
Learning Rich Rankings

Arjun Seshadri
Stanford University
aseshadr@stanford.edu

Stephen Ragain
Stanford University
sragain17@gmail.com

Johan Ugander
Stanford University
jugander@stanford.edu

Abstract

Although the foundations of ranking are well established, the ranking literature has primarily been focused on simple, unimodal models, e.g. the Mallows and Plackett-Luce models, that define distributions centered around a single total ordering. Explicit mixture models have provided some tools for modelling multimodal ranking data, though learning such models from data is often difficult. In this work, we contribute a *contextual repeated selection* (CRS) model that leverages recent advances in choice modeling to bring a natural multimodality and richness to the rankings space. We provide rigorous theoretical guarantees for maximum likelihood estimation under the model through structure-dependent tail risk and expected risk bounds. As a by-product, we also furnish the first tight bounds on the expected risk of maximum likelihood estimators for the multinomial logit (MNL) choice model and the Plackett-Luce (PL) ranking model, as well as the first tail risk bound on the PL ranking model. The CRS model significantly outperforms existing methods for modeling real world ranking data in a variety of settings, from racing to rank choice voting.

1 Introduction

Ranking data is one of the fundamental primitives of statistics, central to the study of recommender systems, search engines, social choice, as well as general data collection across machine learning. The combinatorial nature of ranking data comes with inherent computational and statistical challenges [15], and distributions over the space of rankings (the symmetric group S_n) are very high dimensional objects that are quickly intractable to represent with complete generality. As a result, popular models of ranking data focus on parametric families of distributions in S_n , anchoring the computational and statistical burden of the model to the parameters.

Most popular models of rankings are distance-based or utility-based, where the Mallows [33] and Plackett-Luce [43] models are the two most popular models in each respective category. Both of these models simplistically assume transitivity and center a distribution around a single total ordering, assumptions that are limiting in practice. Intransitivities are frequent in sports competitions and other matchups [12]. The presence of political factions render unimodality an invalid assumption in ranked surveys and ranked voting, and recommender systems audiences often contain subpopulations with significant differences in preferences [26] that also induce multimodal ranking distributions.

A major open challenge in the ranking literature, then, has been to develop rich ranking models that go beyond these assumptions while still being efficiently learnable from data. Work on escaping unimodality is not new—the ranking literature has long considered models that violate unimodality (e.g., Babington Smith [50]), including explicit mixtures of unimodal models [22, 39]. However, such proposals are almost always restricted to theoretical discussions without practical merit.

In Figure 1 we provide a stylized visualization of multimodal data and models on the canonical Cayley graph of S_5 (S_n with $n = 5$), contrasting a bimodal empirical distribution with unimodal

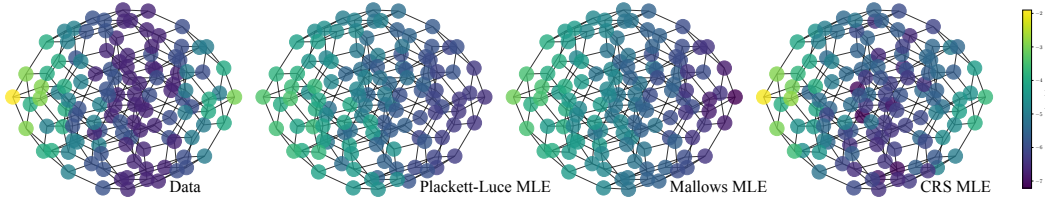


Figure 1: A synthetic multimodal distribution on the canonical Cayley graph of S_5 and the maximum likelihood estimates from the Plackett-Luce, Mallows, and full-rank CRS model classes.

Mallows and Plackett-Luce maximum likelihood estimates, as well the maximum likelihood estimate of the model we introduce in this work, the *contextual repeated selection* (CRS) model.

An important tool for the modelling approach in this work is the transformations of rankings into choice data, where we can then employ tractable choice models to create choice-based models of ranking data. Building on the ranking literature on *L-decomposable distributions* [14], we conceptualize rankings as arising from a “top-down” sequence of choices, allowing us to create novel ranking models from recently introduced choice models. Both Plackett-Luce and Mallows models can be described as arising from such a top-down choice process [18]. We term this generic decomposition *repeated selection*. Estimating such ranking models reduces to estimating choice models on choice data implied by the ranking data, making model inference tractable whenever the underlying choice model inference is tractable.

Our contextual repeated selection (CRS) model arises from applying the recently introduced context-dependent utility model (CDM) [48] to choices arising from repeated selection. The CDM model is a modern recasting of a choice model due to Batsell and Polking [6], an embedding model of choice data similar to popular embedding approaches [38, 42, 47]. By decomposing a ranking into a series of repeated choices and applying the CDM, we obtain ranking models that are straightforward to estimate, with provable estimation guarantees inherited from the CDM.

Our theoretical analysis of the CRS ranking model builds on recent work giving structure-dependent finite-sample risk bounds for the maximum likelihood estimator of the MNL [49] and CDM [48] choice models. As a foundation for our eventual analysis of the CRS model, we improve and generalize several existing results for the MNL choice, CDM choice, and PL ranking models. Our work all but completes the theory of maximum likelihood estimation for the MNL and PL models, with expected risk and tail bounds that match known lower bounds. The tail bounds stem from a new Hanson-Wright-type tail inequality for random quadratic forms [25, 27, 46] with block structure (see Appendix, Lemma 3), itself of potential stand-alone interest. Our tight analysis of the PL tail and expected risk stems from a careful spectral analysis of the (random) Plackett-Luce comparison Laplacian that arises when ranking data is viewed as choice data (see Appendix, Lemma 4).

Our empirical evaluations focus both on predicting out-of-sample rankings as well as predicting sequential entries of rankings as the top entries are revealed. We find that the flexible CRS model we introduce in this work achieves significantly higher out-of-sample likelihood, compared to the PL and Mallows models, across a wide range of applications including ranked choice voting from elections, sushi preferences, Nascar race results, and search engine results. By decomposing the performance to positions in a ranking, we find that while our new model performs similarly to PL on predicting the top entry of a ranking, our model is much better to predict subsequent top entries. Our investigation demonstrates the broad efficacy of our approach across applications as well as dataset characteristics: these datasets differ greatly in size, number of alternatives, how many rankings each alternative appears in, and uniformity of the ranking length.

Other related work. There is an extensive body of work on modeling and learning distributions over the space of rankings, and we do not attempt a complete review here. Early multimodal ranking distributions include Thurstone’s Case II model with correlated noise [51] from the 1920’s and Babington Smith’s model [50] from the 1950’s, though both are intractable [21, 35]. Mixtures of unimodal models have been the most practical approach to multimodality to date [3, 13, 22, 31, 39, 41, 53], but are typically bogged down by expectation maximization (EM) or other difficulties.

Our approach of connecting rankings to choices is not new; repeated selection was first used to connect the MNL model of choice to the PL model of rankings [43]. Choice-based representations

of rankings in terms of pairwise choices are studied in *rank breaking* [5, 28, 40], whereas repeated selection can be thought of as a generalization, “choice breaking” *beyond* pairwise choices. The richness of the CRS model largely stems from the richness of the CDM choice model [48], one of several recent models to inject richness in discrete choice [7, 8, 45].

Our expected risk and risk tail bounds for maximum likelihood estimation stem from prior work for both the MLE for PL [24] and MNL [49] models. For MNL, risk bounds also exist for non-MLE estimators such as those based on rank breaking [4], LSR [37], and ALSR [1]. However, all prior analyses (including for the MLE) fall short of *tight* guarantees (upper bounds that unconditionally match lower bounds). For the MNL model, Shah et al. [49] provides a tail bound for the pairwise setting and a (weak) expected risk bound for larger sets of a uniform size (that grows weaker for larger sets). Our results (tail and risk bounds) for MNL apply to any collection of variable-sized sets, a generalization that is itself necessary for our subsequent analysis of the PL and CRS models in a choice framework. Placing the focus back on rich ranking models, the tail and expected risk results for the CRS ranking model are the first of their kind for ranking models that are not unimodal in nature, meaningfully augmenting the scope of existing theoretical work on rankings.

2 Rankings from choices

We first introduce rankings, then choices, and develop the methodology connecting the two that is crucial to our paper’s framework. Central to all three definitions is the notion of an item universe, \mathcal{X} , denoting a finite collection of n items. Let $[n]$ denote the set of numbers $1, \dots, n$, indexed by i, j, k .

Rankings. A ranking σ orders the items in the universe \mathcal{X} , $\sigma : \mathcal{X} \mapsto [n]$. A ranking is also a bijection, letting us define $\sigma^{-1}(\cdot)$, the inverse mapping of σ . For any item $x \in \mathcal{X}$, $\sigma(x)$ denotes its rank, with a value of 1 indicating the highest position, and n the lowest position. Similarly, the item in the i th rank is $\sigma^{-1}(i)$. A ranking distribution $P(\cdot)$ is a discrete probability distribution over the space of rankings S_n . That is, every ranking $\sigma \in S_n$ is assigned a probability, $0 \leq P(\sigma) \leq 1$, and $\sum_{\sigma \in S_n} P(\sigma) = 1$. A ranking model is then a particular representation of a ranking distribution P , parametric or not, including the Plackett-Luce, Mallows, and Thurstone models.

Discrete choice. Discrete choice modeling concerns itself with the *conditional* probability of a choice from a set $S \subseteq \mathcal{X}$, given that set S . That is, the modeling framework does not account for the process that S is generated from (i.e., the probabilities different subsets may arise), but only the probability of choosing an item from a set, given that set a priori. Given a subset $S \subseteq \mathcal{X}$, a choice of $x \in S$ is denoted by the ordered pair (x, S) . The distribution of probabilities that x is chosen from a given S is denoted by $P(x|S)$, $\forall x \in S, \forall S \subseteq \mathcal{X}, |S| \geq 2$. That is, for every S , each $x \in S$ is assigned a probability $0 \leq P(x|S) \leq 1$, and $\sum_{x \in S} P(x|S) = 1$.

Repeated selection. Repeated selection follows a natural approach to constructing a ranking of the items of a set. Consider first the item that is preferred to all items and assign it rank 1. Then, of the items that remain, the item that is preferred is assigned rank 2. This assignment process is repeated until only one item remains, which is assigned rank n . In this way, a ranking is envisioned as a sequence of repeatedly identifying preferred items from a shrinking slate of options. When the sequence of choices are conditionally independent, we term this approach and its resulting interpretation *repeated selection*. Formally, a ranking distribution $P(\sigma)$ arising from repeated selection has the form

$$P(\sigma(x_1) = 1, \sigma(x_2) = 2, \dots, \sigma(x_n) = n) = P(x_1|\mathcal{X})P(x_2|\mathcal{X} \setminus x_1) \cdots P(x_{n-1}|\{x_{n-1}, x_n\}).$$

It is easy to verify that any such distribution satisfies $\sum_{\sigma \in S_n} P(\sigma) = 1$. Under repeated selection, a ranking is converted into two objects of importance: a collection of choice sets, each a subset of the universe \mathcal{X} , as well as a sequence of independent choices conditioned on the choice sets. The latter (the conditional choice) is the subject of discrete choice modeling while the former (the collection) is a relatively unstudied random object that is a major focus of our analysis. The independence is worth emphasizing: the choices, conditioned on their choice sets, are treated as independent from one another. In contrast, the unconditioned choices are *not* independent from one another: certainly, knowledge of the first ranked item ensures that no other choice is that item.

Decomposing ranking distributions into independent repeated choices this way is not generic; see Critchlow et al. [14] for an extensive treatment of which ranking distributions can be *L-decomposed*

(decomposed from the “left”). As one example of its lack of generality, consider a process of *repeated elimination*, by which a choice model is applied as an elimination process, and the item to be first eliminated from a set is assigned the lowest rank, and the item to be eliminated from the set that remains, the second lowest, and so on. The resulting decomposition of the ranking (the “R-decomposition”) generically induces an entirely different family of ranking distributions for a given family of choice models.

2.1 Popular examples of ranking via repeated selection

We illustrate a few well known ranking models, and how they are a result of repeated selection applied to choice models. These models serve as the basis for much of the literature on modeling rankings. Though pervasive, all of these examples result in families of ranking distributions that center around a single total ordering—that is, the ranking distributions are unimodal.

Plackett-Luce. Perhaps the most popular discrete choice model is the Multinomial Logit (MNL) model, which describes the process of choice from a subset S as simply a choice from the universe \mathcal{X} , conditioned on that choice being in the set S . This statement, along with some regularity conditions, is known as Luce’s Choice Axiom [32]. That is,

$$P(x|S) = P(x|\mathcal{X}, x \in S) = \frac{P(x|\mathcal{X})}{\sum_{y \in S} P(y|\mathcal{X})} = \frac{\exp(\theta_x)}{\sum_{y \in S} \exp(\theta_y)},$$

where the final equality follows from setting $\theta_z = \log(P(z|\mathcal{X}))$, a popular parameterization of the model where $\theta \in \mathbb{R}^n$ are interpretable as utilities. By repeatedly selecting from the Multinomial Logit Model, we arrive at the Plackett-Luce model of rankings [43]:

$$P(\sigma(x_1) = 1, \sigma(x_2) = 2, \dots, \sigma(x_n) = n \mid \theta) = \prod_{i=1}^{n-1} P(x_i | \mathcal{X} \setminus \cup_{j=1}^{i-1} x_j; \theta) = \prod_{i=1}^n \frac{\exp(\theta_{x_i})}{\sum_{j=i}^n \exp(\theta_{x_j})}.$$

The MNL model belongs to the broad class of independent Random Utility Models (RUMs) [34]. Any such RUM can be composed into a *utility-based ranking model* via repeated selection.

Mallows. The Mallows model assigns probabilities to rankings in a manner that decreases exponentially in the number of pairwise disagreements to a reference ranking σ_0 . More precisely, under a Mallows model with concentration parameter θ and reference ranking σ_0 , $P(\sigma; \sigma_0, \theta) \propto \exp(-\theta \tau(\sigma, \sigma_0))$, where $\tau(\cdot, \cdot)$ is Kendall’s τ distance. The model can be fit into the framework of repeated selection via the choice model: $P(x|S) \propto \exp(-\theta |\{y \in S : \sigma_0(y) < \sigma_0(x)\}|)$ [18]. The model’s reliance on a reference ranking σ_0 makes it generally NP-Hard to estimate from data [9, 16]. Mallows also belongs to a broader class of distance-based models, which replace Kendall’s τ with other distance functions between rankings [14].

2.2 Beyond unimodality: contextual ranking with the CRS model

The recently introduced context-dependent utility model (CDM) of discrete choice [48] is both flexible and tractable, making it an attractive choice model to study in a repeated selection framework. The CDM models the probability of selecting an item x from a set S as proportional to a sum of pairwise interaction terms between x and the other items $z \in S$. This strategy of incorporating a “pairwise dependence of alternatives” enables the CDM to subsume the MNL model class while also incorporating a range of context effects. Moreover, the matrix-like parameter structure of the CDM also opens the door for factorized representations that greatly improve the parametric efficiency of the model. The CDM choice probabilities, in full and factorized form, are then:

$$P(x|S) = \frac{\exp(\sum_{z \in S \setminus x} u_{xz})}{\sum_{y \in S} \exp(\sum_{z \in S \setminus y} u_{yz})} = \frac{\exp(\sum_{z \in S \setminus x} c_z^T t_x)}{\sum_{y \in S} \exp(\sum_{z \in S \setminus y} c_z^T t_y)},$$

where $u \in \mathbb{R}^{n(n-1)}$ represents the parameter space of the unfactorized CDM (a parameter for every ordered pair indexed by ordered pairs) and $T \in \mathbb{R}^{n \times r}$, $C \in \mathbb{R}^{n \times r}$ represents the parameter space of the factorized CDM, where r is the dimension of the latent representations. Pushed through the repeated selection framework, we arrive at the CRS model of rankings, in full and factorized form:

$$P(\sigma(x_1) = 1, \dots, \sigma(x_n) = n) = \prod_{i=1}^n \frac{\exp(\sum_{k=i+1}^n u_{x_i x_k})}{\sum_{j=i}^n \exp(\sum_{k=j+1}^n u_{x_j x_k})} = \prod_{i=1}^n \frac{\exp(\sum_{k=i+1}^n c_k^T t_{x_i})}{\sum_{j=i}^n \exp(\sum_{k=j+1}^n c_k^T t_{x_j})}.$$

Just as the factorized CDM subsumes the MNL model for every r , CRS subsumes the PL model. The benefits of a low-rank factorization are often immense in practice. The full CRS can be useful, but its parameter requirements scale quadratically with the number of items n , and is therefore best applied only to settings where n is small. The full CRS is however conveniently amenable to many theoretical analyses, having a smooth and strongly convex likelihood whose landscape looks very similar to the Plackett-Luce likelihood. We thus focus our theoretical analysis of CRS on the full model, noting that all our guarantees that apply to the full CRS also apply to the factorized CRS. The factorized CRS model likely enjoys sharper guarantees for small r .

3 Better guarantees for MNL and Plackett-Luce

Efficient estimation is the main roadblock to employing rich ranking models in practice. This fact alone makes convergence guarantees—the type we provide in this section and the next—immensely valuable when assessing the viability of a model. Such guarantees for repeated selection ranking models involves both an analysis of the process by which a ranking is converted into conditionally independent choices, as well an analysis of the choice model that repeated selection is equipped with. While our efforts were originally focused on risk bounds for the new CRS model, in working to produce the best possible risk bounds for that model we identified several gaps in the analysis of more basic, widely used choice and ranking models. We first provide novel improved guarantees for existing foundational models, specifically, the MNL choice model and the PL ranking model, before proceeding to the CRS model in the next section. Relatively small modifications of the proofs in this section yield results for any utility-based ranking model (Section 2.1) that has a smooth and strongly convex likelihood.

In this section and the next, we focus on a ranking dataset $\mathcal{R} = \{\sigma_1, \dots, \sigma_\ell\}$ of ℓ independent rankings each specified as a total ordering of the set \mathcal{X} where $|\mathcal{X}| = n$. Given a repeated selection model of rankings generically parameterized by θ , the likelihood for the dataset \mathcal{R} becomes:

$$\mathcal{L}(\theta; \mathcal{R}) = \prod_{i=1}^{\ell} p(\sigma_i; \theta) = \prod_{i=1}^{\ell} \prod_{j=1}^n P(x_i | \mathcal{X} \setminus \cup_{k=1}^{j-1} x_k; \theta). \quad (1)$$

We can maximize the likelihood over θ to find the maximum likelihood estimate (MLE). Since the choices within each ranking are conditionally independent, the ranking likelihood reduces to a likelihood of a choice dataset with $\ell(n-1)$ choices. Finding the MLE of a repeated selection ranking model is thus equivalent to finding the MLE of a choice model. Because the PL and full CRS likelihoods are smooth and strongly log concave, we can efficiently find the MLEs in practice.

As a stepping stone to ranking, in Theorem 1 we first provide structure-dependent guarantees on the MLE for the underlying MNL choice models. Then, in Theorem 2 we analyze the set structure induced by repeated selection to provide guarantees on the PL ranking model of ranking data. This two-step process decouples the “choice randomness“, or the randomness inherent to selecting the best item from the remaining set of items, from the “choice set randomness“, the randomness inherent to the set of remaining items. All proofs are found in the Appendix.

Multinomial logit. The following theorem concerns the risk of the MLE for the MNL choice model, which evaluates the proximity of the estimator to the truth in Euclidean distance. We give both a tail bound and a bound on the expected risk.

Theorem 1. *Let θ^* denote the true MNL model from which data is drawn. Let $\hat{\theta}_{MLE}$ denote the maximum likelihood solution. For any $\theta^* \in \Theta_B = \{\theta \in \mathbb{R}^n : \|\theta\|_\infty \leq B, \mathbf{1}^T \theta = 0\}$, $t > 1$, and dataset \mathcal{D} generated by the MNL model,*

$$\mathbb{P} \left[\left\| \hat{\theta}_{MLE}(\mathcal{D}) - \theta^* \right\|_2^2 \geq c_{B, k_{max}} \frac{t}{m \lambda_2(L)^2} \right] \leq e^{-t},$$

where k_{max} is the maximum choice set size in \mathcal{D} , $c_{B, k_{max}}$ is a constant that depends on B and k_{max} , and $\lambda_2(L)$ depends on the spectrum of a Laplacian L formed by \mathcal{D} . For the expected risk,

$$\mathbb{E} \left[\left\| \hat{\theta}_{MLE}(\mathcal{D}) - \theta^* \right\|_2^2 \right] \leq c'_{B, k_{max}} \frac{1}{m \lambda_2(L)^2},$$

where $c'_{B, k_{max}}$ is again a constant that depends on B and k_{max} .

