When the Universe is Too Big: Bounding Consideration Probabilities for Plackett–Luce Rankings

Ben Aoki-Sherwood
University of Colorado Boulder

Catherine Bregou
Carleton College

David Liben-Nowell
Carleton College

Kiran Tomlinson
Microsoft Research

Thomas Zeng
University of Wisconsin–Madison

Abstract

The widely used Plackett–Luce ranking model assumes that individuals rank items by making repeated choices from a universe of items. But in many cases the universe is too big for people to plausibly consider all options. In the choice literature, this issue has been addressed by supposing that individuals first sample a small consideration set and then choose among the considered items. However, inferring unobserved consideration sets (or item consideration probabilities) in this "consider then choose" setting poses significant challenges, because even simple models of consideration with strong independence assumptions are not identifiable, even if item utilities are known. We apply the consider-then-choose framework to top-k rankings, where we assume rankings are constructed according to a Plackett–Luce model after sampling a consideration set. While item consideration probabilities remain non-identified in this setting, we prove that we can infer bounds on the relative values of consideration probabilities. Additionally, given a condition on the expected consideration set size and known item utilities, we derive absolute upper and lower bounds on item consideration probabilities. We also provide algorithms to tighten those bounds on consideration probabilities by propagating inferred constraints. Thus, we show that we can learn useful information about consideration probabilities despite not being able to identify them precisely. We demonstrate our methods on a ranking dataset from a psychology experiment with two different ranking tasks (one with fixed consideration sets and one with unknown consideration sets). This combination of data allows us to estimate utilities and then learn about unknown consideration probabilities using our bounds.

1 INTRODUCTION

Among the wide-ranging topics studied in the behavioral sciences, predicting and explaining human choices is a central challenge across a range of disciplines. (Why *that* entrée at *that* café with *that* person, of all the restaurants you can think of, of all your potential dates, of all the dishes on the menu?) Settings where individuals select an item from a collection of available alternatives are well studied in the literature on *discrete choice* (Train, 2009), with applications ranging from marketing (Chintagunta and Nair, 2011) and voting (Thurner, 2000) to transportation policy (Horne et al., 2005) and recommender systems (Danaf et al., 2019). A closely related line of work studies *rankings* (Alvo and Yu, 2014), rather than single choices, often modeling a ranking as a sequence of discrete choices (selecting the top-ranked item, then the second, etc.). The Plackett–Luce ranking model (Luce, 1959; Plackett, 1975) exemplifies the link between discrete choice and ranking, positing that items at each of k positions are selected in turn according to a logit choice model (McFadden, 1973). Due to its convex negative log-likelihood and easily interpretable parameters, the Plackett–Luce model has seen substantial use in the machine learning literature (Nguyen and Zhang, 2023; Saha and Gopalan, 2020; Seshadri et al., 2020; Zhao et al., 2016; Zhao and Xia, 2019).

Proceedings of the 28$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

However, as Plackett–Luce models are applied to datasets of increasing size, it becomes implausible that individuals are able to weigh all of their options against each other. This issue has been largely ignored in the ranking setting, but has been discussed at length in the discrete choice literature. For instance, in keeping with the framework of bounded rationality (Simon, 1957), a prominent line of work in discrete choice suggests that selection is a two-stage process, where individuals first narrow their options to a small *consideration set* from which their final selection is made (Hauser and Wernerfelt, 1990; Shocker et al., 1991). These so-called "consider then choose" models have shown considerable promise in their explanatory power (Başar and Bhat, 2004; Roberts and Lattin, 1991) and help address the issue of large item universes. But they suffer from a major issue: consideration sets are almost always unobserved (except in carefully controlled experimental settings) and cannot in general be identified from observed choice data (Jagabathula et al., 2023; van Nierop et al., 2010). However, if it were possible to derive meaningful information about consideration probabilities from rankings, this would be very valuable: for instance, in recommender systems, we can think of the core goal as identifying items with high utility but low consideration probability. These are the types of items that users would not discover naturally but still enjoy.

The present work. In this paper, we define the natural consider-then-rank model obtained by augmenting top-k Plackett–Luce with the independent-consideration rule (Manzini and Mariotti, 2014), in which each item advances to the ranking stage randomly and independently with an unknown item-specific *consideration probability.* We term this model *Plackett–Luce with consideration* (PL+C). One might hope that richer observations—a ranking of k items rather than just a single choice—would make it feasible to identify consideration probabilities, at least for large k. Unfortunately, our first result is negative: regardless of k, there are infinite families of consideration probabilities that generate the same distribution of observed data. However, we show that it is possible to derive meaningful bounds on consideration probabilities, despite their non-identifiability. In addition to observations of rankings, some (but not all) of our bounds draw on two types of information for learning about consideration probabilities: (1) known item utilities and (2) a lower bound on expected consideration set size. (Later, we discuss settings where such information is indeed available.) First, we derive relative bounds on the consideration probabilities of different items, of the form "if item i has consideration probability p_i, then j's consideration probability is greater than $f(p_i)$." The intuition is that if i has higher utility then j, then we would expect to see i ranked highly more often than j—but if instead j outperforms i, then consideration must be the culprit, and j's consideration probability must be higher than i's. Quantitatively, the j-vs.-i gap in the frequency of being ranked highly tells us how much more often j is considered than i. To make this bound more useful, we show how to infer that i has higher utility than j despite confounding from consideration, which invalidates standard Plackett–Luce inference.

These relative bounds provide useful information, but, absent additional anchor points, they still admit a large range of possible consideration probabilities. Motivated by this limitation, we then seek absolute upper and lower bounds on consideration probabilities. For lower bounds, naïvely, we might hope that the fraction of observed rankings in which item i appears would be a lower bound on i's consideration probability. However, this idea fails in our setting, as we assume that we only observe rankings of exactly k items, and that any instance where fewer than k items were considered is discarded before we observe it (or, equivalently, a consideration set is resampled). However, we are able to rescue this approach if we can assume a mild bound on expected consideration set size, since we can then upper bound the probability that a sample was discarded. Next, we turn to upper bounds on items' consideration probabilities. Our absolute upper bounds use exogenous knowledge of item utilities, which we learn in our data from survey questions with explicit consideration sets. Thus, if an item i is ranked first less often than we would expect given its utility, then it must not be considered very often relative to the other items. Using a pessimistic upper bound of 1 for other items' consideration probabilities can then yield a nontrivial upper bound on i's consideration probability. To wrap up our theoretical contributions, we provide algorithms to combine our absolute and relative bounds by propagating over a directed acyclic graph induced by our relative bounds.

Finally, we demonstrate how our methods can reveal meaningful information about consideration probabilities on real data. In a psychology experiment about perceptions of U.S. history (Putnam et al., 2018), participants completed several tasks related to their view of the historical importance of particular U.S. states, such as naming the three states they believe contributed most to U.S. history. They were also provided with a random set of 10 states and asked to rate their percentage contribution to U.S. history. (We convert the numerical scores to rankings.) These two settings, *Top-3* and *Random-10*, provide us both observations from fixed consideration sets (Random-10) and from unknown consideration sets (Top-3). We can thus estimate utilities in the absence of consideration from the Random-10 data and then estimate consideration from the Top-3 data, using our bounds. Our results on this data align with expectations about state history: original colonies Massachusetts and Virginia were the states with the highest range of possible consideration probabilities (0.59 to 1.00), while Missouri, an anecdotally often-forgotten state (Cutolo, 2023), had the lowest upper bound on consideration probability (0.11).

2 RELATED WORK

There are several approaches to adding consideration to choice models. The one closest to our work adds a consideration stage to random utility models (Başar and Bhat, 2004; Ben-Akiva and Boccara, 1995; Roberts and Lattin, 1991; van Nierop et al., 2010), following the formulation of Manski (1977). Another line of work uses a different choice model, assuming choosers have a strict preference ordering over items and deterministically pick the highest ranked item they consider (Cattaneo et al., 2020; Manzini and Mariotti, 2014). An extension of this model allows the preference orderings to be stochastic (Jagabathula et al., 2023). There are also a variety of approaches for modeling the consideration stage, including feature-based constraints (Ben-Akiva and Boccara, 1995), rational utility-maximization (Roberts and Lattin, 1991), and the independent consideration model we use (Manzini and Mariotti, 2014).

The issue of non-identifiability likewise has a number of proposed solutions. In the most straightforward approach, experimenters directly ask item availability questions to choosers (e.g., "is item i available?" or "did you consider item i?") (Ben-Akiva and Boccara, 1995; Roberts and Lattin, 1991; Suh, 2009), which is especially useful when trying to estimate constraint-based consideration models. A similar approach in online shopping settings uses data about which items customers view to directly infer consideration sets (Gu et al., 2012; Moe, 2006). Another strategy leans on knowing when a default "outside option" (i.e., making no selection) was selected and assuming that it is only chosen when no items are considered (Jagabathula et al., 2023; Manzini and Mariotti, 2014). Yet another approach supposes we have observations of choices over time as item features vary and uses the idea that demand for highly considered items will respond more to changes in features (Abaluck and Adams-Prassl, 2021). Finally, several papers make assumptions about the parametric form of consideration probabilities as a function of chooser and item features and estimate those parameters from choice data (Başar and Bhat, 2004; van Nierop et al., 2010), although this approach makes inference computationally challenging.

While consideration sets have been extensively explored in single-choice setting, they have received limited attention in the ranking literature; to our knowledge, our work is the first to recover information about consideration probabilities under Plackett–Luce. Palma (2017) discusses the idea of consideration sets in relation to rankings, but only uses them to identify a relevant prefix of rankings in survey data (items ranked above the outside option are treated as the implied consideration set). Fok et al. (2012) mention the idea of applying consideration sets to rankings, but do not pursue this direction as they focus on survey settings where respondents are forced to rank all available items. We note that the Plackett–Luce model is sometimes referred to as the "rank-ordered logit" in the econometrics literature (Beggs et al., 1981; Hausman and Ruud, 1987), while "Plackett–Luce" is more common in computer science (for length-2 rankings, a.k.a. pairwise comparisons, the model is also called "Bradley–Terry" (Bradley and Terry, 1952)). Much of the existing computational work on the Plackett–Luce model has focused on utility inference (Guiver and Snelson, 2009; Liu et al., 2019; Maystre and Grossglauser, 2015; Zhao et al., 2016).

3 PRELIMINARIES

In a ranking setting, we have a universe of items $\mathcal{U} = \{1, \dots, n\}$ and observe a collection of length-k rankings of the form $r = \langle r_1, \dots, r_k \rangle$, where $k \leq n$ is fixed and each $r_i \in \mathcal{U}$ is distinct.

The Plackett–Luce model (Luce, 1959; Plackett, 1975) posits that each item i has a *utility* $u_i \in \mathbb{R}$ and that rankings are formed by a sequence of choices with the probability of selecting $i \in \mathcal{U}$ proportional to $\exp(u_i)$, first choosing a top-ranked item, then a second-ranked item (distinct from the first), etc. We assume these k choices are made from only a subset of items, a *consideration set* $C \subseteq \mathcal{U}$ where $|C| \geq k$. The probability of observing ranking r under a Plackett–Luce (PL) model with consideration set C is

$$\Pr_{\mathrm{PL}}(r \mid C) = \prod_{i=1}^{k} \frac{\exp(u_{r_i})}{\sum_{j \in C \setminus \{r_1, \dots, r_{i-1}\}} \exp(u_j)} \qquad (1)$$

if $\{r_1, \dots, r_k\} \subseteq C$; otherwise, $\Pr_{\mathrm{PL}}(r \mid C) = 0$.

As our model of consideration, we assume each item i has a positive *consideration probability* $p_i \in (0, 1]$ and that items are considered independently, conditioned on the constraint $|C| \geq k$. (We do not observe any cases where fewer than k items are considered—either such instances are thrown out before we observe them, or the chooser considers a new set.) Thus, the probability of considering a set C with $|C| \geq k$ is

$$\Pr_{\mathrm{C}}(C) = \frac{1}{z_{k,p}} \Big(\prod_{i \in C} p_i \Big) \Big(\prod_{j \in \mathcal{U} \setminus C} (1 - p_j) \Big), \qquad (2)$$

where $z_{k,p} = \sum_{C \subseteq \mathcal{U}, |C| \geq k} \left(\prod_{i \in C} p_i \right) \prod_{j \in \mathcal{U} \setminus C} (1 - p_j)$ normalizes the probabilities given the condition $|C| \geq k$. For any set C with $|C| < k$, we define $\Pr_{\mathrm{C}}(C) = 0$.

Combining Equations (1) and (2) yields our full model of rankings, *Plackett–Luce with consideration* (PL+C), in which we sum over all possible consideration sets, weighted by their probabilities, and apply Plackett–Luce:

$$\Pr_{\mathrm{PL+C}}(r) = \sum_{C \subseteq \mathcal{U}} \Pr_{\mathrm{C}}(C) \cdot \Pr_{\mathrm{PL}}(r \mid C). \qquad (3)$$

Finally, we introduce some additional notation. Let $\mathcal{R}$ be the set of length-k rankings over $\mathcal{U}$ and let $\mathcal{R}_{i \leq \ell} \subset \mathcal{R}$ be the set of rankings that contain i in any of the top ℓ positions. For a set of rankings R, let $\Pr_{\mathrm{PL+C}}(R) = \sum_{r \in R} \Pr_{\mathrm{PL+C}}(r)$, and likewise for $\Pr_{\mathrm{PL}}(R \mid C)$.

4 (RELATIVE) CONSIDERATION PROBABILITY INFERENCE

4.1 Inferring exact values is impossible

We begin with an impossibility result: even with complete information about $\Pr_{\mathrm{PL+C}}(r)$ for every ranking r—and even with complete knowledge of the utilities $u_1, \ldots, u_n$ for all items—we cannot infer the consideration probabilities $p_1, \ldots, p_n$ due to non-uniqueness. See Appendix A for all omitted proofs.

Theorem 4.1. *For all $n \geq 1$ and $1 \leq k \leq n$, consideration probabilities are not identifiable in the PL+C model. That is, there are multiple sets of consideration probabilities that generate the same distribution over rankings, with fixed utilities u_i for each $i \in \mathcal{U}$.*

4.2 Inferring relationships among consideration probabilities is possible

Theorem 4.1 says that we cannot hope to compute exact consideration probabilities. Nevertheless, we will show that, in certain circumstances, we can conclude something about the *relative* values of two items' consideration probabilities. To do this, we will first need the following lemma about the probability of observing items in the top ℓ positions under Plackett–Luce.

