ if $y \neq x$. For any $\rho \in \mathcal{P}$, define $\rho' = \sum_{x \in X} \rho(x) \rho^{\pi_x}$. Then, $\rho' \in \mathcal{P}_r$ and $\rho'(x) = \rho(x)$ for any $x \in X$, as desired. Hence, $\mathcal{P} \subset \mathcal{P}_r$.

[10]We do not include monomials of degree 0 (i.e., constants). For example, if $k = 2$ and $d = 2$, then $p_d(x) = (x(1), x(2), x(1)^2, x(1)x(2), x(2)^2)$ where $x = (x(1), x(2))$. The order of elements in $p_d(x)$ does not matter.

[11]The density exists if m is a Radon measure and is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{|X|}$. Notice also that the fact that support includes 0 implies that for any $x, y \in X$, $\mu(\{\varepsilon | \varepsilon_y - \varepsilon_x > 0\}) > 0$ and $\mu(\{\varepsilon | \varepsilon_x - \varepsilon_y > 0\}) > 0$.

[12]The notation s in the subscript indicates utility shock.

general: logit, probit, and nested logit models are all in the class of utility-shock models. For a logit model, μ is a multivariate iid extreme-value type-I distribution; for a probit model, μ is the multivariate standard normal distribution; for a nested logit model, μ is a generalized extreme value distribution (see Train (2009)).

The next definition is a random-coefficient version of the utility-shock models.

Definition 3. Let d be a positive integer and $\eta \in \mathbf{R}^{|X|}$ be a real vector of fixed effects. A stochastic choice function ρ is called a *degree-d random-coefficient utility-shock model* if there exist a probability measure $\mu \in \mathcal{M}$ and a Borel probability measure m such that for all $(D, x) \in \mathcal{D} \times X$, if $x \in D$, then

$$\rho(D,x) = \int \mu(\{\varepsilon | \beta \cdot p_d(x) + \eta(x) + \varepsilon(x) > \beta \cdot p_d(y) + \eta(y) + \varepsilon(y) \forall y \in D \setminus \{x\}\}) dm(\beta).$$

When m has a finite support, then ρ is called a *finite mixture* of degree-d utility-shock models. The set of degree-d random-coefficient utility-shock models with distribution μ and fixed effects η is denoted by $\mathcal{P}_{rs}(d, \eta|\mu)$. When the context makes clear which distribution μ we are considering, we omit the word "with distribution μ."

A widely used special case of the above models is the *mixed logit models*.

Definition 4. Let d be a positive integer and $\eta \in \mathbf{R}^{|X|}$ be a real vector of fixed effects. A stochastic choice function ρ is called a *degree-d mixed logit model* if there exists a Borel probability measure m such that for all $(D, x) \in \mathcal{D} \times X$, if $x \in D$, then

$$\rho(D,x) = \int \frac{\exp(\beta \cdot p_d(x) + \eta(x))}{\sum_{y \in D} \exp(\beta \cdot p_d(y) + \eta(y))} dm(\beta). \tag{2}$$

The set of degree-d mixed logit models with fixed effects η is denoted by $\mathcal{P}_{ml}(d, \eta)$. When m is degenerate (that is, when $m = \delta_\beta$ for some β) in (2), then ρ is called a *logit model.* The set of degree-d logit models with fixed effects η is denoted by $\mathcal{P}_l(d, \eta)$.

As mentioned, if μ is a multivariate iid extreme-value type-I distribution, then $\mathcal{P}_s(d, \eta|\mu) = \mathcal{P}_l(d, \eta)$, and $\mathcal{P}_{rs}(d, \eta|\mu) = \mathcal{P}_{ml}(d, \eta)$, for any (d, η). Note that in most empirical applications of the mixed logit models, the mixing distribution is usually 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.

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 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 aff.C. The convex hull of a set C is denoted by co.C. The *relative interior* of a convex set C is the interior of C in the relative topology with respect to aff.C. The relative interior of C is denoted by rint.C.

3 Main Result

To state the main result of the paper, we review a basic concept in geometry: A set $Y \subset \mathbb{R}^n$ is *affinely independent* if no element in Y can be written as an affine combination of the other elements.[13]

Theorem 1. *Let d be a positive integer.*

(i) Let $\mu \in \mathcal{M}$ be any distribution of utility shock ε. If the set $\{p_d(x)|x \in X\}$ is affinely independent, then any random utility model can be approximated by a degree-d random-coefficient utility-shock model. Moreover, the approximation can be done with a finite mixture of degree-d utility-shock models without fixed effects (i.e., $\eta = 0$). That is,

$$\forall \mu \in \mathcal{M}\ \forall \rho \in \mathcal{P}_r\ \exists \rho_n \in co.\mathcal{P}_s(d, 0|\mu)\ \forall x \in D \in \mathcal{D}\ [\rho_n(D, x) \to \rho(D, x)].$$

(ii) If the set $\{p_d(x)|x \in X\}$ is not affinely independent, then there exists a random utility model that cannot be approximated by any degree-d random-coefficient utility-shock model with any fixed effects and with any distribution $\mu \in \mathcal{M}$. That is,

$$\exists \rho \in \mathcal{P}_r\ \forall \eta \in \mathbf{R}^{|X|}\ \forall \mu \in \mathcal{M}, \rho \notin cl.\mathcal{P}_{rs}(d, \eta|\mu).$$

Notice that the condition (i.e., the affine-independence of $\{p_d(x)|x \in X\}$) does *not* depend on the distribution $\mu \in \mathcal{M}$. This implies that if the affine-independence condition holds, then the approximation is possible with *any* distribution $\mu \in \mathcal{M}$. For example, any random utility model can be approximated by a finite mixture of logit models, probit models, or nested logit models of degree d without using fixed

[13] Formally, for any $y \in Y$, $y \notin$ aff.$(Y \setminus \{y\})$. That is, for any $y \in Y$, there exists no real number $\{\mu_x\}_{x \in X}$ such that $y = \sum_{x \in Y \setminus \{y\}} \mu_x x$ and $\sum_{x \in Y \setminus \{y\}} \mu_x = 1$.

effects. If the affine-independence condition fails, then there exists a random utility model that cannot be approximated by any degree-d random-coefficient utility-shock model, no matter which fixed effects η and no matter which distribution μ we use. (For example, the approximation is impossible using any degree-d mixed logit model or any degree-d random-coefficient probit-model.) In Proposition 2 in Section 4, we will give examples of the random utility models that cannot be approximated.

Although testing for affine-independence is easy, the condition can be simplified further to a generically equivalent condition. To see this, note that, for any $x \in X$ and any positive integer d, $p_d(x)$ is a $\binom{d+k}{k} - 1$ dimensional real vector. Remember this basic fact: for any set $Y \subset \mathbb{R}^n$, (i) if $|Y| > n+1$, then Y is not affinely independent; (ii) if $|Y| \leq n+1$, then Y is generically affinely independent.[14] Given these observations, Theorem 1 implies the following corollary:

Corollary 1. *Let d be a positive integer.*

(i) If $|X| \leq \binom{d+k}{k}$, then the statements in Theorem 1 (i) hold generically.

(ii) If $|X| > \binom{d+k}{k}$, then the statements in Theorem 1 (ii) hold.

We now mention three remarks on the results in order. First, most empirical applications assume that $d = 1$. Hence, the generic necessary and sufficient condition becomes $|X| \leq k+1$. We use this condition for our empirical application in Section 5. Second, note that if X is affinely independent, then it remains to be affinely independent even with small perturbations. This reflects the fact that in the generic condition what is important is the number of alternatives (i.e., $|X|$), not X itself. Third, even though the generic condition holds, the original condition of the affine independence may not hold when explanatory variables include zeroes and ones. In that case, one should check the affine-independence of $\{p_d(x)|x \in X\}$, rather than the generic condition.

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

Proposition 1. *Assume that $\mathcal{D} = \{X\}$. Let d be a positive integer.*

[14] This is a standard concept of genericity in the literature of discrete geometry. Even if a set $Y \subset \mathbb{R}^n$ is not affinely independent, then, as long as $|Y| \leq n+1$, for any $\varepsilon > 0$, there exists an affinely independent set Y', obtained from Y by moving each point by a distance of at most ε. (See Section 3 of Matousek (2013).)

(i) Let $\mu \in \mathcal{M}$ be any distribution of utility shock ε. If the set $\{p_d(x)|x \in X\}$ is convex-independent (i.e., if $p_d(x) \notin co.\{p_d(y)|y \in X \setminus x\}$ for any $x \in X$), then any random utility model can be approximated by a degree-d random-coefficient utility-shock model; moreover, the approximation can be done with a finite mixture of degree-d utility-shock models without fixed effects (i.e., $\eta = 0$). That is,

$$\forall \mu \in \mathcal{M}\ \forall \rho \in \mathcal{P}_r\ \exists \rho_n \in co.\mathcal{P}_s(d, 0|\mu)\ \forall x \in X\ \rho_n(X, x) \to \rho(X, x).$$

(ii) (a) If the set $\{p_d(x)|x \in X\}$ is not convex-independent, then there exists a random utility model that cannot be approximated by any degree-d random-coefficient utility-shock models with any distribution $\mu \in \mathcal{M}$ and without fixed effects (i.e., $\eta = 0$). That is, $\exists \rho \in \mathcal{P}_r\ \forall \mu \in \mathcal{M}\ \rho \notin cl.\mathcal{P}_{rs}(d, 0|\mu)$.
(b) However, if fixed effects are used, then, by Norets and Takahashi (2013), any random utility model can be approximated by a utility-shock model with any distribution $\mu \in \mathcal{M}$.

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

The implications of Theorem 1 and Proposition 1 are similar. One important difference arises when the conditions (i.e., the affine-independence condition in Theorem 1 and the convex-independence condition in Proposition 1) are violated. In both cases, there exists a random utility model that cannot be approximated *without* using fixed effects. However, as stated in Proposition 1 (ii)(b), if fixed effects are used, any random utility model can be approximated.[15] This is in contrast to Theorem 1 (ii), which claims that there exists a random utility model that cannot be approximated even using fixed effects.[16]

Unlike affine-independence, convex-independence does not restrict the number of elements in a convex-independent set.[17] Thus, there exists no counterpart of

[15] This statement directly follows from Norets and Takahashi (2013).

[16] This difference originates from the fact that we require approximation on all $D \in D$ in Theorem 1, while in Proposition 1, we require approximation only on X.

[17] For 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,

Corollary 1.

To conclude this section, we mention the implications of the theorem and the proposition to the empirical literature. Most empirical papers use the linear mixed logit model (i.e., $d = 1$ and μ is a multivariate iid extreme-value type-I distribution). In the papers, the convex-independence condition is usually satisfied. That is, for any alternative x it is likely that x lies outside the convex hull co.$(X \setminus \{x\})$ of the other alternatives. In fact, we will see this is the case in a dataset in section 5.

On the other hand, the condition that $|X| \leq k+1$ is often violated. (Remember that $|X|$ is the number of alternatives and k is the number of characteristics.) There are many choice situations in which $|X|$ is very large such as choices of groceries, hospitals, cars, schools, or restaurants etc. In such a dataset, the condition is likely to be violated. *This means that the linear model is rich enough to describe the choice data from a single choice set; however, the model may not be rich enough to approximate the true substitution pattern, no matter how one chooses parameters and fixed effects.*[18] Thus, researchers might want to increase the degree of polynomial or the number of characteristic variables to satisfy the affine-independence condition.[19] See section 5 for an example.

In the next subsection 3.1, we provide a sketch of proof. The sketch gives geometric insights about our results. Moreover, some concepts (Definitions 5 and 6) will be used in a later empirical section.

3.1 Proof Sketch: Lemmas

We prove Theorem 1 and Proposition 1 by using the five lemmas below. We first consider models without fixed effects (i.e., $\eta = 0$). First we define a notation: for each ranking $\pi \in \Pi$, define

$$\rho^{\pi}(D, x) = \begin{cases} 1 & \text{if } \pi(x) \geq \pi(y) \text{ for all } y \in D; \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

The function ρ^{π} gives probability one to the best alternative x in a choice set D according to the strict preference ranking π. The following fact is elementary but fundamental:

the set of points on the circumference is convex-independent.

[18] By substitution patterns we mean how choice probabilities change in different choice sets.

[19] We address possible over-fitting problems in Section 5 and the Online Appendix.

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

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 rationalizing the random utility model. The hexagons in Figure 2 further below illustrate the polytope.[20]

Lemma 1.

1. *Let $\mathcal{Q}$ be a subset of $\mathcal{P}_r$. Then $\mathcal{P}_r = cl.co.\mathcal{Q}$ if and only if, for any $\pi \in \Pi$, there exists a sequence $\{\rho_n\}_{n=1}^\infty$ of $\mathcal{Q}$ such that $\rho_n \to \rho^\pi$.*
2. *Let $\mathcal{Q}$ be a subset of $rint.\mathcal{P}_r$. Then $rint.\mathcal{P}_r = co.\mathcal{Q}$ if and only if, for any $\pi \in \Pi$, there exists a sequence $\{\rho_n\}_{n=1}^\infty$ of $\mathcal{Q}$ such that $\rho_n \to \rho^\pi$.*
3. *$\mathcal{P}_l \subset rint.\mathcal{P}_r$.*

Parts (1) and (2) of Lemma 1 give conditions under which random utility models can be approximated by a convex combination of elements of $\mathcal{Q}$. We will use the lemma with $\mathcal{Q} = \mathcal{P}_s(d, 0|\mu)$ for some $\mu \in \mathcal{M}$ (the set of degree-d utility-shock models with distribution μ and without fixed effects).

The next lemma makes it easier for us to check the conditions of Lemma 1. First we introduce a definition.

Definition 5. For any positive integer d, a ranking $\pi \in \Pi$ is *degree-d-representable in choice sets $\mathcal{D}$* if there exists a real vector β such that, for all $D \in \mathcal{D}$ and $x \in D$,

$$\pi(x) > \pi(y) \text{ for all } y \in D \setminus \{x\} \Leftrightarrow \beta \cdot p_d(x) > \beta \cdot p_d(y) \text{ for all } y \in D \setminus \{x\}. \quad (4)$$

Lemma 2. *Let d be a positive integer. For any ranking $\pi \in \Pi$, the following statements hold:*

1. *If π is degree-d-representable, then for any $\mu \in \mathcal{M}$, there exists a sequence $\{\rho_n\}_{n=1}^\infty$ of $\mathcal{P}_s(d, 0|\mu)$ such that $\rho_n \to \rho^\pi$.*
2. *If π is not degree-d-representable, then there exists no distribution $\mu \in \mathcal{M}$ and no sequence $\{\rho_n\}_{n=1}^\infty$ of $co.\mathcal{P}_s(d, 0|\mu)$ such that $\rho_n \to \rho^\pi$.*

[20] Although the geometric intuition is useful, it is important to notice that the figure oversimplifies the reality since the number (i.e., $|X|!$) of vertices and the dimension of a random utility model can be very large. To see why the dimension of a random utility model can be very large, notice that it assigns a number for each pair of $(D, x) \in \mathcal{D} \times X$. We calculate the dimension later in Proposition 4.

Notice that for any ranking, checking the degree-d-representability is easy.[21] Thus, Lemma 1 and Lemma 2 provide a testable condition under which the degree-d random-coefficient utility-shock models without fixed effects are flexible enough to approximate any random utility model.

