One for All: Simultaneous Metric and Preference Learning over Multiple Users

Gregory Canal
University of Wisconsin-Madison
Madison, WI
gcanal@wisc.edu

Blake Mason
Rice University
Houston, TX
bm63@rice.edu

Ramya Korlakai Vinayak
University of Wisconsin-Madison
Madison, WI
ramya@ece.wisc.edu

Robert Nowak
University of Wisconsin-Madison
Madison, WI
rdnowak@wisc.edu

Abstract

This paper investigates simultaneous preference and metric learning from a crowd of respondents. A set of items represented by d-dimensional feature vectors and paired comparisons of the form "item i is preferable to item j" made by each user is given. Our model jointly learns a distance metric that characterizes the crowd's general measure of item similarities along with a latent ideal point for each user reflecting their individual preferences. This model has the flexibility to capture individual preferences, while enjoying a metric learning sample cost that is amortized over the crowd. We first study this problem in a noiseless, continuous response setting (i.e., responses equal to differences of item distances) to understand the fundamental limits of learning. Next, we establish prediction error guarantees for noisy, binary measurements such as may be collected from human respondents, and show how the sample complexity improves when the underlying metric is low-rank. Finally, we establish recovery guarantees under assumptions on the response distribution. We demonstrate the performance of our model on both simulated data and on a dataset of color preference judgments across a large number of users.

1 Introduction

In many data-driven recommender systems (e.g., streaming services, online retail), multiple users interact with a set of items (e.g., movies, products) that are common to all users. While each user has their individual preferences over these items, there may exist shared structure in how users *perceive* items when making preference judgments. This is a reasonable assumption, since collections of users typically have shared perceptions of similarity between items regardless of their individual item preferences [1–3]. In this work we develop and analyze models and algorithms for simultaneously learning individual preferences and the common metric by which users make preference judgments.

Specifically, suppose there exists a known, fixed set $\mathcal{X}$ of n items, where each item $i \in 1, \dots, n$ is parameterized by a feature vector $\boldsymbol{x}_i \in \mathbb{R}^d$. We model the crowd's preference judgments between items as corresponding to a common Mahalanobis distance metric $d_{\boldsymbol{M}}(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_{\boldsymbol{M}}$, where $\|\boldsymbol{x}\|_{\boldsymbol{M}} := \sqrt{\boldsymbol{x}^T \boldsymbol{M} \boldsymbol{x}}$ and $\boldsymbol{M}$ is a $d \times d$ positive semidefinite matrix to be learned. Measuring distances with $d_{\boldsymbol{M}}$ has the effect of reweighting individual features as well as capturing pairwise interactions between features. To capture individual preferences amongst the items, we associate with each of K users an *ideal point* $\boldsymbol{u}_k \in \mathbb{R}^d$ for $k \in 1, \dots, K$ such that user k prefers items that are closer to $\boldsymbol{u}_k$ than those items that are farther away, as measured by the common metric $d_{\boldsymbol{M}}$. The ideal point

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

model is attractive since it can capture nonlinear notions of preference, and preference rankings are determined simply by sorting item distances to each user point and can therefore be easily generalized to items outside of $\mathcal{X}$ with known embedding features [4–7] . Furthermore, once a user's point $\boldsymbol{u}_k$ is estimated, in some generative modeling applications it can then be used to synthesize an "ideal" item for the user located exactly at $\boldsymbol{u}_k$, which by definition would be their most preferred item if it existed.

In order to learn the metric and ideal points, we issue a series of *paired comparison* queries to each user in the form "do you prefer item i or item j?" Since such preferences directly correspond to distance rankings in $\mathbb{R}^d$, these comparisons provide a signal from which the user points $\{\boldsymbol{u}_k\}_{k=1}^K$ and common metric $\boldsymbol{M}$ can be estimated. *The main contribution of this work is a series of identifiability, prediction, and recovery guarantees to establish the first theoretical analysis of simultaneous preference and metric learning from paired comparisons over multiple users.* Our key observation is that by modeling a shared metric between all users rather than learning separate metrics for each user, the sample complexity is reduced from $O(d^2)$ paired comparisons per user to only $O(d)$, which is the sample cost otherwise required to learn each ideal point; in essence, when amortizing metric and preference learning over multiple users, the metric comes for free. Our specific contributions include:

- Necessary and sufficient conditions on the number of items and paired comparisons required for exact preference and metric estimation over generic items, when noiseless differences of item distances are known exactly. These results characterize the fundamental limits of our problem in an idealized setting, and demonstrate the benefit of amortized learning over multiple users. Furthermore, when specialized to $K = 1$ our results significantly advance the existing theory of identifiability for single-user simultaneous metric and preference learning [7].
- Prediction guarantees when learning from noisy, one-bit paired comparisons (rather than exact distance comparisons). We present prediction error bounds for two convex algorithms that learn full-rank and low-rank metrics respectively, and again illustrate the sample cost benefits of amortization.
- Recovery guarantees on the metric and ideal points when learning from noisy, binary labels under assumptions on the response distribution. Furthermore, we empirically validate the recovery performance of our multi-user learning algorithms on both synthetic datasets as well as on real psychometrics data studying individual and collective color preferences and perception.

Summary of related work: Metric and preference learning are both extensively studied problems (see [8] and [9] for surveys of each). A common paradigm in metric learning is that by observing distance comparisons, one can learn a linear [10–12], kernelized [13, 14], or deep metric [15, 16] and use it for downstream tasks such as classification. Similarly, it is common in preference learning to use comparisons to learn a ranking or to identify a most preferred item [5, 6, 17–19]. An important family of these algorithms reduces preference learning to identifying an ideal point for a fixed metric [5, 20]. The closest work to ours is [7], who perform metric and preference learning simultaneously from paired comparisons in the single-user case and propose an alternating minimization algorithm that achieves empirical success. However, that work leaves open the question of theoretical guarantees for the simultaneous learning problem, which we address here. A core challenge when establishing such guarantees is that the data are a function of multiple latent parameters (i.e., unknown metric and ideal point(s)) that interact with each other in a nonlinear manner, which complicates standard generalization and identifiability arguments. To this end, we introduce new theoretical tools and advance the techniques of [12] who showed theoretical guarantees for learning a Mahalanobis metric from triplet queries. We survey additional related work more extensively in Appendix B.

Notation: Let $[K] := 1 \dots K$. Unless specified otherwise, $\|\cdot\|$ denotes the ℓ_2 norm when acting on a vector, and the operator norm induced by the ℓ_2 norm when acting on a matrix. Let $\boldsymbol{e}_i$ denote the ith standard basis vector, $\boldsymbol{1}$ the vector of all ones, $\boldsymbol{0}_{a,b}$ the $a \times b$ matrix of all zeros (or $\boldsymbol{0}$ if the dimensions are clear), and $\boldsymbol{I}$ the identity matrix, where the dimensionality is inferred from context. For a symmetric $d \times d$ matrix $\boldsymbol{A}$, let $\operatorname{vec}^*(\boldsymbol{A}) := [\boldsymbol{A}_{1,1}, \boldsymbol{A}_{1,2}, \dots, \boldsymbol{A}_{1,d}, \boldsymbol{A}_{2,2}, \boldsymbol{A}_{2,3}, \dots, \boldsymbol{A}_{2,d}, \dots \boldsymbol{A}_{d,d}]^T$ denote the vectorized upper triangular portion of $\boldsymbol{A}$, which is a D-length vector where $D := d(d+1)/2$. Let $\boldsymbol{u} \otimes_S \boldsymbol{v} := \operatorname{vec}^*(\boldsymbol{u}\boldsymbol{v}^T)$ denote the unique entries of the Kronecker product between vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, and let $\odot$ denote the Hadamard (or element-wise) product between two matrices.

2 Identifiability from unquantized measurements

In this section, we characterize the fundamental limits on the number of items and paired comparisons per user required to identify $\boldsymbol{M}$ and $\{\boldsymbol{u}_k\}_{k=1}^K$ exactly. In order to understand the fundamental hardness of this problem, we begin by presenting identifiability guarantees under the idealized case where we receive *exact, noiseless* difference of distance measurements[1], before deriving similar results in the case of *noisy* realizations of the *sign* of these differences in the following sections.

We formally define our model as follows: if user k responds that they prefer item i to item j, then $\|\boldsymbol{x}_i - \boldsymbol{u}_k\|_{\boldsymbol{M}} < \|\boldsymbol{x}_j - \boldsymbol{u}_k\|_{\boldsymbol{M}}$. Equivalently, by defining

$$\delta_{i,j}^{(k)} := \|\boldsymbol{x}_i - \boldsymbol{u}_k\|_{\boldsymbol{M}}^2 - \|\boldsymbol{x}_j - \boldsymbol{u}_k\|_{\boldsymbol{M}}^2 = \boldsymbol{x}_i^T \boldsymbol{M} \boldsymbol{x}_i - \boldsymbol{x}_j^T \boldsymbol{M} \boldsymbol{x}_j - 2\boldsymbol{u}_k^T \boldsymbol{M}(\boldsymbol{x}_i - \boldsymbol{x}_j), \tag{1}$$

user k prefers item i over item j if $\delta_{i,j}^{(k)} < 0$ (otherwise j is preferred). In this section, we assume that $\delta_{i,j}^{(k)}$ is measured exactly, and refer to this measurement type as an *unquantized* paired comparison. Let m_k denote the number of unquantized paired comparisons answered by user k and let $m_T := \sum_{k=1}^{K} m_k$ denote the total number of comparisons made across all users.

It is not immediately clear if recovery of both $\boldsymbol{M}$ and $\{\boldsymbol{u}_k\}_{k=1}^K$ is possible from such measurements, which depend quadratically on the item vectors. In particular, one can conceive of pathological examples where these parameters are not identifiable (i.e., there exists no unique solution). For instance, suppose $d = n$, $\boldsymbol{M} = \alpha \boldsymbol{I}$ for a scalar $\alpha > 0$, $\boldsymbol{x}_i = \boldsymbol{e}_i$ for $i \in [n]$, and for each user $\boldsymbol{u}_k = \beta_k \boldsymbol{1}$ for a scalar β_k. Then one can show that $\delta_{i,j}^{(k)} = 0$ for all i, j, k, and therefore $\alpha, \beta_1, \dots, \beta_K$ are unidentifiable from any set of paired comparisons over $\mathcal{X}$. In what follows, we derive necessary and sufficient conditions on the number and geometry of items, number of measurements per user, and interactions between measurements and users in order for the latent parameters to be identifiable.

Note that eq. (1) includes a nonlinear interaction between $\boldsymbol{M}$ and $\boldsymbol{u}_k$; however, by defining $\boldsymbol{v}_k := -2\boldsymbol{M}\boldsymbol{u}_k$ (which we refer to as user k's "pseudo-ideal point," reflecting the component of $\boldsymbol{u}_k$ in the column space of $\boldsymbol{M}$) eq. (1) becomes linear in $\boldsymbol{M}$ and $\boldsymbol{v}_k$:

$$\delta_{i,j}^{(k)} = \boldsymbol{x}_i^T \boldsymbol{M} \boldsymbol{x}_i - \boldsymbol{x}_j^T \boldsymbol{M} \boldsymbol{x}_j + (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{v}_k. \tag{2}$$

If $\boldsymbol{M}$ is full-rank and the system of equations admits unique solutions for $\boldsymbol{M}$ and $\{\boldsymbol{v}_k\}_{k=1}^K$, then $\boldsymbol{u}_k$ can be recovered exactly from $\boldsymbol{v}_k$.[2] In other words, the non-convex eq. (1) can be solved in two stages by first solving a linear relaxation (2) in terms of $\boldsymbol{M}$ and $\boldsymbol{v}_k$, and then solving for $\boldsymbol{u}_k$. Note that since $\boldsymbol{M}$ is symmetric, we may write $\boldsymbol{x}_i^T \boldsymbol{M} \boldsymbol{x}_i = \langle \text{vec}^*(2\boldsymbol{M} - \boldsymbol{I} \odot \boldsymbol{M}), \boldsymbol{x}_i \otimes_S \boldsymbol{x}_i \rangle$. Defining $\boldsymbol{\nu}(\boldsymbol{M}) := \text{vec}^*(2\boldsymbol{M} - \boldsymbol{I} \odot \boldsymbol{M})$, from which $\boldsymbol{M}$ can be determined, we have

$$\delta_{i,j}^{(k)} = \begin{bmatrix} (\boldsymbol{x}_i \otimes_S \boldsymbol{x}_i - \boldsymbol{x}_j \otimes_S \boldsymbol{x}_j)^T & (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\nu}(\boldsymbol{M}) \\ \boldsymbol{v}_k \end{bmatrix}.$$

By concatenating all user measurements in a single linear system, we can directly show conditions for identifiability of $\boldsymbol{M}$ and $\{\boldsymbol{v}_k\}_{k=1}^K$ by characterizing when the system admits a unique solution. To do so, we define a class of matrices that will encode the item indices in each pair queried to each user:

Definition 2.1. A $a \times b$ matrix $\boldsymbol{S}$ is a *selection matrix* if for every $i \in [a]$, there exist distinct indices $p_i, q_i \in [b]$ such that $\boldsymbol{S}[i, p_i] = 1$, $\boldsymbol{S}[i, q_i] = -1$, and $\boldsymbol{S}[i, j] = 0$ for $j \in [b] \setminus \{p_i, q_i\}$.

In Appendix C, we characterize several theoretical properties of selection matrices, which will be useful in proving the results that follow.

For each user k, we represent their queried pairs by a $m_k \times n$ selection matrix denoted $\boldsymbol{S}_k$, where each row selects a pair of items corresponding to its nonzero entries. Letting $\boldsymbol{X} := [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n}$, $\boldsymbol{X}_\otimes := [\boldsymbol{x}_1 \otimes_S \boldsymbol{x}_1, \dots, \boldsymbol{x}_n \otimes_S \boldsymbol{x}_n] \in \mathbb{R}^{D \times n}$, and $\boldsymbol{\delta}_k \in \mathbb{R}^{m_k}$ denote the vector of unquantized measurement values for user k, we can write the entire linear system over all users as a set of m_T equations with $D + dK$ variables to be recovered:

$$\boldsymbol{\Gamma} \begin{bmatrix} \boldsymbol{\nu}(\boldsymbol{M}) \\ \boldsymbol{v}_1 \\ \vdots \\ \boldsymbol{v}_K \end{bmatrix} = \begin{bmatrix} \boldsymbol{\delta}_1 \\ \vdots \\ \boldsymbol{\delta}_K \end{bmatrix} \quad \text{where } \boldsymbol{\Gamma} := \begin{bmatrix} \boldsymbol{S}_1 \boldsymbol{X}_\otimes^T & \boldsymbol{S}_1 \boldsymbol{X}^T & \boldsymbol{0}_{m_1,d} & \cdots & \boldsymbol{0}_{m_1,d} \\ \boldsymbol{S}_2 \boldsymbol{X}_\otimes^T & \boldsymbol{0}_{m_2,d} & \boldsymbol{S}_2 \boldsymbol{X}^T & \cdots & \boldsymbol{0}_{m_2,d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{S}_K \boldsymbol{X}_\otimes^T & \boldsymbol{0}_{m_K,d} & \boldsymbol{0}_{m_K,d} & \cdots & \boldsymbol{S}_K \boldsymbol{X}^T \end{bmatrix}. \tag{3}$$

[1] We use the term "measurement" interchangeably with "paired comparison."

[2] If $\boldsymbol{M}$ were rank deficient, only the component of $\boldsymbol{u}_k$ in the row space of $\boldsymbol{M}$ affects $\delta_{i,j}^{(k)}$. In this case, there is an equivalence class of user points that accurately model their responses. We then take $\boldsymbol{u}_k$ to be the minimum norm solution, i.e., $\boldsymbol{u}_k = -\frac{1}{2}\boldsymbol{M}^\dagger \boldsymbol{v}_k$. This generalizes Proposition 1 of [7] for the multiple user case.

From this linear system, it is clear that $\boldsymbol{\nu}(\boldsymbol{M})$ (and hence $\boldsymbol{M}$) and $\{\boldsymbol{v}_k\}_{k=1}^K$ (and hence $\{\boldsymbol{u}_k\}_{k=1}^K$, if $\boldsymbol{M}$ is full-rank) can be recovered exactly if and only if $\boldsymbol{\Gamma}$ has full column rank. In the following sections, we present necessary and sufficient conditions for this to occur.

2.1 Necessary conditions for identifiability

To build intuition, note that the metric $\boldsymbol{M}$ has D degrees of freedom and each of the K pseudo-ideal points $\boldsymbol{v}_k$ has d degrees of freedom. Hence, there must be at least $m_T \geq D + Kd$ measurements in total (i.e., rows of $\boldsymbol{\Gamma}$) to have any hope of identifying $\boldsymbol{M}$ and $\{\boldsymbol{v}_k\}_{k=1}^K$. When amortized over the K users, this corresponds to each user providing at least $d + {}^{D}/_{K}$ measurements on average. In general, d of these measurements are responsible for identifying each user's own pseudo-ideal point (since $\boldsymbol{v}_k$ is purely a function of user k's responses), while the remaining ${}^{D}/_{K}$ contribute towards a *collective* set of D measurements needed to identify the common metric. While these D measurements must be linearly independent from each other and from those used to learn the ideal points, a degree of overlap is acceptable in the additional d measurements each user provides, as the $\boldsymbol{v}_k$'s are independent of one another. We formalize this intuition in the following proposition, where we let $\boldsymbol{S}_T := [\boldsymbol{S}_1^T, \ldots, \boldsymbol{S}_K^T]^T$ denote the concatenation of all user selection matrices.

Proposition 2.1. *If $\boldsymbol{\Gamma}$ has full column rank, then $\sum_{k=1}^K m_k \geq D + dK$ and the following must hold:*

(a) for all $k \in [K]$, $\operatorname{rank}(\boldsymbol{S}_k \boldsymbol{X}^T) = d$, and therefore $\operatorname{rank}(\boldsymbol{S}_k) \geq d$ and $m_k \geq d$

(b) $\sum_{k=1}^K \operatorname{rank}(\boldsymbol{S}_k \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix}) \geq D + dK$, and therefore $\sum_{k=1}^K \operatorname{rank}(\boldsymbol{S}_k) \geq D + dK$

(c) $\operatorname{rank}(\boldsymbol{S}_T \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix}) = D + d$, and therefore $\operatorname{rank}(\boldsymbol{S}_T) \geq D + d$, $\operatorname{rank}(\begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix}) = D + d$, and $n \geq D + d + 1$

If $\sum_{k=1}^K m_k = D + dK$ exactly, then (a) and (b) are equivalent to $m_k \geq d \; \forall \, k$ and each user's selection matrix having full row rank. (c) implies that the number of required items n scales as $\Omega(d^2)$; in higher dimensional feature spaces, this scaling could present a challenge since it might be difficult in practice to collect such a large number of items for querying. Finally, note that the conditions in Proposition 2.1 are *not* sufficient for identifiability: in Appendix C.6, we present a counterexample where these necessary properties are fulfilled, yet the system is not invertible.

2.2 Sufficient condition for identifiability

Next, we present a class of pair selection schemes that are sufficient for parameter identifiability and match the item and measurement count lower bounds in Proposition 2.1. This result leverages the idea that as long the the d measurements each user provides to learn their ideal point do not "overlap" with the D measurements collectively provided to learn the metric, then the set of m_T total measurements is sufficiently rich to ensure a unique solution. First, we define a property of certain selection matrices where each pair introduces at least one new item that has not yet been selected:

Definition 2.2. An $m \times n$ selection matrix $\boldsymbol{S}$ is *incremental* if for all $i \in [m]$, at least one of the following is true, where p_i and q_i are as defined in Definition 2.1: (a) for all $j < i$, $\boldsymbol{S}[j, p_i] = 0$; (b) for all $j < i$, $\boldsymbol{S}[j, q_i] = 0$.

We now present a class of invertible measurement schemes that builds on the definition of incrementality. For simplicity assume that $m_T = D + dK$ exactly, which is the lower bound from Proposition 2.1. Additionally, assume without loss of generality that each $m_k > d$; if instead there existed a user k^* such that $m_{k^*} = d$ exactly, one can show under the necessary conditions in Proposition 2.1 that the system would separate into two subproblems where first the metric would need to be learned from the other $K - 1$ users, and then $\boldsymbol{v}_{k^*}$ is solved for directly from user k^*'s measurements.

Proposition 2.2. *Let $K \geq 1$, and suppose $m_k > d \; \forall \, k \in [K]$, $m_T = D + dK$, and $n \geq D + d + 1$. Suppose that for each $k \in [K]$, there exists a $d \times n$ selection matrix $\boldsymbol{S}_k^{(1)}$ and $m_k - d \times n$ selection matrix $\boldsymbol{S}_k^{(2)}$ such that $\boldsymbol{S}_k = \begin{bmatrix} (\boldsymbol{S}_k^{(1)})^T & (\boldsymbol{S}_k^{(2)})^T \end{bmatrix}^T$, and that the following are true:*

(a) For all $k \in [K]$, $\operatorname{rank}(\boldsymbol{S}_k^{(1)}) = d$

(b) Defining the $D \times n$ selection matrix $\boldsymbol{S}^{(2)}$ as $\boldsymbol{S}^{(2)} := \left[(\boldsymbol{S}_1^{(2)})^T \cdots (\boldsymbol{S}_K^{(2)})^T \right]^T$, there exists a $D \times D$ permutation $\boldsymbol{P}$ such that for each $k \in [K]$, $\begin{bmatrix} \boldsymbol{S}_k^{(1)} \\ \boldsymbol{P}\boldsymbol{S}^{(2)} \end{bmatrix}$ is incremental

Additionally, suppose each item $\boldsymbol{x}_i$ is sampled i.i.d. from a distribution p_X that is absolutely continuous with respect to the Lebesgue measure. Then with probability 1, $\boldsymbol{\Gamma}$ has full column rank.

Remark 2.3. In Appendix C.6 we construct a pair selection scheme that satisfies the conditions[3] in Proposition 2.2 while only using the minimum number of measurements and items, with $m_k = d + {}^{D}/_{K}$ (and therefore $m_T = D + dK$) and $n = D + d + 1$. Importantly, this construction confirms that the lower bounds on the number of measurements and items in Proposition 2.1 are in fact tight. Since $D = O(d^2)$, if $K = \Omega(d)$ then only $m_k = O(d)$ measurements are required per user. This scaling demonstrates the benefit of amortizing metric learning across multiple users, since in the single user case $D + d = \Omega(d^2)$ measurements would be required.

2.3 Single user case

In the case of a single user ($K = 1$), it is straightforward to show that the necessary and sufficient selection conditions in Proposition 2.1 and Proposition 2.2 respectively are *equivalent*, and simplify to the condition that $\operatorname{rank}(\boldsymbol{S}) \geq D + d$ (where we drop the subscript on $\boldsymbol{S}_1$). In a typical use case, a practitioner is unlikely to explicitly select pair indices that result in $\boldsymbol{S}$ being full-rank, and instead would select pairs uniformly at random from the set of $\binom{n}{2}$ unique item pairs. By proving a tail bound on the number of random comparisons required for $\boldsymbol{S}$ to be full-rank, we have with high probability that randomly selected pairs are sufficient for metric and preference identifiability in the single user case. We summarize these results in the following corollary:

Corollary 2.3.1. *When $K = 1$, if $\boldsymbol{\Gamma}$ is full column rank then* $\operatorname{rank}(\boldsymbol{S}) \geq D + d$*. Conversely, for a fixed $\boldsymbol{S}$ satisfying* $\operatorname{rank}(\boldsymbol{S}) \geq D + d$*, if each $\boldsymbol{x}_i$ is sampled i.i.d. according to a distribution p_X that is absolutely continuous with respect to the Lebesgue measure then $\boldsymbol{\Gamma}$ is full column rank with probability 1. If each pair is selected independently and uniformly at random with $n = \Omega(D + d)$ and $m_T = \Omega(D + d)$, then if $\boldsymbol{x}_i$ is drawn i.i.d. from p_X, $\boldsymbol{\Gamma}$ has full column rank with high probability.*

Importantly, the required item and sample complexity for randomly selected pairs matches the lower bounds in Proposition 2.1 up to a constant. As we describe in Appendix C.7, we conjecture that a similar result holds for the multiuser case ($K > 1$), which is left to future work.

3 Prediction and generalization from binary labels

In practice, we do not have access to exact difference of distance measurements. Instead, paired comparisons are one-bit measurements (given by the user preferring one item over the other) that are sometimes noisy due to inconsistent user behavior or from model deviations. In this case, rather than simply solving a linear system, we must optimize a loss function that penalizes incorrect response predictions while enforcing the structure of our model. In this section, we apply a different set of tools from statistical learning theory to characterize the sample complexity of randomly selected paired comparisons under a general noise model, optimized under a general class of loss functions.

We assume that each pair p is sampled uniformly with replacement from the set of $\binom{n}{2}$ pairs, and the user k queried at each iteration is independently and uniformly sampled from the set of K users. For a pair $p = (i, j)$ given to user k, we observe a (possibly noisy) binary response $y_p^{(k)}$ where $y_p^{(k)} = -1$ indicates that user k prefers item i to j, and $y_p^{(k)} = 1$ indicates that j is preferred. Let $\mathcal{S} := \{(p, k, y_p^{(k)})\}_{p=(i,j)}$ be an i.i.d. joint dataset over pairs p, selected users k, and responses $y_p^{(k)}$, where $|\mathcal{S}|$ denotes the number of such data points. We wish to learn $\boldsymbol{M}$ and vectors $\{\boldsymbol{u}_k\}_{k=1}^K$ that

[3] We note that these conditions are not exhaustive: in Appendix C.6 we construct an example where $\boldsymbol{\Gamma}$ is full column rank, yet the conditions in Proposition 2.2 are not met. A general set of matching necessary and sufficient identifiability conditions on $\{\boldsymbol{S}_k\}_{k=1}^K$ has remained elusive; towards this end, in Appendix C.7 we describe a more comprehensive set of conditions that we conjecture are sufficient for identifiability.

predict the responses in $\mathcal{S}$: given a convex, L-Lipschitz loss $\ell\colon \mathbb{R} \to \mathbb{R}_{\geq 0}$,[4] we wish to solve

$$\begin{aligned} \min_{\boldsymbol{M}, \{\boldsymbol{u}_k\}_{k=1}^K} \; & \frac{1}{|\mathcal{S}|} \sum_{\mathcal{S}} \ell\left(y_p^{(k)} \left(\|\boldsymbol{u}_k - \boldsymbol{x}_i\|_{\boldsymbol{M}}^2 - \|\boldsymbol{u}_k - \boldsymbol{x}_j\|_{\boldsymbol{M}}^2 \right) \right) \\ \text{s.t. } & \boldsymbol{M} \succeq 0, \|\boldsymbol{M}\|_F \leq \lambda_F, \|\boldsymbol{u}_k\|_2 \leq \lambda_u \; \forall\, k \in [K], |\delta_{i,j}^{(k)}| \leq \gamma \; \forall\, i,j,k \end{aligned} \tag{4}$$

where $\lambda_F, \lambda_u, \gamma > 0$ are hyperparameters and $\delta_p^{(k)}$ is defined as in eq. (1). The constraint $\boldsymbol{M} \succeq 0$ ensures that $\boldsymbol{M}$ defines a metric, the Frobenius and ℓ_2 norm constraints prevent overfitting, and the constraint on $\delta_p^{(k)}$ is a technical point to avoid pathological cases stemming from coherent $\boldsymbol{x}$ vectors.

The above optimization is non-convex due to the interaction between the $\boldsymbol{M}$ and $\boldsymbol{u}$ terms. Instead, as in Section 2 we define $\boldsymbol{v}_k := -2\boldsymbol{M}\boldsymbol{u}_k$ and solve the relaxation[5]

$$\begin{aligned} & \min_{\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K} \widehat{R}(\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K) \text{ s.t. } \boldsymbol{M} \succeq 0, \|\boldsymbol{M}\|_F \leq \lambda_F, \|\boldsymbol{v}_k\|_2 \leq \lambda_v \; \forall\, k \in [K], |\delta_{i,j}^{(k)}| \leq \gamma \; \forall\, i,j,k \\ & \text{where } \widehat{R}(\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K) := \frac{1}{|\mathcal{S}|} \sum_{\mathcal{S}} \ell\left(y_p^{(k)} \left(\boldsymbol{x}_i^T \boldsymbol{M} \boldsymbol{x}_i - \boldsymbol{x}_j^T \boldsymbol{M} \boldsymbol{x}_j + \boldsymbol{v}_k^T (\boldsymbol{x}_i - \boldsymbol{x}_j) \right) \right). \end{aligned} \tag{5}$$

The quantity $\widehat{R}(\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K)$ is the *empirical risk*, given dataset $\mathcal{S}$. The empirical risk is an unbiased estimate of the true risk given by

$$R(\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K) := \mathbb{E}\left[\ell\left(y_p^{(k)} \left(\boldsymbol{x}_i^T \boldsymbol{M} \boldsymbol{x}_i - \boldsymbol{x}_j^T \boldsymbol{M} \boldsymbol{x}_j + \boldsymbol{v}_k^T (\boldsymbol{x}_i - \boldsymbol{x}_j) \right) \right) \right],$$

where the expectation is with respect to a random draw of $p = (i, j)$, k, and $y_p^{(k)}$ conditioned on the choice of p and k. Let $\widehat{\boldsymbol{M}}$ and $\{\widehat{\boldsymbol{v}}_k\}_{k=1}^K$ denote the minimizers of the empirical risk optimization in eq. (5), and let $\boldsymbol{M}^*$ and $\{\boldsymbol{v}_k^*\}_{k=1}^K$ minimize the *true* risk, subject to the same constraints. The following theorem bounds the excess risk of the empirical optimum $R(\widehat{\boldsymbol{M}}, \{\widehat{\boldsymbol{v}}_k\}_{k=1}^K)$ relative to the optimal true risk $R(\boldsymbol{M}^*, \{\boldsymbol{v}_k^*\}_{k=1}^K)$.

Theorem 3.1. *Suppose $\|\boldsymbol{x}_i\|_2 \leq 1$ for all $i \in [n]$. With probability at least $1 - \delta$,*

$$\begin{aligned} R(\widehat{\boldsymbol{M}}, \{\widehat{\boldsymbol{v}}_k\}_{k=1}^K) - R(\boldsymbol{M}^*, \{\boldsymbol{v}_k^*\}_{k=1}^K) \leq & \sqrt{\frac{256 L^2 (\lambda_F^2 + K\lambda_v^2)}{|\mathcal{S}|} \log(d^2 + d + 1)} \\ & + \frac{\sqrt{128 L^2 (\lambda_F^2 + K\lambda_v^2)}}{3|\mathcal{S}|} \log(d^2 + d + 1) + \sqrt{\frac{8 L^2 \gamma^2 \log(\frac{2}{\delta})}{|\mathcal{S}|}}. \end{aligned} \tag{6}$$

Remark 3.2. To put this result in context, suppose $\|\boldsymbol{M}^*\|_F = d$ so that the average squared magnitude of each entry is a constant, in which case we can set $\lambda_F = d$. Similarly, if each entry of $\boldsymbol{v}_k$ is dimensionless, then $\|\boldsymbol{v}_k\|_2 \propto \sqrt{d}$ and so we can set $\lambda_v = \sqrt{d}$. We then have that the excess risk in eq. (6) is $\widetilde{O}\left(\sqrt{\frac{d^2 + Kd}{|\mathcal{S}|}} \right)$ where $\widetilde{O}$ suppresses logarithmic factors, implying a sample complexity of $d^2 + Kd$ measurements across all users, and therefore an average of $d + d^2/K$ measurements per user. If $K = \Omega(d)$, this is equivalent to $\widetilde{O}(d)$ measurements per user, which corresponds to the parametric rate required per user in order to estimate their pseudo-ideal point $\boldsymbol{v}_k$. Similar to the case of unquantized measurements, the $O(d^2)$ sample cost of estimating the metric from noisy one-bit comparisons has been amortized across all users, demonstrating the benefit of learning multiple user preferences simultaneously when the users share a common metric.

3.1 Low-rank modeling

In many settings, the metric $\boldsymbol{M}$ may be low-rank with rank $r < d$ [8, 12]. In this case, $\boldsymbol{M}$ only has dr degrees of freedom rather than d^2 degrees as in the full-rank case. Therefore if $K = \Omega(d)$, we intuitively expect the sample cost of learning the metric to be amortized to a cost of $O(r)$

[4] We restrict ourselves to the case where the loss is a function of $y_p^{(k)} \left(\|\boldsymbol{u}_k - \boldsymbol{x}_i\|_{\boldsymbol{M}}^2 - \|\boldsymbol{u}_k - \boldsymbol{x}_j\|_{\boldsymbol{M}}^2 \right)$.

[5] Specifically we may choose $\lambda_v = 2\lambda_F \lambda_u$, resulting in a constraint set containing the solution to (4). To see this, let $\boldsymbol{M}^*, \{\boldsymbol{u}_k^*\}_{k=1}^K$ be the solution to (4) and define $\boldsymbol{v}_k^* = -2\boldsymbol{M}^* \boldsymbol{u}_k^*$. Then $\|\boldsymbol{v}_k^*\|_2 = 2\|\boldsymbol{M}^* \boldsymbol{u}_k^*\|_2 \leq 2\|\boldsymbol{M}^*\| \|\boldsymbol{u}_k^*\|_2 \leq 2\|\boldsymbol{M}^*\|_F \|\boldsymbol{u}_k^*\|_2 \leq 2\lambda_F \lambda_u$.

measurements per user. Furthermore, as each $\boldsymbol{v}_k$ is contained in the r-dimensional column space of $\boldsymbol{M}$, we also expect a sample complexity of $O(r)$ to learn each user's pseudo-ideal point. Hence, we expect the amortized sample cost per user to be $O(r)$ in the low-rank setting, which can be a significant improvement over $O(d)$ in the full-rank setting when $r \ll d$.

Algorithmically, ideally one would constrain the $\widehat{\boldsymbol{M}}$ and $\{\widehat{\boldsymbol{v}}_k\}_{k=1}^K$ that minimize the empirical risk such that $\operatorname{rank}(\boldsymbol{M}) = r$ and $\boldsymbol{v}_k \in \operatorname{colsp}(\boldsymbol{M})$; unfortunately, such constraints are not convex. Towards a convex algorithm, note that since $\boldsymbol{v}_k \in \operatorname{colsp}(\boldsymbol{M})$, $\operatorname{rank}([\, \boldsymbol{M}, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_K \,]) = \operatorname{rank}(\boldsymbol{M}) = r$. Thus, it is sufficient to constrain the rank of $[\, \boldsymbol{M}, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_K \,]$. We relax this constraint to a convex constraint on the nuclear norm $\|[\, \boldsymbol{M} \; \boldsymbol{v}_1 \; \cdots \; \boldsymbol{v}_K \,]\|_*$, and solve a similar optimization problem to eq. (5):

$$\min_{\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K} \widehat{R}(\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K) \text{ s.t. } \boldsymbol{M} \succeq 0, \|[\boldsymbol{M} \quad \boldsymbol{v}_1 \quad \cdots \quad \boldsymbol{v}_K]\|_* \leq \lambda_*, |\delta_{i,j}^{(k)}| \leq \gamma \,\forall\, i, j, k. \quad (7)$$

We again let $\boldsymbol{M}^*$ and $\{\boldsymbol{v}_k^*\}_{k=1}^K$ minimize the true risk $R(\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K)$, subject to the same constraints. The following theorem bounds the excess risk over this constraint set:

Theorem 3.3. *Suppose* $\|\boldsymbol{x}_i\|_2 \leq 1$ *for all* $i \in [n]$. *With probability at least* $1 - \delta$,

$$R(\widehat{\boldsymbol{M}}, \{\widehat{\boldsymbol{v}}_k\}_{k=1}^K) - R(\boldsymbol{M}^*, \{\boldsymbol{v}_k^*\}_{k=1}^K) \leq 2L\sqrt{\frac{2\lambda_*^2 \log(2d+K)}{|\mathcal{S}|}\left[\left(8 + \frac{4\min(d,n)}{K}\right)\frac{\|\boldsymbol{X}\|^2}{n} + \frac{16}{\sqrt{K}}\right]} + \frac{8L\lambda_*}{3|\mathcal{S}|}\log(2d+K) + \sqrt{\frac{8L^2\gamma^2\log(2/\delta)}{|\mathcal{S}|}}.$$

To put this result in context, suppose that the items $\boldsymbol{x}_i$ and ideal points $\boldsymbol{u}_k$ are sampled i.i.d. from $\mathcal{N}(\boldsymbol{0}, \frac{1}{d}\boldsymbol{I})$. With this item distribution it is straightforward to show that with high probability, $\|\boldsymbol{X}\|^2 = O(\frac{n}{d})$ (see [21]). For a given $r < d$ let $\boldsymbol{M} = \frac{d}{\sqrt{r}}\boldsymbol{L}\boldsymbol{L}^T$, where $\boldsymbol{L}$ is a $d \times r$ matrix with orthonormal columns sampled uniformly from the Grassmanian. With this choice of scaling we have $\|\boldsymbol{M}\|_F = d$, so that each element of $\boldsymbol{M}$ is dimensionless on average. Furthermore, recalling that $\boldsymbol{v}_k = -2\boldsymbol{M}\boldsymbol{u}_k$, with this choice of scaling $\mathbb{E}[\|\boldsymbol{v}_k\|_2^2] \propto d$ and so each entry of $\boldsymbol{v}_k$ on average is dimensionless. To choose a setting for λ^* recall that $[\boldsymbol{M}, \boldsymbol{v}_1, \ldots \boldsymbol{v}_K]$ has rank r and therefore

$$\|[\boldsymbol{M} \quad \boldsymbol{v}_1 \quad \cdots \quad \boldsymbol{v}_K]\|_* \leq \sqrt{r}\|[\boldsymbol{M} \quad \boldsymbol{v}_1 \quad \cdots \quad \boldsymbol{v}_K]\|_F \leq \sqrt{r(d^2 + K \max_{k \in [K]} \|\boldsymbol{v}_k\|_2^2)},$$

which one can show is $O(\sqrt{r(d^2 + dK\log K)})$ with high probability and so we set $\lambda_* = O\left(\sqrt{r(d^2 + dK\log K)}\right)$. With these term scalings, we have the following corollary:

Corollary 3.3.1. *Let* $n \geq d$, $\boldsymbol{x}_i, \boldsymbol{u}_k \sim \mathcal{N}(\boldsymbol{0}, \frac{1}{d}\boldsymbol{I})$ *and* $\boldsymbol{M} = \frac{d}{\sqrt{r}}\boldsymbol{L}\boldsymbol{L}^T$, *where* $\boldsymbol{L}$ *is a* $d \times r$ *matrix with orthonormal columns. If* $K = \Omega(d^2)$, *then in the same setting as Theorem 3.3 with high probability*

$$R(\widehat{\boldsymbol{M}}, \{\widehat{\boldsymbol{v}}_k\}_{k=1}^K) - R(\boldsymbol{M}^*, \{\boldsymbol{v}_k^*\}_{k=1}^K) = \widetilde{O}\left(\sqrt{\frac{dr + Kr}{|\mathcal{S}|}}\right).$$

Remark 3.4. The scaling $|\mathcal{S}| = O(dr + Kr)$ matches our intuition that $O(dr)$ collective measurements should be made across all users to account for the dr degrees of freedom in $\boldsymbol{M}$, in addition to $O(r)$ measurements per user to resolve their own pseudo-ideal point's r degrees of freedom. If $K = \Omega(d)$, then each user answering $O(r)$ queries is sufficient to amortize the cost of learning the metric with the same order of measurements per user as is required for their ideal point. Although Corollary 3.3.1 requires the even stronger condition that $K = \Omega(d^2)$, we believe this is an artifact of our analysis and that $K = \Omega(d)$ should suffice.[6] Even so, a $\Omega(d^2)$ user count scaling might be reasonable in practice since recommender systems typically operate over large populations of users.

