Mixed Logit and Pure Characteristics Models [*]

Jay Lu[†] Kota Saito[‡]

January 2022

Abstract

Mixed logit or random coefficients logit models are used extensively in empirical work while pure characteristic models feature in much of theoretical work. We provide a theoretical analysis of the relationship between the two classes of models. First, we show an approximation theorem that precisely characterizes the extent and limits of mixed logit approximations of pure characteristic models. Second, we present two conditions that highlight behavioral differences between the two classes of models. The first is a substitutability condition that is satisfied by many pure characteristic models (including the Hotelling model of horizontal differentiation) but is violated by almost all mixed logit models. The second is a continuity condition that is satisfied by all pure characteristic models but is violated by all mixed logit models. Both conditions pertain to choice patterns when product characteristics change or new products are introduced and illustrate the limitations of using mixed logit models for counterfactual analysis.

[*] We thank Giovanni Compiani, Wayne Gao, Yusuke Narita, Matt Shum, Yi Xin and seminar participants at Caltech, Hitsubashi, Academia Sinica and UC Irvine for helpful comments. Financial support from the NSF under awards SES-1558757 (Saito), SES-1919263 (Saito) and SES-1919275 (Lu) are gratefully acknowledged. Jiayin Zhai and Minghao Pan provided outstanding research assistance.

[†] Department of Economics, UCLA; jay@econ.ucla.edu.

[‡] Division of the Humanities and Social Sciences, Caltech; saito@caltech.edu.

1 Introduction

Mixed logit models, also known as random coefficients logit models, have been widely used in empirical work across different fields (McFadden (1973), Rust (1987) and Berry, Levinsohn and Pakes (1995)). In these models, agents' utilities contain iid extreme-value distributed error terms that generate convenient expressions for choice probabilities which are useful for estimation. On the other hand, much of the theoretical literature in decision theory and industrial organization since Hotelling (1929) have focused on pure characteristic models (Berry and Pakes (2007)). In these models, there are no iid error terms and utilities are continuous functions of product characteristics. In this paper, we provide a theoretical analysis of the relationship between these two classes of models.

Our main contribution is two-fold. First, we provide an approximation theorem that precisely characterizes the extent and limit to which mixed logit models can approximate pure characteristic models. This sharpens existing approximation results (e.g. McFadden and Train (2000)) and shows that a pure characteristic model can be approximated by mixed logit models if and only if they belong to the same parametric family. Second, we highlight two novel patterns of choice behavior related to product differentiation, *convex substitutability* and *continuity in characteristics*, that are natural in pure characteristic models but cannot be easily accommodated by mixed logit models.

These results highlight the tradeoffs of using mixed logit models. On the one hand, they are flexible enough to accommodate a wide range of behaviors and can approximate pure characteristic models within the same parametric family with arbitrary precision. This makes them a useful class of models for estimation. However, once a mixed logit model has been estimated using available data, the structure of the model imposes certain restrictions on choice patterns when considering the introduction of new products or changes in product characteristics. This highlights the limitations of using mixed logit models for counterfactual analysis.

Our theoretical results complement the empirical literature by formalizing some of the well-known issues with mixed logit models (see discussion in related literature). Focusing on conditions on choice behavior provides a useful methodology to distinguish between the two classes of models. Moreover, these conditions can also serve more generally as useful criteria when exploring other possible models (beyond mixed logit) as candidates for estimation.

In our model, each choice option (i.e. product) corresponds to a vector in $\mathbb{R}^k$ where k is the number of characteristics. We assume there is a rich set continuous product characteristics. Following most empirical work, we focus on parametric families of polynomials up to some degree $d \geq 1$. For example, polynomials of degree 1 correspond to all linear functions $u(x) = \beta \cdot x$. The main approximation result (Theorem 1) states that a pure characteristic model of degree d can and only can be approximated by mixed logit of degree d. For instance, if the pure characteristic has degree d, then it is in general impossible to approximate the model using mixed logit of degree $d' < d$. In practice, this means that specifying the correct specification of the degree of the parametric family of utilities is important for mixed logit approximations.

In the process of showing this result, we also characterize the universal set of all models that can be approximated by logit and mixed logit models. These results may be of independent interest to researchers. The closure of logit is a class of models which we call *lexicographic-logit*; this is a lexicographic choice rule with logit tie-breaking. Lexicographic-logit is an exceptionally rich class of models and includes some models (e.g. lexicographic choice) that cannot be expressed as random utilities.

While mixed logit models are flexible and can approximate any pure characteristic model, there are inherent differences between the two classes of models which we highlight using two novel conditions on choice behavior. The first condition is *convex substitutability.* Consider two products x and y and a third product $z = \lambda x + (1 - \lambda) y$ with intermediary characteristics of the other two. Convex substitutability says that the demand for x decreases if we replace y with z. The intuition is that since x is more similar to z than to y, agents will substitute away from x when y is replaced with the more similar product. This condition is satisfied by many pure characteristic models, including the classic Hotelling model; however, it is violated by all mixed logit models excepting the special case of uniform choice (Theorem 2).

The second condition is *continuity in characteristics.* This is a continuous version of classic “duplicates problem” Consider two products x and y and series of products y_n with characteristics that converge to those of y, i.e. $y \to y_n$ in $\mathbb{R}^k$. Continuity in characteristics says that the demand for x when both y and y_n are available will eventually converge to the demand for x when only y is available. The intuition is that agents will eventually be unable to distinguish between y and y_n and treat both as the same product. This condition is satisfied by all pure characteristic models but violated by all mixed logit models (Theorem 3).

These conditions show stark differences in choice behavior between mixed logit and pure characteristic models. How significant are these differences for practical purposes? On the one hand, the discrepancies will eventually vanish as mixed logit approximations get arbitrarily close to the pure characteristic model. On the other hand, any mixed logit that eventually emerges from estimation will violate these conditions. Since both conditions pertain to product differentiation, this would complicate counterfactual analysis when product characteristic vary or new products are substituted. The significance and magnitude of these violations will depend on the specific application, and we show in a simulated example that these violations can be severe (see Appendix I). In general, our results highlight the potential issues to consider when using mixed logit approximations.

We focus on the mixed logit due to its prominence in applied work but our results extend to more generally to a larger class of models. For example, all models with iid error terms would have difficulty satisfying convex substitutability and continuity in characteristics. While iid error terms are useful for modeling unobserved heterogeneity, their convenience imposes restrictions on choice behavior that may be undesirable. Importantly, not all classes of models necessarily violate our two conditions; we provide an example of a model with correlated errors that satisfies both (see Example 8). In general, conditions on choice behavior can serve as useful criteria when exploring other classes of models as candidates for estimation.

1.1 Related Literature

Luce (1959) provided an early characterization of multinomial logit. Recent papers in decision theory have considered generalizations of logit. These include mixed logit (Gul, Natenzon, and Pesendorfer (2014), Saito (2018)) and nested logit (Kovach and Tserenjigmid (2020))). Cerreia-Vioglio, Maccheroni, Marinacci and Rustichini (2018a; 2018b) consider the Luce axiom without positivity and obtain a model that is a discrete version of our lexicographic-logit model. Fudenberg and Strzalecki (2015) consider dynamic extensions of logit. Natenzon (2019) studies a Bayesian probit model. Chambers, Cuhadaroglu and Masatlioglu (2020) consider a variation of the logit model in a social setting.

Theoretical work in decision theory has focused on pure characteristic models. These include Gul and Pesendorfer (2006), Ahn and Sarver (2013), Lu (2016), Apesteguia, Ballester and Lu (2017), Lu and Saito (2018), Duraj (2018), Frick, Iijima and Strzalecki (2019), Lu (2019) and Lin (2019). Wilcox (2011) and Apesteguia and Ballester (2018) discuss issues with respect to comparative statics between logit and pure characteristic models while Frick, Iijima, and Strzalecki (2019) highlight issues associated with assessing option values. These results are similar in spirit to ours highlighting the differences in choice behavior between the two class of models.

In the empirical literature, logit-based models have been widely applied for discrete choice analysis. These include McFadden (1973), Rust (1987), Hotz and Miller (1993), Berry, Levinsohn and Pakes (1995), Nevo (2001), Hendel and Nevo (2006), Gowrisankaran and Rysman (2012) and Compiani (2019). McFadden and Train (2000) show that that mixed logit models can approximate any pure characteristic model. Our result sharpens their result and captures the precise extent of this approximation. Narita and Saito (2021) consider the case where the set of characteristics is finite and provide a condition that characterizes when mixed logit can approximate random utility models. They also provide an algorithm for constructing mixed logit models that can approximate random utility arbitrarily well when the condition is satisfied; when it is not satisfied, they find that the size of the approximation error is large.

Other papers comparing logit-based models with pure characteristics models (also known as hedonic models) include Anderson, DePalma and Thisse (1989), Petrin (2002), Bajari and Benkard (2004; 2005), Ackerberg and Rysman (2005) and Berry and Pakes (2007). They consider the implications on price elasticities and welfare when new products are introduced. Logit-based models may imply too much "taste for product" while pure characteristics models may imply competition that is too localized. Our results on convex substitutability and continuity in characteristics are similar in spirit and highlight new patterns in choice behavior differentiating the two classes of models.

2 Setup

There are $k \geq 1$ characteristics and we associate each choice option (i.e. product) with a vector $x \in X \subset \mathbb{R}^k$ of characteristics. We assume that set of characteristics is rich. Formally, X is full-dimensional, compact and convex. A *menu* $A \subset X$ is a finite set of products. Let $\mathcal{A}$ denote the set of all menus. A *stochastic choice* ρ is a mapping on $\mathcal{A}$ such that for any menu $A \in \mathcal{A}$, $\rho_A(\cdot)$ is a probability distribution over elements in A. For binary menus, $A = \{x, y\}$, we use the simpler notation $\rho(x, y) = \rho_A(x)$. The set of all stochastic choice can be thus defined as

$$\mathcal{P} = \prod_{A \in \mathcal{A}} \Delta A$$

where ΔA is the set of all probability distributions over A. We endow $\mathcal{P}$ with the product topology.

We focus on utilities that are continuous in product characteristics. Formally, let $U \subset \mathbb{R}^X$ denote the set of all continuous utility functions $u : X \to \mathbb{R}$. We say $u \in U$ is a *polynomial of degree* $d \geq 1$ if it is a multivariate polynomial where every term has exponents that sum up to at most d, i.e.

$$u(x_1, \ldots, x_k) = \sum_{m_1^i + \cdots + m_k^i \leq d} \beta_i x_1^{m_1^i} \cdots x_k^{m_k^i}$$

for some $\beta \in \mathbb{R}^{m(k,d)}$ where $m(k,d) = \sum_{i=1}^{d} \frac{(k+i-1)!}{i!(k-1)!}$. For example, if $k = d = 2$, then

$$u(x_1, x_2) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$

In general, β represents the coefficients for different characteristics including higher-order and interaction terms. Constant terms β_0 are excluded as they have no bearing on any of the subsequent analysis. Given any $x \in X$, let x^* denote the corresponding vector in polynomial space X^* so

$$u(x) = \beta \cdot x^*.$$

Note that if $d = 1$, then $x = x^*$ and $u(x) = \beta \cdot x$ is just a linear function. Let $U_d \subset U$ denote the set of all polynomials of degree d.

Two prominent classes of models are pure characteristic and mixed logit.

Definition 1. (Pure characteristic) A stochastic choice ρ is *pure characteristic* if there exists a distribution μ on U such that

$$\rho_A(x) = \mu(\{u \in U : u(x) \geq u(y) \text{ for all } y \in A\}).$$

It is *pure characteristic of degree* d if $u \in U_d$ a.s.

Definition 2. (Mixed logit) A stochastic choice ρ is *mixed logit* if there exists a distribution ν on U such that

$$\rho_A(x) = \int_U \frac{e^{v(x)}}{\sum_{y \in A} e^{v(y)}} d\nu$$

It is *mixed logit of degree* d if $v \in U_d$ a.s.

A special case of mixed logit is of course when the distribution $\mu = \delta_u$ is degenerate. In this case, we obtain standard logit where v is its *systematic utility*. We summarize this below.