Lemma 4.2. *For two items $i, j \in \mathcal{U}$ with $u_i > u_j$, let $C \subseteq \mathcal{U}$ with $i \in C$ and $j \notin C$ and let $C_{i \to j} = C \setminus \{i\} \cup \{j\}$ be C with i replaced by j. Let $\ell \leq k \leq |C|$. Under Plackett–Luce, i is more likely to be in the top ℓ positions of a length-k ranking with consideration set C than j is with consideration set $C_{i \to j}$. That is,*

$$\Pr_{\mathrm{PL}}(\mathcal{R}_{i \leq \ell} \mid C) > \Pr_{\mathrm{PL}}(\mathcal{R}_{j \leq \ell} \mid C_{i \to j}).$$

Additionally, if both i and j are considered, it is easy to show that i (the higher utility item) is more likely to be highly ranked than j. Thus, if both are equally likely to be considered, we would expect to see i appearing more often in high rank positions than j. If, to the contrary, we observe that j is chosen more frequently than i, then it must be the case that j is considered more often. That is: if we know the utilities of items, then we can use flips in top-ℓ ranking rates to identify which items are considered more frequently than others. We formalize this intuition in the following theorem.

Theorem 4.3. *Consider two items i, j with $u_i > u_j$ in a PL+C model. If, for some ℓ, $\Pr_{\mathrm{PL+C}}(\mathcal{R}_{i \leq \ell}) \leq \Pr_{\mathrm{PL+C}}(\mathcal{R}_{j \leq \ell})$, then $p_i \leq p_j$.*

We defer the proof of Theorem 4.3, as this result is effectively subsumed by a more powerful theorem that also gives us information about the relative sizes of p_i and p_j:

Theorem 4.4. *Consider two items i, j with $u_i > u_j$ in a PL+C model. If $c = \Pr_{\mathrm{PL+C}}(\mathcal{R}_{i \leq \ell}) / \Pr_{\mathrm{PL+C}}(\mathcal{R}_{j \leq \ell}) \leq 1$ for some ℓ, then*

$$\tfrac{p_i}{1 - p_i} \leq c \cdot \tfrac{p_j}{1 - p_j}. \qquad (4)$$

The proof isolates rankings that contain i and not j (and vice versa) and uses a bijection between those rankings and their associated consideration sets in conjunction with Lemma 4.2 to establish the claim. Intuitively, the terms $p_i(1 - p_j)$ and $p_j(1 - p_i)$ come from considering i and not j or j and not i, which we then cross-divide to group like variables. Taking $c = 1$, Theorem 4.4 implies Theorem 4.3, since $x/(1 - x)$ is a monotonically increasing function of x for $x \in [0, 1)$. Additionally, the smaller c is (i.e., the more we see j ranked higher than i, despite its lower utility), the larger the gap between p_i and p_j must be. We can rearrange (4) to be either an upper bound on p_i in terms of p_j or a lower bound on p_j in terms of p_i:

$$p_i \leq \tfrac{c p_j}{1 - p_j + c p_j} \qquad (5)$$

$$p_j \geq \tfrac{p_i}{c - c p_i + p_i}. \qquad (6)$$

Our relative bounds appear to require exogenous knowledge of utilities, but we only need to know that $u_i > u_j$, not exact values. This can be inferred from rankings: if we only consider rankings in which both i and j appear (so we know both were considered), then whichever is ranked higher more often has higher utility, given enough observations. Formally:

Lemma 4.5. *Let R be a random top-k ranking generated by Plackett–Luce with consideration. For items $i, j \in \mathcal{U}$, let $i \succ_R j$ denote that i is ranked higher in R than j. If $\Pr(i \succ_R j \mid i, j \in R) > 1/2$, then $u_i > u_j$.*

We note that this approach to identifying relative utilities is prone to sampling noise for rare items.

5 LOWER BOUNDS

Suppose that we observe a set of rankings generated by a PL+C model, where the underlying consideration probabilities $p_1, \ldots, p_n$ are unknown. Even if we have *a priori* knowledge of the underlying utilities $u_1, \ldots, u_n$, Theorem 4.1 says that we cannot compute the values of $p_1, \ldots, p_n$. But we might hope that Theorem 4.3 (or the more powerful Theorem 4.4) would allow us to leverage the empirical data into knowledge about their relative consideration probabilities: that theorem might allow us to propagate a lower bound on p_i into a lower bound on p_j. But for this kind of implication to be meaningful, we need a nontrivial starting point—that is, some way to infer that p_i is bounded away from zero, say $p_i > \varepsilon$.

How might we get started? While it is tempting to think that item i's rate of occurrence in the observed length-k rankings would lower bound p_i, the fact that we condition our samples on $|C| \geq k$ means that this relationship may not hold. For example, if each of $n = 5$ items has an identical consideration probability $p > 0$ and identical utility, then each of those five items occurs as the top choice with probability 0.2, by symmetry—but that is true for *any* value of $p > 0$, whether $p \geq 0.2$ or $p = 0.01$. So $p_i \geq \Pr_{\text{PL+C}}(\mathcal{R}_{i \leq k})$ may not hold. However, if we make some relatively mild assumptions about the *expected* consideration set size, then we can use item occurrence rates and a Chernoff bound to get a lower bound on their consideration probability. From a practical perspective, we note that measuring mean consideration set size is a common task in consumer research (Hauser and Wernerfelt, 1990).

Lemma 5.1. *Suppose $\sum_{i \in \mathcal{U}} p_i \geq \alpha k$ for some $\alpha > 1$, and let X be a random variable denoting the size of a random consideration set (prior to conditioning on there being at least k considered items). Then*

$$\Pr(X \leq k) \leq \left(\alpha e^{1-\alpha}\right)^k .$$

Given this bound on the cases where $|C| < k$, we can correct our earlier hopeful argument.

Theorem 5.2. *Suppose $\sum_{i \in \mathcal{U}} p_i \geq \alpha k$ for some $\alpha > 1$. For any $i \in \mathcal{U}$, the consideration probability for i is lower bounded as*

$$p_i \geq \Pr_{\text{PL+C}}(\mathcal{R}_{i \leq k}) \cdot \left[1 - \left(\alpha e^{1-\alpha}\right)^k\right].$$

Algorithm 1 Consideration lower bounds.

Require: items $\mathcal{U}$, (estimates of) utilities u_i, (estimates of) top-ℓ ranking probabilities $\Pr_{\text{PL+C}}(\mathcal{R}_{i \leq \ell})$, lower bound αk on expected consideration set size ($\alpha > 1$)

```
1:  b_i ← Pr_PL+C(R_{i≤k}) · [1 − (α e^{1−α})^k] for each i ∈ U
2:  E ← {⟨i, j⟩ ∈ U × U | u_i > u_j
          and ∃ℓ : Pr_PL+C(R_{i≤ℓ}) < Pr_PL+C(R_{j≤ℓ})}
3:  G ← ⟨U, E⟩
4:  for all i in a topological sort of G do
5:      for all out-neighbors j of i in G do
6:          for ℓ = 1, ..., k do
7:              c ← Pr_PL+C(R_{i≤ℓ}) / Pr_PL+C(R_{j≤ℓ})
8:              if c ≤ 1 then
9:                  b_j ← max{ b_j , b_i / (c − c b_i + b_i) }
10: return b_i for each i ∈ U such that p_i ≥ b_i
```

Algorithmically, we can use Theorem 5.2 to get initial lower bounds on consideration probabilities, and then Theorem 4.4—in particular, in the lower bound form (6)—to propagate them to yield lower bounds on other items' consideration probabilities. Define a directed graph $G = \langle V, E \rangle$ with $V = \mathcal{U}$ and an edge $\langle i, j \rangle$ for each $i, j \in \mathcal{U}$ such that $u_i > u_j$ and $\Pr_{\text{PL+C}}(\mathcal{R}_{i \leq \ell}) < \Pr_{\text{PL+C}}(\mathcal{R}_{j \leq \ell})$ for some ℓ. Each edge $\langle i, j \rangle$ represents a pair of items where Theorem 4.4 gives the lower bound (6) for p_j as a function of p_i. Because utilities strictly decrease along every edge, and thus along every path, the graph G is acyclic. We can therefore topologically sort G and use this ordering to propagate lower bounds on consideration probabilities. This procedure is formalized in Algorithm 1.

Theorem 5.3. *Algorithm 1 returns lower bounds b_i for each $i \in \mathcal{U}$ such that $p_i \geq b_i$ in time $O(kn^2)$.*

In the proof, we show that $\forall i \in \mathcal{U} : b_i \leq p_i$ is an invariant throughout the execution of the algorithm. The initial values of b_i set in Line 1 satisfy the condition by Theorem 5.2; all updates to values of b_i in Line 9 maintain this property by (6), the lower-bound form of Theorem 4.4.

6 UPPER BOUNDS

We now turn to *upper* bounds on consideration probabilities, in the same spirit as for lower bounds in Section 5. We will again be able to use Theorem 4.4 to tighten our initial upper bounds. However, upper bounding consideration probabilities turns out to be a bit trickier, so we will have to build some more technical infrastructure.

Our approach to inferring an upper bound on p_i is based on the intuition that increasing the consideration probability of a different item j only makes i less likely to be chosen; increasing the competition for i (by making j more likely to be an additional competitor) can only hurt i's chances. We can then use this idea to bound i's winning chances against a hypothetical scenario where all other consideration probabilities are 1, in which case the consideration mechanism is easy to analyze. But, surprisingly, that intuition about increasing j's consideration probability turns out to be false (but not by too much). As a result, our lemma statements will need to be a bit more technical. (For example, suppose that $\mathcal{U} = \{1, 2, 3\}$ with $u_1 = u_2 = +\infty$ and $u_3 = -\infty$. Let $k = 2$, and let $p_1 = 1$ and $p_2 = 0.02$. Item 3 never wins, but its presence affects the relative strengths of items 1 and 2. If $p_3 = 0$, then the only consideration set that we ever observe is $\{1, 2\}$, and so item 2 wins 50% of the time. But if p_3 is increased to 1, then 98% of the observed consideration sets are $\{1, 3\}$; the chance that item 2 wins drops from 50% to 1%.) It turns out that the issue that arises here is specifically about cases in which the consideration set is of size *exactly* k. Fortunately, it is rare that a length-k ranking emerges from a consideration set of size exactly equal to k (under our ongoing mild assumptions about $\sum_i p_i$):

Lemma 6.1. *If $\sum_{i \in \mathcal{U}} p_i \geq \alpha k$ for some $\alpha > 1$, then*

$$\sum_{C \subseteq \mathcal{U}, |C| = k} \Pr_{\mathrm{C}}(C) \leq \frac{\left(\alpha e^{1-\alpha}\right)^k}{1 - \left(\alpha e^{1-\alpha}\right)^k}.$$

We use this bound on the probability that $|C| = k$ to bound the amount by which i's chances of winning increases when j's consideration probability decreases.

Lemma 6.2. *Suppose $\sum_{i \in \mathcal{U}} p_i \geq \alpha k$ for some $\alpha > 1$. Let $i, j \in \mathcal{U}$ with $i \neq j$. Let ${\Pr_{\mathrm{PL+C}}}'$ represent PL+C probabilities if we replace p_j with $p_j' > p_j$. Doing so cannot increase the probability i is ranked first, up to an additive error term:*

$${\Pr_{\mathrm{PL+C}}}'(\mathcal{R}_{i=1}) \leq \Pr_{\mathrm{PL+C}}(\mathcal{R}_{i=1}) + \frac{\left(\alpha e^{1-\alpha}\right)^k}{1 - \left(\alpha e^{1-\alpha}\right)^k}.$$

If at least $k + 1$ consideration probabilities are 1 after increasing p_j, the inequality holds with no error:

$${\Pr_{\mathrm{PL+C}}}'(\mathcal{R}_{i=1}) \leq \Pr_{\mathrm{PL+C}}(\mathcal{R}_{i=1}).$$

The idea behind the proof is to group consideration sets containing i into pairs where one contains j and the other does not. Increasing j's consideration probability makes it more likely we observe the consideration set with j, which then makes it less likely we see i ranked first. However, this pairing fails if the first set in the pair has size k, since then its partner without j is infeasible. Lemma 6.1 allows us to bound the error introduced by these failed pairings. On the other hand, if $k + 1$ consideration probabilities are 1, then we never see the problematic size-k consideration sets.

Finally, using Lemma 6.2, we can upper bound p_i by comparing how often i is ranked first under the PL+C model to how often we would expect it to be ranked first if all other consideration probabilities were 1, which is easy to compute as a function of p_i.

Theorem 6.3. *Suppose that $\sum_{i \in \mathcal{U}} p_i \geq \alpha k$ for some $\alpha > 1$. Then for any $i \in \mathcal{U}$,*

$$p_i \leq \frac{\sum_{j \in \mathcal{U}} \exp(u_j)}{\exp(u_i)} \cdot \left(\Pr_{\mathrm{PL+C}}(\mathcal{R}_{i=1}) + \frac{k \left(\alpha e^{1-\alpha}\right)^k}{1 - \left(\alpha e^{1-\alpha}\right)^k} \right).$$

Just as we did with lower bounds, we can use Theorem 6.3 to find initial upper bounds on every p_i and then use Theorem 4.4 to tighten these bounds, this time using the upper bound formulation (5). The algorithm is identical in structure to the lower-bounding procedure; see Appendix A.1 for pseudocode and a proof of correctness.

7 APPLICATION TO DATA

To illustrate how our theoretical machinery can be used to infer consideration behavior, we analyze an existing dataset in which ≈2900 American participants in a psychology experiment were asked to compare contributions of the 50 U.S. states to U.S. history (Putnam et al., 2018). The data[1] include their responses to numerical questions (the *Random-10 question:* estimate the percentage of state X's contribution to U.S. history, for each state X in a set of ten randomly selected states) and ranking questions (the *Top-3 question:* identify, in order, the three states that contributed the most to U.S. history). For Random-10, we convert the numerical ratings into rankings by sorting. Since the set of states being considered is fixed by the Random-10 question itself, we model these rankings using Plackett–Luce without consideration (PL). However, the Top-3 question forces participants to construct their responses without a reference list of candidates, instead having to recall the names of the states that they will list (Brown, 1976); indeed, participants are unlikely to have considered all 50 states, making the Top-3 setting well suited to PL+C.

According to Theorem 4.1, we cannot hope to learn consideration probabilities for the 50 U.S. states. However, Algorithm 1 (and its upper-bound analog; see Algorithm 2 in Appendix A.1) enables us to at least infer

[1]Made available by Putnam et al. (2018) at `https://osf.io/tnjqs/` under a CC-BY 4.0 license.