[6] The required user scaling of $K = \Omega(d^2)$ implies that Corollary 3.3.1 only recovers an amortized scaling of $O(r)$ measurements per user if the user count is very large. Nevertheless, the original statement of Theorem 3.3 *does* apply for any user count, and in this case the only drawback to not invoking $K = \Omega(d^2)$ is that the amortized scaling is larger than $O(r + dr/K)$ measurements per user. We believe this scaling can in fact be tightened to $O(r + dr/K)$ measurements per user for *all* user counts K (not just $K = \Omega(d^2)$), which we leave to future work. See the proof of Corollary 3.3.1 for additional discussion.

4 Recovery guarantees

The results in the previous section give guarantees on the generalization error of a learned metric and ideal points when predicting pair responses over $\mathcal{X}$, but do not bound the recovery error of the learned parameters $\widehat{\boldsymbol{M}}$, $\{\widehat{\boldsymbol{v}}_k\}_{k=1}^K$ with respect to $\boldsymbol{M}^*$ and $\{\boldsymbol{v}_k^*\}_{k=1}^K$. Yet, in some settings such as data generated from human responses [22,23] it may be reasonable to assume that a true $\boldsymbol{M}^*$ and $\{\boldsymbol{v}_k^*\}_{k=1}^K$ do exist that generate the observed data (rather than serving only as a model) and that practitioners may wish to estimate and interpret these latent variables, in which case accurate recovery is critical. Unfortunately, for an arbitrary noise model and loss function, recovering $\boldsymbol{M}^*$ and $\{\boldsymbol{v}_k^*\}_{k=1}^K$ exactly is generally impossible if the model is not identifiable. However, we now show that with a small amount of additional structure, one can ensure that $\widehat{\boldsymbol{M}}$ and $\{\widehat{\boldsymbol{v}}_k\}_{k=1}^K$ accurately approximate $\boldsymbol{M}^*$ and $\{\boldsymbol{v}_k^*\}_{k=1}^K$ if a sufficient number of one-bit comparisons are collected.

We assume a model akin to that of [12] for the case of triplet metric learning. Let $f\colon \mathbb{R} \to [0,1]$ be a strictly monotonically increasing *link function* satisfying $f(x) = 1 - f(-x)$; for example, $f(x) = (1+e^{-x})^{-1}$ is the logistic link and $f(x) = \Phi(x)$ is the probit link where $\Phi(\cdot)$ denotes the CDF of a standard normal distribution. Defining $\delta_p(\boldsymbol{M}, \boldsymbol{v}) := \boldsymbol{x}_i^T \boldsymbol{M} \boldsymbol{x}_i - \boldsymbol{x}_j^T \boldsymbol{M} \boldsymbol{x}_j + \boldsymbol{v}^T(\boldsymbol{x}_i - \boldsymbol{x}_j)$ for $p = (i,j)$, we assume that $\mathbb{P}(y_p^{(k)} = -1) = f\left(-\delta_p(\boldsymbol{M}^*, \boldsymbol{v}_k^*)\right)$ for some $\boldsymbol{M}^* \succeq \boldsymbol{0}$ and $\boldsymbol{v}_k^* \in \operatorname{colsp}(\boldsymbol{M}^*)$. This naturally reflects the idea that some queries are easier to answer (and thus less noisy) than others. For instance, if $\delta_{ij}^{(k)} \ll 0$ such as may occur when $\boldsymbol{x}_i$ very nearly equals user k's ideal point, we may assume that user k almost always prefers item i to j and so $f(-\delta_{ij}^{(k)}) \to 1$ (since f is monotonic). Furthermore, we assume that eq. (5) is optimized with the *negative log-likelihood* loss ℓ_f induced by f: $\ell_f(y_p, p; \boldsymbol{M}, \boldsymbol{v}) := -\log(f(y_p \delta_p(\boldsymbol{M}, \boldsymbol{v}_k)))$. In Appendix E, we show that we may lower bound the excess risk of $\widehat{\boldsymbol{M}}$, $\{\widehat{\boldsymbol{v}}_k\}_{k=1}^K$ by the squared error between the *unquantized* measurements corresponding to $\widehat{\boldsymbol{M}}$, $\{\widehat{\boldsymbol{v}}_k\}_{k=1}^K$ and $\boldsymbol{M}^*$, $\{\boldsymbol{v}_k^*\}_{k=1}^K$. We then utilize tools from Section 2 combined with the results in Section 3 to arrive at the following recovery guarantee.

Theorem 4.1. *Fix a strictly monotonic link function f satisfying $f(x) = 1 - f(-x)$. Suppose for a given item set $\mathcal{X}$ with $n \geq D + d + 1$ and $\|\boldsymbol{x}_i\| \leq 1 \, \forall\, i \in [n]$ that the pairs and users in dataset $\mathcal{S}$ are sampled independently and uniformly at random, and that user responses are sampled according to $\mathbb{P}(y_p^{(k)} = -1) = f\left(-\delta_p(\boldsymbol{M}^*, \boldsymbol{v}_k^*)\right)$ where $\boldsymbol{M}^*, \{\boldsymbol{v}_k^*\}_{k=1}^K$ satisfy the constraints in (5). Let $\widehat{\boldsymbol{M}}$, $\{\widehat{\boldsymbol{v}}_k\}_{k=1}^K$ be the solution to (5) solved using loss ℓ_f. Then with probability at least $1 - \delta$,*

$$\frac{1}{n}\sigma_{\min}\left(\boldsymbol{J}[\boldsymbol{X}_\otimes^T, \boldsymbol{X}^T]\right)^2 \left(\|\widehat{\boldsymbol{M}} - \boldsymbol{M}^*\|_F^2 + \frac{1}{K}\sum_{k=1}^K \|\widehat{\boldsymbol{v}}_k - \boldsymbol{v}_k^*\|^2 \right) \leq$$

$$\frac{4}{C_f^2}\sqrt{\frac{L^2(\lambda_F^2 + K\lambda_v^2)}{|\mathcal{S}|}\log(d^2+d+1)} + \frac{\sqrt{8L^2(\lambda_F^2 + K\lambda_v^2)}}{3C_f^2|\mathcal{S}|}\log(d^2+d+1) + \frac{1}{C_f^2}\sqrt{\frac{L^2\gamma^2\log(\frac{2}{\delta})}{2|\mathcal{S}|}},$$

where $C_f = \min_{z:|z|\leq\gamma} f'(z)$ and $\boldsymbol{J} := \boldsymbol{I}_n - \frac{1}{n}\boldsymbol{1}_n\boldsymbol{1}_n^T$ is the centering matrix. Furthermore, if $\mathcal{X}$ is constructed by sampling each item i.i.d. from a distribution p_X with support on the unit ball that is absolutely continuous with respect to the Lebesgue measure, then with probability 1, $\sigma_{\min}\left(\boldsymbol{J}[\boldsymbol{X}_\otimes^T, \boldsymbol{X}^T]\right) > 0$.

Remark 4.2. The key conclusion from this result is that since $\sigma_{\min}\left(\boldsymbol{J}[\boldsymbol{X}_\otimes^T, \boldsymbol{X}^T]\right) > 0$ almost surely, the recovery error of $\widehat{\boldsymbol{M}}$, $\{\widehat{\boldsymbol{v}}_k\}_{k=1}^K$ with respect to $\boldsymbol{M}^*$, $\{\boldsymbol{v}_k^*\}_{k=1}^K$ is upper bounded by a decreasing function of the number of data points, $|\mathcal{S}|$. In other words, the metric and pseudo-ideal points are identifiable from one-bit paired comparisons under an assumed response distribution. As discussed in Section 2, the ideal points $\{\boldsymbol{u}_k^*\}_{k=1}^K$ are then also identifiable as long as $\boldsymbol{M}^*$ is full-rank. We present an analogous result for the case of a low-rank metric in Appendix E, and leave to future work a study of the scaling of $\sigma_{\min}\left(\boldsymbol{J}[\boldsymbol{X}_\otimes^T, \boldsymbol{X}^T]\right)$ with respect to d and n.

5 Experimental results

Below we analyze the performance of the empirical risk minimizers given in eqs. (5) and (7) on both simulated and real-world data,[7] with further details deferred to Appendix F.

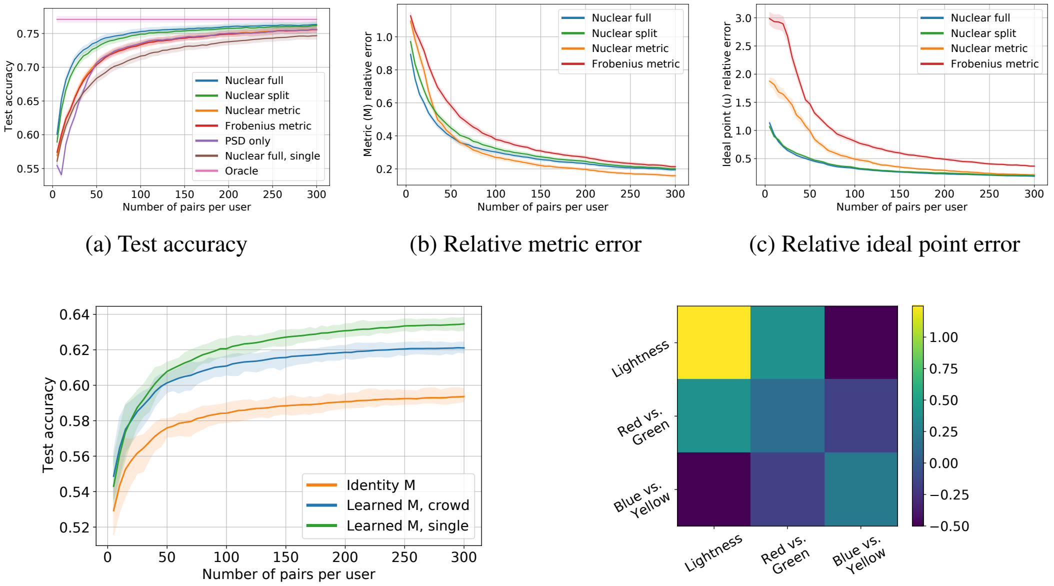
[7] Code available at `https://github.com/gregcanal/multiuser-metric-preference`

Simulated experiments: We first simulate data in a similar setting to Cor. 3.3.1 where $\boldsymbol{x}_i, \boldsymbol{u}_k \sim \mathcal{N}(\boldsymbol{0}, \frac{1}{d}\boldsymbol{I})$ and $\boldsymbol{M}^* = \frac{d}{\sqrt{r}}\boldsymbol{L}\boldsymbol{L}^T$ where $\boldsymbol{L} \in \mathbb{R}^{d \times r}$ is a random orthogonal matrix. To construct the training dataset, we query a fixed number of randomly selected pairs per user and evaluate prediction accuracy on a held-out test set, where all responses are generated according to a logistic link function (injecting response noise). We evaluate the prediction accuracy of the Frobenius norm regularized optimization in eq. (5) (referred to as **Frobenius metric**), designed for full-rank matrix recovery, as well as the nuclear norm regularized optimization in eq. (7) (referred to as **Nuclear full**), designed for low-rank metrics. We also compare against several ablation methods that modify the constraint sets in (5) and (7): **Nuclear metric**, where $\|\boldsymbol{M}\|_*$ and $\|\boldsymbol{v}_k\|_2$ are constrained; **Nuclear split**, where $\|\boldsymbol{M}\|_*$ and $\|[\boldsymbol{v}_1, \cdots, \boldsymbol{v}_K]\|_*$ are constrained; and **PSD only**, where only $\boldsymbol{M} \succeq \boldsymbol{0}$ is enforced. We also compare against **Nuclear full, single**, which is equivalent to **Nuclear full** when applied *separately* to each user (learning a unique metric and ideal point), where test accuracy is averaged over all users. To compare performance under a best-case hyperparameter setting, we tune each method's respective constraints using oracle knowledge of $\boldsymbol{M}^*$ and $\{\boldsymbol{u}_k^*\}_{k=1}^K$. Finally, we also evaluate prediction accuracy when the ground-truth parameters are known exactly (i.e., $\boldsymbol{M} = \boldsymbol{M}^*, \boldsymbol{v}_k = -2\boldsymbol{M}^*\boldsymbol{u}_k^*$), which we call **Oracle**. This accuracy gives the "best case" performance, and reflects the inherent response noise in the generated data.

To test a low-rank setting, we set $d = 10$, $r = 1$, $n = 100$, and $K = 10$. We observe that **Nuclear full** outperforms the baseline methods in terms of test accuracy, and is closely followed by **Nuclear split** (Fig. 1a). Interestingly **Nuclear metric**, which also enforces a nuclear norm constraint on $\boldsymbol{M}$, does not perform as well, possibly because it does not encourage the pseudo-ideal points to lie in the same low-rank subspace. While **Nuclear metric** does demonstrate slightly improved metric recovery (Fig. 1b), **Nuclear full** and **Nuclear split** recover higher quality metrics for lower query counts (which is the typical operating regime for human-in-the-loop systems) and exhibit significantly better ideal point recovery (Fig. 1c), illustrating the importance of proper subspace alignment between the pseudo-ideal points. To this end, unlike **Nuclear split**, **Nuclear full** explicitly encourages the pseudo-ideal points to align with the column space of $\boldsymbol{M}$, which may explain its slight advantage. Finally, we note that **Nuclear full, single** results in the worst prediction accuracy, demonstrating the benefit of a common metric model when the underlying metric is in fact shared. While the single user case is not the focus of this work, in Appendix F.4 we compare the performance of **Nuclear full, single** against the methods proposed in [7], which only considers the single user case.

Color dataset: We also study the performance of our model on a dataset of pairwise color preferences across multiple respondents ($K = 48$) [24]. In this setting, each color ($n = 37$) is represented as a 3-dimensional vector in CIELAB color space (lightness, red vs. green, blue vs. yellow), which was designed as a uniform space for how humans perceive color [25]. Each respondent was asked to order pairs of color by preference, as described in [26, Sec. 3.1]. Since all $2\binom{37}{2}$ possible pairs (including each pair reversal) were queried for each respondent, we may simulate random pair sampling exactly.

As there are only $d = 3$ features, we constrain the Frobenius norm of the metric and optimize eq. (5) using the hinge loss. Varying the number of pairs queried per user, we plot prediction accuracy on a held-out test set (**Learned M, crowd** in Fig. 1d). As CIELAB is designed to be perceptually uniform, we compare against a solution to eq. (5) that *fixes* $\boldsymbol{M} = \boldsymbol{I}$ and only learns the points $\{\boldsymbol{v}_k\}_{k=1}^{48}$ (**Identity M** in Fig. 1d). This method leads to markedly lower prediction accuracy than simultaneously learning the metric and ideal points; this result suggests that although people's *perception* of color is uniform in this space, their *preferences* are not. We also compare against a baseline that solves the same optimization as eq. (5) *separately* for each individual respondent (learning a unique metric and ideal point per user), with prediction accuracy averaged over all respondents (**Learned M, single** in Fig. 1d). Although learning individual metrics appears to result in better prediction after many queries, in the low-query regime (< 20 pairs per user) learning a common metric across all users results in slightly improved performance (see Appendix F for zoomed plot). As $d = 3$ is small relative to the number of queries given to each user, the success of individual metric learning is not unexpected; however, collecting $O(d^2)$ samples per user is generally infeasible for larger d unlike collective metric learning which benefits from crowd amortization. Finally, learning a single metric common to all users allows for insights into the crowd's general measure of color similarity. As can be seen in Fig. 1e, the learned metric is dominated by the "lightness" feature, indicating that people's preferences correspond most strongly to a color's lightness. As an external validation, this is consistent with the findings of Fig. 1 of [24].



(a) Test accuracy

(b) Relative metric error

(c) Relative ideal point error

(d) Color preference prediction accuracy

(e) Average $\widehat{\boldsymbol{M}}$ over all color trials

Figure 1: (a-c) Normally distributed items with $d = 10$, $r = 1$, $n = 100$ and $K = 10$. Error bars indicate ±1 standard error about the sample mean. For visual clarity, **PSD only** and **Nuclear full, single** baselines are omitted from (b-c) due to poor performance. (d) Average color preference prediction accuracy, where error bars indicate 2.5% and 97.5% percentiles. (e) Estimated color preference metric. For (a-e), random train/test splitting was repeated over 30 trials.

6 Discussion

The main contribution of this work is a model for multi-user simultaneous metric and preference learning consisting of a shared metric that captures the crowd's common perceptions between items, as well as user-specific ideal points that characterize individual preferences. Our core result is that when querying paired comparisons over a large number of users, the sample cost of metric learning is distributed over the crowd, with the total number of queries per user scaling with the rank of the underlying metric.

One interesting avenue for future work is to study more flexible metric models, since as with any Mahalanobis metric model it may not be the case that linear weightings of quadratic feature combinations provide enough flexibility to adequately model certain preference judgements. While it may be possible to generalize the linear metric results studied here to a more general Hilbert space, another avenue to increase model flexibility is to acquire a richer set of features in the item set, which would typically involve increasing the ambient dimension. In this setting, an important consideration would be to ameliorate the requirement that the number of items scales as $\Omega(d^2)$. While this condition is necessary for identifiability if the metric is full-rank, we believe that the item count should only need to scale as $\Omega(dr)$ if the metric has rank $r < d$ (see Appendix C.3 for additional discussion). Furthermore, in the case of metric learning from triplet queries it is known that only $\Omega(r)$ items are required for metric recovery in the low-rank case [12]. This suggests a potential strategy of supplementing paired comparisons with triplet queries to reduce the number of required items.

Acknowledgements and Disclosure of Funding

We thank the reviewers for their feedback, as well as Karen Schloss for providing the color preference data used in our experiments and for useful discussions and references about color preference learning. This work was partially supported by AFOSR/AFRL grant FA9550-18-1-0166 and ARMY MURI grant W911NF-15-1-0479.

References

[1] R. N. Shepard, "The analysis of proximities: Multidimensional scaling with an unknown distance function. i." *Psychometrika*, vol. 27, no. 2, pp. 125–140, 1962.

[2] C. H. Coombs, "A theory of data." 1964.

[3] O. Tamuz, C. Liu, S. Belongie, O. Shamir, and A. T. Kalai, "Adaptively learning the crowd kernel," *arXiv preprint arXiv:1105.1033*, 2011.

[4] G. S. Carpenter and K. Nakamoto, "Consumer preference formation and pioneering advantage," *Journal of Marketing research*, vol. 26, no. 3, pp. 285–298, 1989.

[5] K. G. Jamieson and R. Nowak, "Active ranking using pairwise comparisons," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, Eds., vol. 24. Curran Associates, Inc., 2011. [Online]. Available: https://proceedings.neurips.cc/paper/2011/file/6c14da109e294d1e8155be8aa4b1ce8e-Paper.pdf

[6] G. Canal, A. Massimino, M. Davenport, and C. Rozell, "Active embedding search via noisy paired comparisons," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 902–911. [Online]. Available: https://proceedings.mlr.press/v97/canal19a.html

[7] A. Xu and M. Davenport, "Simultaneous preference and metric learning from paired comparisons," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 454–465. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/0561bc7ecba98e39ca7994f93311ba23-Paper.pdf

[8] A. Bellet, A. Habrard, and M. Sebban, "Metric learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 9, no. 1, pp. 1–151, 2015.

[9] J. Fürnkranz and E. Hüllermeier, "Preference learning and ranking by pairwise comparison," in *Preference learning*. Springer, 2010, pp. 65–82.

[10] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in neural information processing systems*, 2006, pp. 1473–1480.

[11] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 209–216.

[12] B. Mason, L. Jain, and R. Nowak, "Learning low-dimensional metrics," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/f12ee9734e1edf70ed02d9829018b3d9-Paper.pdf

[13] R. Chatpatanasiri, T. Korsrilabutr, P. Tangchanachaianan, and B. Kijsirikul, "A new kernelization framework for mahalanobis distance learning algorithms," *Neurocomputing*, vol. 73, no. 10-12, pp. 1570–1579, 2010.

[14] M. Kleindessner and U. von Luxburg, "Kernel functions based on triplet comparisons," *arXiv preprint arXiv:1607.08456*, 2016.

[15] M. Kaya and H. Ş. Bilge, "Deep metric learning: A survey," *Symmetry*, vol. 11, no. 9, p. 1066, 2019.

[16] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International workshop on similarity-based pattern recognition*. Springer, 2015, pp. 84–92.

[17] K. G. Jamieson and R. D. Nowak, "Low-dimensional embedding using adaptively selected ordinal data," in *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2011, pp. 1077–1084.

[18] K.-S. Jun, L. Jain, B. Mason, and H. Nassif, "Improved confidence bounds for the linear logistic model and applications to bandits," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5148–5157.

[19] G. H. Canal, M. R. O'Shaughnessy, C. J. Rozell, and M. A. Davenport, "Joint estimation of trajectory and dynamics from paired comparisons," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2019, pp. 121–125.

[20] A. K. Massimino and M. A. Davenport, "As you like it: Localization via paired comparisons," *Journal of Machine Learning Research*, vol. 22, no. 186, pp. 1–39, 2021.

[21] K. R. Davidson and S. J. Szarek, "Local operator theory, random matrices and banach spaces," *Handbook of the geometry of Banach spaces*, vol. 1, no. 317-366, p. 131, 2001.

[22] B. Mason, M. A. Rau, and R. Nowak, "Cognitive task analysis for implicit knowledge about visual representations with similarity learning methods," *Cognitive Science*, vol. 43, no. 9, p. e12744, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12744

[23] M. A. Rau, B. Mason, and R. Nowak, "How to model implicit knowledge? similarity learning methods to assess perceptions of visual representations." *International Educational Data Mining Society*, 2016.

[24] S. E. Palmer and K. B. Schloss, "An ecological valence theory of human color preference," *Proceedings of the National Academy of Sciences*, vol. 107, no. 19, pp. 8877–8882, 2010.

[25] K. B. Schloss, L. Lessard, C. Racey, and A. C. Hurlbert, "Modeling color preference using color space metrics," *Vision Research*, vol. 151, pp. 99–116, 2018.

[26] S. E. Palmer, K. B. Schloss, and J. Sammartino, "Visual aesthetics and human preference," *Annual review of psychology*, vol. 64, pp. 77–107, 2013.

[27] B. Kulis *et al.*, "Metric learning: A survey," *Foundations and Trends® in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2013.

[28] H.-J. Ye, D.-C. Zhan, and Y. Jiang, "Fast generalization rates for distance metric learning," *Machine Learning*, vol. 108, no. 2, pp. 267–295, 2019.

[29] F. Liu, X. Huang, Y. Chen, and J. Suykens, "Fast learning in reproducing kernel krein spaces via signed measures," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2021, pp. 388–396.

[30] M. Huai, H. Xue, C. Miao, L. Yao, L. Su, C. Chen, and A. Zhang, "Deep metric learning: The generalization analysis and an adaptive algorithm." in *IJCAI*, 2019, pp. 2535–2541.

[31] G. Canal, S. Fenu, and C. Rozell, "Active ordinal querying for tuplewise similarity learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 3332–3340, Apr. 2020. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/5734

[32] B. Mason, A. Tripathy, and R. Nowak, "Learning nearest neighbor graphs from noisy distance samples," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[33] L. Jain, K. G. Jamieson, and R. Nowak, "Finite sample prediction and recovery bounds for ordinal embedding," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper/2016/file/4e0d67e54ad6626e957d15b08ae128a6-Paper.pdf

[34] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs: I. the method of paired comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345, 1952.

[35] P. Rao and L. L. Kupper, "Ties in paired-comparison experiments: A generalization of the bradley-terry model," *Journal of the American Statistical Association*, vol. 62, no. 317, pp. 194–204, 1967.

[36] R. D. Luce, *Individual choice behavior: A theoretical analysis*. Courier Corporation, 2012.

[37] R. L. Plackett, "The analysis of permutations," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 24, no. 2, pp. 193–202, 1975.

[38] L. L. Thurstone, "A law of comparative judgment." *Psychological review*, vol. 34, no. 4, p. 273, 1927.

[39] I. Melekhov, J. Kannala, and E. Rahtu, "Siamese network features for image matching," in *2016 23rd international conference on pattern recognition (ICPR)*. IEEE, 2016, pp. 378–383.

[40] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," *Journal of machine learning research*, vol. 4, no. Nov, pp. 933–969, 2003.

[41] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 89–96.

[42] Z. Zheng, K. Chen, G. Sun, and H. Zha, "A regression framework for learning ranking functions using relative relevance judgments," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, 2007, pp. 287–294.

[43] D. Chumbalov, L. Maystre, and M. Grossglauser, "Scalable and efficient comparison-based search without features," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 1995–2005. [Online]. Available: https://proceedings.mlr.press/v119/chumbalov20a.html

[44] N. Houlsby, F. Huszar, Z. Ghahramani, and J. Hernández-lobato, "Collaborative gaussian processes for preference learning," *Advances in neural information processing systems*, vol. 25, 2012.

[45] B. Dubois, "Ideal point versus attribute models of brand preference: A comparison of predictive validity," *ACR North American Advances*, 1975.

[46] S. Janson, "Tail bounds for sums of geometric and exponential variables," *Statistics & Probability Letters*, vol. 135, pp. 1–6, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167715217303711

[47] M. A. Davenport and J. Romberg, "An overview of low-rank matrix recovery from incomplete observations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 608–622, 2016.

[48] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[49] J. A. Tropp, "An introduction to matrix concentration inequalities," *arXiv preprint arXiv:1501.01571*, 2015.

[50] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *The Annals of Statistics*, vol. 28, no. 5, pp. 1302–1338, 2000. [Online]. Available: http://www.jstor.org/stable/2674095

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] , in Section 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] , in Appendix A
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] , please see linked code repository as well as Sections 5 and F.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] , please see Section F
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] in Section F

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
    (a) If your work uses existing assets, did you cite the creators? [Yes]
    (b) Did you mention the license of the assets? [N/A]
    (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] , please see linked code repository.
    (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
    (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
    (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
    (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
    (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Broader impacts

With the deployment of the ideal point model with a learned metric comes all of the challenges, impacts, and considerations associated with preference learning and recommender systems, such as if the deployed recommender system produces preference estimates and item recommendations that are aligned with the values and goals of the users and society as a whole, and if the item features are selected in a way that is diverse enough to adequately model all users and items. Therefore, we limit our broader impacts discussion to the challenges specific to our model.

As with any statistical model, the utility of our common metric ideal point model is limited by the accuracy to which it captures the patterns observed in the response data. The most immediate question about our model is the appropriateness of the assumption that a single metric is shared among all users. Such a model can prove useful in practice since it directly allows for shared structure between users and shared information between their measurements, and furthermore allows for a direct interpretation of preference at the crowd level, as we demonstrate with color preference data in Section 5. However, with this model comes the implicit assumption that users are homogeneous in their preference judgements. In many recommender systems the assumption of a common preference metric will likely be violated, due to heterogeneous user bases and subpopulations of users.

In general, there are almost certainly individual differences between each user's notion of item similarity and preference, and hence recovery of a single crowd metric should not necessarily be taken to mean that it describes each individual exactly. Although our model does provide a degree of individual flexibility through its use of ideal points (rather than treating the entire crowd's responses as coming from a single user), the result of a common metric violation may be that the learned population metric will fit to the behavior of the majority, or may fail to capture some aspect of each individual user's preference judgements. Our model could certainly be applied individually to each user, such that they each learn their own metric: however, as we demonstrate in our theoretical and empirical results, there is a fundamental sample complexity tradeoff in that to learn individual metrics, many more samples are needed per user, unlike the case of a common metric model where measurements are amortized.

The impacts of such a mismatch on an individual or subpopulation can range from inconvenient, such as in getting poor recommendations for online shopping or streaming services, to actively harmful, such as in receiving poor recommendations for a major decision (e.g., medical) that would otherwise suit the population majority. To prevent such cases, before deploying a common metric model it is important to not only average performance across the entire population (which will reflect the majority), but also evaluate worst-case performance on any given user or subpopulation. Such considerations are especially important if the common metric is not only used for predicting preferences between items, but also used to make inferences about a population by directly examining the metric entries (as we demonstrate for color preferences in Section 5). If the metric only applies to the majority or the population in the aggregate, then such inferences about feature preferences may not be accurate for individual users or subpopulations.

Beyond considering the effects of skewed modeling of individual users or subpopulations, it is important to consider potentially harmful effects of a common metric preference model when arriving at *item rankings*. While the item set examples discussed here include non-human objects such as movies and products, more generally the term "item" may be used in an abstract sense to include people, such as when building recommender systems for an admissions or hiring committee to select between job or school applicants (see [7] for a preference learning example on graduate school admissions data). In this context, the "users" may be a separate population (such as an admissions committee) that is making preference judgements about individual candidates (i.e., the "items"). In such cases, it is critical that extra precautions be taken and considerations be made for any possible biases that may be present across the population of users and reflected in the common metric. For example, if a majority of an admissions committee shared certain implicit biases when making preference decisions between candidates, such biases may be *learned* in a common metric. On the other hand, the existence of a common metric potentially allows for interpretation and insight into the features by which a committee is making its decisions, possibly allowing for intervention and bias mitigation if it is observed that the committee population is sharing a bias with regard to certain candidate features. While our model could be applied to each individual to avoid the challenges described above, there is a fundamental tradeoff in the sample complexity cost per user required to obtain individual models, which our results elucidate.

B Related work

Metric learning has received considerable attention and classical techniques are nicely summarized in the monographs [8, 27]. Efficient algorithms exist for a variety of data sources such as class labels [10, 11] and triplet comparisons [12]. Classical metric learning techniques focus on learning linear (Mahalanobis) metrics parametrized by a positive (semi-)definite matrix. In the case of learning linear metrics from triplet observations, [12, 28] establish tight generalization error guarantees. In practice, to handle increasingly complex learning problems, it is common to leverage more expressive, nonlinear metrics that are parametrized by kernels or deep neural networks and we refer the reader to [13–15] for a survey of nonlinear metric learning techniques. The core idea of many kernelized metric learning algorithms is that one can use Kernelized-PCA to reduce the nonlinear metric learning problem over n items to learning a linear metric in $\mathbb{R}^n$ via a kernel trick on the empirical Gram matrix. The downside to this approach is that the learned metric need not apply to new items other than those contained in the original n.

To circumvent this issue, works such as [16] have proposed deep metric learning. Intuitively, in the linear case one may factor a metric $\boldsymbol{M}$ as $\boldsymbol{M} = \boldsymbol{L}\boldsymbol{L}^T$ and could instead learn a matrix $\boldsymbol{L} \in \mathbb{R}^{d\times r}$. In the case of deep metric learning, the same principle applies except that $\boldsymbol{L}$ is replaced with a map $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}^r$ given by a deep neural network such that the final metric is $d(x, y) = \|\mathcal{L}(x) - \mathcal{L}(y)\|_2$. While the theory of nonlinear metric learning is less mature, [29,30] provide generalization guarantees for deep metric learning using neural tangent kernel and Rademacher analyses respectively. Finally, metric learning is a closely related to the problem of ordinal embedding: [17, 31, 32] propose active sampling techniques for ordinal embedding whereas [33] establishes learning guarantees for passive algorithms.

Preference learning from paired comparisons is a well-studied problem spanning machine learning, psychology, and social sciences, and we refer the reader to [9] for a comprehensive summary of approaches and problem statements. Researchers have proposed a multitude of models ranging from classical techniques such as the Bradley-Terry model [34, 35], Plackett-Luce model [36, 37], and Thurstone model [38] to more modern approaches such as preference learning via Siamese networks [39] to fit the myriad of tailored applications of preference learning. In the linear setting, [7, 40–42] among others propose passive learning algorithms whereas [5, 6, 17–19, 43] propose adaptive sampling procedures. [20] perform localization from paired comparisons, and [44] employ a Gaussian process approach for learning pairwise preferences from multiple users.

The ideal point model we consider here is well-established in the single user setting [4–7], and differs fundamentally from models of preference based only on inner products. In the inner product or "attribute" model of preference, items with larger values of $\langle \boldsymbol{u}, \boldsymbol{x}\rangle$ are preferred, in which case an item's preference can be modified simply by scaling its magnitude. On the other hand, the ideal point (i.e., distance-based) model accounts for item magnitude, and it has been demonstrated that distance-based models can more accurately model user preferences [45].

C Proofs and additional results for identifiability from unquantized measurements

C.1 Properties of selection matrices

In this section, we present several theoretical properties of selection matrices (see Definition 2.1) that will be useful for proving the results that follow. We begin with a lemma upper bounding the rank of selection matrices:

Lemma C.1. *Let $n \geq 2$. For any $m \times n$ selection matrix $\boldsymbol{S}$,* $\operatorname{rank}(\boldsymbol{S}) \leq \min(m, n-1)$.

Proof. Since $\boldsymbol{S}$ has m rows, $\operatorname{rank}(\boldsymbol{S}) \leq m$. By construction, for any selection matrix $\boldsymbol{S}$ note that $\boldsymbol{1}_n \in \ker(\boldsymbol{S})$, where $\boldsymbol{1}_n$ is the vector of all ones in $\mathbb{R}^n$. To see this, for any $\boldsymbol{z} \in \mathbb{R}^n$ the ith element of $\boldsymbol{S}\boldsymbol{z}$ is given by $\boldsymbol{z}[p_i] - \boldsymbol{z}[q_i]$, and so the ith element of $\boldsymbol{S}\boldsymbol{1}_n$ is $1 - 1 = 0$ and hence $\boldsymbol{S}\boldsymbol{z} = \boldsymbol{0}$. Therefore, $\dim(\ker(\boldsymbol{S})) \geq 1$ and so $\operatorname{rank}(\boldsymbol{S}) \leq n - 1$. □

We use this result to show a property of full row rank selection matrices that will be useful for their construction and analysis:

Lemma C.2. *Let $\boldsymbol{S}$ be an $m \times n$ selection matrix with $n \geq m+1$, where for each $i \in [m]$ the nonzero indices of the ith row are given by distinct $p_i, q_i \in [n]$ such that $\boldsymbol{S}[i, p_i] = 1$, $\boldsymbol{S}[i, q_i] = -1$. If $\operatorname{rank}(\boldsymbol{S}) = m$, then for every subset $I \subseteq [m]$ of row indices, there exists $i^* \in I$ such that $\boldsymbol{S}[j, p_{i^*}] = 0$ for all $j \in I \setminus \{i^*\}$, or $\boldsymbol{S}[j, q_{i^*}] = 0$ for all $j \in I \setminus \{i^*\}$.*

Proof. Let $I \subseteq [m]$ be given, and suppose by contradiction that no such i^* exists, i.e., no measurement in I introduces a new item unseen by any other measurements in I. Let $\boldsymbol{S}^{(I)}$ be the $|I| \times n$ selection matrix consisting of the rows in $\boldsymbol{S}$ listed in I. Since $\boldsymbol{S}$ is full row rank, its rows are linearly independent, implying that the rows in $\boldsymbol{S}^{(I)}$ are also linearly independent and therefore $\boldsymbol{S}^{(I)}$ has rank $|I|$.

Let $c \leq n$ be the number of columns that $\boldsymbol{S}^{(I)}$ is supported on (i.e., have at least one nonzero entry). By our contradictory assumption, every item measured in $\boldsymbol{S}^{(I)}$ is measured in at least two rows in I, and therefore each of these c columns must have at least 2 nonzero entries. This implies that $\boldsymbol{S}^{(I)}$ has at least $2c$ nonzero entries in total. Since each measurement adds exactly 2 nonzero entries to $\boldsymbol{S}^{(I)}$, this means that there are least $2c/2 = c$ measurements and so $|I| \geq c$.