Although checking the representability of a particular ranking is easy, checking the representability of all rankings may be computationally prohibitive. This is because the number of rankings equals $|X|!$ and can be large. To overcome this problem, we obtain a simple necessary and sufficient condition for any ranking $\pi \in \Pi$ to be representable:

Lemma 3. *Let d be a positive integer.*

1. *Any ranking is degree-d-representable in $\mathcal{D}$ if and only if the set $\{p_d(x)|x \in X\}$ is affinely independent.*
2. *Any ranking is degree-d-representable in $\{X\}$ if and only if the set $\{p_d(x)|x \in X\}$ is convex-independent.*

To understand Lemma 3 (1) geometrically, see Figure 1. In the figure, we assume that $k = 2$; we consider linear models (i.e., $p_d(x) = x$) in Figure 1 (a) and (b), and quadratic models (i.e., $d = 2$) in Figure 1 (c), respectively. In Figure 1 (a), the set $X = \{x, y, z\}$ is affinely independent. Thus, by Lemma 3 (1) (the "if" part), any ranking is degree-1-representable. For example, the ranking $\pi(x) > \pi(y) > \pi(z)$ is degree-1-representable by $\beta \in \mathbf{R}^2$, which defines the parallel hyperplanes (indifference curves) in Figure 1 (a).

On the other hand, in Figure 1 (b), the set $X = \{x, y, z, w\}$ is not affinely independent. The ranking $\pi(x) > \pi(w) > \pi(z) > \pi(y)$ is not degree-1-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 $\beta \cdot x > \beta \cdot w > \beta \cdot z > \beta \cdot y$. The existence of such a ranking is implied by the "only if" part of Lemma 3 (1).[22]

If we use ellipses as indifference curves, however, we can represent the ranking $\pi(x) > \pi(w) > \pi(z) > \pi(y)$ as in Figure 1 (c). The existence of such curves is again implied by the "if" part of Lemma 3 (1) since ellipses can be defined with at

[21] By definition, checking the represetability condition is equivalent to finding a solution to a system of finite linear inequalties defined by (4).

[22] The slope of the "indifference" line must be steeper than the slope of the line segment (z, y) (because $\pi(z) > \pi(y)$) and less steep than the slope of the line segment (z, x) (because $\pi(x) > \pi(w)$), which together impliy that $\beta \cdot z > \beta \cdot w$.

most degree-2 polynomials and the generic condition with $d = 2$ is satisfied (i.e., $|X| = 4 \leq 6 = \binom{4}{2} = \binom{d+k}{k}$ in this example.[23]

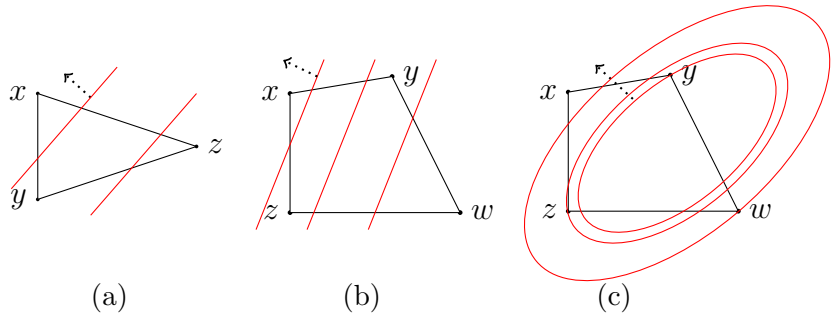

Figure 1: Illustration of the affine-independence condition.

Lemma 3 (2) is more straightforward. To see this, consider $d = 1$ for simplicity and notice that when $\mathcal{D} = \{X\}$, any $\pi \in \Pi$ is representable if and only if, for any $x \in X$, there exists β such that $\beta \cdot x > \beta \cdot y$ for all $y \in X \setminus x$, which means that X is convex-independent. By using Lemmas 1, 2, and 3, we obtain parts (i) of Theorem 1 and Proposition 1.

Remember that so far we have assumed no fixed effects (i.e., $\eta = 0$). In the following, we consider what we can accomplish with fixed effects. First, it is easy to observe that when $\mathcal{D} = \{X\}$, any stochastic choice can be approximated by using fixed effects.[24] Even for general $\mathcal{D}$, the following holds:

Observation: *For any ranking π, ρ^π can be approximated by a utility-shock model with fixed effects.*[25]

However, this is not enough to approximate any random utility model. As an illustration, consider only two fixed effects, η_1 and η_2, and see Figure 2 below. In the figure, given degree d, the two convex sets in the hexagon correspond to $\mathcal{P}_{rs}(d, \eta_1|\mu)$ and $\mathcal{P}_{rs}(d, \eta_2|\mu)$, respectively. Notice that all vertices in the figure can be approximated by elements of $\mathcal{P}_{rs}(d, \eta_1|\mu)$ or $\mathcal{P}_{rs}(d, \eta_2|\mu)$. However, some areas of the hexagon are not covered by either $\mathcal{P}_{rs}(d, \eta_1|\mu)$ or $\mathcal{P}_{rs}(d, \eta_2|\mu)$.

In reality, the problem is more complicated since we need to consider the union of all possible values of fixed effects, and thus the union of the continuum of convex

[23] In fact, we verified that the affine-independence condition is satisfied with $d = 2$.

[24] This can be intuitively understood since we can choose $|X|$ parameters (i.e., $\{\eta(x)\}_{x \in X}$) to fit $|X|$ data points (i.e., $\{\rho(X, x)\}_{x \in X}$).

[25] To see this fix π and choose $\eta \in \mathbf{R}^X$ such that $\eta(x) > \eta(y)$ if and only if $\pi(x) > \pi(y)$. Then, it can be shown that a utility-shock model ρ_n defined by $\rho_n(D, x) = \mu\{n\eta(x) + \varepsilon(x) \geq n\eta(y) + \varepsilon(y)$ for all $y \in D \setminus \{x\}\}$ converges to $\rho^\pi(D, x)$ as $n \to \infty$.

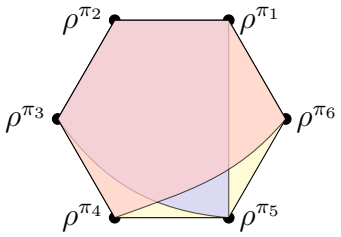

Figure 2: Illustration of $\mathcal{P}_{rs}(d, \eta_1|\mu)$ and $\mathcal{P}_{rs}(d, \eta_2|\mu)$

sets $\mathcal{P}_{rs}(d, \eta|\mu)$ across all values of $\eta \in \mathbf{R}^{|X|}$.[26] Nevertheless, Lemma 4 provides an answer. First we introduce a new notation.

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

Lemma 4. *Let $\alpha \in (0,1)$ and $\mu \in \mathcal{M}$. For each ranking π that is not degree-d-representable, there exists a neighborhood U of $\alpha\rho^{\pi}+(1-\alpha)\rho^{\pi^-}$ such that any random utility that belongs to U cannot be approximated by any degree-d random-coefficient utility-shock model with any fixed effects. That is, $\forall \rho \in \mathcal{P}_r \cap U,\ \rho \notin cl. \bigcup_\eta \mathcal{P}_{rs}(d, \eta|\mu)$.*

To prove the lemma, we need to prove the following two statements: (a) any strict convex combination between ρ^{π} and ρ^{π^-} cannot be approximated by a degenerate utility-shock model with fixed effects; and (b) moreover it cannot be approximated by a nondegenerate random-coefficient utility-shock model even with any fixed effects. We prove statement (a) in the appendix. To show statement (b), we introduce the following concept:

Definition 7. The two rankings π and π' are *adjacent* if there exists $t \in \mathbf{R}^{\mathcal{D}\times X}$ and $a \in \mathbf{R}$ such that (i) $\rho^{\pi} \cdot t = a = \rho^{\pi'} \cdot t$, and (ii) for any $\hat{\pi}$, if $\pi \neq \hat{\pi} \neq \pi'$, then $\rho^{\hat{\pi}} \cdot t > a$.[27]

[26] Mathematically speaking, the difficulty arises because the set of all degree-d utility-shock models with distribution μ and with any fixed effects (i.e., $\bigcup_{\eta \in \mathbf{R}^X} \text{co.}\mathcal{P}_s(d, \eta|\mu)$) may not be convex, although given η, each set $\text{co.}\mathcal{P}_s(d, \eta|\mu)$ is convex. This is because mixtures can be taken only over β but not over η. Thus approximating vertices is not enough for approximation over the polytope of random utility models.

[27] $t \in \mathbf{R}^{\mathcal{D}\times X}$ is a vector that gives a real number for each pair of $(D, x) \in \mathcal{D} \times X$. For any $\rho \in \mathcal{P}$, $\rho \cdot t = \sum_{(D,x)\in\mathcal{D}\times X} \rho(D, x)t(D, x)$.

For example, in Figure 2, ρ^{π_i} and $\rho^{\pi_{i+1}}$ are adjacent for each $i \leq 5$. 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:[28]

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

The characterization of adjacency of vertices for the case $\mathcal{D} = 2^X \setminus \emptyset$ appears in Doignon and Saito (2022). Lemma 5 holds even for the case in which $\mathcal{D} \neq 2^X \setminus \emptyset$ as long as $\mathcal{D}$ contains all binary and trinary sets.[29] Lemma 5 allows us to complete the proof of Lemma 4. If π is not representable, then π^- is also not representable. Although fixed effects are powerful enough to approximate each vertex ρ^{π}, we will prove that it is not powerful enough to approximate both ρ^{π} and ρ^{π^-} by using the same fixed effects. Thus, no strict convex combination of ρ^{π} and ρ^{π^-} can be approximated by the degree-d random-coefficient utility-shock models with distribution μ, no matter which fixed effects we use. Notice that this conclusion does *not* follow if ρ^{π} and ρ^{π^-} are not adjacent since the mixture may be represented in a different way. This proves statement (b) and thus, Lemma 4. Lemmas 1, 2, 3, and 4 prove statement (ii) of Theorem 1, as the proof in the appendix formalizes.

4 Measuring Approximation Errors

In this section, we study the following question: When the condition in the theorem is not satisfied, how large are the approximation errors?

We first define the distance function as follows: For any $\rho, \hat{\rho} \in \mathcal{P}$, define $d(\rho, \hat{\rho}) = \|\rho - \hat{\rho}\|/|\mathcal{D}|$, where $\|\cdot\|$ is the Euclidian norm.[30] In our analysis, $\hat{\rho}$ is a given stochastic choice function; ρ is a random-coefficient utility-shock model by which we appproximate $\hat{\rho}$. The distance $d(\rho, \hat{\rho})$ captures the average distance between ρ and $\hat{\rho}$. For example, $d(\rho, \hat{\rho}) = 0.1$ means that on average across choice sets, the estimated stochastic choice function $\hat{\rho}$ is distant from the given stochastic choice function ρ by 10 percentage points.

Given any approximation target $\hat{\rho} \in \mathcal{P}_r$ and fixing a distribution function μ over the shock ε, the approximation error when one uses the set $\mathcal{P}_{rs}(d, \eta|\mu)$ of degree-d

[28] Our discussions with Jean-Paul Doignon and Haruki Kono were very helpful for obtaining this result.
[29] We are grateful Haruki Kono for pointing out this fact.
[30] Rmember $\|\hat{\rho} - \rho\| = \sqrt{\sum_{(x,D)\in X\times\mathcal{D}}(\hat{\rho}(D,x) - \rho(D,x))^2}$.

random-coefficient utility-shock models with fixed effects η is defined as:

$$\inf_{\rho \in \mathcal{P}_{rs}(d,\eta|\mu)} d(\rho, \hat{\rho}). \tag{5}$$

We call (5) the *approximation error to* $\hat{\rho}$ by degree-d random-coefficient utility-shock models with fixed effects η. To calculate (5), we need to specify a random utility model $\hat{\rho}$ that cannot be approximated. Proposition 2 gives an idea about how to choose $\hat{\rho}$. The following result holds with any distribution of utility shock ε.

Proposition 2. *Let $\mu \in \mathcal{M}$ be any distribution of utility shock ε. Let d be a positive integer. If $\{p_d(x)|x \in X\}$ is not affinely independent, then for each ranking π that is not degree-d-representable, the following statements hold:*[31]

(i) There exists a neighborhood U of ρ^π such that any random utility model that belongs to U cannot be approximated by any degree-d random-coefficient utility-shock model without fixed effects. That is, $\forall \rho \in \mathcal{P}_r \cap U,\ \rho \notin cl.\mathcal{P}_{rs}(d, 0|\mu)$.

(ii) For each $\alpha \in (0,1)$, 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 by any degree-d random-coefficient utility-shock models. That is, $\forall \rho \in \mathcal{P}_r \cap U,\ \rho \notin cl.\bigcup_\eta \mathcal{P}_{rs}(d, \eta|\mu)$.

As for the second statement (ii), remember that by using fixed effects, we can approximate any ρ^π. However, approximating a mixture between ρ^π and ρ^{π^-} is impossible even using fixed effects when π is not representable.

4.1 EM Algorithm

Given $\hat{\rho}$, we propose two algorithms to solve (5) and compute the approximation errors. The first is the standard EM (Expectation-Maximization) algorithm. The second algorithm is a greedy algorithm, as we discuss in next section 4.2.

We use the EM algorithm to estimate a finite mixture logit model that maximizes the likelihood taking $\hat{\rho}$ as the observed choice probabilities. The Online Appendix shows that the resulting finite mixture logit model is indeed a solution to (5) when the affine-independence condition is satisfied. One difficulty to use the EM algorithm in our problem is that it is not clear how many mixtures to include. To overcome this difficulty, in this subsection, we provide a theoretical result that simplifies the set

[31] If $\{p_d(x)|x \in X\}$ is not affinely independent, then such a ranking exists.

of random-coefficient utility-shock models and provides guidance about how many mixtures we need to use.

The first result (Proposition 3) shows that any (continuous) mixture model can be represented as a finite mixture as long as the set of alternatives is finite. The second result (Proposition 4) characterizes the affine hull of the random utility models, which allows us to calculate the dimension of the set of random utility models. This, in turn, gives us an upper bound on the number of mixtures through Caratheodory's theorem.

Proposition 3. *Let $\mathcal{Q}$ be a subset of the set of stochastic choice functions. Then $\left\{ \int \rho dm(\rho) | m \in \Delta(\mathcal{Q}) \right\} = co.\mathcal{Q}$, where $\Delta(\mathcal{Q})$ denotes the set of probability measures over $\mathcal{Q}$.*

This proposition implies that focusing on finite mixtures is without loss of generality as long as X is finite. In particular, it implies that the set $\mathcal{P}_{rs}(d, \eta|\mu) = \text{co.}\mathcal{P}_s(d, \eta|\mu)$ for any d, η, μ.[32] Although Proposition 3 reduces $\mathcal{P}_{rs}(d, \eta|\mu)$ to $\text{co.}\mathcal{P}_s(d, \eta|\mu)$, the set $\text{co.}\mathcal{P}_s(d, \eta|\mu)$ is still big since it contains any (finite) number of mixtures of degree-d utility-shock models. By Caratheodory's theorem, however, there turn out to be at most $\dim \mathcal{P}_r + 1$ elements.

