BLIND REGRESSION VIA NEAREST NEIGHBORS UNDER LATENT VARIABLE MODELS

BY CHRISTINA LEE, YIHUA LI, DEVAVRAT SHAH AND DOGYOON SONG

Massachusetts Institute of Technology

We consider the setup of nonparametric *blind regression* for estimating the entries of a large $m \times n$ matrix, when provided with a small, random fraction of noisy measurements. We assume that all rows $u \in [m]$ and columns $i \in [n]$ of the matrix are associated to latent features $x_1(u)$ and $x_2(i)$ respectively, and the (u, i)-th entry of the matrix, $A(u, i)$ is equal to $f(x_1(u), x_2(i))$ for a latent function f. Given noisy observations of a small, random subset of the matrix entries, our goal is to estimate the unobserved entries of the matrix as well as to "de-noise" the observed entries.

As the main result of this work, we introduce a nearest neighbor based estimation algorithm inspired by the classical Taylor's series expansion. We establish its consistency when the underlying latent function f is Lipschitz, the latent features belong to a compact domain, and the random fraction of observed entries in the matrix is at least $\max\left(m^{-1+\delta}, n^{-1/2+\delta}\right)$, for any $\delta > 0$. As an important byproduct, our analysis sheds light into the performance of the classical collaborative filtering algorithm for matrix completion, which has been widely utilized in practice. Experiments with the MovieLens and Netflix datasets suggest that our algorithm provides a principled improvement over basic collaborative filtering and is competitive with matrix factorization methods.

Our algorithm has a natural extension to the setting of tensor completion. For a t-order balanced tensor with total of N entries, we prove that our approach provides a consistent estimator when at least $N^{-\frac{\lfloor 2t/3 \rfloor}{2t} + \delta}$ fraction of entries are observed, for any $\delta > 0$. When applied to the setting of image inpainting, which is a 3-order tensor, we find that our approach is competitive with respect to state-of-art tensor completion algorithms across benchmark images.

CONTENTS

1. Introduction. The problem of matrix completion has received enormous attention in the past decade: Consider an $m \times n$ matrix A of interest. Suppose we observe a random subset of the entries of an $m \times n$ matrix Z, which is a noisy version of A, such that each (u, i)-th entry $Z(u, i)$ is a random variable with $\mathbb{E}[Z(u, i)] = A(u, i)$. For each $u \in [m]$ and $i \in [n]$, $Z(u, i)$ is observed with probability p and with probability $1 - p$ it is not observed, independently of all other entries. The goal is to recover matrix A given partial observations from Z.

1.1. *Related Literature.* The primary methods used to solve this problem in the literature include neighbor based approaches, such as collaborative filtering, and spectral approaches, which include low-rank matrix factorization or minimization of a loss function with respect to spectral constraints.

Spectral Methods. In the recent years, there have been exciting intellectual developments in the context of spectral approaches such as matrix factorization. All matrices admit a singular-value decomposition, such that they can be uniquely factorized. The goal of the factorization based method is to recover row and column singular vectors accurately from the partially observed, noisy matrix Z and subsequently estimate the matrix A. Srebro, Alon and Jaakkola (2004) was one of the earliest works to suggest the use of low-rank matrix approximation in this context. Subsequently, statistically efficient approaches were suggested using optimization based estimators, proving that matrix factorization can fill in the missing entries with sample complexity as low as $rn \log n$, where r is the rank of the matrix Candès and Recht (2009); Rohde et al. (2011); Keshavan, Montanari and Oh (2009); Negahban and Wainwright (2012); Jain, Netrapalli and Sanghavi (2013). There has been an exciting line of ongoing work to make the resulting algorithms faster and scalable Fazel, Hindi and Boyd (2003); Liu and Vandenberghe (2010); Cai et al. (2008); Lin et al. (2009); Shen, Ji and Ye (2009); Mazumder, Hastie and Tibshirani (2010a).

Xu, Massoulié and Lelarge (2014) proposed a spectral clustering method for inferring the edge label distribution for a network sampled from a generalized stochastic block model. The model is similar to the proposed latent variable model introduced in Section 2, except that the edges are labeled by one of finitely many labels in symmetric setup with $m = n$, and the goal is to estimate the label distribution in addition to the expected label. When the expected function has a finite spectrum decomposition, i.e. low rank, then they provide a consistent estimator for the sparse data regime, with $\Omega(m \log m)$ samples. When the function is only approximately low rank (e.g. the class of general Lipschitz functions), for a fixed rank r approximation, the mean squared error bounds converge to a positive constant which captures the low rank approximation gap. That is, $\Omega(m \log m)$ samples are not sufficient to guarantee consistent estimation for the entire class of Lipschitz

functions.

Many of these approaches are based on the structural assumption that the underlying matrix is *low-rank* and the matrix entries are reasonably "incoherent". Unfortunately, the low-rank assumption may not hold in practice. The recent work Ganti, Balzano and Willett (2015) makes precisely this observation, showing that a simple non-linear, monotonic transformation of a low-rank matrix could easily produce an effectively high-rank matrix, despite few free model parameters. They provide an algorithm and analysis specific to the form of their model, which achieves sample complexity of $O((mn)^{2/3})$. However, their algorithm only applies to functions f which are a nonlinear monotonic transformation of the inner product of the latent features. Lee et al. (2016) propose an algorithm for estimating locally low rank matrices, however their algorithm assumes prior knowledge of the "correct" kernel function between pairs of rows and columns which is not known a priori.

In contrast, Chatterjee (2015) proposes the universal singular value thresholding estimator (USVT) inspired by low-rank matrix approximation. Somewhat interestingly, he argues that under the latent variable model considered in this work (see Section 2), the USVT algorithm provides an accurate estimate for any Lipschitz function. However, to guarantee consistency of the USVT estimator for an $m \times m$ (i.e. $m = n$) matrix, it requires observing $\Omega\left(m^{\frac{2(d+1)}{(d+2)}}\right)$ many entries out of the m^2 total entries, where d is the dimension of the latent space in which the row and column latent features belong. In contrast, our algorithm requires observing $\Omega(m^{\frac{3}{2}+\delta})$ entries of the matrix for any small $\delta > 0$, which is independent of the dimension of the latent space, as long as $d = o(\log m)$ (c.f. Corollary 4.3).

Collaborative Filtering. The term collaborative filtering was coined by Goldberg et al. (1992), and this technique is widely used in practice due to its simplicity and ability to scale. There are two main paradigms in neighborhood-based collaborative filtering: the user-user paradigm and the item-item paradigm. To recommend items to a user in the user-user paradigm, one first looks for similar users, and then recommends items liked by those similar users. In the item-item paradigm, in contrast, items similar to those liked by the user are found and subsequently recommended. Much empirical evidence exists that the item-item paradigm performs well in many cases (Linden, Smith and York, 2003; Koren and Bell, 2011; Ning, Desrosiers and Karypis, 2015). There have also been many heuristic improvements upon the basic algorithm, such as normalizing the data, combining neighbor methods with spectral methods, combining both user and item neighbors, and additionally optimizing over interpolation weights given to each datapoint within the neighborhood when computing the final prediction Bell and Koren (2007); Koren (2008); Wang, de Vries and Reinders (2006).

Despite the widespread success of similarity based collaborative filtering heuristics, the theoretical understanding of these method is very limited. In recent works, latent mixture models have been introduced to explain the collaborative filtering algorithm as well as the empirically observed superior performance of item-item paradigms, c.f. Bresler, Chen and Shah (2014); Bresler, Shah and Voloch (2015). However, these results assume binary ratings and a specific parametric model, such as a mixture distribution model for preferences across users and movies. We hope that by providing an analysis for collaborative filtering within a nonparametric model, we can provide a better understanding of collaborative filtering.

Within the context of dense graphon estimation, when the entries in the data matrix are binary and the sample probability p is constant $O(1)$, there have been a few theoretical results that prove convergence of the mean squared error for similarity based methods Airoldi, Costa and Chan (2013); Zhang, Levina and Zhu (2015). They hinge upon computing similarities between rows or columns by comparing commonly observed entries, similar to collaborative filtering. Similar to our result, they are able to prove convergence for the class of Lipschitz functions. However, Airoldi, Costa and Chan (2013) assumes that the algorithm is given multiple instances of the sampled dataset, which is not available in our formulation. Zhang, Levina and Zhu (2015) is weaker than our result in that it assumes $p = O(1)$, however they are able to handle a more general noise model, when the entries are binary. The similarity between a pair of vertices is computed from the maximum difference between entries in the associated rows of the second power of the data matrix, which is computationally more expensive than directly comparing rows in the original data matrix.

Tensor Completion.   Recently there have been efforts to extend decomposition methods or neighborhood based approaches to the context of tensor completion, however this has proven to be significantly more challenging than matrix completion due to the complication that tensors do not have a canonical decomposition such as the singular value decomposition (SVD) for a matrix. This property makes obtaining a decomposition for a tensor challenging. The survey Kolda and Bader (2009) elaborates on these challenges. There have been recent developments in obtaining efficient tensor decompositions in form of rank-1 tensors (tensors obtained from one vector), presented in Anandkumar et al. (2014). This has been especially effective in learning latent variable models and estimating missing data as shown in, for example Jain and Oh (2014); Oh and Shah (2014). Beyond tensor decomposition, there have been recent developments in the context of learning latent variable models or mixture distributions also called non-negative matrix factorization, c.f. Arora, Ge and Moitra (2012); Arora et al. (2012).

1.2. *Our Contributions.* We provide a similarity based collaborative filtering algorithm with theoretical performance guarantees under the latent variable model. In addition, we extend the algorithm and results to the more challenging setting of tensor completion. To our knowledge, this is the first theoretical analysis for similarity based collaborative filtering algorithms, shedding insight into the widespread success of this popular heuristic for the past two decades. The algorithm we introduce is a simple variant of classical collaborative filtering, in which we compute similarities between pairs of rows and pairs of columns by comparing their common overlapped entries. Our model assumes that each row and column is associated to a latent variable, i.e. hidden features, and that the data is in expectation equal to some unknown function of those latent variables. The key regularity condition that we require is that the function is Lipschitz with respect to the latent space.

Given this latent variable model, we prove that the estimate produced by this algorithm is consistent as long as the fraction of entries that are observed is at least $\max(m^{-1+\delta}, n^{-1/2+\delta})$ for some $\delta > 0$ for an $m \times n$ matrix. We provide experiments using our method to predict ratings in the MovieLens and Netflix datasets. The results suggest that our algorithm improves over basic collaborative filtering and is competitive with factorization based methods.

We also show that the algorithm can be applied to tensor completion by flattening the tensor to a matrix. Additionally we can extend the analysis to prove that our algorithm produces a consistent estimator when at least $N^{-1/3+\delta}$ fraction of the entries are observed, for some $\delta > 0$, where N is the total number of entries in the tensor. We implemented our method for predicting missing pixels in image inpainting, which showed that our method is competitive with existing spectral methods used for tensor completion.

The algorithm that we propose is inspired by local functional approximations, specifically Taylor's series expansion. This work has similarities to classical kernel regression, which also relies on local smoothed approximations, c.f. Mack and Silverman (1982); Wand and Jones (1994). However, since kernel regression and other similar methods use explicit knowledge of the input features, their analysis and proof techniques do not extend to our context. Instead of using distance in the unknown latent space, the algorithm weights datapoints according to similarities that are computed from the data itself. Our analysis shows that although the similarities may not reflect true latent distance, they essentially reflect L_2 functional distances between the expected data function associated to a pair of rows or columns, which is sufficient to guarantee that the datapoints with high similarities are indeed similar in value.

1.3. *Organization of the Paper.* In Section 2 we setup the formal model and problem statement and discuss the assumptions needed for our analysis. In Section

3 we introduce the basic form of our algorithm, which is similar to the user-user variant of collaborative filtering. We show that the algorithm can be intuitively derived from local first order approximations, and we present heuristic variants of the algorithm that perform well in practice. In Section 4 we present the main theoretical results of our paper as they pertain to matrix completion, showing provable convergence of the user-user and item-item variants of our algorithm. We discuss our results alongside the results of the USVT estimator, which highlights that the sample complexity of our method scales well with the latent dimension of the row and column hidden features. In Section 5 we extend the algorithm and theoretical results to tensor completion, including a discussion on the optimal flattening of a tensor which is best suited to our method. In Section 6 we present experimental results from applying our methods to both matrix completion in the context of predicting movie ratings, and tensor completion in the context of image inpainting. Sections 7 and 8 include the detailed proofs and associated lemmas used in the theoretical analysis.

2. Problem Statement.   Suppose that there is an unknown $m \times n$ matrix A which we would like to estimate, and we observe partial observations of a noisy matrix Z. Let $\mathcal{D} \subset [m] \times [n]$ denote the index set of observed entries. Precisely, for any $u \in [m]$ and $i \in [n]$, $Z(u, i)$ is observed and thus $(u, i) \in \mathcal{D}$ with probability $p \in [0, 1]$, and otherwise it is not observed (or missing) and thus $(u, i) \notin \mathcal{D}$ with probability $1 - p$, independent of everything else. When observed, it provides an unbiased noisy signal of $A(u, i)$, such that $\mathbb{E}[Z(u, i)] = A(u, i)$. Concretely,

$$(2.1) \qquad\qquad Z(u, i) = A(u, i) + \eta(u, i),$$

where $\eta(u, i)$ is an independent zero-mean bounded random variable with

$$(2.2) \qquad \mathbb{E}[\eta(u, i)] = 0, \; \text{Var}[\eta(u, i)] = \gamma^2, \; \text{ and } \; |\eta(u, i)| \leq B_e,$$

for some constant B_e.