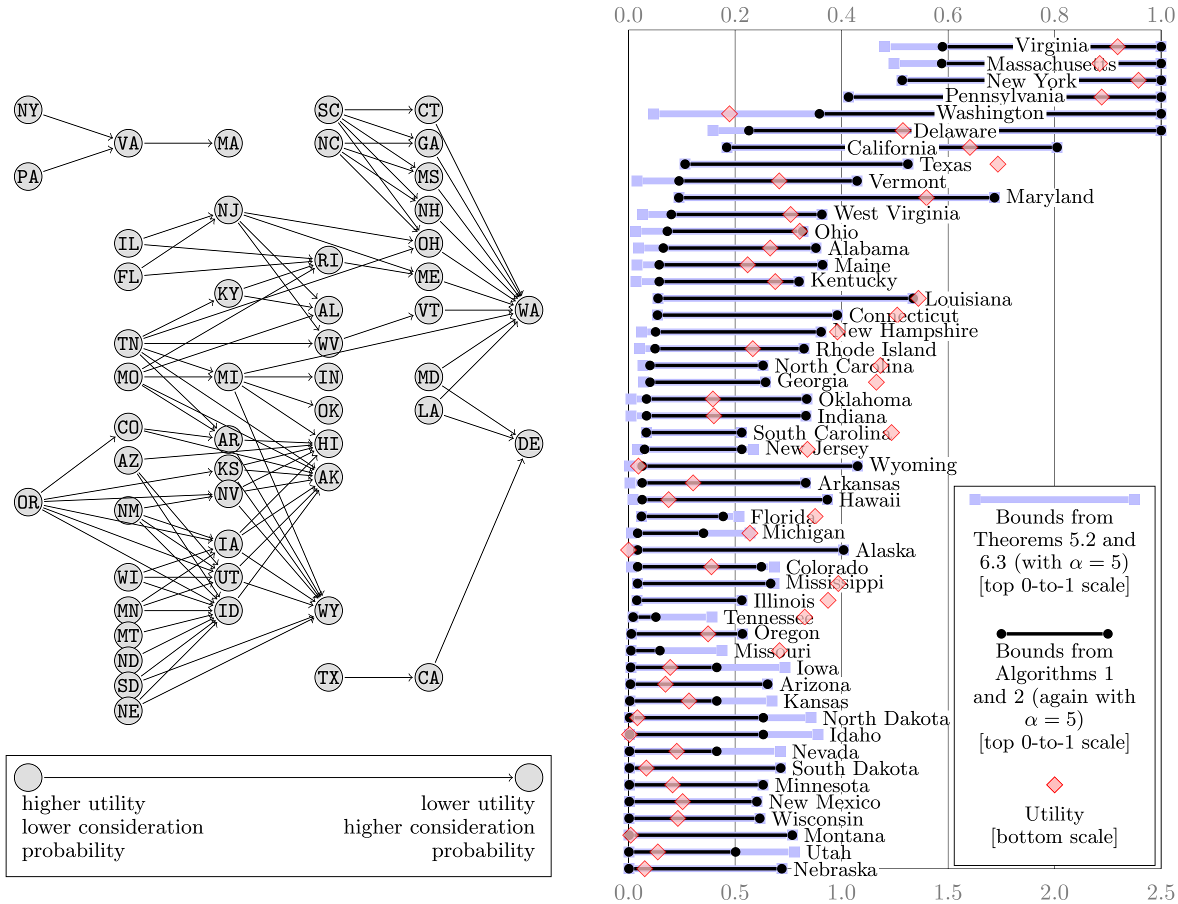

Figure 1: Left: the transitive reduction of the graph G produced by Algorithm 1 on the data from (Putnam et al., 2018); right: feasible consideration probability intervals for the 50 U.S. states. On the left, an edge from state i to state j indicates that i's inferred utility is larger than j's but that j appears in the top ℓ positions more often than i (for some $\ell \in \{1, 2, 3\}$). Thus, we conclude from Theorem 4.3 that $p_i \leq p_j$. Nodes are labeled with states' postal abbreviations. On the right, the light blue intervals show the bounds from Theorems 5.2 and 6.3, while the smaller black intervals show the bounds after tightening with Algorithms 1 and 2. Red diamonds show state utilities learned by PL on the Random-10 data, scaled additively so the minimum utility is 0.

bounds on these consideration probabilities given estimates of top-ℓ appearance rates $\Pr_{\text{PL+C}}(\mathcal{R}_{i \leq \ell})$, utilities u_i, and a lower bound on expected consideration set size αk. Because the Top-3 and Random-10 data feature rankings over the same universe of items (the states), if we model these responses with PL+C and PL, respectively, then it is natural to assume that the underlying utilities are the same between the two questions. Working from this assumption, we fit[2] a PL model (implemented in PyTorch (Paszke et al., 2019)) to the Random-10 data, learning the utility u_i for each state and taking these to be the utilities under the full PL+C model for Top-3. For each state i, we then calculate the empirical estimates of $\Pr_{\text{PL+C}}(\mathcal{R}_{i \leq \ell})$ for $\ell = 1, 2, 3$: the proportion of the Top-3 rankings in which state i appears in the first ℓ positions. Finally, we take $\alpha = 5$, assuming that participants on average consider at least 15 states when ranking their top 3.

Recall that Algorithm 1 (and, similarly, Algorithm 2) relies on pairs of states whose utilities and top-ℓ ranking rates are flipped.[3] There are many such flips in this data, highlighting the importance of consideration in the Top-3 question. We visualize these flips in Fig-

[2]We used Rprop (Riedmiller and Braun, 1993) on the full dataset with initial learning rate 0.05 and L_2 regularization strength 10^{-6}, stopping when the squared gradient magnitude fell below 10^{-8}. Since the PL model has a convex negative log likelihood, any well-tuned or robust optimizer would be expected to converge to a global optimum. While we used a train-test split for initial exploration, the results presented here are from the model fit to the entire dataset, as we do not have meaningful test metrics to evaluate. Given our assumptions, our bounds provably hold. Training takes less than a second on a 2018 MacBook Pro. Our code is available at `https://github.com/aoki-sherwoodb/bounding-consideration-probs`.

[3]To identify these flips, we use utilities learned from Random-10 data rather than Lemma 4.5, avoiding high sampling error on states that rarely appear in the Top-3.

ure 1 (left), which shows the directed acyclic graph G constructed in Algorithm 1. (For clarity, we display the *transitive reduction* (Aho et al., 1972) of G.) For each edge $\langle i, j \rangle$ in this graph, we know that state i has a lower consideration probability than state j, despite having higher utility. For instance, we find that Massachusetts has higher consideration probability than Virginia, which in turn has higher consideration probability than New York and Pennsylvania.

We also compute upper bounds on consideration probabilities using Theorem 6.3, with the same utilities u_i, values of α and k, and top-1 ranking probabilities $\Pr_{\text{PL+C}}(\mathcal{R}_{i=1})$ (again, using empirical estimates). Combining our lower and upper bounds yields feasible intervals on consideration probabilities for each state, which we display in Figure 1 (right). Our upper bounds reveal that, if our assumptions are valid, most states are considered less than 30–40% of the time. Additionally, the bounds on consideration probabilities align with theories about why certain states were highly rated in the data (Putnam et al., 2018). Of the eight states with the highest lower bound, five are states drawn from the thirteen original colonies and commonly associated with the American Revolution (Massachusetts, Virginia, New York, Pennsylvania, and Delaware), two are the largest U.S. states by population (California and Texas), and the last, Washington, was hypothesized by Putnam et al. (2018) to be confused by participants with the U.S. capital city Washington, D.C. and may also be easy to recall in the context of historical judgements due to its namesake.

8 DISCUSSION

We formalized a natural model of ranking with consideration, adding independent consideration to Plackett–Luce. Despite showing that consideration probabilities are not identified in general, we derived relative and absolute bounds that allow us to learn about possible ranges of consideration probabilities from observed ranking data. Our data application demonstrates how these bounds can be used in practice to gain insight into consideration behavior. In addition to providing behavioral insights, approximately recovering consideration probabilities has exciting possible applications in recommender systems, which could be tailored to suggest items that have high utility but low consideration probability.

The astute reader will notice that our bounds use exact top-ℓ ranking probabilities and utilities, but that empirical estimates from observed rankings are necessarily noisy (and may reverse whether $u_i > u_j$ or $u_i < u_j$). Luckily, there are some easy fixes. For Theorem 4.4, we can use upper and lower confidence intervals on top-ℓ ranking rates to compute a range of possible values of c (and, in particular, a high-probability guarantee under sampling error). To handle uncertainty in utility estimates, we can use confidence intervals to find only statistically significant utility-consideration flips.

There remains much to explore regarding the PL+C model, and ranking with consideration sets more generally. While PL+C is clearly at least as powerful as Plackett–Luce, how much expressive power is gained by adding consideration? What kinds of distributions over rankings can and cannot be expressed by a PL+C model? Another natural extension of the Plackett–Luce model allows violations of the independence of irrelevant alternatives (IIA) assumption built into Plackett–Luce, such as the contextual repeated selection (CRS) model (Seshadri et al., 2020). How does the expressive power of CRS compare to PL+C? Does PL+C allow (apparent) IIA violations due to the consideration stage? (We expect so; the three-item example in Section 6 is reminiscent of non-IIA behavior.)

Another interesting question concerns computing PL+C probabilities efficiently. If we know utilities and consideration probabilities, the direct approach to computing $\Pr_{\text{PL+C}}(r)$ using Equation (3) involves a sum over exponentially many consideration sets. Is it possible to exactly compute PL+C probabilities in polynomial time, or is it provably hard? We have derived two efficient approximation algorithms, but not a hardness proof or exact efficient algorithm. The first algorithm samples sufficiently many consideration sets and computes an empirical estimate of $\Pr_{\text{PL+C}}(r)$, while the second groups together consideration sets with similar total utilities (with "similarity" defined by a tolerance ϵ), computes their total consideration probabilities, and approximates their total utilities. Details of the algorithms are in Appendix A.2; we leave further exploration of this problem to future work.

Theorem 8.1. *There is a randomized ϵ-additive approximation algorithm for $\Pr_{\text{PL+C}}(r)$ with runtime $O(kn \log(1/\delta)/(z_{k,p}\epsilon^2))$, where both the approximation and runtime bounds hold with probability at least $1-\delta$.*

Theorem 8.2. *There is a deterministic algorithm approximating $\Pr_{\text{PL+C}}(r)$ with multiplicative error at most $1+\epsilon$ and runtime $O(kn^2 \max\{\log n, m\}/\epsilon)$, where $m = \max_{i \in \overline{r}} u_i$.*

The PL+C model also admits several natural extensions, such as incorporating different models of the consideration stage or allowing rankings of varied length. The latter direction may even simplify some of the consideration probability bounds, as some of the difficulty in our top-k setting arises because of the conditioning that $|C| \geq k$ (but this conditioning is often needed given the ubiquity of top-k data).

Acknowledgments

Portions of this work were carried out while all authors were at Carleton College and while K.T. was at Cornell University. Thanks to Aadi Akyianu, David Chu, Katrina Li, Adam Putnam, Sophie Quinn, Morgan Ross, Laura Soter, and Jeremy Yamashiro for helpful discussions. Comments are welcome.

References

Abaluck, J. and Adams-Prassl, A. (2021). What do consumers consider before they choose? Identification from asymmetric demand responses. *The Quarterly Journal of Economics*, 136(3):1611–1663.

Aho, A. V., Garey, M. R., and Ullman, J. D. (1972). The transitive reduction of a directed graph. *SIAM Journal on Computing*, 1(2):131–137.

Alvo, M. and Yu, P. L. (2014). *Statistical Methods for Ranking Data*. Springer, New York.

Başar, G. and Bhat, C. (2004). A parameterized consideration set model for airport choice: an application to the San Francisco Bay area. *Transportation Research Part B: Methodological*, 38(10):889–904.

Beggs, S., Cardell, S., and Hausman, J. (1981). Assessing the potential demand for electric cars. *Journal of Econometrics*, 17(1):1–19.

Ben-Akiva, M. and Boccara, B. (1995). Discrete choice models with latent choice sets. *International Journal of Research in Marketing*, 12(1):9–24.

Biscarri, W., Zhao, S. D., and Brunner, R. J. (2018). A simple and fast method for computing the Poisson binomial distribution function. *Computational Statistics & Data Analysis*, 122:92–100.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Brown, D. G. (2011). How I wasted too long finding a concentration inequality for sums of geometric variables. `https://cs.uwaterloo.ca/~browndg/negbin.pdf`.

Brown, J., editor (1976). *Recall and Recognition*. John Wiley & Sons, London.

Cattaneo, M. D., Ma, X., Masatlioglu, Y., and Suleymanov, E. (2020). A random attention model. *Journal of Political Economy*, 128(7):2796–2836.

Chintagunta, P. K. and Nair, H. S. (2011). Discrete-choice models of consumer demand in marketing. *Marketing Science*, 30(6):977–996.

Cutolo, M. (2023). The U.S. state everyone forgets when listing all 50. `https://www.rd.com/article/most-forgotten-us-state/`. Accessed 10/9/2023.

Danaf, M., Becker, F., Song, X., Atasoy, B., and Ben-Akiva, M. (2019). Online discrete choice models: Applications in personalized recommendations. *Decision Support Systems*, 119:35–45.

Fok, D., Paap, R., and Van Dijk, B. (2012). A rank-ordered logit model with unobserved heterogeneity in ranking capabilities. *Journal of Applied Econometrics*, 27(5):831–846.

Gu, B., Konana, P., and Chen, H.-W. M. (2012). Identifying consumer consideration set at the purchase time from aggregate purchase data in online retailing. *Decision Support Systems*, 53(3):625–633.

Guiver, J. and Snelson, E. (2009). Bayesian inference for Plackett-Luce ranking models. In *International Conference on Machine Learning*, pages 377–384.

Hauser, J. R. and Wernerfelt, B. (1990). An evaluation cost model of consideration sets. *Journal of Consumer Research*, 16(4):393–408.

Hausman, J. A. and Ruud, P. A. (1987). Specifying and testing econometric models for rank-ordered data. *Journal of Econometrics*, 34(1-2):83–104.

Horne, M., Jaccard, M., and Tiedemann, K. (2005). Improving behavioral realism in hybrid energy-economy models using discrete choice studies of personal transportation decisions. *Energy Economics*, 27(1):59–77.

Jagabathula, S., Mitrofanov, D., and Vulcano, G. (2023). Demand estimation under uncertain consideration sets. *Operations Research*.

Liu, A., Zhao, Z., Liao, C., Lu, P., and Xia, L. (2019). Learning Plackett-Luce mixtures from partial preferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4328–4335.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York.

Manski, C. F. (1977). The structure of random utility models. *Theory and Decision*, 8(3):229.

Manzini, P. and Mariotti, M. (2014). Stochastic choice and consideration sets. *Econometrica*, 82(3):1153–1176.

Maystre, L. and Grossglauser, M. (2015). Fast and accurate inference of Plackett–Luce models. *Advances in Neural Information Processing Systems*, 28.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In Zarembka, P., editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, New York.

Mitzenmacher, M. and Upfal, E. (2017). *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, second edition.

Moe, W. W. (2006). An empirical two-stage choice model with varying decision rules applied to internet clickstream data. *Journal of Marketing Research*, 43(4):680–692.

Nguyen, D. and Zhang, A. Y. (2023). Efficient and accurate learning of mixtures of Plackett-Luce models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9294–9301.