Now consider the $|I| \times c$ matrix $\widetilde{\boldsymbol{S}}^{(I)}$ consisting of $\boldsymbol{S}^{(I)}$ with its zero columns removed. $\operatorname{rank}(\widetilde{\boldsymbol{S}}^{(I)}) = \operatorname{rank}(\boldsymbol{S}^{(I)})$ since $\widetilde{\boldsymbol{S}}^{(I)}$ and $\boldsymbol{S}^{(I)}$ have the same column space. Since $\widetilde{\boldsymbol{S}}^{(I)}$ is itself a $|I| \times c$ selection matrix, we know from Lemma C.1 that $\operatorname{rank}(\widetilde{\boldsymbol{S}}^{(I)}) \leq \min(|I|, c-1)$. Since we know $|I| \geq c$, $\min(|I|, c-1) = c-1$, implying $\operatorname{rank}(\widetilde{\boldsymbol{S}}^{(I)}) \leq c-1$. But this is a contradiction since we already know $\operatorname{rank}(\widetilde{\boldsymbol{S}}^{(I)}) = \operatorname{rank}(\boldsymbol{S}^{(I)}) = |I| \geq c$. □

Intuitively, Lemma C.2 says that if $\boldsymbol{S}$ is full row rank, then every subset of rows contains a row that is supported on a column that is zero for all other rows in the subset, i.e., at least one row measures a new item unmeasured by any other row in the subset. This property is related to a selection matrix being incremental (see Definition 2.2) as follows:

Lemma C.3. *Let $\boldsymbol{S}$ be an $m \times n$ selection matrix with $n \geq m+1$, where for each $i \in [m]$ the nonzero indices of the ith row are given by distinct $p_i, q_i \in [n]$ such that $\boldsymbol{S}[i, p_i] = 1$, $\boldsymbol{S}[i, q_i] = -1$. Suppose for every subset $I \subseteq [m]$ of row indices, there exists $i^* \in I$ such that $\boldsymbol{S}[j, p_{i^*}] = 0$ for all $j \in I \setminus \{i^*\}$, or $\boldsymbol{S}[j, q_{i^*}] = 0$ for all $j \in I \setminus \{i^*\}$. Then there exists an $m \times m$ permutation matrix $\boldsymbol{P}$ such that $\boldsymbol{PS}$ is incremental.*

Proof. We will construct a sequence of row indices such that permuting the rows of $\boldsymbol{S}$ in the sequence order results in an incremental matrix. Let $I_m := [m]$. By assumption, there exists an index $i_m \in I_m$ such that $\boldsymbol{S}[j, p_{i_m}] = 0$ for all $j \in [m] \setminus \{i_m\}$ or $\boldsymbol{S}[j, q_{i_m}] = 0$ for all $j \in [m] \setminus \{i_m\}$. Now let $1 < m' \leq m$ be given, and suppose by induction that there exists a set of distinct indices $\{i_k\}_{k=m'}^m$ such that for all $m' \leq k \leq m$, $\boldsymbol{S}[j, p_{i_k}] = 0$ for all $j \in [m] \setminus \{i_\ell\}_{\ell=k}^m$ or $\boldsymbol{S}[j, q_{i_k}] = 0$ for all $j \in [m] \setminus \{i_\ell\}_{\ell=k}^m$ (we have shown the case of $m' = m$ above). Let $I_{m'-1} := [m] \setminus \{i_k\}_{k=m'}^m$. Then by assumption, there exists an index $i_{m'-1} \in I_{m'-1}$ such that $\boldsymbol{S}[j, p_{i_{m'-1}}] = 0$ for all $j \in [m] \setminus \{i_k\}_{k=m'-1}^m$ or $\boldsymbol{S}[j, q_{i_{m'-1}}] = 0$ for all $j \in [m] \setminus \{i_k\}_{k=m'-1}^m$. Therefore, combined with the fact that $i_{m'-1} \in [m] \setminus \{i_k\}_{k=m'}^m$ along with the inductive assumption on $\{i_k\}_{k=m'}^m$, $\{i_k\}_{k=m'-1}^m$ constitutes an index set where for all $m'-1 \leq k \leq m$, $\boldsymbol{S}[j, p_{i_k}] = 0$ for all $j \in [m] \setminus \{i_\ell\}_{\ell=k}^m$ or $\boldsymbol{S}[j, q_{i_k}] = 0$ for all $j \in [m] \setminus \{i_\ell\}_{\ell=k}^m$.

Taking $m' = 2$, we have proved by induction the existence of an index set $\{i_1, \ldots, i_m\}$ that is a permutation of $[m]$ such that for any $k \in [m]$, $\boldsymbol{S}[j, p_{i_k}] = 0$ for all $j \in [m] \setminus \{i_\ell\}_{\ell=k}^m$ or $\boldsymbol{S}[j, q_{i_k}] = 0$ for all $j \in [m] \setminus \{i_\ell\}_{\ell=k}^m$. By construction, $[m] \setminus \{i_\ell\}_{\ell=k}^m = \{i_\ell\}_{\ell=1}^{k-1}$, so equivalently for any $k \in [m]$, $\boldsymbol{S}[i_j, p_{i_k}] = 0$ for all $j < k$ or $\boldsymbol{S}[i_j, q_{i_k}] = 0$ for all $j < k$.

We can then explicitly construct the $m \times m$ permutation matrix $\boldsymbol{P}$ as

$$\boldsymbol{P}[k, \ell] = \begin{cases} 1 & \ell = i_k \\ 0 & \text{otherwise.} \end{cases}$$

Let $\boldsymbol{S}' = \boldsymbol{PS}$, $p'_k = p_{i_k}$ and $q'_k = q_{i_k}$. p'_k, q'_k are the nonzero column indices of the kth row in the permuted selection matrix $\boldsymbol{S}'$, since for any $\ell \in [n]$, $\boldsymbol{S}'[k, \ell] = \boldsymbol{S}[i_k, \ell]$. We then have for any $k \in [m]$, $\boldsymbol{S}'[j, p'_k] = \boldsymbol{S}[i_j, p_{i_k}] = 0$ for all $j < k$ or $\boldsymbol{S}'[j, q'_k] = \boldsymbol{S}[i_j, q_{i_k}] = 0$ for all $j < k$, and hence $\boldsymbol{S}'$ is incremental. □

Furthermore, if a selection matrix $\boldsymbol{S}$ (more specifically, a permutation thereof) is incremental, then it is also full-rank:

Lemma C.4. *Let $\boldsymbol{S}$ be an $m \times n$ selection matrix with $n \geq m+1$, and suppose there exists an $m \times m$ permutation matrix $\boldsymbol{P}$ such that $\boldsymbol{PS}$ is incremental. Then $\boldsymbol{S}$ is full-rank with* $\operatorname{rank}(\boldsymbol{S}) = m$.

Proof. Denoting the ith row of $\boldsymbol{PS}$ by $\boldsymbol{s}_i$, since $\boldsymbol{PS}$ is incremental for all $i \in [m]$ there exists a j such that $\boldsymbol{s}_i[j] \neq 0$ and $\boldsymbol{s}_\ell[j] = 0$ for all $\ell < i$. Hence, for all $i \in [m]$, $\boldsymbol{s}_i$ does not lie in the span of $\{\boldsymbol{s}_\ell\}_{\ell<i}$. Starting at $i = 2$, this implies that $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ are linearly independent. Let $m' < m$ be given, and assume by induction that $\{\boldsymbol{s}_\ell\}_{\ell \leq m'}$ are linearly independent. Since by assumption $\boldsymbol{s}_{m'+1}$ does not lie in the span of $\{\boldsymbol{s}_\ell\}_{\ell \leq m'}$, the entire set $\{\boldsymbol{s}_\ell\}_{\ell \leq m'+1}$ is linearly independent. Taking $m' = m - 1$, we have by induction that the rows of $\boldsymbol{PS}$ (i.e., $\{\boldsymbol{s}_\ell\}_{\ell \leq m}$) are linearly independent, and since these rows are just a permutation of the rows in $\boldsymbol{S}$, the m rows in $\boldsymbol{S}$ are also linearly independent and so $\operatorname{rank}(\boldsymbol{S}) = m$. □

We summarize the above lemmas in the following corollary:

Corollary C.4.1. *Let $\boldsymbol{S}$ be an $m \times n$ selection matrix with $n \geq m+1$, where for each $i \in [m]$ the nonzero indices of the ith row are given by distinct $p_i, q_i \in [n]$ such that $\boldsymbol{S}[i, p_i] = 1$, $\boldsymbol{S}[i, q_i] = -1$. Then the following are equivalent:*

(a) $\operatorname{rank}(\boldsymbol{S}) = m$.

(b) For every subset $I \subseteq [m]$ of row indices, there exists $i^* \in I$ such that $\boldsymbol{S}[j, p_{i^*}] = 0$ for all $j \in I \setminus \{i^*\}$, or $\boldsymbol{S}[j, q_{i^*}] = 0$ for all $j \in I \setminus \{i^*\}$.

(c) There exists an $m \times m$ permutation matrix $\boldsymbol{P}$ such that $\boldsymbol{PS}$ is incremental.

Proof. By Lemma C.2, $(a) \implies (b)$. By Lemma C.3, $(b) \implies (c)$. By Lemma C.4, $(c) \implies (a)$. Combining these implications, $(a) \iff (b) \iff (c)$. □

Another useful corollary lower bounds the number of columns a selection matrix must be supported on, depending on its rank:

Corollary C.4.2. *Let $\boldsymbol{S}$ be a rank r, $m \times n$ selection matrix with $m \geq r$ and $n \geq r+1$. Then at least $r+1$ columns of $\boldsymbol{S}$ have at least one nonzero entry.*

Proof. Since $\operatorname{rank}(\boldsymbol{S}) = r$, there exists an index set of r linearly independent rows of $\boldsymbol{S}$, which we denote by $I \subseteq [m]$. Let $\boldsymbol{S}'$ be the $r \times n$ submatrix of $\boldsymbol{S}$ consisting of the rows indexed by I: since its rows are linearly independent, $\operatorname{rank}(\boldsymbol{S}') = r$. From Corollary C.4.1, there exists a permutation $\boldsymbol{P}$ of the rows in $\boldsymbol{S}'$ such that $\boldsymbol{PS}'$ is incremental. Since the first row of $\boldsymbol{PS}'$ introduces two items and the remaining $r-1$ rows each introduce at least one new item, $\boldsymbol{PS}'$ must be supported on at least $2 + (r-1) = r+1$ columns. Since the rows in $\boldsymbol{PS}'$ are contained in $\boldsymbol{S}$, $\boldsymbol{S}$ must also be supported on at least $r+1$ columns. □

When studying random selection matrices in Appendix C.2, it will be useful to understand a particular graph constructed from the rows of a selection matrix $\boldsymbol{S}$. For $p, q \in [n]$ and $p \neq q$, let $\boldsymbol{s}_{(p,q)}$ denote a vector in $\mathbb{R}^n$ given by

$$\boldsymbol{s}_{(p,q)}[j] = \begin{cases} 1 & j = p \\ -1 & j = q \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

Consider a set of r vectors $S \coloneqq \{\boldsymbol{s}_i\}_{i=1}^r \subset \mathbb{R}^n$ in the form given by eq. (8). We can construct a graph $G_S = (V_S, E_S)$ from this set as follows: $V_S = [r]$ denotes the vertices of this graph (with vertex $i \in [r]$ corresponding to row $\boldsymbol{s}_i$), and E_S denotes the edge set. We define the connectivity of G_S by an $r \times r$ adjacency matrix $\boldsymbol{A}_S$, where

$$\boldsymbol{A}_S[i,j] = \begin{cases} 1 & \exists k \in [n] \text{ s.t. } \boldsymbol{s}_i[k] \neq 0 \wedge \boldsymbol{s}_j[k] \neq 0 \\ 0 & \text{otherwise.} \end{cases}$$

In other words, vectors $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ are adjacent on G_S if they have overlapping support. We say that vertices i and j are *linked* on G if $A[i,j] = 1$, or if there exists a finite sequence of distinct indices

$\{k_\ell\}_{\ell=1}^T \subseteq [r] \setminus \{i, j\}$ such that $A[i, k_1] = A[k_1, k_2] = \cdots = A[k_{T-1}, k_T] = A[k_T, j] = 1$. Denote the set of linked vertex pairs by

$$C_S = \{(i, j) : i, j \in [r],\ i, j \text{ linked on } G_S\}.$$

We start with a lemma concerning the span of S in how it relates to connectivity on G_S:

Lemma C.5. *Let $S \coloneqq \{\boldsymbol{s}_i\}_{i=1}^r$ denote a set of linearly independent vectors in $\mathbb{R}^n$ in the form eq.* (8)*, with $n \geq r + 1$. For given $p, q \in [n]$ with $p \neq q$, if $\boldsymbol{s}_{(p,q)} \in \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^r)$ then there exists a linked vertex pair $(i_p, i_q) \in C_S$ such that $\boldsymbol{s}_{i_p}[p] \neq 0$ and $\boldsymbol{s}_{i_q}[q] \neq 0$.*

Proof. If $\boldsymbol{s}_{(p,q)} \in \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^r)$, there exist scalars $\{\beta_i\}_{i=1}^r$ not all equal to zero such that $\boldsymbol{s}_{(p,q)} = \sum_{i=1}^r \beta_i \boldsymbol{s}_i$, i.e.,

$$1 = \sum_{i=1}^r \beta_i \boldsymbol{s}_i[p] \tag{9}$$

$$-1 = \sum_{i=1}^r \beta_i \boldsymbol{s}_i[q] \tag{10}$$

$$0 = \sum_{i=1}^r \beta_i \boldsymbol{s}_i[j] \quad j \neq p, q. \tag{11}$$

From eq. (9), we know there exists a $i_p \in [r]$ such that $\beta_{i_p} \neq 0$ and $\boldsymbol{s}_{i_p}[p] \neq 0$: otherwise, $\beta_i \boldsymbol{s}_i[p] = 0$ for all $i \in [r]$ which would result in the summation in eq. (9) being 0. Let j_1 denote the other index supported by $\boldsymbol{s}_{i_p}$, i.e., $j_1 \neq p$ and $\boldsymbol{s}_{i_p}[j_1] \neq 0$. If $j_1 = q$, then there trivially exists $i_q = i_p$ such that $\boldsymbol{s}_{i_q}[q] = \boldsymbol{s}_{i_p}[j_1] \neq 0$. Clearly this choice of (i_p, i_q) is linked on G_S since $i_p = i_q$, which would give us the desired result.

Now suppose $j_1 \neq q$. Recalling that $j_1 \neq p$ as well, from eq. (11) we have

$$\begin{aligned} 0 &= \sum_{i=1}^r \beta_i \boldsymbol{s}_i[j_1] \\ &= \underbrace{\beta_{i_p}}_{\neq 0} \underbrace{\boldsymbol{s}_{i_p}[j_1]}_{\neq 0} + \sum_{i \in [r] \setminus \{i_p\}} \beta_i \boldsymbol{s}_i[j_1], \end{aligned} \tag{12}$$

which implies that there exists $i_1 \in [r] \setminus \{i_p\}$ such that $\beta_{i_1} \neq 0$ and $\boldsymbol{s}_{i_1}[j_1] \neq 0$; otherwise, $\beta_i \boldsymbol{s}_i[j_1] = 0$ for all $i \in [r] \setminus \{i_p\}$ which would result in a contradiction in eq. (12). Note that (i_1, i_p) are linked on G_S, since they are both supported on index j_1.

Now, suppose by induction that for a given $1 \leq T \leq r - 1$ there exist distinct vertices $\{i_1 \ldots, i_T\} \in [r] \setminus \{i_p\}$ and distinct item indices $\{j_1, \ldots, j_T\} \in [n] \setminus \{p, q\}$ such that $\boldsymbol{s}_{i_p}[j_1] \neq 0$, $\boldsymbol{s}_{i_k}[j_k] \neq 0$ and $\boldsymbol{s}_{i_k}[j_{k+1}] \neq 0$ for $k < T - 1$, $\boldsymbol{s}_{i_T}[j_T] \neq 0$, (i_p, i_T) are linked on G_S, and $\beta_{i_k} \neq 0$ for $k \in [T]$. Above we have shown the existence of such sets for the the base case of $T = 1$.

Let j_{T+1} be the other item index supported on $\boldsymbol{s}_{i_T}$, i.e., $j_{T+1} \neq j_T$ and $\boldsymbol{s}_{i_T}[j_{T+1}] \neq 0$. If $j_{T+1} = q$, then we can set $i_q = i_T$ and we have found an i_q linked to i_p on G_S (since $i_T = i_q$ is linked to i_p on G_S by inductive assumption) and $\boldsymbol{s}_{i_q}[q] = \boldsymbol{s}_{i_T}[j_{T+1}] \neq 0$. Otherwise, since $\{\boldsymbol{s}_{i_k}\}_{k=1}^T \cup \{\boldsymbol{s}_{i_p}\}$ are $T + 1$ linearly independent vectors in the form eq. (8), from Corollary C.4.2 we have that $\{\boldsymbol{s}_{i_k}\}_{k=1}^T \cup \{\boldsymbol{s}_{i_p}\}$ are collectively supported on at least $T + 2$ indices in $[n]$. Hence, if $j_{T+1} \neq q$, then we must have $j_{T+1} \in [n] \setminus (\{p, q\} \cup \{j_k\}_{k=1}^T)$. From eq. (11), we then have

$$\begin{aligned} 0 &= \sum_{i=1}^r \beta_i \boldsymbol{s}_i[j_{T+1}] \\ &= \beta_{i_p} \cancelto{0}{\boldsymbol{s}_{i_p}[j_{T+1}]} + \sum_{k=1}^{T-1} \beta_{i_k} \cancelto{0}{\boldsymbol{s}_{i_k}[j_{T+1}]} + \underbrace{\beta_{i_T}}_{\neq 0} \underbrace{\boldsymbol{s}_{i_T}[j_{T+1}]}_{\neq 0} + \sum_{i \in [r] \setminus (i_p \cup \{i_k\}_{k=1}^T)} \beta_i \boldsymbol{s}_i[j_{T+1}], \\ &= \underbrace{\beta_{i_T}}_{\neq 0} \underbrace{\boldsymbol{s}_{i_T}[j_{T+1}]}_{\neq 0} + \sum_{i \in [r] \setminus (i_p \cup \{i_k\}_{k=1}^T)} \beta_i \boldsymbol{s}_i[j_{T+1}], \end{aligned} \tag{13}$$

which implies that there exists $i_{T+1} \in [r] \setminus (i_p \cup \{i_k\}_{k=1}^T)$ such that $\beta_{i_{T+1}} \neq 0$ and $\boldsymbol{s}_{i_{T+1}}[j_{T+1}] \neq 0$; otherwise, $\beta_i \boldsymbol{s}_{i_{T+1}}[j_{T+1}] = 0$ for all $[r] \setminus (i_p \cup \{i_k\}_{k=1}^T)$, which would result in a contradiction in eq. (13). Note that if $T = r-1$, the existence of such an i_{T+1} is impossible since in that case $[r] \setminus (i_p \cup \{i_k\}_{k=1}^T) = \varnothing$; hence, if $T = r-1$, it must be the case that $j_{T+1} = q$ as described above.

If $T < r-1$ and $j_{T+1} \neq q$, then such an $i_{T+1} \in [r] \setminus (i_p \cup \{i_k\}_{k=1}^T)$ exists. Note that i_{T+1} and i_p are linked on G_S, since i_{T+1} and i_T share an item (i.e., j_{T+1}) and i_T and i_p are linked on G_S by inductive assumption. Hence, we have constructed sets $\{i_k\}_{i=1}^{T+1}$ and $\{j_k\}_{i=1}^{T+1}$ that fulfill the inductive assumption for $T' = T+1$.

Therefore, there must exist a $1 \leq T^* \leq r-1$ such that $j_{T^*+1} = q$, in which case we can take $i_q = i_{T^*}$ and thus have identified an i_q that is linked to i_p on G_S (since $i_{T*} = i_q$ is linked to i_p on G_S) and satisfies $\boldsymbol{s}_{i_q}[q] = \boldsymbol{s}_{i_{T^*}}[j_{T^*+1}] \neq 0$. □

C.2 Characterizing random selection matrices

In this section, we explore how many measurements and items are required for a randomly constructed selection matrix to have full-rank.

To answer this question, first we establish a fundamental result concerning how many item pairs sampled uniformly at random are required (on average or with high probability) in order for a selection matrix to be of a certain rank. We start by bounding the probability that, for an existing selection matrix $\boldsymbol{S}$ with rank r, an additional row $\boldsymbol{s}$ constructed by selecting two items uniformly at random lies within the row space of $\boldsymbol{S}$. This is equivalent to the probability that the concatenation of $\boldsymbol{s}$ with $\boldsymbol{S}$ is a rank $r+1$ matrix; bounding this probability will then allow us to bound the number of such appended rows needed to increase the rank of $\boldsymbol{S}$ to some desired value greater than r.

Lemma C.6. *Suppose $\boldsymbol{S}$ is an $m \times n$ selection matrix with rank $r \leq \min(m, n-1)$. Let $\boldsymbol{s} \in \mathbb{R}^n$ be constructed by sampling two integers, p and q, uniformly and without replacement from $[n]$ (and statistically independent of $\boldsymbol{S}$) and setting $\boldsymbol{s} = \boldsymbol{s}_{(p,q)}$. Then*

$$\frac{2r}{n(n-1)} \leq \mathbb{P}(\boldsymbol{s} \in \operatorname{rowsp}(\boldsymbol{S}) \mid \boldsymbol{S}) \leq \frac{(r+1)r}{n(n-1)}. \tag{14}$$

We defer the proof of Lemma C.6 to the end of the section.

With the above result, we can work towards characterizing the probability that a selection matrix with pairs sampled uniformly at random has a particular rank. To make our results as general as possible, assume that we have a known "seed" $m \times n$ selection matrix $\boldsymbol{S}_0$ with rank $r_0 \leq \min(m, n-2)$, and that we append m randomly sampled rows to $\boldsymbol{S}_0$ where each row is constructed by sampling two integers uniformly at random without replacement (and statistically independent from previous measurements and $\boldsymbol{S}_0$) from $[n]$; denote these m rows as $m \times n$ selection matrix $\boldsymbol{S}$. We are interested in characterizing the probability that $\left[\begin{smallmatrix} \boldsymbol{S}_0 \\ \boldsymbol{S} \end{smallmatrix}\right]$ has rank $r > r_0$. We are only interested in $r_0 \leq \min(m, n-2)$, since if $r_0 = n-1$ then from Lemma C.1 $\boldsymbol{S}_0$ already has the maximum rank possible for a selection matrix with n columns, and so we cannot increase its rank with additional random measurements.

Let $\boldsymbol{s}_i$ denote the ith row of $\boldsymbol{S}$. We will take an approach similar in spirit to the coupon collector problem by first defining a notion of a "failure" and "success" with regards to measuring new rows. After having queried $i-1$ random paired comparisons given by rows $\{\boldsymbol{s}_j\}_{j=1}^{i-1}$, we say that sampling a new selection row $\boldsymbol{s}_i$ "fails" if it lies in the span of the selection matrix thus far, and "succeeds" if it lies outside this span. More precisely[8], define failure event $E_i = 0$ if $\boldsymbol{s}_i \in \operatorname{rowsp}(\boldsymbol{S}_0) \cup \operatorname{span}(\{\boldsymbol{s}_j\}_{j=1}^{i-1})$ and success event $E_i = 1$ otherwise. Clearly, $\dim(\operatorname{rowsp}(\boldsymbol{S}_0) \cup \operatorname{span}(\{\boldsymbol{s}_j\}_{j=1}^{i})) = \dim(\operatorname{rowsp}(\boldsymbol{S}_0) \cup \operatorname{span}(\{\boldsymbol{s}_j\}_{j=1}^{i-1})) + 1$ if and only if $E_i = 1$. For $i \geq 1$, let $M_i = \min(\{k : \dim(\operatorname{rowsp}(\boldsymbol{S}_0) \cup \operatorname{span}(\{\boldsymbol{s}_j\}_{j=1}^{k})) = r_0 + i\}) = \min(\{k : \sum_{j=1}^{k} E_j = i\})$. Note that for any $i \geq 1$, $E_{M_i} = 1$; otherwise, $i = \sum_{j=1}^{M_i} E_j = \sum_{j=1}^{M_i - 1} E_j + 0 < i$ by definition of M_i, which would be a contradiction. M_{r-r_0} for $r > r_0$ is exactly the quantity we are interested in, since

[8] In the following statements concerning probability events, $\boldsymbol{S}_0$ is assumed to be fixed and known.

it is the number of random measurements (beyond those already in r_0) needed for $r - r_0$ successes in total, i.e., for the cumulative selection matrix $\left[\begin{smallmatrix} \boldsymbol{S}_0 \\ \boldsymbol{S} \end{smallmatrix}\right]$ to be rank r.

To analyze M_i for $1 \leq i \leq r - r_0$, let $C_i = M_i - M_{i-1}$ denote the number of measurements until the first success after already having had $i - 1$ successes, where $C_1 = M_1$. Then

$$M_i = (M_i - M_{i-1}) + (M_{i-1} - M_{i-2}) + (M_{i-2} + \cdots + (M_2 - M_1) + M_1 = \sum_{j=1}^{i} C_j.$$

Given $C_1, \ldots, C_{i-1}$ (and hence M_{i-1}), we note by definition that for any $c \geq 1$,

$$C_i > c \iff E_j = 0 \text{ for all } M_{i-1} + 1 \leq j \leq M_{i-1} + c. \tag{15}$$

Now suppose we condition on the event $C_1 = c_1, \ldots, C_{i-1} = c_{i-1}$, which we denote for shorthand by $c_1, \ldots, c_{i-1}$. We have

$$\begin{aligned}
\mathbb{P}(C_i > c \mid c_1, \ldots, c_{i-1}) &= \mathbb{P}\left(\bigcap_{k=M_{i-1}+1}^{M_{i-1}+c} E_k = 0 \,\middle|\, c_1, \ldots, c_{i-1} \right) &(16)\\
&= \prod_{k=M_{i-1}+1}^{M_{i-1}+c} \mathbb{P}\left(E_k = 0 \,\middle|\, \bigcap_{\ell=M_{i-1}+1}^{k-1} (E_\ell = 0), c_1, \ldots, c_{i-1} \right), &(17)
\end{aligned}$$

where eq. (16) follows from eq. (15). For a fixed $k \in (M_{i-1} + 1) \ldots (M_{i-1} + c)$ let

$$S_k = \left\{ \{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1} : \bigcap_{\ell=M_{i-1}+1}^{k-1} (E_\ell = 0), C_1 = c_1, \ldots, C_{i-1} = c_{i-1} \right\},$$

i.e., S_k is the set of all possible row sets $\{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1}$ that result in the events $\bigcap_{\ell=M_{i-1}+1}^{k-1} (E_\ell = 0), C_1 = c_1, \ldots, C_{i-1} = c_{i-1}$ (recall that by definition these events are deterministic when conditioned on $\{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1}$). A natural result of this set definition is

$$\{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1} \notin S_k \implies \mathbb{P}\left(\{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1} \,\middle|\, \bigcap_{\ell=M_{i-1}+1}^{k-1} (E_\ell = 0), c_1, \ldots, c_{i-1} \right) = 0, \tag{18}$$

We then have

$$\begin{aligned}
&\mathbb{P}\left(E_k = 0 \,\middle|\, \bigcap_{\ell=M_{i-1}+1}^{k-1} (E_\ell = 0), c_1, \ldots, c_{i-1} \right)\\
&= \sum_{\{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1} \in S_k} \mathbb{P}(E_k = 0 \mid \{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1}, \bigcap_{\ell=M_{i-1}+1}^{k-1} (E_\ell = 0), c_1, \ldots, c_{i-1}) \mathbb{P}(\{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1} \mid \bigcap_{\ell=M_{i-1}+1}^{k-1} (E_\ell = 0), c_1, \ldots, c_{i-1}) &(19)\\
&= \sum_{\{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1} \in S_k} \mathbb{P}(E_k = 0 \mid \{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1}) \mathbb{P}(\{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1} \mid \bigcap_{\ell=M_{i-1}+1}^{k-1} (E_\ell = 0), c_1, \ldots, c_{i-1}) &(20)\\
&= \sum_{\{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1} \in S_k} \mathbb{P}(\boldsymbol{s}_k \in \operatorname{rowsp}(\boldsymbol{S}_0) \cup \operatorname{span}(\{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1}) \mid \{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1}) \mathbb{P}(\{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1} \mid \bigcap_{\ell=M_{i-1}+1}^{k-1} (E_\ell = 0), c_1, \ldots, c_{i-1})\\
&\leq \frac{(r_0 + i)(r_0 + i - 1)}{n(n-1)} \sum_{\{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1} \in S_k} \mathbb{P}(\{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1} \mid \bigcap_{\ell=M_{i-1}+1}^{k-1} (E_\ell = 0), c_1, \ldots, c_{i-1}) &(21)\\
&= \frac{(r_0 + i)(r_0 + i - 1)}{n(n-1)} (1) &(22)
\end{aligned}$$

and so

$$\mathbb{P}\left(E_k = 0 \,\middle|\, \bigcap_{\ell=M_{i-1}+1}^{k-1} (E_\ell = 0), C_1 = c_1, \ldots, C_{i-1} = c_{i-1} \right) \leq \frac{(r_0 + i)(r_0 + i - 1)}{n(n-1)}.$$

In the above, eq. (19) is a result of eq. (18), eq. (20) is since

$$\{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1} \in S_k \implies \bigcap_{\ell=M_{i-1}+1}^{k-1} (E_\ell = 0), c_1, \ldots, c_{i-1},$$

eq. (21) is from Lemma C.6 combined with the fact that since $\{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1} \in S_k$, $\dim(\mathrm{rowsp}(\boldsymbol{S}_0) \cup \mathrm{span}(\{\boldsymbol{s}_\ell\}_{\ell=1}^{k-1})) = r_0 + i - 1$, and eq. (22) is from eq. (18).

Continuing from eq. (17), this implies

$$\mathbb{P}(C_i > c \mid C_1 = c_1, \ldots, C_{i-1} = c_{i-1}) \leq \Big(\frac{(r_0 + i)(r_0 + i - 1)}{n(n-1)}\Big)^c,$$

and so $\mathbb{P}(C_i \leq c \mid C_1 = c_1, \ldots, C_{i-1} = c_{i-1}) \geq 1 - \left(\frac{(r_0+i)(r_0+i-1)}{n(n-1)}\right)^c$.

Now, consider a set of $r - r_0$ *independent* random variables, $B_1, \ldots, B_{r-r_0}$, with each $B_i \in \{1, 2, 3, \ldots\}$ distributed according to a geometric distribution with probability of success given by $p_i = 1 - \frac{(r_0+i)(r_0+i-1)}{n(n-1)}$. We will relate the statistics of B_i to C_i in order to construct a tail bound on M_{r-r_0}, our quantity of interest. Recalling the c.d.f. of geometric distributions, we have for $1 \leq i \leq r - r_0$,

$$\mathbb{P}(B_i \leq c \mid B_1, \ldots, B_{i-1}) = \mathbb{P}(B_i \leq c) = 1 - (1 - p_i)^c = 1 - \left(\frac{(r_0 + i)(r_0 + i - 1)}{n(n-1)}\right)^c,$$

and so $\mathbb{P}(C_i \leq c \mid C_1, \ldots, C_{i-1}) \geq \mathbb{P}(B_i \leq c)$ for all possible $C_1, \ldots, C_{i-1}$.

Let $B \coloneqq \sum_{i=1}^{r-r_0} B_i$. [46] presents a tail bound for the sum of independent geometric random variables, which we can apply to B. Let $X = \sum_{i=1}^{j} X_i$ be the sum of j independent geometric random variables, each with parameter $0 < p_i \leq 1$. Define $\mu \coloneqq \mathbb{E}[X] = \sum_{i=1}^{j} \frac{1}{p_i}$. Then from [46], for any $\lambda \geq 1$,

$$\mathbb{P}(X \geq \lambda\mu) \leq e^{1-\lambda}.$$

In our case, $X_i = B_i$, $j = r - r_0$, and

$$\mu = \mathbb{E}[B] = \sum_{i=1}^{r-r_0} \frac{1}{1 - \frac{(r_0+i)(r_0+i-1)}{n(n-1)}},$$

and so for any $\lambda \geq 1$,

$$\mathbb{P}(B \geq \lambda\mu) \leq e^{1-\lambda}. \tag{23}$$

To translate eq. (23) into a more interpretable tail bound, Let $0 < \delta < 1$ be given. If we choose $\lambda = 1 + \ln \frac{1}{\delta}$ (noting that $\lambda > 1$), then

$$\mathbb{P}\Big(B \geq \Big(1 + \ln\frac{1}{\delta}\Big)\Big(\sum_{i=1}^{r-r_0} \frac{1}{1 - \frac{(r_0+i)(r_0+i-1)}{n(n-1)}}\Big)\Big) \leq \delta. \tag{24}$$

If we can relate the statistics of B to those of M_{r-r_0}, then we can potentially apply eq. (24) to construct a tail bound on M_{r-r_0}; the following lemma will provide the link we need. In the following, for a sequence $\{X_k\}_{k=i}^{j}$ let $X_{i:j} \coloneqq \{X_k\}_{k=i}^{j}$.

Lemma C.7. *Let $\{X_i\}_{i=1}^r$ and $\{Y_i\}_{i=1}^r$ be two sets of random positive integers ($X_i, Y_i \in \mathbb{N}\ \forall i \in [r]$), where for $u_{1:i-1} \in \mathbb{N}^{i-1}$, $\{X_i\}_{i=1}^r$ is characterized by the distribution*

$$F_{X_i|X_{1:i-1}}(u \mid u_{1:i-1}) \coloneqq \mathbb{P}(X_i \leq u \mid X_{1:i-1} = u_{1:i-1})$$

and the $\{Y_i\}_{i=1}^r$ are statistically independent, so that for any $u_{1:i-1} \in \mathbb{N}^{i-1}$

$$\mathbb{P}(Y_i \leq u \mid Y_{1:i-1} = u_{1:i-1}) = \mathbb{P}(Y_i \leq u) \eqqcolon F_{Y_i}(u).$$

Let $X \coloneqq \sum_{i=1}^r X_i$ and $Y \coloneqq \sum_{i=1}^r Y_i$, with $F_X(x) \coloneqq \mathbb{P}(X \leq x)$ and $F_Y(y) \coloneqq \mathbb{P}(Y \leq y)$. Suppose for all $i \in [r]$, $u \in \mathbb{R}$, and $u_{1:i-1} \in \mathbb{N}^{i-1}$, we have $F_{X_i|X_{1:i-1}}(u \mid u_{1:i-1}) \geq F_{Y_i}(u)$. Then $F_X(u) \geq F_Y(u)$ for all $u \in \mathbb{R}$ and $\mathbb{E}[X] \leq \mathbb{E}[Y]$.

We defer the proof of Lemma C.7 to the end of the section.

Corollary C.7.1. *For all $c \in \mathbb{R}$, $\mathbb{P}(M_{r-r_0} > c) \leq \mathbb{P}(B > c)$ and $\mathbb{E}[M_{r-r_0}] \leq \mathbb{E}[B]$.*

Proof. $\{B_i\}_{i=1}^{r-r_0}$ are statistically independent, and we know $\mathbb{P}(C_i \leq c \mid C_1, \ldots, C_{i-1}) \geq \mathbb{P}(B_i \leq c)$ for all c and all possible $C_1, \ldots, C_{i-1}$. Therefore by Lemma C.7, $\mathbb{P}(M_{r-r_0} \leq c) \geq \mathbb{P}(B \leq c)$ and so $\mathbb{P}(M_{r-r_0} > c) \leq \mathbb{P}(B > c)$ and $\mathbb{E}[M_{r-r_0}] \leq \mathbb{E}[B]$. □

Combining Corollary C.7.1 with eq. (24), for any $0 < \delta < 1$ we have

$$\begin{aligned}\mathbb{P}\Big(M_{r-r_0} > \Big(1 + \ln\frac{1}{\delta}\Big)\Big(\sum_{i=1}^{r-r_0} \frac{1}{1 - \frac{(r_0+i)(r_0+i-1)}{n(n-1)}}\Big)\Big) &\leq \mathbb{P}\Big(B > \Big(1 + \ln\frac{1}{\delta}\Big)\Big(\sum_{i=1}^{r-r_0} \frac{1}{1 - \frac{(r_0+i)(r_0+i-1)}{n(n-1)}}\Big)\Big) \\ &\leq \delta\end{aligned}$$

and

$$\mathbb{E}[M_{r-r_0}] \leq \sum_{i=1}^{r-r_0} \frac{1}{1 - \frac{(r_0+i)(r_0+i-1)}{n(n-1)}}.$$

In other words, with probability at least $1 - \delta$, $\Big(1 + \ln\frac{1}{\delta}\Big)\Big(\sum_{i=1}^{r-r_0} \frac{1}{1 - \frac{(r_0+i)(r_0+i-1)}{n(n-1)}}\Big)$ additional random measurements are sufficient to construct a rank r selection matrix from n items and a seed matrix $\boldsymbol{S}_0$ of rank r_0. We formalize the above facts in the following theorem:

Theorem C.8. *Let $\boldsymbol{S}_0$ be a given $m \times n$ selection matrix with rank $1 \leq r_0 \leq \min(m, n-2)$. Let $r_0 < r \leq n-1$ be given. Consider the following random sampling procedure: at sampling time $i \geq 1$, let $\boldsymbol{s}_i = s_{(p,q)} \in \mathbb{R}^n$ where p is sampled uniformly at random from $[n]$, q is sampled uniformly at random from $[n] \setminus \{p\}$, and where each $\boldsymbol{s}_i$ is sampled independently from $\boldsymbol{s}_j$ for $j \neq i$ and from $\boldsymbol{S}_0$. Let $\boldsymbol{S}$ be the selection matrix constructed by concatenating the vectors $\boldsymbol{s}_i$ into rows. Suppose rows are appended to $\boldsymbol{S}$ until $\operatorname{rank}(\big[\begin{smallmatrix}\boldsymbol{S}_0 \\ \boldsymbol{S}\end{smallmatrix}\big]) = r$, at which point sampling halts. Let M be the total number of rows in $\boldsymbol{S}$ resulting from this process. Then for any $0 < \delta < 1$,*

$$\mathbb{P}\Big(M > \Big(1 + \ln\frac{1}{\delta}\Big)\Big(\sum_{i=r_0+1}^{r} \frac{1}{1 - \frac{i(i-1)}{n(n-1)}}\Big)\Big) \leq \delta$$

and

$$\mathbb{E}[M] \leq \sum_{i=r_0+1}^{r} \frac{1}{1 - \frac{i(i-1)}{n(n-1)}}.$$

Proof of Lemma C.6: If $r = n-1$, then by Lemma C.1 $\boldsymbol{S}$ already has maximal rank and so its row space spans $\mathbb{R}^n$. In this case, $\mathbb{P}(\boldsymbol{s} \in \operatorname{rowsp}(\boldsymbol{S}) \mid \boldsymbol{S}) = 1$, and so the upper bound of the inequality is tight at 1. The lower bound is satisfied since, as $n \geq 2$ by definition of selection matrices, $2r/(n(n-1)) = 2/n \leq 1 \leq \mathbb{P}(\boldsymbol{s} \in \operatorname{rowsp}(\boldsymbol{S}) \mid \boldsymbol{S})$ and so is true.

