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:

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

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 $\beta_1 = \mathbf{C}^{-1}Gu_1 \neq Gu_2 = \mathbf{C}^{-1}\beta_2$, 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 $\mathcal{X}$, $\beta_{xy}^{\mathcal{X}} = u_{xy} - u_{yx} + \sum_{z \in \mathcal{X} \setminus \{x,y\}} u_{xy} - u_{yx}$. 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, $\beta_{xy}^{C_w} - \beta_{xy}^{C_z} = [u_{xw} - u_{yw}] - [u_{xz} - u_{yz}]$. Thus, given $[u_{xz} - u_{yz}]$ for a single z, we can set $[u_{xw} - u_{yw}] = \beta_{xy}^{C_w} - \beta_{xy}^{C_z} + [u_{xz} - u_{yz}]$, and set $[u_{xy} - u_{yx}] = \beta_{xy}^{C_z} - \sum_{q \in C_z \setminus \{x,y\}} [u_{xq} - u_{yq}] = \beta_{xy}^{C_z} - \sum_{q \in C_z \setminus \{x,y\}} [\beta_{xy}^{C_z} - \beta_{xy}^{C_q}] - (s - 2)[u_{xz} - u_{yz}]$ 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_{x,C_j}^T 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,*

$$\mathbb{E}\big[\, \|\hat{u}_{MLE}(\mathcal{D}) - u^\star\|_2^2 \,\big] \leq c_{B,k_{max}} \frac{d-1}{m},$$

where k_{max} refers to the maximum choice set size in the dataset, and $c_{B,k_{max}}$ 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:

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

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

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

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* $v \in \mathrm{span}(\mathbf{1})$. 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 $z \in \mathbb{R}^d$,

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

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 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

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

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. Define a matrix $X(\mathcal{D}) = \mathbf{C}_{\mathcal{D}}^{-1} G(\mathcal{D})$, where $\mathbf{C}_{\mathcal{D}}^{-1} \in \mathbb{R}^{\Omega_{\mathcal{D}} \times \Omega_{\mathcal{D}}}$ is the diagonal matrix with values are $\frac{1}{k_j}$, for every datapoint j, for every item $x \in C_j$. Simple calculations show that,

$$L = \frac{1}{m} X(\mathcal{D})^T X(\mathcal{D}) \succeq 0.$$

As a consequence of the properties of matrix rank, we then have that $\text{rank}(L) = \text{rank}(X(\mathcal{D})) = \text{rank}(G(\mathcal{D}))$. 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,

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

which is equivalent to stating that $\ell(u)$ is $\beta_{k_{\max}}$-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

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

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

$$||\hat{u}_{\text{MLE}} - u^\star||_L^2 \leq \frac{1}{\beta_{k_{\max}}^2}||\nabla \ell(u^\star)||_{L^\dagger}^2 = \frac{1}{\beta_{k_{\max}}^2}\nabla \ell(u^\star)^T L^\dagger \nabla \ell(u^\star).$$

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 $V_{j,k_j} \in \mathbb{R}^{k_j}$ as

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

Useful in our analysis will be an alternate expression for the gradient,

$$\nabla \ell(u^\star) = -\frac{1}{m}\sum_{j=1}^{m}E_{j,k_j}V_{j,k_j} = -\frac{1}{m}X(\mathcal{D})^T V,$$

where we have defined $V \in \mathbb{R}^{\Omega_{\mathcal{D}}}$ as the concatenation of all V_{j,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. We note that an immediate consequence of the above result is that $\mathbb{E}[V] = 0$, since V is simply a concatenation of the individual V_{j,k_j}.

Next, we have

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

where the second line follows from identities of the M matrix, the third from the independence of the V_{j,k_j}, the fourth from an upper bound of the quadratic form, the fifth from the properties of trace, the sixth from the definition of the matrix L, and the last from the value of the trace, which is simply the identity matrix with one zero entry in the diagonal. 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. We again note that an immediate consequence of this is that the *absolute value* of every element of V is also upper bounded by 2.

Bringing this back to $\mathbb{E}[\nabla\ell(u^\star)^T L^\dagger\nabla\ell(u^\star)]$, we have that

$$\mathbb{E}[\nabla\ell(u^\star)^T L^\dagger\nabla\ell(u^\star)] \leq \frac{2(d-1)}{m}.$$

This immediately yields a bound on the expected risk in the L semi-norm, which is,

$$\mathbb{E}[||\hat{u}_{\text{MLE}} - u^\star||_L^2] \leq \frac{2(d-1)}{m\beta_{k_{\max}}^2}.$$

By 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 $\hat{u}_{\text{MLE}} - u^\star \perp \text{null}(L)$, we can translate this into the ℓ_2 norm:

$$\mathbb{E}[||\hat{u}_{\text{MLE}} - u^\star||_2^2] \leq \frac{2(d-1)}{m\lambda_2(L)\beta_{k_{\max}}^2}.$$

Now, setting

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

we retrieve the theorem statement,

$$\mathbb{E}\big[\, \|\hat{u}_{\text{MLE}}(\mathcal{D}) - u^\star\|_2^2 \,\big] \le c_{B,k_{\max}} \frac{d-1}{m}.$$

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. □

C. Auxiliary Material

C.1. Removing Constraints from $\mathcal{M}_2$

We restate $\mathcal{M}_2$ for convenience.