Palma, M. A. (2017). Improving the prediction of ranking data. *Empirical Economics*, 53:1681–1710.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32:8026–8037.

Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202.

Putnam, A. L., Ross, M. Q., Soter, L. K., and Roediger III, H. L. (2018). Collective narcissism: Americans exaggerate the role of their home state in appraising US history. *Psychological Science*, 29(9):1414–1422.

Riedmiller, M. and Braun, H. (1993). A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591. IEEE.

Roberts, J. H. and Lattin, J. M. (1991). Development and testing of a model of consideration set composition. *Journal of Marketing Research*, 28(4):429–440.

Saha, A. and Gopalan, A. (2020). From PAC to instance-optimal sample complexity in the Plackett-Luce model. In *International Conference on Machine Learning*, pages 8367–8376. PMLR.

Seshadri, A., Ragain, S., and Ugander, J. (2020). Learning rich rankings. *Advances in Neural Information Processing Systems*, 33:9435–9446.

Shocker, A. D., Ben-Akiva, M., Boccara, B., and Nedungadi, P. (1991). Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Marketing Letters*, 2:181–197.

Simon, H. A. (1957). *Models of Man: Social and Rational*. Wiley, New York.

Suh, J.-C. (2009). The role of consideration sets in brand choice: The moderating role of product characteristics. *Psychology & Marketing*, 26(6):534–550.

Thurner, P. W. (2000). The empirical application of the spatial theory of voting in multiparty systems with random utility models. *Electoral Studies*, 19(4):493–517.

Train, K. E. (2009). *Discrete Choice Methods with Simulation*. Cambridge University Press, Cambridge.

van Nierop, E., Bronnenberg, B., Paap, R., Wedel, M., and Franses, P. H. (2010). Retrieving unobserved consideration sets from household panel data. *Journal of Marketing Research*, 47(1):63–74.

Zhao, Z., Piech, P., and Xia, L. (2016). Learning mixtures of Plackett-Luce models. In *International Conference on Machine Learning*, pages 2906–2914.

Zhao, Z. and Xia, L. (2019). Learning mixtures of Plackett-Luce models from structured partial orders. *Advances in Neural Information Processing Systems*, 32.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. Yes
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. Yes
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. Yes

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. Yes
   (b) Complete proofs of all theoretical results. Yes
   (c) Clear explanations of any assumptions. Yes

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). Yes
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). Yes
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). Not Applicable
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). Yes

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. Yes
   (b) The license information of the assets, if applicable. Yes
   (c) New assets either in the supplemental material or as a URL, if applicable. Yes
   (d) Information about consent from data providers/curators. Not Applicable
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. Not Applicable
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. Not Applicable
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. Not Applicable
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. Not Applicable

A DEFFERED PROOFS

Theorem 4.1. *For all $n \geq 1$ and $1 \leq k \leq n$, consideration probabilities are not identifiable in the PL+C model. That is, there are multiple sets of consideration probabilities that generate the same distribution over rankings, with fixed utilities u_i for each $i \in \mathcal{U}$.*

Proof. We begin with trivial cases. If $n = 1$, then the only observed ranking is $\langle 1 \rangle$, regardless of p_1. If $k = n$, then the only consideration set with nonzero probability is $\mathcal{U}$, so the consideration probabilities do not affect observed rankings. In both cases, arbitrary nonzero consideration probabilities produce the same distribution over rankings, so the probabilities are not identifiable.

Now suppose $n > 1$ and $k < n$. We will define $k-1$ symmetric "excellent" items, $n-k$ symmetric "good" items, and one "bad" item:

- $k-1$ excellent items: $u_1, \dots, u_{k-1} = +\infty$ and $p_1, \dots, p_{k-1} = 1$,
- $n-k$ good items: $u_k, \dots, u_{n-1} = 1$ and $p_k, \dots, p_{n-1} = g$, and
- the one bad item: $u_n = -\infty$ and $p_n = b$,

for some fixed probabilities g and b. Excellent items are always considered and occupy the top $k-1$ positions in the ranking, so only two types of rankings have nonzero PL+C probability: "good" rankings with any of the $n-k$ good items in last (i.e., kth) place, and "bad" rankings with the bad item in last place. We can only observe a bad ranking if the consideration set includes the bad item but none of the good items; we observe a good ranking if at least one good item was considered. (If neither the bad item nor any good item was considered, which happens with probability $(1-b)(1-g)^{n-k}$, the ranking is unobserved.) Thus, after normalization, the probability of observing a bad ranking is

$$c = \frac{b \cdot (1-g)^{n-k}}{1-(1-b)(1-g)^{n-k}}. \tag{7}$$

By symmetry, all bad rankings have equal probability, as do all good rankings; thus the probability mass of c is evenly divided over bad rankings, and the remaining $(1-c)$ mass is evenly divided among good rankings. The distribution of rankings is thus governed by the single parameter c. To get multiple consideration probabilities yielding the same distribution, we can fix an appropriate value of c, select any $g \in (0,1)$, and solve for b in (7). To complete the proof, we verify this equation is feasible (i.e., yields $b \in (0,1]$) for sufficiently small values of c. Write $\lambda = (1-g)^{n-k}$. Then (7) says that $c = b\lambda/(1-(1-b)\lambda)$ or, rearranging, that $b = \frac{c}{1-c} \cdot \frac{1-\lambda}{\lambda}$. Because $g \in (0,1)$ implies $\lambda \in (0,1)$, this formula yields a value of $b \in (0,1]$ so long as $0 < c \leq \lambda$. □

Lemma 4.2. *For two items $i, j \in \mathcal{U}$ with $u_i > u_j$, let $C \subseteq \mathcal{U}$ with $i \in C$ and $j \notin C$ and let $C_{i \to j} = C \setminus \{i\} \cup \{j\}$ be C with i replaced by j. Let $\ell \leq k \leq |C|$. Under Plackett–Luce, i is more likely to be in the top ℓ positions of a length-k ranking with consideration set C than j is with consideration set $C_{i \to j}$. That is,*

$$\Pr_{\mathrm{PL}}(\mathcal{R}_{i \leq \ell} \mid C) > \Pr_{\mathrm{PL}}(\mathcal{R}_{j \leq \ell} \mid C_{i \to j}).$$

Proof. We equivalently show that the probability i is *not* in the top ℓ positions under C is *smaller* than the probability that j is *not* in the top ℓ positions under $C_{i \to j}$. We can find the probability that i is not in the top ℓ positions by summing over every permutation of length-ℓ subsets of $C \setminus \{i\}$ and, for each permutation, multiplying the probabilities yielding that permutation under Plackett–Luce. When we replace i with j, every Plackett–Luce probability strictly increases, since we replace the term $\exp(u_i)$ in every denominator with $\exp(u_j)$, which is strictly smaller since $u_i > u_j$. □

Theorem 4.4. *Consider two items i, j with $u_i > u_j$ in a PL+C model. If $c = \Pr_{\mathrm{PL+C}}(\mathcal{R}_{i \leq \ell})/\Pr_{\mathrm{PL+C}}(\mathcal{R}_{j \leq \ell}) \leq 1$ for some ℓ, then*

$$\frac{p_i}{1-p_i} \leq c \cdot \frac{p_j}{1-p_j}. \tag{4}$$

Proof. Since we are considering a fixed ℓ in this proof, write $\mathcal{R}_i = \mathcal{R}_{i \leq \ell}$ and $\mathcal{R}_j = \mathcal{R}_{j \leq \ell}$ for brevity. Additionally, let $\mathcal{R}_{i \setminus j} = \mathcal{R}_i \setminus \mathcal{R}_j$ and $\mathcal{R}_{i \cap j} = \mathcal{R}_i \cap \mathcal{R}_j$. We can partition $\mathcal{R}_i$ into the rankings that contain i but not j in the first ℓ positions, $\mathcal{R}_{i \setminus j}$, and the rankings that contain both in the first ℓ positions, $\mathcal{R}_{i \cap j}$. Analogously define $\mathcal{R}_{j \setminus i}$. Thus $\Pr_{\mathrm{PL+C}}(\mathcal{R}_i) = \Pr_{\mathrm{PL+C}}(\mathcal{R}_{i \setminus j}) + \Pr_{\mathrm{PL+C}}(\mathcal{R}_{i \cap j})$ and $\Pr_{\mathrm{PL+C}}(\mathcal{R}_j) = \Pr_{\mathrm{PL+C}}(\mathcal{R}_{j \setminus i}) + \Pr_{\mathrm{PL+C}}(\mathcal{R}_{i \cap j})$.

Suppose $c = \Pr_{\text{PL+C}}(\mathcal{R}_i)/\Pr_{\text{PL+C}}(\mathcal{R}_j) \leq 1$. By the above equalities, then, we have

$$\begin{aligned}
&\Pr_{\text{PL+C}}(\mathcal{R}_{i\setminus j}) + \Pr_{\text{PL+C}}(\mathcal{R}_{i\cap j}) = c \cdot \Pr_{\text{PL+C}}(\mathcal{R}_{j\setminus i}) + c \cdot \Pr_{\text{PL+C}}(\mathcal{R}_{i\cap j}) \\
&\Leftrightarrow \Pr_{\text{PL+C}}(\mathcal{R}_{i\setminus j}) + (1-c)\Pr_{\text{PL+C}}(\mathcal{R}_{i\cap j}) = c \cdot \Pr_{\text{PL+C}}(\mathcal{R}_{j\setminus i}) \\
&\Rightarrow \Pr_{\text{PL+C}}(\mathcal{R}_{i\setminus j}) \leq c \cdot \Pr_{\text{PL+C}}(\mathcal{R}_{j\setminus i}). \qquad (8)
\end{aligned}$$

We'll write out both sides of (8) as a sum over consideration sets, starting with $\Pr_{\text{PL+C}}(\mathcal{R}_{i\setminus j})$. Because $\Pr_{\text{PL}}(r \mid C) = 0$ if the items in r are not all in C, for $\Pr_{\text{PL+C}}(\mathcal{R}_{i\setminus j})$ we only need to sum over consideration sets that include i, which either also include j or do not. Denote these two collections of consideration sets $\mathcal{C}_{i\cap j} = \{C \subseteq \mathcal{U} \mid i \in C, j \in C\}$ and $\mathcal{C}_{i\setminus j} = \{C \subseteq \mathcal{U} \mid i \in C, j \notin C\}$. Then

$$\begin{aligned}
\Pr_{\text{PL+C}}(\mathcal{R}_{i\setminus j}) &= \sum_{\substack{r\in\mathcal{R}_{i\setminus j}\\ C\subseteq\mathcal{U}}} \Pr_{\text{C}}(C) \cdot \Pr_{\text{PL}}(r \mid C) \\
&= \sum_{\substack{r\in\mathcal{R}_{i\setminus j}\\ C\subseteq\mathcal{C}_{i\cap j}}} \Pr_{\text{C}}(C) \cdot \Pr_{\text{PL}}(r \mid C) + \sum_{\substack{r\in\mathcal{R}_{i\setminus j}\\ C\subseteq\mathcal{C}_{i\setminus j}}} \Pr_{\text{C}}(C) \cdot \Pr_{\text{PL}}(r \mid C). \qquad (9)
\end{aligned}$$

Now we'll do the same for $\Pr_{\text{PL+C}}(\mathcal{R}_{j\setminus i})$. We can construct a bijection $\phi : \mathcal{R}_{i\setminus j} \to \mathcal{R}_{j\setminus i}$ by taking a ranking $r \in \mathcal{R}_{i\setminus j}$ and replacing i with j. We can also construct a bijection $\psi : \mathcal{C}_{i\setminus j} \to \mathcal{C}_{j\setminus i}$ by taking a set $C \in \mathcal{C}_{i\setminus j}$ and replacing i with j. Using these bijections, we can decompose $\Pr_{\text{PL+C}}(\mathcal{R}_{j\setminus i})$ into sums over the same sets as in (9). For clarity, we'll write $r_{i\to j} = \phi(r)$ and $C_{i\to j} = \psi(C)$. We then have:

$$\begin{aligned}
\Pr_{\text{PL+C}}(\mathcal{R}_{j\setminus i}) &= \sum_{\substack{r\in\mathcal{R}_{j\setminus i}\\ C\subseteq\mathcal{U}}} \Pr_{\text{C}}(C) \cdot \Pr_{\text{PL}}(r \mid C) \\
&= \sum_{\substack{r\in\mathcal{R}_{i\setminus j}\\ C\subseteq\mathcal{U}}} \Pr_{\text{C}}(C) \cdot \Pr_{\text{PL}}(r_{i\to j} \mid C) \\
&= \sum_{\substack{r\in\mathcal{R}_{i\setminus j}\\ C\subseteq\mathcal{C}_{i\cap j}}} \Pr_{\text{C}}(C) \cdot \Pr_{\text{PL}}(r_{i\to j} \mid C) \quad + \quad \sum_{\substack{r\in\mathcal{R}_{i\setminus j}\\ C\subseteq\mathcal{C}_{i\setminus j}}} \Pr_{\text{C}}(C_{i\to j}) \cdot \Pr_{\text{PL}}(r_{i\to j} \mid C_{i\to j}). \qquad (10)
\end{aligned}$$

Combining (8) with (9) and (10), we then have

$$\begin{aligned}
&\sum_{\substack{r\in\mathcal{R}_{i\setminus j}\\ C\subseteq\mathcal{C}_{i\cap j}}} \Pr_{\text{C}}(C) \cdot \Pr_{\text{PL}}(r \mid C) + \sum_{\substack{r\in\mathcal{R}_{i\setminus j}\\ C\subseteq\mathcal{C}_{i\setminus j}}} \Pr_{\text{C}}(C) \cdot \Pr_{\text{PL}}(r \mid C) \\
&\leq c \cdot \sum_{\substack{r\in\mathcal{R}_{i\setminus j}\\ C\subseteq\mathcal{C}_{i\cap j}}} \Pr_{\text{C}}(C) \cdot \Pr_{\text{PL}}(r_{i\to j} \mid C) \quad + \quad c \cdot \sum_{\substack{r\in\mathcal{R}_{i\setminus j}\\ C\subseteq\mathcal{C}_{i\setminus j}}} \Pr_{\text{C}}(C_{i\to j}) \cdot \Pr_{\text{PL}}(r_{i\to j} \mid C_{i\to j}). \qquad (11)
\end{aligned}$$