Corollary 2. *Let $\mu \in M$. For any positive integer d and any $\eta \in \mathbf{R}^{|X|}$, $\mathcal{P}_{rs}(d, \eta|\mu) = co.\mathcal{P}_s(d, \eta|\mu) = \left\{ \sum_{m=1}^{M} \lambda_m \rho_m \middle| \rho_m \in \mathcal{P}_s(d, \eta|\mu), \lambda_m \geq 0 \forall m = 1, \ldots, M, \sum_{m=1}^{M} \lambda_m = 1 \right\}$, where $M = \dim \mathcal{P}_r + 1$.*

To obtain the number $\dim \mathcal{P}_r$, we characterize the affine hull of the set $\mathcal{P}_r$ of random utility models.

Proposition 4. *The affine hull of $\mathcal{P}_r$ is $\mathcal{P}_{\pm} \equiv \left\{ q \in \mathbf{R}^{\mathcal{D} \times X} \middle| (i) \sum_{x \in D} q(D, x) = 1 \forall D \in \mathcal{D}; (ii)\ q(D, x) = 0\ \forall x \notin D \in \mathcal{D} \right\}$. Hence $\dim \mathcal{P}_r \equiv \dim \mathcal{P}_{\pm} = \sum_{D \in \mathcal{D}} (|D| - 1)$.*

Corollary 2 and Proposition 4 imply that in order to analyze the flexibility of the random-coefficient models, 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 data with $|X| = 4$. 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.e., $\mathcal{D} = 2^X \setminus \emptyset$).

[32] This result may not hold when X is continuous. See Lu and Saito (2021).

4.2 Greedy Algorithm

Even with the modification proposed in the previous subsection, the EM algorithm may converge only to a local minimum (Dempster et al., 1977).[33] This concern motivates us to propose a second algorithm, the greedy algorithm, which is inspired by Barron et al. (2008). This algorithm has the useful feature that, given the setup of our problem, it will always return a global minimum (up to small approximation errors which can be made arbitrarily small).

The algorithm takes a stochastic choice function $\hat{\rho}$ and a fixed effects vector η as input and returns a solution to (5). 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 $\rho^1 = \arg\inf_{\rho \in \text{cl.}\mathcal{P}_s(d,\eta|\mu)} ||\hat{\rho} - \rho||^2$.
- **Step n**, $n \geq 2$:
  - Consider a set of grids $\alpha_n = \{\frac{2}{k+1}\}_{k=1}^n$.
  - Find $(\alpha_n^*, \rho_n^*) = \arg\inf_{(\alpha,\rho) \in \alpha_n \times \text{cl.}\mathcal{P}_s(d,\eta|\mu)} ||\hat{\rho} - (1-\alpha)\rho^{n-1} - \alpha\rho||^2$.
  - 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 quickly.

Proposition 5. *Let $\hat{\rho} \in \mathcal{P}$ be any stochastic function and d be a positive integer, $\eta \in \mathbf{R}^{|X|}$, and $\mu \in \mathcal{M}$. Define $d^* = \inf_{\rho \in \mathcal{P}_{rs}(d,\eta|\mu)} d(\rho, \hat{\rho})$. 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 there exists a T such that*

$$d(\rho^n, \hat{\rho}) - d^* \leq \sqrt{\frac{T}{n+1}},$$

where T can be chosen to depend only on $|\mathcal{D}|$.

For our implementation, the terminating criterion is when the number of steps taken reaches 1000. With 1000 steps, (5) implies the margin of error is within 0.026.[34]

[33] To alleviate this problem, in the next section, we run the EM algorithm multiple times, each time with a random initialization.

[34] This is calculated by computing the constant T in Proposition 5. In our case, the T is equal to $88/121$, so the margin of error is $\sqrt{88/(121 \times 1001)} \approx 0.026$. See footnote 48 for the full calculation.

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.

5 Application to Data

In this section, we quantify approximation errors with and without fixed effects, by using data on fishing-site choices from Thomson and Crooke (1991).[35] In the data set, 1182 individuals choose among 4 alternative fishing modes, namely, $X = \{x_{\text{beach}}, x_{\text{boat}}, x_{\text{charter}}, x_{\text{pier}}\}$, which denote fishing from the beach, a private boat, a charter boat or a pier, respectively. Each alternative $x \equiv (x(1), x(2))$ is described by two characteristics (i.e., $k = 2$). The first characteristic $x(1)$ is the fishing mode's price, while the other characteristic $x(2)$ is the *catch rate*, defined as a per-hour-fished basis for each major species by fishing mode.[36]

Throughout this section, we will focus on the mixed logit models (i.e., let μ be a multivariate iid extreme-value type-I distribution) since they are the most widely used models. We also assume that $\mathcal{D} = 2^X \setminus \emptyset$. Thus $\dim \mathcal{P}_r = \sum_{D \in \mathcal{D}} (|D| - 1) = 3 + 2 \times 4 + 1 \times 6 = 17$ by Proposition 4. It is therefore without loss of generality to assume that the number M of mixtures is less than or equal to 18.

5.1 Application of Theorem 1

By Theorem 1, X is affinely independent if and only if the class of linear (degree-1) mixed logit models with fixed effects is flexible enough to approximate any random utility model. In the data set, we have $|X| = 4$ alternatives and $k = 2$ characteristics. Thus, the condition in Corollary 1 is violated and X is not affinely independent. This observation motivates us to compute approximation errors of the linear mixed logit models without fixed effects (in subsection 5.2) and the errors with fixed effects (in subsection 5.3).

On the other hand, with $d = 2$, the generic condition for representability in Corollary 1 is satisfied, since $4 = |X| \leq \binom{2+2}{2} = 6$. In fact, we verified that $\{p_d(x) | x \in X\}$ is affinely independent when $d = 2$. Thus, by Theorem 1, the

[35] The dataset is taken directly from the R package 'mlogit' by Croissant (2020). The data have been used by Herriges and Kling (1999) and Cameron and Trivedi (2005) (p.464).

[36] In the original study, the values of $x(1)$ and $x(2)$ depend on each individual. For our analysis, we aggregate them by taking the average over individuals. The Online Appendix provides more details.

degree-2 mixed logit models should be flexible enough to approximate any random utility model. This theoretical implication is also empirically verified below.

5.2 Approximation Error without Fixed Effects

In this section, we obtain approximation errors without fixed effects. By Proposition 2, there exists a ranking π that is not degree-1-representable and corresponding deterministic choice functions ρ^π that are not approximated by any linear mixed logit model without fixed effects. Since there are four alternatives, there are twenty four possible rankings. Among them, twelve rankings cannot be approximated by linear mixed logit models, as shown in Table 1.

Table 1: Approximation errors to preference rankings ρ^π

Ranking π	$d=1$		$d=2$	
	Greedy	EM	Greedy	EM
(1)	(2)	(3)	(4)	(5)
Unrepresentable Rankings				
$\pi(1)>\pi(2)>\pi(3)>\pi(4)$	0.218	0.227	0.000	0.000
$\pi(1)>\pi(2)>\pi(4)>\pi(3)$	0.202	0.211	0.000	0.000
$\pi(1)>\pi(3)>\pi(2)>\pi(4)$	0.128	0.115	0.000	0.000
$\pi(1)>\pi(4)>\pi(2)>\pi(3)$	0.126	0.165	0.000	0.000
$\pi(2)>\pi(1)>\pi(3)>\pi(4)$	0.138	0.147	0.000	0.000
$\pi(2)>\pi(1)>\pi(4)>\pi(3)$	0.118	0.123	0.000	0.000
$\pi(3)>\pi(2)>\pi(4)>\pi(1)$	0.091	0.096	0.000	0.000
$\pi(3)>\pi(4)>\pi(1)>\pi(2)$	0.121	0.128	0.000	0.000
$\pi(3)>\pi(4)>\pi(2)>\pi(1)$	0.149	0.160	0.000	0.000
$\pi(4)>\pi(2)>\pi(3)>\pi(1)$	0.113	0.115	0.000	0.000
$\pi(4)>\pi(3)>\pi(1)>\pi(2)$	0.157	0.155	0.000	0.000
$\pi(4)>\pi(3)>\pi(2)>\pi(1)$	0.182	0.185	0.000	0.000
Representable Rankings	0.000	0.000	0.000	0.000

Note: The numbers in the table show the approximation errors to each ρ^π, where each preference ranking π is defined in Column (1). The numbers 1, 2, 3, 4 denote beach, boat, charter, pier, respectively. For each ranking, columns (2) and (3) show the approximation errors of the linear mixed logit models computed by the greedy algorithm and the EM algorithm, respectively. Columns (4) and (5) show the approximation errors of the degree-2 (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 table shows the approximation errors of the degree-1 or degree-2 mixed logit models obtained by the greedy algorithm and the EM algorithm. In both algorithms, the approximation errors for unrepresentable rankings π are almost always larger

than 0.1, which means that even the best possible linear mixed logit model deviates from the corresponding choice probabilities ρ^π by 10 percent or more on average. 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 mixed logit models are more than 0.2. Notice that these two rankings are the only rankings in which the alternative 1 (i.e., beach) is the best and the alternative 2 (i.e., private boat) is the second-best. This means that as long as we use the linear mixed logit models, no matter how we choose parameters, it is difficult to capture the *substitution patterns* from the alternative 1 to the alternative 2 (i.e, the change of consumer's choices from the alternative 1 to the alternative 2).

On the other hand, the approximation error for a representable rankings π is almost always zero, as the theorem predicts, as shown in the bottom row of the table. Also, the approximation errors by degree-2 mixed logit models are also almost zero, as the theorem again predicts (column (4) and (5) in the table).

5.3 Approximation Errors with Fixed Effects

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

Ranking π	$d=1$		$d=2$	
	Greedy	EM	Greedy	EM
(1)	(2)	(3)	(4)	(5)
Unrepresentable Rankings				
$\pi(1) > \pi(2) > \pi(3) > \pi(4)$	0.069	0.077	0.000	0.000
$\pi(1) > \pi(2) > \pi(4) > \pi(3)$	0.069	0.079	0.000	0.000
$\pi(1) > \pi(3) > \pi(2) > \pi(4)$	0.049	0.077	0.000	0.000
$\pi(1) > \pi(4) > \pi(2) > \pi(3)$	0.049	0.052	0.000	0.000
$\pi(2) > \pi(1) > \pi(3) > \pi(4)$	0.058	0.075	0.000	0.000
$\pi(2) > \pi(1) > \pi(4) > \pi(3)$	0.058	0.060	0.000	0.000
Representable Rankings	0.000	0.000	0.000	0.000

Notes: 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 Column (1). 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.

In this section, we obtain the approximation errors with fixed effects. By using fixed effects, we can approximate ρ^π for any ranking π. By Proposition 2, however, for each unrepresentable ranking π and each $\alpha \in (0, 1)$, any random utility model

in a neighborhood of $\alpha\rho^{\pi} + (1-\alpha)\rho^{\pi^-}$ cannot be approximated by the linear mixed logit models with fixed effects. In Table 2, we show the approximation error to $\frac{1}{2}\rho^{\pi} + \frac{1}{2}\rho^{\pi^-}$ for each unrepresentable π.

In both algorithms, the approximation errors to $\frac{1}{2}\rho^{\pi} + \frac{1}{2}\rho^{\pi^-}$ are always larger than around 5 percent if π is unrepresentable. This means that even the best possible linear mixed logit model deviates from $\frac{1}{2}\rho^{\pi} + \frac{1}{2}\rho^{\pi^-}$ by 5 percentage points or more on average.

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. Also, the approximation errors by degree-2 mixed logit models are also almost zero, as the theorem again predicts.

Overall, the widely-used linear mixed logit models fail to approximate random utility models in this data set. The approximation errors are also substantial. This result demonstrates how the affine-independence condition in Theorem 1 provides a simple way to check whether a random-coefficient model can provide a good approximation of random utility models. A more practical implication is that researchers might want to increase the degree of polynomial or the number of characteristic variables to satisfy the affine-independence condition.

Given that our model perfectly fits the observed choice data when the condition holds, a natural concern is the over-fitting problem. To address this concern, in the Online Appendix, we evaluate its out-of-sample performance by constructing a mixed logit model that best approximates a part of the data and measuring its approximation performance to the remaining data. We find that our model performs better or equally well compared to standard models models in the simulation, not only in terms of in-sample fit but also in terms of out-of-sample fit. See section B of Online Appendix.

References

Ackerberg, D. A. and M. Rysman (2005): "Unobserved Product Differentiation in Discrete Choice Models: Estimating Price Elasticities and Welfare Effects," *RAND Journal of Economics*, 36, 771–788.

Ahn, D. and T. Sarver (2013): "Preference for Flexibility and Random Choice," *Econometrica*, 81, 341–361.

Apesteguia, J. and M. Ballester (2018): "Monotone Stochastic Choice Mod-

els: The Case of Risk and Time Preferences," *Journal of Political Economy*, 126, 74–106.

Apesteguia, J., M. Ballester, and J. Lu (2017): "Single-Crossing Random Utility Models," *Econometrica*, 85, 661–674.

Athey, S. and G. W. Imbens (2007): "Discrete Choice Models with Multiple Unobserved Choice Characteristics," *International Economic Review*, 48, 1159–1192.

Bajari, P., D. Nekipelov, S. P. Ryan, and M. Yang (2015): "Machine Learning Methods for Demand Estimation," *American Economic Review*, 105, 481–85.

Barron, A. R., A. Cohen, W. Dahmen, and R. A. DeVore (2008): "Approximation and Learning by Greedy Algorithms," *Annals of Statistics*, 36, 64–94.

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

Berry, S. T. and P. A. Haile (2014): "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 82, 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., F. Maccheroni, M. Marinacci, and A. Rustichini (2018): "Multinomial Logit Processes and Preference Discovery: Inside and Outside the Black Box," Working Paper.

——— (2022): "Law of Demand and Stochastic Choice," *Theory and Decision*, 92, 513–529.

Chambers, C. P., T. Cuhadaroglu, and Y. Masatlioglu (2020): "Behavioral Influence," *Journal of the European Economic Association*.

Chambers, C. P., Y. Masatlioglu, and C. Raymond (2021a): "Weighted Linear Discrete Choice," Working Paper.

Chambers, C. P., Y. Masatlioglu, and C. Turansick (2021b): "Correlated Choice," Working Paper.

Compiani, G. (2022): "Market Counterfactuals and the Specification of Multiproduct Demand: A Nonparametric Approach," *Quantitative Economics*, 13, 545–591.

Croissant, Y. (2020): "Estimation of Random Utility Models in R: The mlogit Package," *Journal of Statistical Software*, 95, 1–41.

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, 1–22.

Doignon, J.-P. and K. Saito (2022): "Adjacencies on Random Ordering Polytopes and Flow Polytopes," Working Paper.

Duraj, J. (2018): "Dynamic Random Subjective Expected Utility," Working Paper.

Echenique, F. and K. Saito (2019): "General Luce Model," *Economic Theory*, 68, 811–826.

Frick, M., R. Iijima, and T. Strzalecki (2019): "Dynamic Random Utility," *Econometrica*, 87, 1941–2002.

Fudenberg, D. and T. Strzalecki (2015): "Dynamic Logit with Choice Aversion," *Econometrica*, 83, 651–691.

Ghalanos, A. and S. Theussl (2015): *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*, r package version 1.16.

Gillen, B. J., S. Montero, H. R. Moon, and M. Shum (2019): "BLP-2LASSO for Aggregate Discrete Choice Models with Rich Covariates," *Econometrics Journal*, 21, 1–23.

