Discovering Context Effects from Raw Choice Data

Arjun Seshadri [1] **Alexander Peysakhovich** [2] **Johan Ugander** [1]

Abstract

Many applications in preference learning assume that decisions come from the maximization of a stable utility function. Yet a large experimental literature shows that individual choices and judgements can be affected by "irrelevant" aspects of the context in which they are made. An important class of such contexts is the composition of the choice set. In this work, our goal is to discover such choice set effects from raw choice data. We introduce an extension of the Multinomial Logit (MNL) model, called the context dependent random utility model (CDM), which allows for a particular class of choice set effects. We show that the CDM can be thought of as a second-order approximation to a general choice system, can be inferred optimally using maximum likelihood and, importantly, is easily interpretable. We apply the CDM to both real and simulated choice data to perform principled exploratory analyses for the presence of choice set effects.

1. Introduction

Modeling individual choice is an important component of recommender systems (Resnick and Varian, 1997), search engine ranking (Schapire et al., 1998), analysis of auctions (Athey and Levin, 2001), marketing (Allenby and Rossi, 1998), and demand modeling in diverse domains (Berry et al., 1995; Bruch et al., 2016). The workhorse models used either implicitly or explicitly in these disparate literatures are random utility models (RUMs) (Manski, 1977), which assume that individuals have a numeric utility for each item and that they make choices that maximize noisy observations of these utilities (Luce, 1959; McFadden, 1980; Kreps, 1988).

The most well known and widely used RUM is the conditional multinomial logit (MNL), also called the Luce model, which is the unique RUM that satisfies the axiom known as the *independence of irrelevant alternatives (IIA)* (Luce, 1959). Informally, this axiom states that adding an item to a choice set does not change the relative probabilities of choosing the other items. This assumption is very strong, but it allows analysts to build powerful and interpretable models. However, if one assumes IIA but it is not actually true, predictions for out-of-sample choices could be very wrong. Thus, it is important for analysts to discover whether IIA is approximately true in a given dataset.

At the same time, there is a large amount of experimental evidence showing significant deviations from rational choice across many domains. In particular, the value assigned to an item can strongly depend on the "irrelevant" elements of the context of the choice (Tversky, 1972; Tversky and Simonson, 1993). Attempts to model these context effects in a domain-free manner fall short of being practically valuable, either due to large parameter requirements, inferential intractability, or both (Park and Choi, 2013).

Our contribution addresses both issues. We consider the IIA-satisfying MNL model and make small modifications to subsume a class of IIA violations that we believe are important in practice, while retaining parametric and inferential efficiency. We refer to this model as the context dependent random utility model (CDM). The CDM can be thought of as a "second order" approximation of a general choice system (the MNL model, meanwhile, corresponds to a "first order" approximation). Because the CDM nests MNL, it can fit data that does satisfy IIA just as well, and, importantly, can be used to construct a nested-model hypothesis test for whether a particular dataset is consistent with IIA.

The key assumption of the CDM is that IIA violations come from pairwise interactions between items in the choice set and that larger choice set effects can be approximated additively using all pairwise effects. This assumption means that the CDM has many fewer parameters than a general choice system. We can further reduce the CDM's data dependence by assuming that these underlying effects can be well modeled by latent vectors of a smaller dimensionality than the number of items, resulting in what we call the low-rank CDM. The low-rank CDM can be useful in applications where the number of items is relatively large and where seeing all possible comparisons may be extremely costly.

[1]Stanford University, Stanford, CA [2]Facebook Artificial Intelligence Research, New York, NY. Correspondence to: Arjun Seshadri <aseshadr@stanford.edu>, Alexander Peysakhovich <alexpeys@fb.com>, Johan Ugander <jugander@stanford.edu>.

As a theoretical contribution, as part of this work we furnish formal results for conditions under which the parameters of a CDM can or can not be identified from data. In situations where identifiability is not achieved, we advocate for additive ℓ_2-regularization of the log-likelihood to select the minimum norm solution. We also provide a uniform convergence bound on the expected squared ℓ_2 error of the estimate as well as sample complexity bounds.

As applications, we first test the CDM in synthetic data and show that a nested model likelihood ratio test—between the CDM and MNL models—has good finite sample properties. When IIA holds, a $p < .05$ hypothesis test rejects the null slightly less than 5% of the time. When IIA does not hold the null hypothesis is overwhelmingly rejected even in medium size data-sets. By contrast we see that using a nested model test based on the nested structure of a general choice system and MNL model gives a test that wildly over-rejects the null, even when IIA is true.

We apply the CDM to several real-world datasets. First, we show that we can strongly reject IIA in the popular SFWork and SFShop choice datasets. Second, we consider using the CDM to model choices in the task of Heikinheimo and Ukkonen (2013), where individuals are presented with triplets of items and asked which item is least like the other two. Here the CDM can capture the underlying choice structure quite well while IIA is an extremely unreasonable assumption.

1.1. Related Work

The CDM vaguely resembles the continuous bag of words (CBOW) neural network architecture popularized by word2vec (Mikolov et al., 2013), with two key differences. First, while the CBOW model tries to predict the appearance of a word where candidates are any word in the vocabulary, the CDM models choices from arbitrary subsets. Second, although in principle the word2vec model and its extensions train two embeddings per word (one as target and one as context), these embeddings are typically averaged together at the end of training to obtain a single embedding per word. However, it has been shown that keeping these two embeddings separate does allow one to capture ancillary information not captured by the single embedding (Rudolph et al., 2016) and the two embeddings can be used to model complements and substitutes in supermarket shopping data (Ruiz et al., 2017).

Random utility models are widely used to analyze intertemporal choice (discount functions) (Muraven and Baumeister, 2000; Fudenberg and Levine, 2012), choice under uncertainty (Bordalo et al., 2012; Fox and Tversky, 1995), and choices about cooperation (List, 2007; Liberman et al., 2004; Peysakhovich and Rand, 2015). They are also workhorses in modeling consumer behavior in applied settings. Analysts interested in industrial organization use additive models to try to predict the effects of various policies on markets (Berry et al., 1995). Online recommender systems (Resnick and Varian, 1997), which model user-item interactions as inner products of low rank vectors, can be seen as employing a utility function that is an inner product between item attributes and user weights.

The CDM and low-rank CDM generalize a number of prominent choice models in a unified framework. In a later section, after we introduce the basic mathematical notation, we present connections to the work of Tversky and Simonson (1993), Batsell and Polking (1985), and Chen and Joachims (2016*a*;*b*). Importantly, this means our convergence and identifiability results carry over to these other models, which all previously lacked such results.

2. Modeling choice systems

Let $\mathcal{X}$ be a finite set of n alternatives that we hold fixed and let $\mathcal{C} = \{C : C \subseteq \mathcal{X}, |C| \geq 2\}$ be the set of all subsets of $\mathcal{X}$ of size greater than or equal to two. Throughout this work, we assume there is a single individual that is presented with choice sets and chooses a single item from each choice set. In this setting, the fundamental object of study is a *choice system*, a collection of probability distributions for every $C \in \mathcal{C}$, describing the probability that an item x is chosen from $C, \forall x \in C$. We denote each such probability by $P(x \mid C)$. In general, a choice system can model arbitrary preferences on arbitrary subsets with no further restrictions.

The most commonly assumed restriction on choice systems is that they satisfy the independence of irrelevant alternatives (IIA), which can be stated as follows.

Assumption 1. *A choice system on $\mathcal{X}$ satisfies the independence of irrelevant alternatives (IIA) if for any $x, y \in \mathcal{X}$ and choice sets $A, B \subseteq \mathcal{X}$ with $x, y \in A, B$ we have*

$$\frac{P(x \mid A)}{P(y \mid A)} = \frac{P(x \mid B)}{P(y \mid B)}.$$

In other words, IIA states that the composition of a choice set does not affect the relative attractiveness of items. A main question in this work will be, given a dataset $\mathcal{D}$ of choices from choice sets, can we determine whether $\mathcal{D}$ was generated by a model satisfying IIA or a model of a more general choice system? If not generated by a model satsify IIA, is it possible to define tractable model classes between the class of models satisfying IIA and the class of fully general choice systems? Our answer to this question is a formal truncation of a general choice system that we call the *context dependent random utility model* (CDM).

2.1. Context-dependent random utility models (CDMs)

A trivial model of a general choice system is the *universal logit model* (McFadden et al., 1977), which simply parameterizes the choice system object, defining utilities $u(x \mid C), \forall x \in C$, for each $C \in \mathcal{C}$ that can vary arbitrarily for every item, for every set. The choice probabilities for a universal logit model are then:

$$P(x \mid C) = \frac{\exp(u(x \mid C))}{\sum_{y \in C} \exp(u(y \mid C))}.$$

The above model exhibits scale-invariance on each subset C, and thus we require that $\sum_{y \in C} u(y \mid C) = 0, \forall C \in \mathcal{C}$ for the purposes of identifiability. While relatively uninteresting as a model, the above formulation is the starting point for the following observation about choice systems, first documented by Batsell and Polking (1985).

Lemma 1. *The utilities in the universal logit model, $u(x \mid C), \forall C \in \mathcal{C}, \forall x \in C$, can be uniquely mapped as*

$$u(x \mid C) = \sum_{B \subseteq C \setminus x} v(x \mid B),$$

where $v(x \mid B)$ are values that satisfy the constraints $\sum_{x \notin B} v(x \mid B) = 0, \forall B \subset \mathcal{X}$.

For greater clarity, we expand out the terms individually.

$$u(x \mid C) = \underbrace{v(x)}_{\text{1st order}} + \underbrace{\sum_{y \in C \setminus x} v(x \mid \{y\})}_{\text{2nd order}} + \underbrace{\sum_{\{y,z\} \subseteq C \setminus x} v(x \mid \{y,z\})}_{\text{3rd order}} + \ldots + \underbrace{v(x \mid C \setminus \{x\})}_{|C|\text{th order}},$$

and expand out the first three sets of constraints:

$$\sum_{x \in \mathcal{X}} v(x) = 0, \quad \sum_{x \in \mathcal{X} \setminus y} v(x \mid \{y\}) = 0,$$
$$\sum_{x \in \mathcal{X} \setminus \{y,z\}} v(x \mid \{y,z\}) = 0.$$

We use $v(x) = v(x \mid \emptyset)$ for simplicity. This expansion reveals that arbitrary contextual utilities can be decomposed into intuitive contributions: the first order terms represent the item's intrinsic contribution to the utility, the second order terms represent the contextual contributions from every other item in the choice set, the third order terms the contributions from contextual pairs not modeled by contributions of the pairs constituent items, and so on. The expansion invites one to consider a hierarchy of choice model classes indexed by their order[1]: the pth order model refers to forcing all terms of order greater than p to zero. Denote this class of choice system models by $\mathcal{M}_p$. Clearly, we have $\mathcal{M}_1 \subset \mathcal{M}_2 \subset \ldots \subset \mathcal{M}_{n-1}$, where $\mathcal{M}_{n-1}$ is the universal logit model. We next consider two excercises. First, we write out the 1st order model explicitly:

$$P(x \mid C) = \frac{\exp(v(x))}{\sum_{y \in C} \exp(v(y))}, \quad \sum_{x \in \mathcal{X}} v(x) = 0.$$

This model $\mathcal{M}_1$ is the multinomial logit model, the workhorse model of discrete choice (Luce, 1959; McFadden, 1980). Moving to higher-order models, a counting exercise reveals that the number of free parameters in the pth order model is

$$\sum_{q=1}^{p} \binom{n}{q-1} (n-q),$$

which simplifies to $n-1$ and $(n-2)2^{n-1}+1$ parameters for $\mathcal{M}_1$ (MNL) and $\mathcal{M}_{n-1}$ (universal logit), respectively. Clearly, the parameters grow polynomially in the number of items and exponentially in the order: there are $O(n)$ parameters for the 1st order model, $O(n^2)$ parameters for the 2nd, and so on.