Otherwise, assume $r \leq n-2$. Since $\boldsymbol{S}$ is rank r, there exists a set of r linearly independent rows, which we denote by $\{\boldsymbol{s}_i\}_{i=1}^r$, such that $\operatorname{rowsp}(\boldsymbol{S}) = \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^r)$. Without loss of generality, for each $\boldsymbol{s}_i$ assume that $p_i < q_i$, where $\boldsymbol{s}_i[p_i] = 1$ and $\boldsymbol{s}_i[q_i] = -1$: this assumption does not affect the span of $\{\boldsymbol{s}_i\}_{i=1}^r$, since for $p, q \in [n]$ with $p \neq q$, $\boldsymbol{s}_{(p,q)} = -\boldsymbol{s}_{(q,p)}$. In a slight abuse of notation, let $\boldsymbol{S} \setminus \{\boldsymbol{s}_i\}_{i=1}^r$ denote the remaining rows in $\boldsymbol{S}$. We therefore have

$$\begin{aligned}\mathbb{P}(\boldsymbol{s} \in \operatorname{rowsp}(\boldsymbol{S}) \mid \boldsymbol{S}) = \mathbb{P}(\boldsymbol{s} \in \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^r) \mid \boldsymbol{S}) &= \mathbb{P}(\boldsymbol{s} \in \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^r) \mid \{\boldsymbol{s}_i\}_{i=1}^r, \boldsymbol{S} \setminus \{\boldsymbol{s}_i\}_{i=1}^r) \\ &= \mathbb{P}(\boldsymbol{s} \in \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^r) \mid \{\boldsymbol{s}_i\}_{i=1}^r),\end{aligned}$$

where the last equality follows from the fact that $\boldsymbol{s}$ is statistically independent of $\boldsymbol{S}$.

Without loss of generality, suppose $p < q$. This does not affect our calculation of $\mathbb{P}(\boldsymbol{s} \in \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^r) \mid \{\boldsymbol{s}_i\}_{i=1}^r)$, since $\boldsymbol{s}_{(p,q)} \in \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^r) \iff \boldsymbol{s}_{(q,p)} \in \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^r)$, as $\boldsymbol{s}_{(p,q)} = -\boldsymbol{s}_{(q,p)}$. Therefore, we can calculate $\boldsymbol{s}_{(p,q)} \in \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^r)$ directly by counting which among the $\binom{n}{2}$ equally likely pairs with $p < q$ lies in the span of $\{\boldsymbol{s}_i\}_{i=1}^r$. Precisely, let $Q \coloneqq \{(p,q) : p, q \in [n],\ p < q,\ \boldsymbol{s}_{(p,q)} \in \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^r\})\}$. Then

$$\mathbb{P}(\boldsymbol{s} \in \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^r) \mid \{\boldsymbol{s}_i\}_{i=1}^r) = \frac{|Q|}{\binom{n}{2}}.$$

With these preliminaries established, we can easily lower bound $\mathbb{P}(\boldsymbol{s} \in \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^r) \mid \{\boldsymbol{s}_i\}_{i=1}^r)$: since $\boldsymbol{s}_i \in \operatorname{span}(\{\boldsymbol{s}_j\}_{j=1}^r)$ for each $i \in [r]$, Q contains the item pairs indexing the support of each

$\{\boldsymbol{s}_j\}_{j=1}^r$. Furthermore, since $\{\boldsymbol{s}_i\}_{i=1}^r$ are linearly independent, they must be distinct (i.e., for every $i, j \in [r]$, $\boldsymbol{s}_i \neq \boldsymbol{s}_j$) and so Q contains at least r distinct item pairs, i.e., $|Q| \geq r$. Hence,

$$\mathbb{P}(\boldsymbol{s} \in \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^r) \mid \{\boldsymbol{s}_i\}_{i=1}^r) = \frac{|Q|}{\binom{n}{2}} \geq \frac{r}{\binom{n}{2}} = \frac{2r}{n(n-1)},$$

proving the lower bound in the inequality.

Letting $S := \{\boldsymbol{s}_i\}_{i=1}^r$, we will upper bound $|Q|$ by analyzing the graph G_S (with the graph construction introduced in Appendix C.1) with linked vertex pairs C_S. Let I_S denote the set of distinct *item* pairs corresponding to linked vertices on G_S, i.e.,

$$I_S := \{(p, q) : p, q \in [n],\ p < q,\ \exists (i_p, i_q) \in C_S,\ \boldsymbol{s}_{i_p}[p] \neq 0,\ \boldsymbol{s}_{i_q}[q] \neq 0\}.$$

From Lemma C.5, $(p, q) \in Q \implies (p, q) \in I_S$ and so $|Q| \leq |I_S|$ and

$$\mathbb{P}(\boldsymbol{s} \in \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^r) \mid \{\boldsymbol{s}_i\}_{i=1}^r) = \frac{|Q|}{\binom{n}{2}} \leq \frac{|I_S|}{\binom{n}{2}}. \tag{25}$$

We can therefore upper bound $\mathbb{P}(\boldsymbol{s} \in \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^r) \mid \{\boldsymbol{s}_i\}_{i=1}^r)$ by upper bounding $|I_S|$.

To proceed, without loss of generality that suppose G_S has exactly c distinct subgraphs ($c \in [r]$) $G_1 = (V_1, E_1), \ldots G_c = (V_c, E_c)$ where $V_1, \ldots, V_c$ are a partition of $[r]$, such that for every k and every $i, j \in V_k$, vertices i and j are linked on G_S (and hence G_k), and for every $k, \ell \in [c]$ with $k \neq \ell$ and every $i \in V_k$, $j \in V_\ell$, vertices i and j are not linked on G_S. We next define item pairs according to which subgraph they pertain to: let

$$I_k := \{(p, q) : p, q \in [n],\ p < q,\ \exists (i_p, i_q) \in V_k \text{ s.t. } \boldsymbol{s}_{i_p}[p] \neq 0,\ \boldsymbol{s}_{i_q}[q] \neq 0\}. \tag{26}$$

Note that for any given item pair $(p, q) \in I_S$ with corresponding row indices $(i_p, i_q) \in C_S$ such that $\boldsymbol{s}_{i_p}[p] \neq 0$ and $\boldsymbol{s}_{i_q}[q] \neq 0$, there must exist $k \in [c]$ such that $i_p, i_q \in V_k$. Hence, $I_S = \bigcup_{k=1}^c I_k$ and therefore

$$|I_S| \leq \sum_{k=1}^{c} |I_k|. \tag{27}$$

To calculate $|I_k|$, first define N_k to be the number of items supported by subgraph G_k:

$$N_k := \{i : i \in [n],\ \exists j \in V_k \text{ s.t. } \boldsymbol{s}_j[i] \neq 0\}.$$

Since all vertices in V_k are linked on G_k (by construction), $|I_k|$ is exactly equal to all possible pair permutations of items in N_k, i.e., $|I_k| = \binom{|N_k|}{2}$. Let $r_k := |V_k|$ denote the number of vertices in subgraph G_k, noting that $\sum_{k=1}^c r_k = r$. For each $k \in [c]$ we then have $|N_k| = r_k + 1$: to see this, consider the selection matrix $\boldsymbol{S}_{V_k}$ constructed from the rows $\{\boldsymbol{s}_i\}_{i \in V_k}$, and note that the rows $\{\boldsymbol{s}_i\}_{i \in V_k}$ are linearly independent by construction (since $\{\boldsymbol{s}_i\}_{i \in V_k} \subseteq \{\boldsymbol{s}_i\}_{i=1}^r$). Furthermore, suppose without loss of generality that $\boldsymbol{S}_{V_k}$ is incremental: since $\{\boldsymbol{s}_i\}_{i \in V_k}$ are linearly independent, by Corollary C.4.1 we can always find a permutation of these rows such that the resulting matrix $\boldsymbol{S}_{V_k}$ is incremental. We will show below that each row of $\boldsymbol{S}_{V_k}$ introduces *exactly* one new item.

Let $\boldsymbol{S}_{V_k}^{(t)}$ denote the submatrix of $\boldsymbol{S}_{V_k}$ consisting of the first t rows: note that each $\boldsymbol{S}_{V_k}^{(t)}$ is also incremental. Denote the tth row of $\boldsymbol{S}_{V_k}$ by $\boldsymbol{s}^{(t)}$. Suppose by contradiction that there exists $1 < i \leq r_k$ such that $\boldsymbol{s}^{(i)}$ introduces exactly two new items that are not supported in $\boldsymbol{S}_{V_k}^{(i-1)}$, and consider any $j < i$. Since every row index in V_k is linked on G_k, there must exist at least one finite sequence of distinct indices $\{k_\ell\}_{\ell=1}^T \subseteq [r_k] \setminus \{i, j\}$ such that $\boldsymbol{s}^{(i)}$ and $\boldsymbol{s}^{(k_1)}$ share an item, each $\boldsymbol{s}^{(k_\ell)}$ shares an item with $\boldsymbol{s}^{(k_{\ell-1})}$ for $1 < \ell \leq T$, and $\boldsymbol{s}^{(k_T)}$ shares an item with $\boldsymbol{s}^{(j)}$. Note that $k_1 > i$, since $\boldsymbol{s}^{(i)}$ cannot share an item directly with any row in $\boldsymbol{S}_{V_k}^{(i-1)}$ (due to our contradictory assumption). Let $k^* = \max_{\ell \in [T]} k_\ell$; we know from the above argument that $k^* \geq k_1 > i > j$ and so $k^* > i, j, \{k_\ell\}_{\ell=1}^T \setminus \{k^*\}$. By definition of $\{k_\ell\}_{\ell=1}^T$, $\boldsymbol{s}^{(k^*)}$ shares an item with two distinct indices $k_a, k_b \in (\{k_\ell\}_{\ell=1}^T \cup \{i, j\}) \setminus k^*$: let p_a be the item index shared with k_a and p_b denote the item index shared with k_b. Since $k^* > i, j, \{k_\ell\}_{\ell=1}^T \setminus \{k^*\}$, $k_a < k^*$ and $k_b < k^*$, and therefore both p_a and p_b must appear in $\boldsymbol{S}_{V_k}^{(k^*-1)}$. However, this is a contradiction since $\boldsymbol{S}_{V_k}^{(k^*)}$ is incremental meaning that $\boldsymbol{s}^{(k^*)}$ must introduce at least one new item. Therefore, there cannot exist index $1 < i \leq r_k$ such that

$\boldsymbol{s}^{(i)}$ introduces exactly two new items that are not supported in $\boldsymbol{S}_{V_k}^{(i-1)}$. Hence, hence the first row of $\boldsymbol{S}_{V_k}$ introduces 2 new items and each subsequent row ($r_k - 1$ additional rows in total) introduces exactly one new item, resulting in $2 + r_k - 1 = r_k + 1$ supported columns in total, i.e., $|N_k| = r_k + 1$ and so $|I_k| = \binom{|N_k|}{2} = \binom{r_k+1}{2}$.

Therefore, by eq. (27),

$$|I_S| \leq \sum_{k=1}^{c} \binom{r_k + 1}{2} = \frac{1}{2} \sum_{k=1}^{c} (r_k + 1) r_k,$$

where we recall that $\sum_{k=1}^{c} r_k = r$ and $c \in [r]$. To get an upper bound on $|I_S|$ that only depends on r, we can maximize this bound over the choice of $\{r_k\}_{k=1}^{c}$ and c. We propose that $c = 1$ and hence $r_1 = r$ maximizes this bound: consider any other $c \in [r]$ and $\{r_k\}_{k=1}^{c}$ such that $\sum_{k=1}^{c} r_k = r$. We have

$$\begin{aligned}
(r+1)r - \sum_{k=1}^{c}(r_k + 1)r_k &= r^2 + r - \sum_{k=1}^{c}(r_k^2 + r_k) \\
&= \Big(r^2 - \sum_{k=1}^{c} r_k^2\Big) + r - \sum_{k=1}^{c} r_k \\
&= r^2 - \sum_{k=1}^{c} r_k^2 \\
&\geq r^2 - \sum_{k=1}^{c} r_k^2 - \sum_{k \neq \ell} r_k r_\ell \\
&= r^2 - \left(\sum_{k=1}^{c} r_k\right)^2 \\
&= 0.
\end{aligned}$$

Hence, for any $c \in [r]$ and $\{r_k\}_{k=1}^{c}$ such that $\sum_{k=1}^{c} r_k = r$,

$$\frac{1}{2} \sum_{k=1}^{c} (r_k + 1) r_k \leq \frac{1}{2}(r+1)r,$$

and so $|I_S| \leq \frac{1}{2}(r+1)r$. Recalling eq. (25), we therefore have

$$\mathbb{P}(\boldsymbol{s} \in \operatorname{span}(\{\boldsymbol{s}_i\}_{i=1}^{r}) \mid \{\boldsymbol{s}_i\}_{i=1}^{r}) \leq \frac{1}{2} \frac{(r+1)r}{\binom{n}{2}} = \frac{(r+1)r}{n(n-1)}.$$

Proof of Lemma C.7: Let $X^{(i)} := \sum_{j=i}^{r} X_j$ and $Y^{(i)} := \sum_{j=i}^{r} Y_j$: note that $X = X^{(1)}$ and $Y = Y^{(1)}$. Let

$$F_{X^{(i)}|X_{1:i-1}}(u \mid u_{1:i-1}) = \mathbb{P}(X^{(i)} \leq u \mid X_{1:i-1} = u_{1:i-1})$$

and

$$F_{Y^{(i)}}(u) := \mathbb{P}(Y^{(i)} \leq u).$$

We will prove by induction that $F_{X^{(1)}}(u) \geq F_{Y^{(1)}}(u)$ for all $u \in \mathbb{R}$, and therefore $F_X(u) \geq F_Y(u)$. Starting at $i = r$, we have by assumption that for all $u \in \mathbb{R}$, and $u_{1:r-1} \in \mathbb{N}^{r-1}$,

$$F_{X^{(r)}|X_{1:r-1}}(u \mid u_{1:r-1}) = F_{X_r|X_{1:r-1}}(u \mid u_{1:r-1}) \geq F_{Y_r}(u) = F_{Y^{(r)}}(u).$$

Now, let $1 \leq m < r$ be given, and suppose by induction that for any $u_{1:m} \in \mathbb{N}^m$, we have

$$F_{X^{(m+1)}|X_{1:m}}(u \mid u_{1:m}) \geq F_{Y^{(m+1)}}(u).$$

Expanding $F_{X^{(m)}|X_{1:m-1}}(u \mid u_{1:m-1})$,

$$\begin{aligned}
&F_{X^{(m)}|X_{1:m-1}}(u \mid u_{1:m-1}) = \\
&= \mathbb{P}(X^{(m)} \leq u \mid X_{1:m-1} = u_{1:m-1})
\end{aligned}$$

$$
\begin{align*}
&= \mathbb{P}\Big(\sum_{j=m}^{r} X_j \le u \mid X_{1:m-1} = u_{1:m-1}\Big) \\
&= \mathbb{P}\Big(\sum_{j=m+1}^{r} X_j \le u - X_m \mid X_{1:m-1} = u_{1:m-1}\Big) \\
&= \sum_{v=1}^{\infty} \mathbb{P}\Big(\sum_{j=m+1}^{r} X_j \le u - X_m \mid X_{1:m-1} = u_{1:m-1}, X_m = v\Big)\mathbb{P}(X_m = v \mid X_{1:m-1} = u_{1:m-1}) \\
&= \sum_{v=1}^{\infty} F_{X^{(m+1)}|X_{1:m}}(u - v \mid u_{1:m-1}, v)(F_{X_m|X_{1:m-1}}(v \mid u_{1:i-1}) - F_{X_m|X_{1:m-1}}(v - 1 \mid u_{1:i-1})).
\end{align*}
$$

Similarly we have

$$
\begin{align*}
F_{Y^{(m)}}(u) &= \mathbb{P}(Y^{(m)} \le u) \\
&= \mathbb{P}\Big(\sum_{j=m}^{r} Y_j \le u\Big) \\
&= \mathbb{P}\Big(\sum_{j=m+1}^{r} Y_j \le u - Y_m\Big) \\
&= \sum_{v=1}^{\infty} \mathbb{P}\Big(\sum_{j=m+1}^{r} Y_j \le u - Y_m \mid Y_m = v\Big)\mathbb{P}(Y_m = v) \\
&= \sum_{v=1}^{\infty} \mathbb{P}\Big(\sum_{j=m+1}^{r} Y_j \le u - v\Big)\mathbb{P}(Y_m = v) \tag{28} \\
&= \sum_{v=1}^{\infty} F_{Y^{(m+1)}}(u - v)(F_{Y_m}(v) - F_{Y_m}(v - 1)),
\end{align*}
$$

where eq. (28) follows from the fact that $\{Y_i\}_{i=1}^{r}$ are statistically independent. Therefore, letting $u_{1:m-1} \in \mathbb{N}^{m-1}$ be given,

$$
\begin{align*}
&F_{X^{(m)}|X_{1:m-1}}(u \mid u_{1:m-1}) - F_{Y^{(m)}}(u) \\
&= \sum_{v=1}^{\infty} F_{X^{(m+1)}|X_{1:m}}(u - v \mid u_{1:m-1}, v)(F_{X_m|X_{1:m-1}}(v \mid u_{1:i-1}) \\
&- F_{X_m|X_{1:m-1}}(v - 1 \mid u_{1:i-1})) - F_{Y^{(m+1)}}(u - v)(F_{Y_m}(v) - F_{Y_m}(v - 1) \\
&\ge \sum_{v=1}^{\infty} F_{Y^{(m+1)}}(u - v)(F_{X_m|X_{1:m-1}}(v \mid u_{1:i-1}) - F_{X_m|X_{1:m-1}}(v - 1 \mid u_{1:i-1})) - \\
&F_{Y^{(m+1)}}(u - v)(F_{Y_m}(v) - F_{Y_m}(v - 1)) \tag{29} \\
&= \sum_{v=1}^{\infty} F_{Y^{(m+1)}}(u - v)(F_{X_m|X_{1:m-1}}(v \mid u_{1:i-1}) - F_{Y_m}(v)) \\
&- \sum_{v=1}^{\infty} F_{Y^{(m+1)}}(u - v)(F_{X_m|X_{1:m-1}}(v - 1 \mid u_{1:i-1}) - F_{Y_m}(v - 1)) \\
&= \sum_{v=1}^{\infty} F_{Y^{(m+1)}}(u - v)(F_{X_m|X_{1:m-1}}(v \mid u_{1:i-1}) - F_{Y_m}(v)) \\
&- \sum_{z=0}^{\infty} F_{Y^{(m+1)}}(u - z - 1)(F_{X_m|X_{1:m-1}}(z \mid u_{1:i-1}) - F_{Y_m}(z)) \quad \text{where } z := v - 1 \\
&= \sum_{v=1}^{\infty} (F_{Y^{(m+1)}}(u - v) - F_{Y^{(m+1)}}(u - v - 1))(F_{X_m|X_{1:m-1}}(v \mid u_{1:i-1}) - \\
&F_{Y_m}(v)) - F_{Y^{(m+1)}}(u - 1)(F_{X_m|X_{1:m-1}}(0 \mid u_{1:i-1}) - F_{Y_m}(0)) \\
&= \sum_{v=1}^{\infty} (F_{Y^{(m+1)}}(u - v) - F_{Y^{(m+1)}}(u - v - 1))(F_{X_m|X_{1:m-1}}(v \mid u_{1:i-1}) - F_{Y_m}(v)) \tag{30}
\end{align*}
$$

$$\geq 0 \tag{31}$$

where eq. (29) is by inductive assumption, eq. (30) is because X_m, Y_m are non-negative and so

$$F_{X_m|X_{1:m-1}}(0 \mid u_{1:i-1}) = F_{Y_m|Y_{1:m-1}}(0 \mid v_{1:i-1})) = 0,$$

and eq. (31) is due to the fact that $F_{Y^{(m+1)}}(u)$ is non-decreasing and hence for all $v \geq 1$,

$$F_{Y^{(m+1)}}(u - v) - F_{Y^{(m+1)}}(u - v - 1) \geq 0,$$

and by assumption we have

$$F_{X_m|X_{1:m-1}}(v \mid u_{1:i-1}) - F_{Y_m}(v) \geq 0.$$

Taking $m = 1$, we have $F_{X^{(1)}}(u) - F_{Y^{(1)}}(u) \geq 0$ i.e., $F_X(u) \geq F_Y(u)$.

Using the fact that $\mathbb{E}[X] = \sum_{u=0}^{\infty} \mathbb{P}(X > u) = \sum_{u=0}^{\infty}(1 - F_X(u))$ and similarly $\mathbb{E}[Y] = \sum_{u=0}^{\infty}(1 - F_Y(u))$, we have

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{u=0}^{\infty}(1 - F_X(u)) \\
&\leq \sum_{u=0}^{\infty}(1 - F_Y(u)) \\
&= \mathbb{E}[Y].
\end{aligned}$$

C.3 Proof of Proposition 2.1 and additional discussion

If $\boldsymbol{\Gamma}$ has full column rank, then its $D + dK$ columns are linearly independent and so $\operatorname{rank}(\boldsymbol{\Gamma}) = D + dK$. Since the rank of $\boldsymbol{\Gamma}$ is upper bounded by its number of rows, we require $\sum_{k=1}^{K} m_k \geq D + dK$. Next we will show in turn that each condition in Proposition 2.1 is necessary for $\boldsymbol{\Gamma}$ to have full column rank:

(a) In order for all $D + dK$ columns in $\boldsymbol{\Gamma}$ to be linearly independent, it must be the case that for each $k \in [K]$, the columns corresponding to user k are linearly independent, given by

$$\begin{bmatrix} \mathbf{0}_{m_1,d} \\ \vdots \\ \boldsymbol{S}_k \boldsymbol{X}^T \\ \vdots \\ \mathbf{0}_{m_K,d} \end{bmatrix}.$$

Clearly this is only possible if the d columns in $\boldsymbol{S}_k \boldsymbol{X}^T$ are linearly independent (since padding by zeros does not affect linear independence of columns), i.e., $\operatorname{rank}(\boldsymbol{S}_k \boldsymbol{X}^T) = d$. Since $\operatorname{rank}(\boldsymbol{S}_k \boldsymbol{X}^T) \leq \operatorname{rank}(\boldsymbol{S}_k)$, we require $\operatorname{rank}(\boldsymbol{S}_k) \geq d$, which implies $m_k \geq d$ since $\boldsymbol{S}_k$ has m_k rows.

(b) Since $\operatorname{rank}(\boldsymbol{\Gamma}) = D + dK$, $\boldsymbol{\Gamma}$ must have $D + dK$ linearly independent rows. Observing eq. (3), each user's block of m_k rows is given by

$$\begin{bmatrix} \boldsymbol{S}_k \boldsymbol{X}_\otimes^T & \mathbf{0}_{m_k,d} & \cdots & \boldsymbol{S}_k \boldsymbol{X}^T & \cdots & \mathbf{0}_{m_k,d} \end{bmatrix}, \tag{32}$$

which has the same column space, and therefore the same rank, as $\boldsymbol{S}_k \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix}$. Therefore, the number of linearly independent rows in eq. (32) is equal to the rank of $\boldsymbol{S}_k \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix}$, and so the number of linearly independent rows in $\boldsymbol{\Gamma}$ is upper bounded by $\sum_{k=1}^{K} \operatorname{rank}(\boldsymbol{S}_k \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix})$, which for $\boldsymbol{\Gamma}$ with full column rank must be at least $D + dK$. Since $\operatorname{rank}(\boldsymbol{S}_k \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix}) \leq \operatorname{rank}(\boldsymbol{S}_k)$, we have $\sum_{k=1}^{K} \operatorname{rank}(\boldsymbol{S}_k \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix}) \leq \sum_{k=1}^{K} \operatorname{rank}(\boldsymbol{S}_k)$ and therefore we also require $\sum_{k=1}^{K} \operatorname{rank}(\boldsymbol{S}_k) \geq D + dK$.

(c) Consider any $\boldsymbol{\eta} \in \mathbb{R}^D$ and $\boldsymbol{v} \in \mathbb{R}^d$. Recalling eq. (3), multiplying $\boldsymbol{\Gamma}$ by $[\boldsymbol{\eta}^T \underbrace{\boldsymbol{v}^T \cdots \boldsymbol{v}^T}_{K \text{ times}}]^T$ is equivalent to

$$\boldsymbol{\Gamma} \begin{bmatrix} \boldsymbol{\eta} \\ \boldsymbol{v} \\ \vdots \\ \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \boldsymbol{S}_1 \boldsymbol{X}_\otimes^T \boldsymbol{\eta} + \boldsymbol{S}_1 \boldsymbol{X}^T \boldsymbol{v} \\ \vdots \\ \boldsymbol{S}_K \boldsymbol{X}_\otimes^T \boldsymbol{\eta} + \boldsymbol{S}_K \boldsymbol{X}^T \boldsymbol{v} \end{bmatrix} = \begin{bmatrix} \boldsymbol{S}_1 \\ \vdots \\ \boldsymbol{S}_K \end{bmatrix} \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta} \\ \boldsymbol{v} \end{bmatrix} = \boldsymbol{S}_T \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\eta} \\ \boldsymbol{v} \end{bmatrix}. \quad (33)$$

By the rank-nullity theorem, $\ker(\boldsymbol{S}_T \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix})$ is trivial if and only if $\operatorname{rank}(\boldsymbol{S}_T \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix}) = D + d$ (recall that $\begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix}$ has $D + d$ columns). Therefore if $\operatorname{rank}(\boldsymbol{S}_T \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix}) < D + d$, there exists a $\left[\begin{smallmatrix} \boldsymbol{\eta} \\ \boldsymbol{v} \end{smallmatrix}\right] \neq \boldsymbol{0}$ such that $\boldsymbol{S}_T \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix} \left[\begin{smallmatrix} \boldsymbol{\eta} \\ \boldsymbol{v} \end{smallmatrix}\right] = \boldsymbol{0}$ and therefore exists a nonzero vector in $\mathbb{R}^{D+dK}$ given by $[\boldsymbol{\eta}^T \underbrace{\boldsymbol{v}^T \cdots \boldsymbol{v}^T}_{K \text{ times}}]^T$ such that

$$\boldsymbol{\Gamma} \begin{bmatrix} \boldsymbol{\eta} \\ \boldsymbol{v} \\ \vdots \\ \boldsymbol{v} \end{bmatrix} = \boldsymbol{0},$$

which would imply that $\boldsymbol{\Gamma}$ is rank deficient. Therefore, we require $\operatorname{rank}(\boldsymbol{S}_T \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix}) = D + d$, and since $\operatorname{rank}(\boldsymbol{S}_T \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix}) \leq \min(\operatorname{rank}(\boldsymbol{S}_T), \operatorname{rank}(\begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix}))$ this implies $\operatorname{rank}(\boldsymbol{S}_T) \geq D + d$ and $\operatorname{rank}(\begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix}) \geq D + d$. Since $\boldsymbol{S}_T$ is itself a selection matrix, by Lemma C.1 we require $n \geq D + d + 1$ in order for $\operatorname{rank}(\boldsymbol{S}_T) \geq D + d$. □

In the setting where the metric $\boldsymbol{M}$ is low-rank with rank $r < d$, we conjecture that the required item scaling grows as $O(rd)$ rather than $O(d^2)$, which is a much gentler increase especially when $r \ll d$. Our intuition for this conjecture is as follows: the requirement for $O(d^2)$ items in the full-rank metric case comes from part (c) above, which says that if a hypothetical *single* user existed that answered the queries assigned to *all* users, then such a system would require $O(d^2)$ measurements due to the $O(d^2)$ degrees of freedom in the metric. Due to properties of selection matrices, we require the same or greater order of items as measurements since intuitively one new item is required per independent measurement (see Lemma C.1). If $\boldsymbol{M}$ is rank $r < d$, there are only dr degrees of freedom in $\boldsymbol{M}$, and hence we believe that part (c) above would only require $O(dr)$ independent measurements for the hypothetical single user and therefore only $O(dr)$ items. Concretely, by rewriting the unquantized measurements in eq. (2) as a matrix inner product between $[\, \boldsymbol{M} \; \boldsymbol{v}_1 \; \ldots \; \boldsymbol{v}_K \,]$ and a corresponding measurement matrix only depending on the items (see eq. (40) in the proof of Theorem 3.1 for an example of this technique), and using low-rank matrix recovery techniques such as those described in [47], we believe that one can show only $O(dr)$ independent measurements are required for the hypothetical single user and therefore only $O(dr)$ items are required.

C.4 Proof of Proposition 2.2

We first permute the rows of $\boldsymbol{\Gamma}$ as follows (row permutations do not change matrix rank): first, define $\boldsymbol{\Gamma}^{(1)}$ as

$$\boldsymbol{\Gamma}^{(1)} := \begin{bmatrix} \boldsymbol{S}_1^{(1)} \boldsymbol{X}_\otimes^T & \boldsymbol{S}_1^{(1)} \boldsymbol{X}^T & \boldsymbol{0}_{d,d} & \cdots & \boldsymbol{0}_{d,d} \\ \boldsymbol{S}_2^{(1)} \boldsymbol{X}_\otimes^T & \boldsymbol{0}_{d,d} & \boldsymbol{S}_2^{(1)} \boldsymbol{X}^T & \cdots & \boldsymbol{0}_{d,d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{S}_K^{(1)} \boldsymbol{X}_\otimes^T & \boldsymbol{0}_{d,d} & \boldsymbol{0}_{d,d} & \cdots & \boldsymbol{S}_K^{(1)} \boldsymbol{X}^T \end{bmatrix} \quad (34)$$

and $\boldsymbol{\Gamma}^{(2)}$ as

$$\boldsymbol{\Gamma}^{(2)} := \boldsymbol{P} \begin{bmatrix} \boldsymbol{S}_1^{(2)} \boldsymbol{X}_\otimes^T & \boldsymbol{S}_1^{(2)} \boldsymbol{X}^T & \boldsymbol{0}_{m_1-d,d} & \cdots & \boldsymbol{0}_{m_1-d,d} \\ \boldsymbol{S}_2^{(2)} \boldsymbol{X}_\otimes^T & \boldsymbol{0}_{m_2-d,d} & \boldsymbol{S}_2^{(2)} \boldsymbol{X}^T & \cdots & \boldsymbol{0}_{m_2-d,d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{S}_K^{(2)} \boldsymbol{X}_\otimes^T & \boldsymbol{0}_{m_K-d,d} & \boldsymbol{0}_{m_K-d,d} & \cdots & \boldsymbol{S}_K^{(2)} \boldsymbol{X}^T \end{bmatrix}. \quad (35)$$

Finally, define $\widehat{\boldsymbol{\Gamma}} = \begin{bmatrix} \boldsymbol{\Gamma}^{(1)} \\ \boldsymbol{\Gamma}^{(2)} \end{bmatrix}$. Since $\widehat{\boldsymbol{\Gamma}}$ is simply a permutation of the rows in $\boldsymbol{\Gamma}$, $\operatorname{rank}(\widehat{\boldsymbol{\Gamma}}) = \operatorname{rank}(\boldsymbol{\Gamma})$. Therefore, if we show that the $D + dK$ rows in $\widehat{\boldsymbol{\Gamma}}$ are linearly independent and hence $\operatorname{rank}(\widehat{\boldsymbol{\Gamma}}) = D + dK$, we will have shown that $\operatorname{rank}(\boldsymbol{\Gamma}) = D + dK$ and so $\boldsymbol{\Gamma}$ is full column rank.

We will start by examining the rows in $\boldsymbol{\Gamma}^{(1)}$. For $k \in [K]$, let $\boldsymbol{Q}_k^{(1)} := \boldsymbol{S}_k^{(1)} \boldsymbol{X}^T$; evaluating this matrix product, the ith row of $\boldsymbol{Q}_k^{(1)}$ is given by $\boldsymbol{x}_{p_{k,i}} - \boldsymbol{x}_{q_{k,i}}$, where $p_{k,i}$ indexes the $+1$ entry in the ith row of $\boldsymbol{S}_k^{(1)}$ and $q_{k,i}$ indexes the -1 entry in the ith row of $\boldsymbol{S}_k^{(1)}$. Since each $\boldsymbol{x}_i$ is i.i.d. distributed according to p_X, and p_X is absolutely continuous with respect to the Lebesgue measure, for any $i \neq j$ we have that $\boldsymbol{z}_{i,j} := \boldsymbol{x}_i - \boldsymbol{x}_j$ is distributed according to some distribution p_Z (which does not depend on i or j since $\boldsymbol{x}_i$ is i.i.d. for all i) that is also absolutely continuous with respect to the Lebesgue measure.

Inspecting the first row of $\boldsymbol{Q}_k^{(1)}$, we then have

$$\mathbb{P}(\boldsymbol{x}_{p_{k,1}} - \boldsymbol{x}_{q_{k,1}} = \boldsymbol{0}) = \mathbb{P}_Z(\boldsymbol{z} = \boldsymbol{0}) = 0,$$

where the last equality follows since $\mu(\{\boldsymbol{0}\}) = 0$, where μ is the Lebesgue measure, and p_Z is absolutely continuous. Hence, with probability 1, $\boldsymbol{x}_{p_{k,1}} - \boldsymbol{x}_{q_{k,1}}$ is nonzero and spans a 1-dimensional subspace of $\mathbb{R}^d$.

Let $\boldsymbol{w}$ be a vector orthogonal to $\boldsymbol{x}_{p_{k,1}} - \boldsymbol{x}_{q_{k,1}}$, and consider $\boldsymbol{x}_{p_{k,2}} - \boldsymbol{x}_{q_{k,2}}$, which is the second row of $\boldsymbol{Q}_k^{(1)}$. Since by assumption $\boldsymbol{S}_k^{(1)}$ is incremental, at least one of $p_{k,2}$ or $q_{k,2}$ is not equal to $p_{k,1}$ or $q_{k,1}$. Suppose that both $p_{k,2}, q_{k,2} \notin \{p_{k,1}, q_{k,1}\}$. Then

$$\mathbb{P}(\boldsymbol{w}^T(\boldsymbol{x}_{p_{k,2}} - \boldsymbol{x}_{q_{k,2}}) = 0 \mid \boldsymbol{x}_{p_{k,1}}, \boldsymbol{x}_{q_{k,1}}) = \mathbb{P}_Z(\boldsymbol{w}^T\boldsymbol{z} = 0 \mid \boldsymbol{x}_{p_{k,1}}, \boldsymbol{x}_{q_{k,1}}) = 0,$$

where the last equality follows from the fact that $\mu(\{\boldsymbol{w}^T\boldsymbol{z} = 0 : \boldsymbol{z} \in \mathbb{R}^d\}) = 0$ and p_Z is absolutely continuous.

Now suppose that exactly one of $p_{k,2}$ or $q_{k,2}$ is not equal to $p_{k,1}$ or $q_{k,1}$. Without loss of generality, suppose $q_{k,2}$ is equal to $p_{k,1}$ or $q_{k,1}$ (the same argument holds if this were true for $p_{k,2}$ instead). Then

$$\begin{aligned}
\mathbb{P}(\boldsymbol{w}^T(\boldsymbol{x}_{p_{k,2}} - \boldsymbol{x}_{q_{k,2}}) = 0 \mid \boldsymbol{x}_{p_{k,1}}, \boldsymbol{x}_{q_{k,1}}) &= \mathbb{P}(\boldsymbol{w}^T\boldsymbol{x}_{p_{k,2}} - \boldsymbol{w}^T\boldsymbol{x}_{q_{k,2}} = 0 \mid \boldsymbol{x}_{p_{k,1}}, \boldsymbol{x}_{q_{k,1}}, \boldsymbol{x}_{q_{k,2}}) \\
&= \mathbb{P}_X(\boldsymbol{w}^T\boldsymbol{x} - c = 0 \mid \boldsymbol{x}_{p_{k,1}}, \boldsymbol{x}_{q_{k,1}}, \boldsymbol{x}_{q_{k,2}}) \quad \text{where } c \text{ is a constant.} \\
&= 0,
\end{aligned}$$

where the first equality follows from the fact that $q_{k,2} \in \{p_{k,1}, q_{k,1}\}$, the second equality follows from the fact that when conditioned on $\boldsymbol{x}_{q_{k,2}}$, $\boldsymbol{w}^T\boldsymbol{x}_{q_{k,2}}$ is a constant (which we denote by c) and that $\boldsymbol{x}_{p_{k,2}}$ is distributed as p_X and is independent of other $\boldsymbol{x}_j$ for $j \neq p_{k,2}$, and the final equality follows from the fact that $\mu(\{\boldsymbol{w}^T\boldsymbol{x} - c = 0 : \boldsymbol{x} \in \mathbb{R}^d\}) = 0$ and p_X is absolutely continuous.