Definition 3. (Logit) A stochastic choice ρ is *logit* if there exists a $v \in U$ such that

$$\rho_A(x) = \frac{e^{v(x)}}{\sum_{y \in A} e^{v(y)}}$$

It is *logit of degree d* if $v \in U_d$.

Both pure characteristic and mixed logit belong to a more general class of random utility models. To see why, note that we can rewrite the mixed logit model as

$$\rho_A(x) = \mathbb{P}\left(\{v(x) + \varepsilon(x) \geq v(y) + \varepsilon(y) \text{ for all } y \in A\}\right)$$

where $v(\cdot)$ are distributed according to ν and $\varepsilon(\cdot)$ are extreme-valued distributed and iid across products. On the other hand, the pure characteristic model is a random utility model where the utilities $u(\cdot)$ are continuous in product characteristics. Note that this does not preclude the modeling of unobserved characteristics with error terms; for example, if we let $u(x) = v(x) + \varepsilon(x)$ where $\varepsilon(\cdot)$ follows some Brownian motion, then this is still a pure characteristic model since the error terms are continuous in product characteristics.

Both the space of characteristics X and the set of menus $\mathcal{A}$ we consider are extremely rich. We adopt this approach for two reasons. First, it allows us to consider stochastic choice approximations that are universal in the sense that they apply to all menus.[1] Second, it facilitates counterfactual analysis for when product characteristics vary or additional products are introduced to the market. Our results are thus particularly relevant in settings that involve rich data sets.

3 Mixed Logit Approximations of Pure Characteristic Models

3.1 An Approximation Theorem

This section provides a precise characterization of the extent to which mixed logit models can be used to approximate any pure characteristic model. McFadden and Train (2000)

[1] At the other extreme, one could consider a single menu $A = \{x, y\}$ with only two products. In this case, mixed logit can be used to fit any pure characteristics model exactly.

showed that mixed logit models can be used to approximate any pure characteristic model.[2] We translate their result to our setup. Recall that $\mathcal{P}$ is endowed with the product topology so $\rho^n \to \rho$ iff $\rho^n_A \to \rho_A$ for all $A \in \mathcal{A}$.

Proposition 1. *For any pure characteristic ρ, there exists a sequence of mixed logits ρ^n such that $\rho^n \to \rho$.*

Proof. See Appendix. □

While mixed logit models owe much of their popularity to their computability, the above result also provides some theoretical justification for their use. In practice however, a researcher will usually commit to some class of mixed logit models for approximation. For example, suppose the researcher uses mixed logit of degree $d = 1$, i.e. mixed logit models where the systematic utility $u(x) = \beta \cdot x$ is linear. What is the set of all pure characteristic models that this can approximate? While this clearly includes pure characteristic models with linear utilities, could it include more? It turns out the answer is no.

We now provide a precise characterization of the extent to which mixed logit models can be used to approximate pure characteristic models. We say a stochastic choice ρ can be *approximated* by a set of stochastic choice models if it is in the closure of that set.[3]

Theorem 1. *For any pure characteristic ρ, the following are equivalent:*

(1) *ρ is pure characteristic of degree d*

(2) *ρ can be approximated by mixed logit of degree d.*

Proof. See Appendix. □

The main implication of this result is that it is important to correctly specify the degree of the pure characteristic model. For example, suppose the pure characteristic model has degree 3 where third-order terms matter. In this case, it would be generally impossible to approximate the pure characteristic model if the researcher only uses mixed logits of degree $d < 3$. A special case when a lower-degree mixed logit would suffice is if the utilities in the pure characteristic model is exactly a monotone transformation of a lower-degree polynomial. For instance, if the utilities in the pure characteristic model satisfy

$$u(x) = \beta_1 x_1^3 + \beta_2 x_1^2 x_2 + \beta_3 x_1 x_2^2 + \beta_4 x_2^3 = (\gamma_1 x_1 + \gamma_2 x_2)^3$$

[2] McFadden and Train (2000) also consider unobservable characteristics in mixed logit models. We can apply their approach in our setup as well.

[3] Recall that closure here is with respect to the product topology on $\mathcal{P}$.

where $\beta_1 = \gamma_1^3$, $\beta_2 = 3\gamma_1^2\gamma_2$, $\beta_3 = 3\gamma_1\gamma_2^2$ and $\beta_4 = \gamma_2^3$. In this case, mixed logits of degree 1 can be used to approximate this model.

Theorem 1 sharpens the result from McFadden and Train (2000) in two ways. First, while they only provide sufficiency, we provide necessity as demonstrated in the example above. Second, our result is more precise about the class of utility functions that can be approximated (i.e. the degree of the polynomial utility). This has a practical importance for empirical analysis given that the majority of empirical work assumes linear (i.e. $d = 1$) mixed logit.

There are two other ways in which mixed logit approximations are limited. First, as the proof of Theorem 1 shows (also see Example 1), approximations often require certain mixed logit parameters to explode to infinity. This is not optimal from a practical perspective as it is often convenient to work with compact parameter spaces. In fact, if we restrict the space of mixed logit parameters β to be compact, then our approximation would not hold in general. Second, we define closure with respect to the product topology so the rates of convergence are menu-dependent. In other words, ρ^n may converge much slower for some menus than others. Ideally, it would be nice to show that the rates of convergence are similar for all menus. This involves a notion of uniform convergence across all menus which we show is impossible (see Appendix H).

3.2 Closure of Logit: Lexicographic-logit

In order to demonstrate the reasoning behind Theorem 1, we first characterize the universal set of all models that can be approximated by (mixed) logit. This characterization may be of independent interest to researchers. It shows the full extent in which (mixed) logit models can be used to approximate a rich class of models including some that are not pure characteristic or even random utility. We begin with an example of a model that is the limit of logits.

Example 1 (*Lexicographic choice rule*)**.** *Let $X = [0,1]^2 \subset \mathbb{R}^2$ and $d = 1$. Let ρ^n be logit with $\beta_n = (n^2, n)$ and ρ be the limit of the logits ρ^n. Thus,*

$$\rho_A(x) = \lim_n \frac{e^{\beta_n \cdot x}}{\sum_{y \in A} e^{\beta_n \cdot y}} = \left(\sum_{y \in A} e^{\lim_n \left(n^2(y_1 - x_1) + n(y_2 - x_2) \right)} \right)^{-1}$$

In this case, ρ corresponds to the lexicographic preference $\succ$ on X where $x \succ y$ if $x_1 > y_1$ or

$x_1 = y_1$ and $x_2 > y_2$. To see why, note that if $x_1 > y_1$ or $x_1 = y_1$ and $x_2 > y_2$, then

$$\lim_n \left(n^2 (y_1 - x_1) + n (y_2 - x_2)\right) = -\infty$$

If this is true for all $y \in A$, then $\rho_A(x) = 1$. On the other hand, if $y \succ x$ for some $y \in A$, then $\rho_A(x) = 0$ as desired. Thus, ρ is a lexicographic choice rule.

The above example shows how lexicographic choice rules is one class of models that can be approximated by logit models. Note that in that example, ρ is not pure characteristic or even a random utility. To see why, suppose otherwise and ρ is a random utility model with some distribution over all utility functions $u : X \to \mathbb{R}$. Thus, for each distinct $x, y \in X$, we have $x \succ y$ if $u(x) > u(y)$ with probability one. If we define the average utility $\bar{u}(x) := \mathbb{E}[u(x)]$, then $\bar{u}$ represents $\succ$. This yields a contradiction since it is well-known that no utility representation exists for lexicographic preferences. Thus, although every logit is a random utility, the closure of logit includes models that cannot be expressed as a random utility.[4]

While mixed logit can approximate pure characteristic models and lexicographic choice rules, what is the full class of models that can be approximated? We now characterize that set. Consider a collection of polynomials $(u_1, \dots, u_t)$ where $u_i \in U_d$ for all $i \in \{1, \dots, t\}$. Let $(\beta_1, \dots, \beta_t)$ be their corresponding coefficients so $u_i(x) = \beta_i \cdot x^*$ for all i. We say the collection is *orthogonal* if $\beta_i \cdot \beta_j = 0$ for all $i, j \in \{1, \dots, t\}$.

Given $\omega = (u_1, \dots, u_t)$, let $\succeq_\omega$ be its induced lexicographic preference relation on X. In other words, $x \sim_\omega y$ if $u_i(x) = u_i(y)$ for all $i \in \{1, \dots, t\}$ and $x \succ_\omega y$ if $u_i(x) > u_i(y)$ for some $i \leq t$ and $u_j(x) = u_j(y)$ for all $j < i$. Let Ω_d be the set of all orthogonal polynomials $\omega = (u_1, \dots, u_t)$ for some $t \leq m(k, d)$. Under lexicographic-logit, choices follow a lexicographic preference relation where ties are broken according to logit.

Definition 4. ρ is *lexicographic-logit of degree d* if there exist $\omega \in \Omega_d$ and $v \in U_d$ such that

$$\rho_A(x) = 1\{x \succeq_\omega y \text{ for all } y \in A\} \frac{e^{v(x)}}{\sum_{y \in A, y \sim_\omega x} e^{v(y)}}$$

The following result shows that the closure of logit is exactly lexicographic-logit.

[4] If we allow for only finite-additive distributions or non-measurable utilities, then one could represent lexicographic choice rules using some "random utility" (see Cohen (1980)). However, given that a lexicographic preference has no utility representation, it would be odd for it to have a random utility representation. Moreover, this would exclude the possibility of integrating utilities (e.g. calculating social surplus).

Proposition 2. *The following are equivalent:*

(1) *ρ is lexicographic-logit of degree d*

(2) *ρ can be approximated by logit of degree d.*

Proof. See Appendix. □

How about for mixed logit? We now define the mixed lexicographic-logit model.

Definition 5. ρ is *mixed lexicographic-logit of degree d* if there exists a distribution ν on $\Omega_d \times U_d$ such that

$$\rho_A(x) = \int_{\Omega_d \times U_d} 1\left\{x \succeq_\omega y \text{ for all } y \in A\right\} \frac{e^{v(x)}}{\sum_{y \in A, y \sim_\omega x} e^{v(y)}} d\nu$$

The following result parallels Proposition 2 and shows that mixed lexicographic-logit is exactly the set of all stochastic choices that can be approximated by mixed logit. In fact, it is the smallest set of models containing logit that is closed under mixing and approximations.

Proposition 3. *The following are equivalent:*

(1) *ρ is mixed lexicographic-logit of degree d*

(2) *ρ can be approximated by mixed logit of degree d.*

Proof. See Appendix. □

Mixed lexicographic-logit includes a rich class of models. When ties are universal, i.e. $x \sim_\omega y$ for all $x, y \in X$ a.s., this reduces to mixed logit. When ω only consists of a single polynomial and ties never occur, this reduces to pure characteristic. The following are a few additional special cases:

Example 2 (*Mixed lexicographic*)**.** *Let $\omega = (u_1, \dots, u_t)$ for $t > 1$ and suppose ties never occur, i.e. $x \not\sim_\omega y$ for all $x, y \in X$ a.s. This corresponds to a population of agents where each agent chooses according to a lexicographic preference. As special case of course is Example 1 above.*

Example 3 (*Mixture of logit and pure characteristic*)**.** *Let ν_1 correspond to a mixed logit model, ν_2 correspond to a pure characteristic model and $\nu = \alpha\nu_1 + (1-\alpha)\nu_2$ for $a \in (0,1)$. Here, α parametrizes the degree of iid noise in the model. Note that this is neither a mixed logit model nor a pure characteristic model.*

Example 4 (*Generalized nested-logit*)**.** *Let $A = \{x, y, z\}$ and consider $u_1, u_2 \in U_d$ such that $u_1(x) = u_1(y) > u_1(z)$ and $u_2(x) < u_2(y) = u_2(z)$. Suppose ν is such that $(\omega, v) = (u_1, v_1)$ with probability α and $(\omega, v) = (u_2, v_2)$ with probability $1 - \alpha$. In this case,*

$$\rho_A(x) = \alpha \frac{e^{v_1(y)}}{e^{v_1(y)} + e^{v_1(x)}} + (1 - \alpha) \frac{e^{v_2(y)}}{e^{v_2(y)} + e^{v_2(z)}}$$

This corresponds to an agent who either picks the "nest" $\{x, y\}$ in the first-stage followed by logit with v_1 in the second stage or picks the "nest" $\{y, z\}$ followed by logit v_2. If we interpret the nests as consideration sets, then u_1 and u_2 correspond to salience measures.