The principled next step, then, is to consider the minimal model class that accounts for context effects, $\mathcal{M}_2$:

$$P(x \mid C) = \frac{\exp(v(x) + \sum_{z \in C \setminus x} v(x \mid \{z\}))}{\sum_{y \in C} \exp(v(y) + \sum_{z \in C \setminus y} v(y \mid \{z\}))},$$
$$\text{s.t.} \sum_{x \in \mathcal{X}} v(x) = 0, \quad \sum_{x \in \mathcal{X} \setminus y} v(x \mid \{y\}) = 0.$$

We remove the constraints by introducing new parameters $u_{xz} = v(x \mid \{z\}) - v(z), \forall x, z \in \mathcal{X}$, interpretable as the pairwise push and pull of z on x's utility. We may then rewrite the above as

$$P(x \mid C) = \frac{\exp(\sum_{z \in C \setminus x} u_{xz})}{\sum_{y \in C} \exp(\sum_{z \in C \setminus y} u_{yz})}, \forall C \subseteq \mathcal{X}, \forall x \in C. \tag{1}$$

As in the case of both the MNL and universal logit model, the parameters u_{xz} are still only defined up to a constant shift. We refer to the model $\mathcal{M}_2$, as parameterized in equation (1), as the *context dependent random utility model (CDM)*, and note that it has $n(n-1)-1$ free parameters.

The CDM then corresponds to the following restriction on choice systems.

Assumption 2 (Pairwise linear dependence of irrelevant alternatives)**.** *A choice system on $\mathcal{X}$ satisfies pairwise linear dependence of irrelevant alternatives if, in the universal logit representation of Lemma 1, $v(x \mid B) = 0$ for all $B \subset \mathcal{X}$ for which $|B| \geq 2$.*

[1] This expansion differs from the one used by Batsell and Polking (1985), which expands the log probability ratios of items being chosen instead of the underlying contextual utilities.

This assumption can either be taken literally, or can be justified as an approximation on the grounds of applications: in practice many problems are concerned with choices from relatively small sets, and the linear context effect assumption is then a decent approximation.

2.2. Low-rank CDMs

From equation (1), it is clear that the parameters of the CDM, u_{xz}, $\forall x, z \in \mathcal{X}$, have a matrix-like structure. Note that the parameters do not quite form a matrix, as the diagonal elements u_{xx} are undefined and unused. But given this structure, it is natural ask if the pairwise contextual utilities can be modeled by a lower-dimensional parameterization.

Formally, we define the low-rank CDM as a CDM where the pairwise contextual utilities jointly admit a low-rank factorization $u_{xz} = c_z^T t_x, \forall x, y \in \mathcal{X}$. We call $t_x, c_x \in \mathbb{R}^r$, the *target* and *context* vectors, respectively, for each item $x \in \mathcal{X}$. We can then write the choice probabilities of the low-rank CDM as:

$$P(x \mid C) = \frac{\exp((\sum_{z \in C \setminus x} c_z)^T t_x)}{\sum_{y \in C} \exp((\sum_{z \in C \setminus y} c_z)^T t_y)}, \forall C \subseteq \mathcal{X}, \forall x \in C. \tag{2}$$

The rank-r CDM then has $2nr$ parameters and has at most $\min\{(2n - r)r, n(n-1) - 1\}$ degrees of freedom.

Our low-rank assumption is strongly related to standard additive utility models where one is given a low-dimensional featurization $x \in \mathbb{R}^r$ of each item and an individual's utility is $\beta^T x$. A difference here, other than the notion of contextual utility, is that we assume no featurization is available and that it must be learned.

3. Identifiability and Estimation of the CDM

Consider a dataset $\mathcal{D}$ of choices with generic element (x, C) that correspond to observing element x being chosen from set C. Let $\mathcal{C}_\mathcal{D}$ denote the collection of *unique* subsets of $\mathcal{X}$ represented in $\mathcal{D}$. If we assume that the data was generated by a CDM, it is important to understand conditions under which the parameters of that CDM are identifiable and conditions under which the expected error of a tractable estimation procedure converges to zero as the dataset gets large. In this section we furnish two sufficient conditions and one "insufficient" condition for identifiability. We then bound the expected squared ℓ_2 error of the maximum likelihood estimate (MLE) of a full-rank CDM. Because the log-likelihood of the full-rank CDM is convex (by the convexity of log-sum-exp), we know we can efficiently find this MLE, and that this bound on the error of the full-rank model also bounds the error of any low-rank model.

We consider the dataset $\mathcal{D}$ as being generated in the following hierarchical manner:

1. A choice set A is chosen at random from a distribution on the set of all subsets of $\mathcal{X}$.
2. The chooser chooses an item x from the choice set A according to a CDM with parameters $\theta \in \Theta$.

We can parametrize the utility function by θ referring to it as $u_\theta(\cdot \mid \cdot)$. Given a $\mathcal{D}$ and guess θ we can write the probability of (x, A) as

$$P_\theta(x \mid A) = \frac{\exp(u_\theta(x \mid A)}{\sum_{y \in A} \exp(u_\theta(y \mid A))}.$$

This means we have a well defined likelihood function for the full dataset

$$\mathcal{L}(\mathcal{D} \mid \theta) = \prod_{(x,A) \in \mathcal{D}} P_\theta(x, A). \tag{3}$$

For now we consider a full rank CDM where the parameter vector θ is the set of pairwise contextual utilities u_{xz}, $\forall x, z \in \mathcal{X}$. We will consider $u \in \mathbb{R}^d$ as the parameter vector, where for the full-rank CDM $d = n(n-1) - 1$. Because u can only be identified up to a scale, we consider possible CDMs with the constraint that $\sum_{xz} u_{xz} = 0$.

The likelihood (3) can be maximized using standard techniques. We will say that a dataset *identifies* a CDM if there are no two sets of parameters that have the same distribution $P(x \mid C), \forall C \in \mathcal{C}, \forall x \in C$. We now give a sufficient (but not necessary) condition for identification.

Theorem 1. *A CDM is identifiable from a dataset $\mathcal{D}$ if $\mathcal{C}_\mathcal{D}$ contains comparisons over all choice sets of two sizes k, k', where at least one of k, k' is not 2 or n.*

The proof is given in Appendix A. In the multinomial logit model, the constraints of IIA allow us to identify all parameters given just the probability distribution $P(\cdot \mid \mathcal{X})$, but in the less constrained CDM more information is needed. For a simple demonstration of the theorem, consider a choice system on $\mathcal{X} = \{a, b, c\}$ where

$$P(a \mid \mathcal{X}) = 0.8, \;\; P(b \mid \mathcal{X}) = 0.1, \;\; P(c \mid \mathcal{X}) = 0.1.$$

Here, if we assume IIA we can infer any pairwise choice probability simply by taking the appropriate ratio. However, if we do not assume IIA and only assume Assumption 2 (equivalent to assuming the choice system is a CDM), any set of pairwise probabilities are consistent with what we've observed. Thus the CDM is not identified if we only receive data about choices from $\{a, b, c\}$.

In Appendix A we show that the identifiability of the full-rank CDM for a given dataset $\mathcal{D}$ is equivalent to testing the rank of an integer design matrix $G(\mathcal{D})$ constructed from the dataset (Theorem 4). This characterization of the identifiability of the full-rank CDM also gives a sufficient condition

for the identifiability of low-rank CDMs. The proof of this theorem can be easily expanded to demonstrate an advantage of the CDM instead of a general choice system: for a general choice system to be identified $\mathcal{C}_\mathcal{D}$ would need to include in its support *every* choice set.

In addition to the above sufficient conditions for identifiability, we also have the following result about an important "insufficient" condition.

Theorem 2. *No rank r CDM, $1 \leq r \leq n$, is identifiable from a dataset $\mathcal{D}$ if $\mathcal{C}_\mathcal{D}$ contains only choices from sets of a single size.*

The proof is given in Appendix A. Requiring comparisons over two different choice set sizes is not unique to the CDM; recent results (Chierichetti et al., 2018) demonstrate that even a uniform mixture of two multinomial logit models, a special case of the mixed logit that violates IIA, requires comparisons over two different choice set sizes.

As a result of this theorem, for choice data collected from sets of a fixed size, the parameters of a CDM model that has been fit to data can not be interpreted without some amount of explicit or implicit regularization. This non-identifiability also applies to all blade-chest models (Chen and Joachims, 2016*a*), which (as alluded to in Section 3.3) are CDM models restricted to pairwise choices.

3.1. Uniform convergence

The likelihood function is log-concave and can thus be solved to arbitrary error through standard convex optimization procedures (avoiding shift invariance with the contraint $\sum_{xz} u_{xz} = 0$). We now show that maximum likelihood estimation efficiently recovers the true CDM parameters under mild regularity conditions.

Theorem 3. *Let $u^\star$ denote the true CDM model from which data is drawn. Let $\hat{u}_{MLE}$ denote the maximum likelihood solution. Assume $\mathcal{C}_\mathcal{D}$ identifies the CDM. For any $u^\star \in \mathcal{U}_\mathcal{B} = \{u \in \mathbb{R}^d : \|u\|_\infty \leq B, \mathbf{1}^T u = 0\}$, and expectation taken over the dataset $\mathcal{D}$ generated by the CDM model,*

$$\mathbb{E}\big[\, \|\hat{u}_{MLE}(\mathcal{D}) - u^\star\|_2^2 \,\big] \leq \frac{d}{m} \frac{\alpha k_{max}^2}{k_{min}},$$

where k_{max} and k_{min} respectively refer to the maximum and minimum sizes of choice sets in the dataset, and α is a constant that depends on B, k_{max} and the spectrum of the design matrix $G(\mathcal{D})$.

The proof is given in Appendix B, where we also state the exact relationship of α to the max norm radius B, the maximum choice set size $k_{\max}$ and design matrix $G(\mathcal{D})$. Both the identifiability condition and maximum norm bound are essential to the statement, as the right hand side diverges when the former is violated, and diverges as $B \to \infty$.

Theorem 3 is a generalization of a similar convergence result previously shown for the multinomial logit case (Shah et al., 2016) (the multinomial logit model class is a subset of the CDM model class). The proof follows the same steps, showing first that the objective satisfies a notion of strong convexity, and using that fact to bound the distance between the estimate and the true value. Our contributions augment the notation of Shah et al. (2016) to support multiple set sizes and the more complex structure of the CDM, and carefully bound the role of these deviations in the steps leading to the result.

To our knowledge, this convergence bound furnishes the first tractable sample complexity result for a model that can accommodate deviations from a random utility model (RUM). A comparable lower bound, which we do not furnish in this work, would make clear whether the maximum likelihood procedure is inferentially optimal or not. And while stated for the full-rank model, our convergence bound holds for CDMs of any rank. It is possible that low-rank CDMs admit an improved rank-dependent convergence rate.

