Learning preferences with irrelevant alternatives

Joint work with Alex Peysakhovich, Stephen Ragain, and Arjun Seshadri
Preferences over sets

- Given a universe set $\mathcal{X}$, consider a choice set $C \subseteq \mathcal{X}$. What do you choose?
- **Discrete choice**: learning distributions over items, for all sets $C \subseteq \mathcal{X}$.
- **Ranking**: distributions over permutations of $\mathcal{X}$.

(McFadden, 1974)

(Manski, 1976)
Agenda

• **Choice systems** as mathematical objects.

• The **independence of irrelevant alternatives** (IIA) in discrete choice.

• Tractable **choice models** that forego IIA. (ICML 2019)

• Tractable **rankings models** that forego IIA. (NeurIPS 2020)

• When does data obey IIA? Lower bounds on **hypothesis testing**. (EC 2019)
Probabilistic discrete choice

- Focuses on a peculiar mathematical space, choice systems.

\(\mathbb{R}^n\)

\(P_n\)

\(S^3\)

\(\Delta^n\)

\({0, 1}\)^{n\times n}

\(S_n\)

\(C^k\)

\(T_n\)
Probabilistic discrete choice

• Focuses on a peculiar mathematical space, choice systems.

• Let $P_{x,C}$ denote the probability of choosing $x$ from $C$.

• Definition: Conditional choice system (Falmagne, 1978):

$$\{P_{x,C} \forall C \subseteq X, \forall x \in C\}$$
Probabilistic discrete choice

- Focuses on a peculiar mathematical space, choice systems.
- Let $P_{x,C}$ denote the probability of choosing $x$ from $C$.

- **Definition**: Conditional choice system (Falmagne, 1978):
  \[
  \{P_{x,C}\} \forall C \subseteq \mathcal{X}, \forall x \in C
  \]

- Let $w(C)$ denote the probability of choosing from $C \subseteq \mathcal{X}$. Features in “unconditional choice system”, not part of this talk.
Probabilistic discrete choice

- Consider $\mathcal{X} = \{a, b, c\}$. What is $\{P_{x,C}\}_{\forall C \subseteq \mathcal{X}, \forall x \in C}$?
Consider $\mathcal{X} = \{a, b, c\}$. What is $\{P_{x,C}\}_{\forall C \subseteq \mathcal{X}, \forall x \in C}$?

$C_1 = \{a, b\}$
$C_2 = \{b, c\}$
$C_3 = \{a, c\}$
$C_4 = \{a, b, c\}$

$P_{x, C_1}$:

$P_{x, C_2}$:

$P_{x, C_3}$:

$P_{x, C_4}$:
Independence of Irrelevant Alternatives (IIA)

- Arbitrary choice systems (i.e., McFadden’s *universal logit*) make no assumptions about the relationship between distributions on different sets.

- IIA (Luce, 1959): For every $x \in \mathcal{X}$, $C \subseteq \mathcal{X}$:

$$\frac{P_{x,\{x,y\}}}{P_{y,\{x,y\}}} = \frac{P_{x,\{x,y\} \cup C}}{P_{y,\{x,y\} \cup C}}.$$

- Consequence: the ratio between $x$ and $y$ stays the same, no matter what “irrelevant alternatives” you add to the choice set.
Independence of Irrelevant Alternatives (IIA)

- Arbitrary choice systems (i.e., McFadden’s *universal logit*) make no assumptions about the relationship between distributions on different sets.

- IIA (Luce, 1959): For every \( x \in \mathcal{X}, C \subseteq \mathcal{X} \):
  \[
  \frac{P_{x,\{x,y\}}}{P_{y,\{x,y\}}} = \frac{P_{x,\{x,y\} \cup C}}{P_{y,\{x,y\} \cup C}}.
  \]

- Consequence: the ratio between \( x \) and \( y \) stays the same, no matter what “irrelevant alternatives” you add to the choice set.