Finally, we end this section with a brief outline of how Proposition 3 can be used to prove Theorem 1. Suppose ρ is a pure characteristic model that can also be approximated by mixed logits of degree $d = 1$. Proposition 3 implies that ρ is mixed lexicographic-logit of degree $d = 1$. We first show that $x \sim_\omega y$ can never occur with positive probability so logit tie-breaking never occurs. Let $y_n = \left(1 - \frac{1}{n}\right) y + \frac{1}{n} x$ and note that since $\omega \in \Omega_1$, $x \succeq_\omega y$ iff $x \succeq_\omega y_n$. Let $A_n = \{x, y, y_n\}$ and taking the limit as $n \to \infty$, we have

$$\begin{aligned}
\lim_n \rho_{A_n}(x) &= \lim_n \int_{\Omega_1 \times \mathbb{R}^k} 1\{x \succeq_\omega y\} \frac{e^{\beta \cdot x}}{e^{\beta \cdot x} + 1\{y \sim_\omega x\}\left(e^{\beta \cdot y} + e^{\beta \cdot y_n}\right)} d\nu \\
&= \int_{\Omega_1 \times \mathbb{R}^k} 1\{x \succeq_\omega y\} \frac{e^{\beta \cdot x}}{e^{\beta \cdot x} + 1\{y \sim_\omega x\} 2 e^{\beta \cdot y}} d\nu \\
&\leq \int_{\Omega_1 \times \mathbb{R}^k} 1\{x \succeq_\omega y\} \frac{e^{\beta \cdot x}}{e^{\beta \cdot x} + 1\{y \sim_\omega x\} e^{\beta \cdot y}} d\nu = \rho(x, y)
\end{aligned}$$

Since ρ is also pure characteristic, $\lim_n \rho_{A_n}(x) = \rho(x, y)$ (see Theorem 3.1 below) so it must be that $y \sim_\omega x$ with measure zero. We can thus write

$$\rho_A(x) = \nu\left(\{\omega \in \Omega_1 : x \succeq_\omega y \text{ for all } y \in A\}\right)$$

It is straightforward to show that this satisfies the random linear utility axioms of Gul and Pesendorfer (2006) so ρ is pure characteristic of degree $d = 1$. The full proof of Theorem 1 extends this argument for $d > 1$.

4 Differences in Choice Behavior

This previous section focuses on the extent to which mixed logit models can be used to approximate any pure characteristic model. In this section, we focus on two behavioral differences between the two classes of models. Both pertain to patterns in choice behavior relating to product differentiation that are important for conducting counterfactual analysis. Section 4.1 studies a substitutability condition that is satisfied by many pure characteristic models but is violated by almost all mixed logit models. Section 4.2 studies a continuity condition that is satisfied by all pure characteristic models but is violated by all mixed logit models.

4.1 Convex Substitutability

This section introduces a substitutability condition that is natural in many pure characteristic models but cannot be accommodated by almost all mixed logit models. To illustrate this property, consider the following example.

Example 5 (*Convex substitutes*)**.** *There are two products $x, y \in X \subset \mathbb{R}^k$ where $\rho(x, y)$ denotes the demand for product x over product y. Consider an intermediary product $z = \frac{1}{2}x + \frac{1}{2}y$ that is a convex mixture of the characteristics of x and y. Now, suppose we substitute product y with z. Since product x is more "similar" (in a convex sense) to z than it is to y, more consumers will substitute away from x to z. This means that the demand of x will decrease if we replace y with z, i.e.*

$$\rho(x, y) \geq \rho\left(x, \frac{1}{2}x + \frac{1}{2}y\right).$$

This same reasoning would apply if we considered any intermediary product $z = \lambda x + (1 - \lambda) y$ where $\lambda \in (0, 1)$.

We now extend this property for generic menus. The same relationship would still hold for x and y in any menu A as long as there are no intermediary products $z = \alpha x + (1 - \alpha) y$ already in A. Let $[x, y] \subset X$ denote the set of all convex mixtures $\alpha x + (1 - \alpha) y$ for all $\alpha \in [0, 1]$.

Definition 6. ρ satisfies *convex substitutability* if $\rho_{A \cup \{y\}}(x) \geq \rho_{A \cup \{\lambda x + (1-\lambda) y\}}(x)$ for all $\lambda \in (0, 1)$ and $A \cap [x, y] = \emptyset$.

Convex substitutability says that the demand for an alternative decreases when other alternatives become more similar. This is because similar alternatives serve as substitutes. Importantly, similarity here is measured in terms of convexity in the space of characteristics $\mathbb{R}^k$.

When is convex substitutability satisfied in different models? Consider the classic Hotelling (1929) model of horizontal differentiation.

Example 6 (*Hotelling*). *Each alternative corresponds to a product $x = (\theta, p) \in \mathbb{R}^2$ where θ measures location and p is its price. Each agent i has utility*

$$u_i(\theta, p) = \alpha_i - \gamma_i (\theta - \beta_i)^2 - p$$

This is a pure characteristic model where $\rho(x, y)$ is the demand of product x over y. Given $x = (\theta, p)$ and $y = (\theta', p')$, let $z = \frac{1}{2}x + \frac{1}{2}y$ denote the intermediary product that is a mixture of location and price of the two products. If an agent prefers x to z, then he must also prefer x to y. To see why, note that

$$\begin{aligned} u_i(z) &= \alpha_i - \gamma_i \left(\frac{1}{2}\theta + \frac{1}{2}\theta' - \beta_i\right)^2 - \frac{1}{2}p - \frac{1}{2}p' \\ &\geq \alpha_i - \frac{1}{2}\gamma_i (\theta - \beta_i)^2 - \frac{1}{2}\gamma_i (\theta' - \beta_i)^2 - \frac{1}{2}p - \frac{1}{2}p' \\ &\geq \frac{1}{2}u_i(x) + \frac{1}{2}u_i(y) \end{aligned}$$

so $u_i(x) \geq u_i(z)$ implies $u_i(x) \geq u_i(y)$. This means that $\rho(x, z) \leq \rho(x, y)$ and it is easy to see the convex substitutability is satisfied in general.

In the Hotelling model, convex substitutability is satisfied because utilities are concave. In fact, convex substitutability is satisfied as long as all the utilities in the pure characteristic model are quasiconcave. We say a pure characteristic model is *quasiconcave* if its utility functions $u : X \to R$ are quasiconcave a.s.

Theorem 2.1. *Any quasiconcave pure characteristic ρ satisfies convex substitutability.*

Proof. Consider distinct $x, y \in X$ and let $z = \lambda x + (1 - \lambda) y$. Since u is quasiconcave, if $u(x) > u(z)$, then $u(x) \geq u(y)$. Since ties occur with measure zero, this means that

$$\begin{aligned} \rho_{A \cup \{z\}}(x) &= \mu(\{u \in U : u(x) > u(w) \text{ for all } w \in A \cup \{z\}\}) \\ &\leq \mu(\{u \in U : u(x) > u(w) \text{ for all } w \in A \cup \{y\}\}) = \rho_{A \cup \{y\}}(x) \end{aligned}$$

as desired. □

Note that the proof holds even if utilities are discontinuous. The converse of Theorem

2.1 however is not true (see Example 8 below).[5] A special case of quasiconcave utility is of course linear utility as in the case of expected utility for choice under risk.

Example 7 (*Random expected utility*)**.** *Each alternative corresponds to a lottery $x \in \mathbb{R}^k_+$ over k prizes where $\sum_j x_j = 1$. Each agent i is an expected utility maximizer with utility*

$$u_i(x) = \sum_j x_j \tilde{u}_{ij}$$

where $\tilde{u}_{ij}$ is agent i's Bernoulli utility of prize $j \in \{1, \dots, k\}$. This is the random expected utility model of Gul and Pesendorfer (2006). Since $u_i(x) \geq u_i(\lambda x + (1-\lambda) y)$ iff $u_i(x) \geq u_i(y)$, convex substitutability is in fact satisfied with equality.

While convex substitutability is satisfied by many commonly used pure characteristic models, we now show that it cannot be accommodated by almost all mixed logit models. The only exception is when stochastic choice is *uniform*, i.e. $\rho_A(x) = 1/|A|$ for all A.

Theorem 2.2. *Any non-uniform mixed logit ρ violates convex substitutability.*

Proof. See Appendix. □

To understand the reasoning behind Theorem 2.2, consider some menu A with n products and assume that $[x, y] \cap A = \emptyset$ for all distinct $x, y \in A$. Fix some $x \in A$ so by applying convex substitutability repeatedly, we have

$$\rho_A(x) \geq \rho_{\lambda x + (1-\lambda)A}(x) = \int_U \frac{e^{v(x)}}{\sum_{y \in A} e^{v(\lambda x + (1-\lambda) y)}} d\nu$$

for all $\lambda \in (0, 1)$. Taking the limit as $\lambda \to 1$, we have

$$\rho_A(x) \geq \int_U \frac{e^{v(x)}}{n e^{v(x)}} d\nu = \frac{1}{n}$$

Note that this is true for all $x \in A$. If the inequality is strict for some $x \in A$, then

$$1 = \sum_{y \in A} \rho_A(y) > n\left(\frac{1}{n}\right) = 1$$

yielding a contradiction. The proof shows that this argument holds for all menus so ρ must be uniform choice.

[5] This suggests that convex substitutability is weaker than other related convexity notions in Apesteguia, Ballester and Lu (2017) and Lu (2019).

Convex substitutability can be attractive for various reasons. Descriptively, convex substitutability has certain appeal as it captures the intuitive notion that similar products crowd out demand. Normatively, if one believes that the Hotelling model for instance is the true model, then agents should exhibit such behavior. The above result shows that barring uniform choice, no mixed logit model can accommodate this condition. This is a novel restriction on allowable patterns of choice behavior given mixed logit.

What are the practical implications? Consider a researcher who is using mixed logit models to approximate the Hotelling model. Theorem 2.2 says that all models used for estimation will violate a property that is natural in the Hotelling model. How significant is this violation? On the one hand, this discrepancy will eventually vanish as approximations get arbitrarily close. On the other hand, any mixed logit that eventually emerges from estimation will violate convex substitutability. This would complicate counterfactuals about what would happen to demand as alternatives become more or less similar. Ultimately, the significance and magnitude of this violation will depend on the specific application. In Appendix I, we show via a simulated example that such violations can be severe (up to 20% of all possible menus). In general, Theorem 2.2 highlights a potential issue to take into account when using mixed logit approximations.

Violations of convex substitutability extend more generally to models beyond mixed logit. In fact, any model with additive iid error terms would have difficulty satisfying the condition. To illustrate, consider a model where $X \subset \mathbb{R}$ and $\varepsilon(\cdot)$ represents iid error terms. The probability of choosing x over y is

$$\begin{aligned}\rho(x,y) &= \mathbb{P}\{v(x)+\varepsilon(x) \geq v(y)+\varepsilon(y)\} \\ &= \mathbb{P}\{Z \leq v(x)-v(y)\} \\ &= F(v(x)-v(y))\end{aligned}$$

where Z is the distribution of $\varepsilon(y)-\varepsilon(x)$ with cdf F. This is known as a Fechnarian model in the literature (Debreu (1958), Davidson and Marschak (1959)). Assuming differentiability,

$$\begin{aligned}\frac{\partial \rho(x,y)}{\partial y} &= -f(v(x)-v(y))\,v'(y) \\ \frac{\partial \rho(x,y)}{\partial x} &= f(v(x)-v(y))\,v'(x)\end{aligned}$$

Convex substitutability implies that for all $x < y$, both derivatives must be positive. This implies that $v'(x) \geq 0 \geq v'(y)$ for all $x < y$ which holds only when $v(\cdot)$ is constant and ρ is uniform.