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

Gul, F. and W. Pesendorfer (2006): "Random Expected Utility," *Econometrica*, 74, 121–146.

Herriges, J. A. and C. L. Kling (1999): "Nonlinear Income Effects in Random Utility Models," *Review of Economics and Statistics*, 81, 62–72.

Horan, S. (2018): "Threshold Luce Rules," Working Paper.

Lin, Y. (2019): "Random Non-Expected Utility: Non-Uniqueness," Working Paper.

Lu, J. (2016): "Random Choice and Private Information," *Econometrica*, 84, 1983–2027.

——— (2021): "Random ambiguity," *Theoretical Economics*, 16, 539–570.

Lu, J. and K. Saito (2018): "Random Intertemporal Choice," *Journal of Economic Theory*, 177.

——— (2021a): "Mixed Logit and Pure Characteristic Models," Working Paper.

——— (2021b): "Repeated Choice," Working Paper.

Luce, 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*, 447–470.

Norets, A. and S. Takahashi (2013): "On the Surjectivity of the Mapping between Utilities and Choice Probabilities," *Quantitative Economics*, 4, 149–155.

Rockafellar, R. T. (2015): *Convex Analysis*, Princeton University Press.

Ruiz, F. J., S. Athey, and D. M. Blei (2020): "Shopper: A Probabilistic Model of Consumer Choice with Substitutes and Complements," *Annals of Applied Statistics*, 14, 1–27.

Saito, K. (2018): "Axiomatizations of the Mixed Logit Model," Working Paper.

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 (2019): "Nonparametric Estimates of Demand in the California Health Insurance Exchange," Working Paper.

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.

Tserenjigmid, G. and M. Kovach (2020): "Behavioral Foundations of Nested Stochastic Choice and Nested Logit," Working Paper.

Ye, Y. (1987): "Interior Algorithms for Linear, Quadratic, and Linearly Constrained Non-Linear Programming," Ph.D. thesis, Department of ESS, Stanford University.

A Theoretical Appendix

A.1 Proof of Theorem 1

Lemma 1,2, and 3 –(1) imply statements (i) of Theorem 1 and Proposition 1. Statements (ii) of Theorem 1 and Proposition 1 follow from Lemma 1,2, 3–(1), and 4.

A.2 Proof of Proposition 1

Lemma 1,2, and 3 –(2) imply statement (i) and parts of statement (ii) of Proposition 1. The last statement of statement (ii) can be proved as follows. Consider any stochastic choice function ρ. Then there exists a sequence of stochastic choice functions $\{\rho_n\}$ such that $\rho_n \to \rho$ and $\rho_n(X, x) > 0$ for any $x \in X$. Fix $\mu \in \mathcal{M}$. Note that our assumption of the convexity of the support implies the connectedness. By Corollary 1 of Norets and Takahashi (2013), ρ_n can be represented as the utility-shock models.

A.3 Proof of Lemma 1

A.3.1 Statement (1)

If direction is obvious, we prove only if direction. By assumption, $\mathcal{P}_r = \text{cl.co.}\mathcal{Q} = \text{co.cl.}\mathcal{Q}$, where the last equality holds because $\mathcal{Q}$ is bounded and by Theorem 17.2 of Rockafellar (2015). Since $\mathcal{P}_r = \text{co.cl.}\mathcal{Q}$, for any $\pi \in \Pi$, there exist positive numbers $\{\lambda_i\}_{i=1}^m$ such that $\sum_{i=1}^m \lambda_i = 1$ and a convergent sequence $\{\rho_n^i\}_{n=1}^\infty$ of $\mathcal{Q}$ for each $i \in \{1, \ldots, m\}$ such that $\sum_{i=1}^m \lambda_i \rho_n^i \to \sum_{i=1}^m \lambda_i \rho^i = \rho^\pi$ as $n \to \infty$, where $\rho_n^i \to \rho^i$. Since ρ^π is a vertex of $\mathcal{P}_r$ and thus an exposed point[37], $\rho^i = \rho^\pi$ for all i.

A.3.2 Statement (2)

Let $\mathcal{Q}$ be any subset of $\text{rint.}\mathcal{P}_r$. We will show that $\text{rint.}\mathcal{P}_r = \text{co.}\mathcal{Q}$ if and only if for any $\pi \in \Pi$ there exists a sequence $\{\rho_n\}_{n=1}^\infty$ of $\mathcal{Q}$ such that $\rho_n \to \rho^\pi$ as $n \to \infty$.

Step 1: We will show the if part of the statement. Suppose by way of contradiction that there exists $\rho \in \text{rint.}\mathcal{P}_r \setminus \text{co.}\mathcal{Q}$. Because $\text{co.}\mathcal{Q} \neq \emptyset$, we obtain $\text{rint.co.}\mathcal{Q} \neq \emptyset$. Since $\rho \notin \text{co.}\mathcal{Q}$, by the proper separating hyperplane theorem (Theorem 11.3 of

[37]A 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)

Rockafellar (2015)), there exist $t \in \mathbf{R}^{\mathcal{D}\times X} \setminus \{0\}$ and $a \in \mathbf{R}$ such that

$$\rho \cdot t \geq a \geq \rho' \cdot t \text{ for any } \rho' \in \text{co.}\mathcal{Q} \text{ and } a > \rho'' \cdot t \text{ for some } \rho'' \in \text{co.}\mathcal{Q}. \tag{6}$$

We obtain a contradiction by two substeps.

Step 1.1: We show there exists $\hat{\rho} \in \mathcal{P}_r$ such that $\hat{\rho} \cdot t > \rho \cdot t$. To prove the step, remember that there exists $\rho'' \in \text{co.}\mathcal{Q}$ such that $\rho'' \cdot t < \rho \cdot t$. Moreover, since $\mathcal{Q} \subset \mathcal{P}_r$ and $\mathcal{P}_r$ is convex, it follows that $\rho'' \in \text{co.}\mathcal{Q} \subset \mathcal{P}_r$. Since $\rho \in \text{rint.}\mathcal{P}_r$, there exists $\lambda > 1$ such that $\lambda\rho + (1-\lambda)\rho'' \in \mathcal{P}_r$. Moreover, $(\lambda\rho + (1-\lambda)\rho'') \cdot t = \lambda\rho \cdot t + (1-\lambda)\rho'' \cdot t = \rho \cdot t + (\lambda - 1)(\rho \cdot t - \rho'' \cdot t) > \rho \cdot t$, where the last inequality holds because $\lambda > 1$ and $\rho'' \cdot t < \rho \cdot t$. So $\lambda\rho + (1-\lambda)\rho'' \in \mathcal{P}_r$.

Step 1.2: There exists $\rho' \in \text{co.}\mathcal{Q}$ such that $\rho' \cdot t > \rho \cdot t$, which contradicts with (6). To prove the step, let $\hat{\rho}$ be as in Substep 1.1. Since $\hat{\rho} \in \mathcal{P}_r$, there exist nonnegative numbers $\{\hat{\lambda}_\pi\}_{\pi\in\Pi}$ such that $\hat{\rho} = \sum_{\pi\in\Pi} \hat{\lambda}_\pi \rho^\pi$ and $\sum_{\pi\in\Pi} \hat{\lambda}_\pi = 1$.

By the supposition of the lemma, for any $\pi \in \Pi$, there exists a sequence $\{\rho'_n\}_{n=1}^\infty$ of $\mathcal{Q}$ such that $\rho'_n \to \rho^\pi$ as $n \to \infty$. Therefore, for any $\pi \in \Pi$ and any positive number ε, there exists $\rho'_\pi \in \{\rho'_n\}_{n=1}^\infty$ such that $\|\rho'_\pi - \rho^\pi\| < \varepsilon$. Define $\rho' = \sum_{\pi\in\Pi} \hat{\lambda}_\pi \rho'_\pi$. Then $\rho' \in \text{co.}\mathcal{Q}$ and $\|\rho' - \hat{\rho}\| = \left\|\sum_{\pi\in\Pi} \hat{\lambda}_\pi(\rho'_\pi - \rho^\pi)\right\| \leq \sum_{\pi\in\Pi} \hat{\lambda}_\pi \|\rho'_\pi - \rho^\pi\| \leq \sum_{\pi\in\Pi} \hat{\lambda}_\pi \varepsilon = \varepsilon$. Therefore, $|\rho' \cdot t - \hat{\rho} \cdot t| \leq \|t\|\|\rho' - \hat{\rho}\| \leq \|t\|\varepsilon$. Since $t \cdot \hat{\rho} > t \cdot \rho$, then by choosing ε small enough, we obtain $\rho' \cdot t > \rho \cdot t$.

Step 2: We will show the only if part of the statement. Since $\text{rint.}\mathcal{P}_r = \text{co.}\mathcal{Q}$, we have $\mathcal{P}_r = \text{cl.}\mathcal{P}_r = \text{cl.rint.}\mathcal{P}_r = \text{cl.co.}\mathcal{Q} = \text{co.cl.}\mathcal{Q}$, where the first equality holds because $\mathcal{P}_r$ is closed, the second equality holds by Theorem 6.3 of Rockafellar (2015), and the last equality holds because $\mathcal{Q}$ is bounded and by Theorem 17.2 of Rockafellar (2015). The rest of the proof goes through exactly the same way as in the only if part of Statement (1).

A.3.3 Statement (3)

Fix $\rho \in \mathcal{P}_l$. Let β the coefficient vector associated with ρ. Since $\rho \in \mathcal{P}_r$, there exists $\nu \in \Delta(\Pi)$ such that ν rationalizes ρ. Moreover, in the construction of ν in Block and Marschak (1960), they obtain that for any $\pi \in \Pi$, $\nu(\pi) = \prod_{n=1}^{|X|} \frac{\exp(\beta \cdot p_d(x_n))}{\sum_{l=n}^{|X|} \exp(\beta \cdot p_d(x_l))} > 0$, where $X = \{x_1, x_2, \ldots, x_{|X|}\}$ and $\pi(x_1) > \pi(x_2) > \cdots > \pi(x_{|X|})$. Since $\nu(\pi) > 0$ for all $\pi \in \Pi$, it follows from Theorem 6.9 in Rockafellar (2015) that $\rho \in \text{rint.co.}\{\rho^\pi | \pi \in \Pi\} = \text{rint.}\mathcal{P}_r$, where the last equality holds because $\mathcal{P}_r = \text{co.}\{\rho^\pi | \pi \in \Pi\}$.

A.4 Proof of Lemma 2

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

Proof. Assume that a ranking π is degree-d-representable. Without loss of generality, assume that $X = \{x_1, \ldots, x_{|X|}\}$ and $\pi(x_1) > \pi(x_2) > \cdots > \pi(x_{|X|})$. Then there exists β such that for any $D \in \mathcal{D}$ $\pi(x) > \pi(y)$ for all $y \in D \setminus \{x\}$ if and only if $\beta \cdot p_d(x) > \beta \cdot p_d(y)$ for all $y \in D \setminus \{x\}$. For any positive integer n and any $(D, x) \in \mathcal{D} \times X$ such that $x \in D$, let

$$\begin{aligned}\rho_{n\beta}(D, x) &\equiv \Pr(n\beta \cdot p_d(x) + \varepsilon_x \geq \max_{y \in D\setminus x}\{n\beta \cdot p_d(y) + \varepsilon_y\}) \\ &\geq \Pr(n\beta \cdot p_d(x) + \varepsilon_x \geq \max_{y \in D\setminus x} n\beta \cdot p_d(y) + \max_{y \in D\setminus x} \varepsilon_y) \\ &= \Pr(n(\beta \cdot p_d(x) - \max_{y \in D\setminus x} \beta \cdot p_d(y)) \geq \max_{y \in D\setminus x} \varepsilon_y - \varepsilon_x)\end{aligned}$$

Note $\max_{y \in D\setminus x} \varepsilon_y - \varepsilon_x$ is a well-defined random variable that does not take on infinite values, that is, $\Pr(|\max_{y \in D\setminus x} \varepsilon_y - \varepsilon_x| = \infty) = 0$. Since the distribution of the random variable does not depend on n, we have $n(\beta \cdot p_d(x) - \max_{y \in D\setminus x} \beta \cdot p_d(y)) \to \infty$ as $n \to \infty$, if $\pi(x) > \pi(D \setminus \{x\})$, then $\rho_{n\beta}(D, x) \to 1$. By taking a complement, if $\pi(x) < \pi(D \setminus \{x\})$, then $\rho_{n\beta}(D, x) \to 0$. Hence, $\rho_{n\beta} \to \rho^\pi$ as $n \to \infty$. □

Step 2: If there exists $\mu \in \mathcal{M}$ and a sequence $\{\rho_n\}$ of $\mathcal{P}_s(d, 0|\mu)$ such that $\rho_n \to \rho^\pi$, then π is degree-d-representable.

Proof. Let β_n be the coefficient vector of ρ_n. Consider a particular binary choice set $\{x, y\}$. Without loss of generality, assume $\pi(x) > \pi(y)$, for the binary choice set containing this two elements, we have $\Pr(\beta_n \cdot p_d(x) + \varepsilon(x) > \beta_n \cdot p_d(y) + \varepsilon(y)) \to 1$ as $n \to \infty$. This implies that there exists a $N_{x,y}$ such that for all $n \geq N_{x,y}$ we have $\beta_n \cdot p_d(x) > \beta_n \cdot p_d(y)$. To see this, suppose the previous claim is not true: we have a subsequence of n_k such that $\beta_{n_k} \cdot p_d(x) \leq \beta_{n_k} \cdot p_d(y)$. Then we must have $\Pr(\beta_{n_k} \cdot p_d(x) + \varepsilon(x) > \beta_{n_k} \cdot p_d(y) + \varepsilon(y)) = 1 - \Pr(\beta_{n_k} \cdot p_d(x) - \beta_{n_k} \cdot p_d(y) \leq \varepsilon(y) - \varepsilon(x)) \leq 1 - \Pr(0 < \varepsilon(y) - \varepsilon(x)) < 1$ uniformly in n_k. This contradicts with our hypothesis that $\rho_n \to \rho^\pi$. Finally, although $N_{x,y}$ depend on a particular binary choice sets, we have a finite number of binary choice sets. Taking the maximum of $N_{x,y}$ among all binary choice sets we have the desired result. □

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

Proof. Fix a positive integer d and $\pi \in \Pi$. Suppose that there exists a sequence ρ_n of $\text{co.}\mathcal{P}_s(d, 0|\mu)$ such that $\rho_n \to \rho^\pi$ as $n \to \infty$. Let $M = \dim \text{co.}\mathcal{P}_s(d, 0|\mu)$.