Now, we show the first sum on the left-hand side of (11) is larger than the first sum on the right-hand side (i.e., comparing the summations over $\mathcal{C}_{i\cap j}$, for the moment disregarding the factor of c). To do so, consider some fixed $r^* \in \mathcal{R}_{i\setminus j}$. Recall that $u_i > u_j$. We will show that, for all $C \subseteq \mathcal{C}_{i\cap j}$, we have $\Pr_{\text{PL}}(r^* \mid C) > \Pr_{\text{PL}}(r^*_{i\to j} \mid C)$. Suppose i appears at position h in r^*. We then have

$$\begin{aligned}
&\Pr_{\text{PL}}(r^* \mid C) \\
&= \prod_{s=1}^{h-1} \frac{\exp(u_{r^*_s})}{\sum_{t\in C\setminus\{r^*_1,\ldots,r^*_{s-1}\}} \exp(u_t)} \cdot \frac{\exp(u_i)}{\sum_{t\in C\setminus\{r^*_1,\ldots,r^*_{h-1}\}} \exp(u_t)} \cdot \prod_{s=h+1}^{k} \frac{\exp(u_{r^*_s})}{\sum_{t\in C\setminus\{r^*_1,\ldots,r^*_{s-1}\}} \exp(u_t)}.
\end{aligned}$$

$\Pr_{\text{PL}}(r^*_{i\to j} \mid C)$ is identical, except the middle probability has numerator $\exp(u_j)$ instead of $\exp(u_i)$ and the denominators in the bottom product include $\exp(u_i)$ instead of $\exp(u_j)$. Since $u_i > u_j$, this means $\Pr_{\text{PL}}(r^* \mid C) > \Pr_{\text{PL}}(r^*_{i\to j} \mid C)$. Summing over all r^* and C then yields

$$\sum_{\substack{r\in\mathcal{R}_{i\setminus j}\\ C\subseteq\mathcal{C}_{i\cap j}}} \Pr_{\text{C}}(C) \cdot \Pr_{\text{PL}}(r \mid C) > \sum_{\substack{r\in\mathcal{R}_{i\setminus j}\\ C\subseteq\mathcal{C}_{i\cap j}}} \Pr_{\text{C}}(C) \cdot \Pr_{\text{PL}}(r_{i\to j} \mid C). \qquad (12)$$

Inequality (12) certainly still holds after scaling the right side by $c \leq 1$. We can therefore remove these sums from (11) and maintain the inequality, since we are shrinking the left by more than the right. This leaves us with

$$\sum_{\substack{r \in \mathcal{R}_{i \setminus j} \\ C \subseteq \mathcal{C}_{i \setminus j}}} \Pr_{\mathrm{C}}(C) \cdot \Pr_{\mathrm{PL}}(r \mid C) \leq c \cdot \sum_{\substack{r \in \mathcal{R}_{i \setminus j} \\ C \subseteq \mathcal{C}_{i \setminus j}}} \Pr_{\mathrm{C}}(C_{i \to j}) \cdot \Pr_{\mathrm{PL}}(r_{i \to j} \mid C_{i \to j}). \tag{13}$$

Next, notice that to transform $\Pr_{\mathrm{C}}(C)$ into $\Pr_{\mathrm{C}}(C_{i \to j})$, we don't consider i and instead consider j:

$$\Pr_{\mathrm{C}}(C_{i \to j}) = \Pr_{\mathrm{C}}(C) \cdot \tfrac{p_j(1-p_i)}{p_i(1-p_j)}. \tag{14}$$

Applying (14) to (13) and rearranging the sums yields

$$\begin{aligned} &\sum_{C \subseteq \mathcal{C}_{i \setminus j}} \Big[\Pr_{\mathrm{C}}(C) \sum_{r \in \mathcal{R}_{i \setminus j}} \Pr_{\mathrm{PL}}(r \mid C)\Big] \\ &\leq c \cdot \sum_{C \subseteq \mathcal{C}_{i \setminus j}} \Big[\Pr_{\mathrm{C}}(C_{i \to j}) \sum_{r \in \mathcal{R}_{i \setminus j}} \Pr_{\mathrm{PL}}(r_{i \to j} \mid C_{i \to j})\Big] \\ &= c \cdot \tfrac{p_j(1-p_i)}{p_i(1-p_j)} \cdot \sum_{C \subseteq \mathcal{C}_{i \setminus j}} \Big[\Pr_{\mathrm{C}}(C) \sum_{r \in \mathcal{R}_{i \setminus j}} \Pr_{\mathrm{PL}}(r_{i \to j} \mid C_{i \to j})\Big]. \end{aligned} \tag{15}$$

Now, notice that $\sum_{r \in \mathcal{R}_{i \setminus j}} \Pr_{\mathrm{PL}}(r \mid C)$ is the probability that i is ranked in the top ℓ positions under Plackett–Luce with consideration set C (since $j \notin C$). Similarly, $\sum_{r \in \mathcal{R}_{i \setminus j}} \Pr_{\mathrm{PL}}(r_{i \to j} \mid C_{i \to j})$ is the probability that j is ranked in the top ℓ positions with consideration set $C_{i \to j}$, which is just C with i replaced by j. By Lemma 4.2, we therefore have $\sum_{r \in \mathcal{R}_{i \setminus j}} \Pr_{\mathrm{PL}}(r \mid C) > \sum_{r \in \mathcal{R}_{i \setminus j}} \Pr_{\mathrm{PL}}(r_{i \to j} \mid C_{i \to j})$.

We can therefore increase the right hand side of (15) to find

$$\begin{aligned} &\sum_{C \subseteq \mathcal{C}_{i \setminus j}} \Big[\Pr_{\mathrm{C}}(C) \sum_{r \in \mathcal{R}_{i \setminus j}} \Pr_{\mathrm{PL}}(r \mid C)\Big] \\ &\leq c \cdot \tfrac{p_j(1-p_i)}{p_i(1-p_j)} \cdot \sum_{C \subseteq \mathcal{C}_{i \setminus j}} \Big[\Pr_{\mathrm{C}}(C) \sum_{r \in \mathcal{R}_{i \setminus j}} \Pr_{\mathrm{PL}}(r \mid C)\Big]. \end{aligned}$$

Dividing both sides by the left-hand side then yields $1 \leq c \cdot \frac{p_j(1-p_i)}{p_i(1-p_j)}$, or, rearranging, the desired relationship (4):

$$\tfrac{p_i}{1-p_i} \leq c \cdot \tfrac{p_j}{1-p_j}.$$

□

Lemma 4.5. *Let R be a random top-k ranking generated by Plackett–Luce with consideration. For items $i, j \in \mathcal{U}$, let $i \succ_R j$ denote that i is ranked higher in R than j. If $\Pr(i \succ_R j \mid i, j \in R) > 1/2$, then $u_i > u_j$.*

Proof. Consider the set of top-k rankings containing both i and j, denoted $\mathcal{R}_{i \cap j}$. This set of rankings can be partitioned into two equal-size subsets, one in which i is ranked above j and the other in which j is ranked above i. Moreover, there is a bijection between these sets of rankings defined by swapping the position of i and j. Now suppose (towards proving the contrapositive of the claim) that $u_j \geq u_i$. In each pair of rankings defined by the above bijection, the one with j ranked higher has higher probability under Placett–Luce. To see this, note that in the Plackett–Luce probabilities (Equation (1)) of two corresponding rankings, swapping i and j only affects the ranking probability in the denominator (since the product of numerators commutes). The ranking with j ranked higher has strictly smaller denominators in its Plackett–Luce probability for the choices between j and i's position (including i's position, but not j's), since the higher-utility item is removed from the set of options after it is chosen. Thus it is more likely under Plackett–Luce (or at least as likely in the case that $u_i = u_j$). We then have that in the bijection between the parts of $\mathcal{R}_{i \cap j}$ with j and i ranked higher, the set of rankings with j ranked higher are more probable under Plackett–Luce *with consideration* (for every feasible consideration set, the ranking with j ranked higher is more probable). Thus, $\Pr(j \succ_R i \mid i, j \in R) \geq 1/2$. Taking the contrapositive, we find that if $\Pr(j \succ_R i \mid i, j \in R) < 1/2$ (or equivalently, $\Pr(i \succ_R j \mid i, j \in R) > 1/2$), then $u_i > u_j$. □

Lemma 5.1. *Suppose $\sum_{i\in\mathcal{U}} p_i \geq \alpha k$ for some $\alpha > 1$, and let X be a random variable denoting the size of a random consideration set (prior to conditioning on there being at least k considered items). Then*

$$\Pr(X \leq k) \leq \left(\alpha e^{1-\alpha}\right)^k.$$

Proof. By hypothesis, $\mathrm{E}[X] = \sum_{i\in\mathcal{U}} p_i \geq \alpha k$. Thus,

$$\begin{aligned}
\Pr(X \leq k) &= \Pr(X \leq (1/\alpha)\alpha k) \\
&\leq \Pr\left(X \leq (1/\alpha)E[X]\right) && \text{(since } E[X] \geq \alpha k\text{)} \\
&= \Pr\left(X \leq (1-(1-1/\alpha))E[X]\right) \\
&\leq \left(\frac{e^{-(1-1/\alpha)}}{(1/\alpha)^{1/\alpha}}\right)^{E[X]} && \text{(Chernoff bound (Mitzenmacher and Upfal, 2017))} \\
&\leq \left(\frac{e^{-(1-1/\alpha)}}{(1/\alpha)^{1/\alpha}}\right)^{\alpha k} && (E[X] \geq \alpha k \text{ and the ratio is } \leq 1) \\
&= \left(\frac{e^{-\alpha+1}}{1/\alpha}\right)^{k} \\
&= \left(\alpha e^{1-\alpha}\right)^k.
\end{aligned}$$

□

Theorem 5.2. *Suppose $\sum_{i\in\mathcal{U}} p_i \geq \alpha k$ for some $\alpha > 1$. For any $i \in \mathcal{U}$, the consideration probability for i is lower bounded as*

$$p_i \geq \Pr_{\mathrm{PL+C}}(\mathcal{R}_{i\leq k}) \cdot \left[1 - \left(\alpha e^{1-\alpha}\right)^k\right].$$

Proof. Consider a particular instance where a chooser formed a consideration set. Let $X = \sum_{i\in\mathcal{U}} X_i$ be the random variable representing the size of the consideration set, summing over indicator random variables (with X_i denoting "was item i considered?"). We only observe a ranking from this instance if $X \geq k$. We can bound the probability that the consideration set meets this constraint:

$$\Pr(X < k) \leq \Pr(X \leq k) \leq \left(\alpha e^{1-\alpha}\right)^k, \qquad \text{(by Lemma 5.1)}$$

and therefore

$$\Pr(X \geq k) = 1 - \Pr(X < k) \geq 1 - \left(\alpha e^{1-\alpha}\right)^k. \tag{16}$$

Also, note that if we observe i in a length-k ranking, it must have been the case that $X_i = 1$, and we have conditioned on $X \geq k$. Thus,

$$\Pr_{\mathrm{PL+C}}(\mathcal{R}_{i\leq k}) \leq \Pr(X_i \mid X \geq k). \tag{17}$$

We can use this to lower bound p_i:

$$\begin{aligned}
p_i &= \Pr(X_i) \\
&= \Pr(X_i \mid X \geq k) \cdot \Pr(X \geq k) + \Pr(X_i \mid X < k) \cdot \Pr(X < k) \\
&\geq \Pr(X_i \mid X \geq k) \cdot \Pr(X \geq k) \\
&\geq \Pr(X_i \mid X \geq k) \cdot \left[1 - \left(\alpha e^{1-\alpha}\right)^k\right] && \text{by (16)} \\
&\geq \Pr_{\mathrm{PL+C}}(\mathcal{R}_{i\leq k}) \cdot \left[1 - \left(\alpha e^{1-\alpha}\right)^k\right], && \text{by (17)}
\end{aligned}$$

just as desired. □

Lemma 6.1. *If $\sum_{i\in\mathcal{U}} p_i \geq \alpha k$ for some $\alpha > 1$, then*

$$\sum_{C\subseteq\mathcal{U}, |C|=k} \Pr_{\mathrm{C}}(C) \leq \frac{\left(\alpha e^{1-\alpha}\right)^k}{1-\left(\alpha e^{1-\alpha}\right)^k}.$$

Proof. Let X be the random variable representing the size of the consideration set. Then we can write

$$\begin{aligned}\sum_{C\subseteq\mathcal{U},|C|=k}\Pr_{\mathrm{C}}(C) &= \Pr(X=k\mid X\geq k)\\ &= \frac{\Pr(X=k)}{\Pr(X\geq k)}\\ &\leq \frac{\Pr(X\leq k)}{\Pr(X>k)}\\ &= \frac{\Pr(X\leq k)}{1-\Pr(X\leq k)}\\ &\leq \frac{\left(\alpha e^{1-\alpha}\right)^k}{1-\left(\alpha e^{1-\alpha}\right)^k},\end{aligned}$$

where the last inequality follows by Lemma 5.1. ☐

Lemma 6.2. *Suppose* $\sum_{i\in\mathcal{U}}p_i\geq\alpha k$ *for some* $\alpha>1$*. Let* $i,j\in\mathcal{U}$ *with* $i\neq j$*. Let* ${\Pr_{\mathrm{PL+C}}}'$ *represent PL+C probabilities if we replace* p_j *with* $p'_j>p_j$*. Doing so cannot increase the probability* i *is ranked first, up to an additive error term:*

$${\Pr_{\mathrm{PL+C}}}'(\mathcal{R}_{i=1})\leq\Pr_{\mathrm{PL+C}}(\mathcal{R}_{i=1})+\frac{\left(\alpha e^{1-\alpha}\right)^k}{1-\left(\alpha e^{1-\alpha}\right)^k}.$$

If at least $k+1$ *consideration probabilities are* 1 *after increasing* p_j*, the inequality holds with no error:*

$${\Pr_{\mathrm{PL+C}}}'(\mathcal{R}_{i=1})\leq\Pr_{\mathrm{PL+C}}(\mathcal{R}_{i=1}).$$

Proof. We will be considering the probability that item i is chosen first from a particular consideration set C. For brevity, define

$$\Pr(i\mid C)=\sum_{r\in\mathcal{R}_{i=1}}\Pr_{\mathrm{PL}}(r\mid C),$$

summing over all rankings in which i is chosen first. (We could also have defined $\Pr(i\mid C)$ as $\exp(u_i)/\sum_{h\in C}\exp(u_h)$, by definition.)