Focusing first on the expected risk bound, we see it tends to zero as the dataset size m increases. The underlying set structure, represented in the bound by the object $\lambda_2(L)$, plays a significant role in the rate at which the bound vanishes. Here, L is the Laplacian of the undirected weighted graph formed by the choice sets in \mathcal{D} . The algebraic connectivity of the graph, $\lambda_2(L)$, represents the extent to which there are good cuts in the comparison graph, i.e. whether all items are compared often to each other. Should there be more than one connected component in the graph, $\lambda_2(L)$ would be 0, and the bound would lose meaning. This behavior is not errant—the presence of more than a single connected component in L implies that there is a non trivial partition of \mathcal{X} such that no items in one partition have been compared to another, meaning that the relative ratio of the utilities could be arbitrarily large and the true parameters are *unidentifiable*.

The role of $\lambda_2(L)$ here is similar to Ford’s [19] necessary and sufficient condition for MNL to be identifiable, that the directed comparison graph be strongly connected. The difference, however, is that $\lambda_2(L)$ depends only on the undirected comparison graph constructed only from the choice sets. The apparent gap between directed and undirected structure is filled by B , the bound on the true parameters in θ^* . As is natural, our bound also diverges if B diverges. The remaining terms in the expression regulate the role of set sizes: larger set sizes increase algebraic connectivity, but make the likelihood less smooth, effects that ultimately cancel out for a balanced distribution of sets.

Theorem 1 is the first MNL risk bound that handles multiple set sizes, and is the first to be tight up to constants for set sizes that are not bounded by a constant. Our proof of the expected risk bound sharpens and generalizes the single-set-size proof of Shah et al. [49] to variable sized sets and largely follows the outline of the original proof, albeit with some new machinery (see e.g. Lemma 1, leveraging an old result due to Bunch–Nielsen–Sorensen [10], and the discussion of Lemma 1 in the proof of Theorem 1). A matching lower bound for the expected risk may be found in Shah et al., thus demonstrating the minimax optimality of the MLE at a great level of generality.

The tail bound component of the theorem is the first to go beyond pairwise comparisons. The result relies on a tail bound lemma, Lemma 3, that applies Hoeffding’s inequality in ways that leverage special block structure innate to Laplacians built from choice data. This lemma replaces the use of a Hanson-Wright-type inequality in Shah et al.’s tail bounds for pairwise MNL. Lemma 3 leverages the fact that the constituent random variables are bounded, not merely subgaussian, to furnish a tail bound that is stronger than what Hanson-Wright-type tools deliver for this problem.

Plackett-Luce. With tight guarantees for the MLE of the MNL model, we proceed to analyze the PL ranking model. As Equation (1) demonstrates, the PL likelihood is simply a manifestation of the MNL likelihood. However, for rankings, the MNL tail bound provided so far is a random quantity, owing to the randomness of $\lambda_2(L)$. In choice, only the “choice randomness” is accounted for, and the choice sets are taken as given. In rankings, however, the choice sets themselves are random and we must therefore account for the “choice set randomness” that remains. We give expected risk bounds and tail bounds for the PL model in the following result.

Theorem 2. *Let $\mathcal{R} = \sigma_1, \dots, \sigma_\ell \sim PL(\theta^*)$ be a dataset of full rankings generated from a Plackett-Luce model with true parameter $\theta^* \in \Theta_B = \{\theta \in \mathbb{R}^n : \|\theta\|_\infty \leq B, \mathbf{1}^T \theta = 0\}$ and let $\hat{\theta}_{MLE}$ denote the maximum likelihood solution. Assume that $\ell > 4 \log(\sqrt{\alpha_B n}) / \alpha_B^2$ where α_B is a constant that only depends on B . Then for $t > 1$ and dataset \mathcal{R} generated by the PL model,*

$$\mathbb{P} \left[\left\| \hat{\theta}_{MLE}(\mathcal{R}) - \theta^* \right\|_2^2 \geq c'_B \frac{n}{\ell} t \right] \leq e^{-t} + n^2 \exp(-\ell \alpha_B^2) \exp \left(\frac{-t}{\alpha_B^2 n^2} \right),$$

where c'_B is a constant that depends on B . For the expected risk,

$$\mathbb{E} \left[\left\| \hat{\theta}_{MLE}(\mathcal{R}) - \theta^* \right\|_2^2 \right] \leq c'_B \frac{n^3}{\ell} \mathbb{E} \left[\frac{1}{\lambda_2(L)^2} \right] \leq c_B \frac{n}{\ell},$$

where $c'_B = 4 \exp(4B)$, $c_B = 8 \exp(4B) / \alpha_B^2$, and L is the PL Laplacian constructed from \mathcal{R} .

The expectation in the expected risk is taken over both the choices and choice set randomness, ensuring that the quantity on the final right hand side is deterministic. It is not difficult to show that $\lambda_2(L)$ is always positive (and thus $1/\lambda_2(L) < \infty$) for PL: every ranking contains a choice from the full universe, which is sufficient. Theorem 2 takes additional advantage of the fact that B is often small, which results in subsets that are extraordinarily diverse, giving a considerably larger $\lambda_2(L)$ as

soon as the dataset has a sufficient number of rankings. The technical workhorse of Theorem 2 is Lemma 4, which provides a high probability lower bound on $\lambda_2(L)$ for the (random) Plackett-Luce Laplacian L .

Both our expected risk and tail bounds are the first bounds of their kind for the PL model, which matches a known lower bound on the expected risk (Theorem 1 in Hajek et al. [24]). Though the authors of that work claim to have bounds on expected risk that are weak by a $\log(n)$ factor, a closer inspection reveals that they only furnish upper bounds on a particular quantile of the risk. Much like our MNL tail bound, our PL tail bound integrates to a result on the expected risk that has the same parametric rates as our direct proof of the expected risk bound.

4 Convergence guarantees for the CRS model

The CRS model defines much richer distributions on S_n than the PL model, but we are still able to demonstrate guaranteed convergence, a result that is the first of its kind for a non-simplistic model of ranking data. The focus of our study will be the full CRS model, statistical guarantees for which carry over to factorized CRS models of *any* rank.

Our analysis of the PL model required a generalized (to multiple set sizes) re-analysis of the MNL choice model. Similarly, we improve upon the known guarantees for the CDM choice model [48] that underlies the CRS ranking model by proving a tail bound in Lemma 6. Moreover, the added model complexity of the CDM creates new challenges, notably a notion of (random) “structure”, in the structure-dependent bound, which does not simply reduce to analyzing a (random) Laplacian.

We first consider conditions that ensure the CRS model parameters are not underdetermined, conditions without which the risk can be arbitrarily large. Whereas the MNL model is immediately determined with choices from a single ranking—all the model requires is a single universe choice—a sufficient condition for CDM requires choices from all sets of at least 2 different sizes, with some technical exceptions (see [48], Theorem 1). Meeting this sufficient condition requires that at least n rankings be present, since the two smallest collections of sets are the single set of size n and the n sets of size $n - 1$. We demonstrate in Lemma 5 that, with high probability, $O(n \log(n)^2)$ rankings suffice to meet this sufficient condition. Of course, high probability does not mean always; and for CRS we more strongly rely on the assumption that the true parameter lies in a compact space to ensure that the risk is always bounded. Such assumptions are in fact always necessary for convergence guarantees of any kind, even for the basic MNL model [49].

We are now ready to present our main theoretical result for the CRS ranking model:

Theorem 3. *Let $\mathcal{R} = \sigma_1, \dots, \sigma_\ell \sim CRS(u^*)$ be a dataset of full rankings generated from the full CRS model with true parameter $u^* \in \Theta_B = \{u \in \mathbb{R}^{n(n-1)} : u = [u_1^T, \dots, u_n^T]^T; u_i \in \mathbb{R}^{n-1}, \|u_i\|_1 \leq B, \forall i; \mathbf{1}^T u = 0\}$ and let \hat{u}_{MLE} denote the maximum likelihood solution. Assuming that $\ell > 8ne^{2B} \log(8ne^{2B})^2$, c_B, \dots, c_B''' are constants that depend only on B , and $t > 1$:*

$$\mathbb{P} \left[\|\hat{u}_{MLE}(\mathcal{D}) - u^*\|_2^2 > \frac{c_B''' n^4}{\ell \delta_n} t \right] \leq e^{-t} + n \exp \left(-t \min \left\{ \frac{c_B' n^3}{\ell \delta_n}, 1 \right\} \right) e^{-\ell / (8ne^{2B})}.$$

For the expected risk,

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

where L is a p.s.d. matrix constructed from \mathcal{R} and $\delta_n > 0$ is a quantity that depends only on n .

Similar to Theorem 2, the expectation is taken over both the choices and choice sets, rendering the final bound deterministic. The L in the intermediate expression is not generally a graph Laplacian but rather a block structured matrix that captures the complex dependencies of the CDM parameters. The quantity $\delta_n > 0$, is the second smallest eigenvalue of a fixed block-structured matrix specified in Lemma 5. From simulations, δ_n 's dependence on n is $O(1/n^3)$.

These expected and tail risk bounds may strike the reader as having a disappointing rate in n . On one hand, modeling intransitivity, multimodality, and other richness comes at an inherent cost. The fact that any CRS model subsumes the PL model is also indicative of a larger rate of convergence. Despite

Table 1: Average out-of-sample negative log-likelihood for the MLE of repeated selection ranking models across different datasets (lowercase) or collections of datasets (uppercase), \pm standard errors (of the mean) from five-fold cross-validation. Best results for each dataset appear in bold.

	PL	CRS, $r = 1$	CRS, $r = 4$	CRS, $r = 8$	Mallows (MGA)
sushi	14.24 \pm 0.02	13.94 \pm 0.02	13.57 \pm 0.02	13.47 \pm 0.02	22.23 \pm 0.026
dub-n	8.36 \pm 0.02	8.18 \pm 0.02	7.61 \pm 0.02	7.59 \pm 0.02	11.65 \pm 0.02
dub-w	6.36 \pm 0.02	6.27 \pm 0.02	5.87 \pm 0.02	5.86 \pm 0.01	7.21 \pm 0.02
meath	8.46 \pm 0.02	8.23 \pm 0.02	7.59 \pm 0.02	7.56 \pm 0.02	11.85 \pm 0.07
nascar	113.0 \pm 1.4	112.1 \pm 1.5	103.9 \pm 1.8	102.6 \pm 1.8	238.5 \pm 0.3
LETOR	12.2 \pm 1.0	12.2 \pm 1.0	10.5 \pm 1.1	9.8 \pm 1.1	22.5 \pm 0.5
PREF-SOC	5.52 \pm 0.08	5.53 \pm 0.07	5.55 \pm 0.14	5.54 \pm 0.15	7.05 \pm 1.38
PREF-SOI	4.1 \pm 0.1	4.0 \pm 0.1	3.9 \pm 0.1	3.9 \pm 0.1	6.8 \pm 0.2

these factors, we believe the large rates are likely a result of weakness in our analysis. In practice, as we demonstrate via simulations in Appendix B, the full CRS model appears to converge considerably faster, $O(n^2/\ell)$. The factorized CRS model, used in our empirical work, likely does even better.

The smoothness and strong convexity of the full CRS likelihood render it easy to optimize to obtain the MLE, making our result meaningful in practice. In contrast, MLE risk for ranking mixtures models is difficult to bound [41], and the separate difficulty of finding the MLE for mixtures [3] would question the value of such a result. Our bound on the expected risk extends to factorized CRS models, and despite the non-convexity of factorization models, gradient-based optimization often succeeds in finding global minima in practice and are widely conjectured to generally converge [20, 23, 30].

5 Empirical results

We evaluate the performance of various repeated selection models in learning from and making predictions on empirical datasets, a relative rarity in the theory-focused ranking literature. The datasets span a wide variety of human decision domains including ranked elections and food preferences, while also including (search) rankings made by algorithms. We find across all but one dataset that the novel CRS ranking model outperforms other models in out-of-sample prediction.

We study four widely studied datasets: the `sushi` dataset representing ranked food preferences, the `dub-n`, `dub-w`, and `meath` datasets representing ranked choice voting, the `nascar` dataset representing competitions, and the `LETOR` collection representing search engine rankings. We provide detailed descriptions of the datasets in Appendix A, as well as an explanation of the more complex `PREF-SOC` and `PREF-SOI` collections. Many of these datasets consist of top- k rankings [17] of mixed length, which are fully amenable to decomposition through repeated selection. We use stochastic gradient-based optimization methods [29] to optimize our PL and CRS models. We optimize the Mallows model via the Mallows Greedy Approximation (MGA) [44] approach. We elaborate on the discrepancy in optimization routines and provide more details about our procedures in Appendix A.

In Table 1 we report average out-of-sample negative log-likelihoods for all datasets and collections, averaged over 5 folds. We see that across a range of dimensions r the factorized CRS model typically offers significantly improved performance, or at least no worse performance, than the Mallows and Plackett-Luce models (where the CRS model generalizes the latter). For all datasets, the Mallows Greedy Approximation (MGA)-based model is markedly worse than the other models. A deeper, position-level analysis of model performance is given in Appendix A.

6 Conclusion

We introduce the contextual repeated selection (CRS) model of ranking, a model that can eschew traditional assumptions such as intransitivity and unimodality allowing it to capture nuance in ranking. Our model fits data significantly better than existing models for a wide range of ranking domains including ranked choice voting, food preference surveys, race results, and search engine results. Our theoretical guarantees on the CRS model provide theoretical foundations for the performance we observe. Moreover, our risk analysis of ranking models closes the gap in the theory of maximum likelihood estimation for the multinomial logit (MNL) and Plackett-Luce (PL) models, and opens the door for future rich models and analyses of ranking data.

Broader Impact

Flexible ranking distributions that can be learned with provable guarantees can facilitate more powerful and reliable ranking algorithms inside recommender systems, search engines, and other ranking-based technological products. As a potential adverse consequence, more powerful and reliable learning algorithms can lead to an increased inappropriate reliance on technological solutions to complex problems, where practitioners may not fully grasp the limitations of our work, e.g. independence assumptions, or that our risk bounds hold only under the assumption that the data is generated by the model.

Acknowledgements

This work is supported in part by an NSF Graduate Research Fellowship (AS), a Dantzig-Lieberman Fellowship (SR), a Krishnan Shah Fellowship (SR), a David Morgenthaler II Faculty Fellowship (JU), a Facebook Faculty Award, and a Young Investigator Award from the Army Research Office (73348-NS-YIP).

References

- [1] Shivani Agarwal. On ranking and choice models. In *IJCAI*, pages 4050–4053, 2016.
- [2] Richard Arratia and Louis Gordon. Tutorial on large deviations for the binomial distribution. *Bulletin of mathematical biology*, 51(1):125–131, 1989.
- [3] Pranjal Awasthi, Avrim Blum, Or Sheffet, and Aravindan Vijayaraghavan. Learning mixtures of ranking models. In *Advances in Neural Information Processing Systems*, pages 2609–2617, 2014.
- [4] Hossein Azari Soufiani, William Chen, David C Parkes, and Lirong Xia. Generalized method-of-moments for rank aggregation. In *Advances in Neural Information Processing Systems*, pages 2706–2714, 2013.
- [5] Hossein Azari Soufiani, David C Parkes, and Lirong Xia. Computing parametric ranking models via rank-breaking. In *ICML*, pages 360–368, 2014.
- [6] Richard R Batsell and John C Polking. A new class of market share models. *Marketing Science*, 4(3):177–198, 1985.
- [7] Austin R Benson, Ravi Kumar, and Andrew Tomkins. On the relevance of irrelevant alternatives. In *Proceedings of the 25th International Conference on World Wide Web*, pages 963–973, 2016.
- [8] Jose Blanchet, Guillermo Gallego, and Vineet Goyal. A markov chain approximation to choice modeling. *Operations Research*, 64(4):886–905, 2016.
- [9] Mark Braverman and Elchanan Mossel. Sorting from noisy information. *arXiv preprint arXiv:0910.1191*, 2009.
- [10] James R Bunch, Christopher P Nielsen, and Danny C Sorensen. Rank-one modification of the symmetric eigenproblem. *Numerische Mathematik*, 31(1):31–48, 1978.
- [11] Giuseppe C Calafiore and Laurent El Ghaoui. *Optimization models*. Cambridge university press, 2014.
- [12] Shuo Chen and Thorsten Joachims. Modeling intransitivity in matchup and comparison data. In *Proceedings of the ninth acm international conference on web search and data mining*, pages 227–236. ACM, 2016.
- [13] Flavio Chierichetti, Anirban Dasgupta, Ravi Kumar, and Silvio Lattanzi. On learning mixture models for permutations. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 85–92. ACM, 2015.
- [14] Douglas E Critchlow, Michael A Fligner, and Joseph S Verducci. Probability models on rankings. *Journal of mathematical psychology*, 35(3):294–318, 1991.
- [15] Persi Diaconis. Group representations in probability and statistics. In *Lecture Notes-Monograph Series*. Institute for Mathematical Statistics, 1988.

- [16] Cynthia Dwork, Ravi Kumar, Moni Naor, and Dandapani Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622, 2001.
- [17] Ronald Fagin, Ravi Kumar, and Dakshinamurthi Sivakumar. Comparing top k lists. *SIAM Journal on discrete mathematics*, 17(1):134–160, 2003.
- [18] Michael A Fligner and Joseph S Verducci. Distance based ranking models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 359–369, 1986.
- [19] Lester R Ford Jr. Solution of a ranking problem from binary comparisons. *The American Mathematical Monthly*, 64(8P2):28–33, 1957.
- [20] Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1233–1242. JMLR. org, 2017.
- [21] John Geweke, Michael Keane, and David Runkle. Alternative computational approaches to inference in the multinomial probit model. *The review of economics and statistics*, pages 609–632, 1994.
- [22] Isobel Claire Gormley and Thomas Brendan Murphy. Exploring voting blocs within the irish electorate: A mixture modeling approach. *Journal of the American Statistical Association*, 103(483):1014–1027, 2008.
- [23] Suriya Gunasekar, Blake E Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nati Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.
- [24] Bruce Hajek, Sewoong Oh, and Jiaming Xu. Minimax-optimal inference from partial rankings. In *Advances in Neural Information Processing Systems*, pages 1475–1483, 2014.
- [25] David Lee Hanson and Farroll Tim Wright. A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics*, 42(3):1079–1083, 1971.
- [26] Thomas Hofmann and Jan Puzicha. Latent class models for collaborative filtering. In *IJCAI*, volume 99, 1999.
- [27] Daniel Hsu, Sham Kakade, Tong Zhang, et al. A tail inequality for quadratic forms of subgaussian random vectors. *Electronic Communications in Probability*, 17, 2012.
- [28] Ashish Khetan and Sewoong Oh. Generalized rank-breaking: computational and statistical tradeoffs. *The Journal of Machine Learning Research*, 19(1):983–1024, 2018.
- [29] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [30] Thomas Laurent and James Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *International Conference on Machine Learning*, pages 2902–2907, 2018.
- [31] Ao Liu, Zhibing Zhao, Chao Liao, Pinyan Lu, and Lirong Xia. Learning plackett-luce mixtures from partial preferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4328–4335, 2019.
- [32] R.. Duncan Luce. *Individual Choice Behavior a Theoretical Analysis*. John Wiley and sons, 1959.
- [33] Colin L Mallows. Non-null ranking models. i. *Biometrika*, 44(1/2):114–130, 1957.
- [34] C. F. Manski. The structure of random utility models. *Theory and Decision*, 8(3):229–254, 1977.
- [35] John I Marden. *Analyzing and modeling rank data*. CRC Press, 1996.
- [36] Nicholas Mattei and Toby Walsh. Preflib: A library of preference data [HTTP://PREFLIB.ORG](http://preplib.org). In *Proceedings of the 3rd International Conference on Algorithmic Decision Theory (ADT 2013)*, Lecture Notes in Artificial Intelligence. Springer, 2013.
- [37] Lucas Maystre and Matthias Grossglauser. Fast and accurate inference of Plackett–Luce models. In *Advances in Neural Information Processing Systems*, pages 172–180, 2015.
- [38] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

- [39] Thomas Brendan Murphy and Donal Martin. Mixtures of distance-based models for ranking data. *Computational statistics & data analysis*, 41(3):645–655, 2003.
- [40] Sahand Negahban, Sewoong Oh, Kiran K Thekumparampil, and Jiaming Xu. Learning from comparisons and choices. *Journal of Machine Learning Research*, 19(40), 2018.
- [41] Sewoong Oh and Devavrat Shah. Learning mixed multinomial logit model from ordinal data. In *Advances in Neural Information Processing Systems*, pages 595–603, 2014.
- [42] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710, 2014.
- [43] Robin L Plackett. Random permutations. *Journal of the Royal Statistical Society. Series B*, pages 517–534, 1968.
- [44] Tao Qin, Xiubo Geng, and Tie-Yan Liu. A new probabilistic model for rank aggregation. In *Advances in Neural Information Processing Systems*, pages 1948–1956, 2010.
- [45] Stephen Ragain and Johan Ugander. Pairwise choice markov chains. In *Advances in Neural Information Processing Systems*, pages 3198–3206, 2016.
- [46] Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.
- [47] Maja Rudolph, Francisco Ruiz, Stephan Mandt, and David Blei. Exponential family embeddings. In *Advances in Neural Information Processing Systems*, pages 478–486, 2016.
- [48] Arjun Seshadri, Alex Peysakhovich, and Johan Ugander. Discovering context effects from raw choice data. In *International Conference on Machine Learning*, pages 5660–5669, 2019.
- [49] Nihar B Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin J Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. *The Journal of Machine Learning Research*, 17(1):2049–2095, 2016.
- [50] B Babington Smith. Discussion of professor ross’s paper. *Journal of the Royal Statistical Society B*, 12(1):41–59, 1950.
- [51] Louis L Thurstone. A law of comparative judgment. *Psychological review*, 34(4):273, 1927.
- [52] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.
- [53] Zhibing Zhao, Peter Piech, and Lirong Xia. Learning mixtures of plackett-luce models. In *International Conference on Machine Learning*, pages 2906–2914, 2016.