Then for each ρ_n, by Caratheodory's theorem, there exist $\{\rho_n^i\}_{i=1}^{M+1} \subset \mathcal{P}_s(d,0|\mu)$ 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 $\rho_{n'}^{i^*} \to \rho^\pi$ as $n' \to \infty$. To show the claim, we prove that if $\rho_{n'}^{i^*} \not\to \rho^\pi$, then $\alpha_{n'}^{i^*} \to 0$, which is a contradiction. Assume that $\rho_{n'}^{i^*} \not\to \rho^\pi$. Then there exist $D \in \mathcal{D}$, $x \in D$, and $\varepsilon > 0$ such that for any integer N there exists $n > N$ such that $|\rho_n^{i^*}(D,x) - \rho^\pi(D,x)| > \varepsilon$. This implies that for any N there exists $n > N$ such that $\left| \sum_{i=1}^{M+1} \alpha_{n'}^i \rho_{n'}^i(D,x) - \rho^\pi(D,x) \right| = \sum_{i=1}^{M+1} \alpha_{n'}^i |\rho_{n'}^i(D,x) - \rho^\pi(D,x)| \geq \alpha_{n'}^{i^*} \varepsilon$, where the first equality holds because if $\pi(x) \geq \pi(D)$ then $\rho_{n'}^i(D,x) - \rho^\pi(D,x) \leq 0$ for all i; if not $\pi(x) \geq \pi(D)$ then $\rho_{n'}^i(D,x) - \rho^\pi(D,x) \geq 0$ for all i. Since $\sum_{i=1}^{M+1} \alpha_{n'}^i \rho_{n'}^i(D,x) \to \rho^\pi(D,x)$, it must hold that $\alpha_{n'}^{i^*} \to 0$, which completes the proof of Step 3. □

Steps 2 and 3 show that if there exists a sequence $\{\rho_n\}$ of co.$\mathcal{P}_s(d,0|\mu)$ such that $\rho_n \to \rho^\pi$, then π is degree-d-representable. The contraposition of this statement is the second statement of Lemma 2.

A.5 Proof of Lemma 3

A.5.1 Proof of Statement (1)

We use the following lemma:

Lemma 6. *Let A be an $r \times n$ real matrix, B be an $l \times n$ real matrix, and E be an 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 $\pi \in \mathbf{R}^m$ 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. In the following by using Lemma 6, we prove statement (i).

For any ranking $\pi \in \Pi$ and a positive integer d, consider the following condition: if $\lambda_1 p_d(\pi^{-1}(|X|)) + \sum_{i=2}^{|X|-1}(\lambda_i - \lambda_{i-1})p_d(\pi^{-1}(|X|+1-i)) - \lambda_{|X|-1}p_d(\pi^{-1}(1)) = 0$ and $\lambda_i \geq 0$ for all $i \in \{1, \ldots, |X|-1\}$, then $\lambda_i = 0$ for all $i \in \{1, \ldots, |X|-1\}$. We call this Condition $(*)$.

Step 1: For each $\pi \in \Pi$ and a positive integer d, Condition $(*)$ holds if and only if π is degree-d-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 $x, y \in X$, $\pi(x) > \pi(y) \Leftrightarrow \beta \cdot x > \beta \cdot y$. Fix $\pi \in \Pi$.

$$
\begin{aligned}
&\exists \beta \ \left[\beta \cdot p_d(\pi^{-1}(|X|)) > \beta \cdot p_d(\pi^{-1}(|X|-1)) > \cdots > \beta \cdot p_d(\pi^{-1}(2)) > \beta \cdot p_d(\pi^{-1}(1))\right] \\
&\iff \exists \beta \ \left[\beta \cdot (p_d(\pi^{-1}(|X|)) - p_d(\pi^{-1}(|X|-1))) > 0, \ldots, \beta \cdot (p_d(\pi^{-1}(2)) - p_d(\pi^{-1}(1))) > 0\right] \\
&\iff \exists \beta \ [E\beta \gg 0] \\
&\iff \nexists \lambda \in \mathbf{R}^{|X|-1} \ [\lambda > 0, \ \lambda E = 0] \\
&\iff \nexists \lambda \in \mathbf{R}^{|X|-1} \ \left[\lambda > 0, \sum_{i=1}^{|X|-1} \lambda_i (p_d(\pi^{-1}(|X|+1-i)) - p_d(\pi^{-1}(|X|-i))) = 0\right] \\
&\iff \nexists \lambda \in \mathbf{R}^{|X|-1} \ \big[\lambda > 0, \\
&\quad \lambda_1 p_d(\pi^{-1}(|X|)) + \textstyle\sum_{i=2}^{|X|-1}(\lambda_i - \lambda_{i-1})p_d(\pi^{-1}(|X|+1-i)) - \lambda_{|X|-1}p_d(\pi^{-1}(1)) = 0\big] \\
&\iff \text{Condition}(*),
\end{aligned}
$$

where $\lambda \equiv (\lambda_1, ..., \lambda_{|X|-1})$ and the third equivalence is obtained by using Lemma 6 with $A, B = 0$ and $E^\top \equiv (p_d(\pi^{-1}(|X|)) - p_d(\pi^{-1}(|X|-1)), \ldots, p_d(\pi^{-1}(2)) - p_d(\pi^{-1}(1)))$. □

Step 2: For a given positive integer d, the set $\{p_d(x) | x \in X\}$ is affinely independent if and only if Condition $(*)$ holds for d and any $\pi \in \Pi$.

Proof. We first show that the only if part. Fix any $\pi \in \Pi$. Without loss of generality assume that $\pi(x_i) = |X| + 1 - i$ for all $i \in \{1, \ldots, |X|\}$. Suppose that $\lambda_1 p_d(\pi^{-1}(|X|)) + \sum_{i=2}^{|X|-1}(\lambda_i - \lambda_{i-1})p_d(\pi^{-1}(|X|+1-i)) - \lambda_{|X|-1}p_d(\pi^{-1}(1)) = 0$ and $\lambda_i \geq 0$ for all i. Then, $\lambda_1 p_d(x_1) + \sum_{i=2}^{|X|-1}(\lambda_i - \lambda_{i-1})p_d(x_i) - \lambda_{|X|-1}p_d(x_{|X|}) = 0$. Define $\mu_1 = \lambda_1$, $\mu_i = \lambda_i - \lambda_{i-1}$ for all $i \in \{2, \ldots, |X|-1\}$, and $\mu_{|X|} = -\lambda_{|X|-1}$. Then $\sum_{i=1}^{|X|} \mu_i p_d(x_i) = 0$. Moreover, $\sum_{i=1}^{|X|} \mu_i = \lambda_1 + \sum_{i=2}^{|X|-1}(\lambda_i - \lambda_{i-1}) - \lambda_{|X|-1} = 0$. If $\{p_d(x) | x \in X\}$ is affinely independent, then $\mu_i = 0$ for all $i \in \{1, \ldots, |X|\}$. Hence, $\lambda_i = 0$ for all $i \in \{1, \ldots, |X|-1\}$. This implies Condition $(*)$.

Next we will show the if part. Choose any real numbers $\{\mu_i\}_{i=1}^{|X|}$ such that $\sum_{i=1}^{|X|} \mu_i p_d(x_i) = 0$ and $\sum_{i=1}^{|X|} \mu_i = 0$ to show $\mu_i = 0$ for all $i \in \{1, \ldots, |X|\}$. Without loss of generality, order μ_i by its value so that $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_{|X|}$. Let $\mu \equiv (\mu_1, ..., \mu_{|X|})$. For each $x_i \in X$, define $\pi(x_i) = |X| + 1 - i$. Then $\pi \in \Pi$. Define

$\lambda_1 = \mu_1$ and $\lambda_i = \sum_{j=1}^{i} \mu_j$ for all $i \in \{2, \ldots, |X| - 1\}$.

First we will show that $\lambda_i \geq 0$ for all $i \in \{1, \ldots, |X| - 1\}$. Suppose by way of contradiction that $\lambda_i < 0$ for some $i \in \{1, \ldots, |X| - 1\}$. Then $\mu_i < 0$ because $\mu_1 \geq \cdots \geq \mu_i$. Since $0 > \mu_i \geq \mu_j$ for all $j \geq i$, we have $\sum_{j=i+1}^{|X|} \mu_j < 0$. It follows that $\sum_{j=1}^{|X|} \mu_j = \lambda_i + \sum_{j=i+1}^{|X|} \mu_j < 0$. This contradicts that $\sum_{j=1}^{|X|} \mu_j = 0$. Therefore, $\lambda_i \geq 0$ for all $i \in \{1, \ldots, |X| - 1\}$.

In the following, we will show $\mu = 0$ by using $\lambda_i \geq 0$ for all $i \in \{1, \ldots, |X| - 1\}$. Notice that $\lambda_1 p_d(\pi^{-1}(|X|)) + \sum_{i=2}^{|X|-1} (\lambda_i - \lambda_{i-1}) p_d(\pi^{-1}(|X|+1-i)) - \lambda_{|X|-1} p_d(\pi^{-1}(1)) = \lambda_1 p_d(x_1) + \sum_{i=2}^{|X|-1} (\lambda_i - \lambda_{i-1}) p_d(x_i) - \lambda_{|X|-1} p_d(x_{|X|}) = \mu_1 p_d(x_1) + \sum_{i=2}^{|X|-1} \mu_i p_d(x_i) - \sum_{i=1}^{|X|-1} \mu_i p_d(x_{|X|}) = \sum_{i=1}^{|X|} \mu_i p_d(x_i) = 0$,where the second to the last equality holds because $\sum_{i=1}^{|X|} \mu_i = 0$. Therefore, by Condition $(*)$, $\lambda_i = 0$ for all $i \in \{1, \ldots, |X| - 1\}$. This implies $\mu = 0$. $\square$

A.5.2 Proof of Statement (2)

For any $x \in X$, $p_d(x) \notin \text{co.}(\{p_d(y) | y \in X \setminus \{x\}\}) \Leftrightarrow p_d(x)$ is an extreme point of $\text{co.}(\{p_d(y) | y \in X\}) \Leftrightarrow p_d(x)$ is an exposed point of $\text{co.}(\{p_d(y) | y \in X\}) \Leftrightarrow \exists \beta \; \forall y \in X \setminus \{x\} [\beta \cdot p_d(x) > \beta \cdot p_d(y)] \Leftrightarrow$ all raking π, which best alternative is x, is degree-d-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 $\text{co.}(\{p_d(y) | y \in X\})$ is a polytope.

A.6 Proof of Lemma 5

Let $|X| = n$ and write $X = \{x_1, \ldots, x_n\}$. Let π_n be a ranking over X_n such that $\pi_n(x_i) > \pi_n(x_{i+1})$ for any $i \leq n - 1$. We will prove that ρ^{π_n} and $\rho^{\pi_n^-}$ are adjacent. In particular, we will find $t_n \in \mathbf{R}^{\mathcal{D}_n \times X_n}$ such that $\rho^{\pi_n} \cdot t_n = \rho^{\pi_n^-} \cdot t_n = 0$, and $\rho^{\sigma_n} \cdot t_n > 0$ for any $\sigma_n \in \Pi_n \setminus \{\pi_n, \pi_n^-\}$. The proof is by induction on n.

Induction Base: Let us consider the case of $n = 3$. For $b > a > 0$, let $t_3(\{x_1, x_2\}, x_1) = a$, $t_3(\{x_2, x_3\}, x_2) = -b$, $t_3(\{x_1, x_3\}, x_1) = b - a$, and $t_3(\{x_1, x_2, x_3\}, x_2) = a + b$. For all other $(D, x) \in \mathcal{D} \times X$, $t_3(D, x) = 0$. Then t_3 is defined on $\mathcal{D} \times X$ and satisfies the conditions. By a direct calculation, it can be shown that $\rho^{\pi_3} \cdot t_3 = \rho^{\pi_3^-} \cdot t_3 = 0$, and $\rho^{\pi} \cdot t_3 > 0$ for any $\sigma_3 \in \Pi_3 \setminus \{\pi_3, \pi_3^-\}$.

Assume that $n \geq 4$. For each i such that $3 \leq i \leq n$, define $X_i = \{x_1, x_2, \ldots, x_i\}$ and let Π_i be the set of rankings over X_i. For each i such that $3 \leq i \leq n - 1$, let

$\mathcal{D}_i \subset 2^{X_i} \setminus \emptyset$ such that $\mathcal{D}_i$ is rich and $\mathcal{D}_i \subset \mathcal{D}_{i+1}$ and $\mathcal{D}_n = \mathcal{D}$.[38]

Induction Hypothesis: Let π_{n-1} be the ranking over X_{n-1} such that $\pi_{n-1}(x_i) > \pi_{n-1}(x_{i+1})$ for any $i \leq n-2$. Suppose that there exists $t_{n-1} \in \mathbf{R}^{\mathcal{D}_{n-1} \times X_{n-1}}$ 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 $t_n \in \mathbf{R}^{\mathcal{D}_n \times X_n}$ as follows: For each $(D, x) \in \mathcal{D}_n \times X_n$

$$t_n(D,x) = \begin{cases} t_{n-1}(D,x) & \text{if } (D,x) \in (\mathcal{D}_{n-1} \times X_{n-1}) \setminus \{(\{x_1, x_2\}, x_1)\}, \\ t_{n-1}(D,x) + \varepsilon_{n-1} & \text{if } (D,x) = (\{x_1, x_2\}, x_1), \\ -\varepsilon_{n-1}/(n-1) & \text{if } (D,x) = (\{x_i, x_n\}, x_i) \text{ for some } i \in \{1, \cdots, n-1\}, \\ 2\varepsilon_{n-1} & \text{if } (D,x) = (\{x_{n-2}, x_{n-1}, x_n\}, x_{n-1}), \\ 0 & \text{otherwise.} \end{cases}$$

It is clear that $\rho^{\pi_n} \cdot t_n = \rho^{\pi^-_n} \cdot t_n = 0$. Let $\sigma_n \in \Pi_n \setminus \{\pi_n, \pi^-_n\}$. Let $j \in \{1, \cdots, n\}$ be such that the element x_n be jth best element in σ_n. There exists $\sigma_{n-1} \in \Pi_{n-1}$ such that the ranking σ_n can be written as $(\sigma^{-1}_{n-1}(n-1), \cdots, \sigma^{-1}_{n-1}(n-(j-1)), x_n, \sigma^{-1}_{n-1}(n-j), \cdots, \sigma^{-1}_{n-1}(1))$ in decreasing order of the ranking if $2 \leq j \leq n-1$.[39]

First notice that by the definition of t_n and $\rho^{\sigma_{n-1}} = \rho^{\sigma_n}$ on $\{x_1, x_2\}$, $\rho^{\sigma_n} \cdot t_n = \rho^{\sigma_{n-1}} \cdot t_{n-1} + \varepsilon_{n-1}\rho^{\sigma_{n-1}}(\{x_1, x_2\}, x_1) - \dfrac{\varepsilon_{n-1}}{n-1}(j-1) + 2\varepsilon_{n-1}\rho^{\sigma_n}(\{x_{n-2}, x_{n-1}, x_n\}, x_{n-1})$, where the second term of the right hand side is $(\varepsilon_{n-1}/(n-1))(j-1)$ since in σ_n, there are $j-1$ elements that are better than n.

Case 1: $\sigma_{n-1} = \pi_{n-1}$. Note that $\rho^{\sigma_{n-1}}(\{x_1, x_2\}, x_1) = \rho^{\pi_{n-1}}(\{x_1, x_2\}, x_1) = 1$.[40] Also $\rho^{\sigma_n}(\{x_{n-2}, x_{n-1}, x_n\}, x_{n-1}) = 0$ since ρ^{σ_n} coincide with $\rho^{\sigma_{n-1}} = \rho^{\pi_{n-1}}$ on X_{n-1} and x_{n-2} is better than x_{n-1} in the ranking π_{n-1}. Thus, $\rho^{\sigma_n} \cdot t_n = 0 + \varepsilon_{n-1} - \dfrac{\varepsilon_{n-1}}{n-1}(j-1) + 0 > 0$, where the last inequality holds because $j < n$. (If $j = n$, then $\sigma_{n-1} = \pi_{n-1}$ implies that $\sigma_n = \pi_n$, which is a contradiction.)