3.2. Testing

We can use the CDM to construct a statistical test of whether our data is indeed consistent with the MNL/Luce model, and thus IIA, across the choice sets we observe. Recall that the class of Luce models is nested within the CDM, which is in turn nested within the universal logit, as discussed in Section 2. We can consider the following likelihood ratio statistic,

$$\Lambda(\mathcal{D}) = \frac{\sup_{\theta \in \Theta_{\text{Luce}} \subset \Theta_{\text{CDM}}} \mathcal{L}(\mathcal{D} \mid \theta)}{\sup_{\theta \in \Theta_{\text{CDM}}} \mathcal{L}(\mathcal{D} \mid \theta)},$$

where Θ_{Luce} and Θ_{CDM} respectively refer to the parameter classes of Luce and CDM Models. We then appeal to a classical result from asymptotic statistics (Wilks, 1938) that as the sample size $m \to \infty$, $D = -2\log(\Lambda(\mathcal{D}))$ converges to the χ^2 distribution with degrees of freedom Δ equal to the difference between the number of parameters between the two model classes. For CDM and Luce, $\Delta = n(n-2)$. For a universal logit and Luce, $\Delta = (\sum_{C \in \mathcal{C}_\mathcal{D}} (|C|-1)) - (n-1)$, where $\mathcal{C}_\mathcal{D}$ are the unique subsets in the dataset that the test can reasonably evaluate. Our test then compares the statistic to the value of the χ^2_Δ distribution corresponding to a desired level of statistical significance.

We are keen to note that the CDM test likely enjoys finite sample guarantees when the true distribution is sampled from a CDM, owing to the uniform convergence of the MLE shown in Theorem 3. In experiments that follow, we look at the finite sample performance of this likelihood ratio test, evaluating this claim empirically and comparing the performance of our test to the universal logit test.

3.3. Unifying Existing Choice Models

The CDM and low-rank CDM generalize a number of prominent choice models in a unified framework. In this section we present connections to the work of Tversky and Simonson (1993), Batsell and Polking (1985), and Chen and Joachims (2016*a*;*b*). This means that our convergence and identifiability results carry over to these other models, which all previously lacked such results.

The Tversky-Simonson model. The *additive separable utility model (ASM)* is the cornerstone of random utility modeling in many applications. In the ASM the utility of item x can be written as an inner product $u(x) = w^T t_x$, where t_x is a feature vector of item x (typically known to the analyst, but sometimes latent) and the vector w contains the parameters of the linear model (estimated from data). The parameters w have a real world interpretation: they are the weights that an individual places on each attribute. These can be used to estimate counterfactuals: for example, how much would an individual rank a new item y that we have not seen before?

A seminal experiment by Tversky and Kahneman asks individuals to consider a situation where they are purchasing an object and they learn that the same object is available across town (a 20 minute drive away) for \$5 cheaper (Tversky and Kahneman, 1985). They then ask whether the individuals would drive across town to take advantage of this lower price, essentially a question about their value of time. Individuals are more likely to drive across town when they are considering purchasing a \$10 object compared to when they are purchasing a \$120 object, even though the time/money tradeoff is identical.

The ASM assumes that the weights are constant across contexts, making the choices in the story above impossible if the ASM is indeed the true model. Tversky and Simonson (1993) expand the ASM to allow context to adjust the weights that individuals place on attributes while keeping the attributes fixed. This approach has a particular psychological interpretation: the presence of certain items makes some dimensions of a choice more salient than others, an effect that appears across a variety of decision situations.

This is formalized by setting utility of x in context C to be $u^{TS}(x \mid C) = w(C)^T t_x$. Tversky and Simonson discuss several ways in which some experimental results can be modeled using various forms of weights $w(C)$, though their approach requires both features and context-dependent weight functions to be hand-engineered. They do not formalize any procedure for how one can learn such a model from choice data directly. Thus our CDM can be seen as a method for learning the parameters of a Tversky-Simonsohn model directly from data in an efficient manner.

The Batsell-Polking model. Batsell and Polking (1985) introduces a model of competing product market shares that can also be written as a truncated expansion of the log ratio of choice probabilities. The CDM can be viewed as an alternative parameterization of a *third-order Batsell-Polking model*. There are several significant differences between the way Batsell and Polking viewed their third-order model and how we view the CDM. First, Batsell and Polking advocated for fitting their models to data using a hand-tuned least squares procedure whereas we use more general maximum likelihood techniques. Second, our identifiability and convergence results are entirely new. Their least-squares procedure understandably has no analogous guarantees. Lastly, our restriction to low-rank parameterizations is squarely new and can greatly reduce the model complexity.

The Blade-Chest model. Standard models for competition build on the Elo rating system for chess (Elo, 1978) and the TrueSkill rating system for online gaming (Herbrich et al., 2006). Both of these models assume that individuals have a one-dimensional latent "skill" parameter that can be discovered from matchup data between competitors.

The Blade-Chest model (Chen and Joachims, 2016*a*;*b*) tries to model rock-paper-scissors-type intransitivies in pairwise matchups through a multidimensional latent embedding of skill. In the language of our CDM, the blade-chest "inner product" model (the authors also consider a "distance" model) defines the probability that x beats y as:

$$\Pr(x \mid \{x, y\}) = \frac{\exp(t_x^T c_y)}{\exp(t_x^T c_y) + \exp(t_y^T c_x)},$$

which is precisely a CDM restricted to pairs. We can view the CDM as a natural extension of the Blade-Chest model from pairs to larger sets. Considering our negative identifiability result for choice data consisting of only a single set size (Theorem 2), we conclude that the Blade-Chest model is not identifiable and requires either explicit or implicit regularization in order to make the parameters interpretable.

4. Experiments

We now evaluate the CDM and low-rank CDM on data. Our evaluation includes comparisons with MNL/Luce models and mixed MNL models (McFadden and Train, 2000). MNL and CDM model likelihoods are optimized using Adam (Kingma and Ba, 2014), a stochastic gradient descent algorithm with adaptive moment estimation. Mixed MNL likelihoods are optimized using open source code from (Ragain and Ugander, 2016). The CDM parameter optimization is initialized with values corresponding to a Luce MLE for that dataset. All datasets are pre-existing and public; replication code for all figures will be released at publication time.

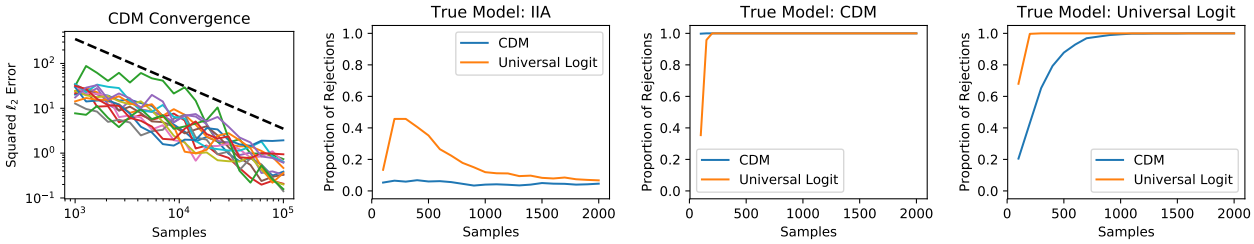

Figure 1. (a) Approximation error of an estimated CDM in 10 growing datasets validates our convergence theorem. The dashed black line is a visual guide of the slope $1/m$. *(b,c,d)* The proportion of rejections for a CDM-based hypothesis test of IIA (at a threshold of $p < .05$) when the data is generated by a MNL, CDM, and general choice system model as a function of the number of samples. When IIA is true, the CDM-based test has a 5% of false rejection rate while the test based on the general choice system is highly anti-conservative. When IIA is false, both tests quickly and correctly reject. All model parameters are described in the main text.

Simulated Data. We begin with simulated data, which allows us to validate our theoretical results regarding the convergence of the MLE in a setting where the underlying data-generating process is known. Since we know whether IIA holds in the simulated data, simulated data is also useful for examining two aspects of the CDM-based hypothesis test. First, we ask about the power of the test, in other words, does the test reject IIA when it is not true? Second, we ask about the conservatism of the test. The nested model likelihood ratio tests are only valid asymptotically; in our simulated data we can check whether the CDM over or under-rejects the null hypothesis of IIA in finite samples.

We consider three data-generating processes: one where the data is generated from a MNL model (where IIA holds), one where the data is generated from a CDM, and one where the data is generated from a general choice system. The universe has $n = 6$ items. In the IIA dataset the underlying MNL has one parameter per item, randomly generated. In the CDM dataset the parameters $U = T^T C$ are generated by sampling elements of both T and C i.i.d. from $N(0, 1)$. We sample choice sets uniformly at random (thus our identification conditions are quickly met) and then a choice according to the underlying model. We fit a Luce, CDM, and universal logit model to the data and look at both the error of the CDM MLE (to evaluate convergence) and the p-value from the nested model likelihood ratio tests. When the p-value falls below .05 we say that the hypothesis of IIA is rejected.

In addition, we compare the CDM-based nested test to another nested model test where we use a general choice system as the alternative model. Recall that the general choice system also nests MNL. However, the general choice system has combinatorially more parameters.

Figure 1 shows our results. The left panel validates the $O(\frac{1}{m})$ convergence result in Theorem 3. The right three panels look at how often the hypothesis of IIA is rejected, out of 1000 independent growing datasets, when the underlying data comes from the three different data generating processes. We see that the CDM rejects the null less than 5% of the time when the data generating process indeed satisfies IIA and rejects IIA when it is not true almost all the time, even with relatively small amounts of data.

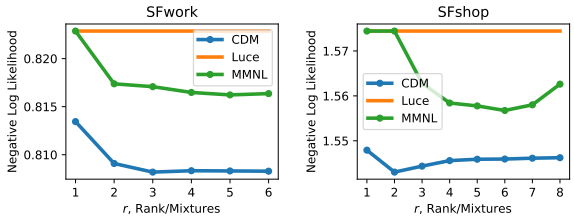

Figure 2. The out of sample negative log-likelihood of the MLE for the SFWork and SFshop datasets under an MNL/Luce model, mixed MNL model with varying number of mixture components, and CDMs of varying rank. The CDM outperforms the other models at all ranks.

By contrast we see that the universal logit requires quite a lot of data to reach the asymptotically valid coverage, even with a universe of only 6 items. For finite samples it is highly anti-conservative, over-rejecting when IIA is true for small and medium amounts of data.

SFwork/SFshop. We now turn to two real-world datasets: SFwork and SFshop. These data are collected from a survey of transportation preferences around the San Francisco Bay Area (Koppelman and Bhat, 2006). SFshop consists of 3,157 observations of a choice set of transportation alternatives available to individuals traveling to and from a shopping center, as well as what transportation that individual actually chose. SFwork is similar, containing 5,029 observations consisting of commuting options and the choice made.