- Models obeying IIA admit a *ratio representation*:
  \[
  P_{x,C} = \frac{\gamma_x}{\sum_{z \in C} \gamma_z}, \forall C \subseteq \mathcal{X}, \forall x \in C.
  \]
Independence of Irrelevant Alternatives (IIA)

- Assuming IIA ⇒ **Multinomial Logit (MNL)** model of discrete choice:
  \[ P_{x,C} = \frac{\exp(u_x)}{\sum_{z\in C} \exp(u_z)}. \]

- Major workhorse of modern machine learning

- If \( u_x = \beta^T f_x \), linear model

#IJALM
Independence of Irrelevant Alternatives (IIA)

- Examples where it (arguably) doesn’t hold:
- Web browsing (Benson-Kumar-Tomkins, 2016)
- Search ads (leong-Mishra-Sheffet 2012, Yin et al. 2014)
- Music (Debreu, 1960)
Three perspectives on IIA, beyond IIA

1. **Random utility model (RUM)** with Gumbel noise (Yellot, 1977)

2. Stationary distribution of a **Markov chain** (Maystre & Grossglauser, 2015)

3. First-order truncation of a **Taylor-like expansion** of a choice system (Batsell & Polking,1985; Seshadri et al. 2019)

Each derivation is its own path to a beyond-IIA model of choice.

(Setting aside mixture/nested models today.)
(1) Random utility models and IIA

- For each $i \in \mathcal{X}$, associate a random variable $X_i = \mu_i + \epsilon_i$.
- Let $P_{i,C} = Pr(X_i = \max_{j \in C} X_j)$.
- Iff $\epsilon_1, \ldots, \epsilon_n$ are independent zero-mean Gumbel, $P_{i,C} = \frac{\exp(\mu_i)}{\sum_{j \in C} \exp(\mu_j)}$. (MNL!)
(1) Random utility models and IIA

- For each \( i \in \mathcal{X} \), associate a random variable \( X_i = \mu_i + \epsilon_i \).
- Let \( P_{i,C} = Pr(X_i = \max_{j \in C} X_j) \).
- Iff \( \epsilon_1, \ldots, \epsilon_n \) are independent zero-mean Gumbel, \( P_{i,C} = \frac{\exp(\mu_i)}{\sum_{j \in C} \exp(\mu_j)} \). (MNL!)
- See Falmagne (1978)'s characterization theorem of RUMs.
- RUMs need not be stochastically transitive! (Makhijani & U, 2019) connects transitivity to log-likelihood concavity of item-level parameterizations.
(2) Choice systems from Markov chains

- Consider a continuous-time Markov chain defined on $\mathcal{X}$, parameterized by $Q$.

$$
\pi^T \begin{bmatrix}
- \sum_{i \neq 1} q_{1i} & q_{12} & q_{13} \\
q_{21} & - \sum_{i \neq 2} q_{2i} & q_{23} \\
q_{31} & q_{32} & - \sum_{i \neq 3} q_{3i}
\end{bmatrix} = 0
$$
(2) Choice systems from Markov chains

- Consider a continuous-time Markov chain defined on $\mathcal{X}$, parameterized by $Q$.

- Define a chain for each subset $C \subseteq \mathcal{X}$ by restricting the rate matrix, e.g.:

$$\pi^T \begin{bmatrix} -\sum_{i \neq 1} q_{1i} & q_{12} \\ q_{21} & -\sum_{i \neq 2} q_{2i} & q_{23} \\ q_{31} & q_{32} & -\sum_{i \neq 3} q_{3i} \end{bmatrix} = 0$$

- These stationary distributions define a choice system (Ragain & U, 2016)

- See also: (Maystre & Grossglauser, 2015)
(3) Truncating choice systems

- Define item-set utilities $u(x|C), \forall x \in C$, such that $\sum_{y \in C} u(y|C) = 0$.
- Arbitrary universal logit model:

$$P_{x,C} = \frac{\exp(u(x|C))}{\sum_{y \in C} \exp(u(y|C))}.$$
(3) Truncating choice systems

• Define item-set utilities $u(x|C), \forall x \in C$, such that $\sum_{y \in C} u(y|C) = 0$.