We posit the following nonparametric model for the matrix A, also known as the Latent Variable Model Chatterjee (2015). Each row u and column i is associated to latent features $x_1(u) \in \mathcal{X}_1$ and $x_2(i) \in \mathcal{X}_2$ for some compact metric spaces $\mathcal{X}_1, \mathcal{X}_2$. The (u, i)-th entry of matrix A takes the form of

$$(2.3) \qquad\qquad A(u, i) = f(x_1(u), x_2(i))$$

for some latent function $f : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathbb{R}$. Note that there does not exist a unique representation, as we can apply a transformation to the latent feature spaces $\mathcal{X}_1$ and $\mathcal{X}_2$, and apply an equivalent transformation to the function f such that the data is exactly equal under the new representation. Therefore, the question of estimating the function f itself is not well defined, and thus we focus our energy on predicting the values $A(u, i)$.

2.1. *Blind Regression.*   We call the problem of interest *Blind Regression* for the following reason. In the setting of *Regression*, one observes data containing features and associated labels; the goal is to learn functional relationship (or model) between features and labels assuming that labels are noisy observation. In our setting, tuple $(x_1(u), x_2(i))$ are features and $Z(u, i)$ are noisy observations of associated labels. In that sense, it is very much like the Regression setting. However, the features $(x_1(u), x_2(i))$ are *latent*. Therefore, the term *Blind Regression*.

2.2. *Operating Assumptions.*   In addition to the assumptions on the additive noise model presented in (2.2), we additionally assume basic regularity conditions on the function f and the latent spaces $\mathcal{X}_1$ and $\mathcal{X}_2$. Assume $\mathcal{X}_1$ and $\mathcal{X}_2$ are endowed with metrics $d_{\mathcal{X}_1}$ and $d_{\mathcal{X}_2}$, and have diameters $D_{\mathcal{X}_1}$ and $D_{\mathcal{X}_2}$:

$$(2.4) \qquad d_{\mathcal{X}_1}(x_1, x_1') \leq D_{\mathcal{X}_1} \text{ for all } x_1, x_1' \in \mathcal{X}_1,$$

$$(2.5) \qquad d_{\mathcal{X}_2}(x_2, x_2') \leq D_{\mathcal{X}_2} \text{ for all } x_2, x_2' \in \mathcal{X}_2.$$

Assume the latent function f is L-Lipschitz:

$$(2.6) \qquad |f(x_1, x_2) - f(x_1', x_2')| \leq L \max \left(d_{\mathcal{X}_1}(x_1, x_1'), d_{\mathcal{X}_2}(x_2, x_2') \right),$$

for all $x_1, x_1' \in \mathcal{X}_1$ and $x_2, x_2' \in \mathcal{X}_2$. While our formal results require Lipschitz continuity, equivalent results can be derived for piecewise Lipschitz functions as well. Assumptions (2.4), (2.5), and (2.6) along with the bounded noise in (2.2) imply that all entries of matrix A and Z are uniformly bounded. Specifically we define parameter

$$(2.7) \qquad B_0 \triangleq LD_{\mathcal{X}_1} + 2B_e,$$

such that for any $u, v \in [m]$ and any $i \in [n]$,

$$|Z(u, i) - Z(v, i)| = |f(x_1(u), x_2(i)) + \eta(u, i) - f(x_1(v), x_2(i)) - \eta(v, i)|$$
$$\leq LD_{\mathcal{X}_1} + 2B_e =: B_0.$$

Assume that for each $u \in [m]$ and $i \in [n]$, the latent features $x_1(u)$ and $x_2(i)$ are sampled independently from $\mathcal{X}_1$ and $\mathcal{X}_2$ according to Borel probability measures $P_{\mathcal{X}_1}$ and $P_{\mathcal{X}_2}$ respectively. For $i \in \{1, 2\}$, define $\phi_i : \mathbb{R}_+ \to [0, 1]$ to be a function lower bounding the cumulative distribution function according to the distance in the latent space. For any radius $r > 0$, we define it according to

$$(2.8) \qquad \phi_i(r) \equiv \text{ess} \inf_{x_0 \in \mathcal{X}_i} P_{\mathcal{X}_i} \left(B_i(x_0, r) \right),$$

where $B_i(x_0, r)$ indicates the ball of radius r centered around point $x_0 \in \mathcal{X}_i$,

$$(2.9) \qquad B(x_0, r) = \{x \in \mathcal{X}_i : d_{\mathcal{X}_i}(x_0, x) \leq r\}.$$

For example, if $P_{\mathcal{X}_1}$ is a uniform distribution over a unit cube in d dimensional Euclidean space, then $\phi_1(r) = \min\left\{1, \left(\frac{r}{2}\right)^d\right\}$. If $P_{\mathcal{X}_1}$ is supported over finitely many points, then $\phi_1(r) \geq \min_{\mathbf{x}\in\text{supp}(P_{\mathcal{X}_1})} P_{\mathcal{X}_1}(\mathbf{x})$ is a positive constant (see Appendix B for a detailed discussion on the function ϕ_1).

2.3. *Connections to Exchangeability.* In fact, the latent variable model is well motivated and arises as a canonical representation for row and column exchangeable data, cf. Aldous (1981) and Hoover (1981). Suppose that our data matrix Z is a particular realization of the first $m \times n$ entries of a random array $\mathbf{Z} = \{\mathbf{Z}(u,i)\}_{(u,i)\in\mathbb{N}\times\mathbb{N}}$, which satisfies

$$(2.10) \qquad \mathbf{Z}(u,i) \stackrel{d}{=} \mathbf{Z}\left(\sigma(u),\tau(i)\right) \ \text{ for all } (u,i),$$

for every pair of permutations[1] σ,τ of $\mathbb{N}$. We use $\stackrel{d}{=}$ to denote equivalently distributed, i.e. the random variables on both sides have the same distribution. Random array $\mathbf{Z}$ satisfying (2.10) is called *separately* row and column exchangeable. That is, a dataset is exchangeable if the distribution is invariant to permutations of the rows and columns. For an interested reader, Austin (2012) and Orbanz and Roy (2015) presents overviews of exchangeable arrays.

In practice, the use of exchangeable arrays as a model is appropriate for variety of reasons. For example, in the setting of a recommendation system with anonymized data, this property may be reasonable if the order of the users in the system does not intrinsically carry information about the type of user; or in other words, a user in the system could equally likely have been located in any row of the dataset.

In addition to exchangeability being quite a reasonable property for a wide variety of applications, it also leads to a convenient latent variable representation. The Aldous-Hoover representation theorem provides a succinct characterization for such exchangeable arrays. According to the theorem (see Corollary 3.3 in Orbanz and Roy (2015) for example), a random data array $\mathbf{Z}$ is exchangeable if and only if it can also be represented as

$$(2.11) \qquad \mathbf{Z}(u,i) \stackrel{d}{=} f_\theta\big(\theta_{row}(u),\theta_{col}(i),\theta(u,i)\big) \ \text{ for all } (u,i)$$

where $\theta, \left\{\theta_{row}(u)\right\}_{u\in\mathbb{N}}, \left\{\theta_{col}(i)\right\}_{i\in\mathbb{N}}, \left\{\theta(u,i)\right\}_{(u,i)\in\mathbb{N}\times\mathbb{N}}$ are independent random variables drawn uniformly from the unit interval $[0,1]$, and f_θ is a measurable function indexed by the realization of θ, such that for any particular realization $\theta = t \in [0,1]$, $f_t : [0,1]^3 \to \mathbb{R}$. As described in Orbanz and Roy (2015), this suggests the following generative model:

---

[1]The permutations over $\mathbb{N}$ are defined in the usual manner where only finitely many indices are *permuted.*

1. Sample an instance of $\theta \sim U[0,1]$ determining the governing function f_θ.
2. Independently sample i.i.d. uniform random variables $\theta_{row}(u) \sim U[0,1]$, $\theta_{col}(i) \sim U[0,1]$, $\theta(u,i) \sim U[0,1]$ for every row $u \in [m]$ and column $i \in [n]$.
3. Compute the realized data matrix Z according to

$$Z(u,i) = f_\theta\big(\theta_{row}(u), \theta_{col}(i), \theta(u,i)\big).$$

By comparing the model from (2.11) with the latent variable model described in (2.1) and (2.3), we can see that the latent variable model considered in this work is a restricted subclass of exchangeable models which additionally impose an additive noise model and regularity conditions on the function f_θ. In our model we have conditioned on the universal index θ, such that given partial observations from matrix Z for a particular f_θ, our goal is to learn predicted outcomes of the realized f_θ. Our model takes the form of

$$(2.12) \qquad Z(u,i) = f(x_1(u), x_2(i)) + \eta(u,i)$$

where $\{x_1(u)\}_{u\in[m]}, \{x_2(i)\}_{i\in[n]}, \{\eta(u,i)\}_{(u,i)\in[m]\times[n]}$ are sampled independently. We can transform this to the form of (2.11) by considering f to be equal to the realized function f_θ, considering the latent variables $x_1(u) \sim P_{\mathcal{X}_1}$ and $x_2(i) \sim P_{\mathcal{X}_2}$ to be higher dimensional representations of $\theta_{row}(u)$ and $\theta_{col}(i)$ in spaces $\mathcal{X}_1$ and $\mathcal{X}_2$, and considering the noise term $\eta(u,i)$ to be generated by applying some transformation to the variable $\theta(u,i)$. Given these transformations, it becomes equivalent that

$$(2.13) \qquad Z(u,i) = f_\theta\big(\theta_{row}(u), \theta_{col}(i), \theta(u,i)\big) = f(x_1(u), x_2(i)) + \eta(u,i).$$

Our model additionally imposes regularity conditions on f by requiring it to be Lipschitz continuous with respect to the higher dimensional representation $\mathcal{X}_1 \times \mathcal{X}_2$ instead of being any arbitrary measurable function over $[0,1] \times [0,1]$. From a modeling perspective, effectively we are transferring the *model complexity* from a potentially complex measurable latent functions over $[0,1] \times [0,1]$ to a simpler Lipschitz latent function over a potentially more complex latent variable space $\mathcal{X}_1 \times \mathcal{X}_2$. The simple functional form provides analytic tractability for establishing theoretical results.

3. Algorithm. Our algorithm builds on intuition from local approximation methods such as kernel regression. Therefore it takes the form of a similarity based method which first defines a kernel, i.e. distances between pairs of rows or columns, and then computes the estimate by averaging over datapoints which are determined to be close according to the estimated distances. The basic user-user

k-nearest neighbor variant which we present next is equivalent to a variant of the classical similarity based collaborative filtering methods. In Section 4 we will provide clear theoretical guarantees showing convergence of the mean squared error of this basic algorithm. We also present another variation of the algorithm which combines both row and column similarities to compute the kernel between datapoints. In Section 6 we will show experimental results that suggest combining row and column similarities improves the estimate.

3.1. *User-User k-Nearest Neighbor Algorithm.* We refer to this algorithm to as the "user-user k-nearest neighbor" variant of our method, because the algorithm computes estimates by exploiting the similarity between rows (users), and averages datapoints over the k most similar rows with available ratings. The algorithm uses parameters $\beta, k \in \mathbb{Z}_+$, and we denote the output estimated matrix as $\hat{A}^k$, which takes as input the observed entries of Z and would like to best approximate A. Steps 1 and 2 set up definitions which are used to compute the row similarities. Step 3 chooses the k most similar rows according to the computed similarities, and Step 4 computes the estimate by averaging over datapoints from the chosen rows.

1. For each row $u \in [m]$, let $\mathcal{O}^u$ be the set of column indices for which $Z(u, i)$ is observed:

   $$(3.1) \qquad \mathcal{O}^u = \{i \ s.t. \ (u, i) \in \mathcal{D}\}.$$

   Define the "overlap" between a pair of rows $(u, v) \in [m] \times [n]$ to be

   $$(3.2) \qquad \mathcal{O}^{uv} := \mathcal{O}^u \cap \mathcal{O}^v.$$

2. For each pair of rows $(u, v) \in [m] \times [m]$ with sufficiently large overlap $|\mathcal{O}^{uv}| \geq \beta$, compute the mean and variance of the difference in their associated observed datapoints:

   $$(3.3) \qquad m_{uv} = \frac{1}{|\mathcal{O}^{uv}|} \left(\sum_{j \in \mathcal{O}^{uv}} Z(u, j) - Z(v, j) \right),$$

   $$(3.4) \qquad s_{uv}^2 = \frac{1}{|\mathcal{O}^{uv}| - 1} \left(\sum_{j \in \mathcal{O}^{uv}} (Z(u, j) - Z(v, j) - m_{uv})^2 \right).$$

   The sample variance s_{uv}^2 acts as an estimated distance between rows u and v. It can equivalently be computed from the expression

   $$s_{uv}^2 = \frac{1}{2|\mathcal{O}^{uv}|(|\mathcal{O}^{uv}| - 1)} \sum_{i, j \in \mathcal{O}^{uv}} ((Z(u, i) - Z(v, i)) - (Z(u, j) - Z(v, j)))^2.$$

3. For each index of the matrix $(u, i) \in [m] \times [n]$, define $\mathcal{S}_u^\beta(i)$ to be the set of rows with sufficient overlap with u that also contain available data about column i:

$$(3.5) \qquad \mathcal{S}_u^\beta(i) = \{v \neq u \in [m] \; s.t. \; (v, i) \in \mathcal{D} \text{ and } |\mathcal{O}^{uv}| \geq \beta\}.$$

Let $\mathcal{S}_u^{\beta,k}(i) \subset \mathcal{S}_u^\beta(i)$ denote the (at most) k rows with minimum sample variance s_{uv}^2 amongst rows $v \in \mathcal{S}_u^\beta(i)$, where ties can be broken in any arbitrary manner.

4. The user-user k-nearest neighbor smoothed estimator of $A(u, i)$ is computed according to:

$$(3.6) \qquad \hat{A}^k(u, i) = \frac{1}{|\mathcal{S}_u^{\beta,k}(i)|} \left(\sum_{v \in \mathcal{S}_u^{\beta,k}(i)} \hat{A}_v(u, i) \right),$$

where

$$(3.7) \qquad \hat{A}_v(u, i) = Z(v, i) + m_{uv}.$$

If $|\mathcal{S}_u^{\beta,k}(i)| = 0$, then define $\hat{A}^k(u, i) = 0$.