In either scenario, $\mathbb{P}(\boldsymbol{w}^T(\boldsymbol{x}_{p_{k,2}} - \boldsymbol{x}_{q_{k,2}}) = 0 \mid \boldsymbol{x}_{p_{k,1}}, \boldsymbol{x}_{q_{k,1}}) = 0$. Hence, when conditioned on $\boldsymbol{x}_{p_{k,1}}$ and $\boldsymbol{x}_{q_{k,1}}$, with probability 1 $\boldsymbol{x}_{p_{k,2}} - \boldsymbol{x}_{q_{k,2}}$ includes a component orthogonal to $\boldsymbol{x}_{p_{k,1}} - \boldsymbol{x}_{q_{k,1}}$ and therefore does not lie in the span of $\boldsymbol{x}_{p_{k,1}} - \boldsymbol{x}_{q_{k,1}}$. Denote the first j rows of $\boldsymbol{Q}_k^{(1)}$ as

$$\boldsymbol{Q}_k^{(1)}[1:j] = \begin{bmatrix} (\boldsymbol{x}_{p_{k,1}} - \boldsymbol{x}_{q_{k,1}})^T \\ \vdots \\ (\boldsymbol{x}_{p_{k,j}} - \boldsymbol{x}_{q_{k,j}})^T \end{bmatrix}.$$

Then, from the above argument, $\mathbb{P}(\operatorname{rank}(\boldsymbol{Q}_k^{(1)}[1:2]) = 2 \mid \boldsymbol{x}_{p_{k,1}}, \boldsymbol{x}_{q_{k,1}}) = 1$. This is true for any $\boldsymbol{x}_{p_{k,1}}, \boldsymbol{x}_{q_{k,1}}$ satisfying $\boldsymbol{x}_{p_{k,1}} - \boldsymbol{x}_{q_{k,1}} \neq \boldsymbol{0}$, which we know occurs with probability 1, and so marginalizing over this event we have $\mathbb{P}(\operatorname{rank}(\boldsymbol{Q}_k^{(1)}[1:2]) = 2) = 1$.

If $d > 2$, let $2 \leq m < d$ be given, and suppose by induction that $\mathbb{P}(\operatorname{rank}(\boldsymbol{Q}_k^{(1)}[1:m]) = m) = 1$. The rows of $\boldsymbol{Q}_k^{(1)}[1:m]$ are constructed from vectors $\boldsymbol{x}_i$ where $i \in M$ and $M := \{p_{k,j}\}_{j=1}^m \cup \{q_{k,j}\}_{j=1}^m$.

Let $\boldsymbol{w}$ be a vector in the orthogonal subspace to $\operatorname{rowsp}(\boldsymbol{Q}_k^{(1)}[1:m])$. Consider row $m+1$ of $\boldsymbol{Q}_k^{(1)}$, given by $\boldsymbol{x}_{p_{k,m+1}} - \boldsymbol{x}_{q_{k,m+1}}$. Since $\boldsymbol{S}_k^{(1)}$ is incremental, at least one of $p_{k,m+1}$ or $q_{k,m+1}$ is not in

M. First suppose this is true for both $p_{k,m+1}$ and $q_{k,m+1}$. Then by similar arguments as above,

$$\mathbb{P}(\boldsymbol{w}^T(\boldsymbol{x}_{p_{k,m+1}} - \boldsymbol{x}_{q_{k,m+1}}) = 0 \mid \{\boldsymbol{x}_i\}_{i \in M}) = \mathbb{P}_Z(\boldsymbol{w}^T\boldsymbol{z} = 0 \mid \{\boldsymbol{x}_i\}_{i \in M}) = 0.$$

Now, instead suppose without loss of generality that only $q_{k,m+1} \in M$ (an identical argument holds for $p_{k,m+1} \in M$). Then by similar arguments as above,

$$\begin{aligned}
&\mathbb{P}(\boldsymbol{w}^T(\boldsymbol{x}_{p_{k,m+1}} - \boldsymbol{x}_{q_{k,m+1}}) = 0 \mid \{\boldsymbol{x}_i\}_{i \in M}) = \mathbb{P}(\boldsymbol{w}^T\boldsymbol{x}_{p_{k,m+1}} - \boldsymbol{w}^T\boldsymbol{x}_{q_{k,m+1}} = 0 \mid \boldsymbol{x}_{q_{k,m+1}}, \{\boldsymbol{x}_i\}_{i \in M}) \\
&= \mathbb{P}_X(\boldsymbol{w}^T\boldsymbol{x} - c = 0 \mid \boldsymbol{x}_{q_{k,m+1}}, \{\boldsymbol{x}_i\}_{i \in M}) \quad \text{where } c \text{ is a constant.} \\
&= 0.
\end{aligned}$$

In either scenario, $\mathbb{P}(\boldsymbol{w}^T(\boldsymbol{x}_{p_{k,m+1}} - \boldsymbol{x}_{q_{k,m+1}}) = 0 \mid \{\boldsymbol{x}_i\}_{i \in M}) = 0$. Hence, when conditioned on $\{\boldsymbol{x}_i\}_{i \in M}$, $\boldsymbol{x}_{p_{k,m+1}} - \boldsymbol{x}_{q_{k,m+1}}$ includes a component orthogonal to the row space of $\boldsymbol{Q}_k^{(1)}[1:m]$ and therefore does not lie in this row space. In other words,

$$\mathbb{P}(\operatorname{rank}(\boldsymbol{Q}_k^{(1)}[1:m+1]) = m+1 \mid \{\boldsymbol{x}_i\}_{i \in M}) = 1.$$

This is true for any $\{\boldsymbol{x}_i\}_{i \in M}$ satisfying $\operatorname{rank}(\boldsymbol{Q}_k^{(1)}[1:m]) = m$, which by inductive assumption is true with probability 1 and so marginalizing over this event we have $\mathbb{P}(\operatorname{rank}(\boldsymbol{Q}_k^{(1)}[1:m+1]) = m+1) = 1$. Taking $m = d-1$, and noting that $\boldsymbol{Q}_k^{(1)}[1:d] = \boldsymbol{Q}_k^{(1)}$, $\mathbb{P}(\operatorname{rank}(\boldsymbol{Q}_k^{(1)}) = d) = 1$. Since this is true for all $k \in [K]$, by the union bound we have

$$\mathbb{P}\Big(\bigcup_{k \in [K]} (\operatorname{rank}(\boldsymbol{Q}_k^{(1)}) < d)\Big) \leq \sum_{k \in [K]} \mathbb{P}(\operatorname{rank}(\boldsymbol{Q}_k^{(1)}) < d) \leq \sum_{k \in [K]} 0 = 0,$$

and therefore with probability 1, $\operatorname{rank}(\boldsymbol{Q}_k^{(1)}) = d$ simultaneously for all $k \in [K]$.

Consider the following matrix:

$$\boldsymbol{Q}^{(1)} := \begin{bmatrix} \boldsymbol{Q}_1^{(1)} & \mathbf{0}_{d,d} & \cdots & \mathbf{0}_{d,d} \\ \mathbf{0}_{d,d} & \boldsymbol{Q}_2^{(1)} & \cdots & \mathbf{0}_{d,d} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{0}_{d,d} & \mathbf{0}_{d,d} & \cdots & \boldsymbol{Q}_K^{(1)} \end{bmatrix}.$$

Each consecutive block of d rows in $\boldsymbol{Q}^{(1)}$ is clearly orthogonal, and since with probability 1 each $\boldsymbol{Q}_k^{(1)}$ is simultaneously full row rank, with probability 1 we have that $\boldsymbol{Q}^{(1)}$ is full row rank (and hence is invertible since it is square). Inspecting eq. (34), we can write $\boldsymbol{\Gamma}^{(1)} = \begin{bmatrix}\boldsymbol{R}^{(1)} & \boldsymbol{Q}^{(1)}\end{bmatrix}$ where $\boldsymbol{R}^{(1)}$ is a $Kd \times D$ submatrix. Since $\boldsymbol{Q}^{(1)}$ is rank Kd, the column space of $\boldsymbol{\Gamma}^{(1)}$ is also of dimension at least Kd and therefore $\boldsymbol{\Gamma}^{(1)}$ is full row rank since it has Kd rows. In other words, we have shown that with probability 1 the rows of $\boldsymbol{\Gamma}^{(1)}$ are linearly independent. We will now show linear independence for the remaining rows in $\widehat{\boldsymbol{\Gamma}}$ (i.e., $\boldsymbol{\Gamma}^{(2)}$), which completes our proof. Specifically, we will proceed through the remaining D measurements row by row, and inductively show how each cumulative set of rows is linearly independent.

First, we define some additional notation: for any vector $\boldsymbol{w} \in \mathbb{R}^{D+Kd}$, let $\psi_k(\boldsymbol{w})$ be the subvector limited to the column indices of $\boldsymbol{\Gamma}$ involving user k, i.e.,

$$\psi_k(\boldsymbol{w}) := \begin{bmatrix} \boldsymbol{w}[1\!:\!D] \\ \boldsymbol{w}[D\!+\!(k\!-\!1)d\!+\!1\!:\!D\!+\!kd] \end{bmatrix}.$$

Let $\boldsymbol{r}_i$ denote the ith row of $\boldsymbol{\Gamma}^{(2)}$, let $k_i \in [K]$ denote the user that this row corresponds to (i.e., $\boldsymbol{r}_i$ is supported on columns $1:D$ and $D\!+\!(k_i\!-\!1)d\!+\!1\!:\!D\!+\!k_i d$), and let $j_{i,1}, j_{i,2}$ denote the first and second items selected at this measurement. We require this flexible definition of the user and items in row $\boldsymbol{r}_i$, since the permutation $\boldsymbol{P}$ has arbitrarily scrambled the users that each row in $\boldsymbol{\Gamma}^{(2)}$ corresponds to. Finally, let

$$\phi(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{x} \otimes_S \boldsymbol{x} \\ \boldsymbol{x} \end{bmatrix},$$

which is a vector in $\mathbb{R}^{D+d}$. For any nonzero vector $\boldsymbol{\mu} \in \mathbb{R}^{D+d}$, $\boldsymbol{\mu}^T \phi(\boldsymbol{x})$ is a nontrivial polynomial in $\boldsymbol{x}$. With this notation defined, for any vector $\boldsymbol{w} \in \mathbb{R}^{D+Kd}$ we see from eq. (35) that

$$\boldsymbol{w}^T \boldsymbol{r}_i = \psi_{k_i}(\boldsymbol{w})^T (\phi(\boldsymbol{x}_{j_{i,1}}) - \phi(\boldsymbol{x}_{j_{i,2}})).$$

Next, we establish a fact about the orthogonal subspace to $\operatorname{rowsp}(\boldsymbol{\Gamma}^{(1)})$. Let $E_0 \coloneqq \operatorname{span}(\{\boldsymbol{e}_i\}_{i=1}^D)$ where $\boldsymbol{e}_i$ is the ith standard basis vector in $\mathbb{R}^{D+dK}$. It is a fact that for every $\boldsymbol{w} \in \operatorname{rowsp}(\boldsymbol{\Gamma}^{(1)})^\perp$, $\operatorname{proj}_{E_0} \boldsymbol{w} \neq \boldsymbol{0}$. In other words, $\boldsymbol{w}$ has at least one nonzero element in its first D entries. Suppose this were not true: then for some nonzero $\boldsymbol{w}' \in \mathbb{R}^{Kd}$, we would have

$$\boldsymbol{w} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{w}' \end{bmatrix}.$$

Since $\boldsymbol{w}' \in \mathbb{R}^{Kd}$ and the rows of $\boldsymbol{Q}^{(1)}$ are a basis for $\mathbb{R}^{Kd}$ (since $\boldsymbol{Q}^{(1)}$ is invertible), we have $\boldsymbol{w}' = (\boldsymbol{Q}^{(1)})^T \boldsymbol{\beta}$ for some $\boldsymbol{\beta} \in \mathbb{R}^{Kd}$. Consider $\boldsymbol{r} = (\boldsymbol{\Gamma}^{(1)})^T \boldsymbol{\beta}$, which is clearly in $\operatorname{rowsp}(\boldsymbol{\Gamma}^{(1)})$. Expanding $\boldsymbol{\Gamma}^{(1)}$, we have

$$\boldsymbol{r} = \begin{bmatrix} (\boldsymbol{R}^{(1)})^T \boldsymbol{\beta} \\ (\boldsymbol{Q}^{(1)})^T \boldsymbol{\beta} \end{bmatrix} = \begin{bmatrix} (\boldsymbol{R}^{(1)})^T \boldsymbol{\beta} \\ \boldsymbol{w}' \end{bmatrix},$$

and so

$$\boldsymbol{w}^T \boldsymbol{r} = \boldsymbol{0}^T (\boldsymbol{R}^{(1)})^T \boldsymbol{\beta} + (\boldsymbol{w}')^T \boldsymbol{w}' = \|\boldsymbol{w}'\|_2^2 > 0,$$

where the last inequality follows since $\boldsymbol{w}' \neq \boldsymbol{0}$. This is a contradiction since by definition, $\boldsymbol{w}^T \boldsymbol{r} = 0$ for every $\boldsymbol{r} \in \operatorname{rowsp}(\boldsymbol{\Gamma}^{(1)})$.

Let $\boldsymbol{w}$ be a vector in $\operatorname{rowsp}(\boldsymbol{\Gamma}^{(1)})^\perp$ not equal to the zero vector, and let J denote the item indices on which each $\boldsymbol{S}_k^{(1)}$ is supported across all $k \in [K]$, i.e.,

$$J = \{j \colon j \in [n], \exists k \in [K], i \in [d] \text{ s.t. } \boldsymbol{S}_k^{(1)}[i,j] \neq 0\}.$$

Consider the first row of $\boldsymbol{\Gamma}^{(2)}$. By the incremental assumption, at least one of $j_{1,1}$ or $j_{1,2}$ is not found in J. First suppose that both are not found in J. Then

$$\mathbb{P}(\boldsymbol{w}^T \boldsymbol{r}_1 = 0 \mid \{\boldsymbol{x}_i\}_{i \in J}) = \mathbb{P}(\psi_{k_1}(\boldsymbol{w})^T (\phi(\boldsymbol{x}_{j_{1,1}}) - \phi(\boldsymbol{x}_{j_{1,2}})) = 0 \mid \{\boldsymbol{x}_i\}_{i \in J}). \tag{36}$$

As an aside, if $\boldsymbol{x}_i, \boldsymbol{x}_j$ are i.i.d. distributed according to absolutely continuous distribution p_X, then the joint distribution $p_{\boldsymbol{x}_i,\boldsymbol{x}_j}(\boldsymbol{x}_i, \boldsymbol{x}_j) = p_X(\boldsymbol{x}_i) p_X(\boldsymbol{x}_j)$ is also absolutely continuous. Also note that for any $\boldsymbol{\mu} \in \mathbb{R}^{D+d}$,

$$\boldsymbol{\mu}^T (\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)) = \begin{bmatrix} \boldsymbol{\mu}^T & -\boldsymbol{\mu}^T \end{bmatrix} \begin{bmatrix} \phi(\boldsymbol{x}_i) \\ \phi(\boldsymbol{x}_j) \end{bmatrix}. \tag{37}$$

Since $\begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{x}_j \end{bmatrix} \otimes_S \begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{x}_j \end{bmatrix}$ contains all terms in $\boldsymbol{x}_i \otimes_S \boldsymbol{x}_i$ and $\boldsymbol{x}_j \otimes_S \boldsymbol{x}_j$, $\phi([\begin{smallmatrix} \boldsymbol{x}_i \\ \boldsymbol{x}_j \end{smallmatrix}])$ contains $\phi(\boldsymbol{x}_i)$ and $\phi(\boldsymbol{x}_j)$. Therefore, we can view eq. (37) as being a polynomial in $[\begin{smallmatrix} \boldsymbol{x}_i \\ \boldsymbol{x}_j \end{smallmatrix}]$. If $\boldsymbol{\mu} \neq \boldsymbol{0}$, then $\begin{bmatrix} \boldsymbol{\mu} \\ -\boldsymbol{\mu} \end{bmatrix} \neq \boldsymbol{0}$ and so $\boldsymbol{\mu}^T (\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j))$ can be viewed as a nontrivial polynomial in $[\begin{smallmatrix} \boldsymbol{x}_i \\ \boldsymbol{x}_j \end{smallmatrix}]$. Since $p_{\boldsymbol{x}_i,\boldsymbol{x}_j}$ is absolutely continuous, for any nonzero $\boldsymbol{\mu}$ we have

$$\mathbb{P}(\boldsymbol{\mu}^T (\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)) = 0) = 0,$$

since the set of roots for a nontrivial polynomial is a set of Lebesgue measure 0.

Returning to eq. (36), we then have

$$\mathbb{P}(\boldsymbol{w}^T \boldsymbol{r}_1 = 0 \mid \{\boldsymbol{x}_i\}_{i \in J}) = \mathbb{P}(\psi_{k_1}(\boldsymbol{w})^T (\phi(\boldsymbol{x}_{j_{1,1}}) - \phi(\boldsymbol{x}_{j_{1,2}})) = 0 \mid \{\boldsymbol{x}_i\}_{i \in J}) = 0,$$

which follows from the fact that $\psi_{k_1}(\boldsymbol{w})$ is nonzero: recall from the fact presented above that $\operatorname{proj}_{E_0}(\boldsymbol{w}) \neq \boldsymbol{0}$, so $\psi_k(\boldsymbol{w}) \neq \boldsymbol{0}$ for any $k \in [K]$.

Now, instead suppose without loss of generality that $j_{1,1} \notin J$ and $j_{1,2} \in J$ (an identical argument holds for $j_{1,1} \in J$ and $j_{1,2} \notin J$). Then by similar arguments as above,

$$\mathbb{P}(\boldsymbol{w}^T \boldsymbol{r}_1 = 0 \mid \{\boldsymbol{x}_i\}_{i \in J}) = \mathbb{P}(\psi_{k_1}(\boldsymbol{w})^T (\phi(\boldsymbol{x}_{j_{1,1}}) - \phi(\boldsymbol{x}_{j_{1,2}})) = 0 \mid \{\boldsymbol{x}_i\}_{i \in J})$$

$$
\begin{aligned}
&= \mathbb{P}(\psi_{k_1}(\boldsymbol{w})^T \phi(\boldsymbol{x}_{j_{1,1}}) - \psi_{k_1}(\boldsymbol{w})^T \phi(\boldsymbol{x}_{j_{1,2}}) = 0 \mid \boldsymbol{x}_{j_{1,2}} \cup \{\boldsymbol{x}_i\}_{i \in J}) \\
&= \mathbb{P}_X(\psi_{k_1}(\boldsymbol{w})^T \phi(\boldsymbol{x}) - c = 0 \mid \boldsymbol{x}_{j_{1,2}} \cup \{\boldsymbol{x}_i\}_{i \in J}) \quad \text{where } c \text{ is a constant.} \\
&= 0.
\end{aligned}
$$

The last equality follows since $\psi_{k_1}(\boldsymbol{w}) \neq \mathbf{0}$ and so $\psi_{k_1}(\boldsymbol{w})^T \phi(\boldsymbol{x}) - c$ is a nontrivial polynomial in $\boldsymbol{x}$, along with the fact that p_X is absolutely continuous.

In either scenario, $\mathbb{P}(\boldsymbol{w}^T \boldsymbol{r}_1 = 0 \mid \{\boldsymbol{x}_i\}_{i \in J}) = 0$. Hence, when conditioned on $\{\boldsymbol{x}_i\}_{i \in J}$, $\boldsymbol{r}_1$ includes a component orthogonal to $\text{rowsp}(\boldsymbol{\Gamma}^{(1)})$ and therefore does not lie in this row space. In other words,

$$
\mathbb{P}\Big(\text{rank}\Big(\big[(\boldsymbol{\Gamma}^{(1)})^T \quad \boldsymbol{r}_1\big]\Big) = Kd + 1 \mid \{\boldsymbol{x}_i\}_{i \in J}\Big) = 1.
$$

This is true for any $\{\boldsymbol{x}_i\}_{i \in J}$ resulting in $\boldsymbol{\Gamma}^{(1)}$ being full-rank, which we know occurs with probability 1 and so by marginalizing we have $\mathbb{P}\Big(\text{rank}\Big(\big[(\boldsymbol{\Gamma}^{(1)})^T \quad \boldsymbol{r}_1\big]\Big) = Kd + 1\Big) = 1$.

If $d > 1$, let $m < D$ be given and suppose by induction that

$$
\mathbb{P}\Big(\text{rank}\Big(\big[(\boldsymbol{\Gamma}^{(1)})^T \quad \boldsymbol{r}_1 \quad \cdots \quad \boldsymbol{r}_m\big]\Big) = Kd + m\Big) = 1.
$$

Let $\boldsymbol{w}$ be a vector in $(\text{rowsp}(\boldsymbol{\Gamma}^{(1)}) \cup \text{span}(\{\boldsymbol{r}_i\}_{i=1}^m))^\perp$ not equal to the zero vector. Note that $\boldsymbol{w} \in \text{rowsp}(\boldsymbol{\Gamma}^{(1)})^\perp$ as well, and so by the above, $\text{proj}_{E_0}(\boldsymbol{w}) \neq \mathbf{0}$ and so $\psi_k(\boldsymbol{w}) \neq \mathbf{0}$ for any $k \in [K]$. Reusing notation, let

$$
J = \{j \colon j \in [n], \exists k \in [K], i \in [d] \text{ s.t. } \boldsymbol{S}_k^{(1)}[i,j] \neq 0\} \cup \{j \colon j \in [n], \exists i \in [m] \; j = j_{i,1} \vee j = j_{i,2}\}
$$

denote the set of all item indices in $\widehat{\boldsymbol{\Gamma}}$ measured up through and including the mth measurement of $\boldsymbol{\Gamma}^{(2)}$. Consider row $m+1$ of $\boldsymbol{\Gamma}^{(2)}$. By the incremental assumption, at least one of $j_{m+1,1}$ or $j_{m+1,2}$ is not found in J. First suppose that both are not found in J. Then

$$
\begin{aligned}
\mathbb{P}(\boldsymbol{w}^T \boldsymbol{r}_{m+1} = 0 \mid \{\boldsymbol{x}_i\}_{i \in J}) &= \mathbb{P}(\psi_{k_{m+1}}(\boldsymbol{w})^T (\phi(\boldsymbol{x}_{j_{m+1,1}}) - \phi(\boldsymbol{x}_{j_{m+1,2}})) = 0 \mid \{\boldsymbol{x}_i\}_{i \in J}) \\
&= 0,
\end{aligned}
$$

due to a similar argument as above.

Now, instead suppose without loss of generality that $j_{m+1,1} \notin J$ and $j_{m+1,2} \in J$ (an identical argument holds for $j_{m+1,1} \in J$ and $j_{m+1,2} \notin J$). Then by similar arguments as above,

$$
\begin{aligned}
&\mathbb{P}(\boldsymbol{w}^T \boldsymbol{r}_{m+1} = 0 \mid \{\boldsymbol{x}_i\}_{i \in J}) = \mathbb{P}(\psi_{k_{m+1}}(\boldsymbol{w})^T (\phi(\boldsymbol{x}_{j_{m+1,1}}) - \phi(\boldsymbol{x}_{j_{m+1,2}})) = 0 \mid \{\boldsymbol{x}_i\}_{i \in J}) \\
&= \mathbb{P}(\psi_{k_{m+1}}(\boldsymbol{w})^T \phi(\boldsymbol{x}_{j_{m+1,1}}) - \psi_{k_{m+1}}(\boldsymbol{w})^T \phi(\boldsymbol{x}_{j_{m+1,2}}) = 0 \mid \boldsymbol{x}_{j_{m+1,2}} \cup \{\boldsymbol{x}_i\}_{i \in J}) \\
&= \mathbb{P}_X(\psi_{k_{m+1}}(\boldsymbol{w})^T \phi(\boldsymbol{x}) - c = 0 \mid \boldsymbol{x}_{j_{m+1,2}} \cup \{\boldsymbol{x}_i\}_{i \in J}) \quad \text{where } c \text{ is a constant.} \\
&= 0.
\end{aligned}
$$

The last equality follows since $\psi_{k_{m+1}}(\boldsymbol{w}) \neq \mathbf{0}$ and so $\psi_{k_{m+1}}(\boldsymbol{w})^T \phi(\boldsymbol{x}) - c$ is a nontrivial polynomial in $\boldsymbol{x}$.

In either scenario, $\mathbb{P}(\boldsymbol{w}^T \boldsymbol{r}_{m+1} = 0 \mid \{\boldsymbol{x}_i\}_{i \in J}) = 0$. Hence, when conditioned on $\{\boldsymbol{x}_i\}_{i \in J}$, $\boldsymbol{r}_{m+1}$ includes a component orthogonal to $\text{rowsp}(\boldsymbol{\Gamma}^{(1)}) \cup \text{span}(\{\boldsymbol{r}_i\}_{i=1}^m)$ and therefore does not lie in the span of the previous rows. In other words,

$$
\mathbb{P}\Big(\text{rank}\Big(\big[(\boldsymbol{\Gamma}^{(1)})^T \quad \boldsymbol{r}_1 \quad \cdots \quad \boldsymbol{r}_{m+1}\big]\Big) = Kd + m + 1 \mid \{\boldsymbol{x}_i\}_{i \in J}\Big) = 1.
$$

This is true for any $\{\boldsymbol{x}_i\}_{i \in J}$ satisfying $\text{rank}\Big(\big[(\boldsymbol{\Gamma}^{(1)})^T \quad \boldsymbol{r}_1 \quad \cdots \quad \boldsymbol{r}_m\big]\Big) = Kd + m$, which by inductive assumption occurs with probability 1 and so by marginalizing we have $\mathbb{P}\Big(\text{rank}\Big(\big[(\boldsymbol{\Gamma}^{(1)})^T \quad \boldsymbol{r}_1 \quad \cdots \quad \boldsymbol{r}_{m+1}\big]\Big) = Kd + m + 1\Big) = 1$. Taking $m = D - 1$, we have with probability 1 that

$$
\big[(\boldsymbol{\Gamma}^{(1)})^T \quad \boldsymbol{r}_1 \quad \cdots \quad \boldsymbol{r}_D\big] = \big[(\boldsymbol{\Gamma}^{(1)})^T \quad (\boldsymbol{\Gamma}^{(2)})^T\big] = \widehat{\boldsymbol{\Gamma}}^T
$$

is full-rank, and so $\boldsymbol{\Gamma}$ has full column rank.

C.5 Proof of results in the single user case

Necessary conditions: When $K = 1$, the three conditions in Proposition 2.1 are equivalent to $\operatorname{rank}(\boldsymbol{S}) \geq D + d$: in the single user case, $\boldsymbol{S}_T = \boldsymbol{S}$, and so (c) directly states that $\operatorname{rank}(\boldsymbol{S}) \geq D + d$. (b) also translates to $\operatorname{rank}(\boldsymbol{S}) \geq D + d$ when $K = 1$. The condition $\operatorname{rank}(\boldsymbol{S}) \geq d$ in (a) is subsumed by $\operatorname{rank}(\boldsymbol{S}) \geq D + d$.

Sufficient conditions: Suppose $\operatorname{rank}(\boldsymbol{S}) \geq D + d$. By definition, there exists a set of $D + d$ linearly independent rows in $\boldsymbol{S}$: denote the $D + d \times n$ submatrix of $\boldsymbol{S}$ defined by these rows as $\boldsymbol{S}'$. Since $\boldsymbol{S}'$ is full row rank by construction, by Corollary C.4.1 there exists a permutation $\boldsymbol{P}$ such that $\boldsymbol{PS}'$ is incremental. Define $\boldsymbol{S}^{(1)}$ as the first d rows of $\boldsymbol{PS}'$, and $\boldsymbol{S}^{(2)}$ as the remaining D rows of $\boldsymbol{PS}'$. Then $\left[\begin{smallmatrix} \boldsymbol{S}^{(1)} \\ \boldsymbol{S}^{(2)} \end{smallmatrix}\right]$ satisfies the conditions of Proposition 2.2 and therefore if each $\boldsymbol{x}_i$ is sampled i.i.d. according to p_X then $\boldsymbol{PS}'[\,\boldsymbol{x}_\otimes^T\ \boldsymbol{x}^T\,]$ has full column rank with probability 1 — and therefore full row rank since it is square. Since the rows of this matrix are simply a permuted subset of the rows in $\boldsymbol{S}[\,\boldsymbol{x}_\otimes^T\ \boldsymbol{x}^T\,]$, we also have that $\boldsymbol{S}[\,\boldsymbol{x}_\otimes^T\ \boldsymbol{x}^T\,]$ has rank $D + d$ and therefore full column rank with probability 1.

Random construction: We will use the results of Appendix C.2 to choose a number of random measurements and items such that a single-user selection matrix $\boldsymbol{S}$ has rank at least $D + d$ with high probability, to satisfy the conditions described above. Let failure probability $0 < \delta < 1$ be given, and suppose $\boldsymbol{S}$ is constructed by drawing m_T item index pairs uniformly and independent at random among n items. After drawing a single measurement, $\boldsymbol{S}$ will immediately have rank 1. According to Theorem C.8 with $r_0 = 1$ and $r = D + d$, with probability at least $1 - \delta$ the total number of additional required measurements M is less than $\left(1 + \ln \frac{1}{\delta}\right)\left(\sum_{i=2}^{D+d} \frac{1}{1 - \frac{i(i-1)}{n(n-1)}}\right)$.

To make this quantity more manageable, note that $\frac{1}{1 - \frac{i(i-1)}{n(n-1)}}$ is an increasing function of i. Hence, if we choose a constant U such that $\frac{1}{1 - \frac{(D+d)(D+d-1)}{n(n-1)}} \leq U$, then for every $2 \leq i \leq D + d$ we also have $\frac{1}{1 - \frac{i(i-1)}{n(n-1)}} \leq U$. To arrive at such a U, suppose that $(n-1)^2 \geq (1+\gamma)(D+d)^2$ for some $\gamma > 0$, i.e., $n \geq \sqrt{1+\gamma}(D+d) + 1$. Then

$$\frac{(D+d)(D+d-1)}{n(n-1)} \leq \frac{(D+d)^2}{(n-1)^2} \leq \frac{1}{1+\gamma},$$

and so

$$\frac{1}{1 - \frac{(D+d)(D+d-1)}{n(n-1)}} \leq \frac{1}{1 - \frac{1}{1+\gamma}} = \frac{1+\gamma}{\gamma},$$

and we can set $U = \frac{1+\gamma}{\gamma}$. Therefore,

$$\sum_{i=2}^{D+d} \frac{1}{1 - \frac{i(i-1)}{n(n-1)}} \leq \frac{1+\gamma}{\gamma}(D+d-1),$$

and so with probability at least $1 - \delta$, $M < \left(1 + \ln \frac{1}{\delta}\right)\frac{1+\gamma}{\gamma}(D+d-1)$. To choose a convenient value for γ, we can let $\gamma = \frac{1}{2}(1+\sqrt{5})$, in which case $\frac{\gamma+1}{\gamma} = \sqrt{1+\gamma} = \frac{1}{2}(1+\sqrt{5}) \approx 1.62$. So, if $n \geq \frac{1}{2}(1+\sqrt{5})(D+d) + 1$, and $m_T \geq \left\lceil \frac{1}{2}(1+\sqrt{5})\left(1 + \ln \frac{1}{\delta}\right)(D+d-1) \right\rceil + 1$ random measurements are taken, then with probability at least $1 - \delta$, $\operatorname{rank}(\boldsymbol{S}) \geq D + d$. Hence, with high probability, $n = \Omega(D+d)$ and $m_T = \Omega(D+d)$ random measurements result in a selection matrix $\boldsymbol{S}$ with rank at least $D + d$. Once such a matrix with rank at least $D + d$ is fixed after sampling, if each $\boldsymbol{x}_i$ is sampled i.i.d. according to p_X then as described above $\boldsymbol{\Gamma}$ will be full column rank with probability 1. Together, the process of independently sampling $\boldsymbol{S}$ and $\{\boldsymbol{x}_i\}_{i=1}^n$ results in $\boldsymbol{\Gamma}$ having full column rank with high probability.

C.6 Constructions and counterexamples

Counterexample for necessary conditions being sufficient: Below we demonstrate a counterexample where the conditions in Proposition 2.1 are met, but the system results in a $\boldsymbol{\Gamma}$ matrix that is not full column rank. In this example, $d = 2$ (and so $D = 3$), $K = 3$, $m_k = d + D/K = 3$, and $n = D + d + 1 = 6$. Consider the selection matrices below:

$$\boldsymbol{S}_1 = \begin{bmatrix} 0 & 0 & 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 \\ 1 & 0 & 0 & -1 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{S}_2 = \begin{bmatrix} 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}$$

$$\boldsymbol{S}_3 = \begin{bmatrix} 0 & 1 & 0 & 0 & -1 & 0 \\ 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix}.$$

By inspection, $\operatorname{rank}(\boldsymbol{S}_k) = 3 \geq d$ for each k, $\sum_k \operatorname{rank}(\boldsymbol{S}_k) = 9 = D + dK$, and $\operatorname{rank}(\boldsymbol{S}_T) = 5 = D + d$. Yet, when we numerically sample $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and verify that $\operatorname{rank}(\boldsymbol{S}_k \boldsymbol{X}^T) = 2 = d$, $\sum_k \operatorname{rank}(\boldsymbol{S}_k \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix}) = D + dK$, and $\operatorname{rank}(\boldsymbol{S}_T \begin{bmatrix} \boldsymbol{X}_\otimes^T & \boldsymbol{X}^T \end{bmatrix}) = 5$, we still find that $\boldsymbol{\Gamma}$ is rank deficient. This counterexample illustrates that the conditions of Proposition 2.1 are *not sufficient* for identifiability.

Incremental condition construction: Here we construct a selection matrix scheme that satisfies the properties of Proposition 2.2 while only using the minimal number of measurements per user (i.e., $m_k = d + D/K$) and items (i.e., $n = D + d + 1$). For each $k \in [K]$, let

$$\boldsymbol{S}_k^{(1)} = d \text{ rows} \left\{ \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 & 0 & \\ 0 & 1 & -1 & \cdots & 0 & 0 & 0 & \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 & \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \boldsymbol{0}_{d,D} \\ 0 & 0 & 0 & \cdots & -1 & 0 & 0 & \\ 0 & 0 & 0 & \cdots & 1 & -1 & 0 & \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 & \end{bmatrix} \right. ,$$

and define

$$\boldsymbol{S}^{(2)} = D \text{ rows} \left\{ \begin{bmatrix} & 1 & -1 & 0 & \cdots & 0 & 0 & 0 \\ & 0 & 1 & -1 & \cdots & 0 & 0 & 0 \\ & 0 & 0 & 1 & \cdots & 0 & 0 & 0 \\ \boldsymbol{0}_{D,d} & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & 0 & 0 & 0 & \cdots & -1 & 0 & 0 \\ & 0 & 0 & 0 & \cdots & 1 & -1 & 0 \\ & 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{bmatrix} \right. .$$

By observation, for each $k \in [K]$ we have

$$\begin{bmatrix} \boldsymbol{S}_k^{(1)} \\ \boldsymbol{S}^{(2)} \end{bmatrix} = D + d \text{ rows} \left\{ \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -1 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 0 & 0 \\ 0 & 0 & 0 & \cdots & 1 & -1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 & -1 \end{bmatrix} \right. ,$$

which by observation is incremental. Assuming for simplicity that ${}^{D}\!/\!{}_{K}$ is an integer, for each $k \in [K]$ let $\boldsymbol{S}_k^{(2)}$ be the submatrix defined by rows $(k-1)({}^{D}\!/\!{}_{K}) + 1$ through $k({}^{D}\!/\!{}_{K})$ of $\boldsymbol{S}^{(2)}$, i.e., each user is allotted ${}^{D}\!/\!{}_{K}$ nonoverlapping rows of $\boldsymbol{S}^{(2)}$. Finally, for each $k \in [K]$ let

$$\boldsymbol{S}_k = \begin{bmatrix} \boldsymbol{S}_k^{(1)} \\ \boldsymbol{S}_k^{(2)} \end{bmatrix}.$$

By observation each $\boldsymbol{S}_k^{(1)}$ has rank d, and by construction each $\left[\begin{smallmatrix} \boldsymbol{S}_k^{(1)} \\ \boldsymbol{S}^{(2)} \end{smallmatrix}\right]$ is incremental, and therefore the conditions of Proposition 2.2 are satisfied.

Counterexample for incremental sufficiency conditions being exhaustive: Below we demonstrate a counterexample where the matrix $\mathbf{\Gamma}$ is full column rank, yet the conditions in Proposition 2.2 are not met, demonstrating that they are not an exhaustive set of sufficiency conditions.

In this example, $d = 2$, $K = 2$, and $n = 6$, with selection matrices given by

$$\boldsymbol{S}_1 = \begin{array}{c} (1a) \\ (1b) \\ (1c) \\ (1d) \end{array} \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \end{bmatrix}$$

$$\boldsymbol{S}_2 = \begin{array}{c} (2a) \\ (2b) \\ (2c) \end{array} \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{bmatrix}.$$

By observation, one cannot partition these selection matrices according to the conditions in Proposition 2.2. To see this, we can attempt to partition these selection matrices according to these conditions. First note that rows (1a) and (2a) are equal, as well as rows (1c) and (2b), which implies that (1a,c) and (2a,b) must belong to $\boldsymbol{S}_1^{(1)}$ and $\boldsymbol{S}_2^{(1)}$ respectively. Otherwise, there would exist at least one repeated pair in the matrix $\left[\begin{smallmatrix} \boldsymbol{S}_k^{(1)} \\ \boldsymbol{S}^{(2)} \end{smallmatrix}\right]$ for $k = 1$ or $k = 2$, which would violate condition (b) of Proposition 2.2.