• Arbitrary universal logit model:

$$P_{x,C} = \frac{\exp(u(x|C))}{\sum_{y \in C} \exp(u(y|C))}.$$  

• Item-set utilities can be uniquely* expanded as (Batsell & Polking, 1985):

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

*with constraints, not shown.
(3) Truncating choice systems

\[ u(x|C) = v(x) + \sum_{\{y\} \subseteq C \setminus x} v(x|\{y\}) + \sum_{\{y,z\} \subseteq C \setminus x} v(x|\{y,z\}) + \ldots + \sum_{|C| \setminus \{x\}} v(x|C \setminus \{x\}) \]

- Call \( p^{th} \) order model \( M_p \). Notice that \( M_1 \subseteq M_2 \subseteq \ldots \subseteq M_{n-1} \).
(3) Truncating choice systems

\[ u(x|C) = \underbrace{v(x)}_{1st\ order} + \sum_{\{y\}\in C\setminus x} v(x|\{y\}) + \sum_{\{y,z\}\subseteq C\setminus x} v(x|\{y,z\}) + \ldots + \underbrace{v(x|C\setminus \{x\})}_{|C|th\ order} \]

- Call \( p^{th} \) order model \( \mathcal{M}_p \). Notice that \( \mathcal{M}_1 \subseteq \mathcal{M}_2 \subseteq \ldots \subseteq \mathcal{M}_{n-1} \).

MNL/IIA \quad \text{Universal logit}
Context-dependent utility model

- For $\mathcal{M}_2$, after manipulations, choice probabilities can be written as:

$$P_{x,C} = \frac{\exp(\sum_{z \in C \setminus x} u_{xz})}{\sum_{z \in C} \exp(\sum_{z \in C \setminus y} u_{yz})}.$$

- Assumes “Pairwise Linear Dependence of Alternatives”

- Negative log likelihood is **convex** in parameters $U$!
Context-dependent utility model

- For $M_2$, after manipulations, choice probabilities can be written as:

$$P_{x,C} = \frac{\exp(\sum_{z \in C \backslash x} u_{xz})}{\sum_{z \in C} \exp(\sum_{z \in C \backslash y} u_{yz})}.$$

- Assumes “Pairwise Linear Dependence of Alternatives”

- Negative log likelihood is **convex** in parameters $U$!

- Can be made **low-rank** (non-convex), essentially a matrix factorization loss:

$$P_{x,C} = \frac{\exp(\sum_{z \in C \backslash x} c_z^T t_x)}{\sum_{z \in C} \exp(\sum_{z \in C \backslash y} c_z^T t_y)}.$$
Structure-dependent convergence rate

- **Identifiability conditions** in choice models are combinatorial (Ford 1957).

- Batsell & Polking used least squares (cleverly!), not MLE.

- Under mild regularity conditions, we show

\[
\mathbb{E} [||\hat{u}_{\text{MLE}}(\mathcal{D}) - u^*||_2^2] \leq \frac{c}{\lambda_2(L(\mathcal{D}))} \frac{n(n-1)}{m}.
\]

where \( n \) is the number of items, \( m \) the size of the data, and \( \mathcal{D} \) a random dataset generated under the model.

- Here \( \lambda_2(L(\mathcal{D})) \) is the second smallest eigenvalue of a Laplacian-like matrix. For pairwise comparisons: Laplacian of comparison graph (Shah et al. 2016).
Broader implications

- Convergence result is for full-rank case; bound still applies when low-rank.

- Analysis also applies to **Blade-Chest** model (Chen & Joachims, 2016a,b) and many **word2vec**-type models (Mikolov et al., 2013).
  
  - For word2vec, the likelihood objective is typically approximated by “negative sampling” the choice set, also changes the objective.

- Recent related work:
  
  - Extension to “salient” features (Bower & Balzano, 2020).
  