Can convex substitutability be satisfied when error terms are not iid? The following illustrates a model with correlated errors that satisfies convex substitutability.

Example 8 (*Correlated errors*)**.** *Let $X = [-1, 1] \subset \mathbb{R}$ and consider a pure characteristic model where*

$$u(x) = -x^2 + \varepsilon(x)$$

The mean-zero errors terms are given by $\varepsilon(x) = \eta(x - \theta)^2$ with $\theta \sim N(0, \sigma^2)$ and η is either $\kappa > 0$ or $-\kappa < 0$ with equal probability. We will show that convex substitutability can be satisfied. Let $A = \{x_1, \dots, x_n\}$ where $x_1 < \cdots < x_n$. Note that $u(x_i) \geq u(x_j)$ iff $2\eta\theta \geq (\eta - 1)(x_i + x_j)$ for any $i < j$. This implies that

$$\rho_A(x_i) = \frac{1}{2\sigma\sqrt{2\pi}} \left(\int_{\sup_{j>i} \frac{\kappa-1}{2\kappa}(x_i+x_j)}^{\inf_{j<i} \frac{\kappa-1}{2\kappa}(x_i+x_j)} e^{-\frac{\theta^2}{2\sigma^2}} d\theta + \int_{\sup_{j<i} \frac{\kappa+1}{2\kappa}(x_i+x_j)}^{\inf_{j>i} \frac{\kappa+1}{2\kappa}(x_i+x_j)} e^{-\frac{\theta^2}{2\sigma^2}} d\theta \right)$$

First, suppose $\kappa \leq 1$. In this case, it is easy to see that both integrals increase when x_{i+1} increases so $\frac{\partial \rho_A(x_i)}{\partial x_{i+1}} \geq 0$. Now, suppose $\kappa > 1$. Note that if $i + 1 < n$, then when x_{i+1} increases, the first integral is unchanged while the second integral increases so $\frac{\partial \rho_A(x_i)}{\partial x_{i+1}} \geq 0$. Finally, consider the case where $i + 1 = n$. Note that $\frac{\partial \rho_A(x_i)}{\partial x_{i+1}} \geq 0$ if

$$e^{-\frac{1}{2}\left(\frac{\kappa+1}{2\sigma\kappa}(x_i+x_{i+1})\right)^2}(\kappa+1) \geq e^{-\frac{1}{2}\left(\frac{\kappa-1}{2\sigma\kappa}(x_i+x_{i+1})\right)^2}(\kappa-1)$$
$$-\frac{1}{2\kappa\sigma^2}(x_i + x_{i+1})^2 \geq \log\left(\frac{\kappa-1}{\kappa+1}\right)$$

This is satisfied as long as $\sigma^2 \geq 2\left(\kappa \log \frac{\kappa+1}{\kappa-1}\right)^{-1}$ since $x_i, x_{i+1} \in X = [-1, 1]$. The case for decreasing x_{i-1} is symmetric so convex substitutability is satisfied. Note that since $u(x)$ may be not be quasiconcave when $\kappa > 1$, convex substitutability does not guarantee utilities are quasiconcave.

In the above model, error terms are correlated depending on how similar alternatives are in the space of characteristics. By decreasing the variance of these error terms (e.g. smaller κ), one could approximate the Hotelling model with arbitrary precision. In contrast to mixed logit models, convex substitutability would be satisfied along the entire path of approximation with this class of models.

4.2 Continuity in Characteristics

This section introduces a continuity condition that is satisfied by all pure characteristic models but violated by all mixed logit models. To illustrate, first consider the classic red-bus/blue-bus example.

Example 9 (*Red-bus/blue-bus*)**.** *Consider the choice of transportation alternatives and let x correspond to traveling by car while y correspond to traveling by a red bus. Suppose car and red bus both have equal market share so $\rho\left(x,y\right)=\frac{1}{2}$. Consider introducing a blue bus y'. Supposing agents are indifferent to color, $\rho\left(y,y'\right)=\frac{1}{2}$ so Luce's independence of irrelevant alternatives (IIA) condition implies that $\rho_{\{x,y,y'\}}\left(x\right)=\frac{1}{3}$. In reality, one would expect the car market share to remain close to* 50% *in violation of IIA.*

The red-bus/blue-bus example is an example of the classic "duplicates problem" as identified by Debreu (1960) and illustrates a limitation of logit. Mixed logit models are not bound by IIA and can accommodate such choice patterns. However, we now present a new variant of the "duplicates problem" that no mixed logit model can accommodate. Suppose we introduce buses y_n with colors that are increasingly closer to red. Eventually, y_n will be indistinguishable from y, so the car market share should approach 50%. In other words,

$$\rho_{\{x,y,y_n\}}\left(x\right)\to\rho\left(x,y\right)$$

We generalize this continuity condition as follows.

Definition 7. ρ satisfies *continuity in characteristics* if $\rho_{A\cup\{y_n\}}\left(x\right)\to\rho_A\left(x\right)$ for all $y_n\to y\in A\setminus\{x\}$.

Intuitively, as alternative y_n becomes increasingly similar to alternative y, the two alternatives will eventually be indistinguishable. When evaluating choice, one can replace two indistinguishable alternatives with the single alternative y in the limit.

In a pure characteristic model, the probability of choosing x over y is given by the probability that the utility of x is greater than that of y. Since utilities are continuous, the utility of y_n will converge to the utility of y as y_n converges to y. As a result, continuity in characteristics will be satisfied.

Theorem 3.1. *Any pure characteristic ρ satisfies continuity in characteristics.*

Proof. See Appendix. □

This applies to the Hotelling model in Example 6 and the random expected utility model in Example 7. It would also apply in models where the set of characteristics is not convex (e.g. Salop (1979)) as long as utilities are continuous in product characteristics. While continuity in characteristics is satisfied by all pure characteristic models, it cannot be accommodated by any mixed logit model.

Theorem 3.2. *Any mixed logit ρ violates continuity in characteristics.*

Proof. Consider distinct $x, y \in A$ and a sequence $y_n \to y$. Let $A_n = A \cup \{y_n\}$. Since ρ is a mixed logit,

$$\begin{aligned} \lim_n \rho_{A_n}(x) &= \lim_n \int_U \frac{e^{v(x)}}{\sum_{z \in A} e^{v(z)} + e^{v(y_n)}} d\nu \\ &= \int_U \frac{e^{v(x)}}{\sum_{z \in A} e^{v(z)} + e^{v(y)}} d\nu \\ &< \int_U \frac{e^{v(x)}}{\sum_{z \in A} e^{v(z)}} d\nu = \rho_A(x) \end{aligned}$$

Thus, $\lim_n \rho_{A_n}(x) < \rho_A(x)$ for any such A and $y_n \to y$. □

The above proof illustrates that any mixed logit model not only violates continuity in characteristics but it violates it in a specific direction. When y_n converges to y, the market share of x in $\{x, y, y_n\}$ converges to a limit that is strictly less than its market share in $\{x, y\}$. The reason is that logit errors force individual market shares for alternatives no matter how similar they are to each other. This is related to well-known limitations of mixed logit models (e.g. Berry and Pakes (2007)) and Theorem 3.2 formalizes such intuition.

The above results imply the following corollary.

Corollary 1. *No mixed logit model is pure characteristic.*

Proof. Follows from Theorem 3. □

Mixed logit and pure characteristic models belong to two very different classes of models; Corollary 1 shows that they have empty intersection. Although Theorem 1 guarantees that a researcher can always approximate pure characteristic models using mixed logit models, this approximation will always be from "outside" the set of pure characteristic models.

What does this mean in practice? Like the results from Section 4.1, the magnitude of these issues depend on the application at hand. Note that the model with correlated errors in Example 8 is a pure characteristic model and thus satisfies continuity in characteristics.

In fact, as long as error terms are continuous in characteristics (e.g. Brownian noise), continuity in characteristics will be satisfied. Theorem 3 illustrate one behavioral condition that separates the two class of models but there may be other differences that would have non-trivial implications for estimation and counterfactual analysis. Ultimately, when deciding whether to use one class of models versus another for estimation, one would need to weigh the importance of these choice patterns versus the burden of computational costs.

Appendix

A Preliminaries

Define the space

$$\mathcal{V} := \prod_{A \in \mathcal{A}} \mathbb{R}^A$$

and note that $\mathcal{P} \subset \mathcal{V}$. We endow $\mathcal{V}$ also with the product topology. Note that $\mathcal{P}$ is compact by Tychonoff's theorem. Since $\mathcal{V}$ is a Hausdorff space (Theorem 19.4 of Munkres (2000)), $\mathcal{P}$ is also closed (Lemma 2.32 of Aliprantis and Border (2006), henceforth AB). Although the space $\mathcal{V}$ is neither metrizable or even first-countable, the next lemma shows that it is locally convex which allows us to use separating hyperplane theorems.

Lemma 1. *$\mathcal{V}$ is a locally convex topological vector space.*

Proof. Since $\mathbb{R}^A$ is a topological vector space for every $A \in \mathcal{A}$, $\mathcal{V}$ is a topological vector space by Theorem 5.2 of AB. Consider the family of seminorms $(r_A)_{A \in \mathcal{A}}$ where $r_A : \mathcal{V} \to \mathbb{R}$ is such that

$$r_A(\tau) = |\tau_A|$$

where $|\cdot|$ is the Euclidean norm in $\mathbb{R}^A$. Since this family of seminorms generates the product topology, $\mathcal{V}$ is locally convex. □

Throughout this appendix, let $m = m(k, d)$, and for every $x \in X$, let x^* denote the corresponding vector in polynomial space $X^* \subset \mathbb{R}^m$. We define the following subsets of $\mathcal{P}$:

- $\mathcal{P}^{pc}$ is the set of pure characteristic and $\mathcal{P}_d^{pc}$ is the set of pure characteristic of degree d
- $\mathcal{P}^{log}$ is the set of logit and $\mathcal{P}_d^{log}$ is the set of logit of degree d
- $\mathcal{P}^{mlog}$ is the set of mixed logit and $\mathcal{P}_d^{mlog}$ is the set of mixed logit of degree d
- $\mathcal{P}_d^{lex}$ is the set of lexicographic-logit of degree d
- $\mathcal{P}_d^{mlex}$ is the set of mixed lexicographic-logit of degree d

Also let $cl(\mathcal{P}')$ denote the closure of any subset $\mathcal{P}' \subset \mathcal{P}$. Note that since $\mathcal{P}$ is closed, $cl(\mathcal{P}') \subset \mathcal{P}$ is compact.

B Proof of Proposition 1

We first prove the following lemma.

Lemma 2. *For $A \in \mathcal{A}$ and $u : X \to \mathbb{R}$ such that $u(x) \neq u(y)$ for all distinct $x, y \in A$,*

$$\lim_n \frac{e^{nu(x)}}{\sum_{y \in A} e^{nu(y)}} = 1\left\{u(x) > u(y) \ \textit{for all } y \in A\right\}$$

Proof. Since $u(x) \neq u(y)$ for all distinct $x, y \in A$, we can rewrite

$$\frac{e^{nu(x)}}{\sum_{y \in A} e^{nu(y)}} = \frac{1}{1 + \sum_{y \in A \backslash x} e^{n(u(y)-u(x))}}$$

Consider $n \to \infty$. First, suppose $u(x) > u(y)$ for all $y \in A \setminus x$. In this case, $e^{n(u(y)-u(x))} \to 0$ for all $y \in A \backslash x$ so the expression above converges to 1. On the other hand, suppose there exists some $y \in A$ such that $u(y) > u(x)$. In this case, $e^{n(u(y)-u(x))} \to \infty$ so the expression converges to 0. The result follows. □

We now prove Proposition 1. Let $\rho \in \mathcal{P}^{pc}$ with distribution μ and define

$$\rho_A^n(x) = \int_U \frac{e^{nu(x)}}{\sum_{y \in A} e^{nu(y)}} d\mu$$

so $\rho^n \in \mathcal{P}^{mlog}$. We will show that $\rho^n \to \rho$. Fix some $A \in \mathcal{A}$ and define

$$U_A = \left\{u \in U \,:\, u(x) \neq u(y) \ \text{ for all distinct } x, y \in A\right\}$$

Since ties never occur for random utility models, $\mu(U_A) = 1$. Now, by Lemma 2 and dominated convergence,

$$\begin{aligned}
\lim_n \rho_A^n(x) &= \lim_n \int_U \frac{e^{nu(x)}}{\sum_{y \in A} e^{nu(y)}} d\mu = \lim_n \int_{U_A} \frac{e^{nu(x)}}{\sum_{y \in A} e^{nu(y)}} d\mu \\
&= \int_{U_A} \lim_n \frac{e^{nu(x)}}{\sum_{y \in A} e^{nu(y)}} d\mu = \int_{U_A} 1\left\{u(x) > u(y) \ \text{ for all } y \in A\right\} d\mu \\
&= \mu\left(\left\{u \in U_A \,:\, u(x) > u(y) \ \text{ for all } y \in A\right\}\right) \\
&= \mu\left(\left\{u \in U \,:\, u(x) \geq u(y) \ \text{ for all } y \in A\right\}\right) = \rho_A(x)
\end{aligned}$$

as desired.