Therefore, $\boldsymbol{S}^{(2)}$ must consist of rows (1b), (1d), and (2c). While (1d) is surely incremental with respect to $\boldsymbol{S}_1^{(1)}$, and (2c) is surely incremental with respect to $\boldsymbol{S}_2^{(1)}$, (1b) overlaps with both (1a) and (1c) and therefore cannot possibly be incremental with respect to $\boldsymbol{S}_1^{(1)}$, and hence the conditions in Proposition 2.2 are not met.

Yet, in simulation we find with normally distributed items that the $\mathbf{\Gamma}$ matrix resulting from the above selection scheme is in fact full column rank.

C.7 Conjectured sufficiency conditions

We conjecture that a set of conditions similar to that of Proposition 2.2 are sufficient for identifiability under items sampled according to a distribution that is absolutely continuous with respect to the Lebesgue measure. We list these conditions below:

Conjectured sufficiency conditions: Let $K \geq 1$, and suppose $m_k > d \, \forall \, k \in [K]$, $m_T = D + dK$, and $n \geq D + d + 1$. Suppose that for each $k \in [K]$, there exists a $d \times n$ selection matrix $\boldsymbol{S}_k^{(1)}$ and $m_k - d \times n$ selection matrix $\boldsymbol{S}_k^{(2)}$ such that $\boldsymbol{S}_k = \left[\, (\boldsymbol{S}_k^{(1)})^T \; (\boldsymbol{S}_k^{(2)})^T \,\right]^T$, and that the following are true:

(a) For all $k \in [K]$, $\operatorname{rank}(\boldsymbol{S}_k^{(1)}) = d$

(b) Defining the $D \times n$ selection matrix $\boldsymbol{S}^{(2)}$ as $\boldsymbol{S}^{(2)} := \left[\, (\boldsymbol{S}_1^{(2)})^T \cdots (\boldsymbol{S}_K^{(2)})^T \,\right]^T$, for each $k \in [K]$, $\left[\begin{smallmatrix} \boldsymbol{S}_k^{(1)} \\ \boldsymbol{S}^{(2)} \end{smallmatrix}\right]$ is full row rank

Intuitively, these conditions replace the permutation condition in Proposition 2.2 with a more general condition concerning only the rank. In fact, due to Corollary C.4.1, condition (b) in Proposition 2.2 implies the second condition above. These conditions capture the intuition that each user is allocated d independent measurements to identity their own pseudo-ideal point (condition (a) above), and then collectively the set of users answers an additional D independent measurements to identify the metric (condition (b) above). As long as the individual measurements do not "overlap" with the collective measurements (captured by the rank condition in condition (b)), then the collective set of measurements should be rich enough to identify the metric and all pseudo-ideal points, even if each individual user has overlapping measurements in their $\boldsymbol{S}_k^{(1)}$ selection matrices. Empirically, we find that the above conditions appear to be sufficient for identifiability, at least with normally distributed items.

Furthermore, the above conditions would provide a convenient avenue to study randomly selected unquantized measurements among multiple users. As we demonstrated in Appendix C.5, we can use Theorem C.8 to bound the number of measurements needed for a selection matrix to be full-rank. We can apply these tools to the multiuser case as follows: first, sample on the order $\Omega(D)$ randomly selected pairs among $\Omega(D+d)$ items to construct a selection matrix $\boldsymbol{S}^{(2)}$ that is rank D with high probability. Then, using $\boldsymbol{S}^{(2)}$ as a seed matrix in Theorem C.8, sample on the order of d additional measurements per user in selection matrix $\boldsymbol{S}_k^{(1)}$, so that for each individual user conditions (a-b) above are satisfied with high probability. The key insight is that if the measurements in $\boldsymbol{S}^{(2)}$ are evenly distributed between users, and the number of samples taken in $\boldsymbol{S}^{(2)}$ and each $\boldsymbol{S}_k^{(1)}$ is fixed ahead of time (and non-adaptive), then the above sampling process is simply equivalent to sampling pairs uniformly at random for each individual user. Yet, with high probability they should satisfy the above conditions, which if are sufficient for identifiability should result in a measurement matrix $\boldsymbol{\Gamma}$ that is full column rank.

D Proofs of prediction and generalization results

D.1 Proof of Theorem 3.1

We start by expanding the excess risk between the empirical and true optimizers:

$$
\begin{aligned}
&R(\widehat{\boldsymbol{M}}, \{\widehat{\boldsymbol{v}}_k\}_{k=1}^K) - R(\boldsymbol{M}^*, \{\boldsymbol{v}_k^*\}_{k=1}^K) \\
&= R(\widehat{\boldsymbol{M}}, \{\widehat{\boldsymbol{v}}_k\}_{k=1}^K) - \widehat{R}(\widehat{\boldsymbol{M}}, \{\widehat{\boldsymbol{v}}_k\}_{k=1}^K) + \widehat{R}(\widehat{\boldsymbol{M}}, \{\widehat{\boldsymbol{v}}_k\}_{k=1}^K) \\
&\qquad - \widehat{R}(\boldsymbol{M}^*, \{\boldsymbol{v}_k^*\}_{k=1}^K) + \widehat{R}(\boldsymbol{M}^*, \{\boldsymbol{v}_k^*\}_{k=1}^K) - R(\boldsymbol{M}^*, \{\boldsymbol{v}_k^*\}_{k=1}^K) \\
&\leq R(\widehat{\boldsymbol{M}}, \{\widehat{\boldsymbol{v}}_k\}_{k=1}^K) - \widehat{R}(\widehat{\boldsymbol{M}}, \{\widehat{\boldsymbol{v}}_k\}_{k=1}^K) + \widehat{R}(\boldsymbol{M}^*, \{\boldsymbol{v}_k^*\}_{k=1}^K) - R(\boldsymbol{M}^*, \{\boldsymbol{v}_k^*\}_{k=1}^K) \qquad (38) \\
&\leq 2 \sup_{\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K} |\widehat{R}(\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K) - R(\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K)| \\
&\leq 2\,\mathbb{E}\left[\sup_{\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K} \left|\widehat{R}(\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K) - R(\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K)\right|\right] + \sqrt{\frac{8L^2\gamma^2 \log(2/\delta)}{|\mathcal{S}|}} \qquad (39)
\end{aligned}
$$

where (38) follows from the fact that $\widehat{\boldsymbol{M}}, \{\boldsymbol{v}_k\}$ are the empirical risk minimizers, and (39) follows from the Bounded differences inequality (also known as McDiarmid's Inequality, see [48]) since for two data points (p, k, y_p) and $(p', k', y_{p'})$ we have that

$$
\begin{aligned}
&\ell\left(y_p(\boldsymbol{x}_i^T \boldsymbol{M} \boldsymbol{x}_i - \boldsymbol{x}_j^T \boldsymbol{M} \boldsymbol{x}_j + (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{v}_k)\right) \\
&\qquad - \ell\left(y_{p'}(\boldsymbol{x}_{i'}^T \boldsymbol{M} \boldsymbol{x}_{i'} - \boldsymbol{x}_{j'}^T \boldsymbol{M} \boldsymbol{x}_{j'} + (\boldsymbol{x}_{i'} - \boldsymbol{x}_{j'})^T \boldsymbol{v}_{k'})\right) \leq 2L\gamma
\end{aligned}
$$

by Lipschitz-ness of ℓ and the definition of γ in eq. (5). The expectation in eq. (39) is with respect to the dataset $\mathcal{S}$. Next, using symmetrization, contraction, and introducing Rademacher random variables ε_p with $\mathbb{P}(\varepsilon_p = 1) = \mathbb{P}(\varepsilon_p = -1) = 1/2$ for all data points in $\mathcal{S}$, we have that (with expectations taken with respect to both $\{\varepsilon_p\}$ and $\mathcal{S}$)

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K} |\widehat{R}(\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K) - R(\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K)|\right] \\
&\leq \frac{2L}{|\mathcal{S}|}\,\mathbb{E}\left[\sup_{\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K} \left|\sum_{\mathcal{S}} \varepsilon_p y_p (\boldsymbol{x}_i^T \boldsymbol{M} \boldsymbol{x}_i - \boldsymbol{x}_j^T \boldsymbol{M} \boldsymbol{x}_j + (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{v}_k)\right|\right] \\
&= \frac{2L}{|\mathcal{S}|}\,\mathbb{E}\left[\sup_{\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K} \left|\sum_{\mathcal{S}} \varepsilon_p (\boldsymbol{x}_i^T \boldsymbol{M} \boldsymbol{x}_i - \boldsymbol{x}_j^T \boldsymbol{M} \boldsymbol{x}_j + (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \boldsymbol{v}_k)\right|\right] \quad \text{since } \mathbb{P}(\varepsilon_p y_p = 1) = 1/2 \\
&= \frac{2L}{|\mathcal{S}|}\,\mathbb{E}\left[\sup_{\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K} \left|\left\langle \sum_{\mathcal{S}} \varepsilon_p \begin{bmatrix} \boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{x}_j\boldsymbol{x}_j^T & \boldsymbol{0} & \cdots & \underbrace{\boldsymbol{x}_i - \boldsymbol{x}_j}_{\text{column } d+k} & \cdots & \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{M} & \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_K \end{bmatrix} \right\rangle\right|\right] \\
&\qquad (40) \\
&\overset{\text{Cauchy-Schwarz}}{\leq} \frac{2L}{|\mathcal{S}|}\,\mathbb{E}\left[\sup_{\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K} \left\|\sum_{\mathcal{S}} \varepsilon_p \begin{bmatrix} \boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{x}_j\boldsymbol{x}_j^T & \boldsymbol{0} & \cdots & \underbrace{\boldsymbol{x}_i - \boldsymbol{x}_j}_{\text{column } d+k} & \cdots & \boldsymbol{0} \end{bmatrix}\right\|_F \left\|\begin{bmatrix} \boldsymbol{M} & \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_K \end{bmatrix}\right\|_F\right] \\
&= \frac{2L}{|\mathcal{S}|}\left(\sup_{\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K} \left\|\begin{bmatrix} \boldsymbol{M} & \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_K \end{bmatrix}\right\|_F\right)\mathbb{E}\left[\left\|\sum_{\mathcal{S}} \varepsilon_p \begin{bmatrix} \operatorname{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{x}_j\boldsymbol{x}_j^T) \\ \boldsymbol{x}_i - \boldsymbol{x}_j \end{bmatrix}\right\|_2\right]
\end{aligned}
$$

$$\le \frac{2L}{|\mathcal{S}|}\sqrt{\lambda_F^2+K\lambda_v^2}\,\mathbb{E}\left[\left\|\sum_{\mathcal{S}}\varepsilon_p\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}\right\|_2\right]. \tag{41}$$

Next we employ Matrix Bernstein to bound

$$\mathbb{E}\left[\left\|\sum_{\mathcal{S}}\varepsilon_p\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}\right\|_2\right],$$

which is a sum of zero-mean random vectors in $\mathbb{R}^{d^2+d}$ (recall that each $\varepsilon_p \in \{-1,1\}$ with equal probability). First note that under the assumption $\|\boldsymbol{x}_i\| \le B\;\forall\, i$, we have

$$\begin{aligned}\left\|\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}\right\|_2^2 &= \|\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T\|_F^2+\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2\\ &\le (\|\boldsymbol{x}_i\boldsymbol{x}_i^T\|_F+\|\boldsymbol{x}_j\boldsymbol{x}_j^T\|_F)^2+(\|\boldsymbol{x}_i\|_2+\|\boldsymbol{x}_j\|_2)^2\\ &\le (\|\boldsymbol{x}_i\|_2^2+\|\boldsymbol{x}_j\|_2^2)^2+(\|\boldsymbol{x}_i\|_2+\|\boldsymbol{x}_j\|_2)^2\\ &\le 4(B^4+B^2),\end{aligned}$$

and therefore

$$\left\|\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}\right\|_2 \le 2B\sqrt{B^2+1} =: C_B.$$

We also have

$$\begin{aligned}&\left\|\mathbb{E}\left[\sum_{\mathcal{S}}\left(\varepsilon_p\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}\right)^T\left(\varepsilon_p\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}\right)\right]\right\|\\ &=\mathbb{E}\left[\sum_{\mathcal{S}}\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}^T\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}\right]\\ &=\mathbb{E}\left[\sum_{\mathcal{S}}(\|\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T\|_F^2+\|\boldsymbol{x}_i-\boldsymbol{x}_j\|_2^2)\right]\\ &\le 4(B^4+B^2)|\mathcal{S}|\\ &=C_B^2|\mathcal{S}|,\end{aligned}$$

and

$$\begin{aligned}&\left\|\mathbb{E}\left[\sum_{\mathcal{S}}\left(\varepsilon_p\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}\right)\left(\varepsilon_p\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}\right)^T\right]\right\|\\ &=\left\|\mathbb{E}\left[\sum_{\mathcal{S}}\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}^T\right]\right\|\\ &\le\mathbb{E}\left[\left\|\sum_{\mathcal{S}}\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}^T\right\|\right] \quad \text{by convexity of } \|\cdot\| \text{ and Jensen's inequality}\\ &\le\mathbb{E}\left[\sum_{\mathcal{S}}\left\|\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}^T\right\|\right] \quad \text{By the triangle inequality}\\ &\le\mathbb{E}\left[\sum_{\mathcal{S}}\left\|\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}\right\|_2^2\right]\\ &\le C_B^2|\mathcal{S}|,\end{aligned}$$

and so

$$\max\left\{\left\|\mathbb{E}\left[\sum_{\mathcal{S}}\left(\varepsilon_p\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}\right)^T\left(\varepsilon_p\begin{bmatrix}\mathrm{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T-\boldsymbol{x}_j\boldsymbol{x}_j^T)\\ \boldsymbol{x}_i-\boldsymbol{x}_j\end{bmatrix}\right)\right]\right\|,\right.$$

$$\left. \left\| \mathbb{E}\left[\sum_{\mathcal{S}} \left(\varepsilon_p \begin{bmatrix} \operatorname{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{x}_j\boldsymbol{x}_j^T) \\ \boldsymbol{x}_i - \boldsymbol{x}_j \end{bmatrix}\right) \left(\varepsilon_p \begin{bmatrix} \operatorname{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{x}_j\boldsymbol{x}_j^T) \\ \boldsymbol{x}_i - \boldsymbol{x}_j \end{bmatrix}\right)^T \right] \right\| \right\} \leq C_B^2 |\mathcal{S}|.$$

Therefore, by Theorem 6.1.1 of [49]

$$\mathbb{E}\left[\left\| \sum_{\mathcal{S}} \varepsilon_p \begin{bmatrix} \operatorname{vec}(\boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{x}_j\boldsymbol{x}_j^T) \\ \boldsymbol{x}_i - \boldsymbol{x}_j \end{bmatrix} \right\|_2 \right] \leq \sqrt{2C_B^2 |\mathcal{S}| \log(d^2 + d + 1)} + \frac{C_B}{3} \log(d^2 + d + 1).$$

Plugging into eq. (41) and then eq. (39), we have with probability greater than $1 - \delta$,

$$\begin{aligned} & R(\widehat{\boldsymbol{M}}, \{\widehat{\boldsymbol{v}}_k\}_{k=1}^K) - R(\boldsymbol{M}^*, \{\boldsymbol{v}_k^*\}_{k=1}^K) \\ & \leq 4L\sqrt{\lambda_F^2 + K\lambda_v^2} \left(\sqrt{\frac{2C_B^2}{|\mathcal{S}|} \log(d^2 + d + 1)} + \frac{C_B}{3|\mathcal{S}|} \log(d^2 + d + 1) \right) + \sqrt{\frac{8L^2\gamma^2 \log(2/\delta)}{|\mathcal{S}|}}. \end{aligned}$$

Taking $B = 1$, we have the desired result.

D.2 Proof of Theorem 3.3

The proof is identical to that of Theorem 3.1 up until eq. (40), where instead of applying Cauchy-Schwarz we apply the matrix Hölder's inequality:

$$\begin{aligned} & \frac{2L}{|\mathcal{S}|} \mathbb{E}\left[\sup_{\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K} \left| \left\langle \sum_{\mathcal{S}} \varepsilon_p \begin{bmatrix} \boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{x}_j\boldsymbol{x}_j^T & \boldsymbol{0} & \cdots & \underbrace{\boldsymbol{x}_i - \boldsymbol{x}_j}_{\text{column } d+k} & \cdots & \boldsymbol{0} \end{bmatrix}, [\boldsymbol{M} \quad \boldsymbol{v}_1 \quad \cdots \quad \boldsymbol{v}_K] \right\rangle \right| \right] \\ & \quad \leq \frac{2L}{|\mathcal{S}|} \mathbb{E}\left[\sup_{\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K} \left\| \sum_{\mathcal{S}} \varepsilon_p \begin{bmatrix} \boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{x}_j\boldsymbol{x}_j^T & \boldsymbol{0} & \cdots & \underbrace{\boldsymbol{x}_i - \boldsymbol{x}_j}_{\text{column } d+k} & \cdots & \boldsymbol{0} \end{bmatrix} \right\| \left\| [\boldsymbol{M} \quad \boldsymbol{v}_1 \quad \cdots \quad \boldsymbol{v}_K] \right\|_* \right] \\ & \quad = \frac{2L}{|\mathcal{S}|} \left(\sup_{\boldsymbol{M}, \{\boldsymbol{v}_k\}_{k=1}^K} \left\| [\boldsymbol{M} \quad \boldsymbol{v}_1 \quad \cdots \quad \boldsymbol{v}_K] \right\|_* \right) \mathbb{E}\left[\left\| \sum_{\mathcal{S}} \varepsilon_p \begin{bmatrix} \boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{x}_j\boldsymbol{x}_j^T & \boldsymbol{0} & \cdots & \underbrace{\boldsymbol{x}_i - \boldsymbol{x}_j}_{\text{column } d+k} & \cdots & \boldsymbol{0} \end{bmatrix} \right\| \right] \\ & \quad \leq \frac{2L\lambda_*}{|\mathcal{S}|} \mathbb{E}\left[\left\| \sum_{p \in \mathcal{S}} \varepsilon_p \begin{bmatrix} \boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{x}_j\boldsymbol{x}_j^T & \boldsymbol{0} & \cdots & \underbrace{\boldsymbol{x}_i - \boldsymbol{x}_j}_{\text{column } d+k} & \cdots & \boldsymbol{0} \end{bmatrix} \right\| \right]. \end{aligned}$$

In a similar manner to Appendix D.1, we can apply Matrix Bernstein to bound $\mathbb{E}\left[\left\|\sum_{\mathcal{S}} \varepsilon_p \boldsymbol{Z}_{ij}^{(k)}\right\|\right]$, where for conciseness we have defined

$$\boldsymbol{Z}_{ij}^{(k)} := \begin{bmatrix} \boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{x}_j\boldsymbol{x}_j^T & \boldsymbol{0} & \cdots & \underbrace{\boldsymbol{x}_i - \boldsymbol{x}_j}_{\text{column } d+k} & \cdots & \boldsymbol{0} \end{bmatrix}.$$

First note that

$$\begin{aligned} \|\varepsilon_p \boldsymbol{Z}_{ij}^{(k)}\| & \leq \|\boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{x}_j\boldsymbol{x}_j^T\| + \|\boldsymbol{x}_i - \boldsymbol{x}_j\| \\ & \leq \|\boldsymbol{x}_i\boldsymbol{x}_i^T\| + \|\boldsymbol{x}_j\boldsymbol{x}_j^T\| + \|\boldsymbol{x}_i\| + \|\boldsymbol{x}_j\| \\ & = \|\boldsymbol{x}_i\|_2^2 + \|\boldsymbol{x}_j\|_2^2 + \|\boldsymbol{x}_i\|_2 + \|\boldsymbol{x}_j\|_2 \\ & \leq 2(B^2 + B), \end{aligned}$$

where we have used the fact that the operator norm of a vector is simply the ℓ_2 norm, along with the assumption that $\|\boldsymbol{x}_i\|_2 \leq B \; \forall i$ for a constant $B > 0$ (taken to be 1 in the theorem statement). We next bound the matrix variance of the sum $\sum_{\mathcal{S}} \varepsilon_p \boldsymbol{Z}_{ij}^{(k)}$, defined as $v := \max\{\|\sum_{\mathcal{S}} \mathbb{E}[\varepsilon_p \boldsymbol{Z}_{ij}^{(k)} (\varepsilon_p \boldsymbol{Z}_{ij}^{(k)})^T]\|, \|\sum_{\mathcal{S}} \mathbb{E}[(\varepsilon_p \boldsymbol{Z}_{ij}^{(k)})^T \varepsilon_p \boldsymbol{Z}_{ij}^{(k)}]\|\}$, which is equal to $|\mathcal{S}| \max\{\|\mathbb{E}[\boldsymbol{Z}_{ij}^{(k)} (\boldsymbol{Z}_{ij}^{(k)})^T]\|, \|\mathbb{E}[(\boldsymbol{Z}_{ij}^{(k)})^T \boldsymbol{Z}_{ij}^{(k)}]\|\}$ since $\varepsilon_p \in \{-1, 1\}$ and each data point in $\mathcal{S}$ is i.i.d.

Towards bounding v, we have the following technical lemma, proved later in this section:

Lemma D.1. *For $k \sim \mathrm{Unif}([K])$ and $1 \le i < j \le n$ such that (i,j) is chosen uniformly at random from the set of $\binom{n}{2}$ unique pairs,*

$$\mathbb{E}\left[\boldsymbol{Z}_{ij}^{(k)}(\boldsymbol{Z}_{ij}^{(k)})^T\right] = \frac{2}{n(n-1)}\left[\boldsymbol{X}(n\boldsymbol{D}-\boldsymbol{G})\boldsymbol{X}^T + n\boldsymbol{X}\boldsymbol{X}^T - n^2\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^T\right]$$

where $\boldsymbol{G} \coloneqq \boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{D} \coloneqq \mathrm{diag}([\|\boldsymbol{x}_1\|^2, \dots, \|\boldsymbol{x}_n\|^2])$, and $\overline{\boldsymbol{x}} \coloneqq \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i$. Furthermore,

$$\mathbb{E}\left[(\boldsymbol{Z}_{ij}^{(k)})^T\boldsymbol{Z}_{ij}^{(k)}\right] = \frac{2}{n(n-1)}\begin{bmatrix} \boldsymbol{X}(n\boldsymbol{D}-\boldsymbol{G})\boldsymbol{X}^T & \frac{n}{K}\cdot\sum_{\ell=1}^n(\boldsymbol{x}_\ell-\overline{\boldsymbol{x}})^T\boldsymbol{x}_\ell\cdot\boldsymbol{x}_\ell\boldsymbol{1}_K^T \\ \frac{n}{K}\cdot\boldsymbol{1}_K\sum_{\ell=1}^n(\boldsymbol{x}_\ell-\overline{\boldsymbol{x}})^T\boldsymbol{x}_\ell\cdot\boldsymbol{x}_\ell^T & \frac{n}{K}\cdot\left(\|\boldsymbol{X}\|_F^2 - n\|\overline{\boldsymbol{x}}\|^2\right)\boldsymbol{I}_K \end{bmatrix},$$

and

$$\max\{\|\mathbb{E}[\boldsymbol{Z}_{ij}^{(k)}(\boldsymbol{Z}_{ij}^{(k)})^T]\|, \|\mathbb{E}[(\boldsymbol{Z}_{ij}^{(k)})^T\boldsymbol{Z}_{ij}^{(k)}]\|\} \le \left(4(B^2+1) + \frac{4\min(d,n)}{K}\right)\frac{\|\boldsymbol{X}\|^2}{n} + \frac{16B^3}{\sqrt{K}}.$$

Therefore, we have

$$v \le |\mathcal{S}|\left[\left(4(B^2+1) + \frac{4\min(d,n)}{K}\right)\frac{\|\boldsymbol{X}\|^2}{n} + \frac{16B^3}{\sqrt{K}}\right].$$

Noting that $\boldsymbol{Z}_{ij}^{(k)}$ is $d \times d + K$, from Theorem 6.1.1 in [49],

$$\begin{aligned}
&\mathbb{E}\left[\left\|\sum_{\mathcal{S}}\varepsilon_p\begin{bmatrix}\boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{x}_j\boldsymbol{x}_j^T & \boldsymbol{0} & \cdots & \underbrace{\boldsymbol{x}_i - \boldsymbol{x}_j}_{\text{column } d+k} & \cdots & \boldsymbol{0}\end{bmatrix}\right\|\right] \\
&\quad = \mathbb{E}\left[\left\|\sum_{\mathcal{S}}\varepsilon_p\boldsymbol{Z}_{ij}^{(k)}\right\|\right] \\
&\quad \le \sqrt{2v\log(2d+K)} + \frac{2(B^2+B)}{3}\log(2d+K) \\
&\quad \le \sqrt{2|\mathcal{S}|\log(2d+K)\left[\left(4(B^2+1) + \frac{4\min(d,n)}{K}\right)\frac{\|\boldsymbol{X}\|^2}{n} + \frac{16B^3}{\sqrt{K}}\right]} + \frac{2(B^2+B)}{3}\log(2d+K),
\end{aligned}$$

and so, continuing where we left off at the proof of Theorem 3.1,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\boldsymbol{M},\{\boldsymbol{v}_k\}_{k=1}^K}|\widehat{R}(\boldsymbol{M},\{\boldsymbol{v}_k\}_{k=1}^K) - R(\boldsymbol{M},\{\boldsymbol{v}_k\}_{k=1}^K)|\right] \\
&\quad \le 2L\sqrt{\frac{2\lambda_*^2\log(2d+K)}{|\mathcal{S}|}\left[\left(4(B^2+1) + \frac{4\min(d,n)}{K}\right)\frac{\|\boldsymbol{X}\|^2}{n} + \frac{16B^3}{\sqrt{K}}\right]} + \frac{4L(B^2+B)\lambda_*}{3|\mathcal{S}|}\log(2d+K).
\end{aligned}$$

Combining this with the first part of the proof of Theorem 3.1 (which, as we mentioned is identical here), with probability at least $1-\delta$

$$\begin{aligned}
&R(\widehat{\boldsymbol{M}},\{\widehat{\boldsymbol{v}}_k\}_{k=1}^K) - R(\boldsymbol{M}^*,\{\boldsymbol{v}_k^*\}_{k=1}^K) \\
&\le 2L\sqrt{\frac{2\lambda_*^2\log(2d+K)}{|\mathcal{S}|}\left[\left(4(B^2+1) + \frac{4\min(d,n)}{K}\right)\frac{\|\boldsymbol{X}\|^2}{n} + \frac{16B^3}{\sqrt{K}}\right]} + \\
&\frac{4L(B^2+B)\lambda_*}{3|\mathcal{S}|}\log(2d+K) + \sqrt{\frac{8L^2\gamma^2\log(2/\delta)}{|\mathcal{S}|}}.
\end{aligned} \tag{42}$$

Taking $B = 1$, we have the desired result.

Proof of Lemma D.1: Break $\boldsymbol{Z}_{i,j}^{(k)}$ into submatrices

$$\boldsymbol{A}_{ij} \coloneqq \boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{x}_j\boldsymbol{x}_j^T$$

and

$$\boldsymbol{B}_{ij}^{(k)} \coloneqq \begin{bmatrix}\boldsymbol{0} & \cdots & \underbrace{\boldsymbol{x}_i - \boldsymbol{x}_j}_{\text{column } k} & \cdots & \boldsymbol{0}\end{bmatrix}.$$

We proceed by computing the expected products of all submatrix combinations. Throughout, we will be summing over all $\binom{n}{2}$ item pairs. To reduce constant factors to track, we will use the fact that $\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\boldsymbol{Q}_{ij} = \frac{1}{2}\sum_{i=1}^{n}\sum_{j\neq i}\boldsymbol{Q}_{ij} = \frac{1}{2}\sum_{i\neq j}\boldsymbol{Q}_{ij}$ for matrices $\boldsymbol{Q}_{ij}$ satisfying $\boldsymbol{Q}_{ij} = \boldsymbol{Q}_{ji}$.

Step 1: Computing $\mathbb{E}\left[\boldsymbol{A}_{ij}^T\boldsymbol{A}_{ij}\right] = \mathbb{E}\left[\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^T\right]$

Note that the above equality holds by symmetry of $\boldsymbol{A}_{ij}$. Define $\boldsymbol{E}_{ij} \coloneqq \boldsymbol{e}_i\boldsymbol{e}_i^T - \boldsymbol{e}_j\boldsymbol{e}_j$, and note that $\boldsymbol{A}_{ij} = \boldsymbol{X}\boldsymbol{E}_{ij}\boldsymbol{X}^T$. Therefore

$$\begin{aligned}
\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\boldsymbol{A}_{ij}^T\boldsymbol{A}_{ij} &= \frac{1}{2}\sum_{i\neq j}\boldsymbol{A}_{ij}^T\boldsymbol{A}_{ij}\\
&= \frac{1}{2}\sum_{i\neq j}\boldsymbol{X}\boldsymbol{E}_{ij}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{E}_{ij}\boldsymbol{X}^T\\
&= \frac{1}{2}\boldsymbol{X}\left(\sum_{i\neq j}\boldsymbol{E}_{ij}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{E}_{ij}\right)\boldsymbol{X}^T\\
&= \frac{1}{2}\boldsymbol{X}\left(\sum_{i\neq j}\boldsymbol{E}_{ij}\boldsymbol{G}\boldsymbol{E}_{ij}\right)\boldsymbol{X}^T\\
&= \frac{1}{2}\boldsymbol{X}\left(\sum_{i\neq j}\|\boldsymbol{x}_i\|^2\boldsymbol{e}_i\boldsymbol{e}_i^T + \|\boldsymbol{x}_j\|^2\boldsymbol{e}_j\boldsymbol{e}_j^T - \boldsymbol{x}_i^T\boldsymbol{x}_j(\boldsymbol{e}_i\boldsymbol{e}_j^T + \boldsymbol{e}_j\boldsymbol{e}_i^T)\right)\boldsymbol{X}^T\\
&= \frac{1}{2}\boldsymbol{X}\left(2(n-1)\boldsymbol{D} - \sum_{i\neq j}\boldsymbol{x}_i^T\boldsymbol{x}_j(\boldsymbol{e}_i\boldsymbol{e}_j^T + \boldsymbol{e}_j\boldsymbol{e}_i^T)\right)\boldsymbol{X}^T\\
&= \boldsymbol{X}\left(n\boldsymbol{D} - \boldsymbol{G}\right)\boldsymbol{X}^T.
\end{aligned}$$

As $\boldsymbol{A}_{ij}$ does not depend on the random variable k,

$$\mathbb{E}\left[\boldsymbol{A}_{ij}^T\boldsymbol{A}_{ij}\right] = \frac{1}{\binom{n}{2}}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\boldsymbol{A}_{ij}^T\boldsymbol{A}_{ij} = \frac{2}{n(n-1)}\boldsymbol{X}\left(n\boldsymbol{D} - \boldsymbol{G}\right)\boldsymbol{X}^T.$$

Step 2: Computing $\mathbb{E}[\boldsymbol{A}_{ij}^T\boldsymbol{B}_{ij}^{(k)}]$

$$\boldsymbol{A}_{ij}^T\boldsymbol{B}_{ij}^{(k)} = \left(\boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{x}_j\boldsymbol{x}_j^T\right)\begin{bmatrix}\boldsymbol{0} & \cdots & \underbrace{\boldsymbol{x}_i - \boldsymbol{x}_j}_{\text{column } k} & \cdots & \boldsymbol{0}\end{bmatrix}.$$

Hence, the product is $\boldsymbol{0}$ for all columns except the k^{th}. As we sum over all $k \in [K]$ to compute the expectation, the resulting submatrix is rank 1 with K copies of this same column. We therefore, compute the expectation of this column first.

$$\begin{aligned}
\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\boldsymbol{x}_i\boldsymbol{x}_i^T - \boldsymbol{x}_j\boldsymbol{x}_j^T\right)(\boldsymbol{x}_i - \boldsymbol{x}_j) &= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\|\boldsymbol{x}_i\|^2\boldsymbol{x}_i + \|\boldsymbol{x}_j\|^2\boldsymbol{x}_j - \boldsymbol{x}_i^T\boldsymbol{x}_j\boldsymbol{x}_i - \boldsymbol{x}_i^T\boldsymbol{x}_j\boldsymbol{x}_j\\
&= \frac{1}{2}\sum_{i}\sum_{j\neq i}\|\boldsymbol{x}_i\|^2\boldsymbol{x}_i + \|\boldsymbol{x}_j\|^2\boldsymbol{x}_j - \boldsymbol{x}_i^T\boldsymbol{x}_j\boldsymbol{x}_i - \boldsymbol{x}_i^T\boldsymbol{x}_j\boldsymbol{x}_j\\
&= (n-1)\sum_{i}\|\boldsymbol{x}_i\|^2\boldsymbol{x}_i - \sum_{i}\sum_{j\neq i}\boldsymbol{x}_i^T\boldsymbol{x}_j\boldsymbol{x}_i\\
&= n\sum_{i}\|\boldsymbol{x}_i\|^2\boldsymbol{x}_i - \sum_{i}\sum_{j}\boldsymbol{x}_i^T\boldsymbol{x}_j\boldsymbol{x}_i\\
&= n\sum_{i}\|\boldsymbol{x}_i\|^2\boldsymbol{x}_i - \sum_{i}\sum_{j}\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{x}_j
\end{aligned}$$

$$
\begin{aligned}
&= n\sum_i \|\boldsymbol{x}_i\|^2\boldsymbol{x}_i - \sum_i \boldsymbol{x}_i\boldsymbol{x}_i^T\sum_j \boldsymbol{x}_j \\
&= n\sum_i \|\boldsymbol{x}_i\|^2\boldsymbol{x}_i - n\sum_i \boldsymbol{x}_i^T\overline{\boldsymbol{x}}\boldsymbol{x}_i \\
&= n\sum_i \left(\|\boldsymbol{x}_i\|^2 - \boldsymbol{x}_i^T\overline{\boldsymbol{x}}\right)\boldsymbol{x}_i \\
&= n\sum_i \left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)^T \boldsymbol{x}_i \cdot \boldsymbol{x}_i.
\end{aligned}
$$

We therefore have

$$
\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \boldsymbol{A}_{ij}^T\boldsymbol{B}_{ij}^{(k)} = \begin{bmatrix} \boldsymbol{0} & \cdots & \underbrace{n\sum_i \left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)^T \boldsymbol{x}_i \cdot \boldsymbol{x}_i}_{\text{column } k} & \cdots & \boldsymbol{0} \end{bmatrix}.
$$

When we then sum over k and normalize by $1/K$, and divide by the $\binom{n}{2}$ unique pairs to finish the expectation computation, we have

$$
\mathbb{E}[\boldsymbol{A}_{ij}^T\boldsymbol{B}_{ij}^{(k)}] = \frac{n}{K\binom{n}{2}}\sum_i \left(\boldsymbol{x}_i - \overline{\boldsymbol{x}}\right)^T \boldsymbol{x}_i \cdot \boldsymbol{x}_i\boldsymbol{1}_K^T = \frac{2n}{K\cdot n\cdot(n-1)}\cdot\sum_{\ell=1}^{n}(\boldsymbol{x}_\ell - \overline{\boldsymbol{x}})^T\boldsymbol{x}_\ell\cdot\boldsymbol{x}_\ell\boldsymbol{1}_K^T,
$$

where the factor of $\boldsymbol{1}_K$ simply generates a matrix with K copies of this same column.

Step 3: Computing $\mathbb{E}[(\boldsymbol{B}_{ij}^{(k)})^T\boldsymbol{B}_{ij}^{(k)}]$

$$
(\boldsymbol{B}_{ij}^{(k)})^T\boldsymbol{B}_{ij}^{(k)} = \begin{bmatrix} \boldsymbol{0}^T \\ \vdots \\ (\boldsymbol{x}_i - \boldsymbol{x}_j)^T \\ \vdots \\ \boldsymbol{0}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{0} & \cdots & \boldsymbol{x}_i - \boldsymbol{x}_j & \cdots & \boldsymbol{0} \end{bmatrix}
$$

Hence,

$$
[(\boldsymbol{B}_{ij}^{(k)})^T\boldsymbol{B}_{ij}^{(k)}]_{p,q} = \begin{cases} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 & \text{if } p = q = k \\ 0 & \text{otherwise} \end{cases}.
$$

Since the non-zero entry in this matrix does not depend on k, $\mathbb{E}[(\boldsymbol{B}_{ij}^{(k)})^T\boldsymbol{B}_{ij}^{(k)}]$ is a equal to a constant times the K-dimensional identity. To compute this constant, first we evaluate

$$
\begin{aligned}
\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 &= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \boldsymbol{G}_{ii} + \boldsymbol{G}_{jj} - 2\boldsymbol{G}_{ij} \\
&= \frac{1}{2}\sum_i\sum_{j\neq i} \boldsymbol{G}_{ii} + \boldsymbol{G}_{jj} - 2\boldsymbol{G}_{ij} \\
&= n\cdot\mathrm{Tr}(\boldsymbol{G}) - \sum_i\sum_j \boldsymbol{G}_{ij} \\
&= n\cdot\mathrm{Tr}(\boldsymbol{G}) - \sum_i\sum_j \boldsymbol{x}_i^T\boldsymbol{x}_j \\
&= n\cdot\mathrm{Tr}(\boldsymbol{G}) - \left(\sum_i \boldsymbol{x}_i\right)^T\sum_j \boldsymbol{x}_j \\
&= n\|\boldsymbol{X}\|_F^2 - n^2\|\overline{\boldsymbol{x}}\|^2
\end{aligned}
$$

The first equality holds since for a Gram matrix $\boldsymbol{G} = \boldsymbol{X}^T\boldsymbol{X}$, we have that $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 = \boldsymbol{G}_{ii} - 2\boldsymbol{G}_{ij} + \boldsymbol{G}_{jj}$. In the final equality, we have used the fact that $\mathrm{Tr}(\boldsymbol{G}) = \|\boldsymbol{X}\|_F^2$.