While the algorithm above chooses the nearest neighbors $\mathcal{S}_u^\beta(i)$ amongst rows $v \neq u$, if the entry $Z(u, i)$ itself is actually observed, we could modify the algorithm to include u into the set $\mathcal{S}_u^{\beta,k}(i)$ chosen in Step 3, such that $Z(u, i)$ itself will be included in compute the estimate $\hat{A}^k(u, i)$. The results which follow will equally hold given this slight modification.

An equivalent expression for $\hat{A}_v(u, i)$ is given by

$$(3.8) \qquad \hat{A}_v(u, i) = \frac{1}{|\mathcal{O}^{uv}|} \left(\sum_{j \in \mathcal{O}^{uv}} \hat{A}_{vj}(u, i) \right)$$

where

$$(3.9) \qquad \hat{A}_{vj}(u, i) = Z(v, i) + Z(u, j) - Z(v, j),$$

which will be a useful form used to define other variations of our algorithm. An equivalent "item-item" variant of the algorithm follows from simply applying the stated algorithm on the transpose of the matrix such that the similarities are computed between columns and estimates are obtained from averaging over similar columns with available data.

This algorithm is asymptotically equivalent to the mean-adjusted variant of the classical user-user k-nearest neighbor collaborative filtering algorithm, since m_{uv} will converge to the difference between the empirical means of each row u and v. Our method uses the row empirical variance instead of the standard cosine similarity.

3.2. *Intuition Derived from First-Order Taylor Approximation.* In fact, since we have a clear model, we can show that the proposed algorithm can be derived from insights related to the classical Taylor approximation of a function. Suppose the latent space $\mathcal{X}_1 \cong \mathcal{X}_2 \cong \mathbb{R}$, and we wish to predict the (u, i)-th entry, $A(u, i) = f(x_1(u), x_2(i))$. According to the first order Taylor approximation of f around $(x_1(v), x_2(j))$ for some $u \neq v \in [m], i \neq j \in [n]$, it follows that

$$f(x_1(u), x_2(i)) \approx f(x_1(v), x_2(j)) + (x_1(u) - x_1(v))\frac{\partial f(x_1(v), x_2(j))}{\partial x_1}$$
$$+ (x_2(i) - x_2(j))\frac{\partial f(x_1(v), x_2(j))}{\partial x_2}.$$

We are not able to directly compute this expression, as we do not know the latent features, the function f, or the partial derivatives of f. However, we can again compute the first order Taylor approximation for $f(x_1(v), x_2(i))$ and $f(x_1(u), x_2(j))$ around $(x_1(v), x_2(j))$, which results in a set of equations with the same unknown terms. It follows from substitution and rearranging the terms that

$$f(x_1(u), x_2(i)) \approx f(x_1(v), x_2(i)) + f(x_1(u), x_2(j)) - f(x_1(v), x_2(j)),$$

as long as the first order Taylor approximation is accurate. Thus if the noise term in (2.1) is small, we can approximate $f(x_1(u), x_2(i))$ by using observed ratings $Z(v, j)$, $Z(u, j)$ and $Z(v, i)$ according to

(3.10) $$\hat{A}_{vj}(u, i) = Z(u, j) + Z(v, i) - Z(v, j).$$

This is precisely (3.9) in the algorithm described in Section 3. However, we only expect this estimate to be close when the first order approximation is valid, i.e. the latent features $x_1(u) \approx x_1(v)$ and $x_2(i) \approx x_2(j)$. Unfortunately we cannot directly verify this because the features are latent. Therefore we need a data-driven *surrogate* to help decide for which entries (v, j) the estimate $\hat{A}_{vj}(u, i)$ is close to $A(u, i)$.

The approximation error for using $\hat{A}_{vj}(u, i)$ to estimate $A(u, i)$ can be directly computed by substituting (2.1) and (2.3) into (3.10),

$$\text{Error} \equiv f(x_1(u), x_2(i) - \hat{A}_{vj}(u, i)$$
$$= (f(x_1(u), x_2(i)) - f(x_1(v), x_2(i)))$$
(3.11) $$- (f(x_1(u), x_2(j) - f(x_1(v), x_2(j))) - \eta(v, i) + \eta(v, j) - \eta(u, j).$$

Therefore, conditioned on the row latent variables $x_1(u), x_1(v)$, the average squared error, with respect to the individual noise terms and the randomly sampled column latent variables $x_2(i)$ and $x_2(j)$, is given by

$$\mathbb{E}\left[(\text{Error})^2 \mid x_1(u), x_1(v)\right]$$
$$= 2\,\text{Var}_{\mathbf{x}_2 \sim \mathcal{X}_2}\left[f(x_1(u), \mathbf{x}_2) - f(x_1(v), \mathbf{x}_2) \mid x_1(u), x_1(v)\right] + 3\gamma^2.$$

This expression shows that the expected squared error conditioned on rows u and v is directly related to the variance of the difference between the entries associated to rows u and v. The good news is that the variance of the row differences can in fact be estimated from the data itself. We can equivalently show that the variance of the column differences is directly related to the expected squared error conditioned on the column latent variables $x_2(i)$ and $x_2(j)$.

This suggests that $\hat{A}_{vj}(u, i)$ is a good estimate for $A(u, i)$ as long as either (a) the empirical variance of the differences between corresponding entries of rows u and v is small, or (b) the empirical variance of the differences between corresponding entries of columns i and j is small. This suggests an algorithm which computes the empirical variances between pairs of rows and pairs of columns, and produces a final estimate for the (u, i)-th entry by averaging over $\hat{A}_{vj}(u, i)$ for (v, j) where either the row variance between u and v is small, or the column variance between i and j is small.

The user-user k-nearest neighbor algorithm presented in section 3.1 follows precisely this format, using the pairwise row variances computed in Step 2, denoted as s_{uv}^2. The choice of a large β guarantees that the empirical variance is a good approximation of the true variance. The at most k rows in $\mathcal{S}_u^{\beta,k}(i)$ are selected to have small empirical variance s_{uv}^2, and a large enough k is chosen to average out the error due to the individual noise terms η, trading off between the bias and variance of the final estimate $\hat{A}^k(u, i)$.

3.3. *General Form of the Algorithm.* The intuition provided characterization of the estimation error as a function of both the row and column pairwise variances, however the user-user k-nearest neighbor algorithm presented in Section 3.1 only used pairwise row variances. Therefore we present a general form of the algorithm in this section which will compute both row and column variances and then compute weights for the datapoints as a function of the row and column variances. These weights are used to predict the final estimate. There are many possible weight functions, i.e. kernel functions, that one could choose. In Section 3.3.1, we show the weight function which leads to the user-user and item-item k-nearest neighbor algorithm which we presented earlier. In Section 3.3.2, we present a weight function which combines both row and column variances. In experiments, we will show that combining both row and column similarities improves the estimate. Let $\hat{A}(u, i)$ denote the estimate our algorithm produces for $A(u, i)$.

1. For rows u and v, compute the row-based expressions $\mathcal{O}^u$ and $\mathcal{O}^{uv}$ from (3.1) and (3.2). For columns i and j, compute the equivalent column-based expressions. The observed entries in column i are

$$(3.12) \qquad \bar{\mathcal{O}}^i = \{u \ s.t. \ (u, i) \in \mathcal{D}\}.$$

The overlap entries between a pair of columns i and j is

$$(3.13) \qquad \bar{\mathcal{O}}^{ij} := \bar{\mathcal{O}}^i \cap \bar{\mathcal{O}}^j.$$

2. For rows u and v, compute the empirical mean m_{uv} and variance s_{uv}^2 of the difference between the entries associated to the two rows, defined in (3.3) and (3.4). Equivalently, for columns i and j, compute the empirical mean and variance of the difference in the entries associated to the two columns according to

$$(3.14) \qquad \bar{m}_{ij} = \frac{1}{\left|\bar{\mathcal{O}}^{ij}\right|} \left(\sum_{u \in \bar{\mathcal{O}}^{ij}} Z(u,i) - Z(u,j) \right),$$

$$(3.15) \qquad \bar{s}_{ij}^2 = \frac{1}{\left|\bar{\mathcal{O}}^{ij}\right| - 1} \left(\sum_{u \in \bar{\mathcal{O}}^{ij}} (Z(u,i) - Z(u,j) - \bar{m}_{ij})^2 \right).$$

3. For an index of the matrix $(u,i) \in [m] \times [n]$ that we would like to estimate, let $\mathcal{B}(u,i)$ denote the set of indices (v,j) such that the entries $Z(v,j)$, $Z(u,j)$ and $Z(v,i)$ are observed, i.e. $\hat{A}_{vj}(u,i)$ as defined in (3.9) is computable from the observations. Formally,

$$(3.16) \qquad \mathcal{B}(u,i) = \{(v,j) \ s.t. \ (v,j) \in \mathcal{D}, (u,j) \in \mathcal{D}, \text{ and } (v,i) \in \mathcal{D}\}.$$

For $(v,j) \in \mathcal{B}(u,i)$, define some weight function $w_{ui}(v,j) \in [0,1]$, which will specify the weight used to incorporate $\hat{A}_{vj}(u,i)$ into the final estimate $\hat{A}(u,i)$. These weights can be any function of the quantities computed in Steps 1 and 2, but according to the intuition presented in Section 3.2, we may want to choose $w_{ui}(v,j)$ to decrease with s_{uv}^2 and $\bar{s}_{ij}^2$.

4. Compute the final prediction for $A(u,i)$ according to a weighted combination

$$(3.17) \qquad \hat{A}(u,i) = \frac{\sum_{(v,j) \in \mathcal{B}(u,i)} w_{ui}(v,j) \hat{A}_{vj}(u,i)}{\sum_{(v,j) \in \mathcal{B}(u,i)} w_{ui}(v,j)},$$

where

$$(3.18) \qquad \hat{A}_{vj}(u,i) = Z(v,i) + Z(u,j) - Z(v,j).$$

We have not yet defined the weight function $w_{ui}(v,j)$ in Step 3, and in the following two sections we show different selections of the weight function and the corresponding algorithm that results from (3.17).

3.3.1. *User-User and Item-Item k-Nearest Neighbor Weights.* We can derive the user-user and item-item k-nearest neighbor algorithms presented in Section 3.1 as an instance of (3.17) by an appropriate selection of the weights. For an index of the matrix $(u, i) \in [m] \times [n]$, recall our definition of set $\mathcal{S}_u^\beta(i)$ from (3.5). We can verify that in fact $\mathcal{S}_u^\beta(i)$ is a subset of $\mathcal{B}(u, i)$, additionally enforcing that the row overlap is larger than β. Recall that $\mathcal{S}_u^{\beta,k}(i) \subset \mathcal{S}_u^\beta(i)$ denotes the (at most) k rows with minimum sample variance s_{uv}^2 amongst rows $v \in \mathcal{S}_u^\beta(i)$. Therefore, the user-user k-nearest neighbor algorithm is equivalent to choosing the following weight function

$$(3.19) \qquad w_{ui}(v, j) = \begin{cases} 1 & \text{if } v \in \mathcal{S}_u^{\beta,k}(i) \\ 0 & \text{otherwise.} \end{cases}$$

This essentially defines a hard threshold where datapoints are included in the final estimate if and only if the row is amongst the k minimum sample variance rows, and evenly weighted amongst included datapoints. We can verify that the estimate computed from (3.17) using this choice of the weight function is equivalent to (3.6).

The item-item k-nearest neighbor algorithm can equivalently be defined using a similar weight function. Let $\bar{\mathcal{S}}_i^\beta(u)$ be the set of columns with sufficient overlap with column i that also contain available data about row u,

$$(3.20) \qquad \bar{\mathcal{S}}_i^\beta(u) = \left\{ j \neq i \in [n] \ s.t. \ (u, j) \in \mathcal{D} \text{ and } |\bar{\mathcal{O}}^{ij}| \geq \beta \right\}.$$

And define the set $\bar{\mathcal{S}}_i^{\beta,k}(u) \subset \bar{\mathcal{S}}_i^\beta(u)$ to be the (at most) k columns with minimum sample variance $\bar{s}_{ij}^2$ amongst columns $j \in \bar{\mathcal{S}}_i^\beta(u)$. Therefore, the item-item k-nearest neighbor algorithm is equivalent to choosing the following weight function

$$(3.21) \qquad w_{ui}(v, j) = \begin{cases} 1 & \text{if } j \in \bar{\mathcal{S}}_i^{\beta,k}(u) \\ 0 & \text{otherwise.} \end{cases}$$

3.3.2. *User-Item Gaussian Kernel Weights.* The previous choice of weights in Section 3.3.1 uses either row or column variances, and implements a hard-threshold kernel, averaging equally amongst the k nearest rows or columns. We introduce a variant of the algorithm which combines both row and column variances s_{uv}^2 and $\bar{s}_{ij}^2$ using a soft-thresholded Gaussian kernel. Since we need the empirical variances to concentrate, we restrict ourselves to the set $\mathcal{B}^\beta(u, i)$, which we define to be entries $(v, j) \in \mathcal{B}(u, i)$ for which rows u and v have overlap of at least β and columns of i and j have overlap of at least β. Formally,

$$(3.22) \qquad \mathcal{B}^\beta(u, i) = \left\{ (v, j) \in \mathcal{B}(u, i) \ s.t. \ |\mathcal{O}^{uv}| \geq \beta \text{ and } |\bar{\mathcal{O}}^{ij}| \geq \beta \right\}.$$

Inspired by kernel regression, we define the weights according to a Gaussian kernel with bandwith parameter $\lambda \in \mathbb{R}_+$, using the minimum of the row and column variances as a proxy for the distance:

$$(3.23) \qquad w_{ui}(v,j) = \begin{cases} \exp\left(-\lambda \min\{s_{uv}^2, \bar{s}_{ij}^2\}\right) & \text{if } (v,j) \in \mathcal{B}^\beta(u,i) \\ 0 & \text{otherwise.} \end{cases}$$

When $\lambda = \infty$, the estimate $\hat{A}(u,i)$ only depends on the basic estimates $\hat{A}_{vj}(u,i)$for entries (v,j) whose row or column has minimum sample variance. When $\lambda = 0$, the algorithm equally averages all the estimates $\hat{A}_{vj}(u,i)$ for $(v,j) \in \mathcal{B}^\beta(u,i)$. Empirically, this variant of the algorithm seem to perform very well with an appropriate selection of the bandwidth parameter λ, which can be tuned using cross validation.