  - Promoting a particular choice (Tomlinson & Benson, 2020).
CDM empirical results

- Predicting transportation choices (Koppelman & Bhat, 2006) with the CDM:
Low-rank factorization of U: embeddings

- “One of these things is not like the other…” triplets (Heikinheimo & Ukkonen, 2013)

CDM Target Vectors, $r = 2$

CDM Context Vectors, $r = 2$
Ranking as choice

- Plackett-Luce: distributions over $S_n$ as “repeated MNL choice”:

\[
\Pr[\pi = 123 \ldots n] = \prod_{i=1}^{n} \frac{\exp(u_i)}{\sum_{j=i}^{n} \exp(u_j)}
\]

- See also: Mallows, mixtures of Mallows/PL.
- What happens if we replace MNL with CDM?
We introduce in this work a novel approach that achieves significantly higher out-of-sample likelihood, compared to the existing methods. The approach is based on the transformations of rankings into choices, which allows for the development of a flexible repeated selection model (CRS) that connects rankings to choices. This approach not only outperforms existing models in terms of likelihood but also provides expected risk and tail bounds that match known lower bounds. The tail bounds stem from a careful spectral analysis of the Plackett-Luce comparison Laplacian, which is a key result in its own right.

Our empirical evaluations focus on predicting out-of-sample rankings as well as predicting performance to positions in a ranking. We find that our new model performs similarly to the Plackett-Luce (PL) model on elections, sushi preferences, Nascar race results, and search engine results. By decomposing a ranking into a series of sequential entries as the top entries are revealed, we observe that the flexible CRS model outperforms the PL model on predicting subsequent top entries. Our approach of connecting rankings to choices is not new; repeated selection was first used to connect the MNL model of choice to the PL model of rankings. However, our tight analysis of the PL tail bounds and the development of the CRS model are novel contributions.

Our contextual repeated selection (CRS) model arises from applying the recently introduced context-dependent utility model (CDM) to choices arising from repeated selection. The CDM model is straightforward to apply to choices, and we can then employ tractable choice models to create choice-based models on choice data implied by the ranking data, making model inference tractable whenever the underlying choice model inference is tractable.

There is an extensive body of work on modeling and learning ranking distributions. The Plackett-Luce and Mallows maximum likelihood estimates, as well as the maximum likelihood estimate from the full-rank CRS model class, are special cases of the Plackett-Luce MLE and Mallows MLE, respectively. Mixtures of Plackett-Luce, Mallows, and full-rank CRS models can be described as arising from such a top-down choice process.

Ranking distributions

- Contextual repeated selection (CRS) can represent rich, multi-modal distributions with the same learning efficiency/guarantees as CDM choice.
Ranking MLE from data

• Similar to choice result, expected risk bound, with \( \ell \) rankings of length \( n \):

\[
\mathbb{E}
\left[
\left\| \hat{u}_{\text{MLE}}(R) - u^* \right\|_2^2
\right]
\leq
\mathbb{E}
\left[
\min\left\{\frac{c_B' n^3}{\ell \lambda_2(L)}, 4B^2 n\right\}
\right]
\leq c_B \frac{n^7}{\ell}
\]

• Notice second eigenvalue can be bounded absolutely.
Ranking MLE from data

- Similar to choice result, expected risk bound, with $\ell$ rankings of length $n$:

$$\mathbb{E}\left[\|\hat{u}_{MLE}(\mathcal{R}) - u^*\|_2^2\right] \leq \mathbb{E}\left[\min\left\{\frac{c' B n^3}{\ell \lambda_2(L)}, 4B^2 n\right\}\right] \leq c_B \frac{n^7}{\ell}$$

- Notice second eigenvalue can be bounded absolutely.

- Paper also has **tail bounds** (not just expected risk).

- Paper also sharpens convergence analysis of vanilla **MNL, Plackett-Luce (!)**
Testing IIA
Why is testing IIA hard?

- **Anna Karenina Principle** of high-dimensional hypothesis testing: “all nulls are alike; deviations from the null all deviate in their own way.”

- **Applied to IIA:** there are only a few ways to be “rational,” there are a many unique ways that people can be “irrational.”

- Follows the burst of work on finite-sample lower bounds on testing:

  (Paninski 2008; Wei & Wainwright 2016; Valiant & Valiant 2017; Daskalakis, Kamath, Wright 2018; Balakrishnan & Wasserman 2018).
Separation and “orthogonal” perturbations

- Begin with the basic formula for lower bounds on minimax risk (and testing):
  - Define separation (TV distance).
  - Simplify to testing uniform choice system $p_0$ vs. composite of other distributions perturbed out of the space of IIA.