A Empirical results supplement

A.1 Datasets

We consider a wide variety of application domains to demonstrate that repeated selection is appropriate and practical for general ranking problems. All the datasets we study can be found in the Preflib repository¹.

In our evaluation we study four widely studied datasets. First, the `sushi` dataset, consisting of 5,000 complete rankings of 10 types of sushi. Next, three election datasets, which consists of ranked choice votes given for three 2002 elections in Irish cities: the `dublin-north` election (abbreviated `dub-n` in tables) had 12 candidates and 43,942 votes for lists of varying length, `meath` had 14 candidates and 64,081 votes, and `dublin-west` (abbreviated `dub-w`) had 9 candidates and 29,988 votes. Third, the `nascar` dataset representing competitions, which consists of the partial ordering given by finishing drivers in each race of the 2002 Winston Cup. The data includes 74 drivers (alternatives) and 35 races (rankings).

The fourth collection we emphasize is the popular LETOR collection of datasets, which consists of ranking data arising from search engines. Although the LETOR data arises from algorithmic rather than human choices, it demonstrates the efficacy of our algorithms in large sparse data regimes. After removing datasets with fewer than 10 rankings and more than 100 alternatives (arbitrary thresholds that exclude small datasets with huge computational costs), the LETOR collection includes 727 datasets with a total of 12,838 rankings of between 3 and 50 alternatives.

Beyond these four emphasized collections, we include analyses of all 51 other Preflib datasets (as of May 2020) that contain partial or complete rankings of up to 10 items and at most 1000 rankings, a total of 11,956 rankings (these thresholds were again decided arbitrarily for computational reasons). We call this collection of datasets `PREF-SOI`, adopting the notation of Mattei and Walsh [36]. We separately study the subset of 10 datasets comprised of complete rankings, referred to herein as `PREF-SOC`, which contain a total of 5,116 rankings. The complete rankings in the `PREF-SOC` collection are suitable for both repeated selection and repeated elimination. While the `sushi` (complete ranking) and election (partial ranking) datasets are part of Preflib, they are comparatively quite large and are excluded from these two collections (`PREF-SOC` and `PREF-SOI`, respectively) by the above thresholds.

A.2 Training

We use the stochastic gradient-based optimization method Adam [29] implemented in Pytorch to train the MNL and CDM-based models in this paper. We run Adam with the default parameters ($lr = 0.001$, $\beta = (0.9, 0.999)$, $\epsilon = 1e - 8$). We use 10 epochs of optimization for the election datasets, where a single epoch converged.

We cannot use Adam (or any simple gradient-based method), for the Mallows model as the reference permutation parameter σ_0 lives in a discrete space. Instead we select the reference permutation via the Mallows Greedy Approximation (MGA) as in [44], and then optimize the concentration parameter numerically, conditional on that reference permutation. Our results broadly show that the Mallows model performs poorly compared to all the other models.

For all datasets we use 5-fold cross validation for evaluating test metrics. Using the `sushi` dataset as an example, for each choice model we train on repeated selection choices for each of 5 folds of the 5,000 rankings in the dataset. Across 5 choice models, 5 folds, and 5,000 rankings of, represents estimation parameters from a total of 125,000 rankings that represent over 1 million choices. The optimization can be easily guided to exploit sparsity, parallelization, and batching. All replication code is included in the supplement, and will be publicly available upon publication.

A.3 Position-level analysis

We measure the error at the k th position of a ranking σ given the set of already ranked items by adding up some distance between the choice probabilities p for the corresponding choice sets and the empirical distribution of those choices in the data. For repeated selection models, we define the

¹Preflib data is available at: <http://www.preflib.org/>

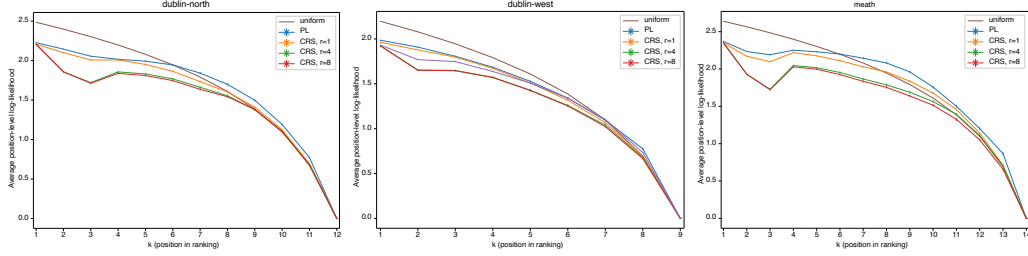


Figure 2: The average position-level log likelihood of choice probabilities for the dublin-north, dublin-west, and meath election datasets.

position-level log-likelihood at each position k as $\ell(k, \theta; \sigma) := \log p_{\theta}(\sigma^{-1}(k), \{\sigma^{-1}(j)\}_{j \geq k})$. When averaging ℓ over a test set T we obtain the average position-level log-likelihood:

$$\ell(k; \theta, T) := \frac{1}{|T|} \sum_{\sigma \in T: \text{len}(\sigma) \geq k} \ell(k, \theta; \sigma), \quad (2)$$

where $\text{len}(\sigma)$ is n for a full ranking and k for a top- k ranking.

In Figure 2 we analyze the election datasets at the position level, where we find that the CRS model makes significant gains relative to Plackett-Luce when predicting candidates near—but not at—the top of the list. We further notice that the performance is not monotonically decreasing in the number of remaining choices. Specifically, it is easier to guess the third-ranked candidate than the fourth, despite having fewer options in the latter scenario. A plausible explanation is that many voters rank candidates from a single political party and then stop ranking others, and the more nuanced choice models are assigning high probability to candidates when other candidates in their political party are removed.

B Auxiliary Material

B.1 Simulation Results

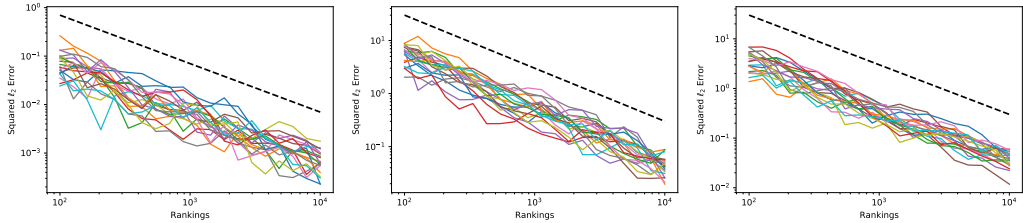


Figure 3: Squared ℓ_2 error of an estimated models in 20 growing datasets, validating our convergence theorems. The dashed black line are a visual guide of the slope $1/\ell$, and the tight bundle formed by the 20 different datasets validates the tail bound. (a) PL estimation error on PL model data, (b) CRS estimation error on PL model data, (c) CRS estimation error on CRS model data.

In this brief supplement we provide simulations that serve to validate our theoretical results. Figure 3 does so in two ways: first, showing that the error rate indeed decreases with $1/\ell$ as suggested by our risk bounds, and second that it does so with seemingly high probability, as shown by our tail bounds. The figure highlights three special cases, a PL model fit on PL data, a CRS model fit on PL data, and a CRS model fit on CRS data. All datasets consist of rankings of $n = 6$ items. For the PL model the number of parameters is $n = 6$. For the CDM model the number of parameters is $d = n(n - 1) = 30$. In both cases, the model parameters were sampled from a truncated standard normal distribution within a B -ball with $B = 1.5$ (per the theorem statements). In all three panels, we generate 20 datasets from the underlying model, and fit cumulative increments 20 times to generate the result. The tight bundle that the 20 datasets form indicates how little the randomness of a given dataset causes the risk to deviate. As in our main empirics, all maximum likelihood estimates were found using gradient-based optimization implemented in Pytorch.

In Figure 4, we demonstrate simulations that suggest that the CRS model’s true convergence rate appears to be $O(n^2/\ell)$, as opposed to the larger n -dependence, $O(n^4/(\delta_n\ell))$, that we were able to guarantee theoretically in Theorem 3. We generate the plot in a manner similar to Figure 3, by generating 20 datasets and fitting them incrementally, this time averaging all 20 datasets performance to produce a single line. We repeat this process for four different model sizes d corresponding to $n \in \{6, 9, 12, 16\}$. We then plot the resulting risk multiplied by ℓ/n^2 . The apparently constant set of lines over the wide range of parameters and dataset sizes indicates that the risk of the model is likely close to $O(n^2/\ell)$ in theory, suggesting room for improvement in our analysis.

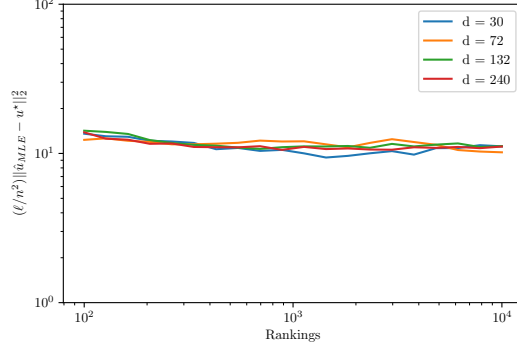


Figure 4: A visualization of the rate of CRS convergence multiplied by ℓ/n^2 . The legend highlights the parameters $d = n(n - 1)$ of the different models. The figure demonstrates that, over a wide range of parameters and rankings, the true rate of convergence for the CRS is likely $O(n^2/\ell)$.

B.2 Code for Reproducibility

Included along with the supplement is code that faithfully reproduces the Tables and Figures in both the main paper and the supplement. The main method in `cayley_code/cayley.py` reproduces Figure 1. The code to generate Table 1 and Figure 2 is contained in the `empirical_code` directory, and we refer the reader to `empirical_code/README.md` for detailed instructions. For the Simulation Code, the `generate_PL_error_plot` function in `simulation_code/laplacian_eigs.py` reproduces Figure 3, panel (a). Panels (b) and (c) are respectively reproduced by the functions `generate_CRS_on_PL_err_plot` and `generate_CRS_err_plot` in `simulation_code/cdm_laplacian_eigs.py`. Finally, Figure 4 can be reproduced using the function `generate_CRS_err_plot_variousn` in `simulation_code/cdm_laplacian_eigs.py`.

C Main Proofs

C.1 Proof of Theorem 1

Theorem 1. *Let θ^* denote the true MNL model from which data is drawn. Let $\hat{\theta}_{MLE}$ denote the maximum likelihood solution. For any $\theta^* \in \Theta_B = \{\theta \in \mathbb{R}^n : \|\theta\|_\infty \leq B, \mathbf{1}^T \theta = 0\}$, $t > 1$, and dataset \mathcal{D} generated by the MNL model,*

$$\mathbb{P} \left[\left\| \hat{\theta}_{MLE}(\mathcal{D}) - \theta^* \right\|_2^2 \geq c_{B, k_{max}} \frac{t}{m \lambda_2(L)^2} \right] \leq e^{-t},$$

where k_{max} is the maximum choice set size in \mathcal{D} , $c_{B, k_{max}}$ is a constant that depends on B and k_{max} , and $\lambda_2(L)$ depends on the spectrum of a Laplacian L formed by \mathcal{D} . For the expected risk,

$$\mathbb{E} \left[\left\| \hat{\theta}_{MLE}(\mathcal{D}) - \theta^* \right\|_2^2 \right] \leq c'_{B, k_{max}} \frac{1}{m \lambda_2(L)^2},$$

where $c'_{B, k_{max}}$ is again a constant that depends on B and k_{max} .

Proof.

We are given some true MNL model with parameters $\theta^* \in \Theta_B$, and for each datapoint $j \in [m]$ we have the probability of choosing item x from set C_j as

$$\mathbb{P}(y_j = x | \theta^*, C_j) = \frac{\exp(\theta_x^*)}{\sum_{y \in C_j} \exp(\theta_y^*)}.$$

We will first introduce notation for analyzing the risk, and then proceed to first give a proof of the expected risk bound. We then carry the technology of that proof forward to give a proof of the tail bound statement.

Notation. We now introduce notation that will let us represent the above expression in a more compact manner. Because our datasets involve choice sets of multiple sizes, we use $k_j \in [k_{\min}, k_{\max}]$ to denote the choice set size for datapoint j , $|C_j|$. Extending a similar concept in [49] to the multiple set sizes, we then define matrices $E_{j,k_j} \in \mathbb{R}^{n \times k_j}$, $\forall j \in [m]$ as follows: E_{j,k_j} has a column for every item $y \in C_j$ (and hence k_j columns), and the column corresponding to item $y \in C_j$ simply has the n -dimensional unit vector e_y . This definition then renders the vector-matrix product $\theta^T E_{j,k_j} = [\theta_{y_1}, \theta_{y_2}, \theta_{y_3}, \dots, \theta_{y_{k_j}}] \in \mathbb{R}^{1 \times k_j}$.

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

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

where the numerator always corresponds to the first entry of the input. These functions F_k have several properties that will become useful later in the proof. First, it is easy to verify that all F_k are shift-invariant, that is, $F_k(x) = F_k(x + c\mathbf{1})$, for any scalar c .

Next, from Lemma 1, we have that $\mathbf{1} \in \text{null}(\nabla^2(-\log(F_k(x))))$ and that

$$\nabla^2(-\log(F_k(x))) \succeq H_k = \beta_k(I - \frac{1}{k}\mathbf{1}\mathbf{1}^T), \quad (3)$$

where

$$\beta_k := \frac{1}{k \exp(2B)}. \quad (4)$$

That is, F_k are strongly log-concave with a null space *only* in the direction of $\mathbf{1}$, since $\nabla^2(-\log(F_k(x))) \succeq H_k$ for some $H_k \in \mathbb{R}^{k \times k}$, $\lambda_2(H_k) > 0$.

As a final notational addition, in the same manner as [49] but accounting for multiple set sizes, we define k permutation matrices $R_{1,k}, \dots, R_{k,k} \in \mathbb{R}^{k,k}$, $\forall k \in [k_{\min}, k_{\max}]$, representing k cyclic shifts in a fixed direction. Specifically, given some vector $x \in \mathbb{R}^k$, $y = x^T R_{l,k}$ is simply x^T cycled (say, clockwise) so $y_1 = x_l$, $y_i = x_{(l+i-1)\%k}$, where $\%$ is the modulo operator. That is, these matrices allow for the cycling of the entries of row vector $v \in \mathbb{R}^{1 \times k}$ so that any entry can become the first entry of the vector, for any of the relevant k . This construction allows us to represent any choice made from the choice set C_j as the first element of the vector x that is input to F , thereby placing it in the numerator.

First, an expected risk bound. Given the notation introduced above, we can now state the probability of choosing the item x from set C_j compactly as:

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

We can then rewrite the MNL likelihood as

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

and the scaled negative log-likelihood as

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

Thus,

$$\hat{\theta}_{\text{MLE}} = \arg \min_{\theta \in \Theta_B} \ell(\theta).$$

The compact notation makes the remainder of the proof a straightforward application of results from convex analysis: we first demonstrate that the scaled negative log-likelihood is strongly convex with respect to a semi-norm², and we use this property to show the proximity of the MLE to the optimal point as desired. The remainder of our expected risk bound proof mirrors that in [49] with a few extra steps of accounting created by the multiple set sizes. Beyond the additional accounting, one technical novelty in this expected risk proof, relative that in [49], is the development of Lemma 1 and its use to give a more careful handling of the Hessian. This handling is built on our observation that the Hessian is a rank-one modification of a symmetric matrix, whereby we can employ an argument due to Bunch–Nielsen–Sorensen [10] that relates the eigenvalues of such a matrix to the eigenvalues of its symmetric part. The tail bound proof (that follows this expected risk bound) is based on technical innovations that depart from previous strategies and will be surveyed there.

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

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

and the Hessian as

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

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

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

The first line follows from applying the definition of the Hessian. The second line follows from pulling the negative sign into the ∇^2 term. The third and fourth line follow from Equation (3), strong log-concavity of all F_k . The fifth line follows recognizing that H_k is invariant to permutation matrices. The sixth line follows from removing the inner sum since the terms are independent of i . The seventh line follows from lower bounding β_{k_j} by $\beta_{k_{\max}}$.

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

Now, defining the matrix L as

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

we first note a few properties of L . First, it is easy to verify that L is the Laplacian of a weighted graph on n vertices, with each vertex corresponding to an item. This conclusion follows because each term in the average corresponds to the Laplacian of an unweighted clique on the subset of nodes C_j , and the average of unweighted Laplacians is a weighted graph Laplacian. Weighted edges of the graph represented by L then denote when nonzero whether a pair of items has been compared in the dataset—that is, whether the pair of items has appeared together in some set C_j for some datapoint j . The weights of the edges then denote the proportion of times the corresponding pairs have been compared in the dataset.

It is now easy to verify that $L\mathbf{1} = 0$, and hence $\text{span}(\mathbf{1}) \subseteq \text{null}(L)$. Moreover, we can show that $\lambda_2(L) > 0$, that is, $\text{null}(L) \subseteq \text{span}(\mathbf{1})$, as long as the weighted graph represented by L is connected. This result follows because the number of zero eigenvalues of a weighted graph Laplacian represents the number of connected components of the graph. Hence, if the graph represented by L is connected, then $\lambda_2(L) > 0$.

We also define the matrix

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

Since $\frac{1}{k_j}$ is strictly positive, \hat{L} has nonzero weighted edges exactly where the graph represented by L does, but different weights. Hence, the two corresponding graphs' number of connected components are identical, and $\text{null}(\hat{L}) \subseteq \text{span}(\mathbf{1})$ if and only if $\text{null}(L) \subseteq \text{span}(\mathbf{1})$. Moreover, since $\hat{L} \succeq \frac{1}{k_{\max}} L$, we also have that $\lambda_2(\hat{L}) \geq \frac{1}{k_{\max}} \lambda_2(L)$. We work with \hat{L} for the remainder of the proof, but state our final results in terms of the eigenvalues of L . We use L in our results to maintain consistency of the final result with that of [49], and use \hat{L} in our proof to produce sharper results for the multiple set size case.