3.3.3. *Cosine Similarity Weights.*   In our proposed algorithm, we selected neighbors and associated weights as a function of the row and column sample variances s_{uv}^2 and $\bar{s}_{ij}^2$, which is equivalent to the squared distance of the mean-adjusted values. In classical collaborative filtering, cosine similarity is commonly used, which can be approximated as a different choice of the weight kernel over the squared difference. Therefore, under the assumption that the variations of users' ratings around their respective means are approximately similar, then weighting the estimates proportional to cosine similarity over mean-adjusted values is equivalent to choosing a polynomially decaying kernel and plugging in the sample variance as a proxy for distance between points.

4. Main Results.   Given an estimator $\hat{A}$ for the unknown matrix $A \in \mathbb{R}^{m \times n}$ of interest, we use the mean-squared error (MSE) to evaluate the performance of the estimator, defined as

$$(4.1) \qquad MSE(\hat{A}) = \mathbb{E}\left[\frac{1}{mn}\sum_{u=1}^{m}\sum_{i=1}^{n}\left(\hat{A}(u,i) - A(u,i)\right)^2\right].$$

We present a theorem which upper bounds the MSE of the estimate produced by the user-user k-nearest neighbor algorithm presented in Section 3.1. Recall from our problem statement in Section 2 that p is the probability each entry is observed, L is the Lipschitz constant of the latent function f, $B_0 = LD_{\mathcal{X}_1} + 2B_e$ is the bound on all entries, and the function ϕ_1 lower bounds the cumulative distribution function of the row latent variable sampled from $\mathcal{X}_1$ according to $P_{\mathcal{X}_1}$, defined in (2.8). Our algorithm uses parameters k and β, where β is the threshold for minimum number of overlapped entries to compute row variances, and k is the smoothing parameter which determines how many nearest neighbor rows are incorporated in

the final estimate. The variable ζ in the formal theorem statement is an analysis parameter used to quantify our error bound.

THEOREM 4.1 (Main theorem; user-user).   *Suppose that*

$$p \geq \max \left\{ m^{-1+\delta}, n^{-\frac{1}{2}+\delta} \right\} \text{ for some } \delta > 0,$$

$$\zeta \text{ satisfies } \phi_1 \left(\sqrt{\frac{\zeta}{L^2}} \right) \geq c_\phi \, (mp)^{-2/3} \text{ for some } c_\phi \geq 0,$$

$$\beta = c_\beta n p^2 \text{ for some } c_\beta \in (0,1), \text{ and}$$

$$k \leq \frac{c_k}{2} (m-1) p \phi_1 \left(\sqrt{\frac{\zeta}{L^2}} \right) \text{ for some } c_k \in [0,1).$$

The mean-squared error (MSE) of the estimate produced by the user-user k-nearest neighbor variant of our algorithm with overlap parameter β is upper bounded by:

(4.2) $$MSE(\hat{A}) \leq 2F_1 \ln \left(\frac{2B_0}{F_1} \right) + (F_1 + F_2)^2 + 2F_2 + 4B_0^2 F_3,$$

where

$$F_1 = \zeta + 2\beta^{-1/3} + \frac{\gamma^2}{k},$$

$$F_2 = \beta^{-1/3}, \text{ and}$$

$$F_3 = 3 \exp \left(-c_1 \, (mp)^{1/3} \right) + \left(m + \frac{9}{2} mp \right) \exp \left(-c_2 \beta^{1/3} \right),$$

with absolute constants c_1 and c_2 defined according to the geometry of the latent spaces,

$$c_1 = \min \left\{ \frac{1}{24}, \frac{c_\phi \, (1 - c_k)^2}{8} \right\},$$

$$c_2 = \min \left\{ \frac{c_\beta^2}{2}, \frac{3}{6B_0^2 + 4B_0}, \frac{1}{8B_0^2 \, (2B_0^2 + 1)} \right\}.$$

To guarantee that F_3 converges to zero, we additionally need to ensure that $\log m < (np^2)^{\delta'/3}$ for some $\delta' \in (0,1)$.

We can prove an equivalent MSE bound for the item-item k-nearest neighbor variant, which essentially follows from taking the transpose of the matrix, thus switching m and n. We state it here for completeness.

THEOREM 4.2 (Main theorem; item-item).   *Suppose that*

$$p \geq \max\left\{m^{-\frac{1}{2}+\delta}, n^{-1+\delta}\right\} \text{ for some } \delta > 0,$$

$$\zeta \text{ satisfies } \phi_2\left(\sqrt{\frac{\zeta}{L^2}}\right) \geq c_\phi \left(np\right)^{-2/3} \text{ for some } c_\phi \geq 0,$$

$$\beta = c_\beta m p^2 \text{ for some } c_\beta \in (0,1), \text{ and}$$

$$k \leq \frac{c_k}{2}(n-1)p\phi_2\left(\sqrt{\frac{\zeta}{L^2}}\right) \text{ for some } c_k \in [0,1).$$

The mean-squared error (MSE) of the estimate produced by the item-item k-nearest neighbor variant of our method with overlap parameter β is upper bounded by:

$$MSE(\hat{A}) \leq 2F_1 \ln\left(\frac{2B_0}{F_1}\right) + (F_1 + F_2)^2 + 2F_2 + 4B_0^2 F_3,$$

where

$$F_1 = \zeta + 2\beta^{-1/3} + \frac{\gamma^2}{k},$$

$$F_2 = \beta^{-1/3}, \text{ and}$$

$$F_3 = 3\exp\left(-c_1 \left(np\right)^{1/3}\right) + \left(n + \frac{9}{2}np\right)\exp\left(-c_2\beta^{1/3}\right),$$

with absolute constants c_1 and c_2 defined according to the geometry of the latent spaces,

$$c_1 = \min\left\{\frac{1}{24}, \frac{c_\phi\left(1-c_k\right)^2}{8}\right\},$$

$$c_2 = \min\left\{\frac{c_\beta^2}{2}, \frac{3}{6B_0^2 + 4B_0}, \frac{1}{8B_0^2\left(2B_0^2 + 1\right)}\right\}.$$

To guarantee that F_3 converges to zero, we additionally need to ensure that $\log n < (mp^2)^{\delta'/3}$ for some $\delta' \in (0,1)$.

Comparing the sample requirements for p in both theorems indicates that the user-user variant is more suitable for a fat matrix, and the item-item variant is more suitable for a tall matrix. This is due to the fact that the limiting condition of our sample complexity requires that there is a sufficiently large overlap between pairs of rows or columns. A fat matrix will naturally satisfy the condition $(np^2)^{\delta'/3} \geq n^{2\delta'\delta/3} \geq \log m$, and a tall matrix will naturally satisfy the condition $(mp^2)^{\delta'/3} \geq$

$m^{2\delta'\delta/3} \geq \log n$, and a square matrix will satisfy both conditions, guaranteeing that the term F_3 decays to 0 exponentially fast.

The parameter β determines the required overlap between rows, such that choosing β to grow with m, n, such as $\beta = np^2/2$, ensures that asymptotically the empirical mean m_{uv} and variance s_{uv}^2 computed in the algorithm converge to the true mean and variance. Given any choice of β, we can derive the rates of convergence of the empirical mean and variance statistics (see Lemmas 7.2 and 7.3), which also impacts the MSE of the final estimate. The parameter k determines the number of nearest neighbor rows which are incorporated in the final estimate in (3.6). We choose k to balance bias and variance, ensuring that k goes to infinity with m and n to drive down the error due to the individual additive noise terms η, yet also controlling that k grows slowly enough to guarantee that the sample variance s_{uv}^2 of the k nearest neighbor rows goes to zero as well.

In the process of proving the bound on the mean squared error, we are able to in fact upper bound the tail probability of error for each entry as presented in Theorem 7.6. This entry-wise error bound is stronger than a MSE bound over the aggregate error, suggesting that the error is evenly spread amongst different entries.

4.1. *Results Given Local Geometry of Latent Probability Measure.* The local "geometry" of the latent probability measure $P_{\mathcal{X}_1}$ through the function ϕ_1 determines the impact of the latent space dimension on the sample complexity and error convergence rate of the user-user k-nearest neighbor algorithm. Since the algorithm is a neighbor-based method, we need to guarantee that for each row $u \in [m]$, there are sufficiently many rows $v \in \mathcal{S}_u^\beta(i)$ such that the variances of their row differences are small, which we showed in section 3.2 intuitively implies bounds on $|\hat{A}_{vj}(u, i) - \hat{A}(u, i)|$ for the average column j. By our model assumption that the matrix A is described by an L-Lipschitz function f, it is sufficient to show that with high probability, there exists sufficiently many rows $v \in \mathcal{S}_u^\beta(i)$ such that $d_{\mathcal{X}_1}(x_1(u), x_1(v))$ is small, which implies the differences in their functional values are small, and thus the sample variance of their differences is also small. This is directly related to the function ϕ_1, since it lower bounds the probability for the latent variable $x_1(v)$ to be sampled within a r-radius ball around $x_1(u)$ such that $d_{\mathcal{X}_1}(x_1(u), x_1(v)) < r$.

For example, suppose $P_{\mathcal{X}_1}$ is a uniform measure over a unit cube in d dimensional Euclidean space. Then

$$\phi_1(r) := \inf_{x_0 \in \mathcal{X}_1} P_{\mathcal{X}_1}\left(d_{\mathcal{X}_1}(\mathbf{x}, x_0) \leq r\right) = \min\left\{1, \left(\frac{r}{2}\right)^d\right\}.$$

Due to the uniform random sampling of latent features, for any x_0, this expressions shows that the number of rows we need to sample, in order to guarantee that there

is at least one row whose latent feature is within distance r of x_0, scales as $\Omega(r^{-d})$. Thus $m \gg r^{-d}$ for $r \to 0$, which implies that we need $d = o(\log m)$.

Corollary 4.3 presents simplified expressions for the upper bound on the MSE when $\mathcal{X}_1$ is a unit cube of $\mathbb{R}^d$ with uniform probability measure $P_{\mathcal{X}_1}$. For readability, we additionally assume that m and n are large enough such that $F_1 \leq 1/2$.

COROLLARY 4.3 (uniform). *When the latent space is a cube in $\mathbb{R}^d$ equipped with the uniform probability measure, as long as $p \geq \max\left\{ m^{-1+\delta}, n^{-\frac{1}{2}+\delta} \right\}$ for some $\delta > 0$, the mean-squared error (MSE) of the estimate produced by the user-user k-nearest neighbor algorithm with $\beta = np^2/2$ and $k = \frac{1}{8}(mp)^{1/3}$ is upper bounded by:*

$$MSE(\hat{A}) \leq \frac{21}{4}F_1 \ln\left(\frac{2B_0}{F_1}\right) + 4B_0^2 F_2,$$

where

$$F_1 = C \max\left\{ (mp)^{-4/3d}, (np^2)^{-1/3}, (mp)^{-1/3} \right\},$$
$$F_2 = 3\exp\left(-c_1(mp)^{1/3}\right) + 6m\exp\left(-c_2(np^2)^{1/3}\right),$$

where C, c_1, c_2 are absolute constants. Additionally assuming that $d = o(\log m)$ and $\log m < (np^2)^{\delta'/3}$ for some $\delta' \in (0,1)$, then the estimator $\hat{A}$ is consistent, i.e., $MSE(\hat{A}) \to 0$ as $m, n \to \infty$.

When our matrix is square , i.e. $m = n$, then it always holds that $mp \geq np^2$, since $p \leq 1$. The bound on the MSE then reduces to

$$MSE(\hat{A}) \leq C \max\left\{ (mp)^{-\frac{4}{3d}}, (mp^2)^{-\frac{1}{3}} \right\},$$

for some constant C up to a logarithmic factor. If we choose the smallest $p = m^{-1/2+\delta}$, we can compare $(mp)^{-4/3d}$ and $(mp^2)^{-1/3}$ to show that $(mp)^{-4/3d} = m^{-\frac{2(1+2\delta)}{3d}} \geq m^{-\frac{2\delta}{3}} = (mp^2)^{-1/3}$ if and only if $\delta \geq \frac{1}{d-2}$.

In an even simpler setting where $P_{\mathcal{X}_1}$ (or $P_{\mathcal{X}_2}$) is supported only on a finite number of points, i.e. there are only finitely many latent row types, then the error convergence rate and the sample complexity have *no* dependency on dimension of $\mathcal{X}_1$ and $\mathcal{X}_2$. This follows from the fact that $\phi_1(r)$ is bounded below by a constant even for $r \to 0$. Corollary 4.4 presents simplified expressions for the MSE upper bounds when $\mathcal{X}_1$ is a unit cube of $\mathbb{R}^d$ with uniform probability measure $P_{\mathcal{X}_1}$. For readability, we additionally assume that m and n are large enough such that $F_1 \leq 1/2$.