First, we'll shrink $\Pr_{\mathrm{PL+C}}(\mathcal{R}_{i=1})$ so it only involves a sum over certain consideration sets (excluding those where $|C\cup\{j\}|=k$):

$$\begin{aligned}\Pr_{\mathrm{PL+C}}(\mathcal{R}_{i=1}) &= \sum_{r\in\mathcal{R}_{i=1}}\sum_{C\subseteq\mathcal{U}}\Pr_{\mathrm{C}}(C)\cdot\Pr_{\mathrm{PL}}(r\mid C)\\ &= \sum_{C\subseteq\mathcal{U}}\left[\Pr_{\mathrm{C}}(C)\sum_{r\in\mathcal{R}_{i=1}}\Pr_{\mathrm{PL}}(r\mid C)\right]\\ &= \sum_{C\subseteq\mathcal{U}}\Pr_{\mathrm{C}}(C)\cdot\Pr(i\mid C)\\ &\geq \sum_{\substack{C\subseteq\mathcal{U}\\|C\cup\{j\}|>k\\ i\in C}}\Pr_{\mathrm{C}}(C)\cdot\Pr(i\mid C).\end{aligned}\tag{18}$$

We partition the consideration sets in this summation into those that contain j and those that don't, using the notation $\mathcal{C}_{i\cap j}=\{C\subseteq\mathcal{U}\mid i\in C,j\in C\}$ and $\mathcal{C}_{i\setminus j}=\{C\subseteq\mathcal{U}\mid i\in C,j\notin C\}$. It will be helpful to pair each consideration set containing j with its corresponding set excluding j, by defining $\phi:\mathcal{C}_{i\cap j}\to\mathcal{C}_{i\setminus j}$ where $\phi(C)=$

$C \setminus \{j\}$. We can then rewrite the summation in (18) as

$$\begin{aligned}
&\sum_{\substack{C \subseteq \mathcal{U} \\ |C \cup \{j\}| > k \\ i \in C}} \Pr_{\mathrm{C}}(C) \cdot \Pr(i \mid C) \\
&= \sum_{\substack{C \in \mathcal{C}_{i\,\cap\,j} \\ |C| > k}} \Pr_{\mathrm{C}}(C) \cdot \Pr(i \mid C) + \sum_{\substack{C \in \mathcal{C}_{i \setminus j} \\ |C \cup \{j\}| > k}} \Pr_{\mathrm{C}}(C) \cdot \Pr(i \mid C) \\
&= \sum_{\substack{C \in \mathcal{C}_{i\,\cap\,j} \\ |C| > k}} \left[\Pr_{\mathrm{C}}(C) \cdot \Pr(i \mid C) + \Pr_{\mathrm{C}}(\phi(C)) \cdot \Pr(i \mid \phi(C))\right]. \qquad (19)
\end{aligned}$$

Let $z_{k,p'}$ be the normalization constant in $\Pr_{\mathrm{C}}$ after replacing p_j with $p'_j > p_j$. Because $z_{k,p}$ denotes the probability that at least k items are considered, we have $z_{k,p'} \geq z_{k,p}$. For any $C \in \mathcal{C}_{i\,\cap\,j}$ with $|C| > k$ (and note that this latter restriction is important for (21), below: if $|C| \leq k$, then $\Pr_{\mathrm{C}}(\phi(C)) = 0$), we then have

$$\begin{aligned}
\Pr_{\mathrm{C}}(C) &= \tfrac{1}{z_{k,p}} \cdot \textstyle\prod_{s \in C} p_s \cdot \prod_{t \in \mathcal{U} \setminus C}(1 - p_t) \\
&= \tfrac{1}{z_{k,p}} \cdot p_j \cdot \textstyle\prod_{s \in \phi(C)} p_s \cdot \prod_{t \in \mathcal{U} \setminus C}(1 - p_t) \\
&\geq \tfrac{1}{z_{k,p'}} \cdot p_j \cdot \textstyle\prod_{s \in \phi(C)} p_s \cdot \prod_{t \in \mathcal{U} \setminus C}(1 - p_t) \qquad (20)
\end{aligned}$$

and, similarly,

$$\begin{aligned}
\Pr_{\mathrm{C}}(\phi(C)) &= \tfrac{1}{z_{k,p}} \cdot \textstyle\prod_{s \in \phi(C)} p_s \cdot \prod_{t \in \mathcal{U} \setminus \phi(C)}(1 - p_t) \\
&= \tfrac{1}{z_{k,p}} \cdot (1 - p_j) \cdot \textstyle\prod_{s \in \phi(C)} p_s \cdot \prod_{t \in \mathcal{U} \setminus C}(1 - p_t) \\
&\geq \tfrac{1}{z_{k,p'}} \cdot (1 - p_j) \cdot \textstyle\prod_{s \in \phi(C)} p_s \cdot \prod_{t \in \mathcal{U} \setminus C}(1 - p_t). \qquad (21)
\end{aligned}$$

Combining (20) and (21), still for $C \in \mathcal{C}_{i\,\cap\,j}$ with $|C| > k$, we have

$$\begin{aligned}
&\Pr_{\mathrm{C}}(C) \cdot \Pr(i \mid C) + \Pr_{\mathrm{C}}(\phi(C)) \cdot \Pr(i \mid \phi(C)) \\
\geq\ &\Pr(i \mid C) \cdot \tfrac{1}{z_{k,p'}} \cdot p_j \cdot \textstyle\prod_{s \in \phi(C)} p_s \cdot \prod_{t \in \mathcal{U} \setminus C}(1 - p_t) \\
&+ \Pr(i \mid \phi(C)) \cdot \tfrac{1}{z_{k,p'}} \cdot (1 - p_j) \cdot \textstyle\prod_{s \in \phi(C)} p_s \cdot \prod_{t \in \mathcal{U} \setminus C}(1 - p_t). \qquad (22)
\end{aligned}$$

Now note that (22) has the form $p_j \cdot a + (1 - p_j) \cdot b$, where $a < b$ (since $\Pr(i \mid C) < \Pr(i \mid \phi(C))$). Thus, when we replace p_j with $p'_j > p_j$, the weighted average shrinks: we place more weight on the smaller term and less weight on the larger one. Using $\Pr_{\mathrm{C}}{}'$ to denote consideration probabilities with p_j replaced by p'_j, and making use of (18) and (19), we then know:

$$\begin{aligned}
&\Pr_{\mathrm{PL+C}}(\mathcal{R}_{i=1}) \\
&\geq \sum_{\substack{C \in \mathcal{C}_{i\,\cap\,j} \\ |C| > k}} [\Pr_{\mathrm{C}}(C) \cdot \Pr(i \mid C) + \Pr_{\mathrm{C}}(\phi(C)) \cdot \Pr(i \mid \phi(C))] \\
&\geq \sum_{\substack{C \in \mathcal{C}_{i\,\cap\,j} \\ |C| > k}} \left[\Pr_{\mathrm{C}}{}'(C) \cdot \Pr(i \mid C) + \Pr_{\mathrm{C}}{}'(\phi(C)) \cdot \Pr(i \mid \phi(C))\right] \\
&= \sum_{\substack{C \subseteq \mathcal{U} \\ |C \cup \{j\}| > k \\ i \in C}} \Pr_{\mathrm{C}}{}'(C) \cdot \Pr(i \mid C). \qquad (23)
\end{aligned}$$

Finally, we add the additive error term to recover the consideration sets missing from (23)—those containing j and with size exactly k—and arrive at $\Pr_{\mathrm{PL+C}}{}'(\mathcal{R}_{i=1})$. Recall that $\sum_{i \in \mathcal{U}} p_i \geq \alpha k$; this inequality is still true after replacing p_j with $p'_j > p_j$, so by Lemma 6.1

$$\sum_{C \subseteq \mathcal{U}, |C| = k} \Pr_{\mathrm{C}}{}'(C) \leq \frac{\left(\alpha e^{1-\alpha}\right)^k}{1 - \left(\alpha e^{1-\alpha}\right)^k}. \qquad (24)$$

Thus, by (23) and (24), we have

$$\begin{aligned}
&\Pr_{\mathrm{PL+C}}{}'(\mathcal{R}_{i=1}) \\
&= \sum_{\substack{C\subseteq\mathcal{U}\\ |C\cup\{j\}|>k\\ i\in C}} \Pr_{\mathrm{C}}{}'(C)\cdot\Pr(i\mid C) + \sum_{\substack{C\subseteq\mathcal{U}\\ |C|=k\\ j\in C}} \Pr_{\mathrm{C}}{}'(C)\cdot\Pr(i\mid C) \\
&\le \sum_{\substack{C\subseteq\mathcal{U}\\ |C\cup\{j\}|>k\\ i\in C}} \Pr_{\mathrm{C}}{}'(C)\cdot\Pr(i\mid C) + \sum_{\substack{C\subseteq\mathcal{U}\\ |C|=k}} \Pr_{\mathrm{C}}{}'(C) \\
&\le \sum_{\substack{C\subseteq\mathcal{U}\\ |C\cup\{j\}|>k\\ i\in C}} \Pr_{\mathrm{C}}{}'(C)\cdot\Pr(i\mid C) + \tfrac{\left(\alpha e^{1-\alpha}\right)^k}{1-\left(\alpha e^{1-\alpha}\right)^k} \\
&\le \Pr_{\mathrm{PL+C}}(\mathcal{R}_{i=1}) + \tfrac{\left(\alpha e^{1-\alpha}\right)^k}{1-\left(\alpha e^{1-\alpha}\right)^k},
\end{aligned}$$

which is exactly the desired inequality. Finally, if at least $k+1$ of the consideration probabilities (including p_j') are 1, then the sum in (24) equals 0, since we can never have a size-k consideration set. The same argument as above—identical up to (23), and with the tightened (24)—then yields the inequality with no error term. ☐

Theorem 6.3. *Suppose that $\sum_{i\in\mathcal{U}} p_i \ge \alpha k$ for some $\alpha > 1$. Then for any $i \in \mathcal{U}$,*

$$p_i \le \frac{\sum_{j\in\mathcal{U}} \exp(u_j)}{\exp(u_i)} \cdot \left(\Pr_{\mathrm{PL+C}}(\mathcal{R}_{i=1}) + \frac{k\left(\alpha e^{1-\alpha}\right)^k}{1-\left(\alpha e^{1-\alpha}\right)^k}\right).$$

Proof. We'll apply Lemma 6.2 to $\Pr_{\mathrm{PL+C}}(\mathcal{R}_{i=1})$ $n-1$ times, each time increasing p_j to 1 (for every $j \ne i$). Note that since we only increase consideration probabilities, αk remains a valid lower bound on $\sum_{i\in\mathcal{U}} p_i$. The first k times we do this, we incur the additive error from Lemma 6.2. After k consideration probabilities have been increased to 1, we no longer incur additional error from subsequent increases. Thus, using $\Pr_{\mathrm{PL+C}}{}'(\mathcal{R}_{i=1})$ to denote the probability that i is ranked first if all consideration probabilities except p_i are 1,

$$\Pr_{\mathrm{PL+C}}{}'(\mathcal{R}_{i=1}) \le \Pr_{\mathrm{PL+C}}(\mathcal{R}_{i=1}) + k\tfrac{\left(\alpha e^{1-\alpha}\right)^k}{1-\left(\alpha e^{1-\alpha}\right)^k}. \tag{25}$$

Now, we can simplify $\Pr_{\mathrm{PL+C}}{}'(\mathcal{R}_{i=1})$, since $\mathcal{U}$ (occurring w.p. p_i) is the only feasible consideration set that includes i, since all other consideration probabilities equal 1:

$$\begin{aligned}
\Pr_{\mathrm{PL+C}}{}'(\mathcal{R}_{i=1}) = \sum_{r\in\mathcal{R}_{i=1}} p_i\cdot \Pr_{\mathrm{PL}}(r\mid\mathcal{U}) &= p_i \sum_{r\in\mathcal{R}_{i=1}} \Pr_{\mathrm{PL}}(r\mid\mathcal{U}) \\
&= p_i\cdot\frac{\exp(u_i)}{\sum_{j\in\mathcal{U}}\exp(u_j)}.
\end{aligned}$$

We can combine this with (25) and solve for p_i:

$$\begin{aligned}
&p_i\cdot\frac{\exp(u_i)}{\sum_{j\in\mathcal{U}}\exp(u_j)} \le \Pr_{\mathrm{PL+C}}(\mathcal{R}_{i=1}) + k\tfrac{\left(\alpha e^{1-\alpha}\right)^k}{1-\left(\alpha e^{1-\alpha}\right)^k} \\
\Rightarrow\ &p_i \le \frac{\sum_{j\in\mathcal{U}}\exp(u_j)}{\exp(u_i)}\cdot\left(\Pr_{\mathrm{PL+C}}(\mathcal{R}_{i=1}) + k\tfrac{\left(\alpha e^{1-\alpha}\right)^k}{1-\left(\alpha e^{1-\alpha}\right)^k}\right).
\end{aligned}$$

☐

A.1 Algorithms for Lower and Upper Bounds

Here is the deferred proof of Theorem 5.3, and the omitted upper-bound algorithm (and its proof of correctness).

Theorem 5.3. *Algorithm 1 returns lower bounds b_i for each $i \in \mathcal{U}$ such that $p_i \ge b_i$ in time $O(kn^2)$.*

Algorithm 2 Upper-bounding consideration probabilities.

Require: items $\mathcal{U}$, (estimates of) utilities u_i, (estimates of) top-ℓ ranking probabilities $\Pr_{\mathrm{PL+C}}(\mathcal{R}_{i\le\ell})$, lower bound αk on expected consideration set size ($\alpha > 1$)

$b_i \leftarrow \frac{\sum_{j\in\mathcal{U}} \exp(u_j)}{\exp(u_i)} \cdot \left(\Pr_{\mathrm{PL+C}}(\mathcal{R}_{i=1}) + \frac{k\left(\alpha e^{1-\alpha}\right)^k}{1-(\alpha e^{1-\alpha})^k}\right)$ for all $i \in \mathcal{U}$
$E \leftarrow \{\langle j, i\rangle \subset \mathcal{U} \times \mathcal{U} \mid u_i > u_j \text{ and } \exists \ell : \Pr_{\mathrm{PL+C}}(\mathcal{R}_{i\le\ell}) < \Pr_{\mathrm{PL+C}}(\mathcal{R}_{j\le\ell})\}$
$G \leftarrow \langle \mathcal{U}, E\rangle$
for all j in a topological sort of G **do**
  for all out-neighbors i of j in G **do**
    for $\ell = 1, \ldots, k$ **do**
      $c \leftarrow \Pr_{\mathrm{PL+C}}(\mathcal{R}_{i\le\ell}) / \Pr_{\mathrm{PL+C}}(\mathcal{R}_{j\le\ell})$
      if $c \le 1$ **then**
        $b_i \leftarrow \min\left\{b_i, \frac{cb_j}{1-b_j+cb_j}\right\}$
return b_i for each $i \in \mathcal{U}$ such that $p_i \le b_i$

Proof. We begin with correctness. We'll show that $b_i \le p_i$ is an invariant of the nested for loop for every $i \in \mathcal{U}$. First, the initial values of b_i satisfy $b_i \le p_i$ by Theorem 5.2. Now, suppose inductively that $b_j \le p_j$ for all $j \in \mathcal{U}$ before an iteration of the innermost loop. We only update b_j if $c = \Pr_{\mathrm{PL+C}}(\mathcal{R}_{i\le\ell})/\Pr_{\mathrm{PL+C}}(\mathcal{R}_{j\le\ell})$. Additionally, j must be an outneighbor of i in G. By definition of G, we then have $u_i > u_j$. Thus, in any case where b_j is updated, we know from Theorem 4.4 that $\frac{p_i}{1-p_i} \le c \cdot \frac{p_j}{1-p_j}$. We can rewrite this as a lower bound on p_j, as in (6):

$$p_j \ge \frac{p_i}{c - cp_i + p_i}.$$

The right side of this lower bound is an increasing function of p_i for $p_i, c \in (0, 1]$, so we can replace p_i with a smaller value (namely, b_i, which is smaller by inductive hypothesis) and maintain the inequality. Thus, we know $p_j \ge \frac{b_i}{c-cb_i+b_i}$. Therefore, if we update b_j to be equal to this lower bound in the loop, we still maintain the invariant $b_j \le p_j$. By induction, we then have $b_i \le p_i$ for all $i \in \mathcal{U}$ when the algorithm terminates.