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

Here, a counting exercise reveals that there are n^2 variables (n from the $v(x)$ and $n(n-1)$ from the $v(x \mid \{u\})$ and there are $n+1$ linear equality constraints (1 from the constraint on $v(x)$, and n from the constraints on $v(x \mid \{u\})$). Our goal in this step is to find a parameterization such that there remains only one equality constraint and $n(n-1)$ variables. To do this, we define the variable $u_{xz} \forall x \neq z \in \mathcal{X}$, and subject it to the constraint that $\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}\setminus x} u_{xy} = 0$. Set $v(z) = -\frac{1}{n-1} \sum_{x \in \mathcal{X}\setminus z} u_{xz}$, $\forall z$ and set $v(x \mid \{z\}) = u_{xz} - \frac{1}{n-1} \sum_{y \in \mathcal{X}\setminus z} u_{yz}$. We may then verify that $\sum_{z \in \mathcal{X}} v(z) = \frac{1}{n-1} \sum_{z \in \mathcal{X}} \sum_{x \in \mathcal{X}\setminus z} u_{xz} = 0$ because of the constraint on u. We can also verify that

$$\sum_{x \in \mathcal{X}\setminus z} v(x \mid \{z\}) = \sum_{x \in \mathcal{X}\setminus z} [u_{xz} - \frac{1}{n-1} \sum_{y \in \mathcal{X}\setminus z} u_{yz}] = 0.$$

Thus, the assignment is feasible for any u satisfying its sum constraint. Substituting the assignments into the expression for the probability, we have,

$$\begin{aligned} P(x \mid C) &= \frac{\exp(-\frac{1}{n-1} \sum_{w \in \mathcal{X}\setminus x} u_{wx} + \sum_{z \in C\setminus x} [u_{xz} - \frac{1}{n-1} \sum_{w \in \mathcal{X}\setminus z} u_{wz}])}{\sum_{y \in C} \exp(-\frac{1}{n-1} \sum_{w \in \mathcal{X}\setminus y} u_{wy} + \sum_{z \in C\setminus y} [u_{yz} - \frac{1}{n-1} \sum_{w \in \mathcal{X}\setminus z} u_{wz}])} \\ &= \frac{\exp(-\sum_{z \in C} \frac{1}{n-1} \sum_{w \in \mathcal{X}\setminus z} u_{wz} + \sum_{z \in C\setminus x} u_{xz})}{\sum_{y \in C} \exp(-\sum_{z \in C} \frac{1}{n-1} \sum_{w \in \mathcal{X}\setminus z} u_{wz} + \sum_{z \in C\setminus y} u_{yz})} \\ &= \frac{\exp(\sum_{z \in C\setminus x} u_{xz})}{\sum_{y \in C} \exp(\sum_{z \in C\setminus y} u_{yz})} \end{aligned}$$

where the third step follows from $\sum_{z \in C} v(z)$ terms cancelling out across the numerator and denominator. Thus, every u that satisfies the constraint $\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{X}\setminus x} u_{xy} = 0$ always satisfies the constraints on $v(x)$ and $v(x \mid \{z\})$, and hence the new $P(x \mid C)$ is a valid reparameterization.

C.2. Examples of IIA Violations Handled by CDM

Copying over the example from the main text, consider a choice system on $\mathcal{X} = \{a, b, c\}$ where

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

Assuming IIA implies that we can immediately infer the parameters. Using the notation from model $\mathcal{M}_1$, we have that $v(a) = 1.386$, $v(b) = v(c) = -.693$. These three values sum to zero, as per the constraint. We may then state the three

relevant pairwise probabilities using these parameters:

$$P(a \mid \{a,b\}) = 0.89, \;\; P(b \mid \{b,c\}) = 0.50, \;\; P(c \mid \{a,c\}) = 0.11$$

Thus, IIA is full specified and constrained this way. This is in contrast to the CDM, which can specify any arbitrary pairwise probability. As an example, we can model an extreme preference reversal as follows:

$$P(a \mid \{a,b\}) = 0.11, \;\; P(b \mid \{b,c\}) = 0.50, \;\; P(c \mid \{a,c\}) = 0.89$$