C Proof of Theorem 3.1

First, we define continuity for stochastic choice. We endow $\mathcal{A}$ with the Hausdorff metric.

Definition 8. ρ satisfies *continuity* if $\rho_{A_k} \to \rho_A$ for all $A_k \to A$.

We now prove a stronger version of Theorem 3.1 below.

Theorem 3.1*. *Any pure characteristic ρ satisfies continuity.*

Let $\rho \in \mathcal{P}^{pc}$ with distribution μ. Consider $A_k \to A$. Note that $u(x) = u(y)$ with μ-measure zero. Now, define

$$I := \bigcup_{\{x,y\} \subset A_k \cup A} \{u \in U : u(x) = u(y)\}$$

which is measurable and $\mu(I) = 0$. Let $U^* := U \backslash I$ so $\mu(U^*) = 1$. Let μ^* be the restriction of μ on U^*.

We will now define random variables $\xi_k : U^* \to X$ and $\xi : U^* \to X$ that have distributions ρ_{A_k} and ρ_A respectively. For each A_k, let $\xi_k : U^* \to X$ be such that

$$\xi_k(u) := \arg\max_{x \in A_k} u(x)$$

and define ξ similarly for A. Note that these are well-defined because there exists a unique maximizer for every $u \in U^*$. Now, for any measurable set $E \subset X$,

$$\begin{aligned} \xi_k^{-1}(E) &= \{u \in U^* : \xi_k(u) \in E \cap A_k\} \\ &= \bigcup_{x \in E \cap A_k} \{u \in U^* : u(x) > u(y) \text{ for all } y \in A_k\} \end{aligned}$$

which is measurable. Hence, ξ_k and ξ are random variables. Note that

$$\begin{aligned} \mu^*\left(\xi_k^{-1}(E)\right) &= \sum_{x \in E \cap A_k} \mu^* \{u \in U^* : u(x) > u(y) \text{ for all } y \in A_k\} \\ &= \sum_{x \in E \cap A_k} \mu \{u \in U^* : u(x) > u(y) \text{ for all } y \in A_k\} \\ &= \rho_{A_k}(E \cap A_k) = \rho_{A_k}(E) \end{aligned}$$

so ρ_{A_k} and ρ_A are the distributions of ξ_k and ξ respectively. Since every $u \in U^* \subset U$ is continuous, by the Maximum Theorem, $\xi_k(u) = \arg\max_{x \in A_k} u(x)$ is upper hemicontinuous in A_k and thus continuous as ξ_k is singleton-valued. Since $A_k \to A$, $\xi_k \to \xi$ μ^*-a.s. and since a.s. convergence implies convergence in distribution, $\rho_{A_k} \to \rho_A$ as desired.

D Proof of Proposition 2

Note that since $\mathcal{P}$ is not metrizable under the product topology, it may not be sequential. However, the following result shows shows that the sequential limit points of logit still coincide with the closure of logit.

Proposition 2*. *The following are equivalent:*

(1) $\rho \in \mathcal{P}_d^{lex}$

(2) *there exists a sequence* $\rho^n \in \mathcal{P}_d^{log}$ *such that* $\rho^n \to \rho$.

(3) *there exists a net* $\rho^a \in \mathcal{P}_d^{log}$ *such that* $\rho^a \to \rho$.

(4) $\rho \in cl\left(\mathcal{P}_d^{log}\right)$

The equivalence of (3) and (4) is standard (Theorem 2.14 of AB). Since every sequence is also a net, (2) implies (3). We will show that (3) implies (1) implies (2). This will then establish Proposition 2. First, we prove the following useful lemma.

Lemma 3. *Consider any net* $\{\beta_a\}_{a\in D}$ *where* $\beta_a \in \mathbb{R}^m$.

(1) *If* $\limsup_a |\beta_a| < \infty$, *then there exists a subnet* β_i *such that* $\beta_i \to \beta$.

(2) *If* $\limsup_a |\beta_a| = \infty$, *then there exists a subsequence* β_i *such that* $\frac{\beta_i}{|\beta_i|} \to \gamma \neq 0$. *Moreover, for any* $z \in \mathbb{R}^m$, *(i)* $\gamma \cdot z > 0$ *implies* $\lim_i \beta_i \cdot z = \infty$, *and (ii)* $\gamma \cdot z < 0$ *implies* $\lim_i \beta_i \cdot z = -\infty$

Proof. Let $\geqslant$ be the preorder associated with D. First consider (1). Since

$$\infty > \limsup_a |\beta_a| = \inf_a \sup_{b\geqslant a} |\beta_b|$$

there must exist some $a^* \in D$ such that $\sup_{b\geqslant a^*} |\beta_b| < \infty$. Thus, $\{\beta_b\}_{b\geqslant a^*}$ is a bounded net in $\mathbb{R}^m$ so it must have a convergent subnet $\beta_i \to \beta$. Since $\{\beta_b\}_{b\geqslant a^*}$ is a subnet of β_a, β_i is also subnet of β_a as desired.

Now, consider (2). First, we show that there exists a subsequence β_j such that $|\beta_j| \to \infty$. Since $\infty = \limsup_a |\beta_a| = \inf_a \sup_{b\geqslant a} |\beta_b|$, it must be that $\sup_{b\geqslant a} |\beta_b| = \infty$ for all $a \in D$. Fix some $a_0 \in D$ and note that for any $j \in \mathbb{N}$, we can find some $a_1 \geqslant a_0$ such that $|\beta_{a_1}| > j$.

By induction, we can create a sequence a_j such that $\left|\beta_{a_j}\right| > j$. Moreover, since $a_{j+1} \geqslant a_j$, $\beta_j = \beta_{a_j}$ is a subnet of β_a and $|\beta_j| \to \infty$.

Let $S \subset \mathbb{R}^m$ be the unit sphere and let $\hat{\beta}_j = \frac{\beta_j}{|\beta_j|} \in S$ be the normalized unit vector. Since S is compact, there must exist a convergent subsequence β_i such that $\hat{\beta}_i \to \gamma \in S$ as desired. Since $|\beta_i| \to \infty$, if $\gamma \cdot z > 0$, then

$$\lim_i \beta_i \cdot z = \lim_i |\beta_i| \left(\hat{\beta}_i \cdot z\right) = \infty$$

The case for $\gamma \cdot z < 0$ is symmetric. □

We now prove (3) implies (1) in Proposition 2*. Let $\rho^a \in \mathcal{P}_d^{log}$ with $\beta_a \in \mathbb{R}^m$ and $\rho^a \to \rho$. Now,

$$\rho_A(x) = \lim_a \rho_A^a(x) = \lim_a \frac{e^{\beta_a \cdot x^*}}{\sum_{y \in A} e^{\beta_a \cdot y^*}} = \left(\sum_{y \in A} e^{\lim_a \beta_a \cdot (y^* - x^*)}\right)^{-1}$$

Since $\rho(y, x)$ is well-defined, so is $\lim_a \beta_a \cdot (y^* - x^*)$ on $\bar{\mathbb{R}}$, the extended real line. Let

$$Z := \{y^* - x^* \,:\, x, y \in X\} \subset \mathbb{R}^m$$

First, suppose $\limsup_a |\beta_a| < \infty$ so by Lemma 3, there exists a convergent subnet $\beta_i \to \beta \in \mathbb{R}^m$. Thus,

$$\rho_A(x) = \frac{e^{\beta \cdot x^*}}{\sum_{y \in A} e^{\beta \cdot y^*}} = \frac{e^{u(x)}}{\sum_{y \in A} e^{u(y)}}$$

where $u \in U_d$. This means that $\rho \in \mathcal{P}_d^{log} \subset \mathcal{P}_d^{lex}$ as desired.

Now, suppose $\limsup_a |\beta_a| = \infty$ so by Lemma 3, there exists a convergent subsequence $\hat{\beta}_i := \frac{\beta_i}{|\beta_i|} \to \gamma_1 \in \mathbb{R}^m$. Moreover, for any $z \in Z$, $\lim_n \beta_n \cdot z = \lim_i \beta_i \cdot z = \infty$ if $\gamma_1 \cdot z > 0$ and $\lim_n \beta_n \cdot z = -\infty$ if $\gamma_1 \cdot z < 0$. Let $H_1 \subset \mathbb{R}^m$ denote the $(m-1)$-dimensional hyperplane such that $\gamma_1 \cdot z = 0$ for all $z \in \mathbb{R}^m$. Thus

$$\rho_A(x) = 1\{\gamma_1 \cdot x^* \geq \gamma_1 \cdot y^* \text{ for all } y \in A\} \left(\sum_{y \in A, y^* - x^* \in H_1} e^{\lim_i \beta_i \cdot (y^* - x^*)}\right)^{-1}$$

Let $T_1 : \mathbb{R}^m \to H_1$ be the projection mapping onto H_1, that is

$$T_1(\beta) := \beta - (\beta \cdot \gamma_1)\gamma_1$$

Now, for any $z \in H_1$,

$$\beta \cdot z = \left(T_1\left(\beta\right) + \left(\beta \cdot \gamma_1\right)\gamma_1\right) \cdot z = T_1\left(\beta\right) \cdot z$$

We thus have

$$\rho_A\left(x\right) = 1\left\{\gamma_1 \cdot x^* \geq \gamma_1 \cdot y^* \text{ for all } y \in A\right\}\left(\sum_{y \in A, y^* - x^* \in H_1} e^{\lim_i T_1\left(\beta_i\right) \cdot \left(y^* - x^*\right)}\right)^{-1}$$

Now, for any $y^* - x^* \in H_1$, so we can repeat the same arguments as above. If $\limsup_i \left|T_1\left(\beta_i\right)\right| < \infty$, then by Lemma 3, we can assume $T_1\left(\beta_i\right) \to \beta \in \mathbb{R}^m$ and

$$\rho_A\left(x\right) = 1\left\{\gamma_1 \cdot x^* \geq \gamma_1 \cdot y^* \text{ for all } y \in A\right\}\frac{e^{\beta \cdot x^*}}{\sum_{y \in A, \gamma_1 \cdot y^* = \gamma_1 \cdot x^*} e^{\beta \cdot y^*}}$$

On the other hand, if $\limsup_i \left|T_1\left(\beta_i\right)\right| = \infty$, then by Lemma 3, we can assume $\frac{T_1\left(\beta_i\right)}{\left|T_1\left(\beta_i\right)\right|} \to \gamma_2 \in H_1$. Let $H_2 \subset \mathbb{R}^m$ denote the $(m-2)$-dimensional hyperplane such that $\gamma_1 \cdot z = \gamma_2 \cdot z = 0$ for all $z \in \mathbb{R}^m$ and note that $\gamma_1 \cdot \gamma_2 = 0$. If we let $T_2 : \mathbb{R}^m \to H_2$ be the projection mapping onto H_2, then by the same arguments as above,

$$\rho_A\left(x\right) = 1\left\{x \succeq_{\left(\gamma_1, \gamma_2\right)} y \text{ for all } y \in A\right\}\left(\sum_{y \in A, y^* - x^* \in H_2} e^{\lim_i T_2\left(\beta_i\right) \cdot \left(y^* - x^*\right)}\right)^{-1}$$

where $\succeq_{\left(\gamma_1, \gamma_2\right)}$ is the lexicographic preference induced by γ_1 and γ_2.