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 $\rho^{\sigma_{n-1}}(\{x_1, x_2\}, x_1) = \rho^{\pi^-_{n-1}}(\{x_1, x_2\}, x_1) = 0$. Note also that $\rho^{\sigma_n}(\{x_{n-2}, x_{n-1}, x_n\}, x_{n-1}) = 1$ since (i) $\rho^{\sigma_n}(\{x_{n-2}, x_{n-1}, x_n\}, x_{n-2}) = 0$; (ii) $\rho^{\sigma_n}(\{x_{n-2}, x_{n-1}, x_n\}, x_n) = 0$. (i) holds because $\sigma_{n-1} = \pi^-_{n-1}$ and (ii) holds because $\sigma_n \neq \pi^-_n$ and $\sigma_{n-1} = \pi^-_{n-1}$. Thus, $\rho^{\sigma_n} \cdot t_n = 0 + 0 - \dfrac{\varepsilon_{n-1}}{n-1}(j-1) + 2\varepsilon_{n-1} > 0$.

[38] Remember that the richness of $\mathcal{D}_i$ means that $\{x, y\} \in \mathcal{D}_i$ and $\{x, y, z\} \in \mathcal{D}_i$ for any $x, y, z \in X_i$

[39] If $j = 1$, then x_n is the best element in σ_n. If $j = n$, then x_n is the worst element in σ_n.

[40] Remember that $\pi_{n-1}(x_i) > \pi_{n-1}(x_{i+1})$ for any $i \leq n-1$

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 $j \leq n$.

A.7 Proof of Lemma 4

To prove the lemma, we prove following lemmas. Fix a ranking π that is not degree-d-representable. For any $\alpha \in (0,1)$, define $\rho_\alpha^\pi \equiv \alpha\rho^\pi + (1-\alpha)\rho^{\pi^-}$. We first will show statement (a) mentioned after Lemma 4 in Section 3.1.

Lemma 7. *Let $\mu \in \mathcal{M}$. For any $\alpha \in (0,1)$, $\rho_\alpha^\pi \notin cl. \bigcup_\eta \mathcal{P}_s(d, \eta|\mu)$.*

Proof. Choose any $x, y, z \in X$ such that $\pi(x) > \pi(y) > \pi(z)$. Consider ρ_α^π. Suppose by way of contradiction that $\rho_\alpha^\pi \in \text{cl.} \bigcup_\eta \mathcal{P}_s(\eta|\mu)$. This implies there exists $\rho_n \in \bigcup_\eta \mathcal{P}_s(d, \eta|\mu)$ such that $\rho_n \to \rho_\alpha^\pi$. Let β_n be corresponding to ρ_n. Consider the sequence $\{\beta_n \cdot (p_d(x) - p_d(y))\}$.

Step 1: There exists a convergent subsequence $\{\beta_{n'} \cdot (p_d(x) - p_d(y))\}$.

Proof. To see this notice the indices of the original sequence $\beta_n \cdot (p_d(x) - p_d(y))$ must be bounded. If not, for each large number N, there exists $|\beta_n \cdot (p_d(x) - p_d(y))| > N$ for some n. This implies $\rho_n(\{x,y\}, x)$ will be close to either 0 or 1 infinitely often, violating the convergence assumption. Since $\beta_n \cdot (p_d(x) - p_d(y))$ lies in a bounded set, we can extract a convergent subsequence $\{\beta_{n'} \cdot (p_d(x) - p_d(y))\}$. □

Given Step 1, we fix one convergent subsequence $\{\beta_{n'} \cdot (p_d(x) - p_d(y))\}$. We consider corresponding stochastic choice functions $\rho_{n'}$. Note that, by definition, $\lim_{n'} \rho_{n'} = \rho_\alpha^\pi$. To make our notation simple, in the following, we write ρ_n and β_n instead of $\rho_{n'}$ and β_n.

For any $s, t \in \{x, y, z\}$ and $n \in N$, define $E_{n,st} = \{\varepsilon | \beta_n \cdot p_d(s) + \eta(s) + \varepsilon(s) > \beta_n \cdot p_d(t) + \eta(t) + \varepsilon(t)\}$ and $E_{st} = \{\varepsilon | \lim_n \beta_n \cdot p_d(s) + \eta(s) + \varepsilon(s) \geq \lim_n \beta_n \cdot p_d(t) + \eta(t) + \varepsilon(t)\}$. (In short, E_{st} is the event s is preferred to t.)

Step 2: $E_{xy} = E_{xz}$ and $E_{zy} = E_{zx}$ up to a measure zero set.

Proof. By Fatou's lemma $\alpha = \rho_\alpha^\pi(\{x,y,z\}, x) = \limsup \rho_n(\{x,y,z\}, x) = \limsup \mu(E_{n,xy} \cap E_{n,xz}) \leq \mu(\limsup(E_{n,xy} \cap E_{n,xz})) = \mu(E_{xy} \cap E_{xz})$. Moreover, looking at the binary choice sets $\{x,y\}$, $\alpha = \rho_\alpha^\pi(\{x,y\}, x) = \liminf \mu(E_{n,xy}) \geq \mu(\liminf E_{n,xy}) = \mu(E_{xy})$, where the inequality holds by Fatou's lemma and the last equality holds because $\mu\{\varepsilon | \beta_n \cdot p_d(x) + \eta(x) + \varepsilon(x) = \beta_n \cdot p_d(y) + \eta(y) + \varepsilon(y)\} = 0$. (This holds because μ is absolutely continuity with respect to the Lebesgue measure since its

density exists.) Thus we have the identity $\mu(E_{xy} \cap E_{xz}) \geq \alpha \geq \mu(E_{xy})$. Thus $E_{xy} \subset E_{xz}$ up to a measure zero set. By symmetry, we have the opposite inclusion and by combining them we obtain $E_{xy} = E_{xz}$ up to a measure zero set.

In the same way, we obtain $E_{zy} = E_{zx}$ up to a measure zero set. □

Define two events $A = \left\{(\varepsilon(x), \varepsilon(y), \varepsilon(z)) \equiv \varepsilon \in \mathbf{R}^3 \,\middle|\, U\varepsilon^T \geq c\right\}$ as well as $B = \left\{(\varepsilon(x), \varepsilon(y), \varepsilon(z)) \equiv \varepsilon \in \mathbf{R}^3 \,\middle|\, U\varepsilon^T \leq c\right\}$, where

$$U = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}, \quad c = \begin{bmatrix} \lim \beta_n \cdot (p_d(y) - p_d(x)) + \eta(y) - \eta(x) \\ \lim \beta_n \cdot (p_d(z) - p_d(y)) + \eta(z) - \eta(y) \end{bmatrix}. \tag{7}$$

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

Proof. Notice E_{xy} is the event x is chosen over y in the binary set $\{x, y\}$. Step 2 implies the event $E_{xy} \cap E_{xz}$ has measure α.

Similarly, we have the event $E_{zy} \cap E_{zx}$ has probability measure $1 - \alpha$. Notice $E_{xy} \cap E_{xz}$ and $E_{zy} \cap E_{zx}$ have measure zero intersections, so the two events partition the probability space (ignoring measure zero events). The event $E_{xy} \cap E_{xz}$ has probability α and the event $E_{zy} \cap E_{zx}$ has probability $1 - \alpha$, so the two events partition the probability space (ignoring measure zero events).

Now notice that $\alpha = \rho^{\pi}_{\alpha}(\{y, z\}, y) = \mu(E_{yz})$. Since the event E_{yz} is incompatible with the event $E_{zy} \cap E_{zx}$ it must completely lie within the event $E_{xy} \cap E_{xz}$. Moreover, since the probability of the event E_{yz} is α. Thus, the event E_{yz} coincides with event in $E_{xy} \cap E_{xz}$ (ignoring measure zero events). Finally notice the event A is the intersection of $E_{xy} \cap E_{xz}$ and E_{yz}, up to a measure zero events.

Similarly notice that $1 - \alpha = \rho^{\pi}_{\alpha}(\{x, y\}, y) = \mu(E_{yx})$. Since the event E_{yx} is not compatible with the event $E_{xy} \cap E_{xz}$, it must completely coincide $E_{zy} \cap E_{zx}$. Finally notice the event B is the intersection of $E_{zy} \cap E_{zx}$ and E_{yx}, up to a measure zero events. □

Fix some ε_a and ε_b in events A and B and in the support. Define $a \equiv U\varepsilon_a^T$ and $b \equiv U\varepsilon_b^T$. Without loss of generality, we suppose there does not exist $t \in \mathbb{R}$ such that $(a - c) = t(b - c)$, where c is defined as in (7). This can be achieved by the open support condition.[41] For any $\lambda \in (0, 1)$, define $\varepsilon_\lambda \equiv \lambda\varepsilon_a + (1 - \lambda)\varepsilon_b$.

Step 4: There exists some $\lambda^* \in (0, 1)$ such that $\varepsilon_{\lambda^*} \in (A \cup B)^c$.

Proof. First notice that for any $\lambda \in (0, 1)$, $U\varepsilon_\lambda^T \neq c$. This is because if $U\varepsilon_\lambda^T = c$ for

[41] Notice first U has full column rank. If such t exists, we can perturb the points ε_a and ε_b so that such t does not exist.

some λ, then $\lambda a + (1-\lambda)b = c$ implies $\lambda(a-c) = -(1-\lambda)(b-c)$, which contradicts the non existence of t above.

Now to show Step 4, it suffices to show that there exists some $\lambda^* \in (0,1)$ such that $\varepsilon_{\lambda^*} \notin A' \cup B'$, where $A' = \left\{\varepsilon \in \mathbf{R}^3 \left| U\varepsilon^T > c \right.\right\}$ and $B' = \left\{\varepsilon \in \mathbf{R}^3 \left| U\varepsilon^T < c \right.\right\}$, where c is defined as in (7). Note that the set $\{U\varepsilon_\lambda^T\}_{\lambda\in(0,1)}$ is connected, and A' and B' are disjoint open sets. If $\{U\varepsilon_\lambda^T\}_{\lambda\in(0,1)} \subset A' \cup B'$ is true, this implies $\{U\varepsilon_\lambda^T\}_{\lambda\in(0,1)}$ is not connected. □

Since the support is convex, ε_{λ^*} belongs to the support. Moreover, the support is open and $(A\cup B)^c$ is open, it follows from Step 4 that there exists a non-negligible mass around ε_{λ^*} outside $A \cup B$. This contradicts with Step 3, which implies the probabilities of $A \cup B$ is one. ■

Given Lemma 7, in order to show Lemma 4, it suffices to show that even with using mixtures, it is impossible to approximate ρ_α^π. For this purpose, we need two more lemma.

Lemma 8. *(i)For any $\rho \in cl.\bigcup_\eta \mathcal{P}_s(d,\eta|\mu)$, if $\rho \notin \{\rho^\pi, \rho^{\pi^-}\}$ then $\rho \notin \{\rho_\alpha^\pi | \alpha \in [0,1]\}$; (ii) Let (t,a) be as in Definition 7 with a pair $(\rho^\pi, \rho^{\pi-})$ of adjacent rankings. For any $\rho \in cl.\bigcup_\eta \mathcal{P}_s(d,\eta|\mu)$, if $\rho \notin \{\rho^\pi, \rho^{\pi^-}\}$ then $\rho \cdot t > a$.*

Proof. To show (i), suppose by way of contradiction that $\rho \in \{\rho_\alpha^\pi | \alpha \in [0,1]\}$. Since $\rho \notin \{\rho^\pi, \rho^{\pi^-}\}$, $\rho = \rho_\alpha^\pi$ for some $\alpha \in (0,1)$. By Lemma 7, $\rho \notin \text{cl.}\bigcup_\eta \mathcal{P}_s(d,\eta|\mu)$, which is a contradiction.

Now we will show (ii) by using (i). Since $\rho \in \mathcal{P}_r$, it can be written as a convex combination of ρ^πs: $\rho = \sum_{\sigma\in\Pi} \mu(\sigma)\rho^\sigma = \mu(\pi)\rho^\pi + \mu(\pi^-)\rho^{\pi^-} + \sum_{\sigma\in\Pi\setminus\{\pi,\pi^-\}} \mu(\sigma)\rho^\sigma$. By (i), $\rho \in \text{cl.}\bigcup_\eta \mathcal{P}_s(d,\eta|\mu)$ and $\rho \notin \{\rho^\pi, \rho^{\pi^-}\}$ implies that $\rho \notin \{\rho_\alpha^\pi | \alpha \in [0,1]\}$. Thus, we must have one of the $\mu(\sigma)$ in the third term positive. Moreover, by definition and the fact that ρ^π and ρ^{π^-} are adjacent, for any $\sigma \notin \{\pi, \pi^-\}$, $\rho^\sigma \cdot t > a = \rho^\pi \cdot t = \rho^{\pi^-} \cdot t$. Thus we can conclude that $\rho \cdot t > a$. ■

Lemma 9. *Let $\mu \in \mathcal{M}$. If there exists a sequence of degree-d random-coefficient utility-shock models with fixed effects η_n converging to ρ_α^π for some $\alpha \in (0,1)$, then there exists two sequences $\rho_{(\eta_n,\beta_n)}$ and $\rho_{(\eta_n,\beta_n')}$ of degree-d utility-shock models with fixed effects η_n that converges to ρ^π and ρ^{π^-}, respectively.*

Proof. Since $\bigcup_\eta \mathcal{P}_{ra}(d,\eta|\mu) = \bigcup_\eta co.\mathcal{P}_a(d,\eta|\mu)$, there exists $\sum_{i=1}^{M+1} \mu_n(i)\rho_{(\beta_n(i),\eta_n)} \to \rho_\alpha^\pi$, where $M = dim\mathcal{P}_r + 1$ (allowing $\mu_n(i) = 0$ for some i). For all i, we first extract converging subsequences $\mu_m(i)$ and $\rho_{(\beta_m(i),\eta_m)}$ of $\mu_n(i)$ and $\rho_{(\beta_n(i),\eta_n)}$, respectively.

We can do this sequentially. Notice that for each i, $\mu_m(i)$ is a bounded sequence in a compact set: thus, it has a convergent subsequence. We denote the limit by $\mu^*(i)$. Similarly, $\rho_{(\beta_n(i),\eta_n)}$ belongs to a compact set of random utility models:thus it has a convergent subsequence. We denote the limit $\rho^*(i)$. A diagonal argument gives us desirable subsequences such that for all i, $\mu_m(i) \to \mu^*(i)$ and $\rho_{(\beta_m(i),\eta_m)} \to \rho^*(i)$ as $m \to \infty$. Thus, $\sum_i \mu_n(i)\rho_{(\beta_n(i),\eta_n)} \to \sum_i \mu^*(i)\rho^*(i) = \rho^\pi_\alpha$. Moreover, since $\rho_{(\beta_m(i),\eta_m)} \in \bigcup_\eta \mathcal{P}_s(d,\eta|\mu)$, we have $\rho^*(i) \in \text{cl.} \bigcup_\eta \mathcal{P}_s(d,\eta|\mu)$.