![Diagram](image)
Structure-dependent lower bounds

• In a strict sense, if data doesn’t contain choices from every subset, the full implications of IIA can’t be tested.

• Instead: let \( \mathcal{C} \) be the set of subsets being compared.

• **Example:** \( \mathcal{X} = \{1, 2, 3, 4\} \)

\[
\mathcal{C} = \{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \{1, 2, 3, 4\}\}
\]

\[
\begin{align*}
C_1 & = \{1, 2\} \\
C_2 & = \{1, 3\} \\
C_3 & = \{1, 4\} \\
C_4 & = \{2, 3\} \\
C_5 & = \{2, 4\} \\
C_6 & = \{3, 4\} \\
C_7 & = \{1, 2, 3, 4\}
\end{align*}
\]
Structure-dependent lower bounds

• In a strict sense, if data doesn’t contain choices from every subset, the full implications of IIA can’t be tested.

• Instead: let $\mathcal{C}$ be the set of subsets being compared.

• Example: $\mathcal{X} = \{1, 2, 3, 4\}$
  
  $\mathcal{C} = \{\{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \{1, 2, 3, 4\}\}$

• Consider: bipartite comparison incidence graph $G_{\mathcal{C}} = (\mathcal{X}, \mathcal{C}, E)$:
Constructing perturbations

• Starting at uniform, want perturbations out of IIA space that all still **project back** onto uniform.

• Want as many perturbations as possible.
Constructing perturbations

• Starting at uniform, want perturbations out of IIA space that all still **project back** onto uniform.

• Want as many perturbations as possible.

• Sketch of construction:
  
  • Need **sets** to maintain their frequency,
    **items** to maintain their choice frequency.
  
  • Seek perturbations of parameters that keep overall item probabilities fixed, set probabilities fixed.
  
  • Seek a **cycle decomposition** of $G_C = (\mathcal{X}, C, E)$ into many cycles!
Structure-dependent lower bounds

- Let $\mu(\sigma)$ and $\alpha(\sigma)$ be properties of some cycle decomposition $\sigma$ of $G_C = (\mathcal{X}, C, E)$. Then for $N$ choices:

<table>
<thead>
<tr>
<th>Structure of $C$</th>
<th>$R_{N, \delta}(\mathcal{P}_{C}^{IIA})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>General</td>
<td>$\geq \frac{1}{2} - \frac{1}{4} \left( \exp \left( \frac{8\mu(\sigma)^4 \alpha(\sigma)N^2 \delta^4}{d} \right) - 1 \right)^{\frac{1}{2}}$</td>
</tr>
<tr>
<td>All subsets, $d = n2^{n-1}$</td>
<td>$\geq \frac{1}{2} - \frac{1}{4} \left( \exp \left( \frac{c \log(n)N^2 \delta^4}{n2^{n-1}} \right) - 1 \right)^{\frac{1}{2}}$</td>
</tr>
<tr>
<td>All pairs, $d = n(n-1)$</td>
<td>$\geq \frac{1}{2} - \frac{1}{4} \left( \exp \left( \frac{cN^2 \delta^4}{n(n-1)} \right) - 1 \right)^{\frac{1}{2}}$</td>
</tr>
</tbody>
</table>

- $R_{N, \delta} \geq 0$ means lower bound has fallen away.

- No upper bounds, no tests analyzed.
Thank you!

- **Choice systems** are beautiful things.
- Doors have recently opened to introduce and analyze tractable models beyond IIA based on **Markov chains**, based on **truncations**.
- **Testing IIA**: we replace ambiguity with rigorous pessimism.
- **Papers**:  
  PCMC: Ragain & Ugander, NeurIPS 2016  
  CDM: Seshadri, Peysakhovich, Ugander, ICML 2019  
  Testing: Seshadri & Ugander, EC 2019  
  Choice models of networks: Overgoor et al. WWW 2019, KDD 2020  
  Ranking: Seshadri, Ragain, Ugander, NeurIPS 2020