We can continue this argument by induction, and since m is finite, we can find a sequence $\left(\gamma_1, \dots, \gamma_t, \beta\right)$ such that

$$\begin{aligned}\rho_A\left(x\right) &= 1\left\{x \succeq_{\left(\gamma_1, \dots, \gamma_t\right)} y \text{ for all } y \in A\right\}\frac{e^{\beta \cdot x^*}}{\sum_{y \in A, \gamma_j \cdot y^* = \gamma_j \cdot x^*, j \in \{1, \dots, t\}} e^{\beta \cdot y^*}} \\ &= 1\left\{x \succeq_{\left(v_1, \dots, v_t\right)} y \text{ for all } y \in A\right\}\frac{e^{u(x)}}{\sum_{y \in A, v_j(y) = v_j(x), j \in \{1, \dots, t\}} e^{u(y)}}\end{aligned}$$

for $v_1, \dots, v_t, u \in U_d$ where $v_1, \dots, v_t$ are all orthogonal. This means that $\rho \in \mathcal{P}_d^{lex}$ as desired.

Finally, we prove (1) implies (2) in Proposition 2^*. Suppose $\rho \in \mathcal{P}_d^{lex}$ with $\omega = \left(v_1, \dots, v_t\right)$ for $t \leq m$ and $u \in U_d$. Let β^* be the polynomial vector corresponding to u and $\gamma_1, \dots, \gamma_t \in \mathbb{R}^m$ be the polynomial vectors corresponding to ω which are orthogonal. Without loss of generality, we can assume that $\gamma_i \neq 0$. By a change of basis, we can also assume without loss that they correspond to the standard basis in $\mathbb{R}^m$. Now, define

$$\beta_n = \left(n^k, n^{k-1} \dots, n^2, n\beta_t^*, n\beta_{t+1}^*, \dots, n\beta_m^*\right)$$

Let ρ^n be the logit corresponding to β_n so $\rho^n \in \mathcal{P}_d^{log}$ and it is straightforward to see that $\rho^n \to \rho$.

E Proof of Proposition 3

Let $\bar{\mathcal{P}}_d^{lex}$ denote the closed convex hull of $\mathcal{P}_d^{lex}$, that is

$$\bar{\mathcal{P}}_d^{lex} := cl\left(co\left(\mathcal{P}_d^{lex}\right)\right) \subset \mathcal{P}$$

which is compact. We first show the following which will imply that $\mathcal{P}_d^{mlex}$ is closed.

Lemma 4. $\bar{\mathcal{P}}_d^{lex} = \mathcal{P}_d^{mlex}$

Proof. Since $co\left(\mathcal{P}_d^{lex}\right)$ is convex, $\bar{\mathcal{P}}_d^{lex}$ is also convex (Lemma 5.27 of AB). We first show that $\mathcal{P}_d^{mlex} \subset \bar{\mathcal{P}}_d^{lex}$. Let $\rho \in \mathcal{P}_d^{mlex}$ so there exists a distribution ν on $\Omega_d \times U_d$ such that

$$\rho = \int_{\Omega_d \times U_d} \rho_{(\omega,u)} d\nu$$

where $\rho_{(\omega,u)} \in \mathcal{P}_d^{lex}$ is the lexicographic-logit stochastic choice corresponding to $(\omega, u) \in \Omega_d \times U_d$. Suppose $\rho \notin \bar{\mathcal{P}}_d^{lex}$. Since $\mathcal{V}$ is locally convex (Lemma 1), continuous linear functionals separates points in $\mathcal{V}$. Thus, we can apply the strict separating hyperplane theorem (Theorem 3.5 of Rudin (1991)) and find a continuous linear functional Λ such that for all $\tau \in \bar{\mathcal{P}}_d^{lex}$,

$$\Lambda(\rho) = 1 > 0 = \Lambda(\tau)$$

Now,

$$1 = \Lambda(\rho) = \Lambda\left(\int_{\Omega_d \times U_d} \rho_{(\omega,u)} d\nu\right) = \int_{\Omega_d \times U_d} \Lambda\left(\rho_{(\omega,u)}\right) d\nu = 0$$

as $\rho_{(\omega,u)} \in \bar{\mathcal{P}}_d^{lex}$. This yields a contradiction so $\mathcal{P}_d^{mlex} \subset \bar{\mathcal{P}}_d^{lex}$.

Next, we show that $\bar{\mathcal{P}}_d^{lex} \subset \mathcal{P}_d^{mlex}$. Fix some $\rho \in \bar{\mathcal{P}}_d^{lex}$. Since $\mathcal{P}_d^{lex} = cl\left(\mathcal{P}_d^{log}\right)$ by Proposition 2, it is closed and thus compact. By Theorem 3.28 of Rudin (1991), there exists a Borel probability measure π on $\mathcal{P}_d^{lex}$ such that

$$\rho = \int_{\mathcal{P}_d^{lex}} \tau \; d\pi$$

We now show ρ must be mixed lexicographic-logit. Consider the mapping $\varphi : \Omega_d \times U_d \to \mathcal{P}_d^{lex}$ such that $\varphi = \rho_{(\omega,u)}$. Let $\mathcal{G}$ be be the σ-algebra on $\Omega_d \times U_d$ generated by φ. We can thus define a measure ν on $\mathcal{G}$ such that

$$\pi = \nu \circ \varphi^{-1}$$

Thus, by a change of variables (Theorem 13.46 of AB),

$$\rho = \int_{\mathcal{P}_d^{lex}} \tau \; d\pi = \int_{\Omega_d \times U_d} \varphi\left(\omega, u\right) d\nu = \int_{\Omega_d \times U_d} \rho_{(\omega,u)} d\nu$$

so $\rho \in \mathcal{P}_d^{mlex}$ as desired. □

We now prove Proposition 3. Since $\mathcal{P}_d^{log} \subset \mathcal{P}_d^{lex}$, we have that $\mathcal{P}_d^{mlog} \subset \mathcal{P}_d^{mlex}$. Thus,

$$cl\left(\mathcal{P}_d^{mlog}\right) \subset cl\left(\mathcal{P}_d^{mlex}\right) = \mathcal{P}_d^{mlex}$$

where the last equality follows from Lemma 4. Now, consider $\rho \in \mathcal{P}_d^{mlex}$ and suppose $\rho \notin cl\left(\mathcal{P}_d^{mlog}\right)$. Applying the strict separating hyperplane theorem again, there exists a continuous linear functional Λ such that for all $\tau \in cl\left(\mathcal{P}_d^{mlog}\right)$,

$$\Lambda\left(\rho\right) = 1 > 0 = \Lambda\left(\tau\right)$$

Now

$$1 = \Lambda\left(\rho\right) = \Lambda\left(\int_{\Omega_d \times U_d} \rho_{(\omega,u)} d\nu\right) = \int_{\Omega_d \times U_d} \Lambda\left(\rho_{(\omega,u)}\right) d\nu = 0$$

where the last equality follows from the fact that $\rho_{(\omega,u)} \in \mathcal{P}_d^{lex} = cl\left(\mathcal{P}_d^{log}\right) \subset cl\left(\mathcal{P}_d^{mlog}\right)$. Thus, we have a contradiction. This shows that $cl\left(\mathcal{P}_d^{mlog}\right) = \mathcal{P}_d^{mlex}$ as desired.

F Proof of Theorem 1

We first prove a technical lemma.

Lemma 5. *Let $x_n = \left(1 - \frac{1}{n}\right) x + \frac{1}{n} y$ for $x, y \in X$. For any $\omega \in \Omega_d$, exactly one of the following holds*

(1) $\lim_n 1\left\{y \succ_\omega x_n\right\} = 1$

(2) $\lim_n 1\left\{y \sim_\omega x_n\right\} = 1$

(3) $\lim_n 1\left\{y \prec_\omega x_n\right\} = 1$

Proof. For any $v \in U_d$, let

$$g\left(\alpha\right) = v\left(y\right) - v\left(\left(1 - \alpha\right) x + \alpha y\right)$$

and note that g is a polynomial. Since any non-zero polynomial has a finite number of roots, it means that we can find some N such that either (i) $g\left(\frac{1}{n}\right) > 0$ for all $n > N$, (ii) $g\left(\frac{1}{n}\right) = 0$ for all $n > N$, or (iii) $g\left(\frac{1}{n}\right) < 0$ for all $n > N$. Let $U_d^{\succ}$, $U_d^{\sim}$ and $U_d^{\prec}$ be the partition of U_d corresponding to these three conditions. It is straightforward to see that for $\omega = (v_1, \ldots, v_t) \in \Omega_d$,

(1) If $v_i \in U_d^{\sim}$ for all $0 \leq i < j \leq t$ and $v_j \in U_d^{\succ}$, then $\lim_n 1\left\{y \succ_\omega x_n\right\} = 1$.

(2) If $v_i \in U_d^{\sim}$ for all $1 \leq i \leq t$, then $\lim_n 1\left\{y \sim_\omega x_n\right\} = 1$.

(3) Otherwise, $\lim_n 1\left\{y \prec_\omega x_n\right\} = 1$

The result follows. □

We now prove Theorem 1. Let $\rho \in \mathcal{P}^{pc}$ and suppose $\rho \in cl\left(\mathcal{P}_d^{mlog}\right)$. By Proposition 3, $\rho \in \mathcal{P}_d^{mlex}$ so there exists some distribution ν on $\Omega_d \times U_d$ such that

$$\rho_A(x) = \int_{\Omega_d \times U_d} 1\left\{x \succeq_\omega y \text{ for all } y \in A\right\} \frac{e^{u(x)}}{\sum_{y \in A, y \sim_\omega x} e^{u(y)}} d\nu$$

We will show that any distinct $x, y \in X$, $x \sim_\omega y$ with ν-measure zero. Let $x_n = \left(1 - \frac{1}{n}\right) x + \frac{1}{n} y$ and define the following sets

$$\begin{aligned}
\Omega_1 &= \left\{\omega \in \Omega_d : y \succ_\omega x\right\} & \Omega_1' &= \left\{\omega \in \Omega_d : \lim_n 1\left\{y \succ_\omega x_n\right\} = 1\right\} \\
\Omega_2 &= \left\{\omega \in \Omega_d : y \sim_\omega x\right\} & \Omega_2' &= \left\{\omega \in \Omega_d : \lim_n 1\left\{y \sim_\omega x_n\right\} = 1\right\} \\
\Omega_3 &= \left\{\omega \in \Omega_d : y \prec_\omega x\right\} & \Omega_3' &= \left\{\omega \in \Omega_d : \lim_n 1\left\{y \prec_\omega x_n\right\} = 1\right\}
\end{aligned}$$

Note that $\{\Omega_1, \Omega_2, \Omega_3\}$ and $\{\Omega_1', \Omega_2', \Omega_3'\}$ are both partitions of Ω_d, where the latter follows from Lemma 5. Suppose $\omega = (v_1, \ldots, v_t) \in \Omega_2'$, so $v_i(y) = v_i(x_n)$ for sufficiently large n and all $1 \leq i \leq t$. This implies that $v_i(y) = v_i(x)$ for all $1 \leq i \leq t$ or $y \sim_\omega x$. Thus, $\Omega_2' \subset \Omega_2$. This implies that

$$\Omega_1 \cap \Omega_2' = \Omega_3 \cap \Omega_2' = \emptyset \tag{1}$$

Let $A_n = \{y, x, x_n\}$ and note that $A_n \to \{y, x\}$. Since $\rho \in \mathcal{P}^{pc}$, by Theorem 3.1,

$$\rho(y, x) = \lim_n \rho(y, x_n) = \lim_n \rho_{A_n}(y)$$

For ease of notation, we suppress the dependence on U_d; for instance, we let $\nu(\Omega_i)$ denote