These datasets are similar to those employed in many demand estimation applications. For example, Berry et al. (1995) fit a MNL model to aggregate data in order to estimate the utility function of an average consumer for automobiles as well as how individuals (on average) trade off various qualities of a car (e.g., gas mileage vs. price). With access to underlying parameters, the analyst can then make counterfactual estimates such as, for example, what would be the sales of a hypothetical cheaper and higher gas

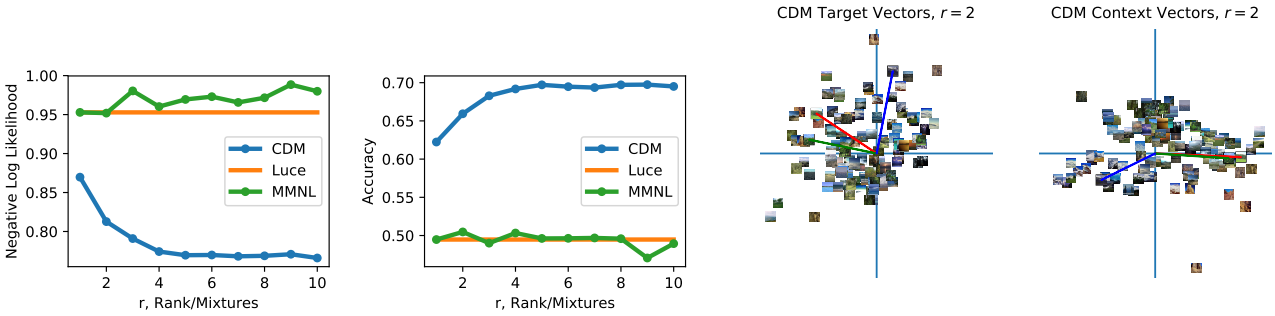

Figure 3. (Left) The out of sample negative log-likelihood and accuracy of the MLE for the nature photo dataset under an MNL/Luce model, mixed MNL model with varying number of mixture components, and CDMs of varying rank. The CDM outperforms the other models at all ranks in terms of both likelihood and accuracy. The target (center) and context (right) vector embeddings for the nature photo dataset with a rank-2 CDM, with three sample vectors highlighted. An item's context and target vectors have, on average, a negative dot product, showing that the addition of an item makes similar items already in the choice set less likely to be chosen.

mileage car? With the SFWork/SFShop data, we can ask questions like: what would happen if we made certain types of transit more or less available? Of course, if the underlying assumption (IIA) of the MNL model is wrong, then we may expect our counterfactual answers to also be wrong.

We run both hypothesis test for IIA (asking: "is there IIA in the data?") and examine the out-of-sample performance of the low-rank CDM (asking: "does the violation of IIA have meaningful consequence for prediction?"). From the hypothesis test we obtain a p-value of 10^{-7} and can strongly reject IIA. Figure 2 shows the out of sample fit on a held out 20% of the data for low-rank CDMs, mixed MNLs, and an MNL model, again showing that IIA is not satisfied in this data.

Not Like The Other. We turn to a slightly different dataset to demonstrate another way the CDM can be used. We consider the task introduced by Heikinheimo and Ukkonen (2013) where individuals are shown triplets of nature photographs and asked to choose the one photo that is most unlike the other two. This task involves comparison-based choices (Kleinberg et al., 2017) where there are no "irrelevant alternatives" and IIA is clearly violated: consider two example task where the choice set is two mountains and a beach vs. two beaches and a mountain.

The dataset is comprised of $m = 3355$ triplets spanning $n = 120$ photos. Because the dataset only has choice sets of a fixed size, the CDM is not directly identifiable (Theorem 2). We resolve this issue by adding an ℓ_2 regularization term to the log-likelihood. For a small positive regularization penalty, the optimizer then selects the least norm solution within the null space. We choose a non-negligible penalty, chosen through cross-validation, to serve the additional purpose of improving model generalization.

We fit low-rank CDMs and see that they handily outperforms a MNL model (i.e. just item-level utilities) and mixed MNL models on a 20% held-out test set (Figure 3, left). In addition, we plot the vectors learned in the low-rank CDM (Figure 3, right). We see that similar images are grouped together both as targets and as contexts. We also see an intuitive property of the dataset: for most items x, t_x and c_x have a negative inner product. Essentially, having two copies of the same item in a choice set makes each copy less likely to be chosen.

5. Conclusion

Existing work has argued that context dependence, and in particular choice-set dependence, is an important part of human decision-making (Ariely et al., 2003; Slovic, 1995; Tversky and Simonson, 1993). For example, the addition of an item that is strictly dominated can shift individual choices (Huber et al., 1982), an effect impossible under IIA. Further IIA violations are often seen in intertemporal choice, choice under uncertainty, and choices about cooperation. Applying a CDM to these domains is an important area of future work.

There is separate experimental evidence that human choices are intransitive in some settings, where people may prefer A to B and B to C but then C to A. This evidence has given rise to a theoretical literature on relaxing the transitivity axiom of rational choice or the regularity axiom of random utility (Tversky, 1969; Ragain and Ugander, 2016; Benson et al., 2016). We note that context dependence can imply intransitivity but it implies a particular kind of intransitivity. Beyond the scope of this paper, it would be interesting to provide an axiomatic characterization of the CDM and the kinds of violations of rational choice that the model can or can not represent.

Understanding human decision-making is an important endeavor for both basic and applied science and is becoming increasingly important in human-centered machine learning and artificial intelligence. We view the introduction

of techniques from machine learning and AI into behavioral science and the flow of realistic models of human behavior in the other direction as crucial and beneficial for both fields (Wager and Athey, 2015; Fudenberg and Peysakhovich, 2014; Naecker, 2015; Epstein et al., 2016; Peysakhovich and Rand, 2017). We hope that our work contributes to this important conversation.

Acknowledgements

We thank Fred Feinberg and Stephen Ragain for their helpful comments and feedback. AS was supported in part by an NSF Graduate Research Fellowship. JU was supported in part by an ARO Young Investigator Award.

References

Allenby, G. M. and Rossi, P. E. (1998), 'Marketing models of consumer heterogeneity', *Journal of econometrics* **89**(1), 57–78.

Ariely, D., Loewenstein, G. and Prelec, D. (2003), "coherent arbitrariness': Stable demand curves without stable preferences', *The Quarterly Journal of Economics* **118**(1), 73–106.

Athey, S. and Levin, J. (2001), 'Information and competition in us forest service timber auctions', *Journal of Political economy* **109**(2), 375–417.

Batsell, R. R. and Polking, J. C. (1985), 'A new class of market share models', *Marketing Science* **4**(3), 177–198.

Benson, A. R., Kumar, R. and Tomkins, A. (2016), On the relevance of irrelevant alternatives, *in* 'Proceedings of the 25th International Conference on World Wide Web', International World Wide Web Conferences Steering Committee, pp. 963–973.

Berry, S., Levinsohn, J. and Pakes, A. (1995), 'Automobile prices in market equilibrium', *Econometrica: Journal of the Econometric Society* pp. 841–890.

Bordalo, P., Gennaioli, N. and Shleifer, A. (2012), 'Salience theory of choice under risk', *The Quarterly journal of economics* p. qjs018.

Bruch, E., Feinberg, F. and Lee, K. Y. (2016), 'Extracting multistage screening rules from online dating activity data', *Proceedings of the National Academy of Sciences* **113**(38), 10530–10535.

Chen, S. and Joachims, T. (2016*a*), Modeling intransitivity in matchup and comparison data, *in* 'Proceedings of the Ninth ACM International Conference on Web Search and Data Mining', ACM, pp. 227–236.

Chen, S. and Joachims, T. (2016*b*), Predicting matchups and preferences in context, *in* 'Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, pp. 775–784.

Chierichetti, F., Kumar, R. and Tomkins, A. (2018), Learning a mixture of two multinomial logits, *in* 'International Conference on Machine Learning', pp. 960–968.

Elo, A. E. (1978), *The rating of chessplayers, past and present*, Arco Pub.

Epstein, Z. G., Peysakhovich, A. and Rand, D. G. (2016), 'The good, the bad, and the unflinchingly selfish: Cooperative decision-making can be predicted with high accuracy using only three behavioral types', *Proceedings of the 17th Economics and Computation Conference (EC17)* .

Fox, C. R. and Tversky, A. (1995), 'Ambiguity aversion and comparative ignorance', *The Quarterly Journal of Economics* **110**(3), 585–603.

Fudenberg, D. and Levine, D. K. (2012), 'Timing and self-control', *Econometrica* pp. 1–42.

Fudenberg, D. and Peysakhovich, A. (2014), Recency, records and recaps: Learning and non-equilibrium behavior in a simple decision problem, *in* 'Proceedings of the fifteenth ACM conference on Economics and Computation', ACM, pp. 971–986.

Heikinheimo, H. and Ukkonen, A. (2013), The crowd-median algorithm, *in* 'First AAAI Conference on Human Computation and Crowdsourcing'.

Herbrich, R., Minka, T. and Graepel, T. (2006), Trueskill^{TM}: a bayesian skill rating system, *in* 'Proceedings of the 19th International Conference on Neural Information Processing Systems', MIT Press, pp. 569–576.

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

Kingma, D. and Ba, J. (2014), 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980* .

Kleinberg, J., Mullainathan, S. and Ugander, J. (2017), Comparison-based choices, *in* 'Proceedings of the 2017 ACM Conference on Economics and Computation', ACM, pp. 127–144.

Koppelman, F. S. and Bhat, C. (2006), 'A self instructing course in mode choice modeling: multinomial and nested logit models'.

Kreps, D. (1988), *Notes on the Theory of Choice*, Westview press.

Liberman, V., Samuels, S. M. and Ross, L. (2004), 'The name of the game: Predictive power of reputations versus situational labels in determining prisoner?s dilemma game moves', *Personality and social psychology bulletin* **30**(9), 1175–1185.

List, J. A. (2007), 'On the interpretation of giving in dictator games', *Journal of Political economy* **115**(3), 482–493.

Luce, R. D. (1959), *Individual Choice Behavior a Theoretical Analysis*, John Wiley and sons.

Manski, C. F. (1977), 'The structure of random utility models', *Theory and decision* **8**(3), 229–254.

McFadden, D. (1980), 'Econometric models for probabilistic choice among products', *Journal of Business* pp. S13–S29.

McFadden, D. and Train, K. (2000), 'Mixed mnl models for discrete response', *Journal of applied Econometrics* **15**(5), 447–470.

McFadden, D., Tye, W. B. and Train, K. (1977), *An application of diagnostic tests for the independence from irrelevant alternatives property of the multinomial logit model*, Institute of Transportation Studies, University of California.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. and Dean, J. (2013), Distributed representations of words and phrases and their compositionality, *in* 'Advances in neural information processing systems', pp. 3111–3119.

Muraven, M. and Baumeister, R. F. (2000), 'Self-regulation and depletion of limited resources: Does self-control resemble a muscle?', *Psychological bulletin* **126**(2), 247.

Naecker, J. (2015), 'The lives of others: Predicting donations with non-choice responses'.

Park, S.-J. and Choi, S. (2013), 'A theoretical note on the number of free parameters in the elimination-by-aspects model', *Journal of Mathematical Psychology* **57**(5), 255–259.

Peysakhovich, A. and Rand, D. G. (2015), 'Habits of virtue: Creating norms of cooperation and defection in the laboratory', *Management Science* **62**(3), 631–647.

Peysakhovich, A. and Rand, D. G. (2017), 'In-group favoritism caused by pokemon go and the use of machine learning to learn its mechanisms', *SSRN* .

Ragain, S. and Ugander, J. (2016), Pairwise choice markov chains, *in* 'Advances in Neural Information Processing Systems', pp. 3198–3206.

Resnick, P. and Varian, H. R. (1997), 'Recommender systems', *Communications of the ACM* **40**(3), 56–58.

Rudolph, M., Ruiz, F., Mandt, S. and Blei, D. (2016), Exponential family embeddings, *in* 'Advances in Neural Information Processing Systems', pp. 478–486.

Ruiz, F. J., Athey, S. and Blei, D. M. (2017), 'Shopper: A probabilistic model of consumer choice with substitutes and complements', *arXiv preprint arXiv:1711.03560* .

Schapire, R. E., Cohen, W. W. and Singer, Y. (1998), 'Learning to order things', *Advances in Neural Information Processing Systems* **10**(451), 24.

Shah, N. B., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K. and Wainwright, M. J. (2016), 'Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence', *The Journal of Machine Learning Research* **17**(1), 2049–2095.