With the matrix \hat{L} , we can write,

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

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

$$\beta_{k_{\max}} \|\hat{\theta}_{\text{MLE}} - \theta^*\|_{\hat{L}}^2 \leq \langle \nabla \ell(\hat{\theta}_{\text{MLE}}) - \nabla \ell(\theta^*), \hat{\theta}_{\text{MLE}} - \theta^* \rangle.$$

Further, we have

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

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

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

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

$$\nabla \ell(\theta^*) = -\frac{1}{m} \sum_{j=1}^m \sum_{i=1}^{k_j} \mathbf{1}[y_j = i] E_{j,k_j} R_{i,k_j} \nabla \log(F_{k_j}(\theta^{*T} E_{j,k_j} R_{i,k_j})) = -\frac{1}{m} \sum_{j=1}^m E_{j,k_j} V_{j,k_j}, \quad (5)$$

where in the equality we have defined $V_{j,k_j} \in \mathbb{R}^{k_j}$ as

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

Now, taking expectations over the dataset, we have,

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

Here, the third equality follows from applying the expectation to the indicator and retrieving the true probability. The fourth line follows from applying the definition of gradient of log, and the final line from performing a change of variables $z = \theta^{*T} E_{j,k_j}$, pulling out the gradient and undoing the chain rule, and finally, recognizing that the expression sums to 1 for any z , thus resulting in a 0 gradient. We note that an immediate consequence of the above result is that $\mathbb{E}[V] = 0$, since V is simply a concatenation of the individual V_{j,k_j} .

Next, we have

$$\begin{aligned} \mathbb{E}[\nabla \ell(\theta^*)^T \hat{L}^\dagger \nabla \ell(\theta^*)] &= \frac{1}{m^2} \mathbb{E}\left[\sum_{j=1}^m \sum_{l=1}^m V_{j,k_j}^T E_{j,k_j}^T \hat{L}^\dagger E_{l,k_l} V_{l,k_l}\right] \\ &= \frac{1}{m^2} \mathbb{E}\left[\sum_{j=1}^m V_{j,k_j}^T E_{j,k_j}^T \hat{L}^\dagger E_{j,k_j} V_{j,k_j}\right] \\ &\leq \frac{\lambda_n(\hat{L}^\dagger)}{m^2} \mathbb{E}\left[\sum_{j=1}^m V_{j,k_j}^T E_{j,k_j}^T E_{j,k_j} V_{j,k_j}\right] \\ &= \frac{1}{m \lambda_2(\hat{L})} \mathbb{E}\left[\frac{1}{m} \sum_{j=1}^m V_{j,k_j}^T V_{j,k_j}\right] \\ &\leq \frac{1}{m \lambda_2(\hat{L})} \mathbb{E}\left[\sup_{l \in [m]} \|V_{l,k_l}\|_2^2\right], \end{aligned}$$

where the second line follows from the mean zero and independence of the V_{j,k_j} , the third from an upper bound of the quadratic form, the fourth from observing that the E_{j,k_j} do not change the norm

of the V_{j,k_j} , and the last from averages being upper bound by maxima. We then have that,

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

where $R_{i,k_j}^T R_{i,k_j}$ in the first line is simply the identity matrix. For the final line, recalling the expression for the log gradient of F_k ,

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

it is straightforward to show that $\sup_{v \in [-(k_{\max}-1)B, (k_{\max}-1)B]^{k_{\max}}} \|\nabla \log(F_{k_{\max}}(v))\|_2^2$ is always upper bounded by 2.

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

$$\mathbb{E}[\nabla \ell(\theta^*)^T \hat{L}^\dagger \nabla \ell(\theta^*)] \leq \frac{2}{m\lambda_2(\hat{L})}.$$

This expression in turn yields a bound on the expected risk in the \hat{L} semi-norm, which is,

$$\mathbb{E}\left[\beta_{k_{\max}}^2 \|\hat{\theta}_{\text{MLE}} - \theta^*\|_{\hat{L}}^2\right] \leq \frac{2}{m\lambda_2(\hat{L})}.$$

By noting that $\|\hat{\theta}_{\text{MLE}} - \theta^*\|_{\hat{L}}^2 = (\hat{\theta}_{\text{MLE}} - \theta^*)^T \hat{L} (\hat{\theta}_{\text{MLE}} - \theta^*) \geq \lambda_2(\hat{L}) \|\hat{\theta}_{\text{MLE}} - \theta^*\|_2^2$, since $\hat{\theta}_{\text{MLE}} - \theta^* \perp \text{null}(\hat{L})$, we can translate our finding into the ℓ_2 norm:

$$\mathbb{E}\left[\beta_{k_{\max}}^2 \|\hat{\theta}_{\text{MLE}} - \theta^*\|_2^2\right] \leq \frac{2}{m\lambda_2(\hat{L})^2}.$$

Applying the fact that $\lambda_2(\hat{L}) \geq \frac{1}{k_{\max}} \lambda_2(L)$, we get:

$$\mathbb{E}\left[\|\hat{\theta}_{\text{MLE}} - \theta^*\|_2^2\right] \leq \frac{2k_{\max}^2}{m\lambda_2(L)^2 \beta_{k_{\max}}^2}.$$

Now, setting

$$c'_{B,k_{\max}} := \frac{2k_{\max}^2}{\beta_{k_{\max}}^2} = 2 \exp(4B) k_{\max}^4,$$

we retrieve the expected risk bound in the theorem statement,

$$\mathbb{E}\left[\left\|\hat{\theta}_{\text{MLE}}(\mathcal{D}) - \theta^*\right\|_2^2\right] \leq c'_{B,k_{\max}} \frac{1}{m\lambda_2(L)^2}.$$

We close the expected risk portion of this proof with some remarks about $c_{B,k_{\max}}$. The quantity $\beta_{k_{\max}}$, defined in equation (4), serves as the important term that approaches 0 as a function of B and k_{\max} , requiring that the former be bounded. Finally, $\lambda_2(L)$ is a parallel to the requirements on the algebraic connectivity of the comparison graph in [49] for the pairwise setting.

From expected risk to tail bound. Our proof of the tail bound is a continuation of the expected risk bound proof. While the expected risk bound closely followed the expected risk proof of [49], our tail bound proof contains significant novel machinery. Our presentation seem somewhat circular, given that the tail bound itself integrates out to an expected risk bound with the same parametric rates

(albeit worse constants), but we felt that to first state the expected risk bound was clearer, given that it arises as a stepping stone to the tail bound.

Recall again the expression for the gradient in Equation (5). Useful in our analysis will be an alternate expression:

$$\nabla \ell(\theta^*) = -\frac{1}{m} \sum_{j=1}^m E_{j,k_j} V_{j,k_j} = -\frac{1}{m} E^T V,$$

where we have defined $V \in \mathbb{R}^{\Omega_{\mathcal{D}}}$ as the concatenation of all V_{j,k_j} , and $E \in \mathbb{R}^{\Omega_{\mathcal{D}} \times n}$, the vertical concatenation of all the E_{j,k_j} . Here, $\Omega_{\mathcal{D}} = \sum_{i=1}^m k_i$.

For the expected risk bound, we showed that V_{j,k_j} have expectation zero, are independent, and $\|V_{j,k_j}\|_2^2 \leq 2$. Next, we have

$$(\nabla \log(F_k(x)))_l = \mathbf{1}[l = 1] - \frac{\exp(x_l)}{\sum_{p=1}^k \exp(x_p)}, \quad (6)$$

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

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

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

Thus we have $M_{k_j} V_{j,k_j} = M_{k_j}^{\frac{1}{2}} M_{k_j}^{\frac{1}{2}} V_{j,k_j} = M_{k_j} M_{k_j}^\dagger V_{j,k_j} = V_{j,k_j}$, where the last equality follows since V_{j,k_j} is orthogonal to the nullspace of M_{k_j} . We may now again revisit the expression for the gradient:

$$\nabla \ell(\theta^*) = -\frac{1}{m} \sum_{j=1}^m E_{j,k_j} V_{j,k_j} = -\frac{1}{m} \sum_{j=1}^m E_{j,k_j} M_{k_j}^{1/2} V_{j,k_j} := -\frac{1}{m} X(\mathcal{D})^T V,$$

where we have defined $X(\mathcal{D}) \in \mathbb{R}^{\Omega_{\mathcal{D}} \times n}$ as the vertical concatenation of all the $E_{j,k_j} M_{k_j}^{1/2}$. As an aside, $X(\mathcal{D})$ is the design matrix in the terminology of generalized linear models (and is thus named fancifully).

Now, consider that

$$\nabla \ell(\theta^*)^T \hat{L}^\dagger \nabla \ell(\theta^*) = \frac{1}{m^2} V^T X(\mathcal{D}) \hat{L}^\dagger X(\mathcal{D})^T V.$$

We apply Lemma 3, a modified Hanson-Wright-type tail bound for random quadratic forms. This lemma follows from simpler technologies (largely Hoeffding's inequality) given that the random variables are bounded while also carefully handling the block structure of the problem.

In the language of Lemma 3 we have V_{j,k_j} playing the role of $x^{(j)}$ and $\Sigma_{\mathcal{D}} := \frac{1}{m^2} X(\mathcal{D}) \hat{L}^\dagger X(\mathcal{D})^T$ plays the role of A . The invocation of this lemma is possible because V_{j,k_j} is mean zero, $\|V_{j,k_j}\|_2 \leq \sqrt{2}$, and because $\Sigma_{\mathcal{D}}$ is positive semi-definite. We sweep $K^4 = 4$ from the lemma statement into the constant c of the right hand side. Stating the result of Lemma 3 we have, for all $t > 0$,

$$\mathbb{P}(V^T \Sigma_{\mathcal{D}} V - \sum_{i=1}^m \lambda_{\max}(\Sigma_{\mathcal{D}}^{(i,i)}) \mathbb{E}[V^{(i)T} V^{(i)}] \geq t) \leq 2 \exp\left(-c \frac{t^2}{\sum_{i,j} \sigma_{\max}(\Sigma_{\mathcal{D}}^{(i,j)})^2}\right). \quad (7)$$

We note that

$$\begin{aligned}
\sigma_{\max}(\Sigma_{\mathcal{D}}^{(i,j)}) &= \sigma_{\max}\left(\frac{1}{m^2} M_{k_i}^{1/2} E_{i,k_i}^T \hat{L}^\dagger E_{j,k_j} M_{k_j}^{1/2}\right) \\
&= \frac{1}{m^2} y_{\max}^T M_{k_i}^{1/2} E_{i,k_i}^T \hat{L}^\dagger E_{j,k_j} M_{k_j}^{1/2} z_{\max} \\
&\leq \frac{1}{m^2} \lambda_{\max}(\hat{L}^\dagger) \|E_{i,k_i} M_{k_i}^{1/2} y_{\max}\|_2 \|E_{j,k_j} M_{k_j}^{1/2} z_{\max}\|_2 \\
&= \frac{1}{m^2} \lambda_{\max}(\hat{L}^\dagger) \|M_{k_i}^{1/2} y_{\max}\|_2 \|M_{k_j}^{1/2} z_{\max}\|_2 \\
&\leq \frac{1}{m^2 \lambda_2(\hat{L})},
\end{aligned}$$

for all i, j , where the second line follows because y_{\max} and z_{\max} are the maximum left and right singular vectors of unit norm, the third line from an upper bound on quadratic forms, the fourth because E_{i,k_i} is a re-indexing that does not change Euclidean norm, and the final one because centering matrices can only lower the norm of a vector. This result has two consequences:

$$\lambda_{\max}(\Sigma_{\mathcal{D}}^{(i,i)}) = \sigma_{\max}(\Sigma_{\mathcal{D}}^{(i,i)}) \leq \frac{1}{m^2 \lambda_2(\hat{L})},$$

and

$$\sum_{i,j} \sigma_{\max}(\Sigma_{\mathcal{D}}^{(i,j)})^2 \leq \frac{m^2}{\lambda_2(\hat{L})^2 m^4} = \frac{1}{\lambda_2(\hat{L})^2 m^2}.$$

Now, noting that the norm of V_{i,k_i} is bounded (thus $\mathbb{E}[V^{(i)T} V^{(i)}] \leq 2$), and substituting in the relevant values into Equation (7), we have for all $t > 0$:

$$\mathbb{P}\left(\nabla \ell(\theta^*)^T \hat{L}^\dagger \nabla \ell(\theta^*) - \frac{2}{m \lambda_2(\hat{L})} \geq t\right) \leq 2 \exp\left(-cm^2 \lambda_2(\hat{L})^2 t^2\right).$$

A variable substitution and simple algebra transforms this expression to

$$\mathbb{P}\left[\nabla \ell(\theta^*)^T \hat{L}^\dagger \nabla \ell(\theta^*) \geq c_2 \frac{t}{m \lambda_2(\hat{L})}\right] \leq e^{-t} \quad \text{for all } t > 1,$$

where c_2 is an absolute constant. We may then make the same substitutions as before with expected risk, to obtain,

$$\mathbb{P}\left[\|\hat{\theta}_{\text{MLE}}(\mathcal{D}) - \theta^*\|_2^2 > c_2 \frac{tk_{\max}^2}{m \lambda_2(L)^2 \beta_{k_{\max}}^2}\right] \leq e^{-t} \quad \text{for all } t > 1.$$

Making the appropriate substitution with $c_{B,k_{\max}}$, we retrieve the second theorem statement, for another absolute constant c .

$$\mathbb{P}\left[\|\hat{\theta}_{\text{MLE}}(\mathcal{D}) - \theta^*\|_2^2 \geq c_{B,k_{\max}} \frac{t}{m \lambda_2(L)^2}\right] \leq e^{-t} \quad \text{for all } t > 1.$$

Integrating the above tail bound leads to a similar bound on the expected risk (same parametric rates), albeit with a less sharp constants due to the added presence of c . \square

C.2 Proof of Theorem 2

Theorem 2. Let $\mathcal{R} = \sigma_1, \dots, \sigma_\ell \sim PL(\theta^*)$ be a dataset of full rankings generated from a Plackett-Luce model with true parameter $\theta^* \in \Theta_B = \{\theta \in \mathbb{R}^n : \|\theta\|_\infty \leq B, \mathbf{1}^T \theta = 0\}$ and let $\hat{\theta}_{\text{MLE}}$ denote the maximum likelihood solution. Assume that $\ell > 4 \log(\sqrt{\alpha_B n}) / \alpha_B^2$ where α_B is a constant that only depends on B . Then for $t > 1$ and any dataset \mathcal{R} generated by the PL model,

$$\mathbb{P}\left[\|\hat{\theta}_{\text{MLE}}(\mathcal{R}) - \theta^*\|_2^2 \geq c_B'' \frac{n}{\ell} t\right] \leq e^{-t} + n^2 \exp(-\ell \alpha_B^2) \exp\left(\frac{-t}{\alpha_B^2 n^2}\right),$$

where c'_B is a constant that depends on B . For the expected risk,

$$\mathbb{E} \left[\left\| \hat{\theta}_{MLE}(\mathcal{R}) - \theta^* \right\|_2^2 \right] \leq c'_B \frac{n^3}{\ell} \mathbb{E} \left[\frac{1}{\lambda_2(L)^2} \right] \leq c_B \frac{n}{\ell},$$

where $c'_B = 4 \exp(4B)$ and $c_B = 8 \exp(4B)/\alpha_B^2$.

Proof.

As with the proof for the MNL model, we first give an expected risk bound, and then proceed to carry that technology forward to give a tail bound. The tail bound will again integrate out to give an expected risk bound with the same parametric rates as the direct proof, albeit with weaker constants.

Expected risk bound. We exploit the fact that the PL likelihood is the MNL likelihood with $\ell(n-1)$ choices. We thus begin with the result of Theorem 1, unpacking $c_{B,k_{\max}}$ and applying $k_{\max} = n$ and $m = (n-1)\ell$:

$$\mathbb{E}[\|\hat{\theta}_{MLE} - \theta^*\|_2^2] \leq \frac{2 \exp(4B)n^4}{\ell(n-1)\lambda_2(L)^2}.$$

We remind the reader that since the choice sets are assumed fixed in the proof of 1, the expectation above is taken *only* over the choices, conditioned on the choice sets, and not over the choice sets themselves. Since we are now working with rankings, there is randomness over the choice sets themselves. The randomness manifests itself as an expectation *conditional* on the choice sets on the left hand side and in the randomness of $\lambda_2(L)$ on the right hand side. We may rewrite the expression to reflect this fact:

$$\mathbb{E}[\|\hat{\theta}_{MLE} - \theta^*\|_2^2 \mid S_1, S_2, \dots, S_{\ell(n-1)}] \leq \frac{2 \exp(4B)n^4}{\ell(n-1)\lambda_2(L)^2}.$$

and make progress towards the theorem statement, by take expectations over the choice sets S_i on both sides and apply the law of iterated expectations:

$$\begin{aligned} \mathbb{E}[\|\hat{\theta}_{MLE} - \theta^*\|_2^2] &= \mathbb{E}[\mathbb{E}[\|\hat{\theta}_{MLE} - \theta^*\|_2^2 \mid S_1, S_2, \dots, S_{\ell(n-1)}]] \\ &\leq \mathbb{E} \left[\frac{2 \exp(4B)n^4}{\ell(n-1)\lambda_2(L)^2} \right] \\ &= 4 \exp(4B) \frac{n^3}{\ell} \mathbb{E} \left[\frac{1}{\lambda_2(L)^2} \right], \end{aligned}$$

where in the last line we have bounded $n/(n-1)$ by 2. We have reached the intermediate form of the expected risk bound theorem statement.

What now remains is upper bounding $\mathbb{E}[1/\lambda_2(L)^2]$. Recall that L is the Laplacian of a weighted comparison graph. A crude bound for $\lambda_2(L)$ comes from noting that choice set \mathcal{X} appears at least ℓ times, each time adding $\frac{1}{\ell(n-1)}(nI - \mathbf{1}\mathbf{1}^T)$ to the Laplacian so that we get

$$\lambda_2(L) \geq \lambda_2\left(\frac{1}{n-1}(nI - \mathbf{1}\mathbf{1}^T)\right) = \lambda_2\left(\frac{n}{n-1}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\right) = \frac{n}{n-1} \geq 1 \quad (8)$$

as $I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is simply the centering matrix, and where the first inequality follows from properties of sums of PSD matrices [11][See pg. 128, Corollary (4.2)].

We will use a more sophisticated bound that comes from a careful study of the graph that the random Plackett-Luce Laplacian represents. We have packaged this analysis inside Lemma 4, which says that

$$\alpha_B n \leq \lambda_2(L) \quad \text{with probability at least } 1 - n^2 \exp(-\alpha_B^2 \ell), \quad (9)$$

where $\alpha_B = 1/(4(1 + 2e^{3B}))$.

We can use Lemma 4 to upper bound the expectation of $1/\lambda_2(L)^2$ as follows:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{\lambda_2(L)^2}\right] &= \mathbb{E}\left[\frac{1}{\lambda_2(L)^2} \mid \frac{1}{\lambda_2(L)} \leq \frac{1}{\alpha_B n}\right] \mathbb{P}\left\{\frac{1}{\lambda_2(L)} \leq \frac{1}{\alpha_B n}\right\} \\ &\quad + \mathbb{E}\left[\frac{1}{\lambda_2(L)^2} \mid \frac{1}{\lambda_2(L)} > \frac{1}{\alpha_B n}\right] \mathbb{P}\left\{\frac{1}{\lambda_2(L)} > \frac{1}{\alpha_B n}\right\} \\ &\leq \frac{1}{\alpha_B^2 n^2} + \mathbb{P}\left\{\frac{1}{\lambda_2(L)} > \frac{1}{\alpha_B n}\right\} \\ &\leq \frac{1}{\alpha_B^2 n^2} + n^2 \exp(-\alpha_B^2 \ell),\end{aligned}$$

where the first inequality follows from applying the bound of $1/(\alpha_B^2 n^2)$ to the first expectation and a bound of 1 to the second expectation (which comes from Equation (8)). The second inequality follows from applying the tail bound. Now, we need that $\ell > 4 \log(\sqrt{\alpha_B n})/\alpha_B^2$ to ensure that

$$\mathbb{E}\left[\frac{1}{\lambda_2(L)^2}\right] \leq \frac{2}{\alpha_B^2 n^2}.$$

We can now circle back to the start of the proof to apply this result:

$$\mathbb{E}[\|\hat{\theta}_{\text{MLE}} - \theta^*\|_2^2] \leq 4 \exp(4B) \frac{n^3}{\ell} \mathbb{E}\left[\frac{1}{\lambda_2(L)^2}\right] \leq \frac{8 \exp(4B) n}{\alpha_B^2 \ell},$$

so long as $\ell > 4 \log(\sqrt{\alpha_B n})/\alpha_B^2$. Defining c_B as

$$c_B := \frac{8 \exp(4B)}{\alpha_B^2},$$

we arrive at the expected risk bound in the theorem statement.