Concerning the runtime, initializing the lower bounds takes time $O(n)$ and constructing G takes times $O(kn^2)$ (by looping through all pairs of items and all k values of ℓ). Additionally, topologically sorting G takes time $O(n^2)$ and the nested loop takes time $O(kn^2)$, for a total running time of $O(kn^2)$. □

Theorem A.1. *Suppose $\sum_{i\in\mathcal{U}} p_i \ge \alpha k$. Algorithm 2 returns upper bounds b_i for each $i \in \mathcal{U}$ such that $p_i \le b_i$ in time $O(kn^2)$.*

Proof. The analysis is nearly identical to Algorithm 1, so we only highlight the differences. First, the initial values of b_i are valid upper bounds by Theorem 6.3. Then, we use Theorem 4.4 in its upper bound form (5): $p_i \le \frac{cp_j}{1-p_j+cp_j}$. This bound is an increasing function of p_j, so we can replace p_j by a larger quantity (namely, an upper bound $b_j \ge p_j$) and maintain the inequality. Whenever we update b_i, the conditions of Theorem 4.4 are satisfied, so b_i remains a valid upper bound on p_i. Correctness then follows by induction and the runtime analysis is the same as for Algorithm 1. □

A.2 Algorithms for approximating PL+C probabilities

Theorem 8.1. *There is a randomized ϵ-additive approximation algorithm for $\Pr_{\mathrm{PL+C}}(r)$ with runtime $O(kn \log(1/\delta)/(z_{k,p}\epsilon^2))$, where both the approximation and runtime bounds hold with probability at least $1 - \delta$.*

Proof. Let R be the random variable that takes value $\Pr_{\mathrm{PL}}(r|C)$ with probability $\Pr_{\mathrm{C}}(C)$. It follows by (3) that $\Pr_{\mathrm{PL+C}}(r) = \mathrm{E}[R]$. Let $\hat{R} = \frac{1}{s}\sum_{i=1}^{s} R$ be the empirical mean of s i.i.d. samples from R. As $R \in [0, 1]$ since it outputs probabilities, by Hoeffding's inequality (Mitzenmacher and Upfal, 2017), we get

$$\Pr(|\hat{R} - \Pr_{\mathrm{PL+C}}(r)| \ge \epsilon) \le 2\exp(-2s\epsilon^2).$$

Setting $\delta/2 = 2\exp(-2s\epsilon^2)$ and solving for s, we find that we need $s = \log(4/\delta)/(2\epsilon^2) = O(\log(1/\delta)/\epsilon^2)$ samples from R to estimate $\Pr_{\mathrm{PL+C}}(r)$ with ϵ-additive error and failure probability at most $\delta/2$.

Algorithm 3 Probably approximately computing PL+C probabilities.

Require: ranking r, utilities $u_i > 0$ and consideration probabilities p_i for all $i \in \mathcal{U}$, error tolerance ϵ, failure probability δ

$\hat{R} \leftarrow 0$
$s \leftarrow \lceil \log(4/\delta)/(2\epsilon^2) \rceil$
for s iterations **do**
  $C \leftarrow \emptyset$
  while $|C| < k$ **do**
    sample consideration set C according to consideration probabilities p_i
  $\hat{R} \leftarrow \hat{R} + \Pr_{\mathrm{PL}}(r|C)$
$\hat{R} \leftarrow \hat{R}/s$
return $\hat{R}$

In order to sample from R, Algorithm 3 uses rejection sampling to discard cases where $|C| < k$. We can bound the number of consideration sets T we need to sample to get s consideration sets of size at least k—and therefore s useful samples from R. The probability of drawing a sufficiently large consideration set is exactly $z_{k,p}$, so T is distributed according to a negative binomial distribution (counting the total number of trials, not just failures), namely $T \sim \mathrm{NB}(s, z_{k,p})$, which has expectation $s/z_{k,p}$. Using a tail bound for this type of negative binomial distribution (Brown, 2011), we have that for any $c > 1$,

$$\Pr(T > c\,\mathrm{E}[T]) \leq \exp(-cs(1 - 1/c)^2/2).$$

We want to find a c such that this failure probability is at most $\delta/2$, since then if we run $cs/z_{k,p}$ times, we will only fail to get s successes with probability at most $\delta/2$. Setting $\delta/2 \geq \exp(-cs(1 - 1/c)^2/2)$ and solving for c, we find that we need

$$c + 1/c \geq \frac{2\log(2/\delta)}{s} + 2.$$

Thus $c = \frac{2\log(2/\delta)}{s} + 2$ suffices. That is, if we sample $\left(\frac{2\log(2/\delta)}{s} + 2\right) s/z_{k,p} = O((\log(1/\delta) + s)/z_{k,p})$ sets, then we will have at least s sufficiently large sets with probability at least $1 - \delta/2$. We can plug in our previously found value of s to find that the total number of iterations we need is $O((\log(1/\delta) + \log(1/\delta)/\epsilon^2)/z_{k,p}) = O(\log(1/\delta)/(z_{k,p}\epsilon^2))$.

By the union bound, one or both of the approximation and runtime guarantees fail with probability at most δ (either one fails w.p. at most $\delta/2$), so our overall success probability is at least $1 - \delta$.

Lastly, it takes $O(n)$ time to sample one consideration set, and $O(kn)$ time to compute $\Pr_{\mathrm{PL}}(r|C)$ from one consideration set C, so the overall runtime (with probability at least $1 - \delta$) is $O(kn\log(1/\delta)/(z_{k,p}\epsilon^2))$. □

Next, we provide our deterministic algorithm for approximating PL+C probabilities. The algorithm clusters possible consideration sets into bins according to their total exp-utility and stores the total consideration probability of the sets in each bin (without storing the sets themselves). These sets are built up item-by-item, each time adding the consideration probability of new sets to the appropriate bin. At the end of the algorithm, we know the approximate exp-utility of sets in each bin and their exact consideration probabilities. This allows us to estimate the PL+C probability of a ranking r. Since we know all items in r must have been considered, the algorithm only builds up consideration sets over the items $\overline{r} = \mathcal{U} \setminus r$. Finally, we note that the normalizing constant $z_{k,p}$ can actually be computed efficiently. The DC ("direct computation") algorithm (Biscarri et al., 2018) we use as a subroutine in Algorithm 4 computes the full PMF of the Poisson binomial distribution in time $O(n^2)$, which we can sum over to get $z_{k,p}$.

Lemma A.2. *After processing items $\{1, \dots, i\} \subseteq \overline{r}$ in the outer for loop, there is a partition $\{C_{i,-1}, \dots, C_{i,s}\}$ of $\mathcal{P}(\{1, \dots, i\})$ such that:*

1. $C_{i,-1} = \{\emptyset\}$.
2. *For all* $-1 \leq j \leq s$, $A[j][0] = \sum_{C \in C_{i,j}} \left(\prod_{a \in C} p_a\right) \left(\prod_{b \in \{1,\dots,i\} \setminus C} (1 - p_b)\right)$.

Algorithm 4 Approximately computing PL+C probabilities.

Require: ranking r, utilities $u_i > 0$ and consideration probs. p_i for all $i \in \mathcal{U}$, error tolerance ϵ

1: **if** $k = n$ **then**
2:     **return** $\Pr_{\text{PL}}(r \mid \mathcal{U})$
3: $\delta \leftarrow \epsilon/(2kn)$
4: $\overline{r} \leftarrow \mathcal{U} \setminus r$
5: $s \leftarrow \lfloor \log_{1+\delta} \left(\sum_{i \in \overline{r}} e^{u_i} \right) \rfloor$
6: $A \leftarrow$ array of size $s+1$ indexed $-1, 0, \ldots, s$
7: $A[-1] \leftarrow (1, 0)$
8: $A[0, \ldots, s] \leftarrow (0, 0)$
9: $A' \leftarrow A$
10: **for all** $i \in \overline{r}$ **do**
11:     $A'[j] \leftarrow (A[j][0] \cdot (1 - p_i), A[j][1])$ (for every $j = -1, 0, \ldots, s$)
12:     **for** $j = -1, \ldots, s$ **do**
13:         **if** $A[j][0] > 0$ **then**
14:             $h_j \leftarrow \lfloor \log_{1+\delta}(A[j][1] + e^{u_i}) \rfloor$
15:             $A'[h_j] \leftarrow (A'[h_j][0] + A[j][0] \cdot p_i, \max\{A'[h_j][1], A[j][1] + e^{u_i}\})$
16:     $A \leftarrow A'$
17: $z_{k,p} \leftarrow \sum_{i=k}^{n} \mathsf{DC}(p_1, \ldots, p_n)[i]$
18: **return** $\frac{\prod_{i \in r} p_i}{z_{k,p}} \sum_{j=-1}^{s} A[j][0] \cdot \prod_{i=1}^{k} \frac{e^{u_i}}{(1+\delta)^j + \sum_{a \in \{r_i, \ldots, r_k\}} e^{u_a}}$

3. *For every* $0 \leq j \leq s$, *every* $C \in C_{i,j}$ *satisfies* $(1+\delta)^{j-i} \leq \sum_{a \in C} e^{u_a} < (1+\delta)^{j+1}$.

4. $A[j][1] = \max_{C \in C_{i,j}} \sum_{a \in C} e^{u_a}$.

Proof. By induction on i.

Base case: $i = 0$. Before processing any items, define the partition to be $C_{0,-1} = \{\emptyset\}$ and every other $C_{0,j} = \emptyset$, satisfying claim 1. By lines 7 and 8, we have $A[-1][0] = 1$ and every other $A[j][0] = 0$, satisfying claim 1. Claim 3 is vacuously true, since all the $C_{0,j}$ with $0 \leq j \leq s$ are empty. Finally, claim 4 also holds since all the $A[j][1] = 0$ (either $C = \emptyset$ or $C_{0,j} = \emptyset$).

Inductive case: $i \geq 1$. By inductive hypothesis, the $C_{i-1,j}$ form a partition of $\mathcal{P}(\{1, \ldots, i-1\})$. Thus, every set in $\mathcal{P}(\{1, \ldots, i\})$ is either in some $C_{i-1,j}$ or is obtained by adding item i to a set in some $C_{i-1,j}$. To make the new partition, we'll start with the old one and, for each $C_{i-1,j}$, add all of its sets plus i to the collection C_{i,h_j}, with h_j defined in line 14. We can verify that h_j is always a valid index:

- Since the utilities are all positive and $A[j][1]$ is non-negative, $\lfloor \log_{1+\delta}(A[j][1] + e^{u_i}) \rfloor \geq 0$.
- Since $A[j][1]$ is the sum of exp-utilities of some set C in $C_{i-1,j}$ by inductive hypothesis (of claim 4), adding a new exp-utility cannot cause h to exceed s, which is defined using the sum of all item exp-utilities.

Formally, the new partition $C_{i,j}$ can be written as follows:

$$C_{i,j} = C_{i-1,j} \cup \{C \cup \{i\}, \text{ for all } C \in C_{i-1,\ell}, \text{ for all } \ell \text{ s.t. } h_\ell = j\}.$$

In this new partition, claim 1 still holds: since $h_j \geq 0$, we leave $C_{i,-1}$ unchanged. We can show claim 2 still holds as well. Just before executing line 16, lines 11 and 15 have resulted in:

$$A'[j][0] = A[j][0] \cdot (1 - p_i) + \sum_{\ell : h_\ell = j, A[\ell][0] > 0} A[\ell][0] \cdot p_i.$$

(Note that the condition $A[\ell][0] > 0$ prevents us from adding i to empty collections of sets.) By inductive hypothesis and our construction of $C_{i,j}$,

$$\begin{aligned}
A'[j][0] = &\sum_{C \in C_{i-1,j}} \left(\prod_{a \in C} p_a\right) \left(\prod_{b \in \{1,\ldots,i-1\} \setminus C} (1-p_b)\right) \cdot (1-p_i) \\
&+ \sum_{\substack{\ell: h_\ell = j, \\ A[\ell][0] > 0}} \sum_{C \in C_{i-1,\ell}} \left(\prod_{a \in C} p_a\right) \left(\prod_{b \in \{1,\ldots,i-1\} \setminus C} (1-p_b)\right) \cdot p_i \\
= &\sum_{C \in C_{i,j}, i \notin C} \left(\prod_{a \in C} p_a\right) \left(\prod_{b \in \{1,\ldots,i\} \setminus C} (1-p_b)\right) \\
&+ \sum_{C \in C_{i,j}, i \in C} \left(\prod_{a \in C} p_a\right) \left(\prod_{b \in \{1,\ldots,i\} \setminus C} (1-p_b)\right) \\
= &\sum_{C \in C_{i,j}} \left(\prod_{a \in C} p_a\right) \left(\prod_{b \in \{1,\ldots,i\} \setminus C} (1-p_b)\right).
\end{aligned}$$

Thus claim 2 is satisfied at the end of the for loop iteration. Claim 4 follows from our construction of $C_{i,j}$, inductive hypothesis, and lines 11 and 15: whenever we add new sets to $C_{i,j}$ with higher exp-utility, we also update $A[j][1]$ to be the larger of the new and old exp-utility sums. Finally, we show claim 3.

Any set $C \in C_{i,j}$ was either (a) in $C_{i-1,j}$ or (b) was just added by including i. For carried-over sets, we have $(1+\delta)^{j-(i-1)} \leq \sum_{a \in C} e^{u_a} < (1+\delta)^{j+1}$ by inductive hypothesis, so the claim still holds with i. For newly added sets, we know they are of the form $C = C' \cup \{i\}$, where $C' \in C_{i-1,\ell}$ such that $h_\ell = j$.

If $C' = \emptyset$, then $j = h_\ell = \lfloor \log_{1+\delta}(e^{u_i}) \rfloor$. Thus $\log_{1+\delta} e^{u_i} - 1 < j$ and $j \leq \log_{1+\delta} e^{u_i}$. We can rewrite these inequalities as $e^{u_i} < (1+\delta)^{j+1}$ and $(1+\delta)^j \leq e^{u_i}$. Since $i \geq 0$, this shows claim 3 is satisfied.