Slovic, P. (1995), 'The construction of preference.', *American psychologist* **50**(5), 364.

Tversky, A. (1969), 'Intransitivity of preferences', *Preference, Belief, and Similarity* p. 433.

Tversky, A. (1972), 'Elimination by aspects: A theory of choice.', *Psychological review* **79**(4), 281.

Tversky, A. and Kahneman, D. (1985), The framing of decisions and the psychology of choice, *in* 'Environmental Impact Assessment, Technology Assessment, and Risk Analysis', Springer, pp. 107–129.

Tversky, A. and Simonson, I. (1993), 'Context-dependent preferences', *Management science* **39**(10), 1179–1189.

Wager, S. and Athey, S. (2015), 'Estimation and inference of heterogeneous treatment effects using random forests', *arXiv preprint arXiv:1510.04342* .

Wilks, S. S. (1938), 'The large-sample distribution of the likelihood ratio for testing composite hypotheses', *The Annals of Mathematical Statistics* **9**(1), 60–62.

A. Proofs of Identifiability

There are three main theorems proven in this section of the appendix. The first two are given in the main text.

Theorem 1. *A CDM is identifiable from a dataset $\mathcal{D}$ if $\mathcal{C}_\mathcal{D}$ contains comparisons over all choice sets of two sizes k, k', where at least one of k, k' is not 2 or n.*

Theorem 2. *No rank r CDM, $1 \leq r \leq n$, is identifiable from a dataset $\mathcal{D}$ if $\mathcal{C}_\mathcal{D}$ contains only choices from sets of a single size.*

Theorem 4. *A full rank CDM is identifiable from a dataset $\mathcal{D}$ if and only if the rank of an integer design matrix $G(\mathcal{D})$, properly constructed, is $n(n-1)-1$.*

We begin with a few definitions and simple facts, providing proofs for clarity. Given these facts, main workhorse for proving our identifiability theorems is Lemma 2.

Since the CDM parameters are invariant to constant offsets, we choose (for the full rank case) an offset such that

$$\sum_{x \in \mathcal{X}} \exp\Big(\sum_{z \in \mathcal{X} \setminus x} u_{xz} \Big) = 1. \tag{4}$$

Note that this implies $P_{x,\mathcal{X}} = \exp(\sum_{z \in \mathcal{X}\setminus x} u_{xz})$.

Because the CDM is a logit-based model, it will be much easier to work with log probability ratios. To that end, we define, for a choice set $C \ni x$,

$$\beta_{x,C} = \log(P_{x,C}/\bar{P_C}), \tag{5}$$

where $\bar{P_C} = (\prod_{y \in C} P_{y,C})^{\frac{1}{|C|}}$, the geometric average of the probabilities.

Fact 1. *Given a choice set C of size s, there is a 1-to-1 mapping between the set of log probability ratios $\{\beta_{x,C} : x \in C\}$ and the set of probabilities $\{P_{x,C} : x \in C\}$.*

Proof. Uniquely find $\beta_{x,C}$ $\forall x \in C$ using the mapping in equation (5). Now, for the other direction, observe that $\frac{\exp \beta_{x,C}}{\sum_{y \in C} \exp \beta_{y,C}} = \frac{P_{x,C}/\bar{P_C}}{\sum_{y \in C} P_{y,C}/\bar{P_C}} = P_{x,C}$ $\forall x \in C$. □

Hence, statements regarding identifiability between CDM parameters and the β's can be mapped to statements about identifiability between CDM parameters and probabilities. It will also be much easier to relate differences in CDM parameters of the following pattern, $u_{xy} - u_{yx}$ and $u_{xz} - u_{yz}$ $\forall x \neq y \neq z$, to the β's. Because CDM is shift invariant, these differences between parameters uniquely identify the parameters when the offset constraint (4) is applied.

Fact 2. *Under the offset constraint (4), CDM parameter differences $u_{xy} - u_{yx}$ and $u_{xz} - u_{yz}$, $\forall x \neq y \neq z$, have a 1-to-1 mapping with CDM parameters u_{xy} $\forall x \neq y$.*

Proof. It is immediately obvious that given the parameters, we can uniquely construct the differences. For the other direction, consider that

$$\begin{aligned}
u_{xy} &= u_{xy} + \frac{1}{n-1} \log\Big(\sum_{w \in \mathcal{X}} \exp\Big(\sum_{z \in \mathcal{X}\setminus z} u_{wz} \Big)\Big) \\
&= \frac{1}{n-1} \log\Big(\sum_{w \in \mathcal{X}} \exp\Big(\sum_{z \in \mathcal{X}\setminus w} u_{wz} - u_{xy} \Big)\Big) \\
&= \frac{1}{n-1} \log\Big(\sum_{w \in \mathcal{X}} \exp\Big([u_{wy} - u_{xy}]\mathbf{1}(w \neq y) + \sum_{z \in \mathcal{X}\setminus w,y} u_{wz} - u_{xy} \Big)\Big) \\
&= \frac{1}{n-1} \log\Big(\sum_{w \in \mathcal{X}} \exp\Big([u_{wy} - u_{xy}]\mathbf{1}(w \neq y) + \sum_{z \in \mathcal{X}\setminus w,y} [u_{zy} - u_{xy}] + [u_{yz} - u_{zy}] + [u_{wz} - u_{yz}] \Big)\Big).
\end{aligned}$$

Here the first equality follows because the second term on the right hand size is 0, by the offset constraint (4). The remaining equalities are simply algebraic manipulations. The last equality is purely a function of differences following the aforementioned statement, therefore proving the claim. □

Hence, statements regarding identifiability between CDM parameter differences of the pattern $u_{xy} - u_{yx}$ and $u_{xz} - u_{yz}$ $\forall x \neq y \neq z$ and the β's can be mapped to statements about identifiability between CDM parameters and probabilities.

We now link the above facts with the following: the β's can be conveniently represented in terms of these CDM parameter differences. Using $u \in \mathbb{R}^{n(n-1)}$ to refer to a vectorization of the parameters, with elements of the vector indexed as we have so far (i.e., u_{xy} finds the subset of $(n-1)$ entries associated with item x, and finds the contextual role of item y within those entries), we have the following fact.

Fact 3. *For any set C and any $x \in C$, $\beta_{x,C} = \frac{1}{|C|} \sum_{y \in C \setminus x} \Big([u_{xy} - u_{yx}] + \sum_{z \in C \setminus \{x,y\}} [u_{xz} - u_{yz}] \Big)$.*

Proof. From the definition of $\beta_{x,C}$ in equation (5) we have:

$$\begin{aligned}
\beta_{x,C} &= \log(\frac{P_{x,C}}{P_C}) \\
&= \sum_{z \in C \setminus x} u_{xz} - \frac{1}{|C|} \sum_{y \in C} \sum_{z \in C \setminus y} u_{yz} \\
&= \frac{1}{|C|} \sum_{y \in C \setminus x} \Big([u_{xy} - u_{yx}] + \sum_{z \in C \setminus \{x,y\}} [u_{xz} - u_{yz}] \Big)
\end{aligned}$$

Here the final equality is a rearrangement of terms into the parameter differences of interest. □

We introduce an indicator vector $g_{x,C} \in \mathbb{Z}^{n(n-1)}$ that contains non-zero values at the relevant indices of u so that the final equality can be rewritten as

$$\frac{1}{|C|} \sum_{y \in C \setminus x} \Big([u_{xy} - u_{yx}] + \sum_{z \in C \setminus \{x,y\}} [u_{xz} - u_{yz}] \Big) = \frac{1}{|C|} g_{x,C}^T u. \tag{6}$$

Lastly, we state and prove the following lemma, which will serve as the departure point for the three proofs. Consider a collection $\mathcal{C}_\mathcal{D}$ of unique subsets of the universe $\mathcal{X}$ of sizes 2 or greater, and let $\Omega = \sum_{C \in \mathcal{C}_\mathcal{D}} |C|$ be the sum of the sizes of all the sets. We then refer to a system design matrix $G(\mathcal{C}_\mathcal{D}) \in \mathbb{Z}^{\Omega \times n(n-1)}$ as the linear system relating the parameters u to the scaled log probability ratios $|C|\beta_{x,C}$. We construct such a matrix by concatenating, for each set $C \in \mathcal{C}_\mathcal{D}$, for every item $x \in C$, the indicator vector $g_{x,C}^T$, as defined in (6), as a row.

Lemma 2. *The full rank CDM is identifiable up to a shift for collection $\mathcal{C}_\mathcal{D}$ iff rank$(G(\mathcal{C}_\mathcal{D})) = n(n-1) - 1$.*

Proof. Clearly, rank$(G(\mathcal{C}_\mathcal{D})) \leq n(n-1) - 1$, due to the shift invariance of u. That is, G is only specified in terms of differences of elements in u, and hence null$(G(\mathcal{C}_\mathcal{D})) \ni \mathbf{1}$.

Suppose first that rank$(G(\mathcal{C}_\mathcal{D})) = n(n-1) - 1$. Then, for two vectors $u_1, u_2 \in \mathbb{R}^{n(n-1)}$, if $u_1 \neq \alpha\mathbf{1} + u_2$ for any $\alpha \in \mathbb{R}$ then $\beta_1 = \mathbf{C}^{-1} G u_1 \neq G u_2 = \mathbf{C}^{-1}\beta_2$, where $\mathbf{C}^{-1} \in \mathbb{R}^{\Omega \times \Omega}$ is the diagonal matrix with values are $\frac{1}{|C|}, \forall C \in \mathcal{C}_\mathcal{D}$ (which undoes the scaling of the scaled log probability ratios). Since Fact 1 states that β's have a unique mapping with the choice system probabilities over the collection $\mathcal{C}_\mathcal{D}$, u vectors are identifiable up to a shift for a given set of probabilities over the collection $\mathcal{C}_\mathcal{D}$.

Suppose now that rank$(G(\mathcal{C}_\mathcal{D})) < n(n-1) - 1$. Then, there exists some vector $v \in$ null$(G(\mathcal{C}_\mathcal{D})), v \neq \alpha\mathbf{1}$ for any α, for which $\mathbf{C}^{-1} G(\mathcal{C}_\mathcal{D})(u_1) = \mathbf{C}^{-1} G(\mathcal{C}_\mathcal{D})(u_1 + v)$. Again since the β's uniquely map to the probabilities, there exist two u vectors different beyond a shift that map to the same set of choice system probabilities. Hence, u is not identifiable up to a shift. □

We add as an additional note that under the offset constraint (4), the CDM parameters are uniquely identifiable, following the analysis of Fact 2. Now we proceed to proving the individual theorems, each of which essentially boils down to analyzing the rank of the system design matrix $G(\mathcal{C}_\mathcal{D})$ of collections $\mathcal{C}_\mathcal{D}$ comprised of sets of a single size, of collections $\mathcal{C}_\mathcal{D}$ comprised of sets of multiple sizes, and formalizing the calculation of $G(\mathcal{C}_\mathcal{D})$ for a given dataset.

A.1. Proof of Theorem 1

Proof. It is sufficient to show that the statement holds for the full rank case, as further constraining the parameters using rank conditions does not affect identifiability. Note that the statement of the theorem is a sufficient condition for identifiability, and for low-rank CDMs in particular it is possibly an overly strong requirement.