COROLLARY 4.4 (discrete). *When the latent space consists of a finite number of points, in other words, $P_{\mathcal{X}_1}$ is supported only on a finitely many points, as long as $p \geq \max\left\{m^{-1+\delta}, n^{-\frac{1}{2}+\delta}\right\}$ for some $\delta > 0$, the mean-squared error (MSE) of the estimate produced by the user-user k-nearest neighbor algorithm with $\beta = np^2/2$ and $k = \frac{1}{8}(mp)^{1/3}$ is upper bounded by:*

$$MSE(\hat{A}) \leq \frac{21}{4}F_1 \ln\left(\frac{2B_0}{F_1}\right) + 4B_0^2 F_2,$$

where

$$F_1 = C \max\left\{(np^2)^{-1/3}, (mp)^{-1/3}\right\},$$
$$F_2 = 3\exp\left(-c_1 (mp)^{1/3}\right) + 6m\exp\left(-c_2 (np^2)^{1/3}\right),$$

where C, c_1, c_2 are absolute constants. Additionally assuming that $\log m < (np^2)^{\delta'/3}$ for some $\delta' \in (0,1)$, then the estimator $\hat{A}$ is consistent, i.e., $MSE(\hat{A}) \to 0$ as $m, n \to \infty$.

4.2. *Comparison with the USVT estimator.* Our result can be compared with the upper bound on the MSE for the UVST estimator as presented in Theorem 2.7 of Chatterjee (2015). For simplicity, consider the setting of a square matrix, i.e. $m = n$. For a matrix sampled from the latent variable model with latent variable dimension d, their theorem guarantees that

$$(4.3) \qquad MSE(\hat{A}^{USVT}) \leq C\frac{m^{-\frac{1}{d+2}}}{\sqrt{p}}$$

for some constant C as long as $p \geq m^{-1+\delta}$. This upper bound is meaningful only when $p > m^{-\frac{2}{d+2}}$, because the MSE bound in (4.3) is bounded below by C when $p \leq m^{-\frac{2}{d+2}}$. However, requiring $p > m^{-\frac{2}{d+2}}$ can be too restrictive when the latent dimension d is large since it means that we need to sample almost every entry to achieve a nontrivial bound.

In contrast, when $d = o(\log m)$, our algorithm and analysis provides a vanishing upper bound on the MSE whenever $p \geq \max\left\{m^{-1+\delta}, n^{-1/2+\delta}\right\}$, surprisingly independent of the latent dimension. In fact, our analysis guarantees that our algorithm achieves a vanishing MSE even as d grows with m as long as $d = o(\log m)$.

This difference in the provided sample complexity for the USVT spectral method and our similarity based method is likely due to the fact that our analysis essentially relies on "local" structure. Even if the latent dimension increases, we only need to ensure that there are sufficiently many close neighbor points. On the other hand,

Chatterjee's result stems from showing that a Lipschitz function can be approximated by a piecewise constant function, which upper bounds the rank of the (approximate) target matrix. This global discretization results in a large penalty with regards to the dimension of the latent space.

In other aspects, the results of Chatterjee (2015) are also more general in that they do not require that latent features are generated i.i.d. according to an underlying distribution, i.e. they could be arbitrarily generated. They additionally can handle more general noise models, as long as the data is bounded, while our analysis requires that the noise terms have uniform variance equal to γ^2.

4.3. *Discussion.*   We discuss strengths and limitations of our results followed by some natural directions for future work. One limitation of our proof is that we assumed an additive noise model in (2.1) and (2.2), where the individual noise terms $\eta(u, i)$ are independent with zero mean, identical variance and bounded support. The result of Chatterjee (2015) holds in the independent zero mean bounded noise setting, allowing the variance of the noise across entries to be different. In that sense, our assumption on the noise model is restrictive.

It is also not clear if our result is tight or not, as we do not know of information theoretic lower bounds for the MSE under the general latent variable model considered. For specific settings such as when the function f when considered as an integral operator has finite spectrum, it is equivalent to low-rank models, for which lower bounds have been characterized. For specific noise models such as the binary observation model which corresponds to the graphon generative model for random graphs, Gao, Lu and Zhou (2015); Klopp, Tsybakov and Verzelen (2015) show that variants of the least squares estimator achieve optimal rates, but unfortunately they are not polynomial time computable.

From an implementation perspective, the similarity based algorithm proposed in this work, similar to classical collaborative filtering methods, is easy to implement and scales extremely well to large datasets, as it naturally enjoys a parallelizable implementation. Furthermore, the operation of finding k nearest neighbors can benefit from computational advances in building scalable approximate nearest neighbor indices, cf. Indyk (2001, 2004).

Next we discuss some natural extensions and directions for future work. In our model, the latent function f is assumed to be Lipschitz. However, the proof only truly utilizes the fact that "locally" the function value does not oscillate too wildly. Intuitively, this suggests that the result may extend to a broader class of functions, beyond Lipschitz functions. For example, a function with bounded Fourier coefficients does not oscillate too wildly, and thus it may behave well for the purposes of analyzing our algorithm.

Another possible direction for extension is related to the measurement of simi-

larity and the sample complexity. Our current algorithm measures the similarity of rows u and v from their overlapping observed entries, which critically determines the sample complexity requirement of $np^2 \gg 1$. However, for sparser datasets without overlaps, we may be able to reveal the similarity by instead comparing distribution signatures such as moments or comparing them through their "extended" neighborhoods.

As a concluding remark, we would like to mention that the latent variable model is a fairly general model and there is a large body of related applications. Some of the popular recent examples, which are special cases of latent variable model, include Stochastic blockmodels for community detection, the Bradley-Terry model for ranking from pair-wise comparison data and the Dawid-Skene model for low-cost crowd sourcing. Another prominent example of latent variable model is the generative model for random graphs referred to as a Graphon, which has been shown to be the limit of a sequence of graphs. We refer interested readers to (Chatterjee, 2015, Section 2.4) for an excellent overview on the broad applicability of the latent variable model.

5. Extending Beyond Matrices to Tensors. A natural extension beyond matrix completion is to higher order tensor completion. Given an unknown t-order tensor T of interest with dimensions $n_1 \times \cdots \times n_t$, suppose that we observe a random subset of noisy observations of its entries. Similar to matrix completion, the goal in tensor completion is to estimate the missing entries in the tensor from the noisy partial observations, as well as to "de-noise" the observed entries. The tensor completion problem is important within a wide variety of applications, including recommendation systems, multi-aspect data mining Kolda and Sun (2008); Sun et al. (2009), and machine vision Liu et al. (2013a); Zhang et al. (2014); Ravi et al. (2013).

Although tensor completion has been widely studied, there is still a wide gap in understanding, unlike matrix completion. This gap partially stems from the hardness of tensor decomposition, as most recovery methods rely on retrieving hidden algebraic structure through the framework of low-rank factorization. Tensors do not have a canonical decomposition such as the singular value decomposition (SVD) for a matrix.

There is a factorization scheme, namely the CANDECOMP/PARAFAC (CP) decomposition, which factorizes the tensor as a sum of rank-1 tensors (outer product of vectors). However, it is known that finding the rank of a tensor is NP-Complete, which makes it computationally intractable. Also, there are known ill-posedness De Silva and Lim (2008) issues with CP-based low-rank approximation.

There are other kinds of decompositions such as the Tucker decomposition. Approaches based on Tucker decomposition essentially unfold (matricize or flatten)

the tensor, and make use of matrix completion theory and methods Gandy, Recht and Yamada (2011); Signoretto et al. (2011); Tomioka et al. (2011); Liu et al. (2013a); Mu et al. (2014).

5.1. *Latent Variable Model for Tensor.* The nonparametric blind regression setup presented in Section 2 naturally extends beyond bivariate functions which correspond to matrices, to higher dimensional functions encompassing higher-order tensors. We set up the formal latent variable model for a tensor following similar assumptions as stated in Section 2.

Consider a t-order tensor $T_A \in \mathbb{R}^{n_1 \times n_2 \times \dots n_t}$. Consider a vector $\vec{\alpha} = (\alpha_1, \dots, \alpha_t) \in [n_1] \times \dots \times [n_t]$ indexing a position in the tensor T_A. Let the coordinate α_q in the qth dimension of the tensor be associated to a latent feature $x_q(\alpha_q)$ drawn i.i.d from the space $\mathcal{X}_q$ according to probability measure $P_{\mathcal{X}_q}$, for $q \in [t]$ and $\alpha_q \in [n_q]$. Assume $\mathcal{X}_q$ is a compact metric space with metric $d_{\mathcal{X}_q}$ and diameter $D_{\mathcal{X}_q}$. Define $\phi_q : \mathbb{R}_+ \to [0, 1]$ to be a function lower bounding the cumulative distribution function according to the distance in the latent space. For any radius $r > 0$, we define it according to

$$(5.1) \qquad \phi_q(r) := \text{ess} \inf_{x_0 \in \mathcal{X}_q} P_{\mathcal{X}_q}(B_q(x_0, r)),$$

where $B_q(x_0, r) := \{x \in \mathcal{X}_q : d_{\mathcal{X}_q}(x_0, x) \leq r\}$. Then $T_A(\vec{\alpha})$, the value in tensor T_A corresponding to position $\vec{\alpha}$, is equal to

$$(5.2) \qquad T_A(\vec{\alpha}) = f(x_1(\alpha_1), \dots, x_t(\alpha_t))$$

for a latent function $f : \mathcal{X}_1 \times \dots \times \mathcal{X}_t \to \mathbb{R}$. Assume that the function f is L-Lipschitz over the latent spaces:

$$(5.3) \quad |f(x_1, \dots x_t) - f(x'_1, \dots x'_t)| \leq L \max \left(d_{\mathcal{X}_1}(x_1, x'_1), \dots d_{\mathcal{X}_t}(x_t, x'_t) \right),$$

for all $(x_1, \dots x_t) \in \mathcal{X}_1 \times \dots \times \mathcal{X}_t$ and $(x'_1, \dots x'_t) \in \mathcal{X}_1 \times \dots \times \mathcal{X}_t$. Let $T_Z \in \mathbb{R}^{n_1 \times n_2 \times \dots n_t}$ be a noisy tensor derived from T_A by adding independent noise to each entry $\vec{\alpha} \in [n_1] \times \dots \times [n_t]$ according to

$$(5.4) \qquad T_Z(\vec{\alpha}) = T_A(\vec{\alpha}) + \eta(\vec{\alpha}),$$

where the additive noise terms $\eta(\vec{\alpha})$ are independent with bounded support, zero mean, and variance equal to γ^2, as assumed in the setup presented in Section 2. Let $\mathcal{D} \subset [n_1] \times \dots \times [n_t]$ denote the index set of observed entries. We assume that each entry $T_Z(\vec{\alpha})$ is observed independently with probability $p \in [0, 1]$, and otherwise unobserved or missing. Therefore for any $\vec{\alpha} \in [n_1] \times \dots \times [n_t]$, $\vec{\alpha} \in \mathcal{D}$ with probability p, and $\vec{\alpha} \notin \mathcal{D}$ with probability $1 - p$.

Exchangeability Revisited. Our data structure is a tensor of order t which satisfies t-order exchangeability, i.e. for all $(\alpha_1 \ldots \alpha_t) \in [n_1] \times \cdots \times [n_t]$,

$$(5.5) \qquad T_Z(\alpha_1, \ldots \alpha_t) \stackrel{d}{=} T_Z(\pi_1(\alpha_1), \ldots \pi_t(\alpha_t)),$$

for all permutations $\pi_1 \ldots \pi_t$. This follows from our assumption that the latent variables associated to each coordinate of the tensor are drawn i.i.d. according to the probability measures $P_{\mathcal{X}_1} \ldots P_{\mathcal{X}_t}$. Aldous and Hoover's representation theorem states than a t-order exchangeable array can be represented as a measurable function of 2^t independent random sources, each of which corresponds to a member in the power set of $[t]$, cf. see Austin (2012).

The additive noise model which we assume in (5.2) and (5.4) enforces a smaller subclass within t-order exchangeable models, reducing the function to only depend upon $t + 2$ independent sources of randomness according to

$$(5.6) \qquad T_Z(\vec{\alpha}) \stackrel{d}{=} f_\theta(\theta_1, \ldots \theta_t, \theta_{\vec{\alpha}}),$$

where θ is a universal parameter indexing the function f_θ, $(\theta_1, \ldots \theta_t)$ are latent features associated to each dimension, and $\theta_{\vec{\alpha}}$ is an individual noise term. Given a particular instance of the dataset, θ is realized and our goal is to estimate the outputs of the function f_θ. Our model additionally restricts the class to consider Lipschitz functions f_θ and additive noise model, such that

$$(5.7) \qquad f_\theta(\theta_1, \ldots \theta_t, \theta_{\vec{\alpha}}) = f(\theta_1, \ldots \theta_t) + \eta(\vec{\alpha}).$$

5.2. *Tensor Completion by Flattening to Matrix.* Given the setup, one could imagine similarly constructing a neighborhood based algorithm exploiting local approximations, exploiting the intricate structure within higher-order tensors to provide a meaningful estimation algorithm. In this section we suggest a simple algorithm for tensor completion which follows naturally from flattening the tensor to matrix and applying our matrix completion algorithm. While this algorithm itself is not a surprise, we are additionally able to provide theoretical guarantees showing consistency of the resulting estimator. This provides a starting point for analyzing similarity based algorithms for tensor completion, however, it may be possible to improve the efficiency of the estimator by designing an algorithm which directly acts on the tensor itself.

Given a tensor T_A, we discuss how to flatten it to a corresponding matrix A. Let $(\mathcal{I}_1, \mathcal{I}_2)$ be a bi-partition of $[t]$ such that $\mathcal{I}_1 = \{\pi(1), \ldots, \pi(t_1)\}$ and $\mathcal{I}_2 = \{\pi(t_1 + 1), \ldots, \pi(t)\}$ for some $1 \leq t_1 \leq t - 1$ and some permutation $\pi : [t] \to [t]$. We can reduce the tensor T_A to a matrix A by "flattening", i.e. taking the cartesian product of all dimensions in $\mathcal{I}_1$ to be the rows of the matrix and taking the cartesian

product of all dimensions in $\mathcal{I}_2$ to be the columns of the matrix. The resulting flattened matrix has dimension $m \times n$ where $m = \prod_{q \in \mathcal{I}_1} n_q$ and $n = \prod_{q \in \mathcal{I}_2} n_q$. Given this matrix or flattened tensor, the resulting latent row features belong to space $\mathcal{X}_1^\pi = \times_{q \in \mathcal{I}_1} \mathcal{X}_q$, and the latent column features belong to $\mathcal{X}_2^\pi = \times_{q \in \mathcal{I}_2} \mathcal{X}_q$. The underestimator function of the product probability measure for the row features is denoted by $\phi_1^\pi(r) = \prod_{q \in \mathcal{I}_1} \phi_q(r)$. Let the observation matrix Z be equivalently obtained by flattening tensor T_Z according to the same partition $(\mathcal{I}_1, \mathcal{I}_2)$.