Although b is disproportionately preferred over a in the pair setting, the story almost reverses in the triplet setting. The CDM parameters corresponding to this example are: $[u_{ab}, u_{ac}, u_{ba}, u_{bc}, u_{ca}, u_{cb}] = [.693, .693, 2.784, -3.477, 2.784, -3.477]$, where the sum to 0 constraint is being enforced. This notion of preference reversal, and CDM's ability to accommodate it, is actually fairly versatile. Indeed, many of the storied effects in discrete choice, such as those of Similarity Aversion, Asymmetric Dominance, and the Compromise Effect are simply instances of preference reversal. We illustrate this using the following table, adapted from (Srivastava and Schrater, 2012). $P_{x,A}$ is used to denote the probability of choosing an item x from a set A.

Table 1. An Overview of the Various Effects

Name	Effect	Constraints
Preference Reversal	$P_{x,\{x,y\}} > P_{y,\{x,y\}}$, *but* $P_{x,\{x,y,z\}} < P_{y,\{x,y,z\}}$	None
Similarity Aversion	$P_{x,\{x,y\}} > P_{y,\{x,y\}}$, *but* $P_{x,\{x,y,z\}} < P_{y,\{x,y,z\}}$	$z \approx x$, splits share
Compromise Effect	$P_{x,\{x,y\}} > P_{y,\{x,y\}}$, *but* $P_{x,\{x,y,z\}} < P_{y,\{x,y,z\}}$	$x > y, x > z, y > z$
Asymmetric Dominance	$P_{x,\{x,y\}} > P_{y,\{x,y\}}$, *but* $P_{x,\{x,y,z\}} < P_{y,\{x,y,z\}}$	$x \approx y, y \geq z$

Table 1 provides an overview of the idea that the famous observations of IIA violations in discrete choices are simply instances of preference reversals. Since the CDM can help model such reversals, it can consequently model these effects.

C.3. Identifiability and Regularization

In this section, we further explore the concepts developed in the main text about identifiability and regularization. Intricate conditions of identifiability are not unique to the CDM, but are rather widespread in the embeddings literature. These conditions, however, are not very well described or stated anywhere, and especially matter in the embedding setting because regularization is often omitted. Here, we explore a few different models, starting first with the Blade Chest model.

C.3.1. Blade Chest

As stated before, we may treat the Blade Chest model as the CDM applied only to the pairwise comparisons. But Theorem 2 demonstrates that the CDM is not identified in this setting, hence, neither is the Blade Chest Model. We make this clear as follows. Consider first the full rank case, $d = n$. If $\hat{U}$ is a solution to the problem, then $\tilde{U} = \hat{U} + A$ for any symmetric matrix A. Using this, we can consider $d < n$. A subset of solutions when $d < n$ is $\hat{T} + X, \hat{C} + Y$, where $X = \beta\hat{C} + \gamma_1\alpha\beta\hat{T}$, and $Y = \alpha\hat{T} + \gamma_2\alpha\beta\hat{C}$ where $\alpha, \beta \in \mathbb{R}$, $\gamma_1, \gamma_2 \in \{0, 1\}, \gamma_1 \neq \gamma_2$.

We note that this, however, is only an illustrative small subset to a more general set of solutions that could be better explored through heuristic approaches to the computationally hard affine rank minimization problem.

C.3.2. Shopper

Yet another model that suffers from identifiability issues is the Shopper model (Ruiz et al., 2017). We refer the reader to the orignal work for a review on the model in order to keep the dicussion here terse. Consider first the full rank case, $d = n$. If $\hat{U}$ is a solution to the problem, then $\tilde{U} = \hat{U} + \mathbf{1}z^T + \text{diag}(a)$ for any vectors $z, a \in \mathbb{R}^n$. A subset of solutions when $d < n$ is $\hat{T} + x\mathbf{1}^T$, or the origin of the target vector. Though mere shifts of the origin might seem trivial in visualizing the underlying embeddings, these shifts become significant under a measure like cosine distance, or the embeddings use in any absolute, as opposed to relative setting.

C.3.3. CONTINUOUS BAG OF WORDS (CBOW)

Here, we describe the original CBOW, not the version with negative sampling that is an entirely different objective (Rudolph et al., 2016). Consider first the full rank case, $d = n$. If $\hat{U}$ is a solution to the problem, then $\tilde{U} = \hat{U} + \mathbf{1}z^T$ for any vector $z \in \mathbb{R}^n$. A subset of solutions when $d < n$ is $\hat{T} + x\mathbf{1}^T$, or the origin of the target vector. Yet again, when the underlying measure of comparing word similarity is cosine distance—which it frequently is in natural language processing—an origin discrepancy make a difference in underlying task performance.