In the following, we will argue that there exists some i, j such that $\rho^*(i) = \rho^\pi$ and $\rho^*(j) = \rho^{\pi^-}$. By way of contradiction and without loss of generality, suppose that $\rho^*(i) \neq \rho^\pi$ for any i.[42]

Let (t, a) be as in Definition 7 with a pair (ρ^π, ρ^{π^-}) of adjacent rankings.

We will consider two cases.

Case 1: $\rho^*(i) \neq \rho^{\pi^-}$ for any i. For all i, $\rho^*(i) \in \text{cl.} \bigcup_\eta \mathcal{P}_s(d,\eta|\mu)$ and $\rho^*(i) \notin \{\rho^\pi, \rho^{\pi-}\}$. Then, by Lemma 8 (ii), $\rho^*(i) \cdot t > a$ for all i. Thus, $\left(\sum_i \mu^*(i)\rho^*(i)\right) \cdot t = \sum_i \mu^*(i)\rho^*(i) \cdot t > a$. On the other hand by Definition 7, $\rho^\pi_\alpha \cdot t = a$. This is a contradiction to $\sum_i \mu^*(i)\rho^*(i) = \rho^\pi_\alpha$.

Case 2: $\rho^*(i) = \rho^{\pi^-}$ for some i. Define $J = \{i \in \{1, \ldots, M+1\} | \rho^*(i) = \rho^{\pi^-}\}$. First notice that there exists $i \in \{1, \ldots, M+1\} \setminus J$ such that $\mu^*(i) > 0$. (If such i does not exist, then $\rho^{\pi^-} = \sum_i \mu^*(i)\rho^*(i) = \rho^\pi_\alpha$, which contradicts with $\alpha \notin \{0, 1\}$.) Then, $a = \rho^\pi_\alpha \cdot t = \sum_i \mu^*(i)\rho^*(i) \cdot t = \sum_{i \in J} \mu^*(i)\rho^{\pi^-} \cdot t + \sum_{i \notin J} \mu^*(i)\rho^*(i) \cdot t > a$, where the last inequality holds since as in Case 1, by Lemma 8 (ii), $\rho^{\pi^-} \cdot t = a$ and $\rho^*(i) \cdot t > a$ for all $i \notin J$. ■

A.7.1 Main Proof of Lemma 4 by Using Lemma 7, 8, 9

As mentioned, given Lemma 7, it suffices to show that even with using mixtures, it is impossible to approximate ρ^π_α.

Let π be a raking that is not degree-d-representable. By Lemma 5, ρ^π and ρ^{π^-} are adjacent. Now suppose by way of contradiction that there exists a sequence of degree-d mixed logit models with fixed effects that approximates ρ^π_α for some $\alpha \in (0, 1)$. Then by Lemma 9, there exist sequences $\{(\eta_n, \beta_n, \beta'_n)\}$ such that (i) $\rho_{(\eta_n,\beta_n)} \to \rho^\pi$ and (ii) $\rho_{(\eta_n,\beta'_n)} \to \rho^{\pi^-}$. Given statement (i), by exactly the same way as Step 2 of Lemma 2, we can prove that there exists large N_1 such that for any $n \geq N_1$, we have $\beta_n \cdot p_d(x) + \eta_x > \beta_n \cdot p_d(y) + \eta_y$ for any $x, y \in X$ such that

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

$\pi(x) > \pi(y)$. Similarly from statement (ii), there exists large N_2 such that for any $n \geq N_2$, we have $\beta'_n \cdot p_d(x) + \eta_x > \beta'_n \cdot p_d(y) + \eta_y$ for any $x, y \in X$ such that $\pi^-(x) > \pi^-(y)$.

Now for any $x, y \in X$ such that $\pi(x) > \pi(y)$. Then for any $n_{xy} \geq \max\{N_1, N_2\}$, we have $\beta_n \cdot p_d(x) + \eta_x > \beta_n \cdot p_d(y) + \eta_y$. Since $\pi^-(y) > \pi^-(x)$, we have $-\beta_n \cdot p_d(x) - \eta(x) > -\beta_n \cdot p_d(y) - \eta(y)$. Summing the two inequalities, we have $(\beta_n - \beta'_n) \cdot p_d(x) > (\beta_n - \beta'_n) \cdot p_d(y)$. Because the number of binary choice sets is finite, we can find $n^* > n_{x,y}$ for any $x, y \in X$. We shows that if $\pi(x) > \pi(y)$ implies $(\beta_{n^*} - \beta'_{n^*}) \cdot p_d(x) > (\beta_{n^*} - \beta'_{n^*}) \cdot p_d(y)$. This contradicts with the fact that π not degree-d-representable.

A.8 Proof of Proposition 3

To prove the proposition, we will prove the following general claim. The claim is trivial when the set C is closed. Proposition 3 follows from the claim with $C = \mathcal{P}_s$, where $\mathcal{P}_s$ may not be closed. (For example, when $\mathcal{P}_s = \mathcal{P}_l$)

Claim: For any set bounded $C \subset \mathbf{R}^k$, let $\Delta(C)$ denote the set of Borel probability measures over C. Then, $\text{co}.C = \left\{ \int x dm(x) | m \in \Delta(C) \right\}$, where $\int x dm(x)$ denotes k-dimensional vector whose l-th element is $\int x(k) dm(x)$ for any $l \in \{1, \ldots, k\}$.

Proof. By definition, we immediately obtain $\text{co}.C \subset \{\int x dm(x) | m \in \Delta(C)\}$. In the following, we will show the statement $(*)$: $\left\{ \int x dm(x) | m \in \Delta(C) \right\} \subset \text{co}.C$. First we will show the satement $(**)$: $\left\{ \int x dm(x) | m \in \Delta(C) \right\} \subset \text{cl.co}.C$.

To prove this statement, suppose by way of contradiction that $\int x dm(x) \notin \text{cl.co}.C$ for some $m \in \Delta(C)$. By the strict separating hyperplane theorem (Corollary 11.4.2 of Rockafellar (2015)), there exist $t \in \mathbf{R}^k \setminus \{0\}$ 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 will show $(*)$ by the induction on the dimension of co.C.

Induction Base: If $\dim \text{co}.C = 1$, then there must exist y and z such that co.C is the line segment between y and z. In the following, we assume that the line segment does not contain both y and z but the proof for the other cases are similar. Then for any $x \in C$, there exists unique $\alpha(x) \in (0, 1)$ such that $x = \alpha(x)y + (1 - \alpha(x))z$. Notice that the function α is continuous in x and hence measurable. Moreover, the function α is integrable because α is bounded and nonnegative. Choose any $m \in \Delta(C)$, then $\int \alpha dm = \int \alpha(x) dm(x)$ exists.

Moreover, since $0 < \alpha(x) < 1$, it follows from the monotonicity of integral that $0 < \int \alpha(x)dm(x) < 1$. Denote the value of the integral by $\beta \in (0,1)$. Then, $\int xdm(x) = \int \alpha(x)y + (1-\alpha(x))zdm(x) = \beta y + (1-\beta)z \in \text{co.}C$, as desired.

Choose any integer $l \geq 3$.

Induction Hypothesis: Now suppose that $(*)$ holds for any C such that $\dim C \leq l$.

Induction Step: For any C such that $\dim C = l+1$, $(*)$ holds. To prove the step, choose any $m \in \Delta(C)$. By $(**)$, we have $\int xdm(x) \in \text{cl.co.}C$.

First consider the case where $\int xdm(x) \in \text{rint.cl.co.}C$. Since $\text{rint.cl.co.}C = \text{rint.co.}C$ (by Theorem 6.3 of Rockafellar (2015)), we have $\int xdm(x) \in \text{co.}C$, as desired.

Next consider the case where $\int xdm(x) \notin \text{rint.cl.co.}C$. Then, $\int xdm(x) \in \partial\text{cl.co.}C \equiv \text{cl.co.}C \setminus \text{rint.co.}C$. By the strict separating hyperplane theorem (Corollary 11.4.2 of Rockafellar (2015)), there exists a supporting hyperplane H of $\text{cl.co.}C$ at $\int xdm(x)$. There exist $t \in \mathbf{R}^k \setminus \{0\}$ and $\alpha \in \mathbf{R}$ such that $H = \{x | x \cdot t = \alpha\}$ and $\int xdm(x) \cdot t = \alpha > x \cdot t$ for any $x \in \text{cl.co.}C \cap H^c$. This implies that $m(H) = 1$. 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, $\dim(H \cap C) \leq l$. Therefore, the induction hypothesis shows that $\int xdm(x) \in \text{co.}(H \cap C) \subset \text{co.}C$, as desired. □

The claim above implies Proposition 3. The result is not true in an infinite dimensional space.[43]

A.9 Proof of Proposition 4

To prove Proposition 4, we prove one lemma.

Lemma 10. *For any $t \in \mathbf{R}^{\mathcal{D}\times X}$, $\rho^\pi \cdot t = \rho^{\pi'} \cdot t$ for all $\pi, \pi' \in \Pi$ if and only if $t(D,x) = t(D,y)$ for all $D \in \mathcal{D}$ and $x, y \in D$.*[44]

Proof. For notational convenience, for any $\pi \in \Pi$ and $D \in \mathcal{D}$ with $D = \{x_1, \ldots, x_{|D|}\}$, we write $\rho^\pi(D) = \big(\rho^\pi(D,x_1), \ldots, \rho^\pi(D,x_{|D|})\big)$. To prove the if part, assume $t(D,x) = t(D,y)$ for all $D \in \mathcal{D}$ and $x, y \in D$. Define $t(D) = t(D,x)$ for any $x \in D$. Then for

[43] 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. $\int xdm$ cannot be represented as any finite mixture of elements of C. For any $y \in \text{co.}C$, there exists i such that $y(e_i) = 0$.

[44] We are identifying each $\rho \in \mathcal{P}$ as an element of $\mathbf{R}^{\mathcal{D}\times X}$.

any $\pi \in \Pi$, $\rho^\pi \cdot t = \sum_{D\in\mathcal{D}} \sum_{x\in D} \rho^\pi(D,x)t(D,x) = \sum_{D\in\mathcal{D}} t(D) \sum_{x\in D} \rho^\pi(D,x) = \sum_{D\in\mathcal{D}} t(D)$, completing the proof of the if part.

The only if part is obvious for any D such that $|D| = 1$. Consider any D such that $|D| \geq 2$. Let l be the maximal integer such that $|D| \geq l+1$ for any $D \in \mathcal{D}$. Then $l \geq 1$.[45]

Claim: For any $D \in \mathcal{D}$ such that $|D| = l+1$ and any $x, y \in D$, $t(D,x) = t(D,y)$.

Proof. To prove the claim, denote D by $\{x, y, w_1, \ldots, w_{l-1}\}$. (If $l \leq 1$, then w_i's are not included in D and remove w_i's in the following proof.) Choose any $\pi, \pi' \in \Pi$ such that for any $z \in X \setminus \{x, y, w_1, \ldots, w_{l-1}\}$ and any $i \in \{1, \ldots, l-1\}$, $\pi(z) = \pi'(z)$, $\pi(z) > \pi(x) > \pi(y) > \pi(w_i)$, $\pi'(z) > \pi'(y) > \pi'(x) > \pi'(w_i)$, and $\pi(w_i) = \pi'(w_i)$.

To show the claim, we will show the following two facts: (a) For any $E \in \mathcal{D}$, $\rho^\pi(E) \neq \rho^{\pi'}(E)$ if and only if $\{x, y\} \subset E$ and $\pi(x) \geq \pi(E)$; (b) If $E \in \mathcal{D}$, $\{x, y\} \subset E$ and $\pi(x) \geq \pi(E)$, then $\rho^\pi(E, x) = 1$, $\rho^\pi(E, z) = 0$ for any $z \in D \setminus \{x\}$ and $\rho^{\pi'}(E, y) = 1$, $\rho^{\pi'}(E, z) = 0$ for any $z \in E \setminus \{y\}$.

It is easy to see statement (b) and the only if part of statement (a). To show the if part of statement (a), assume $\{x, y\} \not\subset E$ or $\pi(x) < \pi(z)$ for some $z \in E$. First consider the case where $\{x, y\} \not\subset E$. If both x and y do not belong to E, then $\rho^\pi(E) = \rho^{\pi'}(E)$ because the ranking over $X \setminus \{x, y\}$ is the same for π and π'. If only one of them, say x, belongs to E, then $\rho^\pi(E) = \rho^{\pi'}(E)$ because the ranking over $X \setminus \{y\}$ is the same for π and π'.

Next consider the case where $\pi(x) < \pi(z)$ for some $z \in E$. By the definition of π, we obtain $z \in X \setminus \{x, y, w_1, \ldots, w_{l-1}\}$. Therefore, $\pi'(y) < \pi'(z)$. Hence, $\rho^\pi(E, z) = 1 = \rho^{\pi'}(E, z)$ and $\rho^\pi(E, z') = 0 = \rho^{\pi'}(E, z')$ for all $z' \in E \setminus \{z\}$.

Now, we will prove the claim. Since $t \cdot \rho^\pi = t \cdot \rho^{\pi'}$,

$$
\begin{aligned}
0 &= \textstyle\sum_{(E,z)\in\mathcal{D}\times X} t(E,z)(\rho^\pi(E,z) - \rho^{\pi'}(E,z)) \\
&= \textstyle\sum_{(E,z)\in\mathcal{D}\times X:\{x,y\}\subset E,\pi(x)\geq\pi(E)} t(E,z)(\rho^\pi(E,z) - \rho^{\pi'}(E,z)) && (\because \text{(a)}) \\
&= \textstyle\sum_{E\in\mathcal{D}:\pi(x)\geq\pi(E),\{x,y\}\subset E}(t(E,x) - t(E,y)) && (\because \text{(b)}) \\
&= \textstyle\sum_{E\in\mathcal{D}:\pi(x)\geq\pi(E),\{x,y\}\subset E,|E|\geq l+1}(t(E,x) - t(E,y)) \\
&\quad + \textstyle\sum_{E\in\mathcal{D}:\pi(x)\geq\pi(E),\{x,y\}\subset E,|E|\leq l}(t(E,x) - t(E,y)) \\
&= t(D,x) - t(D,y) + \textstyle\sum_{E\in\mathcal{D}:\pi(x)\geq\pi(E),\{x,y\}\subset E,|E|\leq l}(t(E,x) - t(E,y)),
\end{aligned}
$$

where the last equality holds because if E contains both x and y, $\pi(x) \geq \pi(E)$, and $|E| \geq l+1$ then $|E| = l+1$, and hence E must be equal to D. The second

[45] If $\mathcal{D} = 2^X \setminus \emptyset$, then $l = 1$. If $\mathcal{D} \subsetneq 2^X \setminus \emptyset$, then l can be larger than 1.

term is zero because there is no $D \in \mathcal{D}$ such that $|D| \leq l$ by the definition of l. So $t(D,x) = t(D,y)$. This completes the proof of the claim. □

The general case can be proved by the induction on $|D|$. Choose any D such that $|D| = l' + 1$, where $l' > l$. Choose any $x, y \in D$. As an induction hypothesis, suppose that for any $E \in \mathcal{D}$, if $|E| \leq l'$ then $t(E,x) = t(E,y)$ for any $x, y \in E$. By the same argument (with l' in place of l) in the proof of the claim, we have

$$0 = t(D,x) - t(D,y) + \sum_{E \in \mathcal{D}: \pi(x) \geq \pi(E), \{x,y\} \subset E, |E| \leq l'} (t(E,x) - t(E,y)).$$

Since the second term is zero by the induction hypothesis, $t(D,x) = t(D,y)$. ■

A.9.1 Main Proof of Proposition 4 by using Lemma 10

The set $\{q \in \mathbf{R}^{\mathcal{D} \times X} | \text{(i) and (ii)}\}$ is affine. So it suffices to show that for any affine set A, if $\mathcal{P}_r \subset A$, then $\{q \in \mathbf{R}^{\mathcal{D} \times X} | \text{(i) and (ii)}\} \subset A$. Since the set is affine, then by Theorem 1.4 of Rockafellar (2015), there exist a positive integer L, $L \times (|\mathcal{D}| \times |X|)$ matrix B, and $L \times 1$ vector b such that $A = \{q \in \mathbf{R}^{\mathcal{D} \times X} | Bq = b\}$. For any $l \in \{1, \ldots, L\}$, $B_l(D,x)$ denotes $(l, (D,x))$ entry of B. (Remember that B has a column vector for each $(D,x) \in \mathcal{D} \times X$.) So $Bq = b$ means that for any $l \in \{1, \ldots, L\}$,

$$\sum_{D \in \mathcal{D}} \sum_{x \in X} B_l(D,x) q(D,x) = b_l. \tag{8}$$

By assuming $\mathcal{P}_r \subset \{q \in \mathbf{R}^{\mathcal{D} \times X} | Bq = b\}$, we will show that if q satisfies (i) and (ii), then (8) holds for any $l \in \{1, \ldots, L\}$.