Consider two different subset sizes s and t, and assume wlog that t is within $[3, n-1]$. For any $\{x, y\}$, consider $C_{wz} \ni \{x, y\}$, $|C_{wz}| = t - 1$, indexed by items $\{w, z\} \in \mathcal{X}$, $\{w, z\} \notin C_{wz}$. Let $A_{wz} = C_{wz} \cup \{w\}$ and $B_{wz} = C_{wz} \cup \{z\}$. Using β^C_{xy} as shorthand for $\beta_{x,C} - \beta_{y,C}$., we have that

$$\beta^{A_{wz}}_{xy} - \beta^{B_{wz}}_{xy} = [u_{xw} - u_{yw}] - [u_{xz} - u_{yz}].$$

Now, if $s < t$, Take $D \ni \{x, y\}$ of size s and A (of size t) such that $D \subset A$. Now,

$$\beta^A_{xy} - \beta^D_{xy} = \sum_{q \in A \setminus D} [u_{xq} - u_{yq}].$$

Then, we can solve for $[u_{xw} - u_{yw}]$ as follows:

$$[u_{xw} - u_{yw}] = \frac{1}{t-s}(\beta^A_{xy} - \beta^D_{xy} + \sum_{q \in A \setminus D} \beta^{A_{wq}}_{xy} - \beta^{B_{wq}}_{xy}).$$

With this relation we see that $[u_{xy} - u_{yx}] = \beta^A_{xy} - \sum_{q \in A \setminus \{x,y\}} [u_{xq} - u_{yq}]$.

If $s > t$, Take D of size s such that $A \subset D$. We then see that $\beta^D_{xy} - \beta^A_{xy} = \sum_{q \in D \setminus A} [u_{xq} - u_{yq}]$, and as before, we can solve for $[u_{xw} - u_{yw}]$ as:

$$[u_{xw} - u_{yw}] = \frac{1}{s-t}(\beta^D_{xy} - \beta^A_{xy} + \sum_{q \in D \setminus A} \beta^{A_{wq}}_{xy} - \beta^{B_{wq}}_{xy}).$$

With this relation we see that $[u_{xy} - u_{yx}] = \beta^D_{xy} - \sum_{q \in D \setminus \{x,y\}} [u_{xq} - u_{yq}]$.

Applying Facts 1 and 2, statements regarding identifiability between CDM parameter differences of the pattern $u_{xy} - u_{yx}$ and $u_{xz} - u_{yz}$ $\forall x \neq y \neq z$ and the β's can be mapped to statements about identifiability between CDM parameters and probabilities. We then conclude that the CDM parameters can be uniquely recovered from probabilities over two choice sets. Thus, comparisons over all choice sets of two sizes uniquely identify the CDM. □

A.2. Proof of Theorem 2

Proof. To prove this claim, we separately consider three conditions on the set size s: $s = 2$, $s = n$, and $3 \leq s \leq n - 1$. For each case, we first demonstrate the result for the full rank CDM and then show that every low rank CDM suffers from the same problem.

In terms of notation, we consider a U "matrix", $U \in \mathbb{R}^{n \times n}$, organizing the parameters u_{xy}, $\forall x \neq y$, with the matrix diagonal taking on arbitrary unused values. For the low rank case, the U matrix is the dot product of the matrix of target vectors $T \in \mathbb{R}^{n \times r}$ and the matrix of context vector $C \in \mathbb{R}^{n \times r}$. Here, the diagonal formed by $t_x \cdot c_x$ can be arbitrary and is unused. We also use β^C_{xy} as shorthand for $\beta_{x,C} - \beta_{y,C}$.

(i) $s = 2$

For any pair $C = \{x, y\}$, $\beta^C_{xy} = u_{xy} - u_{yx}$. Thus, increasing both u_{xy} and u_{yx} by the same value leaves the pairwise probabilities unchanged. Thus the CDM parameter U matrix is only specified up to a symmetric matrix A, where $U + A$ produces the same pairwise probabilities as U.

Any rank r matrix also suffers from the same identifiability issue: consider $T + B$ and $C + F$, where $B = \beta C + \gamma_1 \alpha \beta T$, and $F = \alpha T + \gamma_2 \alpha \beta C$ for $\alpha, \beta \in \mathbb{R}$, $\gamma_1, \gamma_2 \in \{0, 1\}, \gamma_1 \neq \gamma_2$. These scalar parameters form a subset of perturbations that modify the dot product $U = TC^T$ only by a symmetric matrix, thereby leaving the pairwise probabilities unchanged.

(ii) $s = n$

For the full universe $\mathcal{X}$, $\beta^{\mathcal{X}}_{xy} = u_{xy} - u_{yx} + \sum_{z \in \mathcal{X} \setminus \{x,y\}} u_{xy} - u_{yx}$. Consider then any matrix $A \in \mathbb{R}^{n \times n}$ that has $(A - \text{diag(A)})\mathbf{1} = g\mathbf{1}$, where g is a constant and $\mathbf{1} \in \mathbb{R}^{n \times 1}$ is the vector of all ones. This is, any matrix A where the rows (not including the diagonal) all sum to the same constant. Then U and $U + A$ have the same choice probabilities on the full universe set.

For the identifiability problem to transfer to the rank r case, we find $T + \gamma_1 B$ and $C + \gamma_2 F$ where $\gamma_1, \gamma_2 \in \{0, 1\}, \gamma_1 \neq \gamma_2$ such that the perturbation to a U matrix follows the same properties as the matrix A in the full rank case above. We show how to find such a matrix for the rank 1 case, which is sufficient for all rank r. Consider $U = tc^T$, where $t, c \in \mathbb{R}^{n \times 1}$. We may perturb t by a vector $b \in \mathbb{R}^{n \times 1}$ where $b_x = \frac{g}{(c^T \mathbf{1} - c_x)}$, $\forall x$, for any constant g, as long as $(c^T \mathbf{1} - c_x) \neq 0 \; \forall x$. In case $(c^T \mathbf{1} - c_y) = 0$ for any y, set $g = 0$, $b_x = 0 \; \forall x \neq y$, and b_y to any arbitrary value. The perturbation to U is then bc^T, and we leave the reader to verify $((bc^T) - \text{diag}(bc^T))\mathbf{1} = g\mathbf{1}$, thereby not changing the universe probabilities. Similarly, we may perturb c by a vector f, where $f_x = g[\frac{1}{n-1} \sum_z (\frac{1}{t_z}) - \frac{1}{t_x}]$ if $t_x \neq 0$, $\forall x$. In case $t_y = 0$ for some y, set $g = 0$, $f_x = 0$, $\forall x \neq y$, and f_y to any arbitrary value. The perturbation to U is then $t^T f$, and we have $((tf^T) - \text{diag}(tf^T))\mathbf{1} = g\mathbf{1}$, thereby not changing the universe probabilities.

(iii) $3 \leq s \leq n - 1$

For all other set sizes, we again show the identifiability issue for the full rank case, and show that the null space in the parameters also transfers over to the rank r case. Consider any $C \ni \{x, y\}$, $\{w, z\} \notin C$ of size $s - 1$ for any $\{x, y, w, z\}$. Take $C_w = C \cup \{w\}$, and $C_z = C \cup \{z\}$. Note that we can always identify such sets because we are in the size regime $3 \leq s \leq n - 1$. Then, $\beta^{C_w}_{xy} - \beta^{C_z}_{xy} = [u_{xw} - u_{yw}] - [u_{xz} - u_{yz}]$. Thus, given $[u_{xz} - u_{yz}]$ for a single z, we can set $[u_{xw} - u_{yw}] = \beta^{C_w}_{xy} - \beta^{C_z}_{xy} + [u_{xz} - u_{yz}]$, and set $[u_{xy} - u_{yx}] = \beta^{C_z}_{xy} - \sum_{q \in C_z \setminus \{x,y\}} [u_{xq} - u_{yq}] = \beta^{C_z}_{xy} - \sum_{q \in C_z \setminus \{x,y\}} [\beta^{C_z}_{xy} - \beta^{C_q}_{xy}] - (s - 2)[u_{xz} - u_{yz}]$ to keep the choice probabilities unchanged. This invariance implies that the U matrix can be perturbed by the rank-1 matrix $a\mathbf{1}^T$ where $a \in \mathbb{R}^{n \times 1}$ is any vector and the choice probabilities are unchanged.

We can now show that such perturbations to U can be produced in the rank r case by modifiying C. Consider $C + \mathbf{1}b^T$ where $b \in \mathbb{R}^{r \times 1}$. Then, $U = T(C + \mathbf{1}b^T)^T = TC^T + (Tb)\mathbf{1}^T$, which is a perturbation to U of the proper form. Through these three cases, we have now shown that every rank r CDM cannot be uniquely identified even when provided all comparisons of a single choice set size. □

A.3. Proof of Theorem 4

Proof. Consider a dataset of the form $\mathcal{D} = \{(x_j, C_j)\}_{j=1}^m$ of a decision maker making choices: a datapoint j represents a decision scenario, and contains C_j, the context provided in that decision, and $x_j \in C_j$, the item chosen in the context. Recall that $\Omega_{\mathcal{D}} = \sum_{j=1}^m |C_j|$. Construct then a matrix $G(\mathcal{D}) \in \mathbb{Z}^{\Omega_{\mathcal{D}} \times n(n-1)}$ by concatenating, for every datapoint j, for every item $x \in C_j$, the indicator vector g^T_{x,C_j} as defined in equation (6) as a row. Denoting $\mathcal{C}_{\mathcal{D}}$ as the collection of unique choice sets in dataset $\mathcal{D}$, it is clear that $\text{rank}(G(\mathcal{D})) = \text{rank}(G(\mathcal{C}_{\mathcal{D}}))$, where the latter matrix is defined as in Lemma 2 for the collection $\mathcal{C}_{\mathcal{D}}$. This equality of ranks follows from the fact that the set of unique rows of $G(\mathcal{D})$ are the same as those in $G(\mathcal{C}_{\mathcal{D}})$, and repeated rows do not change the rank of a matrix. Thus, we can directly test whether a dataset results in an identifiable CDM by testing the rank of $G(\mathcal{D})$. □

B. Convergence proof

We restate and then prove Theorem 3.

Theorem 3. *Let $u^\star$ denote the true CDM model from which data is drawn. Let $\hat{u}_{MLE}$ denote the maximum likelihood solution. Assume $\mathcal{C}_\mathcal{D}$ identifies the CDM. For any $u^\star \in \mathcal{U}_\mathcal{B} = \{u \in \mathbb{R}^d : \|u\|_\infty \leq B, \boldsymbol{1}^T u = 0\}$, and expectation taken over the dataset $\mathcal{D}$ generated by the CDM model,*

$$\mathbb{E}\big[\, \|\hat{u}_{MLE}(\mathcal{D}) - u^\star\|_2^2 \,\big] \leq \frac{d}{m} \frac{\alpha k_{max}^2}{k_{min}},$$

where k_{max} and k_{min} respectively refer to the maximum and minimum sizes of choice sets in the dataset, and α is a constant that depends on B, k_{max} and the spectrum of the design matrix $G(\mathcal{D})$.