Tail bound. Our tail bound proof proceeds very similarly to that of the risk bound. To start, we again exploit the fact that the PL likelihood is the MNL likelihood with $\ell(n-1)$ choices. We thus begin with the result of Theorem 1, unpacking $c_{B, k_{\max}}$ and applying $k_{\max} = n$ and $m = (n-1)\ell$:

$$\mathbb{P}\left[\|\hat{\theta}_{\text{MLE}}(\mathcal{D}) - \theta^*\|_2^2 > c_2 \frac{tn^3 \exp(4B)}{\ell \lambda_2(L)^2}\right] \leq e^{-t} \quad \text{for all } t > 1,$$

where c_2 is some absolute constant (note we have lower bounded $n/(n-1)$ by 2). Like before, we remind the reader that because the choice sets are assumed fixed in the proof of Theorem 1, the probabilistic statement *only* accounts for the randomness in the choices. Since we now are working with rankings, we must additionally account for the randomness over the choice sets, and the above statement is more clearly stated as a conditional probability over the sets:

$$\mathbb{P}\left[\|\hat{\theta}_{\text{MLE}}(\mathcal{D}) - \theta^*\|_2^2 > c_2 \frac{tn^3 \exp(4B)}{\ell \lambda_2(L)^2} \mid S_1, S_2, \dots, S_{\ell(n-1)}\right] \leq e^{-t} \quad \text{for all } t > 1.$$

In order to obtain an unconditional statement, we now account for the choice sets S . Notice first that the expression depends only on the choice sets via the matrix L (and more specifically its second smallest eigenvalue), and so:

$$\mathbb{P}\left[\|\hat{\theta}_{\text{MLE}}(\mathcal{D}) - \theta^*\|_2^2 > c_2 \frac{tn^3 \exp(4B)}{\ell \lambda_2(L)^2} \mid \lambda_2(L)\right] = \mathbb{P}\left[\|\hat{\theta}_{\text{MLE}}(\mathcal{D}) - \theta^*\|_2^2 > c_2 \frac{tn^3 \exp(4B)}{\ell \lambda_2(L)^2} \mid S_1, \dots, S_{\ell(n-1)}\right]$$

We may additionally perform a change of variables, and rewrite the tail bound as

$$\mathbb{P}\left[\|\hat{\theta}_{\text{MLE}}(\mathcal{D}) - \theta^*\|_2^2 > c_2 \frac{\delta n^3 \exp(4B)}{\ell} \mid \lambda_2(L)\right] \leq e^{-\delta \lambda_2(L)^2} \quad \text{for all } \delta > 1/\lambda_2(L)^2. \quad (10)$$

We can now control $\lambda_2(L)$ using the same steps taken in the expected risk bound. Using Lemma 4, also stated above in (9), we can integrate Equation 10 over $\lambda_2(L)$:

$$\mathbb{E}_{\lambda_2(L)} \left[\mathbb{P} \left[\|\hat{\theta}_{\text{MLE}}(\mathcal{D}) - \theta^*\|_2^2 > c_2 \frac{\delta n^3 \exp(4B)}{\ell} \middle| \lambda_2(L) \right] \right] \leq \mathbb{E}_{\lambda_2(L)} \left[\frac{1}{\exp(\delta \lambda_2(L)^2)} \right].$$

We may use the same trick to upper bound the right hand side just as we did the expectation of $1/\lambda_2(L)^2$ in the expected risk portion of our proof:

$$\begin{aligned} \mathbb{E}_{\lambda_2(L)} \left[\frac{1}{\exp(\delta \lambda_2(L)^2)} \right] &= \mathbb{E} \left[\frac{1}{\exp(\delta \lambda_2(L)^2)} \middle| \frac{1}{\lambda_2(L)} \leq \frac{1}{\alpha_B n} \right] \mathbb{P} \left\{ \frac{1}{\lambda_2(L)} \leq \frac{1}{\alpha_B n} \right\} \\ &\quad + \mathbb{E} \left[\frac{1}{\exp(\delta \lambda_2(L)^2)} \middle| \frac{1}{\lambda_2(L)} > \frac{1}{\alpha_B n} \right] \mathbb{P} \left\{ \frac{1}{\lambda_2(L)} > \frac{1}{\alpha_B n} \right\} \\ &\leq \frac{1}{\exp(\delta \alpha_B^2 n^2)} + \frac{1}{\exp(\delta)} \mathbb{P} \left\{ \frac{1}{\lambda_2(L)} > \frac{1}{\alpha_B n} \right\} \\ &\leq \exp(-\delta \alpha_B^2 n^2) + \exp(-\delta) n^2 \exp(-\alpha_B^2 \ell), \end{aligned}$$

where the first inequality follows from applying the bound of $1/\exp(\alpha_B^2 n^2)$ to the first expectation and a bound of $1/\exp(\delta)$ to the second expectation (which follows from Equation (8)). The second inequality follows from applying the tail bounds. Returning to the tail expression we have:

$$\mathbb{P} \left[\|\hat{\theta}_{\text{MLE}}(\mathcal{D}) - \theta^*\|_2^2 > c_2 \frac{\delta n^3 \exp(4B)}{\ell} \right] \leq \exp(-\delta \alpha_B^2 n^2) + \exp(-\delta) \exp(-\ell \alpha_B^2).$$

Setting $t = \delta(\alpha_B^2 n^2)$, we obtain,

$$\mathbb{P} \left[\|\hat{\theta}_{\text{MLE}}(\mathcal{D}) - \theta^*\|_2^2 > c_2 \frac{tn^3 \exp(4B)}{\ell \alpha_B^2 n^2} \right] \leq \exp(-t) + \exp\left(\frac{-t}{\alpha_B^2 n^2}\right) n^2 \exp(-\ell \alpha_B^2) \quad \text{for all } t > 1.$$

Canceling terms we have,

$$\mathbb{P} \left[\|\hat{\theta}_{\text{MLE}}(\mathcal{D}) - \theta^*\|_2^2 > c_2 \frac{\exp(4B)}{\alpha_B^2} \frac{n}{\ell} t \right] \leq e^{-t} + n^2 \exp(-\ell \alpha_B^2) \exp\left(\frac{-t}{\alpha_B^2 n^2}\right) \quad \text{for all } t > 1.$$

Defining c_B as

$$c_B := c_2 \frac{\exp(4B)}{\alpha_B^2},$$

we arrive at the tail bound in the theorem statement.

Integrating the above tail bound leads to a similar bound on the expected risk as the direct proof, albeit with less sharp constants.

C.3 Proof of Theorem 3

Theorem 3 *Let $\mathcal{R} = \sigma_1, \dots, \sigma_\ell \sim \text{CRS}(u^*)$ be a dataset of full rankings generated from the full CRS model with true parameter $u^* \in \Theta_B = \{u \in \mathbb{R}^{n(n-1)} : u = [u_1^T, \dots, u_n^T]^T; u_i \in \mathbb{R}^{n-1}, \|u_i\|_1 \leq B, \forall i; \mathbf{1}^T u = 0\}$ and let \hat{u}_{MLE} denote the maximum likelihood solution. Assuming that $\ell > 8ne^{2B} \log(8ne^{2B})^2$, and $t > 1$:*

$$\mathbb{P} \left[\|\hat{u}_{\text{MLE}}(\mathcal{D}) - u^*\|_2^2 > \frac{c_B'' n^4}{\ell \delta_n} t \right] \leq e^{-t} + n \exp\left(-t \min\left\{\frac{c_B'' n^3}{\ell \delta_n}, 1\right\}\right) e^{-\ell/(8ne^{2B})},$$

where c_B', c_B'' is a constant that depends on B , and δ_n is a strictly positive quantity that depends only on n . For the expected risk,

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

where c_B', c_B are constants that depend only on B .

Proof of Theorem 3.

As with the proof for the PL model, we first give an expected risk bound, and then proceed to carry that technology forward to give a tail bound. The tail bound will again integrate out to give an expected risk bound with the same parametric rates as the direct proof, albeit with weaker constants.

Expected risk bound. Our proof leverages the fact that the CRS likelihood is the CDM likelihood with $\ell(n-1)$ choices, just as our analysis of the PL model leveraged the relationship between the PL and MNL likelihoods. We thus begin with the result of Lemma 6, our adaptation of an existing CDM risk bound. Unpacking $c_{B,k_{\max}}$ and applying $k_{\max} = n$ and $m = (n-1)\ell$:

$$\mathbb{E} [\|\hat{u}_{MLE}(\mathcal{D}) - u^*\|_2^2] \leq \frac{2n(n-1)}{m\lambda_2(L)\beta_{k_{\max}}^2} \leq \frac{n^3(n-1)2\exp(4B)}{\ell(n-1)\lambda_2(L)} = \frac{n^3 2\exp(4B)}{\ell\lambda_2(L)}.$$

Working with rankings, we must handle the randomness over the choice sets themselves. The randomness manifests itself as an expectation *conditional* on the choice sets on the left hand side and in the randomness of $\lambda_2(L)$ on the right hand side. We may rewrite the expression in Lemma 6 to reflect this fact:

$$\mathbb{E} [\|\hat{u}_{MLE}(\mathcal{D}) - u^*\|_2^2 | S_1, S_2, \dots, S_{\ell(n-1)}] \leq \frac{n^3 2\exp(4B)}{\ell\lambda_2(L)}.$$

In Theorem 2 we proceeded to use a law of iterated expectations and then bound $\lambda_2(L)$. For the PL model, $\lambda_2(L)$ was always at least 1, and with high probability much larger. For the CRS model, however, $\lambda_2(L)$ can sometimes be 0. This result holds because non-trivial conditions on the choice set structure are required for the CDM model's *identifiability*. We refer the reader to [48] for more details. In our ranking setting, these conditions are never met with one ranking's worth of choices, and hence results in the CRS model parameters being underdetermined. As an aside, this claim should not be confused with the CRS model being unidentifiable. In fact, the CRS model with true parameter $u^* \in \Theta_B$ is *always* identifiable. That is, it is always determined with sufficiently many rankings, as we will later see.

Nevertheless, $1/\lambda_2(L)$ is difficult to meaningfully upper bound directly since $\lambda_2(L)$ can sometimes be 0. However, when the model is not identifiable the risk under our assumptions is *not* infinity. Because the true parameters live in a Θ_B , a norm bounded space, we may bound the error of any guess \hat{u} in that space:

$$\|u^* - \hat{u}\|_2^2 = \sum_i \|u_i^* - \hat{u}_i\|_2^2 \leq \sum_i \|u_i^* - \hat{u}_i\|_1^2 \leq \sum_i (\|u_i^*\|_1 + \|\hat{u}_i\|_1)^2 \leq 4B^2 n.$$

We may thus bound the expected risk as

$$\mathbb{E}[\|\hat{u}_{MLE} - u^*\|_2^2 | S_1, S_2, \dots, S_{\ell(n-1)}] \leq \min \left\{ \frac{2\exp(4B)n^3}{\ell\lambda_2(L)}, 4B^2 n \right\}.$$

and use a bound of $4B^2 n$ whenever $\lambda_2(L) = 0$. Now, we work towards the theorem statement by take expectations over the choice sets S_i on both sides and apply the law of iterated expectations:

$$\begin{aligned} \mathbb{E}[\|\hat{u}_{MLE} - u^*\|_2^2] &= \mathbb{E}[\mathbb{E}[\|\hat{u}_{MLE} - u^*\|_2^2 | S_1, S_2, \dots, S_{\ell(n-1)}]] \\ &\leq \mathbb{E} \left[\min \left\{ \frac{2\exp(4B)n^3}{\ell\lambda_2(L)}, 4B^2 n \right\} \right]. \end{aligned}$$

The above bound is the intermediate bound of the theorem statement, where $c'_B = 2\exp(4B)$.

We now use Lemma 5, which says that

$$\frac{\delta_n}{(n-1)e^{2B}} \leq \lambda_2(L) \quad \text{with probability at least } 1 - n \exp\left(-\frac{\ell}{8ne^{2B}}\right), \quad (11)$$

where $0 < \delta_n \leq \frac{1}{2n}$ is a fixed value that depends only on n .

We can use Lemma 5 to upper bound the expectation of the risk as follows:

$$\begin{aligned}
\mathbb{E} \left[\min \left\{ \frac{2 \exp(4B)n^3}{\ell \lambda_2(L)}, 4B^2 n \right\} \right] &= \mathbb{E} \left[\min \left\{ \frac{2 \exp(4B)n^3}{\ell \lambda_2(L)}, 4B^2 n \right\} \middle| \frac{1}{\lambda_2(L)} \leq \frac{(n-1)e^{2B}}{\delta_n} \right] \\
&\quad \times \mathbb{P} \left\{ \frac{1}{\lambda_2(L)} \leq \frac{(n-1)e^{2B}}{\delta_n} \right\} \\
&\quad + \mathbb{E} \left[\min \left\{ \frac{2 \exp(4B)n^3}{\ell \lambda_2(L)}, 4B^2 n \right\} \middle| \frac{1}{\lambda_2(L)} > \frac{(n-1)e^{2B}}{\delta_n} \right] \\
&\quad \times \mathbb{P} \left\{ \frac{1}{\lambda_2(L)} > \frac{(n-1)e^{2B}}{\delta_n} \right\} \\
&\leq \frac{2 \exp(4B)n^3}{\ell} \frac{(n-1)e^{2B}}{\delta_n} + 4B^2 n \mathbb{P} \left\{ \frac{1}{\lambda_2(L)} > \frac{(n-1)e^{2B}}{\delta_n} \right\} \\
&\leq \frac{2 \exp(6B)n^4}{\ell \delta_n} + 4B^2 n^2 \exp \left(-\frac{\ell}{8ne^{2B}} \right),
\end{aligned}$$

where the first inequality follows from selecting the first value in the min and applying the bound on $\lambda_2(L)$ to the first expectation; and a bound of $4B^2 n$ to the second expectation. The second inequality follows from applying the tail bound. Now, as long as $\ell > 8ne^{2B} \log(8ne^{2B})^2$, we may upper bound the second term as follows

$$4B^2 n^2 \exp \left(-\frac{\ell}{8ne^{2B}} \right) \leq \frac{4B^2 n^2}{\ell},$$

and so

$$\mathbb{E}[\|\hat{u}_{MLE} - u^*\|_2^2] \leq \mathbb{E} \left[\min \left\{ \frac{2 \exp(4B)n^3}{\ell \lambda_2(L)}, 4B^2 n \right\} \right] \leq \frac{2 \exp(6B)n^4}{\ell \delta_n} + \frac{4B^2 n^2}{\ell} \leq \frac{3 \exp(6B)n^4}{\ell \delta_n},$$

so long as $\ell > 8ne^{2B} \log(8ne^{2B})^2$, where the final inequality follows because $\delta_n \leq \frac{1}{2n}$ per Lemma 5. Define $c_B := 3 \exp(6B)$ to arrive at the theorem statement.

Tail bound. Our tail bound proof proceeds very similarly to that of the risk bound. To start, we again exploit the fact that the CRS likelihood is the CDM likelihood with $\ell(n-1)$ choices. We thus begin again with the result of Lemma 6, unpacking $c_{B, k_{\max}}$ and applying $k_{\max} = n$ and $m = (n-1)\ell$:

$$\mathbb{P} \left[\|\hat{u}_{MLE}(\mathcal{D}) - u^*\|_2^2 > c_2 \frac{tn^3 \exp(4B)}{\ell \lambda_2(L)} \right] \leq e^{-t} \quad \text{for all } t > 1.$$

where c_2 is some absolute constant. Like before, we remind the reader that because the choice sets are assumed fixed in the proof of Lemma 6, the probabilistic statement *only* accounts for the randomness in the choices. Since we now are working with rankings, we must additionally account for the randomness over the choice sets, and the above statement is more clearly stated as a conditional probability over the sets:

$$\mathbb{P} \left[\|\hat{u}_{MLE}(\mathcal{D}) - u^*\|_2^2 > c_2 \frac{tn^3 \exp(4B)}{\ell \lambda_2(L)} \middle| S_1, S_2, \dots, S_{\ell(n-1)} \right] \leq e^{-t} \quad \text{for all } t > 1.$$

In order to obtain an unconditional statement, we now account for the choice sets S . Notice first that the expression depends only on the choice sets via the matrix L (and more specifically its second smallest eigenvalue), and so:

$$\mathbb{P} \left[\|\hat{u}_{MLE}(\mathcal{D}) - u^*\|_2^2 > c_2 \frac{tn^3 \exp(4B)}{\ell \lambda_2(L)} \middle| \lambda_2(L) \right] = \mathbb{P} \left[\|\hat{u}_{MLE}(\mathcal{D}) - u^*\|_2^2 > c_2 \frac{tn^3 \exp(4B)}{\ell \lambda_2(L)} \middle| S_1, \dots, S_{\ell(n-1)} \right]$$

Now, note additionally that

$$\mathbb{P} \left[\|\hat{u}_{MLE}(\mathcal{D}) - u^*\|_2^2 > t4B^2 n \middle| \lambda_2(L) \right] = 0 \leq e^{-t} \quad \text{for all } t \geq 1.$$

and so

$$\mathbb{P}\left[\|\hat{u}_{\text{MLE}}(\mathcal{D}) - u^*\|_2^2 > t \frac{c_2 n^3 \exp(4B)}{\ell} \min\left\{\frac{1}{\lambda_2(L)}, \frac{4B^2 \ell}{c_2 n^2 \exp(4B)}\right\} \middle| \lambda_2(L)\right] \leq e^{-t} \text{ for all } t \geq 0.$$

We may additionally perform a change of variables, and rewrite the tail bound as

$$\mathbb{P}\left[\|\hat{u}_{\text{MLE}}(\mathcal{D}) - u^*\|_2^2 > c_2 \frac{\delta n^3 \exp(4B)}{\ell} \middle| \lambda_2(L)\right] \leq e^{-\delta \max\left\{\lambda_2(L), \frac{c_2 n^2 \exp(4B)}{4B^2 \ell}\right\}}. \quad (12)$$

$$\text{for all } \delta > \min\left\{\frac{1}{\lambda_2(L)}, \frac{4B^2 \ell}{c_2 n^2 \exp(4B)}\right\}$$

. We can now control $\lambda_2(L)$ using the same steps taken in the expected risk bound. Using Lemma 5, also stated above in (11), we can integrate Equation (12) over $\lambda_2(L)$:

$$\mathbb{E}_{\lambda_2(L)} \left[\mathbb{P}\left[\|\hat{u}_{\text{MLE}}(\mathcal{D}) - u^*\|_2^2 > c_2 \frac{\delta n^3 \exp(4B)}{\ell} \middle| \lambda_2(L)\right] \right] \leq \mathbb{E}_{\lambda_2(L)} \left[\frac{1}{\exp(\delta \max\left\{\lambda_2(L), \frac{c_2 n^2 \exp(4B)}{4B^2 \ell}\right\})} \right].$$