Step 1: $B_l(D,x) = B_l(D,y)$ for any $l \in \{1, \ldots, L\}$, $D \in \mathcal{D}$, and $x, y \in D$. To prove Step 1, fix any l. For any $\pi \in \Pi$, $\rho^\pi \in \mathcal{P}_r \subset \{q \in \mathbf{R}^{\mathcal{D} \times X} | Bq = b\}$. Hence, (8) holds with $q = \rho^\pi$ for any $\pi \in \Pi$. Thus $\rho^\pi \cdot B_l = \rho^{\pi'} \cdot B_l$ for any $\pi, \pi' \in \Pi$. By Lemma 10, this implies that $B_l(D,x) = B_l(D,y)$ for any $D \in \mathcal{D}$, and $x, y \in D$.

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

Step 2: If q satisfies (i) and (ii), then $Bq = b$, i.e., $\sum_{D \in \mathcal{D}} \sum_{x \in X} B_l(D,x) q(D,x) = b_l$ for any $l \in \{1, \ldots, L\}$. To prove Step 2, choose any $\pi \in \Pi$ and $l \in \{1, \ldots, L\}$. Since $\rho^\pi \in \mathcal{P}_r \subset \{q \in \mathbf{R}^{\mathcal{D} \times X} | Bq = b\}$, then by (8),

$$b_l = \sum_{D \in \mathcal{D}} \sum_{x \in X} B_l(D,x) \rho^\pi(D,x) = \sum_{D \in \mathcal{D}} B_l(D), \tag{9}$$

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

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

$$\begin{aligned}
\textstyle\sum_{D\in\mathcal{D}}\sum_{z\in X}B_l(D,z)q(D,z) &= \textstyle\sum_{D\in\mathcal{D}}\sum_{z\in D}B_l(D,z)q(D,z) && (\because \text{(ii)})\\
&= \textstyle\sum_{D\in\mathcal{D}}\sum_{z\in D}B_l(D)q(D,z) && (\because \text{Step 1})\\
&= \textstyle\sum_{D\in\mathcal{D}}B_l(D)\sum_{z\in D}q(D,z) &&\\
&= \textstyle\sum_{D\in\mathcal{D}}B_l(D) && (\because \text{(i)})\\
&= b_l. && (\because (9))
\end{aligned}$$

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

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

A.10 Proof of Proposition 5

Since the set $\mathrm{cl.co}.\mathcal{P}_s(d,\eta|\mu)$ is compact and convex, $\rho^*=\arg\inf_{\rho\in\mathrm{cl.co}.\mathcal{P}_s(d,\eta|\mu)}d(\rho,\hat{\rho})=\arg\inf_{\rho\in\mathrm{cl.co}.\mathcal{P}_s(d,\eta|\mu)}||\rho-\hat{\rho}||_2^2$ exists and it can be written as a convex combination of elements of $\mathrm{cl}.\mathcal{P}_s(d,\eta|\mu)$. 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 $\rho_i\in\mathrm{cl}.\mathcal{P}_s(d,\eta)$, $M=\dim\mathcal{P}_r+1$. For each step n, define $E_n=\|\hat{\rho}-\rho^n\|^2-\|\hat{\rho}-\rho^*\|^2$. For each step n, let α_n^* and ρ_n^* be the minimizers over the grids $\{\alpha_n\}$ and $\mathrm{cl}.\mathcal{P}_s(d,\eta|\mu)$, respectively. Define $C=\sum_i\lambda_i\|\rho_i-\rho^*\|^2$ and $T=\max\{2E_1,4C\}$. Note that E_1 can be upper bounded by $2||\hat{\rho}||^2+2||\rho^n||^2\leq 4|\mathcal{D}|$ and similarly $C\leq 4|\mathcal{D}|$. Thus we can chooce $T=16|\mathcal{D}|$.

Then, for each step n and each α_n,

$$E_n \leq (1-\alpha_n)E_{n-1} + C\alpha_n^2.^{46} \tag{10}$$

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 (10), we have (i): $E_n \leq \frac{T}{n+1}$.[47] Let $d^n = d(\hat{\rho}, \rho^n)$ and $d^* = d(\hat{\rho}, \rho^*)$. Since $E_n = (\sum_{D\in\mathcal{D}} 1)^2((d^n)^2 - (d^*)^2)$, we have (ii): $(d^n)^2 - (d^*)^2 \leq \frac{T'}{n+1}$, where $T' = \frac{T}{(\sum_{D\in\mathcal{D}} 1)^2}$. Then we have $(d^n - d^*)^2 \leq (d^n - d^*)(d^n - d^*) + (d^n - d^*)2d^* = (d^n - d^*)(d^n + d^*) \leq \frac{T'}{n+1}$, where we use the fact that $d^n \geq d^*$ and $d^* \geq 0$. This implies $d^n - d^* \leq \sqrt{\frac{T'}{n+1}}$[48].

[46]Full calculation is as follows:

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

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

[48]We comment that we can upper bound E_1 and C by the squared diameter of the random utility polytope. For example $C \leq \sum_i \lambda_i||\rho_i - \rho^*||^2 \leq \sup_{x,y\in\mathcal{P}_r} ||x-y||_2^2 = 2 \times$ the number of choice sets. The extremum is achieved by selecting x to be a degenerate preference ranking and y its reverse ranking. Similarly we can bound E_1. Notice this implies $T = 8 \times$ the number of choice sets. Thus T' should be $\frac{8}{\text{the number of choice sets}}$

Online Appendix for "Approximating Choice Data by Discrete Choice Models"

Haoge Chang, Yusuke Narita, and Kota Saito

A EM Algorithm: Details

To compute approximation errors for unrepresentable rankings in Table 1, we fit finite-mixture models to each deterministic preference ranking by the method of maximum likelihood. The data input is the observed stochastic choice function $\widehat{\rho}(D,x)$ and covariates of each alternative. We choose the number of mixtures, M, according to the theoretical upper bound using Corollary 2. 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

$$\mathcal{L} \equiv \sum_{D\in\mathcal{D}} \sum_{x\in D} \widehat{\rho}(D,x) \log \sum_{i=1}^{M} \lambda_i \frac{\exp(\beta_i \cdot x)}{\sum_{y\in D}\exp(\beta_i \cdot y)}.$$

We estimate the parameters by the EM algorithm (Dempster et al. (1977), Train (2009)). We implement the algorithm according to Chapter 14 in Train (2009). We terminate the algorithm when the change of the implied L2 distance between the estimated choice probability and the target choice probability is smaller than $\frac{1}{10^6}$ between two successive runs.

Our use of Maximum Likelihood Estimation with the EM algorithm is partially motivated by the following observation: If the sufficient condition in Theorem 1-(i) 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. Maximum Likelihood Estimation therefore minimizes the approximation error metric in (5).

To see this, notice that under the sufficient condition in Theorem 1-(i), Proposition 4 and Corollary 2 imply that any interior random utility model can be represented by a finite mixture of logit models with $M = \sum_{D\in\mathcal{D}}(|D|-1)$ mixtures. That is, there exists a set of parameters $\{\beta_i^*, \lambda_i^*\}_{i=1}^M$ such that $\sum_{i=1}^M \lambda_i^* \frac{\exp(\beta_i^* \cdot x)}{\sum_{y\in D}\exp(\beta_i^* \cdot y)} =$

$\hat{\rho}(D,x)$ for any $D \in \mathcal{D}$, $x \in D$ and $M = \sum_{D\in\mathcal{D}}(|D|-1)$. This set of parameters maximizes the likelihood. Recall that for any other choice probability vector ρ, the likelihood is:

$$\sum_{D\in\mathcal{D}}\sum_{x\in D}\hat{\rho}(D,x)\log(\rho(D,x)).$$

$\hat{\rho}$ maximizes the likelihood since

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

where we use the fact $-\log(x) \geq -(x-1)$ for the inequality. Finally observe that this set of parameters yields a perfect fit of the target probability.

B In-sample and Out-of-sample Fit

In this section, we evaluate in-sample and out-of-sample fit of our model to address the possible concern of over-fitting. We use the same fishing choice dataset used in Section 5 and predict choice probabilities using aggregated characteristics. We find that our model performs better or equally well compared to standard models models, not only in terms of in-sample fit but also in terms of out-of-sample fit.

Our model is a random coefficient model with arbitrary mixing distributions. In the dataset, we have four alternatives and we consider only one choice set $\mathcal{D} = \{X\}$. Thus by Proposition 4 and Corollary 2, it suffices to mix four logit models without fixed effects to represent any random utility model.[49] We also estimate several standard models for comparision. They include a multinomial logit model; a nested logit model with two nests (charter and the rest); a nested logit model with two nests (boat and the rest); a random coefficient logit model with a log-normal mix-

[49] In this case, it is easy to show mixing three logit models is enough.

ing distribution for each variable; a multinomial logit model with alternative fixed effects; and a random coefficient logit model with log-normal mixing distributions and alternative fixed effects. We detail the definition of each specification in Section B.1.

To evaluate in-sample and out-of-sample fit, 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 $\max_{\theta \in \Theta} \sum_{j=1}^{|X|} \hat{\rho}_j \log \rho(x_j, \theta)$, 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 $\rho(x_j, \theta)$ is the model-predicted choice probability for fishing mode j with characteristic vector x_j. See Section B.1 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 in Section A of Online Appendix.[50]

To evaluate the in-sample fit performance, we compute the predicted choice probabilities in the training sample ($\hat{\rho}_{train} \in \mathbf{R}^{|X|}$) and compare it with the observed choice probabilities in the training sample ($\rho_{train} \in \mathbf{R}^{|X|}$) .[51] For this comparison, we calculate the l2 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 ($\hat{\rho}_{test} \in \mathbf{R}^{|X|}$) and compare it with the aggregated observed choice probabilities in the testing sample ($\rho_{test} \in \mathbf{R}^{|X|}$). We use the l2 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.1. The results for out-of-sample fits are in Table A.2.

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 l2 prediction error 0.038.

[50] We prefer the EM algorithm over the greedy algorithm here because the EM algorithm is faster.

[51] We only consider the single choice set case in this simulation. So the choice probability vector has length $|X|$.

Table A.2 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 the fixed effects, our model performs better or equally well compared to standard models in this simulation, not only in terms of in-sample fit but also in terms of out-of-sample fit.

Table A.1: In-Sample Fit

	Choice probabilities				Prediction error
	Beach	Boat	Charter	Pier	
(1)	(2)	(3)	(4)	(5)	(6)
Our method	0.114	0.353	0.383	0.151	0.000
	(0.009)	(0.015)	(0.017)	(0.010)	(0.000)
Multinomial logit	0.141	0.355	0.378	0.126	0.038
	(0.007)	(0.015)	(0.017)	(0.007)	(0.009)
Nested logit	0.114	0.353	0.383	0.150	0.001
(charter and others)	(0.010)	(0.015)	(0.017)	(0.011)	(0.002)
Nested logit	0.141	0.355	0.378	0.126	0.038
(boat and others)	(0.007)	(0.015)	(0.017)	(0.007)	(0.009)
Mixed logit with	0.142	0.354	0.378	0.126	0.038
log normal distribution	(0.008)	(0.015)	(0.017)	(0.007)	(0.009)
Multinomial logit	0.113	0.353	0.383	0.151	0.000
with fixed effects	(0.009)	(0.015)	(0.017)	(0.010)	(0.000)
Mixed logit with log normal	0.114	0.353	0.383	0.151	0.000
distribution and fixed effects	(0.009)	(0.015)	(0.017)	(0.010)	(0.000)

Note: Table A.1 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.

B.1 Definitions of Other Models

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

Table A.2: Out-Sample Fit

	Choice probabilities				Prediction error
	Beach	Boat	Charter	Pier	
(1)	(2)	(3)	(4)	(5)	(6)
Our method	0.114	0.353	0.383	0.151	0.049
	(0.009)	(0.015)	(0.017)	(0.010)	(0.017)
Multinomial logit	0.143	0.353	0.377	0.127	0.058
	(0.011)	(0.017)	(0.022)	(0.011)	(0.019)
Nested logit	0.115	0.350	0.383	0.152	0.050
(charter and others)	(0.012)	(0.022)	(0.018)	(0.014)	(0.020)
Nested logit	0.143	0.353	0.377	0.127	0.058
(boat and others)	(0.011)	(0.017)	(0.022)	(0.011)	(0.019)
Mixed logit with	0.143	0.352	0.377	0.127	0.058
log normal distribution	(0.010)	(0.016)	(0.021)	(0.010)	(0.018)
Multinomial logit	0.116	0.350	0.380	0.154	0.048
with fixed effects	(0.014)	(0.019)	(0.022)	(0.019)	(0.022)
Mixed logit with log normal	0.113	0.353	0.383	0.151	0.048
distribution and fixed effects	(0.008)	(0.015)	(0.018)	(0.010)	(0.017)

Note: Table A.2 summarizes the out-sample fit of different models. The row "our method" presents choice probabilities predicted by the four-mixture mixed logit model and the prediction error (5). 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: $\rho_j = \dfrac{\exp(x_j'\beta)}{\sum_{j' \in J} \exp(x_{j'}'\beta)}$
- Nested logit (charter and others): the choice probability of alternative j that belongs to nest g is specified as

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

  The nest is defined by the partition $G = \{\{\text{charter}\}, \{\text{beach, boat, pier}\}\}$.
- Nested logit (boat and others): the nested logit model specified above, with the nest defined by $G = \{\{\text{boat}\}, \{\text{beach, charter, pier}\}\}$.
- Mixed logit: $\rho_j = \int \dfrac{\exp(x_j'\beta)}{\sum_{j' \in J} \exp(x_{j'}'\beta)} f(\beta) d\beta$ where f is the density of the dis-

tribution of random coefficients. We use independent log-normal distributions for each coefficient. To evaluate the integral, we random draw 100 realizations from the random coefficient distribution.

- 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 random draw 100 realizations from the random coefficient distribution.