For notational simplicity, we rearrange and relabel the indices to map between the tensor and matrix indices. For a tensor index $\vec{\alpha} = (\alpha_1, \ldots, \alpha_t) \in [n_1] \times \cdots \times [n_t]$, we equate it to a matrix row and column index pair $\vec{\alpha} = (\vec{u}, \vec{i})$, where $\vec{u} = (u_1, \ldots, u_{t_1}) = (\alpha_{\pi(1)}, \ldots, \alpha_{\pi(t_1)})$ and $\vec{i} = (i_1, \ldots, i_{t_2}) = (\alpha_{\pi(t_1+1)}, \ldots, \alpha_{\pi(t_1+t_2)})$ with $t_2 := t - t_1$ ($1 \leq t_1 \leq t - 1$). Although $\vec{u} \in \times_{q \in \mathcal{I}_1} [n_q]$ is a t_1-tuple of positive integers, we will sometimes identify the vector $\vec{u}$ with an integer $u \in [m]$ according to

$$u = 1 + \sum_{\tau \in [t_1]} \left((u_\tau - 1) \prod_{s=\tau+1}^{t_1} n_{\pi(s)} \right).$$

We use this notion when u refers to a row in the matrix which corresponds to the flattened tensor. Equivalently, i will be sometimes identified with an integer $i \in [n]$ according to

$$i = 1 + \sum_{\tau \in [t_2]} \left((i_\tau - 1) \prod_{s=\tau+1}^{t_1} n_{\pi(t_1+s)} \right).$$

The corresponding row (i.e. "user") and column (i.e. "item") features are denoted by

$$\vec{x}_1^\pi(\vec{u}) = \left(x_{1,1}^\pi(u_1), \ldots, x_{1,t_1}^\pi(u_{t_1}) \right) = \left(x_{\pi(1)}(u_1), \ldots, x_{\pi(t_1)}(u_{t_1}) \right),$$
$$\vec{x}_2^\pi(\vec{i}) = \left(x_{2,1}^\pi(i_1), \ldots, x_{2,t_2}^\pi(i_{t_2}) \right) = \left(x_{\pi(t_1+1)}(i_1), \ldots, x_{\pi(t_1+t_2)}(i_{t_2}) \right).$$

We let $\mathcal{X}_{1,k}^\pi$ denote $\mathcal{X}_{\pi(k)}$ for $k \in [t_1]$, and we let $\mathcal{X}_{2,k}^\pi$ denote $\mathcal{X}_{\pi(t_1+k)}$ for $k \in [t_2]$.

It follows that if the tensors T_A and T_Z are drawn from the tensor latent variable model described in Section 5.1, then the associated matrices A and Z, obtained by flatting the tensors T_A and T_Z, follow a similar form as introduced in Section 2:

$$A(u, i) = f(\vec{x}_1^\pi(\vec{u}), \vec{x}_2^\pi(\vec{i}))$$
$$Z(u, i) = A(u, i) + \eta(u, i),$$

where $\eta(u, i) = \eta(\vec{\alpha})$ for the correct index mapping $\vec{\alpha} = (\vec{u}, \vec{i})$. We can verify that A and Z satisfy all the assumptions needed in Section 2 except for the assumption that the row and column latent variables are independently sampled. This is due to the fact that the associated latent variables of $\vec{x}_1^\pi(\vec{u})$ and $\vec{x}_1^\pi(\vec{u}')$ are correlated

if any component of $\vec{u}$ and $\vec{u}'$ are the same. If $u_\tau = u'_\tau$ for any $\tau \in [t_1]$, then $x^\pi_{1,\tau}(u_\tau) = x^\pi_{1,\tau}(u'_\tau)$. However if all the components of $\vec{u}$ and $\vec{u}'$ are distinct, then $\vec{x}^\pi_1(\vec{u})$ and $\vec{x}^\pi_1(\vec{u}')$ will be independent. In fact the structure of the correlations is very specific and follows from the flattening of the tensor to a matrix. Therefore, we will show in the following sections that the estimate which results from applying our matrix completion algorithm to the flattened tensor in fact leads to a consistent tensor completion algorithm.

5.3. *Tensor Completion Algorithm.* Given partial observations from the noisy tensor T_Z, we construct an estimator for the desired tensor T_A by flattening the observation tensor to a matrix Z and applying the user-user k-nearest neighbor algorithm presented in Section 3 on the observed entries of Z to construct an estimated matrix $\hat{A}$ of the corresponding flattened matrix A. The only modification we introduce, purely for the purposes of our analysis, is that when we compute the overlap entries, we remove any entries which share a coordinate in the original tensor representation. Formally speaking, we define a set $\mathcal{O}^{uv}_i \subset \mathcal{O}^{uv}$ such that

(5.8) $$\mathcal{O}^{uv}_i := \{j \in \mathcal{O}^{uv} \ s.t. \ j_k \neq i_k \text{ for all } k \in [t_2]\}.$$

Let $\mathcal{N}_i \subset [n]$ denote the set of columns which do not share a tensor coordinate with i,

$$\mathcal{N}_i := \{j \in [n] \ s.t. \ j_k \neq i_k \text{ for all } k \in [t_2]\}.$$

Recall that we are overloading the notation such that a column index i is associated to a corresponding vector $\vec{i} = (i_1, \ldots i_{t_2})$ which denotes the original tensor coordinates. Therefore, the set $\mathcal{O}^{uv}_i$ restricts to columns j which do not share any coordinates in the original tensor representation, which translates to the condition of $j_k \neq i_k$ for all $k \in [t_2]$. Then the set $\mathcal{O}^{uv}_i$ is used instead of $\mathcal{O}^{uv}$ for computing the sample means and variances. Since the calculation now depends on i, we will denote the means and variances computed from the set $\mathcal{O}^{uv}_i$ by $m_{uv}(i)$ and $s^2_{uv}(i)$.

We will also change the set of row indices that are chosen in Step 3 to be part of the computation. In order to estimate entry (u, i) we find all rows v such that the overlap $|\mathcal{O}^{uv}_i|$ is larger than β_l and smaller than β_h, and the row contains information about column i. Formally, we replace the definition in (3.5) with the set $\mathcal{S}^{\beta_l,\beta_h}_u(i)$:

(5.9) $$\mathcal{S}^{\beta_l,\beta_h}_u(i) = \{v \in [m] : Z(v,i) \text{ is observed, and } \beta_l \leq |\mathcal{O}^{uv}_i| \leq \beta_h\}.$$

Then the final estimate is computed by averaging over the i-th entry of the k rows in $\mathcal{S}^{\beta_l,\beta_h}_u(i)$ which have minimum sample variance $s^2_{uv}(i)$ amongst rows $v \in \mathcal{S}^{\beta_l,\beta_h}_u(i)$.

The difference is that we now require that $|\mathcal{O}_i^{uv}|$ is also upper bounded by a parameter β_h. Intuitively having a larger overlap should only help, but this is introduced purely for an analytical purpose, when showing concentration of the sample means and variances. This upper bound is not restrictive, because we can always subsample from the overlap when our data matrix is dense, and then apply boosting to construct an averaged estimator. Since each individual estimator obtained from the subsampled data is consistent, the averaged estimator is also consistent. This argument also implies that the upper bound on the sample probability p in Theorem 5.1 can be relaxed.

The first modification of removing columns in $\mathcal{O}^{uv}$ which are correlated with i is needed for the current analysis so that the choice of the k nearest neighbor rows are independent from the latent variables associated to the column i. This increases the computation complexity, as the means and variances for a pair of rows (u, v) must be computed for each i. In a practical implementation, we would simply use the original matrix algorithm, and we show in experiments presented in Section 6 that this performs well. In fact, we believe that a modified proof would be able to circumvent this fix by showing that the latent variables associated to column i are only marginally correlated with the set of k nearest neighbor rows chosen, since the fraction of columns correlated to column i goes to zero as the size of the tensor increases.

5.4. *Results.* We provide bounds on the mean squared error for the estimate produced by applying our method to a flattened tensor, drawn from the latent variable model. The results show that in fact the estimate is consistent, i.e. the mean squared error converges to zero.

Recall that $(\mathcal{I}_1, \mathcal{I}_2)$ denote the partition used in flattening the matrix. Therefore the matrix dimensions are $m = \prod_{q \in \mathcal{I}_1} n_q$ and $n = \prod_{q \in \mathcal{I}_2} n_q$, and the number of columns after removing one coordinate from each tensor dimension is denoted by $n' = \prod_{q \in \mathcal{I}_2} (n_q - 1)$. The algorithm uses parameters β_l and β_h to set the upper and lower thresholds for the overlap between pairs of users, and it uses k as a smoothing parameter, specifying the number of neighbors the algorithm averages over to obtain the final estimates. p is the probability that each entry of the tensor is observed, and ϕ_q is the underestimator function for the latent probability space associated to the latent features of dimension q in the original tensor. ϕ_1^π is the underestimator function associated to the matrix row product space, taking the form of $\phi_1^\pi(r) = \prod_{q \in \mathcal{I}_1} \phi_q(r)$.

THEOREM 5.1 (Main theorem for tensor completion). *Suppose that*

$$\max\left\{m^{-1+\delta}, n'^{-\frac{1}{2}+\delta}\right\} \leq p \leq n'^{-\frac{1}{6}-\delta} \quad \textit{for some } \delta > 0,$$

$$\forall q \in \mathcal{I}_1, \ \zeta \textit{ satisfies } \phi_q\left(\sqrt{\frac{\zeta}{L^2}}\right) \geq c_q n_q^{-\frac{\log mp}{2 \log m}} \textit{ for some } c_q > 0,$$

$$2 \leq \beta_l \leq c_l \min\left\{n'p^2, n'^{1/2}\right\} \textit{ for some } c_l \in (0,1),$$

$$c_h \max\left\{n'^{1/2}, n'p^2\right\} \leq \beta_h \leq n'^{\frac{2}{3}-\delta} \textit{ for some } c_h > 1 \textit{ and}$$

$$k \leq \frac{1}{8}mp\phi_1^\pi\left(\sqrt{\frac{\zeta}{L^2}}\right) = \frac{p}{8}\prod_{q \in \mathcal{I}_1} n_q \phi_q\left(\sqrt{\frac{\zeta}{L^2}}\right).$$

The mean squared error (MSE) of the estimate produced by applying the user-user k-nearest neighbor variant of our method on a flattened tensor, using overlap parameters β_l and β_h, is upper bounded by:

$$(5.10) \qquad MSE(\hat{A}) \leq 2F_1' \ln\left(\frac{2B_0}{F_1'}\right) + \left(F_1' + F_2'\right)^2 + 2F_2' + 4B_0^2 F_3',$$

where

$$F_1' = \frac{(1+\theta)\zeta + 2F_2'}{1-\theta} + \frac{\gamma^2}{k},$$

$$F_2' = \max\left\{(n_{q^*} - 1)^{-1/3}, \left(\frac{n'^2}{\beta_h^3}\right)^{-1/3}, \beta_l^{-1/3}\right\}, \textit{ and}$$

$$F_3' = 4(m-1)\exp\left(-c_1 n'p^2\right) + 2\exp\left(-\frac{1}{24}mp\right)$$

$$+ 6(m-1)p\exp\left(-c_2(n_{q^*}-1)^{1/3}\right) + 6(m-1)p\exp\left(-c_3 \min\left\{\frac{n'^{2/3}}{4\beta_h}, \frac{\beta_l^{1/3}}{4}\right\}\right)$$

$$+ t_1 \exp\left(-c_4 n_{q^*}^{1/2}\right) + \exp\left(-\frac{k}{8}\right),$$

with $q^* := \arg\min_{q \in \mathcal{I}_1} n_q$ and

$$c_1 := \min\left\{ \frac{(1-c_l)^2}{2}, \frac{(c_h-1)^2}{3} \right\},$$

$$c_2 := \min\left\{ \frac{1}{8L^2D^2t_2 + 16B_e^2}, \frac{1}{32L^2D^2(3LD+4B_e)^2t_2 + 64B_e^2(2LD+5B_e)^2} \right\},$$

$$c_3 := \min\left\{ \frac{1}{32(LD+2B_e)^2}, \frac{1}{128(LD+2B_e)^4} \right\},$$

$$c_4 := \min_{q \in \mathcal{I}_q}\left\{ \frac{(1-2^{-1/t_1})^2}{2} c_q \right\}$$

being some absolute constants, which may depend only on the geometry of the latent spaces. Also, $\theta = \sum_{q \in \mathcal{I}_2} \frac{1}{n_q-1}$ is a quantity which depends only on the shape of the given tensor and vanishes to 0 as $n_q \to \infty, \forall q \in \mathcal{I}_2$.

The statement is similar to the matrix completion result stated in Theorem 4.1, but with slightly more restrictive conditions and more complex error bound terms. This complication mainly stems from a two-step analysis used to prove the concentration of the sample means and variances **?**. This is introduced to handle the correlations amongst the latent features in flattened tensor. The additional upper bound on the sample complexity, $p \leq n'^{-\frac{1}{6}-\delta}$, is not restrictive, because we can always build multiple sparser matrices by subsampling the data to satisfy the constraint, and then using boosting to combine the outputs from each of the subsampled datasets to obtain a final estimate.