We may use the same trick to upper bound the right hand side just as we did the expectation in the expected risk portion of our proof:

$$\begin{aligned} \mathbb{E}_{\lambda_2(L)} \left[\frac{1}{\exp(\delta \max\left\{\lambda_2(L), \frac{c_2 n^2 \exp(4B)}{4B^2 \ell}\right\})} \right] &= \mathbb{E} \left[\frac{1}{\exp(\delta \max\left\{\lambda_2(L), \frac{c_2 n^2 \exp(4B)}{4B^2 \ell}\right\})} \middle| \frac{1}{\lambda_2(L)} \leq \frac{(n-1)e^{2B}}{\delta_n} \right] \\ &\quad \times \mathbb{P}\left\{ \frac{1}{\lambda_2(L)} \leq \frac{(n-1)e^{2B}}{\delta_n} \right\} \\ &\quad + \mathbb{E} \left[\frac{1}{\exp(\delta \max\left\{\lambda_2(L), \frac{c_2 n^2 \exp(4B)}{4B^2 \ell}\right\})} \middle| \frac{1}{\lambda_2(L)} > \frac{(n-1)e^{2B}}{\delta_n} \right] \\ &\quad \times \mathbb{P}\left\{ \frac{1}{\lambda_2(L)} > \frac{(n-1)e^{2B}}{\delta_n} \right\} \\ &\leq \frac{1}{\exp(\delta \max\left\{\frac{\delta_n}{(n-1)e^{2B}}, \frac{c_2 n^2 \exp(4B)}{4B^2 \ell}\right\})} \\ &\quad + \frac{1}{\exp(\delta \frac{c_2 n^2 \exp(4B)}{4B^2 \ell})} \mathbb{P}\left\{ \frac{1}{\lambda_2(L)} > \frac{(n-1)e^{2B}}{\delta_n} \right\} \\ &\leq \exp(-\delta \max\left\{\frac{\delta_n}{(n-1)e^{2B}}, \frac{c_2 n^2 \exp(4B)}{4B^2 \ell}\right\}) \\ &\quad + \exp(-\delta \frac{c_2 n^2 \exp(4B)}{4B^2 \ell}) n \exp\left(-\frac{\ell}{8ne^{2B}}\right), \end{aligned}$$

where the first inequality follows from applying the bound on $1/\lambda_2(L)$ to the first expectation and a bound of the second term in the max to the second expectation. The second inequality follows from applying the tail bound probability from Lemma 5. Returning to the tail expression we have:

$$\begin{aligned} \mathbb{P}\left[\|\hat{u}_{\text{MLE}}(\mathcal{D}) - u^*\|_2^2 > c_2 \frac{\delta n^3 \exp(4B)}{\ell}\right] &\leq \exp(-\delta \max\left\{\frac{\delta_n}{(n-1)e^{2B}}, \frac{c_2 n^2 \exp(4B)}{4B^2 \ell}\right\}) \\ &\quad + \exp(-\delta \frac{c_2 n^2 \exp(4B)}{4B^2 \ell}) n \exp\left(-\frac{\ell}{8ne^{2B}}\right). \end{aligned}$$

Setting

$$t = \delta \max\left\{\frac{\delta_n}{(n-1)e^{2B}}, \frac{c_2 n^2 \exp(4B)}{4B^2 \ell}\right\},$$

we obtain

$$\mathbb{P} \left[\|\hat{u}_{\text{MLE}}(\mathcal{D}) - u^*\|_2^2 > t \min \left\{ \frac{c_2 \exp(6B)n^3(n-1)}{\ell\delta_n}, 4B^2n \right\} \right] \leq \exp(-t) + \exp \left(-t \min \left\{ \frac{c_2 \exp(6B)n^2(n-1)}{4B^2\ell\delta_n}, 1 \right\} \right) n \exp \left(-\frac{\ell}{8ne^{2B}} \right) \quad \text{for all } t > 1.$$

Canceling terms we have, for all $t > 1$,

$$\mathbb{P} \left[\|\hat{u}_{\text{MLE}}(\mathcal{D}) - u^*\|_2^2 > t \frac{c_2 \exp(6B)n^4}{\ell\delta_n} \right] \leq e^{-t} + n \exp \left(-t \min \left\{ \frac{c_2 \exp(6B)n^3}{4B^2\ell\delta_n}, 1 \right\} \right) e^{-\ell/(8ne^{2B})}.$$

Defining c_B, c'_B as

$$c_B := c_2 \frac{\exp(6B)}{4B^2}, c'_B := c_2 \exp(6B),$$

we arrive at the tail bound in the theorem statement:

$$\mathbb{P} \left[\|\hat{u}_{\text{MLE}}(\mathcal{D}) - u^*\|_2^2 > \frac{c'_B n^4}{\ell\delta_n} t \right] \leq e^{-t} + n \exp \left(-t \min \left\{ \frac{c_B n^3}{\ell\delta_n}, 1 \right\} \right) e^{-\ell/(8ne^{2B})}.$$

Integrating the above tail bound leads to a similar bound on the expected risk as the direct proof, albeit with less sharp constants. \square

C.4 Auxiliary Lemmas

Lemma 1. For the collection of functions $F_k : \mathbb{R}^k \mapsto [0, 1]$, $\forall k \geq 2$ defined as

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

where $x \in [-B, B]^k$, we have that

$$\nabla^2(-\log(F_k(x)))\mathbf{1} = 0,$$

and

$$\nabla^2(-\log(F_k(x))) \succeq \frac{1}{k \exp(2B)} \left(I - \frac{1}{k} \mathbf{1}\mathbf{1}^T \right).$$

Proof. We first compute the Hessian as:

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

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

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

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

$$\nabla^2(-\log(F_k(x))) \succeq H_k = \beta_k \left(I - \frac{1}{k} \mathbf{1}\mathbf{1}^T \right), \quad (13)$$

where

$$\beta_k := \min_{x \in [-B, B]^k} \lambda_2(\nabla^2(-\log(F_k(x)))) \quad (14)$$

and it's clear that $\beta_k > 0$. The minimization is taken over $x \in [-B, B]^k$ since each x_i is simply an entry of the θ vector, each entry of which is in $[-B, B]$. We next show that

$$\beta_k \geq \frac{1}{k \exp(2B)},$$

to complete the result.

First, a definition:

$$p(x) := \frac{\exp(x)}{\langle \exp(x), \mathbf{1} \rangle}.$$

Evidently, $p(x) \in \Delta_k$, and since $x \in [-B, B]^k$, $p(x) \succ 0$. We may also write the Hessian as

$$\nabla^2(-\log(F_k(x))) = \text{diag}(p(x)) - p(x)p(x)^T.$$

In this format, we may now directly apply Theorem 1 from [10], which lower bounds the second eigenvalue of the Hessian by the minimum probability in $p(x)$. Thus,

$$\lambda_2(\nabla^2(-\log(F_k(x)))) \geq \min_i p(x)_i$$

A simple calculation reveals then that

$$\begin{aligned}\beta_k &= \min_{x \in [-B, B]^k} \lambda_2(\nabla^2(-\log(F_k(x)))) \geq \min_{x \in [-B, B]^k} \min_i p(x)_i \\ &= \frac{1}{1 + (k-1)\exp(2B)} \\ &\geq \frac{1}{k\exp(2B)},\end{aligned}$$

which completes the proof. \square

Lemma 2. For $\Sigma_{\mathcal{D}} := \frac{1}{m^2} X(\mathcal{D}) \hat{L}^\dagger X(\mathcal{D})^T$, where the constituent quantities are defined in the proof of Lemma 6, we have,

$$\text{tr}(\Sigma_{\mathcal{D}}) = \frac{d-1}{m}, \quad \|\Sigma_{\mathcal{D}}\|_F^2 = \frac{d-1}{m^2}.$$

Proof. Consider first that

$$\frac{1}{m} X(\mathcal{D})^T X(\mathcal{D}) = \frac{1}{m} \sum_{j=1}^m E_{j,k_j} M_{k_j}^{1/2} M_{k_j}^{1/2} E_{j,k_j}^T = \frac{1}{m} \sum_{j=1}^m E_{j,k_j} (I - \frac{1}{k_j} \mathbf{1}\mathbf{1}^T) E_{j,k_j}^T = \hat{L}.$$

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

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

That is, $\Sigma_{\mathcal{D}}$ is a positive semi-definite matrix with spectra corresponding to $d-1$ values equaling $\frac{1}{m}$, and the last equaling 0. The result about the trace immediately follows. The equality about the Frobenius norm comes from the observation that the Frobenius norm of a positive semi-definite matrix is the squared sum of its eigenvalues. \square

Lemma 3. Suppose we have a collection of mean zero independent random vectors $X^{(i)} \in \mathbb{R}^{k_i}$, $i = 1, \dots, m$, of bounded Euclidean norm $\|X^{(i)}\|_2 \leq K$ stacked together into a single vector $X \in \mathbb{R}^d$, where $d = \sum_{i=1}^m k_i$. Additionally suppose we have a real positive semidefinite matrix $A \in \mathbb{R}^{d \times d}$ and denote by $A^{(i,j)} \in \mathbb{R}^{k_i \times k_j}$ the submatrix of A whose rows align with the index position of $X^{(i)}$ in X and whose columns align with the index position of $X^{(j)}$ in X . Then, for every $t \geq 0$,

$$\mathbb{P}(X^T A X - \sum_{i=1}^m \lambda_{\max}(A^{(i,i)}) \mathbb{E}[X^{(i)T} X^{(i)}] \geq t) \leq 2 \exp\left(-c \frac{t^2}{K^4 \sum_{i,j} \sigma_{\max}(A^{(i,j)})^2}\right),$$

where $\lambda_{\max}(\cdot)$, $\sigma_{\max}(\cdot)$ refer to the maximum eigenvalue/singular value of a matrix, $\|\cdot\|_F$ and $\|\cdot\|_{op}$ respectively refer to the Frobenius and operator norm of a matrix, and $c > 0$ is an absolute positive constant.

Proof. We heavily reference preliminary concepts about sub-Gaussian random variables as they are described in [52]. We assume without loss of generality that $K = 1$ ($\|X^{(i)}\|_2 \leq 1$), since we may substitute X/K in place of any X to satisfy the assumption, and rearrange terms to produce the result for any K .

First, note that

$$\begin{aligned}
X^T A X - \sum_{i=1}^m \lambda_{\max}(A^{(i,i)}) \mathbb{E}[X^{(i)T} X^{(i)}] &= \sum_{i,j} X^{(i)T} A^{(i,j)} X^{(j)} - \sum_{i=1}^m \lambda_{\max}(A^{(i,i)}) \mathbb{E}[X^{(i)T} X^{(i)}] \\
&= \sum_i X^{(i)T} A^{(i,i)} X^{(i)} - \sum_{i=1}^m \lambda_{\max}(A^{(i,i)}) \mathbb{E}[X^{(i)T} X^{(i)}] \\
&\quad + \sum_{i,j \neq i} X^{(i)T} A^{(i,j)} X^{(j)} \\
&\leq \sum_{i=1}^m \lambda_{\max}(A^{(i,i)}) (X^{(i)T} X^{(i)} - \mathbb{E}[X^{(i)T} X^{(i)}]) \\
&\quad + \sum_{i,j \neq i} X^{(i)T} A^{(i,j)} X^{(j)}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{P}(X^T A X - \sum_{i=1}^m \lambda_{\max}(A^{(i,i)}) \mathbb{E}[X^{(i)T} X^{(i)}] \geq t) &\leq \mathbb{P}\left(\sum_{i=1}^m \lambda_{\max}(A^{(i,i)}) (X^{(i)T} X^{(i)} - \mathbb{E}[X^{(i)T} X^{(i)}]) \right. \\
&\quad \left. + \sum_{i,j \neq i} X^{(i)T} A^{(i,j)} X^{(j)} \geq t\right),
\end{aligned}$$

and we may upper bound the right hand side to obtain an upper bound on the left hand side. We will in fact individually bound from above an expression corresponding to the block diagonal entries

$$p_1 = \mathbb{P}\left(\sum_{i=1}^m \lambda_{\max}(A^{(i,i)}) (X^{(i)T} X^{(i)} - \mathbb{E}[X^{(i)T} X^{(i)}]) \geq \frac{t}{2}\right),$$

and an expression corresponding to the off block diagonal entries

$$p_2 = \mathbb{P}\left(\sum_{i \neq j} X^{(i)T} A^{(i,j)} X^{(j)} \geq \frac{t}{2}\right),$$

and use the union bound to obtain the desired result.

Before we proceed, we remind the reader that a random Z is sub-Gaussian if and only if

$$(\mathbb{E}|Z|^p)^{1/p} \leq K_2 \sqrt{p}, \quad \forall p \geq 1,$$

where K_2 is a positive constant. We define the sub-Gaussian norm of a random variable, $\|Z\|_{\psi_2}$, as the smallest K_2 that satisfies the above expression, i.e.,

$$\|Z\|_{\psi_2} = \sup_{p \geq 1} p^{-1/2} (\mathbb{E}|Z|^p)^{1/p}.$$

The sub-Gaussian norm recovers the familiar bound on the moment generating function for centered Z :

$$\mathbb{E}[\exp(tZ)] \leq \exp(Ct^2 \|Z\|_{\psi_2}), \quad \forall t,$$

where C is an absolute positive constant. Similarly, we have that a random variable W is sub-exponential if and only if

$$(\mathbb{E}|W|^p)^{1/p} \leq K_2 p, \quad \forall p \geq 1,$$

and similarly sub-exponential norm $\|W\|_{\psi_1}$ is defined as

$$\|W\|_{\psi_1} = \sup_{p \geq 1} p^{-1} (\mathbb{E}|W|^p)^{1/p}.$$

Part 1: The Block Diagonal Entries. Define $y_i := \|X^{(i)}\|_2^2$ and $a_i = \lambda_{\max}(A^{(i,i)})$, and let the vector $a \in \mathbb{R}^m$ consist of elements a_i . Recall that the norms of all $X^{(i)}$ are bounded by K by assumption and $K = 1$, WLOG. Thus all y_i are sub-Gaussian where $\|y_i\|_{\psi_2} \leq K^2 \leq 1$. Using this

notation we may then write the expression corresponding to the block diagonal entries in a more compact manner:

$$\sum_{i=1}^m \lambda_{\max}(A^{(i,i)})(X^{(i)T} X^{(i)} - \mathbb{E}[X^{(i)T} X^{(i)}]) = \sum_{i=1}^m a_i(y_i - \mathbb{E}[y_i]).$$

Note that centering does not change subgaussianity, and that $y_i^2 - \mathbb{E}[y_i^2]$ is a centered sub-gaussian random variable where

$$\|y_i - \mathbb{E}[y_i]\|_{\psi_2} \leq 2\|y_i\|_{\psi_2} \leq 2.$$

These inequalities follow from Remark 5.18 in [52]. Hence, we may apply a one-sided Hoeffding-type inequality (Proposition 5.10 in [52]) to state that

$$\mathbb{P}\left(\sum_{i=1}^m a_i(y_i - \mathbb{E}[y_i]) \geq t\right) \leq \exp\left[-c_1\left(\frac{t^2}{4\|a\|_2^2}\right)\right],$$

where c_1 is some absolute positive constant.

Now, examining $\|a\|_2^2$ we see that

$$\|a\|_2^2 = \sum_{i=1}^m \lambda_{\max}(A^{(i,i)})^2 = \sum_{i=1}^m \sigma_{\max}(A^{(i,i)})^2 \leq \sum_{i,j} \sigma_{\max}(A^{(i,j)})^2 := \sigma. \quad (15)$$

Thus assembling all the pieces together and lumping together absolute positive constants, we can conclude,

$$p_1 = \mathbb{P}\left(\sum_{i=1}^m \lambda_{\max}(A^{(i,i)})(X^{(i)T} X^{(i)} - \mathbb{E}[X^{(i)T} X^{(i)}]) \geq \frac{t}{2}\right) \leq \exp\left[-c_2\left(\frac{t^2}{\sigma}\right)\right],$$

where c_2 is another absolute positive constant.

Part 2: The Off Block Diagonal Entries. In this section, we are attempting to bound:

Note that

$$|X^{(i)T} A^{(i,j)} X^{(j)}| \leq \sigma_{\max}(A^{(i,j)}) \|X^{(i)}\|_2 \|X^{(j)}\|_2.$$

Thus, since $\|X^{(i)}\| \leq K = 1$, $X^{(i)T} A^{(i,j)} X^{(j)}$ is a mean zero sub-Gaussian random variable for all $i \neq j$ with sub-Gaussian norm:

$$\|X^{(i)T} A^{(i,j)} X^{(j)}\|_{\psi_2} \leq \sigma_{\max}(A^{(i,j)}) K^2 \leq \sigma_{\max}(A^{(i,j)})$$

We may then again apply a one sided Hoeffding-type inequality to state that

$$\mathbb{P}\left(\sum_{i \neq j} X^{(i)T} A^{(i,j)} X^{(j)} \geq t\right) \leq \exp\left[-c_1\left(\frac{t^2}{\sum_{i \neq j} \sigma_{\max}(A^{(i,j)})^2}\right)\right].$$

Now, we have

$$\sum_{i \neq j} \sigma_{\max}(A^{(i,j)})^2 \leq \sum_{i,j} \sigma_{\max}(A^{(i,j)})^2 = \sigma, \quad (16)$$

Thus assembling all the pieces together and lumping together absolute positive constants, we can conclude,

$$p_2 = \mathbb{P}\left(\sum_{i \neq j} X^{(i)T} A^{(i,j)} X^{(j)} \geq \frac{t}{2}\right) \leq \exp\left[-c_3\left(\frac{t^2}{\sigma}\right)\right],$$

This bound on p_2 is the same, up to the constant, as the bound on p_1 from Part 1.

To finish the proof, we can just use the union bound over the block diagonal and block off-diagonal results and lump together constants:

$$\begin{aligned} \mathbb{P}\left(\sum_{i=1}^m \lambda_{\max}(A^{(i,i)})(X^{(i)T} X^{(i)} - \mathbb{E}[X^{(i)T} X^{(i)}]) + \sum_{i,j \neq i} X^{(i)T} A^{(i,j)} X^{(j)} \geq t\right) &\leq p_1 + p_2 \\ &\leq 2 \exp\left[-c\left(\frac{t^2}{\sigma_{\max}(A^{(i,j)})^2}\right)\right], \end{aligned}$$

where c is an absolute positive constant. Since the left hand side upper bounds the left hand side of the expression in the lemma statement, we can conclude the proof. \square

Lemma 4. Let $\sigma^{(1)}, \sigma^{(2)}, \dots, \sigma^{(\ell)}$ be drawn iid $PL(\theta^*)$, for some parameter $\theta^* \in \Theta_B = \{\theta \in \mathbb{R}^d : \|\theta\|_\infty \leq B, \mathbf{1}^T \theta = 0\}$. Let $\lambda_2(L)$ be the second smallest eigenvalue of the (random) Plackett-Luce Laplacian obtained from ℓ samples of $PL(\theta^*)$. Then

$$\alpha_B n \leq \lambda_2(L),$$

with probability at least $1 - n^2 \exp(-\alpha_B^2 \ell)$, where $\alpha_B = 1/(4(1 + 2e^{3B}))$.

Proof. The edge weights of the comparison graph, denoted by $\bar{w}_{ij}, \forall i \neq j$, for a collection of ℓ rankings can be described as:

$$\bar{w}_{ij} = \frac{1}{(n-1)} \frac{1}{\ell} \sum_{k=1}^{\ell} \min\{\sigma_k(i), \sigma_k(j)\}.$$

Where $L_{ij} = -\bar{w}_{ij}, \forall i \neq j$. The implication of considering the edge weights of the graph represented by L is that we can lower bound $\lambda_2(L)$ in a simple way:

$$\lambda_2(L) \geq \lambda_2(K_n) \min_{ij} \bar{w}_{ij} = n \min_{ij} \bar{w}_{ij} = \frac{n}{n-1} \min_{ij} \left[\frac{1}{\ell} \sum_{k=1}^{\ell} \min\{\sigma_k(i), \sigma_k(j)\} \right]. \quad (17)$$

For the inequality, we use the fact that the algebraic connectivity of a graph on n vertices G must be lower bounded the algebraic connectivity of a complete graph on n vertices K_n whose edge weights are the smallest edge weight of G . The equality follows from the algebraic connectivity of a complete graph K_n .