Now suppose $C' \neq \emptyset$. By inductive hypothesis (claim 3), we then have:

$$(1+\delta)^{\ell-(i-1)} \leq \sum_{a \in C'} e^{u_a} < (1+\delta)^{\ell+1}$$

We also know by inductive hypothesis (claim 4) that $A[\ell][1] = \max_{C \in C_{i-1,j}} \sum_{a \in C} e^{u_a}$. C' is one such C, so we have:

$$(1+\delta)^{\ell-(i-1)} \leq \sum_{a \in C'} e^{u_a} \leq A[\ell][1] < (1+\delta)^{\ell+1}$$

Using these bounds, we can then bound how much larger $A[\ell][1]$ is than the exp-utility sum of C':

$$\begin{aligned}
&\frac{A[\ell][1]}{\sum_{a \in C'} e^{u_a}} < \frac{(1+\delta)^{\ell+1}}{(1+\delta)^{\ell-i+1}} = (1+\delta)^i \\
\Rightarrow \quad &A[\ell][1] < (1+\delta)^i \sum_{a \in C'} e^{u_a}.
\end{aligned}$$

We will now use this fact to establish the bounds in claim 3. For the lower bound:

$$
\begin{aligned}
& j = h_\ell = \lfloor \log_{1+\delta}(A[\ell][1] + e^{u_i}) \rfloor \\
\Rightarrow\quad & j \le \log_{1+\delta}((1+\delta)^i \sum_{a \in C'} e^{u_a} + e^{u_i}) && \text{(IH, claim 4)} \\
\Rightarrow\quad & (1+\delta)^j \le (1+\delta)^i \sum_{a \in C'} e^{u_a} + e^{u_i} && \text{(exponentiate)} \\
\Rightarrow\quad & (1+\delta)^j \le (1+\delta)^i \sum_{a \in C'} e^{u_a} + (1+\delta)^i e^{u_i} && \text{(increase RHS)} \\
\Rightarrow\quad & (1+\delta)^j \le (1+\delta)^i \sum_{a \in C} e^{u_a} && \text{(fold into sum)} \\
\Rightarrow\quad & (1+\delta)^{j-i} \le \sum_{a \in C} e^{u_a}.
\end{aligned}
$$

For the upper bound:

$$
\begin{aligned}
& j = h_\ell = \lfloor \log_{1+\delta}(A[\ell][1] + e^{u_i}) \rfloor \\
\Rightarrow\quad & j + 1 > \log_{1+\delta}(A[\ell][1] + e^{u_i}) \\
\Rightarrow\quad & (1+\delta)^{j+1} > A[\ell][1] + e^{u_i} \ge \sum_{a \in C'} e^{u_a} + e^{u_i} = \sum_{a \in C} e^{u_a}.
\end{aligned}
$$

Thus all of the claims still hold after iteration i. □

We can now prove Algorithm 4 is correct.

Theorem 8.2. *There is a deterministic algorithm approximating* $\Pr_{\mathrm{PL+C}}(r)$ *with multiplicative error at most* $1 + \epsilon$ *and runtime* $O(kn^2 \max\{\log n, m\}/\epsilon)$*, where* $m = \max_{i \in \overline{r}} u_i$*.*

Proof. By Lemma A.2, at the end of the nested loops of Algorithm 4, there is a partition $\{C_{-1}, \ldots, C_s\}$ of the subsets of $\overline{r}$ such that $A[j][0]$ stores the total consideration probabilities of the sets in C_j and such that the exp-utility sums of every set in C_j is between $(1+\delta)^{j-n}$ and $(1+\delta)^{j+1}$ (since $|\overline{r}| \le n$). Thus,

$$
\begin{aligned}
& \frac{\prod_{i \in r} p_i}{z_{k,p}} \sum_{j=-1}^{s} A[j] \prod_{i=1}^{k} \frac{e^{u_i}}{(1+\delta)^j + \sum_{a \in \{r_i, \ldots, r_k\}} e^{u_a}} \\
&= \frac{\prod_{i \in r} p_i}{z_{k,p}} \sum_{j=-1}^{s} \sum_{C \in C_j} \left(\prod_{a \in C} p_a \right) \left(\prod_{b \in \overline{r} \setminus C} (1 - p_b) \right) \prod_{i=1}^{k} \frac{e^{u_i}}{(1+\delta)^j + \sum_{a \in \{r_i, \ldots, r_k\}} e^{u_a}} \\
&= \sum_{j=-1}^{s} \sum_{C \in C_j} \frac{1}{z_{k,p}} \left(\prod_{a \in C \cup r} p_a \right) \left(\prod_{b \in \overline{r} \setminus C} (1 - p_b) \right) \prod_{i=1}^{k} \frac{e^{u_i}}{(1+\delta)^j + \sum_{a \in \{r_i, \ldots, r_k\}} e^{u_a}} \\
&= \sum_{j=-1}^{s} \sum_{C \in C_j} \Pr_{\mathrm{C}}(C \cup r) \prod_{i=1}^{k} \frac{e^{u_i}}{(1+\delta)^j + \sum_{a \in \{r_i, \ldots, r_k\}} e^{u_a}}.
\end{aligned}
$$

Note that the product is very close to the desired Plackett–Luce probability:

$$
\begin{aligned}
\Pr_{\mathrm{PL}}(r \mid C \cup r) &= \prod_{i=1}^{k} \frac{e^{u_i}}{\sum_{a \in C \cup \{r_i, \ldots, r_k\}} e^{u_a}} \\
&= \prod_{i=1}^{k} \frac{e^{u_i}}{\sum_{a \in C} e^{u_a} + \sum_{a \in \{r_i, \ldots, r_k\}} e^{u_a}}
\end{aligned}
$$

We'll show that the error introduced by estimating the exp-utility of C by $(1+\delta)^j$ is not too large. Consider the size of the maximum possible underestimation factor (when $\sum_{a\in C} e^{u_a}$ achieves its upper bound $(1+\delta)^{j+1}$). Let $e_i = \sum_{a\in\{r_i,\dots,r_k\}} e^{u_a}$

$$\begin{aligned}\frac{e^{u_i}}{(1+\delta)^j + e_i} \Big/ \frac{e^{u_i}}{(1+\delta)^{j+1} + e_i} &= \frac{(1+\delta)^{j+1} + e_i}{(1+\delta)^j + e_i} \\ &\le \frac{(1+\delta)^{j+1}}{(1+\delta)^j} \\ &= (1+\delta).\end{aligned}$$

Now consider the maximum possible overestimation factor (when $\sum_{a\in C} e^{u_a}$ achieves its lower bound $(1+\delta)^{j-n}$):

$$\begin{aligned}\frac{e^{u_i}}{(1+\delta)^{j-n} + e_i} \Big/ \frac{e^{u_i}}{(1+\delta)^{j} + e_i} &= \frac{(1+\delta)^{j} + e_i}{(1+\delta)^{j-n} + e_i} \\ &\le \frac{(1+\delta)^{j}}{(1+\delta)^{j-n}} \\ &= (1+\delta)^n.\end{aligned}$$

Thus, combining the largest possible under- and overestimation errors for the PL probability (incurred k times, once for each term of the product),

$$\begin{aligned}&\sum_{j=-1}^{s} \sum_{C\in C_j} \Pr_{\mathrm{C}}(C\cup r)\Pr_{\mathrm{PL}}(r \mid r\cup C)(1+\delta)^{-k} \\ &\le \sum_{j=-1}^{s} \sum_{C\in C_j} \Pr_{\mathrm{C}}(C\cup r) \prod_{i=1}^{k} \frac{e^{u_i}}{(1+\delta)^j + \sum_{a\in\{r_i,\dots,r_k\}} e^{u_a}} \\ &\le \sum_{j=-1}^{s} \sum_{C\in C_j} \Pr_{\mathrm{C}}(C\cup r)\Pr_{\mathrm{PL}}(r \mid r\cup C)(1+\delta)^{nk}.\end{aligned}$$

Rewriting the sums, since we know the C_j partition the subsets of $\overline{r}$,

$$\begin{aligned}&(1+\delta)^{-k}\cdot \Pr_{\mathrm{PL+C}}(r) = \sum_{C\subseteq\mathcal{U}} \Pr_{\mathrm{C}}(C)\Pr_{\mathrm{PL}}(r \mid C)(1+\delta)^{-k} \\ &\le \sum_{j=-1}^{s} \sum_{C\in C_j} \Pr_{\mathrm{C}}(C\cup r) \prod_{i=1}^{k} \frac{e^{u_i}}{(1+\delta)^j + \sum_{a\in\{r_i,\dots,r_k\}} e^{u_a}} \\ &\le \sum_{C\subseteq\mathcal{U}} \Pr_{\mathrm{C}}(C)\Pr_{\mathrm{PL}}(r \mid C)\Pr_{\mathrm{PL}}(r \mid r\cup C)(1+\delta)^{nk} = (1+\delta)^{nk}\cdot \Pr_{\mathrm{PL+C}}(r).\end{aligned}$$

Since $e^x \ge (1+x/n)^n$ for $x \le n$, we have that $(1+\delta)^{nk} = (1+\epsilon/(2nk))^{nk} \le e^{\epsilon/2} \le 1+\epsilon$ (the final inequality follows from the identity $e^x \le 1+x+x^2$ for $x\le 1$). We therefore return an estimate $\hat{p}$ satisfying:

$$\frac{\Pr_{\mathrm{PL+C}}(r)}{1+\epsilon} \le \hat{p} \le (1+\epsilon)\cdot \Pr_{\mathrm{PL+C}}(r).$$

Finally, we establish the runtime of the algorithm. The outer loop runs $n-k = O(n)$ times. Line 11 takes $O(s)$ time, the inner loop runs $O(s)$ times, and lines 13-15 take constant time per iteration. Thus the full nested loop takes $O(ns)$ time. The DC algorithm takes time $O(n^2)$. Line 18 takes $O(ns)$ time, for a total runtime of

$O(n^2 + ns)$. Now we can bound s :

$$
\begin{aligned}
s &= \left\lfloor \log_{1+\delta} \sum_{i \in \overline{r}} e^{u_i} \right\rfloor \\
&\leq \frac{\ln \sum_{i \in \overline{r}} e^{u_i}}{\ln(1+\delta)} \\
&\leq (1+\delta) \frac{\ln \sum_{i \in \overline{r}} e^{u_i}}{\delta} && \text{(since } \ln(1+x) \geq \tfrac{x}{1+x} \text{ for } x > -1\text{)} \\
&= \frac{(2kn) \ln \sum_{i \in \overline{r}} e^{u_i}}{\epsilon} + \ln \sum_{i \in \overline{r}} e^{u_i} && \text{(since } \delta = \epsilon/(2kn)\text{)} \\
&\leq \frac{(2kn)(\max_{i \in \overline{r}} u_i + \ln n)}{\epsilon} + (\max_{i \in \overline{r}} u_i + \ln n) && \text{(property of LogSumExp)}
\end{aligned}
$$

Thus $s = O(kn \max\{m, \log n\}/\epsilon)$ and the claimed runtime follows. □

B ADDITIONAL FIGURES

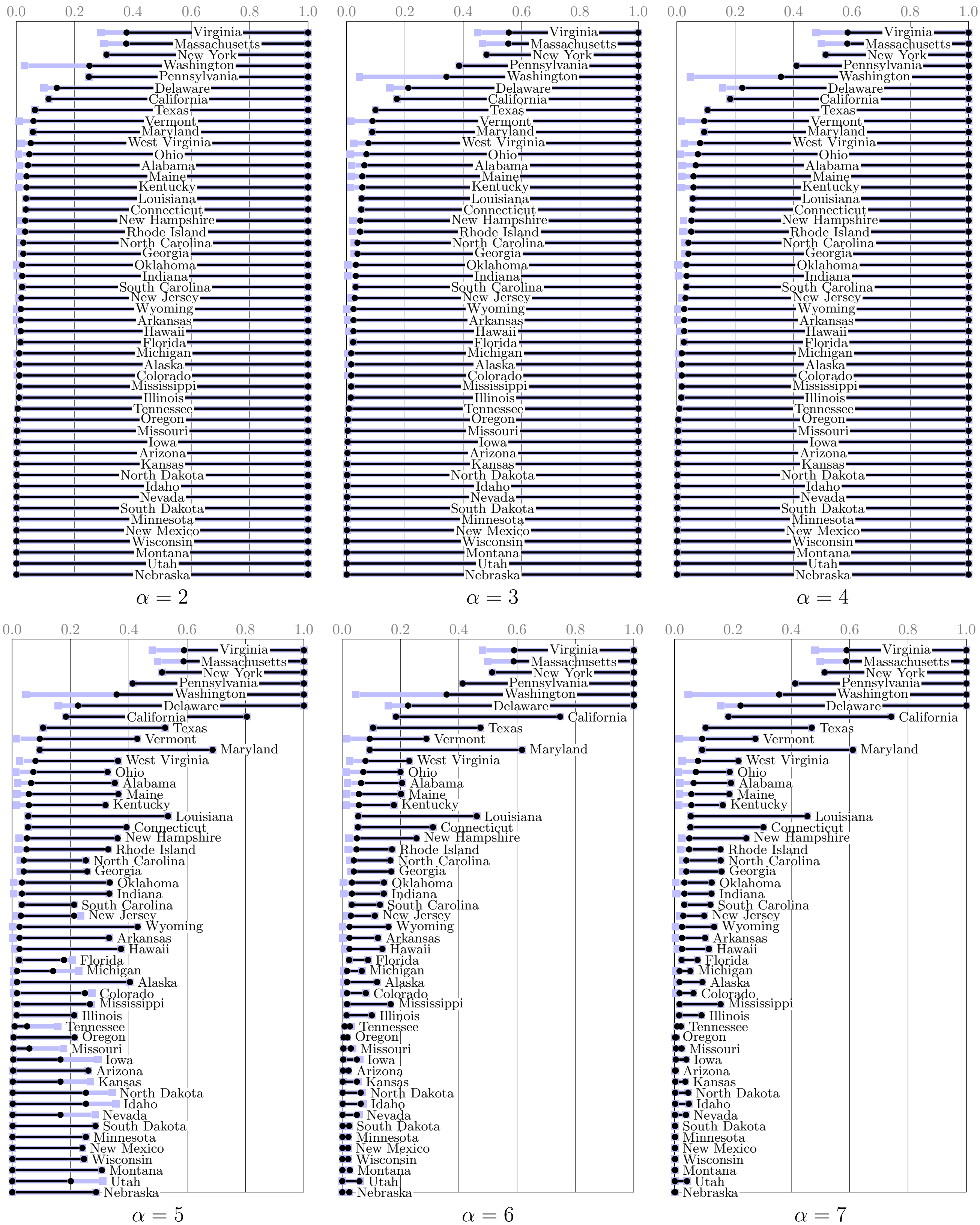

Figure 2: Versions of Figure 1 (right) with α varying from 2 to 7, showing how our bounds change as the lower bound on average number of states considered grows from $\alpha k = 6$ to $\alpha k = 21$.