5.5. *Optimal Flattening to Balance Matrix Dimensions.* The stated theorem results depend on the choice of partitions $(\mathcal{I}_1, \mathcal{I}_2)$ used for flattening the tensor to a matrix. Different choices of partitions will affect the row and column dimensions of the resulting matrix, i.e. m and n, which impact both the convergence rates and sample complexity stated in Theorem 5.1. Therefore, it is natural to ask if there is an optimal partition for flattening.

We will specifically define a partition $(\mathcal{I}_1, \mathcal{I}_2)$ to be user-optimal if it minimizes the required sample complexity of the user-user algorithm. In order to guarantee that the user-user method produces a consistent estimator, Theorem 5.1 requires that the fraction of observed datapoints $p \geq \max\left\{m^{-1+\delta}, n'^{-\frac{1}{2}+\delta}\right\}$. Since $n' = \Theta(n)$, this is equivalent to requiring that $p \geq \max\left\{m^{-1+\delta}, n^{-\frac{1}{2}+\delta}\right\}$, as $\frac{n}{n'} \to 1$ as $n \to \infty$. The equivalent theorem for the item-item method would require $p \geq \max\left\{n^{-1+\delta}, m^{-\frac{1}{2}+\delta}\right\}$. An optimal partition would minimize the lower bound required on p to guarantee consistency.

DEFINITION 5.1 (optimal flattening). *Given a t-order tensor $T \in \mathbb{R}^{n_1 \times \cdots \times n_t}$, a partition $(\mathcal{I}_1^*, \mathcal{I}_2^*)$ of $[t]$ is user-optimal if for any other partition $(\mathcal{I}_1, \mathcal{I}_2)$ of $[t]$,*

$$\max\left\{\left(\prod_{q \in \mathcal{I}_1^*} n_q\right)^{-1}, \left(\prod_{q \in \mathcal{I}_2^*} n_q\right)^{-1/2}\right\} \leq \max\left\{\left(\prod_{q \in \mathcal{I}_1} n_q\right)^{-1}, \left(\prod_{q \in \mathcal{I}_2} n_q\right)^{-1/2}\right\}.$$

Switching the roles of $\mathcal{I}_1^*$ and $\mathcal{I}_2^*$, we define $(\mathcal{I}_2^*, \mathcal{I}_1^*)$ to be item-optimal if $(\mathcal{I}_1^*, \mathcal{I}_2^*)$ is user-optimal.

As an optimal partition only depends on the product of the dimensions in the two partitions, there may not be a unique optimal partition. For example, consider an equilateral t-order tensor with dimension l, i.e., a tensor with $n_q = l, \forall q \in [t]$. The following lemma states that any partition $(\mathcal{I}_1, \mathcal{I}_2)$ of $[t]$ which satisfies $|\mathcal{I}_2| = \lfloor \frac{2t}{3} \rfloor$ is user-optimal.

LEMMA 5.2. *For an equilateral t-order tensor, any partition $(\mathcal{I}_1, \mathcal{I}_2)$ of $[t]$ which satisfies $|\mathcal{I}_2| = \lfloor \frac{2t}{3} \rfloor$ is user-optimal.*

PROOF. Let l denote the dimension of the given equilateral tensor. The lower bound (threshold) on the required sample complexity for the user-user method is given by $p^* = \max\left\{m^{-1}, n^{-1/2}\right\} = \left\{l^{-|\mathcal{I}_1|}, l^{-\frac{1}{2}|\mathcal{I}_2|}\right\}$. Taking log with respect to l yields $\log_l p^* = \max\left\{-|\mathcal{I}_1|, -\frac{1}{2}|\mathcal{I}_2|\right\}$. Since t must be a positive integer, either $t \equiv 0, 1,$ or $2 \pmod 3$.

If $t = 3k$ for some $k \in \mathbb{Z}_+$, $|\mathcal{I}_2| = \lfloor \frac{2t}{3} \rfloor = 2k$, and $|\mathcal{I}_1| = k$.

If $t = 3k + 1$ for some $k \in \mathbb{Z}_+$, $|\mathcal{I}_2| = \lfloor \frac{2t}{3} \rfloor = 2k$, and $|\mathcal{I}_1| = k + 1$.

If $t = 3k + 2$ for some $k \in \mathbb{Z}_+$, $|\mathcal{I}_2| = \lfloor \frac{2t}{3} \rfloor = 2k + 1$, and $|\mathcal{I}_1| = k + 1$.

It is easy to observe that perturbing $|\mathcal{I}_2|$ from this value results in increasing the threshold from the expression of $\log_l p^*$ above. Therefore, the choice of $|\mathcal{I}_2| = \lfloor \frac{2t}{3} \rfloor$ is user-optimal. $\qquad \square$

Given an equilateral tensor of order t, i.e., $n_q = l, \forall q \in [t]$, the total number of entries in the tensor is $N = l^t$. The above lemma implies that if we apply the user-user algorithm on the user-optimally flattened tensor, the required fraction of observed samples in order to guarantee consistency is lower bounded by

$$(5.11) \qquad p^* = n^{-1/2} = (l)^{-\frac{1}{2}|\mathcal{I}_2|} = N^{-\frac{\lfloor \frac{2t}{3} \rfloor}{2t}},$$

which converges to $N^{-1/3}$ as t goes to infinity. In the following corollary, we state the simplified error convergence rate for the user-user algorithm applied to a user-optimally flattened equilateral tensor, in the case when the latent features are sampled uniformly from a cube in $\mathbb{R}^d$.

COROLLARY 5.3 (uniform, user-optimally flattened, equilateral tensor).   *Given an equilateral t-order tensor where $n_q = l$ for all $q \in [t]$, when each latent space is a cube in $\mathbb{R}^d$ equipped with the uniform probability measure, as long as $\max\left\{m^{-1+\delta}, n'^{-\frac{1}{2}+\delta}\right\} \leq p \leq n'^{-\frac{1}{6}-\delta}$ for some $\delta > 0$, the user-user k-smoothed variant of our method applied to the user-optimally flattened tensor with $\beta_l = \frac{1}{2}\min\left\{n'p^2, \sqrt{n'}\right\}$, $\beta_h = 2\max\left\{n'p^2, \sqrt{n'}\right\}$ and $k = \frac{1}{8}(mp)^{1/2}$ is consistent. Moreover, its mean squared error is bounded by*

$$(5.12) \qquad MSE(\hat{A}) \leq \frac{21}{4}F_1' \ln\left(\frac{2B_0}{F_1'}\right) + 4B_0^2 F_2',$$

where

$$F_1' = C\max\left\{(mp)^{-\frac{1}{d|\mathcal{I}_1|}}, (l-1)^{-\frac{1}{3}}, \frac{\beta_h}{(l-1)^{\frac{2}{3}|\mathcal{I}_2|}}, \beta_l^{-\frac{1}{3}}, (mp)^{-\frac{1}{2}}\right\},$$

$F_2' = $ *same as the exponentially decaying term F_3' defined in Theorem 5.1*

with some absolute constant C.

6. Experiments.   In this section we present experimental results from applying the User-Item Gaussian Kernel variant of our algorithm to real datasets.

6.1. *Matrix completion.*   We evaluated the performance of our algorithm on predicting user-movie ratings for the MovieLens 1M and Netflix datasets. We chose the overlap parameter $\beta = 2$ to ensure the algorithm is able to compute an estimate for all missing entries. When β is larger, the algorithm enforces rows (or columns) to have more commonly rated movies (or users). Although this increases the reliability of the estimates, it also reduces the fraction of entries for which the estimate is defined.

We compared our method with user-user collaborative filtering, item-item collaborative filtering, and SoftImpute from Mazumder, Hastie and Tibshirani (2010b). We chose the classic mean-adjusted collaborative filtering method, in which the weights are proportional to the cosine similarity of pairs of users or items (i.e. movies). SoftImpute is a matrix-factorization-based method which iteratively replaces missing elements in the matrix with those obtained from a soft-thresholded SVD.

The MovieLens 1M data set contains about 1 million ratings by 6000 users of 4000 movies. The Netflix data set consists of about 100 million movie ratings by 480,189 users of about 17,770 movies. For both MovieLens and Netflix data sets, the ratings are integers from 1 to 5. From each dataset, we generated 100 smaller user-movie rating matrices, in which we randomly subsampled 2000 users and

2000 movies. For each rating matrix, we randomly select and withhold a percentage of the known ratings for the test set, while the remaining portion of the data set is revealed to the algorithm for computing the estimates (or training). After the algorithm computes its predictions for all the missing user-movie pairs, we evaluate the Root Mean Squared Error (RMSE) of the predictions compared to the ratings from the withheld test set. Figure 1 plots the RMSE of our method along with classic collaborative filtering and SoftImpute evaluated against 10%, 30%, 50%, and 70% withheld test sets. The RMSE is averaged over 100 subsampled rating matrices, and 95% confidence intervals are provided.
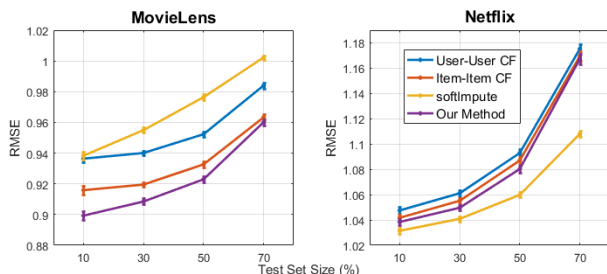

Fig 1: Performance of algorithms on Netflix and MovieLens datasets with 95% confidence interval. λ values used by our algorithm are 2.8 (10%), 2.3 (30%), 1.7 (50%), 1 (70%) for MovieLens, and 1.8 (10%), 1.7 (30%), 1.6 (50%), 1.5 (70%) for Netflix.

Figure 1 suggests that our algorithm achieves a systematic improvement over classical user-user and item-item collaborative filtering. SoftImpute performs worse than all methods on the MovieLens dataset, but it performs better than all methods on the Netflix dataset.

6.2. *Tensor completion.*   We consider the problem of image inpainting for evaluating the performance of tensor completion algorithm. Inpainting is the process of reconstructing lost or deteriorated parts of image or videos. Such methods, in particular, have revitalized the process of recovery old artifacts in museum world which was historically done by conservators or art restorers. An interested reader is referred to a recent survey Ravi et al. (2013) for summary of the state of art on methods and techniques. We compare performance of our algorithm against existing methods in the literature on the image inpainting problem.

An image can be represented as a 3rd-order tensor where the dimensions are rows $\times$ columns $\times$ RGB. In particular we used three images (Lenna, Pepper, and Facade) of dimensions $256 \times 256 \times 3$. For each image, a percentage of the pixels are randomly removed, and the missing entries are filled in by various tensor

completion algorithms.

For the implementation of our tensor completion method, we collapsed the last two dimensions of the tensor (columns and RGB) to reduce the image to a matrix, and applied our method. We set the overlap parameter $\beta = 2$. We compared our method against fast low rank tensor completion (FaLRTC) Liu et al. (2013b), alternating minimization for tensor completion (TenAlt) Jain and Oh (2014), and fully Bayesian CP factorization (FBCP) Zhao, Zhang and Cichocki (2015), which extends the CANDECOMP/PARAFAC(CP) tensor factorization with automatic tensor rank determination.

To evaluate the outputs produced by each method, we computed the relative squared error (RSE), defined as

$$\text{RSE} = \frac{\sum_{i,j,k \in E}(\hat{Z}(i,j,k) - Z(i,j,k))^2}{\sum_{i,j,k \in E}(Z(i,j,k) - \bar{Z})^2},$$

where $\bar{Z}$ is the average value of the true entries. Figure 2 plots the RSE achieved by each tensor completion method on the three images, as a function of the percentage of pixels removed. The results demonstrate that our tensor completion method is competitive with existing tensor factorization based approaches, while maintaining a naive simplicity. In short, a simple algorithm works nearly as good as the best algorithm for this problem!

Fig 2: Performance comparison between different tensor completion algorithms based on RSE vs testing set size. For our method, we set overlap parameter β to 2.

Figure 3 shows a sample of the image inpainting results for the facade and pepper images when 70% of the pixels are removed.

7. Proofs: Matrix Completion. The proof of the main theorem corresponds to showing that the estimates associated to the k-nearest neighbors selected by the algorithm are indeed close to the true target value. Theorem 7.6 provides such a probabilistic tail bound for each (u, i), and integrating this leads to the conclusion of Theorem 4.1.

Original	Degraded	FaLRTC	TenAlt	FBCP	Our Method
		0.0924	0.099	0.12	0.0869
		0.1101	0.1182	0.154	0.109

Fig 3: Recovery results for Facade and Pepper images with 70% of missing entries. RSE is reported under the recovery images.