$\nu\left(\Omega_i \times U_d\right)$. Now,

$$\begin{aligned}\rho\left(y,x\right) &= \nu\left(\Omega_1\right) + \int_{\Omega_2} \frac{e^{u(y)}}{e^{u(y)}+e^{u(x)}} d\nu \\ &= \nu\left(\Omega_1 \cap \Omega_1'\right) + \nu\left(\Omega_1 \cap \Omega_3'\right) + \int_{\Omega_2} \frac{e^{u(y)}}{e^{u(y)}+e^{u(x)}} d\nu \qquad (2)\end{aligned}$$

where the second equality follows from equation (1). By dominated convergence

$$\begin{aligned}\lim_n \rho\left(y,x_n\right) &= \nu\left(\Omega_1'\right) + \int_{\Omega_2'} \frac{e^{u(y)}}{e^{u(y)}+e^{u(x)}} d\nu \\ &= \nu\left(\Omega_1 \cap \Omega_1'\right) + \nu\left(\Omega_2 \cap \Omega_1'\right) + \nu\left(\Omega_3 \cap \Omega_1'\right) + \int_{\Omega_2 \cap \Omega_2'} \frac{e^{u(y)}}{e^{u(y)}+e^{u(x)}} d\nu \qquad (3)\end{aligned}$$

where the last equality follows from (1) again. Applying dominated convergence again,

$$\lim_n \rho_{A_n}\left(y\right) = \nu\left(\Omega_1 \cap \Omega_1'\right) + \int_{\Omega_2 \cap \Omega_1'} \frac{e^{u(y)}}{e^{u(y)}+e^{u(x)}} d\nu + \int_{\Omega_2 \cap \Omega_2'} \frac{e^{u(y)}}{e^{u(y)}+2e^{u(x)}} d\nu \qquad (4)$$

Subtracting equation (4) from (3), we have

$$0 = \int_{\Omega_2 \cap \Omega_1'} \left(1 - \frac{e^{u(y)}}{e^{u(y)}+e^{u(x)}}\right) d\nu + \nu\left(\Omega_3 \cap \Omega_1'\right) + \int_{\Omega_2 \cap \Omega_2'} \left(\frac{e^{u(y)}}{e^{u(y)}+e^{u(x)}} - \frac{e^{u(y)}}{e^{u(y)}+2e^{u(x)}}\right) d\nu$$

This implies that $\Omega_2 \cap \Omega_1'$, $\Omega_3 \cap \Omega_1'$ and $\Omega_2 \cap \Omega_2'$ are all ν-measure zero sets. Combining equations (2) and (3), we have

$$0 = \nu\left(\Omega_1 \cap \Omega_3'\right) + \int_{\Omega_2} \frac{e^{u(y)}}{e^{u(y)}+e^{u(x)}} d\nu$$

so Ω_2 must also be a ν-measure zero set as desired.

We thus have

$$\begin{aligned}\rho_A\left(x\right) &= \int_{\Omega_d} 1\left\{x \succeq_\omega y \text{ for all } y \in A\right\} d\nu \\ &= \nu\left(\left\{\omega \in \Omega_d : x \succeq_\omega y \text{ for all } y \in A\right\}\right)\end{aligned}$$

Now, for every $\omega = \left(v_1, \ldots, v_t\right) \in \Omega_d$, let $\theta_\omega = \left(\beta_1^\omega, \ldots, \beta_t^\omega\right)$ denote the collection of coefficients such that

$$v_i\left(x\right) = \beta_i^\omega \cdot x^*$$

We can thus extend ρ to a stochastic choice ρ^* in $\mathbb{R}^m$ such that $\rho_A\left(x\right) = \rho_{A^*}^*\left(x^*\right)$ and for any finite $D \subset \mathbb{R}^m$,

$$\rho_D^* (z) = \nu \left(\{ \omega \in \Omega_d : z \succeq_{\theta_\omega} w \text{ for all } w \in D \} \right)$$

Moreover, since for all $x, y \in X$, $x \sim_\omega y$ with ν-measure zero, without loss of generality, we can assume that for all $z, w \in \mathbb{R}^m$, $z \sim_{\theta_\omega} w$ with ν-measure zero as well.

We now show that ρ^* satisfies the Gul and Pesendorfer (2006) axioms. Since $\succeq_{\theta_\omega}$ satisfies independence, ρ^* satisfies linearity and since $y \sim_{\theta_\omega} x$ with ν-measure zero, ρ also satisfies extremeness. Mixture continuity follows from the same argument as Lemma 3 in the Supplement of Gul and Pesendorfer (2006). This means that we can find some finitely-additive μ^* on $\mathbb{R}^m$ such that

$$\rho_D^* (z) = \mu^* \left(\{ \beta \in \mathbb{R}^m : \beta \cdot z \geq \beta \cdot w \text{ for all } w \in D \} \right)$$

Since $\rho_A (x) = \rho_{A^*}^* (x^*)$,

$$\begin{aligned} \rho_A (x) &= \mu^* \left(\{ \beta \in \mathbb{R}^m : \beta \cdot x^* \geq \beta \cdot y^* \text{ for all } y \in A \} \right) \\ &= \mu \left(\{ u \in U_d : u(x) \geq u(y) \text{ for all } y \in A \} \right) \end{aligned}$$

Finally, since ρ is continuous, the countable additivity of μ follows from the same argument as Lemma 6 in Gul and Pesendorfer (2006).

G Proof of Theorem 2.2

First, suppose $k \geq 2$. Let $A = \{x_1, \ldots, x_n\}$ and consider some $x_i \in A$. Since X is full-dimensional, we can find $x_i' = x_i + \varepsilon$ such that $[x_i', x_j] \cap A' = \emptyset$ for all $j \neq i$ where $A' = A \cup \{x_i'\} \setminus \{x_i\}$. By applying convex substitutability repeatedly, we obtain

$$\rho_{A'} (x_i') \geq \rho_{\lambda x_i' + (1-\lambda) A'} (x_i')$$

for all $\lambda \in (0, 1)$. Since ρ is mixed logit, this means that

$$\int_U \frac{e^{v\left(x_i'\right)}}{e^{v\left(x_i'\right)} + \sum_{j \neq i} e^{v(x_j)}} d\nu \geq \int_U \frac{e^{v\left(x_i'\right)}}{e^{v\left(x_i'\right)} + \sum_{j \neq i} e^{v\left(\lambda x_i' + (1-\lambda) x_j\right)}} d\nu$$

Taking the limit as $\lambda \to 1$, we have

$$\int_U \frac{e^{v\left(x_i'\right)}}{e^{v\left(x_i'\right)} + \sum_{j\neq i} e^{v(x_j)}} d\nu \geq \int_U \frac{e^{v\left(x_i'\right)}}{ne^{v\left(x_i'\right)}} d\nu = \frac{1}{n}$$

Thus,

$$\rho_A\left(x_i\right) = \int_U \frac{e^{v(x_i)}}{\sum_j e^{v(x_j)}} d\nu = \lim_{\varepsilon\to 0} \int_U \frac{e^{v(x_i+\varepsilon)}}{e^{v(x_i+\varepsilon)} + \sum_{j\neq i} e^{v(x_j)}} d\nu \geq \frac{1}{n}$$

This is true for any $x_i \in A$. Note that if the inequality is strict for some $x_i \in A$, then

$$1 = \sum_i \rho_A\left(x_i\right) > \sum_i \frac{1}{n} = 1$$

yielding a contradiction. Thus, ρ must be uniform choice.

Now, suppose $k = 1$. Let $A = \{x_1, \ldots, x_n\}$ and since $k = 1$, we can order the alternatives $x_1 < \cdots < x_n$. Consider some $x_i \in A$ and let $z = \lambda x_i + (1-\lambda)\, x_{i+1}$. By convex substitutability,

$$\begin{aligned}
\rho_A\left(x_i\right) &\geq \rho_{A\cup\{z\}\setminus\{x_{i+1}\}}\left(x_i\right) = \int_U \frac{e^{v(x_i)}}{e^{v(z)} + \sum_{j\neq i+1} e^{v(x_j)}} d\nu \\
\rho_A\left(x_i\right) &\geq \int_U \frac{e^{v(x_i)}}{e^{v(x_i)} + \sum_{j\neq i+1} e^{v(x_j)}} d\nu
\end{aligned}$$

where the second inequality follows from taking the limit as $\lambda \to 1$. Let $A' = A \cup \{x_i - \varepsilon_1\} \setminus \{x_{i+1}\}$ where $x_{i-1} < x_i - \varepsilon_1 < x_i$ and consider $z = \lambda x_i + (1-\lambda)\, x_{i+2}$. By convex substitutability,

$$\int_U \frac{e^{v(x_i)}}{e^{v(x_i-\varepsilon_1)} + \sum_{j\neq i+1} e^{v(x_j)}} d\nu \geq \int_U \frac{e^{v(x_i)}}{e^{v(x_i-\varepsilon_1)} + e^{v(z)} + \sum_{j\notin\{i+1,i+2\}} e^{v(x_j)}} d\nu$$

Taking limits as $\lambda \to 1$,

$$\begin{aligned}
\int_U \frac{e^{v(x_i)}}{e^{v(x_i-\varepsilon)} + \sum_{j\neq i+1} e^{v(x_j)}} d\nu &\geq \int_U \frac{e^{v(x_i)}}{e^{v(x_i-\varepsilon_1)} + e^{v(x_i)} + \sum_{j\notin\{i+1,i+2\}} e^{v(x_j)}} d\nu \\
\int_U \frac{e^{v(x_i)}}{e^{v(x_i)} + \sum_{j\neq i+1} e^{v(x_j)}} d\nu &\geq \int_U \frac{e^{v(x_i)}}{2e^{v(x_i)} + \sum_{j\notin\{i+1,i+2\}} e^{v(x_j)}} d\nu
\end{aligned}$$

where the last inequality follows from $\varepsilon_1 \to 0$. By replacing each x_j with $x_i - \varepsilon_j$ for $j > i$, we can apply this argument repeatedly and obtain

$$\rho_A\left(x_i\right) \geq \int_U \frac{e^{v(x_i)}}{(n-i)\, e^{v(x_i)} + \sum_{j\leq i} e^{v(x_j)}} d\nu$$

By symmetric reasoning, we can apply the same argument for all $j < i$ and obtain

$$\rho_A(x_i) \geq \int_U \frac{e^{v(x_i)}}{(n-i)\, e^{v(x_i)} + i e^{v(x_i)}} d\nu = \frac{1}{n}$$

Thus, ρ must be uniform choice by the same reasoning as the case for $k \geq 2$.