By summing $(\boldsymbol{B}_{ij}^{(k)})^T\boldsymbol{B}_{ij}^{(k)}$ over all users and unique pairs and then dividing by K and then $\binom{n}{2}$ to compute the expectation, we have:

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{B}_{ij}^{(k)})^T\boldsymbol{B}_{ij}^{(k)}] &= \frac{1}{K\binom{n}{2}}\sum_{k=1}^{K}\sum_{i\neq j}(\boldsymbol{B}_{ij}^{(k)})^T\boldsymbol{B}_{ij}^{(k)} \\
&= \frac{2}{K\cdot n\cdot(n-1)}\cdot\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2\boldsymbol{I}_K \\
&= \frac{2n}{K\cdot n\cdot(n-1)}\cdot\left(\|\boldsymbol{X}\|_F^2-n\|\overline{\boldsymbol{x}}\|^2\right)\boldsymbol{I}_K.
\end{aligned}$$

Steps 1-3 establish the second claim regarding $\mathbb{E}\left[(\boldsymbol{Z}_{ij}^{(k)})^T\boldsymbol{Z}_{ij}^{(k)}\right]$ by noting that

$$\mathbb{E}\left[(\boldsymbol{Z}_{ij}^{(k)})^T\boldsymbol{Z}_{ij}^{(k)}\right] = \begin{bmatrix}\mathbb{E}\left[\boldsymbol{A}_{ij}^T\boldsymbol{A}_{ij}\right] & \mathbb{E}[\boldsymbol{A}_{ij}^T\boldsymbol{B}_{ij}^{(k)}] \\ \mathbb{E}[(\boldsymbol{B}_{ij}^{(k)})^T\boldsymbol{A}_{ij}] & \mathbb{E}[(\boldsymbol{B}_{ij}^{(k)})^T\boldsymbol{B}_{ij}^{(k)}]\end{bmatrix},$$

where we have used both the result of Step 2 and its transpose.

Step 4: Computing $\mathbb{E}[\boldsymbol{B}_{ij}^{(k)}(\boldsymbol{B}_{ij}^{(k)})^T]$

$$\boldsymbol{B}_{ij}^{(k)}(\boldsymbol{B}_{ij}^{(k)})^T = [\boldsymbol{0}\quad\cdots\quad\boldsymbol{x}_i-\boldsymbol{x}_j\quad\cdots\quad\boldsymbol{0}]\begin{bmatrix}\boldsymbol{0}^T\\ \vdots \\ (\boldsymbol{x}_i-\boldsymbol{x}_j)^T \\ \vdots \\ \boldsymbol{0}^T\end{bmatrix} = (\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T.$$

We compute

$$\begin{aligned}
\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T &= \frac{1}{2}\sum_{i}\sum_{j\neq i}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T \\
&= (n-1)\sum_i\boldsymbol{x}_i\boldsymbol{x}_i^T-\sum_i\sum_{j\neq i}\boldsymbol{x}_i\boldsymbol{x}_j^T \\
&= n\sum_i\boldsymbol{x}_i\boldsymbol{x}_i^T-\sum_i\sum_j\boldsymbol{x}_i\boldsymbol{x}_j^T \\
&= n\boldsymbol{X}\boldsymbol{X}^T-\sum_i\boldsymbol{x}_i\sum_j\boldsymbol{x}_j^T \\
&= n\boldsymbol{X}\boldsymbol{X}^T-n^2\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^T.
\end{aligned}$$

Note that this expression does not depend on K, and so

$$\mathbb{E}[\boldsymbol{B}_{ij}^{(k)}(\boldsymbol{B}_{ij}^{(k)})^T] = \frac{1}{\binom{n}{2}}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^T = \frac{2}{n\cdot(n-1)}\left(n\boldsymbol{X}\boldsymbol{X}^T-n^2\overline{\boldsymbol{x}}\overline{\boldsymbol{x}}^T\right).$$

Steps 1 and 4 establish the claim regarding $\mathbb{E}\left[\boldsymbol{Z}_{ij}^{(k)}(\boldsymbol{Z}_{ij}^{(k)})^T\right]$ by noting that

$$\mathbb{E}\left[\boldsymbol{Z}_{ij}^{(k)}(\boldsymbol{Z}_{ij}^{(k)})^T\right] = \mathbb{E}\left[\boldsymbol{A}_{ij}\boldsymbol{A}_{ij}^T\right]+\mathbb{E}[\boldsymbol{B}_{ij}^{(k)}(\boldsymbol{B}_{ij}^{(k)})^T].$$

To bound $\|\mathbb{E}[\boldsymbol{Z}_{ij}^{(k)}(\boldsymbol{Z}_{ij}^{(k)})^T]\|$, note that

$$\mathbb{E}\left[\boldsymbol{Z}_{ij}^{(k)}(\boldsymbol{Z}_{ij}^{(k)})^T\right] = \frac{2}{n(n-1)}\boldsymbol{X}\left[n\boldsymbol{D}-\boldsymbol{G}+n\boldsymbol{I}-\boldsymbol{1}\boldsymbol{1}^T\right]\boldsymbol{X}^T.$$

We can expand the center term as

$$n\boldsymbol{D} - \boldsymbol{G} + n\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T = \begin{cases} (n-1)(\|\boldsymbol{x}_i\|^2 + 1) & i = j \\ -\boldsymbol{x}_i^T \boldsymbol{x}_j - 1 & i \neq j \end{cases}.$$

From the Gershgorin Circle Theorem we then have

$$\begin{aligned} \|n\boldsymbol{D} - \boldsymbol{G} + n\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T\| &\leq \max_i [(n-1)(\|\boldsymbol{x}_i\|^2 + 1) + \sum_{j \neq i} |\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + 1|] \\ &\leq \max_i [(n-1)(\|\boldsymbol{x}_i\|^2 + 1) + n - 1 + \|\boldsymbol{x}_i\| \sum_{j \neq i} \|\boldsymbol{x}_j\|] \\ &\leq (n-1)(B^2+1) + n - 1 + (n-1)B^2 \\ &= 2(n-1)(B^2+1). \end{aligned}$$

We then have

$$\begin{aligned} \left\| \mathbb{E}\left[\boldsymbol{Z}_{ij}^{(k)} (\boldsymbol{Z}_{ij}^{(k)})^T \right] \right\| &\leq \frac{2}{n(n-1)} \|\boldsymbol{X}\|^2 \|n\boldsymbol{D} - \boldsymbol{G} + n\boldsymbol{I} - \boldsymbol{1}\boldsymbol{1}^T\| \\ &\leq \frac{4(B^2+1)}{n} \|\boldsymbol{X}\|^2. \end{aligned}$$

We take a slightly different approach in bounding $\left\| \mathbb{E}\left[(\boldsymbol{Z}_{ij}^{(k)})^T \boldsymbol{Z}_{ij}^{(k)} \right] \right\|$. First, we decompose $\mathbb{E}\left[(\boldsymbol{Z}_{ij}^{(k)})^T \boldsymbol{Z}_{ij}^{(k)} \right]$ as

$$\mathbb{E}\left[(\boldsymbol{Z}_{ij}^{(k)})^T \boldsymbol{Z}_{ij}^{(k)} \right] = \frac{2}{n(n-1)} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^T & \boldsymbol{C} \end{bmatrix},$$

where $\boldsymbol{A} = \boldsymbol{X}(n\boldsymbol{D} - \boldsymbol{G})\boldsymbol{X}^T$, $\boldsymbol{B} = \frac{n}{K} \cdot \sum_{\ell=1}^n (\boldsymbol{x}_\ell - \overline{\boldsymbol{x}})^T \boldsymbol{x}_\ell \cdot \boldsymbol{x}_\ell \boldsymbol{1}_K^T$, and $\boldsymbol{C} = \frac{n}{K} \cdot \left(\|\boldsymbol{X}\|_F^2 - n\|\overline{\boldsymbol{x}}\|^2 \right) \boldsymbol{I}_K$. By repeated applications of the triangle inequality,

$$\begin{aligned} \left\| \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}^T & \boldsymbol{C} \end{bmatrix} \right\| &\leq \left\| \begin{bmatrix} \boldsymbol{A} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \right\| + \left\| \begin{bmatrix} \boldsymbol{0} & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \right\| + \left\| \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{B}^T & \boldsymbol{0} \end{bmatrix} \right\| + \left\| \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C} \end{bmatrix} \right\| \\ &= \|\boldsymbol{A}\| + 2\|\boldsymbol{B}\| + \|\boldsymbol{C}\|. \end{aligned}$$

We bound each of these terms in turn. Noting that $\boldsymbol{A} = \boldsymbol{X}(n\boldsymbol{D} - \boldsymbol{G})\boldsymbol{X}^T$ and that

$$n\boldsymbol{D} - \boldsymbol{G} = \begin{cases} (n-1)\|\boldsymbol{x}_i\|^2 & i = j \\ -\boldsymbol{x}_i^T \boldsymbol{x}_j & i \neq j \end{cases},$$

and so by applying the Gershgorin Circle Theorem as above we have

$$\begin{aligned} \|n\boldsymbol{D} - \boldsymbol{G}\| &\leq \max_i [(n-1)\|\boldsymbol{x}_i\|^2 + \sum_{j \neq i} |\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle|] \\ &\leq \max_i [(n-1)\|\boldsymbol{x}_i\|^2 + \|\boldsymbol{x}_i\| \sum_{j \neq i} \|\boldsymbol{x}_j\|] \\ &\leq (n-1)B^2 + (n-1)B^2 \\ &= 2(n-1)B^2, \end{aligned}$$

and so

$$\|\boldsymbol{A}\| = \|\boldsymbol{X}(n\boldsymbol{D} - \boldsymbol{G})\boldsymbol{X}^T\| \leq \|\boldsymbol{X}\|^2 \|n\boldsymbol{D} - \boldsymbol{G}\| \leq 2(n-1)B^2 \|\boldsymbol{X}\|^2.$$

Next, we bound $\|\boldsymbol{B}\|$. First note that for any matrix of the form $\boldsymbol{z}\boldsymbol{1}_K^T$ where $\boldsymbol{z} \in \mathbb{R}^K$,

$$\boldsymbol{z}\boldsymbol{1}_K^T = \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_2} (\|\boldsymbol{z}\|_2 \sqrt{K}) \frac{\boldsymbol{1}_K^T}{\sqrt{K}}$$

and so $\|\boldsymbol{z}\boldsymbol{1}_K^T\| = \|\boldsymbol{z}\|_2 \sqrt{K}$. Applying this result to $\boldsymbol{B}$,

$$\|\boldsymbol{B}\| = \left\| \frac{n}{K} \cdot \sum_{\ell=1}^n (\boldsymbol{x}_\ell - \overline{\boldsymbol{x}})^T \boldsymbol{x}_\ell \cdot \boldsymbol{x}_\ell \boldsymbol{1}_K^T \right\|$$

$$
\begin{aligned}
&= \frac{n}{\sqrt{K}} \left\| \sum_{\ell=1}^{n} (\boldsymbol{x}_\ell - \overline{\boldsymbol{x}})^T \boldsymbol{x}_\ell \cdot \boldsymbol{x}_\ell \right\|_2 \\
&\le \frac{n}{\sqrt{K}} \sum_{\ell=1}^{n} \left|(\boldsymbol{x}_\ell - \overline{\boldsymbol{x}})^T \boldsymbol{x}_\ell\right| \|\boldsymbol{x}_\ell\|_2 \\
&\le \frac{2n^2 B^3}{\sqrt{K}}.
\end{aligned}
$$

Finally,

$$
\begin{aligned}
\|\boldsymbol{D}\| &= \left\| \frac{n}{K} \cdot \left(\|\boldsymbol{X}\|_F^2 - n\|\overline{\boldsymbol{x}}\|^2\right) \boldsymbol{I}_K \right\| \\
&= \frac{n}{K} \left| \|\boldsymbol{X}\|_F^2 - n\|\overline{\boldsymbol{x}}\|^2 \right| \\
&= \frac{n}{K} \left| \|\boldsymbol{X}\|_F^2 - n\overline{\boldsymbol{x}}^T\overline{\boldsymbol{x}} \right| \\
&= \frac{n}{K} \left| \|\boldsymbol{X}\|_F^2 - \frac{1}{n} \left(\sum_{i=1}^{n} \boldsymbol{x}_i \right)^T \left(\sum_{i=1}^{n} \boldsymbol{x}_i \right) \right| \\
&= \frac{n}{K} \left| \|\boldsymbol{X}\|_F^2 - \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{x}_i^T \boldsymbol{x}_j \right| \\
&= \frac{n}{K} \left| \|\boldsymbol{X}\|_F^2 - \frac{1}{n} \boldsymbol{1}_n^T \boldsymbol{G} \boldsymbol{1}_n \right| \\
&= \frac{n}{K} \left| \mathrm{Tr}(\boldsymbol{G}) - \frac{1}{n} \boldsymbol{1}_n^T \boldsymbol{G} \boldsymbol{1}_n \right| \\
&= \frac{n}{K} \left| \mathrm{Tr}(\boldsymbol{G}) - \frac{\boldsymbol{1}_n}{\sqrt{n}}^T \boldsymbol{G} \frac{\boldsymbol{1}_n}{\sqrt{n}} \right|
\end{aligned}
$$

Let λ_i, $i \in [n]$ denote the eigenvalues of $\boldsymbol{G}$, sorted in decreasing order. Each $\lambda_i \ge 0$ since $\boldsymbol{G}$ is positive semidefinite by construction. We can then rewrite $\mathrm{Tr}(\boldsymbol{G}) = \sum_i \lambda_i$, and $\frac{\boldsymbol{1}_n}{\sqrt{n}}^T \boldsymbol{G} \frac{\boldsymbol{1}_n}{\sqrt{n}} \le \max_{\|\boldsymbol{x}\|_2=1} \boldsymbol{x}^T \boldsymbol{G} \boldsymbol{x} = \lambda_1$, and so $\mathrm{Tr}(\boldsymbol{G}) - \frac{\boldsymbol{1}_n}{\sqrt{n}}^T \boldsymbol{G} \frac{\boldsymbol{1}_n}{\sqrt{n}} \ge \mathrm{Tr}(\boldsymbol{G}) - \lambda_1 = \sum_{i>1} \lambda_i \le 0$, and so

$$
\left| \mathrm{Tr}(\boldsymbol{G}) - \frac{\boldsymbol{1}_n}{\sqrt{n}}^T \boldsymbol{G} \frac{\boldsymbol{1}_n}{\sqrt{n}} \right| = \mathrm{Tr}(\boldsymbol{G}) - \frac{\boldsymbol{1}_n}{\sqrt{n}}^T \boldsymbol{G} \frac{\boldsymbol{1}_n}{\sqrt{n}} \le \mathrm{Tr}(\boldsymbol{G}) = \|\boldsymbol{X}\|_F^2,
$$

where the final inequality follows since $\boldsymbol{G}$ is positive semidefinite. Therefore, $\|\boldsymbol{D}\| \le ({}^{n}\!/\!_{K})\|\boldsymbol{X}\|_F^2$.

Combining these bounds, we have

$$
\begin{aligned}
\left\| \mathbb{E}\left[(\boldsymbol{Z}_{ij}^{(k)})^T \boldsymbol{Z}_{ij}^{(k)} \right] \right\| &\le \frac{2}{n(n-1)} (\|\boldsymbol{A}\| + 2\|\boldsymbol{B}\| + \|\boldsymbol{C}\|) \\
&\le \frac{2}{n(n-1)} \left(2(n-1)B^2\|\boldsymbol{X}\|^2 + \frac{4n^2B^3}{\sqrt{K}} + \frac{n}{K}\|\boldsymbol{X}\|_F^2 \right) \\
&= \frac{4B^2}{n}\|\boldsymbol{X}\|^2 + \frac{8B^3 n}{(n-1)\sqrt{K}} + \frac{2n}{K(n-1)} \frac{\|\boldsymbol{X}\|_F^2}{n} \\
&\le \frac{4B^2}{n}\|\boldsymbol{X}\|^2 + \frac{16B^3}{\sqrt{K}} + \frac{4}{K} \frac{\|\boldsymbol{X}\|_F^2}{n} \quad \text{since } n \ge 2 \implies {}^{n}\!/\!_{n-1} \le 2 \\
&\le \left(4B^2 + \frac{4\min(d,n)}{K} \right) \frac{\|\boldsymbol{X}\|^2}{n} + \frac{16B^3}{\sqrt{K}} \quad \text{since } \|\boldsymbol{X}\|_F^2 \le \mathrm{rank}(\boldsymbol{X})\|\boldsymbol{X}\|^2 \le \min(d,n)\|\boldsymbol{X}\|^2
\end{aligned}
$$

Therefore,

$$
\max\{\| \mathbb{E}[\boldsymbol{Z}_{ij}^{(k)} (\boldsymbol{Z}_{ij}^{(k)})^T]\|, \|\mathbb{E}[(\boldsymbol{Z}_{ij}^{(k)})^T \boldsymbol{Z}_{ij}^{(k)}]\|\} \le
$$

$$
\begin{aligned}
&\max\left\{\frac{4(B^2+1)}{n}\|\boldsymbol{X}\|^2, \left(4B^2 + \frac{4\min(d,n)}{K}\right)\frac{\|\boldsymbol{X}\|^2}{n} + \frac{16B^3}{\sqrt{K}}\right\} \\
&\leq \left(4(B^2+1) + \frac{4\min(d,n)}{K}\right)\frac{\|\boldsymbol{X}\|^2}{n} + \frac{16B^3}{\sqrt{K}},
\end{aligned}
$$

completing the proof.

As an aside, we believe that this analysis can be tightened. Specifically, we believe that the final $O(1/\sqrt{K})$ term can be sharpened and that the maximum above should scale as $\|\boldsymbol{X}\|^2/n$, which would eliminate the requirement that $K = \Omega(d^2)$ in Corollary 3.3.1.

D.3 Proof of Corollary 3.3.1

To prove the corollary, we will select values for the constants in Theorem 3.3 that hold with high probability, based on our assumed item, user, and metric distribution. We will derive our result from the extended version of Theorem 3.3 in eq. (42), which holds with probability at least $1-\delta$ for a constant B such that $\|\boldsymbol{x}_i\|_2 \leq B$ for all $i \in [n]$ and a specification of $0 < \delta < 1$.

To arrive at a setting for B, we can rewrite our item vectors as $\boldsymbol{x}_i = \frac{1}{\sqrt{d}}\boldsymbol{\eta}_i$, where $\boldsymbol{\eta}_i$ are i.i.d. $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, and so $\|\boldsymbol{x}_i\|_2^2 = \frac{1}{d}\|\boldsymbol{\eta}_i\|_2^2$. Since $\|\boldsymbol{\eta}_i\|_2^2$ is a chi-squared random variable with d degrees of freedom, from [50] we have that for any $t > 0$, $\mathbb{P}(\|\boldsymbol{\eta}_i\|_2^2 \geq d + 2\sqrt{dt} + 2t) \leq e^{-t}$, and therefore by the union bound we have

$$
\mathbb{P}(\max_{i\in[n]}\|\boldsymbol{\eta}_i\|_2^2 \geq d + 2\sqrt{dt} + 2t) \leq \sum_{i\in[n]}\mathbb{P}(\|\boldsymbol{\eta}_i\|_2^2 \geq d + 2\sqrt{dt} + 2t) \leq ne^{-t}.
$$

Setting $t = \log\frac{n}{\delta_1}$ for any given $0 < \delta_1 < 1$, we have with probability greater than $1-\delta_1$ that $\max_{i\in[n]}\|\boldsymbol{\eta}_i\|_2^2 < d + 2\sqrt{d\log\frac{n}{\delta_1}} + 2\log\frac{n}{\delta_1}$, which implies $\max_{i\in[n]}\|\boldsymbol{x}_i\|_2^2 < 1 + 2\sqrt{\frac{1}{d}\log\frac{n}{\delta_1}} + \frac{2}{d}\log\frac{n}{\delta_1}$. To get a more interpretable bound, we note for $n \geq 3$ that $\log\frac{n}{\delta} > 1$ and therefore with probability at least $1-\delta_1$, $\max_{i\in[n]}\|\boldsymbol{x}_i\|_2^2 < 5\log\frac{n}{\delta_1}$. We can therefore set $B = \sqrt{5\log\frac{n}{\delta_1}}$.

Towards a setting for γ as defined in eq. (7), let $\boldsymbol{z}_i := \sqrt{\frac{d}{\sqrt{r}}}\boldsymbol{L}^T\boldsymbol{x}_i$, in which case $\boldsymbol{x}_i^T\boldsymbol{M}\boldsymbol{x}_i = \|\boldsymbol{z}_i\|_2^2$. $\boldsymbol{z}_i$ is normally distributed with $\mathbb{E}[\boldsymbol{z}_i] = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{z}_i) = \mathbb{E}[\boldsymbol{z}_i\boldsymbol{z}_i^T] = \frac{d}{\sqrt{r}}\boldsymbol{L}^T\,\mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_i^T]\boldsymbol{L} = \frac{d}{\sqrt{r}}\boldsymbol{L}^T\left(\frac{1}{d}\boldsymbol{I}\right)\boldsymbol{L} = \frac{1}{\sqrt{r}}\boldsymbol{I}_r$ and therefore $\boldsymbol{z}_i \sim \mathcal{N}(\mathbf{0}, \frac{1}{\sqrt{r}}\boldsymbol{I}_r)$ where we notated the identity as $\boldsymbol{I}_r$ as a reminder that $\boldsymbol{z}_i \in \mathbb{R}^r$. Reusing notation, if $\boldsymbol{\eta}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_r)$, we can write $\boldsymbol{z}_i = r^{-\frac{1}{4}}\boldsymbol{\eta}_i$ and so $\|\boldsymbol{z}_i\|_2^2 = \frac{1}{\sqrt{r}}\|\boldsymbol{\eta}_i\|_2^2$. By the same arguments as above, for a given $0 < \delta_2 < 1$ we have with probability at least $1-\delta_2$ that $\max_{i\in[n]}\|\boldsymbol{\eta}_i\|_2^2 < 5r\log\frac{n}{\delta_2}$ and therefore $\max_{i\in[n]}\|\boldsymbol{z}_i\|_2^2 < 5\sqrt{r}\log\frac{n}{\delta_2}$. Applying a similar argument to the user points, letting $\boldsymbol{w}_k := \sqrt{\frac{d}{\sqrt{r}}}\boldsymbol{L}^T\boldsymbol{u}_k$ we have for a given $0 < \delta_3 < 1$ that with probability at least $1-\delta_3$, $\max_{k\in[K]}\|\boldsymbol{w}_k\|_2^2 < 5\sqrt{r}\log\frac{K}{\delta_3}$. We then have:

$$
\begin{aligned}
\max_{i,j\in[n],k\in[K]}|\delta_{i,j}^{(k)}| &= \max_{i,j\in[n],k\in[K]}|\boldsymbol{x}_i^T\boldsymbol{M}\boldsymbol{x}_i - \boldsymbol{x}_j^T\boldsymbol{M}\boldsymbol{x}_j + (\boldsymbol{x}_i - \boldsymbol{x}_j)^T\boldsymbol{v}_k| \\
&= \max_{i,j\in[n],k\in[K]}|\|\boldsymbol{z}_i\|_2^2 - \|\boldsymbol{z}_j\|_2^2 - 2(\boldsymbol{x}_i - \boldsymbol{x}_j)^T\frac{d}{\sqrt{r}}\boldsymbol{L}\boldsymbol{L}^T\boldsymbol{u}_k| \\
&= \max_{i,j\in[n],k\in[K]}|\|\boldsymbol{z}_i\|_2^2 - \|\boldsymbol{z}_j\|_2^2 - 2(\boldsymbol{z}_i - \boldsymbol{z}_j)^T\boldsymbol{w}_k| \\
&\leq \max_{i,j\in[n],k\in[K]}\|\boldsymbol{z}_i\|_2^2 + \|\boldsymbol{z}_j\|_2^2 + 2|(\boldsymbol{z}_i - \boldsymbol{z}_j)^T\boldsymbol{w}_k| \\
&\leq 2\max_{i\in[n]}\|\boldsymbol{z}_i\|_2^2 + 2\max_{i,j\in[n],k\in[K]}|(\boldsymbol{z}_i - \boldsymbol{z}_j)^T\boldsymbol{w}_k| \\
&\leq 2\max_{i\in[n]}\|\boldsymbol{z}_i\|_2^2 + 2\max_{i,j\in[n],k\in[K]}\|\boldsymbol{z}_i - \boldsymbol{z}_j\|_2\|\boldsymbol{w}_k\|_2 \\
&\leq 2\max_{i\in[n]}\|\boldsymbol{z}_i\|_2^2 + 4\max_{i\in[K]}\|\boldsymbol{z}_i\|_2\max_{k\in[K]}\|\boldsymbol{w}_k\|_2.
\end{aligned}
$$

Taking a union bound over the events $\max_{i\in[n]}\|\boldsymbol{z}_i\|_2^2 \geq 5\sqrt{r}\log\frac{n}{\delta_2}$ and $\max_{k\in[K]}\|\boldsymbol{w}_k\|_2^2 \geq 5\sqrt{r}\log\frac{K}{\delta_3}$ along with a failure of $\|\boldsymbol{x}_i\|_2 \leq B$ for all $i \in [n]$, and setting $\delta_1 = \delta_2 = \delta_3 =: \delta$ for convenience, we have with probability at least $1-3\delta$ that these failure events do not occur and so

$$\begin{aligned}\max_{i,j\in[n],k\in[K]}|\delta_{i,j}^{(k)}| &\leq 2\max_{i\in[n]}\|\boldsymbol{z}_i\|_2^2 + 4\max_{i\in[K]}\|\boldsymbol{z}_i\|_2\max_{k\in[K]}\|\boldsymbol{w}_k\|_2 \\ &\leq 10\sqrt{r}\log\frac{n}{\delta} + 20\sqrt{r\log\frac{n}{\delta}\log\frac{K}{\delta}} \\ &\leq 30\sqrt{r}\log\frac{\max\{n,K\}}{\delta},\end{aligned}$$

and so we can set $\gamma = 30\sqrt{r}\log\frac{\max\{n,K\}}{\delta}$.

We next bound $\|\boldsymbol{X}\|$. For convenience, let $\widetilde{\boldsymbol{x}}_i := \sqrt{d}\boldsymbol{x}_i$ such that $\widetilde{\boldsymbol{x}}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, and $\widetilde{\boldsymbol{X}} := [\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_n] = \sqrt{d}\boldsymbol{X}$, so that we have $\|\widetilde{\boldsymbol{X}}\| = \sqrt{d}\|\boldsymbol{X}\|$. From [21], we know that for any $t > 0$,

$$\mathbb{P}(\|\widetilde{\boldsymbol{X}}\| \geq \sqrt{n} + \sqrt{d} + t) < e^{-\frac{t^2}{2}}.$$

For any given $0 < \delta_4 < 1$, let $t = \sqrt{2\log\frac{1}{\delta_4}}$ in which case $\mathbb{P}(\|\widetilde{\boldsymbol{X}}\| \geq \sqrt{n} + \sqrt{d} + \sqrt{2\log\frac{1}{\delta_4}}) < \delta_4$. Therefore, with probability at least $1-\delta_4$, $\|\boldsymbol{X}\| = \frac{1}{\sqrt{d}}\|\widetilde{\boldsymbol{X}}\| \leq \sqrt{\frac{n}{d}} + 1 + \sqrt{\frac{2}{d}\log\frac{1}{\delta_4}} \leq \sqrt{3(\frac{n}{d} + 1 + \frac{2}{d}\log\frac{1}{\delta_4})}$ since from Jensen's inequality, for any m non-negative scalars $a_1, \ldots, a_m$ we have $\sum_i\sqrt{a_i} \leq \sqrt{m\sum_i a_i}$. Therefore, with probability at least $1-\delta_4$, $\|\boldsymbol{X}\|^2 \leq 3(\frac{n}{d}+1+\frac{2}{d}\log\frac{1}{\delta_4})$.

Finally, we select a setting for λ^*. Note that by definition each $\boldsymbol{v}_k$ lies in the column space of $\boldsymbol{M}$; hence, $[\boldsymbol{M}, \boldsymbol{v}_1, \ldots \boldsymbol{v}_K]$ is a rank r matrix. By norm equivalence, we have that

$$\begin{aligned}\|[\boldsymbol{M} \quad \boldsymbol{v}_1 \quad \cdots \quad \boldsymbol{v}_K]\|_* &\leq \sqrt{r}\|[\boldsymbol{M} \quad \boldsymbol{v}_1 \quad \cdots \quad \boldsymbol{v}_K]\|_F \\ &= \sqrt{r(\|\boldsymbol{M}\|_F^2 + \sum_k\|\boldsymbol{v}_k\|_2^2)} \\ &= \sqrt{r(d^2 + \sum_k\|\boldsymbol{v}_k\|_2^2)} \\ &\leq \sqrt{r(d^2 + K\max_{k\in[K]}\|\boldsymbol{v}_k\|_2^2)}.\end{aligned}$$

Note that $\|\boldsymbol{v}_k\|_2^2 = \|-2\frac{d}{\sqrt{r}}\boldsymbol{L}\boldsymbol{L}^T\boldsymbol{u}_k\|_2^2 = 4\frac{d^2}{r}\boldsymbol{u}_k^T\boldsymbol{L}\boldsymbol{L}^T\boldsymbol{L}\boldsymbol{L}^T\boldsymbol{u}_k = 4\frac{d^2}{r}\boldsymbol{u}_k^T\boldsymbol{L}\boldsymbol{L}^T\boldsymbol{u}_k = 4\frac{d}{\sqrt{r}}\|\boldsymbol{w}_k\|_2^2$. Recall that with probability at least $1-\delta_3$, $\max_{k\in[K]}\|\boldsymbol{w}_k\|_2^2 < 5\sqrt{r}\log\frac{K}{\delta_3}$. Therefore with probability at least $1-\delta_3$, $\max_{k\in[K]}\|\boldsymbol{v}_k\|_2^2 < 20d\log\frac{K}{\delta_3}$, i.e.,

$$\|[\boldsymbol{M} \quad \boldsymbol{v}_1 \quad \cdots \quad \boldsymbol{v}_K]\|_* \leq \sqrt{r\left(d^2 + 20dK\log\frac{K}{\delta_3}\right)},$$

and so we can set $\lambda_* = \sqrt{r(d^2 + 20dK\log\frac{K}{\delta_3})}$.

Taking a union bound over all of the event failures described above and taking $\delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta$ for simplicity, where δ is the same as as in Theorem 3.3, we have with probability greater than $1-5\delta$,

$$\begin{aligned}&R(\widehat{\boldsymbol{M}}, \{\widehat{\boldsymbol{v}}_k\}_{k=1}^K) - R(\boldsymbol{M}^*, \{\boldsymbol{v}_k^*\}_{k=1}^K) \\ &\leq 2L\sqrt{\frac{2\lambda_*^2\log(2d+K)}{|\mathcal{S}|}}\left[\left(4(B^2+1) + \frac{4\min(d,n)}{K}\right)\frac{\|\boldsymbol{X}\|^2}{n} + \frac{16B^3}{\sqrt{K}}\right] + \\ &\frac{4L(B^2+B)\lambda_*}{3|\mathcal{S}|}\log(2d+K) + \sqrt{\frac{8L^2\gamma^2\log(2/\delta)}{|\mathcal{S}|}},\end{aligned} \tag{43}$$

with the selection of $\lambda_*,\ B,\ \|\boldsymbol{X}\|$, and γ as described above. To arrive at an order of magnitude statement for this expression, we begin with the term under the first square root:

$$\frac{2\lambda_*^2 \log(2d+K)}{|\mathcal{S}|}\left[\left(4(B^2+1)+\frac{4\min(d,n)}{K}\right)\frac{\|\boldsymbol{X}\|^2}{n}+\frac{16B^3}{\sqrt{K}}\right]$$
$$=\frac{2r(d^2+20dK\log\frac{K}{\delta})\log(2d+K)}{|\mathcal{S}|}\left[3\left(20\log\frac{n}{\delta}+4+\frac{4\min(d,n)}{K}\right)\left(\frac{1}{d}+\frac{1}{n}+\frac{2}{dn}\log\frac{1}{\delta}\right)+\frac{16\left(5\log\frac{n}{\delta}\right)^{\frac{3}{2}}}{\sqrt{K}}\right].$$

If $K=\Omega(d^2)$ and $n\geq d$, then

$$\begin{aligned}
&3\left(20\log\frac{n}{\delta}+4+\frac{4\min(d,n)}{K}\right)\left(\frac{1}{d}+\frac{1}{n}+\frac{2}{dn}\log\frac{1}{\delta}\right)+\frac{16\left(5\log\frac{n}{\delta}\right)^{\frac{3}{2}}}{\sqrt{K}}\\
&\leq\frac{6}{d}\left(20\log\frac{n}{\delta}+4+\frac{4d}{K}\right)\left(1+\log\frac{1}{\delta}\right)+\frac{16\left(5\log\frac{n}{\delta}\right)^{\frac{3}{2}}}{\sqrt{K}}\\
&=O\left(\frac{1}{d}\right)O\left(\log n+\frac{d}{K}+1\right)+O\left(\frac{(\log n)^{\frac{3}{2}}}{\sqrt{K}}\right)\\
&=O\left(\frac{1}{d}\right)O\left(\log n+\frac{1}{d}+1\right)+O\left(\frac{(\log n)^{\frac{3}{2}}}{d}\right)\\
&=O\left(\frac{1}{d}\right)\left[O\left(\log n+(\log n)^{\frac{3}{2}}+1\right)\right]
\end{aligned}$$

where we have treated δ as a constant, and so in total the first square root term in eq. (43) scales as

$$O\left(\sqrt{\left[\frac{r(d+K\log\frac{K}{\delta})\log(2d+K)}{|\mathcal{S}|}\right]\left[\log n+(\log n)^{\frac{3}{2}}+1\right]}\right).$$

We ignore the second term in eq. (43), since it decays faster than $\sqrt{\frac{1}{|\mathcal{S}|}}$. Plugging in our selection for γ, the third term in eq. (43) scales as

$$\sqrt{\frac{7200L^2r\left(\log\frac{\max\{n,K\}}{\delta}\right)^2\log(2/\delta)}{|\mathcal{S}|}}=O\left(\sqrt{\frac{r\left(\log\max\{n,K\}\right)^2}{|\mathcal{S}|}}\right),$$

where we have treated δ as a constant. By slightly loosening each term's scaling and combining terms, we have

$$\begin{aligned}
&R(\widehat{\boldsymbol{M}},\{\widehat{\boldsymbol{v}}_k\}_{k=1}^K)-R(\boldsymbol{M}^*,\{\boldsymbol{v}_k^*\}_{k=1}^K)\\
&=O\left(\sqrt{\left[\frac{r(d+K\log\frac{K}{\delta})\log(2d+K)}{|\mathcal{S}|}\right]\left[(\log(\max\{n,K\}))^2+(\log n)^{\frac{3}{2}}+1\right]}\right).
\end{aligned}$$

Suppressing log terms, this equals

$$\widetilde{O}\left(\sqrt{\frac{rd+rK}{|\mathcal{S}|}}\right).$$

E Proofs and additional results for recovery guarantees

We can also demonstrate a recovery result in the low-rank setting.

Theorem E.1. *Assume the data is gathered as in Theorem 4.1, where $\boldsymbol{M}^*,\{\boldsymbol{v}_k^*\}_{k=1}^K$ satisfy the constraints in (7), and that (7) is solved with loss ℓ_f. With probability at least $1-\delta$*

$$\frac{1}{n}\sigma_{\min}\left(\boldsymbol{J}[\boldsymbol{X}_{\otimes}^T,\boldsymbol{X}^T]\right)^2\left(\|\widehat{\boldsymbol{M}}-\boldsymbol{M}^*\|_F^2+\frac{1}{K}\sum_{k=1}^K\|\hat{\boldsymbol{v}}_k-\boldsymbol{v}_k^*\|^2\right)$$

$$\le \frac{L}{C_f^2}\sqrt{\frac{\lambda_*^2\log(2d+K)}{2|\mathcal{S}|}\left[\left(8+\frac{4\min(d,n)}{K}\right)\frac{\|\boldsymbol{X}\|^2}{n}+\frac{16}{\sqrt{K}}\right]}+ \tag{44}$$

$$\frac{2L\lambda_*}{3C_f^2|\mathcal{S}|}\log(2d+K)+\frac{L}{C_f^2}\sqrt{\frac{\gamma^2\log(2/\delta)}{2|\mathcal{S}|}}. \tag{45}$$

To prove both Theorems 4.1 and E.1, we begin with a helpful lemma that in conjunction with Theorems 3.1 and 3.3 establishes the recovery upper bounds in Theorems 4.1 and E.1.