C.3.4. REGULARIZATION

A clean solution to issues of uniqueness is to add regularization. Specifically, any amount of ℓ_2 regularization immediately guarantees identifiability, whereas the same cannot be said of ℓ_1 regularization. We consider the impact of regularization on the CDM in two specific instances.

ℓ_1 regularization on exponentiated variables. Because the CDM is shift invariant, we may set the shift such that the sum of the exponentiated sum of all the rows may be set to 1. That is, $\sum_{y\in\mathcal{X}} \exp(\sum_{x\in\mathcal{X}\setminus y} u_{xy}) = 1$. With such a shift, applying ℓ_1 regularization to the exponentiated entries may be reformulated as adding a uniform prior of choices from the Universe. Such an idea is described in (Ragain et al., 2018) for the MNL model. This regularization is a valuable addition when the set of observations is small or the comparison graph is irregular. In these settings, the regularization plays a balancing role that is also interpretable for any dataset: additional choices from the universe. However, we know that such an addition alone will not uniquely identify the CDM - especially if the dataset only contains pairwise comparisons, where the CDM will not be identified even with an arbitrarily large sample size. Even with datasets of a choice set size greater than 2, the dataset still requires samples from a diverse range of choice sets within that size before it is identifiable with the regularization. This is consistent with the view that ℓ_1 does not always identify the CDM.

ℓ_2 regularization on the U matrix. As stated earlier, any small amount of ℓ_2 regularization immediately identifies the CDM. Since the "pairwise comparisons only" setting suffers in a rather extreme way from identifiability issues, understanding the role ℓ_2 regularization plays there is important. We recall from earlier than in the setting of pairwise comparisons, the CDM matrix U is only specified up to a symmetric matrix A when inferred from pairwise comparisons. Since ℓ_2 regularization will minimize the entrywise norm of the U matrix, A will be chosen to be zero. That is, the U matrix will be antisymmetric. We may then use this property to solve for parameter u_{xy} as a function of the pairwise probabilities:

$$u_{xy} = \frac{1}{2}\log\left(\frac{P_{x,\{x,y\}}}{P_{y,\{x,y\}}}\right)$$

It is most interesting to look at

$$u_{xz} - u_{yz} = \frac{1}{2}\log\left(\frac{P_{z,\{y,z\}}P_{x,\{x,z\}}}{P_{y,\{y,z\}}P_{z,\{x,z\}}}\right).$$

Since $u_{xz} - u_{yz}$ corresponds to the influence a third item z's presence has on the choice between x and y, it is interesting that the relative intransitivities of the three items in their respective pairwise settings are leveraged to describe this influence in the triplet case. This is quite possibly the best outcome one could hope for having just pairwise comparisons, and demonstrates the value of regularization.

C.4. Auxiliary Lemmas

Lemma 3. *For $\Sigma_{\mathcal{D}} := \frac{1}{m^2}X(\mathcal{D})L^\dagger X(\mathcal{D})^T$, where the remaining quantities are defined in the proof of Theorem 3, we have,*

$$\textbf{\textit{tr}}(\Sigma_{\mathcal{D}}) = \frac{d-1}{m} \qquad \textbf{\textit{tr}}(\Sigma_{\mathcal{D}}^2) = \frac{(d-1)^2}{m^2} \qquad ||\Sigma_{\mathcal{D}}||_{op} = \frac{1}{m}.$$

Proof. Consider first that $L = \frac{1}{m}X(\mathcal{D})^T X(\mathcal{D})$. Since L is symmetric and positive semidefinite, it has an eigenvalue decomposition of $U\Lambda U^T$. By definition, the Moore-Penrose inverse is $L^\dagger = U\Lambda^\dagger U^T$. We must have that $X(\mathcal{D}) = \sqrt{m}V\Lambda^{\frac{1}{2}}U^T$ for some orthogonal matrix V in order for L to equal $\frac{1}{m}X(\mathcal{D})^T X(\mathcal{D})$. With these facts, we have

$$\begin{aligned}\frac{1}{m^2}X(\mathcal{D})L^\dagger X(\mathcal{D})^T &= \frac{1}{m^2}\sqrt{m}V\Lambda^{\frac{1}{2}}U^T U\Lambda^\dagger U^T U\Lambda^{\frac{1}{2}}V^T\sqrt{m}\\ &= \frac{1}{m}V\Lambda\Lambda^\dagger V^T.\end{aligned}$$

That is, $\Sigma_{\mathcal{D}}$ is a positive semi-definite matrix with spectra corresponding to $d-1$ values equaling $\frac{1}{m}$, and the last equaling 0. The three results about the traces and the operator norm immediately follow.