Proof. We describe the sampling process as follows using the same notation as before. Given some true CDM $u^\star \in \mathcal{U}_B$, for each datapoint $j \in [m]$ we have the probability of choosing item x from set C_j as

$$\mathbb{P}(y_j = x | u^\star, C_j) = \frac{\exp(\sum_{z \in C_j \setminus x} u^\star_{xz})}{\sum_{y \in C_j} \exp(\sum_{z \in C_j \setminus y} u^\star_{yz}))}.$$

We now introduce notation that will let us represent the above expression in a more compact manner. Because our datasets involve choice sets of multiple sizes, we use $k_j \in [k_{\min}, k_{\max}]$ to denote the choice set size for datapoint j. Extending a similar concept in (Shah et al., 2016) to the multiple set sizes, and the more complex structure of the CDM, we then define matrices $E_{j,k_j} \in \mathbb{R}^{d \times k_j}$, $\forall j \in [m]$ as follows: E_{j,k_j} has a column for every item $y \in C_j$ (and hence k_j columns), and the column corresponding to item $y \in C_j$ has a one at the position of each u_{yz} for $z \in C_j \setminus y$, and zero otherwise. This construction allows us to write the familiar expressions $\sum_{z \in C_j \setminus y} u_{yz}$, for each y, simply as a single vector-matrix product $u^T E_{j,k_j} = [\sum_{z \in C_j \setminus y_1} u_{y_1 z}, \sum_{z \in C_j \setminus y_2} u_{y_2 z}, \dots \sum_{z \in C_j \setminus y_{k_j}} u_{y_{k_j} z}] \in \mathbb{R}^{1 \times k_j}$.

Next, we define a collection of functions $F_k : \mathbb{R}^k \mapsto [0, 1], \forall k \in [k_{\min}, k_{\max}]$ as

$$F_k([x_1, x_2, \dots, x_k]) = \frac{\exp(x_1)}{\sum_{l=1}^k \exp(x_l)},$$

where the numerator always corresponds to the first entry of the input. These functions F_k have several properties that will become useful later in the proof. First, it is easy to verify that all F_k are shift-invariant, that is, $F_k(x) = F_k(x + c\mathbf{1})$, for any scalar c. Next, we show that all F_k are strongly log-concave, that is, $\nabla^2(-\log(F_k(x))) \succeq H_k$ for some $H_k \in \mathbb{R}^{k \times k}$, $\lambda_2(H_k) > 0$. The proof for this property stems directly from its counterpart in (Shah et al., 2016), as multiple set sizes does not affect the result. We compute the Hessian as:

$$\nabla^2(-\log(F_k(x))) = \frac{\exp(x_1)}{(\langle \exp(x), 1 \rangle)^4} (\langle \exp(x), 1 \rangle \mathbf{diag}(\exp(x)) - \exp(x) \exp(x)^T),$$

where $\exp(x) = [e^{x_1}, \dots, e^{x_k}]$. Note that

$$\begin{aligned} v^T \nabla^2(-\log(F_k(x))) v &= \frac{\exp(x_1)}{(\langle \exp(x), 1 \rangle)^4} v^T (\langle \exp(x), 1 \rangle \mathbf{diag}(\exp(x)) - \exp(x) \exp(x)^T) v \\ &= \frac{\exp(x_1)}{(\langle \exp(x), 1 \rangle)^4} (\langle \exp(x), 1 \rangle \langle \exp(x), v^2 \rangle - \langle \exp(x), v \rangle^2) \\ &\geq 0, \end{aligned}$$

where v^2 refers to the element-wise square operation on vector v. While the final inequality is an expected consequence of the positive semidefiniteness of the Hessian, we note that it also follows from an application of Cauchy-Schwarz to the vectors $\sqrt{\exp(x)}$ and $\sqrt{\exp(x)} \odot v$, and is thus an equality *if and only if* $v \in \text{span}(\mathbf{1})$. Thus, we have that the smallest eigenvalue $\lambda_1(\nabla^2(-\log(F_k(x)))) = 0$ is associated with the vector $\mathbf{1}$, a property we expect from shift invariance, and that the second smallest eigenvalue $\lambda_2(\nabla^2(-\log(F_k(x)))) > 0$. Thus, we can state that

$$\nabla^2(-\log(F_k(x))) \succeq H_k = \beta_k (I - \frac{1}{k} \mathbf{1}\mathbf{1}^T), \tag{7}$$

where

$$\beta_k := \min_{x \in [-(k-1)B,(k-1)B]^k} \lambda_2(\nabla^2(-\log(F_k(x)))), \tag{8}$$

and it's clear that $\beta_k > 0$. The minimization is taken over $x \in [-(k-1)B, (k-1)B]^k$ since each x_i is a sum of $k-1$ values of the u vector, each entry of which is in $[-B, B]$. We conclude that all F_k are strongly log-concave.

As a final notational addition, in the same manner as (Shah et al., 2016) but accounting for multiple set sizes, we define k permutation matrices $R_{1,k}, \ldots, R_{k,k} \in \mathbb{R}^{k,k}, \forall k \in [k_{\min}, k_{\max}]$, representing k cyclic shifts in a fixed direction. That is, these matrices allow for the cycling of the entries of row vector $v \in \mathbb{R}^{1\times k}$ so that any entry can become the first entry of the vector, for any of the relevant k. This construction allows us to represent any choice made from the choice set C_j as the first element of the vector x that is input to F, thereby placing it in the numerator.

Given the notation introduced above, we can now state the probability of choosing the item x from set C_j compactly as:

$$\mathbb{P}(y_j = x | u^\star, C_j) = \mathbb{P}(y_j = x | u^\star, k_j, E_{j,k_j}) = F_{k_j}(u^{\star T} E_{j,k_j} R_{x,k_j}).$$

We can then rewrite the full-rank CDM likelihood as

$$\sup_{u \in \mathcal{U}_B} \prod_{(x_j,k_j,E_{j,k_j}) \in \mathcal{D}} F_{k_j}(u^T E_{j,k_j} R_{x_j,k_j}),$$

and the scaled negative log-likelihood as

$$\ell(u) = -\frac{1}{m} \sum_{(x_j,k_j,E_{j,k_j}) \in \mathcal{D}} \log(F_{k_j}(u^T E_{j,k_j} R_{x_j,k_j})) = -\frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{k_j} \mathbf{1}[y_j = i] \log(F_{k_j}(u^T E_{j,k_j} R_{i,k_j})).$$

Thus,

$$\hat{u}_{\text{MLE}} = \arg\max_{u \in \mathcal{U}_B} \ell(u).$$

The compact notation makes the remainder of the proof a straightforward application of results from convex analysis: we first demonstrate that the scaled negative log-likelihood is strongly convex with respect to a semi-norm[2], and we use this property to show the proximity of the MLE to the optimal point as desired. The remainder of the proof exactly mirrors that in (Shah et al., 2016) with a few extra steps of accounting created by the multiple set sizes. The notable exception is in the definition of L, and conditions about its eigenvalues that tie back to the previous results about identifiability. While in (Shah et al., 2016) there is a clear connection of L to the graph Laplacian matrix of the item comparison graph, it is unclear here how to interpret L as a graph Laplacian.

First, we have the gradient of the negative log-likelihood as

$$\nabla \ell(u) = -\frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{k_j} \mathbf{1}[y_j = i] E_{j,k_j} R_{i,k_j} \nabla \log(F_{k_j}(u^T E_{j,k_j} R_{i,k_j})),$$

and the Hessian as

$$\nabla^2 \ell(u) = -\frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{k_j} \mathbf{1}[y_j = i] E_{j,k_j} R_{i,k_j} \nabla^2 \log(F_{k_j}(u^T E_{j,k_j} R_{i,k_j})) R_{i,k_j}^T E_{j,k_j}^T.$$

[2] A semi-norm is a norm that allows non-zero vectors to have zero norm.

We then have, for any vector $z \in \mathbb{R}^d$,

$$\begin{aligned}
z^T \nabla^2 \ell(u) z &= -\frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{k_j} \mathbf{1}[y_j = i] z^T E_{j,k_j} R_{i,k_j} \nabla^2 \log(F_{k_j}(u^T E_{j,k_j} R_{i,k_j})) R_{i,k_j}^T E_{j,k_j}^T z \\
&= \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{k_j} \mathbf{1}[y_j = i] z^T E_{j,k_j} R_{i,k_j} \nabla^2 (-\log(F_{k_j}(u^T E_{j,k_j} R_{i,k_j}))) R_{i,k_j}^T E_{j,k_j}^T z \\
&\geq \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{k_j} \mathbf{1}[y_j = i] z^T E_{j,k_j} R_{i,k_j} H_k R_{i,k_j}^T E_{j,k_j}^T z \\
&= \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{k_j} \mathbf{1}[y_j = i] z^T E_{j,k_j} R_{i,k_j} \beta_{k_j} (I - \frac{1}{k_j} \mathbf{1}\mathbf{1}^T) R_{i,k_j}^T E_{j,k_j}^T z \\
&\geq \frac{\beta_{k_{\max}}}{k_{\max}} \frac{1}{m} \sum_{j=1}^{m} \sum_{i=1}^{k_j} \mathbf{1}[y_j = i] z^T E_{j,k_j} (k_j I - \mathbf{1}\mathbf{1}^T) E_{j,k_j}^T z \\
&= \frac{\beta_{k_{\max}}}{k_{\max}} \frac{1}{m} \sum_{j=1}^{m} z^T E_{j,k_j} (k_j I - \mathbf{1}\mathbf{1}^T) E_{j,k_j}^T z.
\end{aligned}$$

The first line follows from applying the definition of the Hessian. The second line follows from pulling the negative sign into the ∇^2 term. The third and fourth line follow from (7), strong log-concavity of all F_k. The fifth line follows from the pulling out k_j and lower bounding it with $k_{\max}$, pulling out β_{k_j} and lower bounding it with $\beta_{k_{\max}}$ and recognizing that H_k is invariant to permutation matrices. The sixth line follows from removing the inner sum since the terms are independent of i. Now, defining the matrix L as

$$L = \frac{1}{m} \sum_{j=1}^{m} E_{j,k_j} (k_j I - \mathbf{1}\mathbf{1}^T) E_{j,k_j}^T,$$

we first note a few properties of L. First, it is easy to verify that $L\mathbf{1} = 0$, and hence $\text{span}(\mathbf{1}) \subseteq \text{null}(L)$. Moreover, we now show that $\lambda_2(L) > 0$, that is, $\text{null}(L) \subseteq \text{span}(\mathbf{1})$. Consider the matrix $G(\mathcal{D})$ in Theorem 4 []This is referring to matrix rank test theorem in the identifiability doc, will show up when merged/want to avoid using xr. Simple calculations show that $L = G(\mathcal{D})^T G(\mathcal{D}) \succeq 0$. As a consequence of the properties of matrix rank, we then have that $\text{rank}(G(\mathcal{D})) = \text{rank}(L)$. Thus, from Theorem 4, we have that if the dataset $\mathcal{D}$ identifies the CDM, $\text{rank}(L) = d - 1$, and hence $\lambda_2(L) > 0$. With this matrix, we can write,

$$z^T \nabla^2 \ell(u) z \geq \frac{\beta_{k_{\max}}}{k_{\max}} z^T L z = \frac{\beta_{k_{\max}}}{k_{\max}} ||z||_L^2,$$

which is equivalent to stating that $\ell(u)$ is $(\frac{\beta_{k_{\max}}}{k_{\max}})$-strongly convex with respect to the L semi-norm at all $u \in \mathcal{U}_B$. Since $u^\star, \hat{u}_{\text{MLE}} \in \mathcal{U}_B$, strong convexity implies that

$$\frac{\beta_{k_{\max}}}{k_{\max}} ||\hat{u}_{\text{MLE}} - u_\star||_L^2 \leq \ell(\hat{u}_{\text{MLE}}) - \ell(u^\star) - \langle \nabla \ell(u^\star), \hat{u}_{\text{MLE}} - u^\star \rangle.$$