Lemma E.2. *In the same settings as Theorems 4.1 and E.1, we have that*

$$\frac{1}{n}\sigma_{\min}\left(\boldsymbol{J}[\boldsymbol{X}_\otimes^T,\boldsymbol{X}^T]\right)^2\left(\|\widehat{\boldsymbol{M}}-\boldsymbol{M}^*\|_F^2+\frac{1}{K}\sum_{k=1}^{K}\|\widehat{\boldsymbol{v}}_k-\boldsymbol{v}_k^*\|^2\right)$$
$$\le \frac{1}{4C_f^2}(R(\boldsymbol{M},\{\boldsymbol{v}_k\}_{k=1}^K)-R(\boldsymbol{M}^*,\{\boldsymbol{v}_k^*\}_{k=1}^K)).$$

Proof of Lemma E.2. Recall that $\ell_f(y_p,p;\boldsymbol{M},\boldsymbol{v}) = -\log(f(y_p\delta_p(\boldsymbol{M},\boldsymbol{v}_k)))$, and we have that $\mathbb{P}(y_p^{(k)}=-1)=f\left(-\delta_p(\boldsymbol{M}^*,\boldsymbol{v}_k^*)\right)$. Furthermore, recall that we have taken a uniform distribution over pairs p and users k. Hence, it is straightforward to show that we may write the excess risk of any metric $\boldsymbol{M}$ and points $\{\boldsymbol{v}_k\}_{k=1}^K$ as

$$R(\boldsymbol{M},\{\boldsymbol{v}_k\}_{k=1}^K)-R(\boldsymbol{M}^*,\{\boldsymbol{v}_k^*\}_{k=1}^K)=\frac{1}{K\binom{n}{2}}\sum_{i<j}\sum_{k=1}^{K}KL(f\left(-\delta_{ij}(\boldsymbol{M}^*,\boldsymbol{v}_k^*)\right)\|f\left(-\delta_{ij}(\boldsymbol{M},\boldsymbol{v}_k)\right)),$$

where $KL(p\,\|\,q)=p\log(p/q)+(1-p)\log((1-p)/(1-q))$. Define $\boldsymbol{\Delta}(\boldsymbol{M},\{\boldsymbol{v}_k\}_{k=1}^K)\in\mathbb{R}^{\binom{n}{2}\times K}$ such that $[\boldsymbol{\Delta}(\boldsymbol{M},\boldsymbol{v}_k)]_{p,k}=\delta_p(\boldsymbol{M},\boldsymbol{v}_k)$, where we slightly abuse notation to let p denote the row of $\boldsymbol{\Delta}$ corresponding to pair p. We have the following result (proved at the end of the section):

Proposition E.3. *Let $C_f := \min_{x:|x|\le\gamma} f'(x)$. Then,*

$$\frac{2C_f^2}{K\binom{n}{2}}\left\|\boldsymbol{\Delta}\left(\boldsymbol{M},\{\boldsymbol{v}_k\}_{k=1}^K\right)-\boldsymbol{\Delta}\left(\boldsymbol{M}^*,\{\boldsymbol{v}_k^*\}_{k=1}^K\right)\right\|_F^2\le R(\boldsymbol{M},\{\boldsymbol{v}_k\}_{k=1}^K)-R(\boldsymbol{M}^*,\{\boldsymbol{v}_k^*\}_{k=1}^K).$$

Next, define $\overline{\boldsymbol{S}}\in\{0,1\}^{\binom{n}{2}\times n}$ to be the *complete* selection matrix of all $\binom{n}{2}$ unique pairs of items such that the i,jth row is 1 in the ith column, -1 in the jth column, and 0 otherwise. Note that $\Delta(\cdot,\cdot)$ is linear in both terms. Therefore, we may factor

$$\left\|\Delta\left(\boldsymbol{M},\{\boldsymbol{v}_k\}_{k=1}^K\right)-\Delta\left(\boldsymbol{M}^*,\{\boldsymbol{v}_k^*\}_{k=1}^K\right)\right\|_F^2=\sum_{k=1}^{K}\left\|\overline{\boldsymbol{S}}[\boldsymbol{X}_\otimes^T,\boldsymbol{X}^T]\begin{bmatrix}\boldsymbol{\nu}\left(\boldsymbol{M}-\boldsymbol{M}^*\right)\\ \boldsymbol{v}_k-\boldsymbol{v}_k^*\end{bmatrix}\right\|^2.$$

Hence, we may lower bound the above as

$$\begin{aligned}
&\sum_{k=1}^{K}\left\|\overline{\boldsymbol{S}}[\boldsymbol{X}_\otimes^T,\boldsymbol{X}^T]\begin{bmatrix}\boldsymbol{\nu}\left(\boldsymbol{M}-\boldsymbol{M}^*\right)\\ \boldsymbol{v}_k-\boldsymbol{v}_k*\end{bmatrix}\right\|^2\\
&\quad\ge\sigma_{\min}\left(\overline{\boldsymbol{S}}[\boldsymbol{X}_\otimes^T,\boldsymbol{X}^T]\right)^2\sum_{k=1}^{K}\left\|\begin{bmatrix}\boldsymbol{\nu}\left(\boldsymbol{M}-\boldsymbol{M}^*\right)\\ \boldsymbol{v}_k-\boldsymbol{v}_k^*\end{bmatrix}\right\|^2\\
&\quad=\sigma_{\min}\left(\overline{\boldsymbol{S}}[\boldsymbol{X}_\otimes^T,\boldsymbol{X}^T]\right)^2\left(K\|\boldsymbol{\nu}(\boldsymbol{M}-\boldsymbol{M}^*)\|^2+\sum_{k=1}^{K}\|\boldsymbol{v}_k-\boldsymbol{v}_k^*\|^2\right)\\
&\quad\ge\sigma_{\min}\left(\overline{\boldsymbol{S}}[\boldsymbol{X}_\otimes^T,\boldsymbol{X}^T]\right)^2\left(K\|\boldsymbol{M}-\boldsymbol{M}^*\|_F^2+\sum_{k=1}^{K}\|\boldsymbol{v}_k-\boldsymbol{v}_k^*\|^2\right).
\end{aligned}$$

where $\sigma_{\min}(\boldsymbol{A})$ denotes the smallest singular value of a matrix $\boldsymbol{A}$ and the final inequality follows from the fact that for symmetric $d\times d$ matrix $\boldsymbol{A}$,

$$\|\boldsymbol{\nu}(\boldsymbol{A})\|_2^2=\|\mathrm{vec}^*(2\boldsymbol{A}-\boldsymbol{I}\odot\boldsymbol{A})\|_2^2=\sum_{j\ge i}((2-\mathbb{1}_{i=j})\boldsymbol{A}_{i,j})^2$$

$$
\begin{aligned}
&= 4\sum_{j>i} \boldsymbol{A}_{i,j}^2 + \sum_i \boldsymbol{A}_{i,i}^2 \\
&\geq 2\sum_{j>i} \boldsymbol{A}_{i,j}^2 + \sum_i \boldsymbol{A}_{i,i}^2 \\
&= \sum_{i>j} \boldsymbol{A}_{i,j}^2 + \sum_{i<j} \boldsymbol{A}_{i,j}^2 + \sum_i \boldsymbol{A}_{i,i}^2 \\
&= \|\boldsymbol{A}\|_F^2.
\end{aligned}
$$

Since $n \geq D + d + 1$, surely the selection matrix of all possible paired comparisons $\overline{\boldsymbol{S}}$ contains the construction presented in Appendix C.6 which satisfies the conditions of Proposition 2.2, and so $\overline{\boldsymbol{S}}[\boldsymbol{X}_\otimes^T, \boldsymbol{X}^T]$ is a tall matrix that is full column rank if the items $\boldsymbol{x}_i$ are drawn i.i.d. according to a distribution that is absolutely continuous with respect to the Lebesgue measure, in which case $\sigma_{\min}\left(\overline{\boldsymbol{S}}[\boldsymbol{X}_\otimes^T, \boldsymbol{X}^T]\right)^2 > 0$. To simply this expression even further, we have the following result (proved at the end of the section):

Proposition E.4. *For $\overline{\boldsymbol{S}} \in \mathbb{R}^{\binom{n}{2} \times n}$, $\overline{\boldsymbol{S}}^T \overline{\boldsymbol{S}} = n\boldsymbol{J}$ for $\boldsymbol{J} := \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$.*

Therefore,

$$
\sigma_{\min}\left(\overline{\boldsymbol{S}}[\boldsymbol{X}_\otimes^T, \boldsymbol{X}^T]\right)^2 = n\lambda_{\min}\left(\begin{bmatrix}\boldsymbol{X}_\otimes \\ \boldsymbol{X}\end{bmatrix}\boldsymbol{J}^2[\boldsymbol{X}_\otimes^T, \boldsymbol{X}^T]\right) = n\sigma_{\min}\left(\boldsymbol{J}[\boldsymbol{X}_\otimes^T, \boldsymbol{X}^T]\right)^2.
$$

Finally, note that,

$$
\frac{2nC_f^2}{K\binom{n}{2}} \geq \frac{4C_f^2}{Kn}.
$$

The proof follows by rearranging terms. □

Proof of Proposition E.3. By Lemma 5.2 of [12], for $(y, z) \in (0, 1)$, $KL(y||z) \geq 2(y - z)^2$. Now, let $y = f(x)$ and $z = f(x')$, for a continuously differentiable function f. Then $2(y - z)^2 \geq 2(\min_a f'(a))^2(x - x')^2$ since f is monotonic. Applying this to the decomposition of the excess risk alongside the definition of $\Delta(\boldsymbol{M}, \boldsymbol{v}_k)$ establishes the result. □

Proof of Proposition E.4. Note that for $\overline{\boldsymbol{S}} \in \mathbb{R}^{\binom{n}{2} \times n}$, by construction the i^{th} column corresponds to item i and each row maps to a pair (e.g., (i, j)) of the possible $\binom{n}{2}$ unique pairs such that exactly 2 elements are non-zero in each row, with a 1 in the column of one item in the pair and a -1 in the other. Since $\overline{\boldsymbol{S}}^T\overline{\boldsymbol{S}}$ is a Gram matrix, it is sufficient to characterize the inner products between any two columns of $\overline{\boldsymbol{S}}$. First, for the i^{th}, note that i can be paired with $n - 1$ other items uniquely. Hence, there are exactly $n - 1$ non-zero entries in each column all of which are 1 or -1. Hence, every diagonal entry of $\overline{\boldsymbol{S}}^T\overline{\boldsymbol{S}}$ is $n - 1$. For the off diagonal entries, consider a pair $i \neq j$. As each row corresponds to a unique pair, the supports of i and j overlap in a single entry corresponding to the (i, j) pair. By construction one column has a 1 at this entry and the other has a -1. Hence the inner product between these two columns is -1 and all off diagonal entries of $\overline{\boldsymbol{S}}^T\overline{\boldsymbol{S}}$ are -1. Hence, $\overline{\boldsymbol{S}}^T\overline{\boldsymbol{S}} = n\boldsymbol{I} - \mathbf{1}_n\mathbf{1}_n^T = n\boldsymbol{J}$. □

F Additional experimental details

In this section we provide additional experimental details and results.

F.1 Datasets

The color preference data was originally collected by [24] and we include a .mat file of the dataset in the paper supplement along with a full description in the code README document. Before running our learning algorithms, we centered the 3×37 item matrix $\boldsymbol{X}$ of CIELAB coordinates, and normalized the centered coordinates by the magnitude of the largest norm color in CIELAB space such that $\max_{i\in[n]}\|\boldsymbol{x}_i\|_2 = 1$ after centering and normalization.

For both the normally distributed items and color experiments, we performed train and test dataset splits over multiple simulation runs, and averaged results across each run. For each simulation run, we *blocked* the train/test splitting by user, in that all users were queried equally in both training and test data. Specifically, during each run the dataset was randomly shuffled *within* each user's responses, and then a train/test split created. For the normally distributed data, this consisted of 300 training comparisons per user, and 300 test comparisons. For the color dataset, each user provided $37 \times 36 = 1332$ responses, which was partitioned into a training set of 300 pairs and a test set of 1032 pairs. To vary the number of training pairs per user, for both datasets we trained incrementally on the 300 pairs per user in a randomly permuted order (while evaluating on the full test set).

For both normally distributed items and color preference data, we repeat 30 independent trials. In the color preference data, since the dataset is fixed ahead of time the only difference between each trial is the train/test splitting as detailed above. In the normally distributed experiments, we generate responses as $\mathbb{P}(y_{i,j}^{(k)} = -1) = (1 + e^{\beta(\boldsymbol{x}_i^T\boldsymbol{M}^*\boldsymbol{x}_i - \boldsymbol{x}_j^T\boldsymbol{M}^*\boldsymbol{x}_j + (\boldsymbol{v}_k^*)^T(\boldsymbol{x}_i - \boldsymbol{x}_j))})^{-1}$ for a noise scaling parameter $\beta > 0$.

F.2 Implementation and computation

In out experiments, we did not enforce the γ constraints required for our theoretical results (i.e., the γ constraints in eqs. (5) and (7)). This constraint is added to the theory to guard against highly coherent $\boldsymbol{x}_i$ vectors. In the simulated instances, this quantity appears to be controlled by the isotropic nature of both the normally distributed and color datasets, along with the fact that we are constraining the norms of the latent parameters to be learned.

For all simulated experiments, we leveraged ground-truth knowledge of $\boldsymbol{M}^*$ and $\boldsymbol{v}_k^*$ to set the hyperparameter constraints. This was done to compare each method under its best possible hyperparameter tuning — namely, the smallest norm balls that still contained the true solution. Specifically, we set hyperparameters for the normally distributed items experiment as follows:

- **Frobenius metric**: $\|\boldsymbol{M}\|_F \leq \|\boldsymbol{M}^*\|_F$, for all $k \in [K]$, $\|\boldsymbol{v}_k\|_2 \leq 2\max_{k\in[K]}\|\boldsymbol{M}^*\boldsymbol{u}_k^*\|$
- **Nuclear full**: $\|[\boldsymbol{M}, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_K]\|_* \leq \|[\boldsymbol{M}^*, -2\boldsymbol{M}^*\boldsymbol{u}_1^*, \cdots, -2\boldsymbol{M}^*\boldsymbol{u}_K^*]\|_*$
- **Nuclear metric**: $\|\boldsymbol{M}\|_* \leq \|\boldsymbol{M}^*\|_*$, for all $k \in [K]$, $\|\boldsymbol{v}_k\|_2 \leq 2\max_{k\in[K]}\|\boldsymbol{M}^*\boldsymbol{u}_k^*\|$
- **Nuclear split**: $\|\boldsymbol{M}\|_* \leq \|\boldsymbol{M}^*\|_*$, $\|[\boldsymbol{v}_1, \cdots, \boldsymbol{v}_K]\|_* \leq 2\|\boldsymbol{M}^*[\boldsymbol{u}_1^*, \cdots, \boldsymbol{u}_K^*]\|_*$
- **Nuclear full, single**: for each $k \in [K]$, $\|[\boldsymbol{M}, \boldsymbol{v}_k]\|_* \leq \|[\boldsymbol{M}^*, -2\boldsymbol{M}^*\boldsymbol{u}_k^*]\|_*$.

For the color preferences experiment, we set all hyperparameters under an a priori estimate of $\boldsymbol{M}^* = \boldsymbol{I}$ (due to the assumed perceptual uniformity of CIELAB space). Specifically, we constrained $\|\boldsymbol{M}\|_F \leq \|\boldsymbol{I}\|_F = \sqrt{3}$ (since $d = 3$), and constrained $\|\boldsymbol{v}_k\|_2 \leq 2$, since under the heuristic assumption that $\boldsymbol{M}^* = \boldsymbol{I}$ we have $\|\boldsymbol{v}_k\|_2 = \|-2\boldsymbol{M}\boldsymbol{u}_k\|_2 = 2\|\boldsymbol{u}_k\| \lessapprox 2\max_{i\in[n]}\|\boldsymbol{x}_i\|_2 = 2$, where in the last inequality we have approximated the distribution of ideal points $\boldsymbol{u}_k$ with the empirical item distribution over centered, scaled CIELAB colors (which have maximal norm of 1 as described above).

To solve these optimizations, we leveraged CVXPY[9] with different solvers. When learning on normally distributed items, we set $\ell(x) = \log(1 + \exp(-\beta x))$ to be the logistic loss, where $\beta > 0$ is the same parameter used to generate the response noise. We used the Splitting Conic Solver with a convergence tolerance set to $1e - 6$ to balance between accuracy and computation time. For the color preference data, we used the hinge loss $\ell(x) = \max\{0, 1 - x\}$, solved using the CVXOPT solver with default parameters, which we found performed more stably than SCS. As an additional safeguard for numerical stability due to the presence of negative eigenvalues near machine precision, we project all learned metrics back onto the positive semidefinite cone after solving with CVXPY. All additional code was written in Python, and all experiments were computed on three Dell 740 servers with 36, 3.1 GHz Xeon Gold 6254 CPUs.

[9] `https://www.cvxpy.org/`

When estimating ideal points $\widehat{\boldsymbol{u}}_k$ from $\widehat{\boldsymbol{M}}$ and $\widehat{\boldsymbol{v}}_k$, rather than using the exact pseudo-inverse we perform a regularized recovery as in [7], since $\widehat{\boldsymbol{M}}$ may have recovery errors. Specifically, for regularization parameter $\alpha > 0$ we estimate $\widehat{\boldsymbol{u}}_k$ as

$$\widehat{\boldsymbol{u}}_k = -2(4\widehat{\boldsymbol{M}}^2 + \alpha \boldsymbol{I})^{-1}\widehat{\boldsymbol{M}}^T\widehat{\boldsymbol{v}}_k. \tag{46}$$

We only perform ideal point recovery in the normally distributed item experiments, since no ground-truth ideal points are available in the real-world color preference data. Since we know a priori that $\boldsymbol{u}_k \sim \mathcal{N}(\boldsymbol{0}, \frac{1}{d}\boldsymbol{I})$, we can leverage the interpretation of the recovery estimate in eq. (46) as the maximum a posteriori estimator under a Gaussian prior over $\boldsymbol{u}_k$ with Gaussian observations in order to set $\alpha = d$.

F.3 Additional experiments and details

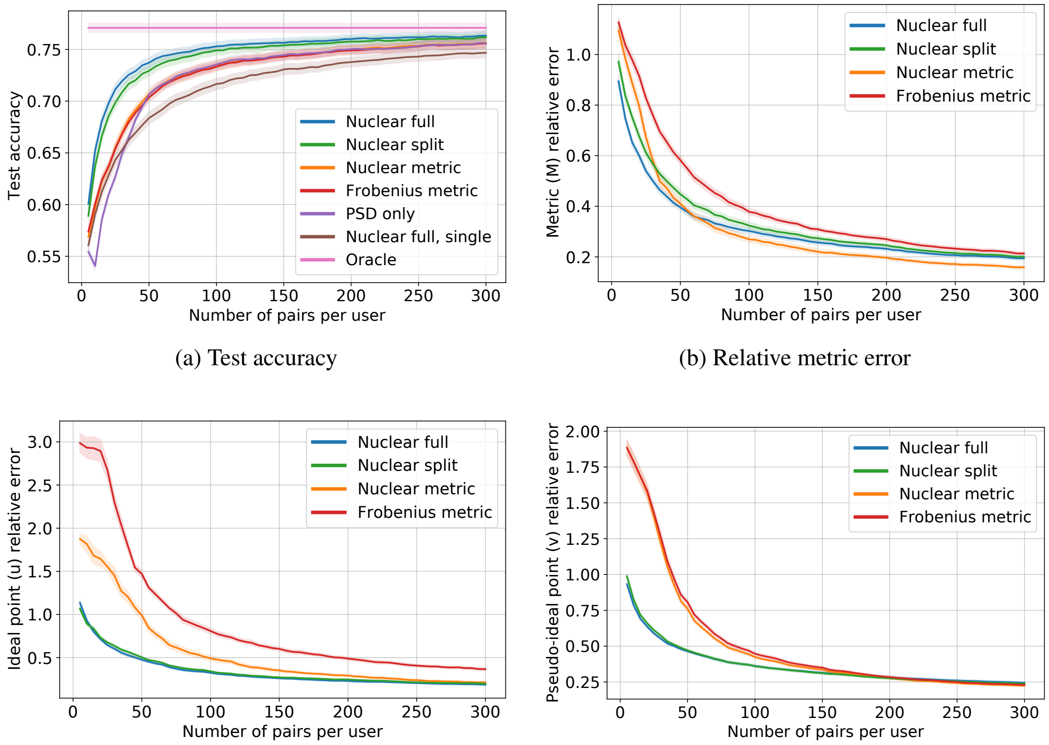
Below we present experimental results in additional simulation settings, as well as supplementary figures for the data presented in the main paper body. We detail specific performance metrics below: in the following, let $\widehat{\boldsymbol{V}} := [\widehat{\boldsymbol{v}}_1, \cdots, \widehat{\boldsymbol{v}}_K]$, $\boldsymbol{U}^* := [\boldsymbol{u}_1^*, \cdots, \boldsymbol{u}_K^*]$ $\boldsymbol{V}^* := -2\boldsymbol{M}^*\boldsymbol{U}^*$, and $(\boldsymbol{M}^*)^\dagger$ denote the pseudoinverse of the ground-truth metric.

- *Test accuracy*: fraction of test data responses predicted correctly from $\mathrm{sign}(\widehat{\delta}_{i,j}^{(k)})$, where $\widehat{\delta}_{i,j}^{(k)}$ is computed as in eq. (2) using $\widehat{\boldsymbol{M}}$ and $\widehat{\boldsymbol{v}}_k$ as parameter estimates.
- *Relative metric error*: $\frac{\|\widehat{\boldsymbol{M}}-\boldsymbol{M}^*\|_F}{\|\boldsymbol{M}^*\|_F}$
- *Relative ideal point error*: $\frac{\|\widehat{\boldsymbol{U}}-(\boldsymbol{M}^*)^\dagger\boldsymbol{M}^*\boldsymbol{U}^*\|_F}{\|(\boldsymbol{M}^*)^\dagger\boldsymbol{M}^*\boldsymbol{U}^*\|_F}$. We compare recovery error against $(\boldsymbol{M}^*)^\dagger\boldsymbol{M}^*\boldsymbol{U}^*$ rather than $\boldsymbol{U}^*$ since if $\boldsymbol{M}^*$ is low-rank, then the components of $\boldsymbol{U}^*$ in the kernel of $\boldsymbol{M}^*$ are not recoverable.
- *Relative pseudo-ideal point error*: $\frac{\|\widehat{\boldsymbol{V}}-\boldsymbol{V}^*\|_F}{\|\boldsymbol{V}^*\|_F}$

In Figure 1e, we compute the heatmap of the crowd's metric as the empirical average of the learned metrics over all independent trials.

We present additional simulation results with normally distributed items in two noise regimes — "high" noise with $\beta = 1$ in the logistic model, and "medium" noise with $\beta = 4$ in the logistic model. We generate the dataset in the same manner as in Section 5. We present results for the low-rank case as in the main paper body ($d = 10$, $r = 1$) as well as the full-rank case ($d = r = 10$). In the full-rank case, to generate a ground-truth metric $\boldsymbol{M}^*$ we generate a $d \times r$ matrix $\boldsymbol{L}$ whose entries are sampled independently according to the standard normal distribution, compute $\boldsymbol{M}^* = \boldsymbol{L}\boldsymbol{L}^T$, and normalize $\boldsymbol{M}^*$ such that it has a Frobenius norm of d. Otherwise, if we generated $\boldsymbol{L}$ as in the low-rank experiments, $\boldsymbol{M}^*$ would simply become a scaled identity matrix.

Fig. 2 repeats the main results in the paper body, with an additional subfigure depicting recovery error for pseudo-ideal points. This is a "high" noise, low-rank setting. The remaining figures are: "high" noise full-rank metric (Fig. 3); "medium" noise low-rank metric (Fig. 4); and "medium" noise full-rank metric (Fig. 5). In Fig. 6 we analyze the color prediction results from Fig. 1d in the low query count regime.



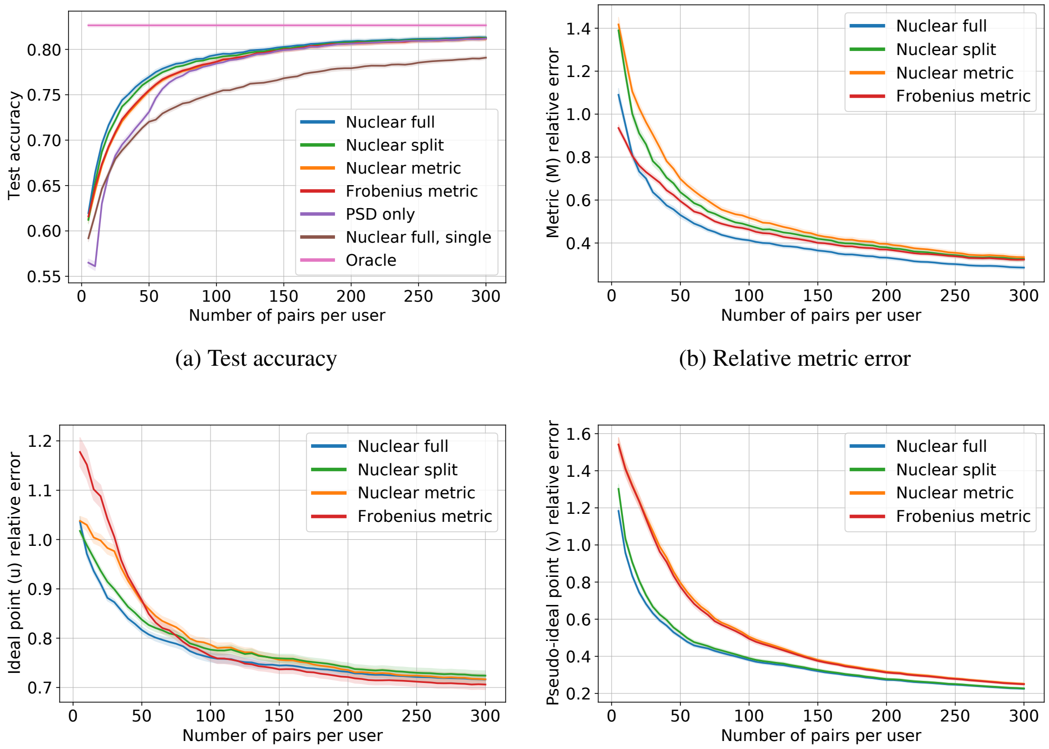
(a) Test accuracy

(b) Relative metric error

(c) Relative ideal point error

(d) Relative pseudo-ideal point error

Figure 2: Full prediction and recovery results for a high noise setting ($\beta = 1$) with a low-rank metric ($r = 1$). Error bars indicate ± 1 standard error about the sample mean. (a-c) appear in the main paper body, with (d) added here for completeness. **Nuclear full** gives the best performance on prediction and **Nuclear split** is a close second (subfigure a), reflecting the necessity of modeling the low-rank nature of the $\boldsymbol{M}^*$ and $\boldsymbol{V}^*$. While the **Nuclear metric** method that only places a nuclear norm constraint on $\boldsymbol{M}$ performs well in terms of relative error for recovering $\boldsymbol{M}^*$ (subfigure b), it achieves far worse performance for estimating $\boldsymbol{U}^*$ and $\boldsymbol{V}^*$ as shown in subfigures (c) and (d). This reflects the importance of enforcing that $\widehat{\boldsymbol{M}}$ and $\widehat{\boldsymbol{V}}$ share a column space.



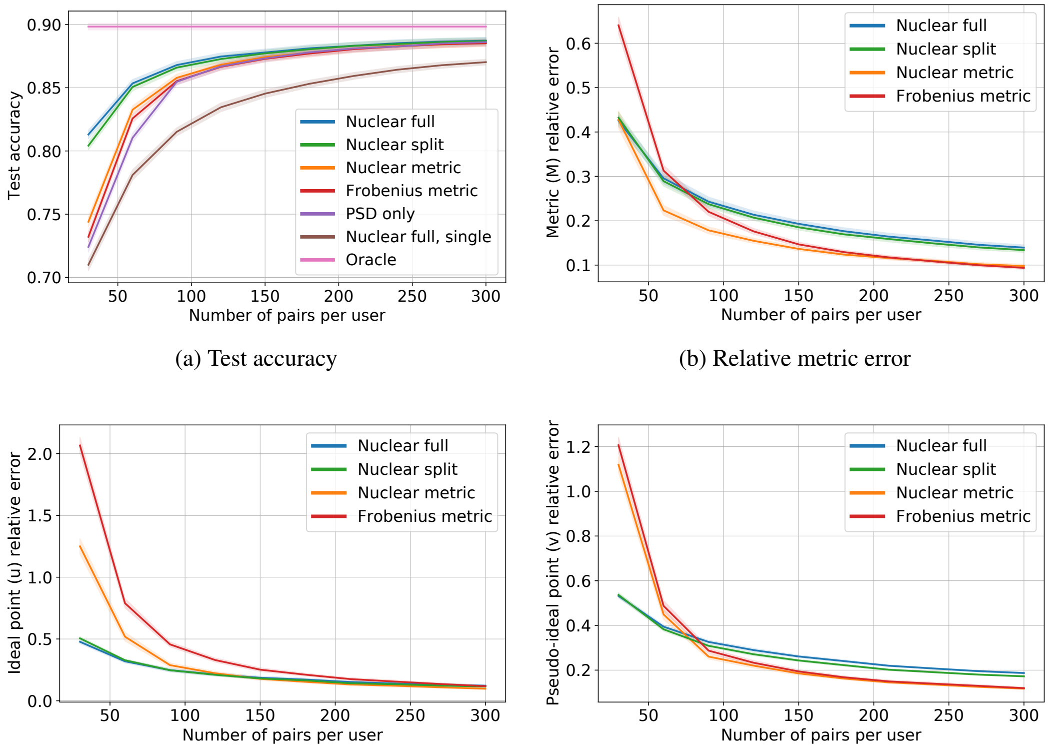
(a) Test accuracy

(b) Relative metric error

(c) Relative ideal point error

(d) Relative pseudo-ideal point error

Figure 3: Prediction and recovery results for a high noise setting ($\beta = 1$) with a full-rank metric ($d = r = 10$). Error bars indicate ± 1 standard error about the sample mean. Surprisingly, even in the full-rank scenario, **Nuclear full** demonstrates the highest prediction performance, even in comparison to **Frobenius metric** which is designed for full-rank metrics. That said, the difference is less stark for estimating both $\boldsymbol{M}^*$ and $\boldsymbol{U}^*$.



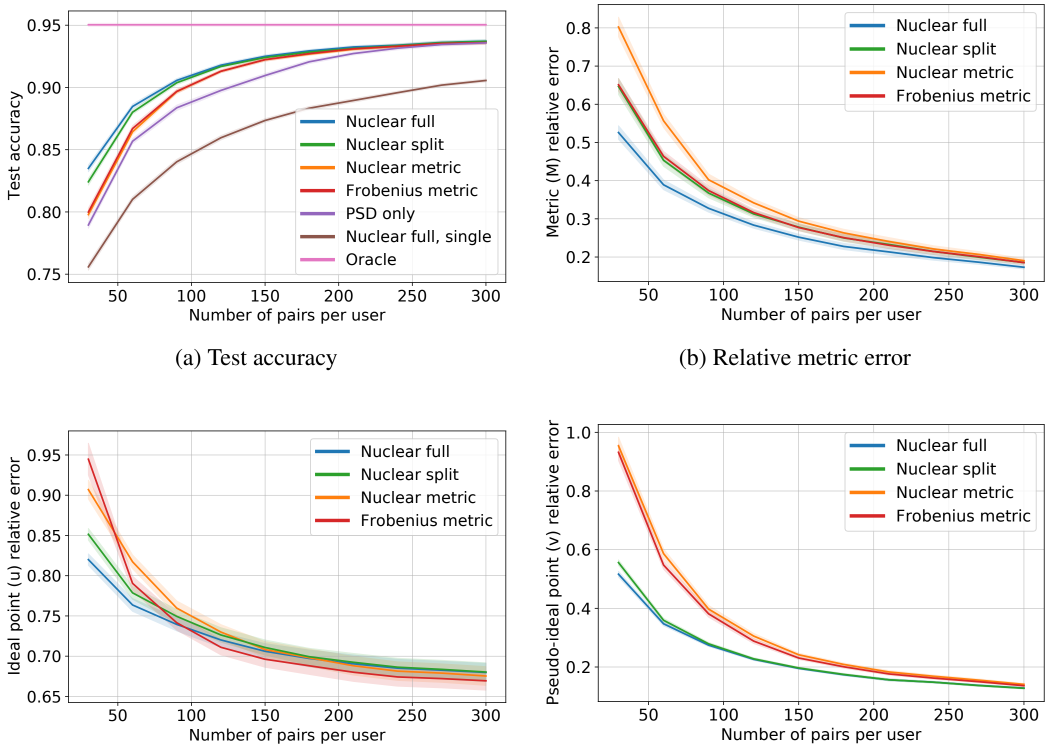
(a) Test accuracy

(b) Relative metric error

(c) Relative ideal point error

(d) Relative pseudo-ideal point error

Figure 4: Prediction and recovery results for a medium noise setting ($\beta = 4$) with a low-rank metric ($r = 1$). Error bars indicate ± 1 standard error about the sample mean. Similar trends as Figure 2 hold except that the differences are less pronounced. Interestingly, for large numbers of samples per user, **Nuclear metric** and **Frobenius metric** appear to achieve better performance on estimating $\boldsymbol{M}^*$ and therefore achieve better performance for estimating $\boldsymbol{V}^*$, though they do not achieve as strong of performance for estimating $\boldsymbol{U}^*$, perhaps because these methods do not enforce that $\widehat{\boldsymbol{M}}$ and $\widehat{\boldsymbol{V}}$ share a column space.



(a) Test accuracy

(b) Relative metric error

(c) Relative ideal point error

(d) Relative pseudo-ideal point error

Figure 5: Prediction and recovery results for a medium noise setting ($\beta = 4$) with a full-rank metric ($d = r = 10$). Error bars indicate ± 1 standard error about the sample mean. Similar trends as Figure 3 hold in this case.

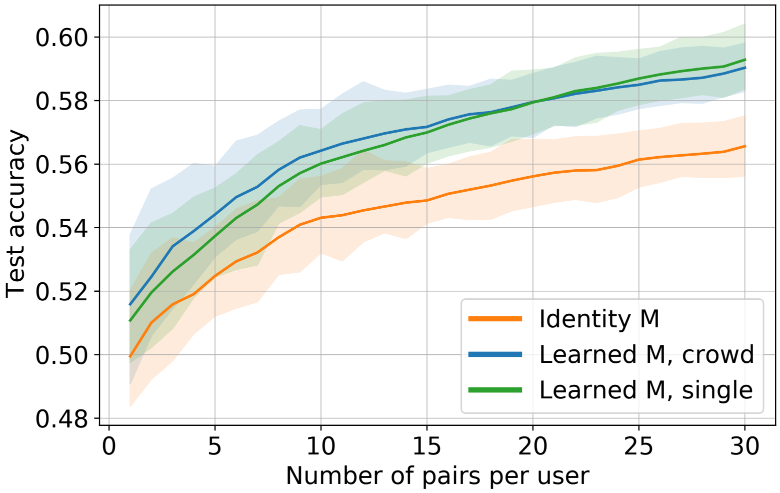


Figure 6: Comparison of methods on color preference data in the low query regime, with error bars representing 2.5% and 97.5% percentiles. The identity metric performs more poorly than the methods that learn a metric tuned to user judgments. The method that learns a single M for the crowd and method that learns a M for each individual perform similarly in most of the range of number of pairs. There is a slight advantage to the method that learns a metric for the crowd when very few pairs have been given to each user. This stems from the fact that the crowd metric can amortize the cost of learning the metric over the responses given by all users.

F.4 Comparison with [7] in the single user case

Although the focus of this work is developing and analyzing algorithms for simultaneous metric and preference learning over *multiple* users, for completeness we compare the performance of our approach against the algorithms presented in [7], which only study simultaneous metric and preference learning in the *single* user case ($K = 1$, in our notation). Xu and Davenport propose both “single-step” and “alternating” estimation algorithms, referred to here as **XDsingle** and **XDalt** respectively (we direct the reader to [7] for the details of these algorithms). Since [7] does not specifically consider the low-rank metric setting, we evaluate these methods on the full-rank simulated data described in Fig. 3. To run their algorithms, we use the default hyperparameter values described in their paper and found in their code (see [7] for details and code link).

In the table below we evaluate their algorithms' performance using the same procedure as in Fig. 3, at sampled intervals of 50 queries per user. For reference, we also copy the corresponding values from **Nuclear full, single** in Fig. 3. We evaluate **XDsingle** and **XDalt** in an analogous manner to **Nuclear full, single**: the crowd of $K = 10$ simulated users generates responses according to a common metric, but the single user algorithms are solved separately, learning a separate metric for each user. Test accuracy, metric recovery error, and ideal point recovery error are then averaged over all users.

Number of queries per user	50	100	150	200	250	300
Nuclear full, single	0.720	0.752	0.768	0.779	0.785	0.791
XDsingle	0.641	0.682	0.701	0.712	0.727	0.731
XDalt	0.683	0.715	0.737	0.745	0.757	0.762

(a) Test accuracy

Number of queries per user	50	100	150	200	250	300
Nuclear full, single	1.134	0.975	0.874	0.797	0.747	0.704
XDsingle	1.003	1.889	1.816	1.513	1.233	1.085
XDalt	0.791	0.757	0.733	0.715	0.680	0.674

(b) Relative metric error

Number of queries per user	50	100	150	200	250	300
Nuclear full, single	0.985	0.879	0.841	0.819	0.794	0.782
XDsingle	1.000	0.963	0.910	0.882	0.862	0.866
XDalt	0.922	0.906	0.887	0.887	0.875	0.875

(c) Relative ideal point error

Table 1: Comparison of **Nuclear full, single** against the methods proposed in [7] for the *single* user case. We evaluate prediction accuracy and recovery error (averaged over 30 trials) for a high noise setting ($\beta = 1$) with a full-rank metric ($d = r = 10$). Simulation details are otherwise identical to those in Fig. 1a-c.

Overall, among the single user methods **Nuclear full, single** recovers the highest test prediction accuracy and lowest ideal point recovery error, while **XDalt** results in the lowest metric estimation error. While our work and [7] share the same model in the single user case (as a reminder, [7] does not consider the multi-user case), each frames the problem differently and solves a different optimization problem, so it is intuitively unclear how each single user algorithm differs theoretically.