We now need to unpack the right hand side of this bound. For each alternative pair $i, j \in \mathcal{X}$ and a ranking $\sigma \sim PL(\theta^*)$, define the random variables $X_{ij} = \min\{\sigma(i), \sigma(j)\}$. Extend this notation so that for $k = 1, \dots, \ell$, let $X_{ij}^{(k)} = \min\{\sigma(i)^{(k)}, \sigma(j)^{(k)}\}$ and additionally,

$$\bar{X}_{ij} = \frac{1}{\ell} \sum_{k=1}^{\ell} X_{ij}^{(k)} = \frac{1}{\ell} \sum_{k=1}^{\ell} \min\{\sigma(i)^{(k)}, \sigma(j)^{(k)}\}.$$

We are thus aiming to show that

$$\frac{1}{4} \left(\frac{n}{1 + 2e^{3B}} + 2 \right) \leq \min_{i,j \in \mathcal{X}} \bar{X}_{ij}.$$

Intuitively, \bar{X}_{ij} is the mean of ℓ iid variables X_{ij} , for each pair i, j . We have that $P(X_{ij} \geq k)$ is the probability that neither i nor j were chosen in the first $k-1$ MNL choices of repeated selection. Of course, $P(X_{ij} \geq 1) = 1$ and $P(X_{ij} \geq n) = 0$. For $k = 2, \dots, n-1$, we have,

$$P(X_{ij} \geq k) = \underbrace{\left(\sum_{(i_1, \dots, i_{k-2}) \in \mathcal{X} \setminus \{i, j\}} \prod_{m=1}^{k-2} \frac{e^{\theta_{i_m}}}{\sum_{x \in \mathcal{X} \setminus \cup_{q=1}^{m-1} \{i_q\}} e^{\theta_x}} \right)}_{\text{Neither } i \text{ nor } j \text{ are choice } 1, \dots, k-2} \underbrace{\left(1 - \frac{e^{\theta_i} + e^{\theta_j}}{\sum_{x \in \mathcal{X} \setminus \{i_1, \dots, i_{k-2}\}} e^{\theta_x}} \right)}_{\text{Neither } i \text{ nor } j \text{ are choice } k-1},$$

where we define the terms so that the first underbrace term just becomes 1 when $k = 2$.

We will now lower bounded the probability of not choosing item i or j in steps $1, \dots, k-1$ as the probability of not choosing two items each with utility B when all other items have utility $-B$. We

proceed as follows:

$$\begin{aligned}
P(X_{ij} \geq k) &= \sum_{(i_1, \dots, i_{k-2}) \in \mathcal{X} \setminus \{i, j\}} \prod_{m=1}^{k-2} \frac{e^{\theta_{i_m}}}{\sum_{x \in \mathcal{X} \setminus \cup_{q=1}^{m-1} \{i_q\}} e^{\theta_x}} \left(1 - \frac{e^{\theta_i} + e^{\theta_j}}{\sum_{x \in \mathcal{X} \setminus \{i_1, \dots, i_{k-2}\}} e^{\theta_x}}\right) \\
&\geq \sum_{(i_1, \dots, i_{k-2}) \in \mathcal{X} \setminus \{i, j\}} \prod_{m=1}^{k-2} \frac{e^{\theta_{i_m}}}{\sum_{x \in \mathcal{X} \setminus \cup_{q=1}^{m-1} \{i_q\}} e^{\theta_x}} \left(1 - \frac{e^{\theta_i} + e^{\theta_j}}{e^{\theta_i} + e^{\theta_j} + (n-k)e^{-B}}\right) \\
&= \sum_{(i_1, \dots, i_{k-3}) \in \mathcal{X} \setminus \{i, j\}} \prod_{m=1}^{k-3} \frac{e^{\theta_{i_m}}}{\sum_{x \in \mathcal{X} \setminus \cup_{q=1}^{m-1} \{i_q\}} e^{\theta_x}} \left(1 - \frac{e^{\theta_i} + e^{\theta_j}}{\sum_{x \in \mathcal{X} \setminus \{i_1, \dots, i_{k-3}\}} e^{\theta_x}}\right) \\
&\quad \times \left(1 - \frac{e^{\theta_i} + e^{\theta_j}}{e^{\theta_i} + e^{\theta_j} + (n-k)e^{-B}}\right) \\
&\geq \sum_{(i_1, \dots, i_{k-3}) \in \mathcal{X} \setminus \{i, j\}} \prod_{m=1}^{k-3} \frac{e^{\theta_{i_m}}}{\sum_{x \in \mathcal{X} \setminus \cup_{q=1}^{m-1} \{i_q\}} e^{\theta_x}} \left(1 - \frac{e^{\theta_i} + e^{\theta_j}}{e^{\theta_i} + e^{\theta_j} + (n-k-1)e^{-B}}\right) \\
&\quad \times \left(1 - \frac{e^{\theta_i} + e^{\theta_j}}{e^{\theta_i} + e^{\theta_j} + (n-k)e^{-B}}\right) \\
&\dots \\
&\geq \prod_{m=1}^{k-1} \left(1 - \frac{e^{\theta_i} + e^{\theta_j}}{e^{\theta_i} + e^{\theta_j} + (n-m-1)e^{-B}}\right) \\
&\geq \prod_{m=1}^{k-1} \left(1 - \frac{2e^B}{2e^B + (n-m-1)e^{-B}}\right).
\end{aligned}$$

The first line restates the equation, the second lower bounds the final term by decreasing the logits of all variables but θ_i and θ_j to $-B$, their smallest possible value. The lower bound is then restated in the third line, where an explicit specification of the probability of choice $k-2$ is stated implicitly as the probability of any choice in the available universe but i and j . We may perform this step after the lower bound, but not before, because the probability of choice $k-1$ in the lower bound is unaffected by the choice made in $k-2$, whereas it would be affected in the original bound. In the fourth line, we lower bound the $k-2$ choice probability in a similar manner to the second line, and repeat the restate-bound procedure over and over until we arrive at the final inequality, stated in the second to last line. This expression is again lower bounded in the last line by raising the utility of items i and j to make the lower bound independent of i and j .

Next, we have,

$$\begin{aligned}
P(X_{ij} \geq k) &\geq \prod_{m=1}^{k-1} \left(1 - \frac{2e^B}{2e^B + (n-m-1)e^{-B}}\right) \\
&\geq \left(1 - \frac{2e^B}{2e^B + (n-k)e^{-B}}\right)^{k-1} \\
&\geq 1 - \frac{2(k-1)e^B}{2e^B + (n-k)e^{-B}} \\
&\geq 1 - \frac{2ke^B}{2e^B + (n-k)e^{-B}},
\end{aligned}$$

and so for $\delta \in [0, 1]$,

$$\begin{aligned}
P(X_{ij} \geq \delta n) &\geq 1 - \frac{2e^{2B}\delta n}{2e^{2B} + (1-\delta)ne^{-B}} \\
&= 1 - \frac{\delta n}{1 + (1-\delta)n\frac{e^{-3B}}{2}} \\
&\geq 1 - \frac{2e^{3B}\delta}{(1-\delta)}.
\end{aligned}$$

Now, setting

$$\delta = \frac{1-c}{2e^{3B} + 1 - c},$$

we have that

$$P\left(X_{ij} \geq \frac{1-c}{2e^{3B} + 1 - c}n\right) \geq c.$$

This expression claims that X_{ij} is at least linear in n with however large a probability we desire. We can use these expressions to also lower bound the expectation:

$$\begin{aligned} \mathbb{E}[X_{ij}] &= 1 + \sum_{k=2}^{n-1} P(X_{ij} \geq k) \\ &\geq 1 + \sum_{k=2}^{\frac{n}{1+2e^{3B}}} \left(1 - \frac{2e^{3B}k}{n-k}\right) \\ &= 1 + \left(\frac{n}{1+2e^{3B}} - 1\right) - 2e^{3B} \sum_{k=2}^{\frac{n}{1+2e^{3B}}} \frac{k}{n-k} \\ &\geq 1 + \left(\frac{n}{1+2e^{3B}} - 1\right) - 2e^{3B} \sum_{k=2}^{\frac{n}{1+2e^{3B}}} \frac{k}{n - \frac{n}{1+2e^{3B}}} \\ &= 1 + \left(\frac{n}{1+2e^{3B}} - 1\right) - 2e^{3B}(1+2e^{3B}) \sum_{k=2}^{\frac{n}{1+2e^{3B}}} \frac{k}{(1+2e^{3B})n - n} \\ &= 1 + \left(\frac{n}{1+2e^{3B}} - 1\right) - \frac{1+2e^{3B}}{n} \sum_{k=2}^{\frac{n}{1+2e^{3B}}} k \\ &= 1 + \left(\frac{n}{1+2e^{3B}} - 1\right) - \frac{1+2e^{3B}}{n} \frac{1}{2} \left(\frac{n}{1+2e^{3B}}\right) \left(\frac{n}{1+2e^{3B}} + 1\right) + 1 \\ &= \frac{1}{2} \left(\frac{n}{1+2e^{3B}} + 1\right), \end{aligned}$$

where the first expression follows from the substitution of the tail bound only for the regime where it is non-zero (ignoring, for simplicity, the matter of the ceiling operators), and using zero otherwise, the next follows from lower bounding the second expression with the lowest value achieved in the sum. The remaining steps perform the necessary algebra to arrive at the final expression, which is clearly affine in n .

Now, recall that \bar{X}_{ij} is the mean of ℓ independent variables X_{ij} . Certainly, the expectation of \bar{X}_{ij} is the same as that of X_{ij} . Since X_{ij} is always bounded between 1 and $n-1$, we may use the one-sided Hoeffding's inequality to make the following claim, for any i, j :

$$P(\mathbb{E}[X_{ij}] - \bar{X}_{ij} \geq \delta n) \leq \exp\left(-\frac{\ell^2 \delta^2 n^2}{\ell(n-2)^2}\right) \leq \exp(-\ell \delta^2).$$

That is, for any i, j , with probability at least $1 - \exp(-\ell \delta^2)$ we have that $\mathbb{E}[X_{ij}] - \bar{X}_{ij} \leq \delta n$. Using our lower bound from before, we have that

$$\frac{1}{2} \left(\frac{n}{1+2e^{3B}} + 1\right) - \delta n \leq \bar{X}_{ij}.$$

A strong upper bound on the failure probability of \bar{X}_{ij} attaining the left-hand-side value allows us to easily union bound the failure probability of $\min_{ij} \bar{X}_{ij}$ attaining the same left-hand-side value. Namely,

$$\min_{ij} \bar{X}_{ij} < \frac{1}{2} \left(\frac{n}{1+2e^{3B}} + 1\right) - \delta n$$

if any \bar{X}_{ij} is less than the right hand side. Thus,

$$\mathbb{P}\left[\min_{ij} \bar{X}_{ij} < \frac{1}{2}\left(\frac{n}{1+2e^{3B}}+1\right)-\delta n\right] \leq \sum_{ij} \mathbb{P}\left[\bar{X}_{ij} < \frac{1}{2}\left(\frac{n}{1+2e^{3B}}+1\right)-\delta n\right] \leq n^2 \exp(-\ell\delta^2),$$

where the first inequality is the union bound, and the second inequality applies the preceding bound on the failure probability of any \bar{X}_{ij} and additionally upper bounds as $\binom{n}{2}$ by n^2 .

Now, Setting δ to $\frac{1}{4}\frac{1}{1+2e^{3B}}$, we may conclude the following: with probability at least

$$1 - n^2 \exp\left(-\ell \frac{1}{16} \frac{1}{(1+2e^{3B})^2}\right), \quad (18)$$

we have that

$$\frac{1}{4}\left(\frac{n}{1+2e^{3B}}+2\right) \leq \min_{ij} \bar{X}_{ij}.$$

Returning all the way to Equation (17), we have that:

$$\lambda_2(L) \geq \frac{n}{n-1} \min_{ij} \bar{X}_{ij} \geq \frac{n}{n-1} \frac{1}{4}\left(\frac{n}{1+2e^{3B}}+2\right).$$

Defining α_B as

$$\alpha_B := \frac{1}{4(1+2e^{3B})},$$

which simplifies to:

$$\lambda_2(L) \geq \frac{n}{n-1}\left(\alpha_B n + \frac{1}{2}\right) \geq \alpha_B n, \quad (19)$$

where we lower bound $\frac{n}{n-1}$ by 1 and drop the $1/2$.

Rewriting the probability of this inequality (from (18)) in terms of α_B , we have that Equation (19) occurs with probability

$$1 - n^2 \exp(-\ell\alpha_B), \quad (20)$$

completing the proof. \square

Lemma 5. Let $\sigma_1, \dots, \sigma_\ell$ be drawn iid $\text{CRS}(u^*)$, the full CRS model with true parameter $u^* \in \Theta_B = \{u \in \mathbb{R}^{n(n-1)} : u = [u_1^T, \dots, u_n^T]^T; u_i \in \mathbb{R}^{n-1}, \|u_i\|_1 \leq B, \forall i; \mathbf{1}^T u = 0\}$. Let $\lambda_2(X^T X)$ be the second smallest eigenvalue of the scaled (random) design matrix X obtained from ℓ samples of $\text{CRS}(u^*)$. Then

$$\frac{\delta_n}{(n-1)e^{2B}} < \lambda_2(X^T X),$$

with probability at least $1 - n \exp(-\frac{\ell}{8ne^{2B}})$, where $0 < \delta_n \leq \frac{1}{2n}$ is a fixed number that depends only on n .

Proof. Define $L = X^T X$, denote by \mathcal{D} the collection of $\ell(n-1)$ choices constructed from the ℓ rankings by repeated selection, and let $C_{\mathcal{D}}$ be the set of choice sets in \mathcal{D} . We also use $X(\mathcal{D})$ to refer to X . The CDM is identifiable (and so $\lambda_2(L) > 0$) the moment $C_{\mathcal{D}}$ contains the set \mathcal{X} and $\mathcal{X} \setminus i, \forall i$. We can see this observation because when we have all $\mathcal{X} \setminus i$, we have all sets of size n and $n-1$, and thus can invoke Theorem 4 of [48] to say that $\text{rank}(X(\mathcal{D})) = n(n-1) - 1$ and thus $\text{rank}(L) \leq \text{rank}(X(\mathcal{D})^T X(\mathcal{D})) = \text{rank}(X(\mathcal{D})) = n(n-1) - 1$, so $\lambda_2(L) > 0$.

The universe set \mathcal{X} always appears with every ranking, so we could aim to lower bound the probability that each set of size $n-1$ appears in $C_{\mathcal{D}}$ at least once in order to get identifiability. We can however, do more if each set of size $n-1$ appears in $C_{\mathcal{D}}$ at least r times.

Suppose, for instance, that all $\mathcal{X} \setminus i$ are in $C_{\mathcal{D}}$. Consider just the choices corresponding to the n unique sets of size $n-1$, as well as the n universe sets that came along with the sets of $n-1$ in the same ranking, and refer to this collection of choices by the dataset $\bar{\mathcal{D}}$. For this dataset, the corresponding L matrix (which we will label $L_{\bar{\mathcal{D}}}$) is, using the notation of [48],

$$L_{\bar{\mathcal{D}}} = \frac{1}{2n} \sum_{i=1}^{2n} E_{i,k_i}^T \left(I - \frac{1}{k_i} \mathbf{1}\mathbf{1}^T\right) E_{i,k_i}.$$

where the matrices $E_{i,k_i} \in \{0, 1\}^{n(n-1) \times k_i}$ depend only on the sets. Hence, it is easy to see that

$$0 < \delta_n = \lambda_2(L_{\tilde{\mathcal{D}}}),$$

where δ_n depends only on n (since $C_{\tilde{\mathcal{D}}}$ identifies the CDM, $\delta_n > 0$). Note that δ_n is deterministic since the sets in $\tilde{\mathcal{D}}$ are fixed. Moreover, since

$$L_{\tilde{\mathcal{D}}} = \frac{1}{2n} \sum_{i=1}^{2n} E_{i,k_i}^T (I - \frac{1}{k_i} \mathbf{1}\mathbf{1}^T) E_{i,k_i} \preceq \frac{1}{2n} \sum_{i=1}^{2n} E_{i,k_i}^T E_{i,k_i} =: T,$$

where T is block diagonal matrix with n blocks $T_i \in \mathbb{R}^{n-1 \times n-1}$, $\forall i = 1..n$. Specifically, $\forall i$,

$$T_i = \frac{1}{2n} [(2n-3)\mathbf{1}\mathbf{1}^T + \text{Diag}(\mathbf{1})]$$

Since T_i 's $n-2$ smallest eigenvalues are of value $1/(2n)$, T 's $n(n-2)$ smallest eigenvalues are of value $1/(2n)$ by the properties of block diagonal matrices. Hence, by the eigenvalue interlacing theorem [11][See pg. 128, Corollary (4.2)],

$$\delta_n = \lambda_2(L_{\tilde{\mathcal{D}}}) \leq \lambda_2(T) \leq \frac{1}{2n}.$$

Now, look back at \mathcal{D} , which has ℓ rankings, and the $n-1$ choices that come from each. L is thus:

$$L = \frac{1}{\ell(n-1)} \sum_{i=1}^{\ell(n-1)} E_{i,k_i} (I - \frac{1}{k_i} \mathbf{1}\mathbf{1}^T) E_{i,k_i}^T$$

Since \mathcal{D} contains the choices in $\tilde{\mathcal{D}}$, and since each term in the sum of L is PSD,

$$\frac{2n}{\ell(n-1)} \delta_n \leq \lambda_2(L),$$

as a result of properties of sums of PSD matrices [11][See pg. 128, Corollary (4.2)]. In fact, if \mathcal{D} contains the choices $\tilde{\mathcal{D}}$ in r copies, then we can say

$$\frac{2nr}{\ell(n-1)} \delta_n \leq \lambda_2(L).$$

With this understanding, we may now proceed to study the probability that each set of size $n-1$ appears in $C_{\mathcal{D}}$ at least r times. Let $\sigma \sim CRS(u^*)$. The subsets $\mathcal{X} \setminus i$ appear in the repeated selection decomposition of a ranking if $\sigma^{-1}(i) = 1$, that is, if i is ranked first, which happens with probability

$$P(i|\mathcal{X}; u^*) \propto \exp\left(\sum_{j \in \mathcal{X} \setminus i} u_{ij}^*\right).$$

Noting that $\|u_i^*\|_1 \leq B$, it follows that $\sum_{j \in \mathcal{X} \setminus i} u_{ij}^* \geq -B$ and that for every $i' \neq i$, $\sum_{j \in \mathcal{X} \setminus i'} u_{i'j}^* \leq B$. It follows that

$$P(i|\mathcal{X}; u^*) \geq \frac{e^{-B}}{e^{-B} + (n-1)e^B} \geq \frac{1}{ne^{2B}}.$$

Consider now \mathcal{D} , which are ℓ rankings from the model, $\sigma_1, \dots, \sigma_{\ell_0} \sim CRS(u^*)$. The probability that $\mathcal{X} \setminus i \in C_{\mathcal{D}}$ at most r_i times given ℓ trials is the CDF of a Binomial distribution with success probability $P(i|\mathcal{X}; u^*)$. We may use a Chernoff bound to upper bound the CDF [2]:

$$\begin{aligned} P(\mathcal{X} \setminus i \in C_{\mathcal{D}} \text{ at most } r_i \text{ times}) &\leq \exp\left(-\ell D\left(\frac{r_i}{\ell} \parallel P(i|\mathcal{X}; u^*)\right)\right) \\ &\leq \exp(-\ell D(.5P(i|\mathcal{X}; u^*) \parallel P(i|\mathcal{X}; u^*))) \\ &\leq \exp(-\ell .125P(i|\mathcal{X}; u^*)) \\ &\leq \exp\left(-\frac{\ell}{8ne^{2B}}\right), \end{aligned}$$

where the second inequality follows by setting $r_i = .5P(i|\mathcal{X}; u^*)\ell$, the third from the fact that $D(.5p \parallel p) \geq .125p$, and the last from lower bounding $P(i|\mathcal{X}; u^*)$. For each $i \in \mathcal{X}$, let A_i be the

event that $\mathcal{X} \setminus i \in C_{\mathcal{D}}$ at most r_i times. Then all $\mathcal{X} \setminus i$ are in $C_{\mathcal{D}}$ at least $r = \min_i r_i$ times whenever we are *not* in $\cup_{i \in \mathcal{X}} A_i^C$. With a union bound and the previous bound we have

$$\begin{aligned} P(\cup_{i \in \mathcal{X}} A_i^C) &\leq \sum_i P(A_i^C) \\ &\leq n \exp\left(-\frac{\ell}{8ne^{2B}}\right). \end{aligned}$$

With this result, we can see that with probability at least $1 - n \exp(-\frac{\ell}{8ne^{2B}})$

$$\lambda_2(L) \geq \frac{2nr}{\ell(n-1)} \delta_n \geq \frac{2n}{\ell(n-1)} \frac{\ell}{2ne^{2B}} \delta_n \geq \frac{\delta_n}{(n-1)e^{2B}}.$$

This completes the proof. \square

Lemma 6. *Let u^* denote the true CDM model from which choice data \mathcal{D} with m choices is drawn. Let \hat{u}_{MLE} denote the maximum likelihood solution. For any $u^* \in \mathcal{U}_B = \{u \in \mathbb{R}^{n(n-1)} : u = [u_1^T, \dots, u_n^T]^T; u_i \in \mathbb{R}^{n-1}, \|u_i\|_1 \leq B, \forall i; \mathbf{I}^T u = 0\}$, and $t > 1$,*

$$\mathbb{P}\left[\|\hat{u}_{MLE}(\mathcal{D}) - u^*\|_2^2 \geq c_{B, k_{\max}} \frac{ctn(n-1)}{m\lambda_2(L)}\right] \leq e^{-t},$$

where k_{\max} is the maximum choice set size in \mathcal{D} , $c_{B, k_{\max}}$ is a constant that depends on B and k_{\max} , and $\lambda_2(L)$ the spectrum of $L = X(\mathcal{D})^T X(\mathcal{D})$ with scaled design matrix $X(\mathcal{D})$. For the expected risk,

$$\mathbb{E}[\|\hat{u}_{MLE}(\mathcal{D}) - u^*\|_2^2] \leq c'_{B, k_{\max}} \frac{n(n-1)}{m\lambda_2(L)},$$

where the expectation is taken over the dataset \mathcal{D} generated by the choice model and $c'_{B, k_{\max}}$ is again a constant that depends on B and k_{\max} .