Further, we have

$$\begin{aligned}
\ell(\hat{u}_{\text{MLE}}) - \ell(u^\star) - \langle \nabla \ell(u^\star), \hat{u}_{\text{MLE}} - u^\star \rangle &\leq -\langle \nabla \ell(u^\star), \hat{u}_{\text{MLE}} - u^\star \rangle \\
&\leq |(\hat{u}_{\text{MLE}} - u^\star)^T \nabla \ell(u^\star)| \\
&= |(\hat{u}_{\text{MLE}} - u^\star)^T L^{\frac{1}{2}} L^{\frac{1}{2}\dagger} \nabla \ell(u^\star)| \\
&\leq ||L^{\frac{1}{2}} (\hat{u}_{\text{MLE}} - u^\star)||_2 ||L^{\frac{1}{2}\dagger} \nabla \ell(u^\star)||_2 \\
&= ||\hat{u}_{\text{MLE}} - u^\star||_L ||\nabla \ell(u^\star)||_{L^\dagger}.
\end{aligned}$$

Here the third line follows from the fact that $\mathbf{1}^T(\hat{u}_{\text{MLE}} - u^\star) = 0$, and so $(\hat{u}_{\text{MLE}} - u^\star) \perp \text{null}(L)$, which also implies that $(\hat{u}_{\text{MLE}} - u^\star) \perp \text{null}(L^{\frac{1}{2}})$, and so $(\hat{u}_{\text{MLE}} - u^\star)L^{\frac{1}{2}}L^{\frac{1}{2}\,\dagger} = (\hat{u}_{\text{MLE}} - u^\star)$. The fourth line follows from Cauchy-Schwarz. Thus, we can conclude that

$$||\hat{u}_{\text{MLE}} - u^\star||_L^2 \leq \frac{k_{\text{max}}^2}{\beta_{k_{\text{max}}}^2}||\nabla\ell(u^\star)||_{L^\dagger}^2 = \frac{k_{\text{max}}^2}{\beta_{k_{\text{max}}}^2}\nabla\ell(u^\star)^T L^\dagger \nabla\ell(u^\star).$$

Now, all that remains is bounding the term on the right hand side. Recall the expression for the gradient

$$\nabla\ell(u^\star) = -\frac{1}{m}\sum_{j=1}^{m}\sum_{i=1}^{k_j}\mathbf{1}[y_j = i]E_{j,k_j}R_{i,k_j}\nabla\log(F_{k_j}(u^{\star T}E_{j,k_j}R_{i,k_j})) = -\frac{1}{m}\sum_{j=1}^{m}E_{j,k_j}V_{j,k_j},$$

where in the equality we have defined

$$V_{j,k_j} := \sum_{i=1}^{k_j}\mathbf{1}[y_j = i]R_{i,k_j}\nabla\log(F_{k_j}(u^{\star T}E_{j,k_j}R_{i,k_j})).$$

Now, we have

$$(\nabla\log(F_k(x)))_l = \mathbf{1}[l = 1] - \frac{\exp(x_l)}{\sum_{p=1}^{k}\exp(x_p)}, \tag{9}$$

and so $\langle\nabla\log(F_k(x)), \mathbf{1}\rangle = \frac{1}{F_k(x)}\langle\nabla F_k(x), \mathbf{1}\rangle = \sum_{l=1}^{k}(\mathbf{1}[l = 1] - \frac{\exp(x_l)}{\sum_{p=1}^{k}\exp(x_p)}) = 0$, and hence, $V_{j,k_j}^T\mathbf{1} = 0$.

We now consider the matrix $M_k = (I - \frac{1}{k}\mathbf{1}\mathbf{1}^T)$. We note that M_k has rank $k-1$, with its nullspace corresponding to the span of the ones vector. We state the following identities:

$$M_k = M_k^\dagger = M_k^{\frac{1}{2}} = M_k^{\dagger\,\frac{1}{2}}.$$

Thus we have $M_{k_j}V_{j,k_j} = M_{k_j}{}^{\frac{1}{2}}M_{k_j}^{\frac{1}{2}}V_{j,k_j} = M_kM_k^\dagger V_{j,k_j} = V_{j,k_j}$, where the last equality follows since V_{j,k_j} is orthogonal to the nullspace of M_{k_j}. Now, taking expectations over the dataset, we have,

$$\begin{aligned}
\mathbb{E}[V_{j,k_j}] &= \mathbb{E}\Big[\sum_{i=1}^{k_j}\mathbf{1}[y_j = i]R_{i,k_j}\nabla\log(F_{k_j}(u^{\star T}E_{j,k_j}R_{i,k_j}))\Big] \\
&= \sum_{i=1}^{k_j}\mathbb{E}\Big[\mathbf{1}[y_j = i]\Big]R_{i,k_j}\nabla\log(F_{k_j}(u^{\star T}E_{j,k_j}R_{i,k_j})) \\
&= \sum_{i=1}^{k_j}F_{k_j}(u^{\star T}E_{j,k_j}R_{i,k_j})R_{i,k_j}\nabla\log(F_{k_j}(u^{\star T}E_{j,k_j}R_{i,k_j})) \\
&= \sum_{i=1}^{k_j}R_{i,k_j}\nabla F_{k_j}(u^{\star T}E_{j,k_j}R_{i,k_j}) \\
&= \nabla_z\Big(\sum_{i=1}^{k_j}F_{k_j}(z^TR_{i,k_j})\Big) = \nabla_z(1) = 0.
\end{aligned}$$

Here, the third equality follows from applying the expectation to the indicator and retrieving the true probability. The fourth line follows from applying the definition of gradient of log, and the final line from performing a change of variables $z = u^{\star T}E_{j,k_j}$, pulling out the gradient and undoing the chain rule, and finally, recognizing that the expression sums to 1 for any z, thus resulting in a 0 gradient.

Next, we have

$$\begin{aligned}
\mathbb{E}[\ell(u^\star)^T L^\dagger \nabla \ell(u^\star)] &= \frac{1}{m^2}\mathbb{E}\Big[\sum_{j=1}^m \sum_{l=1}^m V_{j,k_j}^T E_{j,k_j}^T L^\dagger E_{l,k_l} V_{l,k_l}\Big] \\
&= \frac{1}{m^2}\mathbb{E}\Big[\sum_{j=1}^m \sum_{l=1}^m V_{j,k_j}^T M_{k_j}{}^{\frac{1}{2}} E_{j,k_j}^T L^\dagger E_{l,k_l} M_{k_l}{}^{\frac{1}{2}} V_{l,k_l}\Big] \\
&= \frac{1}{m^2}\mathbb{E}\Big[\sum_{j=1}^m V_{j,k_j}^T M_{k_j}{}^{\frac{1}{2}} E_{j,k_j}^T L^\dagger E_{j,k_j} M_{k_j}{}^{\frac{1}{2}} V_{j,k_j}\Big] \\
&\leq \frac{1}{m}\mathbb{E}\Big[\sup_{l\in[m]} ||V_{l,k_l}||_2^2\Big]\mathbf{tr}\Big(\frac{1}{m}\sum_{j=1}^m M_{k_j}{}^{\frac{1}{2}} E_{j,k_j}^T L^\dagger E_{j,k_j} M_{k_j}{}^{\frac{1}{2}}\Big),
\end{aligned}$$

where the second line follows from identities of the M matrix, the third from the independence of the V_{j,k_j}, and the fourth from an upper bound of the quadratic form. We then have that,

$$\begin{aligned}
\sup_{j\in[m]} ||V_{j,k_j}||_2^2 &= \sup_{j\in[m]} \sum_{i=1}^{k_j} \mathbf{1}[y_j = i]\nabla \log(F_{k_j}(u^T E_{j,k_j} R_{i,k_j}))^T R_{i,k_j}^T R_{i,k_j} \nabla \log(F_{k_j}(u^T E_{j,k_j} R_{i,k_j})) \\
&= \sup_{j\in[m]} \sum_{i=1}^{k_j} \mathbf{1}[y_j = i]\nabla \log(F_{k_j}(u^T E_{j,k_j} R_{i,k_j}))^T \nabla \log(F_{k_j}(u^T E_{j,k_j} R_{i,k_j})) \\
&= \sup_{j\in[m]} \sum_{i=1}^{k_j} \mathbf{1}[y_j = i]||\nabla \log(F_{k_j}(u^T E_{j,k_j} R_{i,k_j}))||_2^2 \\
&\leq \sup_{v\in[-(k_{\max}-1)B,(k_{\max}-1)B]^{k_{\max}}} ||\nabla \log(F_{k_{\max}}(v))||_2^2 \leq 2,
\end{aligned}$$

where $R_{i,k_j}^T R_{i,k_j}$ in the first line is simply the identity matrix. For the final line, recalling the expression for the log gradient of F_k in equation (9), it is straightforward to show that $\sup_{v\in[-(k_{\max}-1)B,(k_{\max}-1)B]^{k_{\max}}} ||\nabla \log(F_{k_{\max}}(v))||_2^2$ is always upper bounded by 2.

Next we note that

$$\begin{aligned}
\mathbf{tr}\Big(\frac{1}{m}\sum_{j=1}^m M_{k_j}{}^{\frac{1}{2}} E_{j,k_j}^T L^\dagger E_{j,k_j} M_{k_j}{}^{\frac{1}{2}}\Big) &= \frac{1}{m}\sum_{j=1}^m \mathbf{tr}\Big(M_{k_j}{}^{\frac{1}{2}} E_{j,k_j}^T L^\dagger E_{j,k_j} M_{k_j}{}^{\frac{1}{2}}\Big) \\
&= \frac{1}{m}\sum_{j=1}^m \mathbf{tr}\Big(L^\dagger E_{j,k_j} M_{k_j}{}^{\frac{1}{2}} M_{k_j}{}^{\frac{1}{2}} E_{j,k_j}^T\Big) \\
&\leq \frac{1}{k_{\min}}\mathbf{tr}\Big(L^\dagger L\Big) \\
&= \frac{d-1}{k_{\min}} \leq \frac{d}{k_{\min}}.
\end{aligned}$$

Putting it all together, we have that

$$\mathbb{E}[||\hat{u}_{\text{MLE}} - u^\star||_L^2] \leq \frac{2dk_{\max}^2}{mk_{\min}\beta_{k_{\max}}^2}$$

The final step is noting that $||\hat{u}_{\text{MLE}} - u^\star||_L^2 = (\hat{u}_{\text{MLE}} - u^\star)L(\hat{u}_{\text{MLE}} - u^\star) \geq \lambda_2(L)||\hat{u}_{\text{MLE}} - u^\star||_L^2$, since $\hat{u}_{\text{MLE}} - u^\star \perp \text{null}(L)$. Then, we have:

$$\mathbb{E}[||\hat{u}_{\text{MLE}} - u^\star||_2^2] \leq \frac{d}{m}\frac{k_{\max}^2}{k_{\min}}\frac{2}{\lambda_2(L)\beta_{k_{\max}}^2}.$$

Now, setting

$$c_{B,k_{\max}} := \frac{2}{\lambda_2(L)\beta^2_{k_{\max}}},$$

we retrieve the theorem statement,

$$\mathbb{E}[||\hat{u}_{\text{MLE}} - u^\star||_2^2] \leq \frac{d}{m}\frac{c_{B,k_{\max}}k^2_{\max}}{k_{\min}}.$$

We close with some remarks about $c_{B,k_{\max}}$. The quantity $\beta_{k_{\max}}$, defined in equation (8), serves as the important term that approaches 0 as a function of B and $k_{\max}$, requiring that the former be bounded. Finally, $\lambda_2(L)$ is a parallel to the requirements on the algebraic connectivity of the comparison graph in (Shah et al., 2016) for the multinomial setting. Though the object L here appears similar to the graph Laplacian L in that work, there are major differences that are most worthy of further study. □