H Uniform convergence

In this section, we show that mixed logit cannot approximate pure characteristics models under uniform convergence. Define $\rho^n \to \rho$ *uniformly* if for any $\varepsilon > 0$, there exists some N such that for all $n > N$,

$$\sup_{A \in \mathcal{A}} |\rho_A^n - \rho_A| < \varepsilon$$

Proposition 4. *For any pure characteristic ρ, there exists no sequence of mixed logits ρ^n such that $\rho^n \to \rho$ uniformly.*

Proof. Fix distinct $x, y \in X$ and let

$$x_n = \left(1 - \frac{1}{n}\right) x + \frac{1}{n} y$$

Note that $x_n \in X$ as X is convex. Let $A_n = \{y, x_n, x_{n+1}\}$ so $A_n \to \{y, x\}$ and thus $\rho(y, x) = \lim_n \rho_{A_n}(y)$ by Theorem 3.1. Suppose $\rho^m \to \rho$ uniformly so for any $\varepsilon > 0$, we can find some M such that $\sup_{A \in \mathcal{A}} |\rho_A^m - \rho_A| < \varepsilon$ for all $m > M$. We thus have

$$\begin{aligned}
\int_U \frac{e^{v(y)}}{e^{v(y)} + 2e^{v(x)}} d\nu_m &= \lim_n \int_U \frac{e^{v(y)}}{e^{v(y)} + e^{v(x_n)} + e^{v(x_{n+1})}} d\nu_m = \lim_n \rho_{A_n}^m(y) \\
&\geq \lim_n \rho_{A_n}(y) - \varepsilon = \rho(y, x) - \varepsilon \geq \rho^m(y, x) - 2\varepsilon \\
&\geq \int_U \frac{e^{v(y)}}{e^{v(y)} + e^{v(x)}} d\nu_m - 2\varepsilon
\end{aligned}$$

Since this is true for all $x \neq y$, we can consider $x_t \to y$ and taking limits,

$$\frac{1}{3} = \int_U \frac{e^{v(y)}}{e^{v(y)} + 2e^{v(y)}} d\nu_m \geq \int_U \frac{e^{v(y)}}{e^{v(y)} + e^{v(y)}} d\nu_m - 2\varepsilon = \frac{1}{2} - 2\varepsilon$$

yielding a contradiction for small ε. □

I Empirical exercise on convex substitutability

In this section, we use simulated data in a simple example to demonstrate the degree to which an estimated mixed logit model can violate convex substitutability. Set $k = 1$ and $X = [0, 10] \subset \mathbb{R}$. Consider a pure characteristic model where each type has utility

$$u_\theta (x) = -(x - \theta)^2$$

To simulate choice data, we consider 2000 draws of θ from $N(5, 2.5^2)$. We consider two data treatments:

- Case 1: 5 binary menus $\{x, y\}$ are draw uniformly from X
- Case 2: 20 binary menus $\{x, y\}$ are draw uniformly from X

In each case, simulated choices from these binary menus are used to estimate (via maximum likelihood) a mixed logit model

$$\int_{\mathbb{R}^2} \frac{e^{\beta_1 x^2 + \beta_2 x}}{e^{\beta_1 x^2 + \beta_2 x} + e^{\beta_1 y^2 + \beta_2 y}} dF(\beta)$$

where $(\beta_1, \beta_2) \sim N(\mu, \Sigma)$ is multivariate Normal.

We then assess the extent to which the estimated mixed logit satisfies convex substitutability. Figure 1 shows the results for Case 1 where 5 menus are used for estimation. It shows the plots for $\rho(x, 10)$ (where x varies from 0 to 10) and $\rho(0, y)$ (where y varies from 0 to 10) under the original pure characteristic model and estimated mixed logit model. Note that convex substitutability implies that $\rho(\cdot, 10)$ and $\rho(0, \cdot)$ should both be increasing functions. The plots for the estimated mixed logit model however demonstrate large regions of non-monotonicity: in fact, they are strictly decreasing for up to 20% of the $[0, 10]$ domain. Note that the violations are most severe when x and y are most similar (closest). Figure 2 shows the same plots for Case 2 when there is more data (20 menus) and the violations of convex substitutability are less severe.

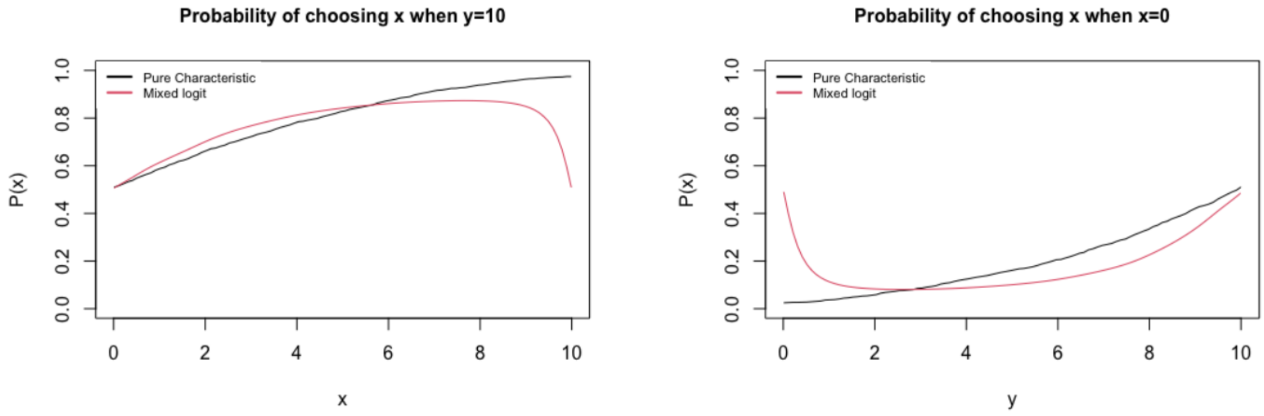

Figure 1: Plots of $\rho(\cdot, 10)$ and $\rho(0, \cdot)$ under Case 1

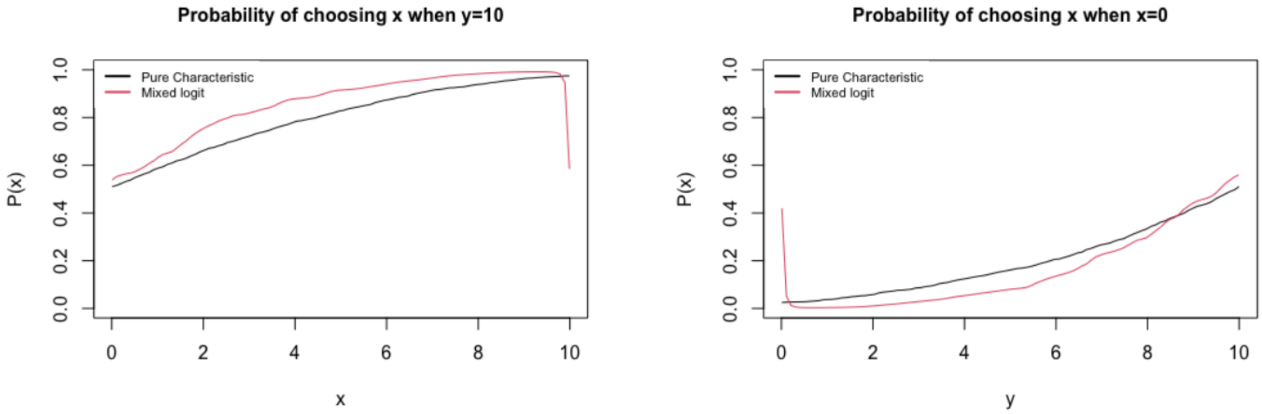

Figure 2: Plots of $\rho(\cdot, 10)$ and $\rho(0, \cdot)$ under Case 2

References

Ackerberg, D., and M. Rysman (2005): "Unobserved Product Differentiation in Discrete-Choice Models: Estimating Price Elasticities and Welfare Effects," *RAND Journal of Economics*, 36(4), 1–19.

Ahn, D., and T. Sarver (2013): "Preference for Flexibility and Random Choice," *Econometrica*, 81(1), 341–361.

Aliprantis, C., and K. Border (2006): *Infinite Dimensional Analysis*. Springer.

Anderson, S., A. De Palma, and J. Thisse (1989): "Demand for Differentiated Products, Discrete Choice Models, and the Characteristics Approach," *The Review of Economic Studies*, 56(1), 21–35.

Apesteguia, J., and M. Ballester (2018): "Monotone Stochastic Choice Models: The Case of Risk and Time Preferences," *Journal of Political Economy*, 126(1), 74–106.

Apesteguia, J., M. Ballester, and J. Lu (2017): "Single-Crossing Random Utility Models," *Econometrica*, 85(2), 661–674.

Bajari, P., and C. Benkard (2004): "Comparing Hedonic and Random Utility Models of Demand with an Application to PC's," Mimeo.

——— (2005): "Demand Estimation with Heterogeneous Consumers and Unobserved Product Characteristics: A Hedonic Approach," *Journal of Political Economy*, 113(6), 1239–1276.

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

Berry, S., and A. Pakes (2007): "The Pure Characteristics Demand Model," *International Economic Review*, 48(4), 1193–1225.

Cerreia-Vioglio, S., F. Maccheroni, M. Marinacci, and A. Rustichini (2018a): "Law of Demand and Stochastic Choice," Mimeo.

——— (2018b): "Multinomial Logit Processes and Preference Discovery: Inside and Outside the Black Box," Mimeo.

Chambers, C., T. Cuhadaroglu, and Y. Masatlioglu (2020): "Behavioral Influence," Mimeo.

Cohen, M. (1980): "Random Utility Systems - The Infinite Case," *Journal of Mathematical Psychology*, 22, 1–23.

Compiani, G. (2019): "Market Counterfactuals and the Specification of Multi-Product Demand: a Nonparametric Approach," Mimeo.

Davidson, D., and J. Marschak (1959): "Experimental Tests of Stochastic Decision Theory," in *Measurement: Definitions and Theories*, ed. by C. W. Churchman. Wiley.

Debreu, G. (1958): "Stochastic Choice and Cardinal Utility," *Econometrica*, 26(3), 440–444.

——— (1960): "Individual Choice Behavior: A Theoretical Analysis by R. Duncan Luce (review)," *American Economic Review*, 50(1), 186–188.

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

Frick, M., R. Iijima, and T. Strzalecki (2019): "Dynamic Random Utility," *Econometrica*, 87(6), 1941–2002.

Fudenberg, D., and T. Strzalecki (2015): "Dynamic Logit with Choice Aversion," *Econometrica*, 83(2), 651–691.

Gowrisankaran, G., and M. Rysman (2012): "Dynamics of Consumer Demand for New Durable Goods," *Journal of Political Economy*, 120(6), 1173–1219.

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

Gul, F., and W. Pesendorfer (2006): "Random Expected Utility," *Econometrica*, 74(1), 121–146.

Hendel, I., and A. Nevo (2006): "Measuring the Implications of Sales and Consumer Inventory Behavior," *Econometrica*, 74(6), 1637–1673.

Hotelling, H. (1929): "Stability in Competition," *Economic Journal*, 39, 41–57.

HOTZ, V. J., AND R. MILLER (1993): "Conditional Choice Probabilities and the Estimation of Dynamic Models," *Review of Economic Studies*, 60, 497–529.

LIN, Y. (2019): "Random Non-Expected Utility: Non-Uniqueness," Mimeo.

LU, J. (2016): "Random Choice and Private Information," *Econometrica*, 84(6), 1983–2027.

——— (2019): "Random Ambiguity," Mimeo.

LU, J., AND K. SAITO (2018): "Random Intertemporal Choice," *Journal of Economic Theory*, 177.

LUCE, D. (1959): *Individual Choice Behavior*. New York: Wiley.

MCFADDEN, D. (1973): "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, ed. by P. Zarembka. Academic Press.

MCFADDEN, D., AND K. TRAIN (2000): "Mixed MNL Models for Discrete Response," *Journal of Applied Econometrics*, pp. 447–470.

MUNKRES, J. (2000): *Topology*. Prentice Hall.

NARITA, Y., AND K. SAITO (2021): "Approximating Choice Data by Discrete Choice Models," Mimeo.

NATENZON, P. (2019): "Random Choice and Learning," *Journal of Political Economy*, 127(1), 419–457.

NEVO, A. (2001): "Measuring Market Power in the Ready-to-Eat Cereal Industry," *Econometrica*, 69(2), 307–342.

PETRIN, A. (2002): "Quantifying the Benefits of New Products: The Case of the Minivan," *Journal of Political Economy*, 110(4), 705–729.

RUDIN, W. (1991): *Functional Analysis*. McGraw-Hill.

RUST, J. (1987): "Optimal replacement of GMC bus engines, an empirical model of Harold Zurcher," *Econometrica*, 55(5), 999–1033.

SAITO, K. (2018): "Axiomatizations of the Mixed Logit Model," Mimeo.

Salop, S. (1979): “Monopolistic Competition with Outside Goods,” *The Bell Journal of Economics*, 10(1), 141–156.

Tserenjigmid, G., and M. Kovach (2020): “Behavioral Foundations of Nested Stochastic Choice and Nested Logit,” Mimeo.

Wilcox, N. (2011): “Stochastically more risk averse: A contextual theory of stochastic discrete choice under risk,” *Journal of Econometrics*, 162, 89–104.