Proof.

Our proof is very similar to the proof of Theorem 1. Much like that proof, we will first introduction notation for analyzing the risk, and then proceed to first give a proof of the expected risk bound. We then carry the technology of that proof forward to give a proof of the tail bound statement. The great similarity to the proof of Theorem 1 demands that we repeat some of the same arguments here – in lieu of doing that, we jump to conclusions and refer the reader to the respective sections of Theorem 1. The resulting risk bound portion is a nearly exact replica of the risk bound in [48] (see Theorem 1), with the main difference being that our statement here employs a more restrictive assumption on \mathcal{U}_B than its counterpart in [48], based on an infinity-norm vs. the 1-norm above. We constrain \mathcal{U}_B so we may apply Lemma 1 to efficiently bound a $\beta_{k_{\max}}$ term that arises in the proof. Unpacking the proof for the risk bound additionally yields the right tools to prove the tail bound result above, an entirely novel contribution of our work. For both the tail and risk bound sections, the notation we use for the CDM model often overloads the notation used in the MNL model’s proof. This overloading is intentional, and is meant to convey the high degree of similarity between the two models, and the proof techniques used to provide guarantees for them.

We are given some true CDM model with parameters $u^* \in \mathcal{U}_B$, and for each datapoint $j \in [m]$ we have the probability of choosing item x from set C_j as

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

Notation. We now introduce notation that will let us represent the above expression in a more compact manner. In this proof, we will use $d = n(n-1)$ to refer to the CDM parameter space. Because our datasets involve choice sets of multiple sizes, we use $k_j \in [k_{\min}, k_{\max}]$ to denote the choice set size for datapoint j , $|C_j|$. Extending a similar concept in [49] to the multiple set sizes, and the more complex structure of the CDM, we then define matrices $E_{j, k_j} \in \mathbb{R}^{d \times k_j}$, $\forall j \in [m]$ as follows: E_{j, k_j} has a column for every item $y \in C_j$ (and hence k_j columns), and the column corresponding to item $y \in C_j$ has a one at the position of each u_{yz} for $z \in C_j \setminus y$, and zero otherwise. This construction

allows us to write the familiar expressions $\sum_{z \in C_j \setminus y} u_{yz}$, for each y , simply as a single vector-matrix product $u^T E_{j,k_j} = [\sum_{z \in C_j \setminus y_1} u_{y_1 z}, \sum_{z \in C_j \setminus y_2} u_{y_2 z}, \dots, \sum_{z \in C_j \setminus y_{k_j}} u_{y_{k_j} z}] \in \mathbb{R}^{1 \times k_j}$.

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

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

where the numerator always corresponds to the first entry of the input. These functions F_k have several properties that will become useful later in the proof. First, it is easy to verify that all F_k are shift-invariant, that is, $F_k(x) = F_k(x + c\mathbf{1})$, for any scalar c . The purpose of introducing F_k is to write CDM probabilities compactly. Since the inputs x_i are sums of $k - 1$ values of the u vector, the inputs $x \in [-B, B]^n$.

We may thus apply Lemma 1, and obtain that that $\mathbf{1} \in \text{null}(\nabla^2(-\log(F_k(x))))$ and that

$$\nabla^2(-\log(F_k(x))) \succeq H_k = \beta_k(I - \frac{1}{k}\mathbf{1}\mathbf{1}^T), \quad (21)$$

where

$$\beta_k := \frac{1}{k \exp(2B)}. \quad (22)$$

That is, F_k are strongly log-concave with a null space *only* in the direction of $\mathbf{1}$, since $\nabla^2(-\log(F_k(x))) \succeq H_k$ for some $H_k \in \mathbb{R}^{k \times k}, \lambda_2(H_k) > 0$.

As a final notational addition, in the same manner as [49] but accounting for multiple set sizes, we define k permutation matrices $R_{1,k}, \dots, R_{k,k} \in \mathbb{R}^{k,k}, \forall k \in [k_{\min}, k_{\max}]$, representing k cyclic shifts in a fixed direction. Specifically, given some vector $x \in \mathbb{R}^k, y = x^T R_{l,k}$ is simply x^T cycled (say, clockwise) so $y_1 = x_l, y_i = x_{(l+i-1)\%k}$, where $\%$ is the modulo operator. That is, these matrices allow for the cycling of the entries of row vector $v \in \mathbb{R}^{1 \times k}$ so that any entry can become the first entry of the vector, for any of the relevant k . This construction allows us to represent any choice made from the choice set C_j as the first element of the vector x that is input to F , thereby placing it in the numerator.

First, an expected risk bound. Given the notation introduced above, we can now state the probability of choosing the item x from set C_j compactly as:

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

We can then rewrite the CDM likelihood as

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

and the scaled negative log-likelihood as

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

Thus,

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

At this point, it should be clear to the reader that the problem formulation is almost exactly the same as that of Theorem 1, the only difference being that u belongs in a higher dimensional space than θ , and that the respective proofs' E_{j,k_j} is defined differently. Due to \mathcal{U}_B 's definition restricting the ℓ_1 norm of certain entries of u , we see that the inputs $u^T E_{j,k_j} R_{x_j, k_j}$ to the functions F_{k_j} live within $[-B, B]^{k_j}$, much like the restrictions on Θ_B also resulted in $\theta^T E_{j,k_j} R_{x_j, k_j} \in [-B, B]^{k_j}$. The similarity of the problems allow us to port over certain steps used in the proof without additional justification.

To begin, we have the gradient of the negative log-likelihood as

$$\nabla \ell(u) = -\frac{1}{m} \sum_{j=1}^m \sum_{i=1}^{k_j} \mathbf{1}[y_j = i] E_{j,k_j} R_{i,k_j} \nabla \log(F_{k_j}(u^T E_{j,k_j} R_{i,k_j})), \quad (23)$$

and the Hessian as

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

Proceeding identically as Theorem 1, we have, for any vector $z \in \mathbb{R}^d$,

$$z^T \nabla^2 \ell(u) z \geq \beta_{k_{\max}} \frac{1}{m} \sum_{j=1}^m z^T E_{j,k_j} (I - \frac{1}{k_j} \mathbf{1}\mathbf{1}^T) E_{j,k_j}^T z,$$

where we have followed the same steps taken in Theorem 1 to bound $z^T \nabla^2 \ell(\theta) z$. Now, defining the matrix L as

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

we first note a few properties of L . First, it is easy to verify that, like in Theorem 1, $L\mathbf{1} = 0$, and hence $\text{span}(\mathbf{1}) \subseteq \text{null}(L)$. It is also straightforward to see that L is **not** a Laplacian, unlike the MNL case, since every non pair edge creates positive entries in the off diagonal. Moreover, we follow an argument in [48] to show that $\lambda_2(L) > 0$, that is, $\text{null}(L) \subseteq \text{span}(\mathbf{1})$. Consider the matrix $G(\mathcal{D})$, the design matrix of the CDM for the given dataset \mathcal{D} described in detail in Theorem 4 of [48]. Define a matrix $X(\mathcal{D}) = \mathbf{C}_{\mathcal{D}}^{-1} G(\mathcal{D})$, where $\mathbf{C}_{\mathcal{D}}^{-1} \in \mathbb{R}^{\Omega_{\mathcal{D}} \times \Omega_{\mathcal{D}}}$ is the diagonal matrix with values are $\frac{1}{k_j}$, for every datapoint j , for every item $x \in C_j$, and where $\Omega_{\mathcal{D}} = \sum_{i=1}^m k_i$. $X(\mathcal{D})$ should therefore be thought of as a scaled design matrix. Simple calculations show that,

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

As a consequence of the properties of matrix rank, we then have that $\text{rank}(L) = \text{rank}(X(\mathcal{D})) = \text{rank}(G(\mathcal{D}))$. Thus, from Theorem 4 of [48], we have that if the dataset \mathcal{D} identifies the CDM, $\text{rank}(L) = d - 1$, and hence $\lambda_2(L) > 0$. We may then leverage conditions for identifiability from [48] (Theorem 1 and 2) to determine when L is positive definite.

With this matrix, we can write,

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

which is equivalent to stating that $\ell(u)$ is $\beta_{k_{\max}}$ -strongly convex with respect to the L semi-norm at all $u \in \mathcal{U}_B$. Since $u^*, \hat{u}_{\text{MLE}} \in \mathcal{U}_B$, we can now follow the implications of strong convexity with respect to the L semi-norm used in Theorem 1, of course now with a different L , and conclude that:

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

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

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

where in the equality we have defined $V_{j,k_j} \in \mathbb{R}^{k_j}$ as

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

It is not difficult to see that V_{j,k_j} is defined identically to the one in Theorem 1. We may thus borrow three results directly from that proof:

$$V_{j,k_j}^T \mathbf{1} = 0 \quad \mathbb{E}[V_{j,k_j}] = 0 \quad \sup_{j \in [m]} \|V_{j,k_j}\|_2^2 \leq 2.$$

The remainder of the proof departs from the corresponding sections of 1, and we thus carefully detail every step. Consider the matrix $M_k = (I - \frac{1}{k} \mathbf{1}\mathbf{1}^T)$. We note that M_k has rank $k - 1$, with its nullspace corresponding to the span of the ones vector. We state the following identities:

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

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

Next, we have

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

where the second line follows from identities of the M matrix, the third from the independence of the V_{j,k_j} , the fourth from an upper bound of the quadratic form, the fifth from the properties of trace, the sixth from the definition of the matrix L , and the last from the value of the trace, which is simply the identity matrix with one zero entry in the diagonal.

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

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

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

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

By noting that $\|\hat{u}_{\text{MLE}} - u^*\|_L^2 = (\hat{u}_{\text{MLE}} - u^*)^T L (\hat{u}_{\text{MLE}} - u^*) \geq \lambda_2(L) \|\hat{u}_{\text{MLE}} - u^*\|_2^2$, since $\hat{u}_{\text{MLE}} - u^* \perp \text{null}(L)$, we can translate our above result into the ℓ_2 norm:

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

Now, setting

$$c'_{B,k_{\max}} := \frac{2}{\beta_{k_{\max}}^2} = 2 \exp(4B) k_{\max}^2,$$

we retrieve the expected risk bound in the theorem statement,

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

We close the expected risk portion of this proof with some remarks about $c_{B, k_{\max}}$. The quantity $\beta_{k_{\max}}$, defined in equation (22), serves as the important term that approaches 0 as a function of B and k_{\max} , requiring that the former be bounded. Finally, $\lambda_2(L)$ is a parallel to the requirements on the algebraic connectivity of the comparison graph in [49] for the pairwise setting. Though the object L here appears similar to the graph Laplacian L in that work, there are major differences that are most worthy of further study.

From expected risk to tail bound. Our proof of the tail bound is a continuation of the expected risk bound proof. While the expected risk bound closely followed the expected risk proof of [48] and [49], our tail bound proof contains significant novel machinery. Our presentation seem somewhat circular, given that the tail bound itself integrates out to an expected risk bound with the same parametric rates (albeit worse constants), but we felt that to first state the expected risk bound was clearer, given that it arises as a stepping stone to the tail bound.

Recall again the expression for the gradient in Equation (23). Useful in our analysis will be an alternate expression:

$$\nabla \ell(u^*) = -\frac{1}{m} \sum_{j=1}^m E_{j, k_j} V_{j, k_j} = -\frac{1}{m} E^T V,$$

where we have defined $V \in \mathbb{R}^{\Omega_{\mathcal{D}}}$ as the concatenation of all V_{j, k_j} , and $E \in \mathbb{R}^{\Omega_{\mathcal{D}} \times n}$, the vertical concatenation of all the E_{j, k_j} . Recall that $\Omega_{\mathcal{D}} = \sum_{i=1}^m k_i$.

For the expected risk bound, we showed that V_{j, k_j} have expectation zero, are independent, and $\|V_{j, k_j}\|_2^2 \leq 2$. Considering again the matrix M_k , recall that we have $M_{k_j} V_{j, k_j} = M_{k_j}^{\frac{1}{2}} M_{k_j}^{\frac{1}{2}} V_{j, k_j} = M_{k_j} M_{k_j}^{\dagger} V_{j, k_j} = V_{j, k_j}$, where the last equality follows since V_{j, k_j} is orthogonal to the nullspace of M_{k_j} . We may now again revisit the expression for the gradient:

$$\nabla \ell(\theta^*) = -\frac{1}{m} \sum_{j=1}^m E_{j, k_j} V_{j, k_j} = -\frac{1}{m} \sum_{j=1}^m E_{j, k_j} M_{k_j}^{1/2} V_{j, k_j} := -\frac{1}{m} X(\mathcal{D})^T V,$$

where we have defined $X(\mathcal{D}) \in \mathbb{R}^{\Omega_{\mathcal{D}} \times n}$ as the vertical concatenation of all the $E_{j, k_j} M_{k_j}^{1/2}$, and the scaled design matrix described before.

Now, consider that

$$\nabla \ell(u^*)^T \hat{L}^{\dagger} \nabla \ell(u^*) = \frac{1}{m^2} V^T X(\mathcal{D}) \hat{L}^{\dagger} X(\mathcal{D})^T V.$$

We apply Lemma 3, a modified Hanson-Wright-type tail bound for random quadratic forms. This lemma follows from simpler technologies (largely Hoeffding's inequality) given that the random variables are bounded while also carefully handling the block structure of the problem.

In the language of Lemma 3 we have V_{j, k_j} playing the role of $x^{(j)}$ and $\Sigma_{\mathcal{D}} := \frac{1}{m^2} X(\mathcal{D}) \hat{L}^{\dagger} X(\mathcal{D})^T$ plays the role of A . The invocation of this lemma is possible because V_{j, k_j} is mean zero, $\|V_{j, k_j}\|_2 \leq \sqrt{2}$, and because $\Sigma_{\mathcal{D}}$ is positive semi-definite. We sweep $K^4 = 4$ from the lemma statement into the constant c of the right hand side. Stating the result of Lemma 3 we have, for all $t > 0$,

$$\mathbb{P}(V^T \Sigma_{\mathcal{D}} V - \sum_{i=1}^m \lambda_{\max}(\Sigma_{\mathcal{D}}^{(i, i)}) \mathbb{E}[V^{(i)T} V^{(i)}] \geq t) \leq 2 \exp\left(-c \frac{t^2}{\sum_{i, j} \sigma_{\max}(\Sigma_{\mathcal{D}}^{(i, j)})^2}\right). \quad (24)$$

We note that

$$\sum_{i, j} \sigma_{\max}(\Sigma_{\mathcal{D}}^{(i, j)})^2 \leq \sum_{i, j} \sum_k \sigma_k(\Sigma_{\mathcal{D}}^{(i, j)})^2 = \sum_{i, j} \|\Sigma_{\mathcal{D}}^{(i, j)}\|_F^2 = \|\Sigma_{\mathcal{D}}\|_F^2,$$

where the first inequality follows because the max is less than the sum of positive values, and the first equality from the definition of Frobenius norm, and second from the Frobenius norm of blocks being

the Frobenius norm of the whole. This inequality allows us to conclude that, for the right hand side of Equation (24):

$$2 \exp \left(-c \frac{t^2}{\sum_{i,j} \sigma_{\max}(\Sigma_{\mathcal{D}}^{(i,j)})^2} \right) \leq 2 \exp \left(-c \frac{t^2}{\|\Sigma_{\mathcal{D}}\|_F^2} \right).$$

Next, within the left hand side of Equation (24) we have,

$$\sum_{i=1}^m \lambda_{\max}(\Sigma_{\mathcal{D}}^{(i,i)}) \leq \sum_{i=1}^m \text{tr}(\Sigma_{\mathcal{D}}^{(i,i)}) \leq \text{tr}(\Sigma_{\mathcal{D}}),$$

and so:

$$\mathbb{P}(V^T \Sigma_{\mathcal{D}} V - \text{tr}(\Sigma_{\mathcal{D}}) \sup_{i \in [m]} \mathbb{E}[V^{(i)T} V^{(i)}] \geq t) \leq \mathbb{P}(V^T \Sigma_{\mathcal{D}} V - \sum_{i=1}^m \lambda_{\max}(\Sigma_{\mathcal{D}}^{(i,i)}) \mathbb{E}[V^{(i)T} V^{(i)}] \geq t).$$

We may now combine these improvements for a much more compact version of Equation (24):

$$\mathbb{P}(V^T \Sigma_{\mathcal{D}} V - \text{tr}(\Sigma_{\mathcal{D}}) \sup_{i \in [m]} \mathbb{E}[V^{(i)T} V^{(i)}] \geq t) \leq 2 \exp \left(-c \frac{t^2}{\|\Sigma_{\mathcal{D}}\|_F^2} \right). \quad (25)$$

Now, some algebra (see Lemma 2) reveals that:

$$\text{tr}(\Sigma_{\mathcal{D}}) = \frac{d-1}{m}, \quad \|\Sigma_{\mathcal{D}}\|_F^2 = \frac{(d-1)}{m^2}, \quad (26)$$

where we have used the fact that $\hat{L} = \frac{1}{m} X(\mathcal{D})^T X(\mathcal{D})$ and hence its pseudoinverse cancels out the other terms in $\Sigma_{\mathcal{D}}$. Now, noting that the norm of V_{i,k_i} is bounded (thus $\mathbb{E}[V^{(i)T} V^{(i)}] \leq 2$), and substituting in the relevant values into Equation (25), we have for all $t > 0$:

$$\mathbb{P} \left(\nabla \ell(u^*)^T \hat{L}^\dagger \nabla \ell(u^*) - \frac{2(d-1)}{m} \geq t \right) \leq 2 \exp \left(-c \frac{m^2 t^2}{d-1} \right).$$

A variable substitution and simple algebra transforms this expression to

$$\mathbb{P} \left[\nabla \ell(u^*)^T \hat{L}^\dagger \nabla \ell(u^*) \geq c_2 \frac{t(d-1)}{m} \right] \leq e^{-t} \quad \text{for all } t > 1,$$

where c_2 is an absolute constant. We may then make the same substitutions as before with expected risk, to obtain,

$$\mathbb{P} \left[\|\hat{u}_{\text{MLE}}(\mathcal{D}) - u^*\|_2^2 > c_2 \frac{t(d-1)}{m \lambda_2(L) \beta_{k_{\max}}^2} \right] \leq e^{-t} \quad \text{for all } t > 1.$$

Setting $d-1 = n(n-1) - 1$, dropping the final minus one term, and making the appropriate substitution with $c_{B,k_{\max}}$, we retrieve the desired tail bound statement, for another absolute constant c .

$$\mathbb{P} \left[\|\hat{u}_{\text{MLE}}(\mathcal{D}) - u^*\|_2^2 \geq c_{B,k_{\max}} \frac{tn(n-1)}{m \lambda_2(L)} \right] \leq e^{-t} \quad \text{for all } t > 1.$$

Integrating the above tail bound leads to a similar bound on the expected risk (same parametric rates), albeit with a less sharp constants due to the added presence of c . \square