7.1. *Proof Outline.* For any fixed row latent features $a, b \in \mathcal{X}_1$, and randomly sampled column latent feature variable $\mathbf{x}_2 \sim P_{\mathcal{X}_2}$, we denote the mean and variance of the difference $f(a, \mathbf{x}_2) - f(b, \mathbf{x}_2)$ with respect to $\mathbf{x}_2$ according to

$$\mu_{ab} \triangleq \mathbb{E}_{\mathbf{x}_2}[f(a, \mathbf{x}_2) - f(b, \mathbf{x}_2)] = \mathbb{E}[m_{uv}|\mathbf{x}_1(u) = a, \mathbf{x}_1(v) = b],$$
$$\sigma_{ab}^2 \triangleq \mathrm{Var}_{\mathbf{x}_2}[f(a, \mathbf{x}_2) - f(b, \mathbf{x}_2)] = \mathbb{E}[s_{uv}^2|\mathbf{x}_1(u) = a, \mathbf{x}_1(v) = b] - 2\gamma^2.$$
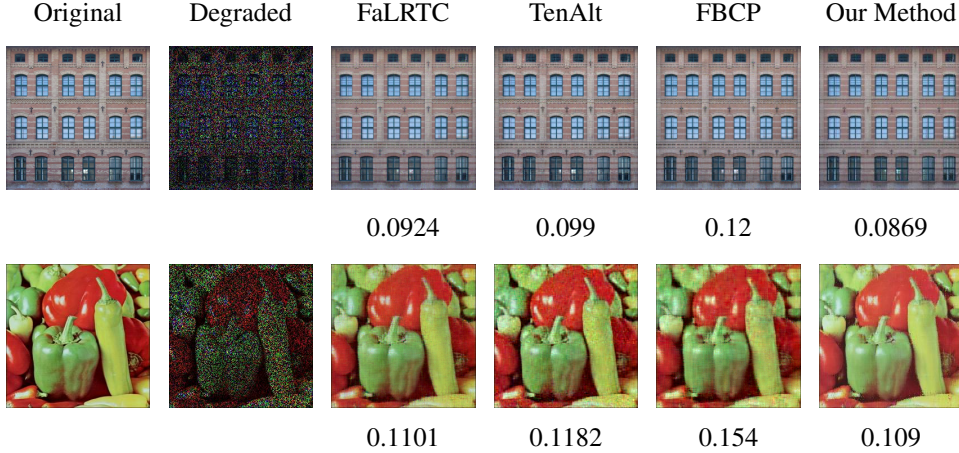
This is also equivalent to the expectation of the empirical means and variances computed by the algorithm when we condition on the latent representations of the users. The computation of $\hat{A}_{ui}$ involves two steps: first the algorithm determines the k rows in $\mathcal{S}_u^{\beta}(i)$ with minimum sample variance, and then it computes the estimate by averaging over the k rows, adjusting each according to the empirical row mean.

The proof involves three key steps, each stated within a lemma. We first prove that every pair of rows has sufficient overlap with high probability. Specifically, Lemma 7.1 proves that for any (u, i), the number of the candidate rows, $|\mathcal{S}_u^{\beta}(i)|$, concentrates around $(m-1)p$. This relies on concentration of Binomial random variables via Chernoff's bound, using the fact that every entry is independently observed uniformly at random.

LEMMA 7.1. *Given $p > 0$, $2 \leq \beta < np^2$, and $\theta \in (0, 1)$, for any $(u, i) \in$*

$[m] \times [n]$,

$$\mathbb{P}\left(|\mathcal{S}_u^\beta(i) - (m-1)p| > \theta(m-1)p\right)$$
$$\leq (m-1)\exp\left(-\frac{(np^2-\beta)^2}{2np^2}\right) + 2\exp\left(-\frac{\theta^2}{3}(m-1)p\right).$$

Next, assuming the overlap between two rows u and v is sufficiently large, Lemmas 7.2 and 7.3 prove that the sample mean and variance of the difference $Z(u,i) - Z(v,i)$, denoted by m_{uv} and s_{uv}^2 and defined in Section 3.1 (main article), concentrate around their expectations $\mu_{x_1(u)x_1(v)}$ and $\sigma_{x_1(u)x_1(v)}^2$ with high probability. Recall that $B_0 = LD_{\mathcal{X}_1} + 2B_e$, uniformly upper bounds $|Z(u,i) - Z(v,i)|$ for any $u,v \in [m]$ and $i \in [n]$, due to the properties that f is Lipschitz, $\mathcal{X}_1$ is compact, and the noise terms are bounded by B_e. We will use the Bernstein and Maurer-Pontil inequalities along with the boundedness of the variables to show the following concentration results for the mean and variance.

LEMMA 7.2.   *Given $u \in [m], i \in [n]$, and $\beta \geq 2$, for any $\nu > 0$,*

$$\mathbb{P}\left(\left|\mu_{x_1(u)x_1(v)} - m_{uv}\right| > \nu \,\Big|\, v \in \mathcal{S}_u^\beta(i)\right) \leq \exp\left(-\frac{3\beta\nu^2}{6B_0^2 + 4B_0\nu}\right).$$

LEMMA 7.3.   *Given $u \in [m], i \in [n]$, and $\beta \geq 2$, for any $\tau > 0$,*

$$\mathbb{P}\left(\left|s_{uv}^2 - \left(\sigma_{x_1(u)x_1(v)}^2 + 2\gamma^2\right)\right| > \tau \,\Big|\, v \in \mathcal{S}_u^\beta(i)\right) \leq 2\exp\left(-\frac{\beta\tau^2}{8B_0^2(2B_0^2+\tau)}\right).$$

Next, Lemma 7.4 proves that since the latent features are sampled iid from a bounded metric space, for any index pair (u,i), there exists k "good" neighboring rows in $\mathcal{S}_u^\beta(i)$, whose $\sigma_{x_1(u)x_1(v)}^2$ is sufficiently small. Let $\left\{\sigma_{x_1(u)x_1(v)}^2\right\}_{v\in\mathcal{S}_u^\beta(i)}^{(k)}$ denote the value of the k-th minimum element in the set $\left\{\sigma_{x_1(u)x_1(v)}^2\right\}_{v\in\mathcal{S}_u^\beta(i)}$.

LEMMA 7.4.   *Given $u \in [m]$, $i \in [n]$, for any $\zeta > 0$ and for any integer $k \in (0, N_0]$,*

$$\mathbb{P}\left(\left\{\sigma_{x_1(u)x_1(v)}^2\right\}_{v\in\mathcal{S}_u^\beta(i)}^{(k)} > \zeta \,\bigg|\, |\mathcal{S}_u^\beta(i)| \in \left[\frac{1}{2}(m-1)p, \frac{3}{2}(m-1)p\right]\right)$$
$$\leq \exp\left(-\frac{(N_0-k)^2}{2N_0}\right),$$

where $N_0 = \frac{1}{2}(m-1)p\phi_1\left(\sqrt{\frac{\zeta}{L^2}}\right)$ and $\phi_1(r) := ess\inf_{x'\in\mathcal{X}_1} P_{\mathcal{X}_1}(d_{\mathcal{X}_1}(\mathbf{x}, x') \leq r)$.

Given that there exist k good neighbors in $\mathcal{S}_u^\beta(i)$ whose variance is small, and conditioned on the event that all the sample variances concentrate, it follows that the true variance between u and its k nearest neighbors are small with high probability. Therefore, we can provide a bound on the tail probability of the estimation error conditioned on these good events by using Chebyshev's inequality.

LEMMA 7.5. *Given $\nu > 0$, $\tau > 0$, for any $\varepsilon > \nu$, $\zeta \geq 0$ and any integer $k \in (0, N_0]$, the estimate produced by the user-user k-nearest neighbor with overlap parameter β satisfies*

$$\mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon \;\middle|\; E\right) \leq \frac{1}{(\varepsilon - \nu)^2}\left(\zeta + 2\tau + \frac{\gamma^2}{k}\right),$$

where the events E, E_1, E_2, E_3 and E_4 are defined as follows

$$E := E_1 \cap E_2 \cap E_3 \cap E_4,$$

$$E_1 := \left\{|\mathcal{S}_u^\beta(i)| \in \left[\frac{1}{2}(m-1)p, \frac{3}{2}(m-1)p\right]\right\},$$

$$E_2 := \left\{\left|\mu_{x_1(u)x_1(v)} - m_{uv}\right| \leq \nu,\; \forall\, v \in \mathcal{S}_u^\beta(i)\right\},$$

$$E_3 := \left\{\left|s_{uv}^2 - (\sigma_{x_1(u)x_1(v)}^2 + 2\gamma^2)\right| \leq \tau,\; \forall\, v \in \mathcal{S}_u^\beta(i)\right\},$$

$$E_4 := \left\{\left\{\sigma_{x_1(u)x_1(v)}^2\right\}_{v \in \mathcal{S}_u^\beta(i)}^{(k)} \leq \zeta\right\}.$$

Finally we combine the lemmas which bound each of the deviating events to obtain a bound on the tail probability of the error.

THEOREM 7.6. *Suppose that*

$$p \geq \max\left\{m^{-1+\delta}, n^{-\frac{1}{2}+\delta}\right\} \textit{ for some } \delta > 0,$$

$$\zeta \textit{ satisfies } \phi_1\left(\sqrt{\frac{\zeta}{L^2}}\right) \geq c_\phi\,(mp)^{-2/3} \textit{ for some } c_\phi \geq 0,$$

$$2 \leq \beta \leq cnp^2 \textit{for some } c \in (0,1),\textit{ and}$$

$$k \leq \frac{c_k}{2}(m-1)p\phi_1\left(\sqrt{\frac{\zeta}{L^2}}\right) \textit{ for some } c_k \in [0,1).$$

For any $\varepsilon > \left(np^2\right)^{-1/3}$, the tail probability of the error of the estimate produced by the user-user k-nearest neighbor algorithm with overlap parameter β is upper

bounded by:

$$\mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon\right)$$

$$\leq \frac{1}{\left(\varepsilon - \beta^{-1/3}\right)^2}\left(\zeta + 2\beta^{-1/3} + \frac{\gamma^2}{k}\right)$$

$$+ 3\exp\left(-c_1\left(mp\right)^{1/3}\right) + \left(m + \frac{9}{2}mp\right)\exp\left(-c_2\beta^{1/3}\right),$$

where $c_1 := \min\left\{\frac{1}{24}, \frac{c_\phi(1-c_k)^2}{8}\right\}$ *and* $c_2 := \min\left\{\frac{c^2}{2}, \frac{3}{6B_0^2+4B_0}, \frac{1}{8B_0^2\left(2B_0^2+1\right)}\right\}$ *are absolute constants, which depend only on the geometry of the latent spaces.*

Note that we can obtain a similar theorem for the item-item k-nearest neighbor variant by exchanging m and n in Theorem 7.6. Given an upper bound on the tail of the error probability, we can prove an upper bound on the MSE by integrating the tail bound. The results for the user-user variant is stated in the final Theorem 4.1 (main article). Corresponding results for the item-item variant is stated in Theorem 4.2 (main article), which follows from taking the transpose of the matrix and simply exchanging m and n.

7.2. *Proof of Key Lemmas.*   In this section we prove the five key lemmas introduced in the proof outline.

Sufficiently many rows with large overlap.   Using the fact that every entry is independently observed uniformly at random, we can prove Lemma 7.1. It states that with high probability, for every entry (u, i) which we might want to estimate, there are sufficiently many other rows v for which entry (v, i) is observed and there is a sufficiently large overlap $|\mathcal{O}^{uv}|$ between rows u and v.

PROOF OF LEMMA 7.1.   The set $\mathcal{S}_u^\beta(i)$ consists of all rows v such that (a) entry (v, i) is observed, and (b) $|\mathcal{O}^{uv}| \geq \beta$. With a fixed u, for each v, define binary random variables Q_v and R_{uv}, such that $Q_v = 1$ if $(v, i) \in \mathcal{D}$, i.e. $Z(v, i)$ is observed, and 0 otherwise; $R_{uv} = 1$ if $|\mathcal{O}^{uv}| \geq \beta$ and 0 otherwise. Then we can equivalently express the cardinality of the set as a sum over $m - 1$ binary random variables

$$|\mathcal{S}_u^\beta(i)| = \sum_{v\neq u} Q_v R_{uv}.$$

If the events $\sum_{v\neq u} Q_v \in [a, b]$ and $\sum_{v\neq u} R_{uv} = m - 1$ are satisfied, it implies $|\mathcal{S}_u^\beta(i)| = \sum_{v\neq u} Q_v R_{uv} \in [a, b]$. It follows from the contrapositive of this state-

ment that

$$\left\{ |\mathcal{S}_u^\beta(i)| \notin [a,b] \right\} \Rightarrow \left\{ \sum_{v \neq u} Q_v \notin [a,b] \right\} \vee \left\{ \sum_{v \neq u} R_{uv} < m - 1 \right\}.$$

Therefore, by applying the union bound, we obtain that for any $0 \leq a < b \leq m-1$,

$$(7.1) \quad \mathbb{P}\left(|\mathcal{S}_u^\beta(i)| \notin [a,b] \right) \leq \mathbb{P}\left(\sum_{v \neq u} Q_v \notin [a,b] \right) + \mathbb{P}\left(\sum_{v \neq u} R_{uv} < m - 1 \right).$$

Given that the entries within each row v are sampled independently with probability p, it follows that $\sum_{v \neq u} Q_v$ is Binomial with parameters $(m-1)$ and p. We will choose the interval specified by endpoints $a = (1 - \theta)(m-1)p$ and $b = (1 + \theta)(m-1)p$. Directly applying Chernoff's bound (see Theorem A.1 in Appendix A) implies that for any $\theta \in (0,1)$,

$$(7.2) \quad \mathbb{P}\left(\left| \sum_{v \neq u} Q_v - (m-1)p \right| > \theta(m-1)p \right) \leq 2 \exp\left(-\frac{\theta^2(m-1)p}{3} \right).$$

Next recall that the variable $R_{uv} \triangleq \mathbb{I}(|\mathcal{O}^{uv}| \geq \beta)$. Again due to the assumption that each entry is observed independently with probability p, it follows that $|\mathcal{O}^{uv}|$ is Binomial with parameters n and p^2. Therefore, for any $\beta \in [2, np^2)$, by an application of Chernoff's bound for lower tails, it follows that for each $v \neq u$,

$$(7.3) \qquad\qquad \mathbb{P}\left(R_{uv} = 0 \right) \leq \exp\left(-\frac{\left(np^2 - \beta \right)^2}{2np^2} \right).$$

By the union bound, it follows that

$$(7.4)$$
$$\mathbb{P}\left(\sum_{v \neq u} R_{uv} < m - 1 \right) \leq \sum_{v \neq u} \mathbb{P}\left(R_{uv} = 0 \right) \leq (m-1) \exp\left(-\frac{\left(np^2 - \beta \right)^2}{2np^2} \right).$$

By combining (7.1)-(7.4), we obtain the desired result. $\qquad\square$

Concentration of sample mean and sample variance. Assuming that the overlap between two rows u and v is sufficiently large, the sample mean and variance of the difference between the two rows will concentrates around their expected values.

PROOF OF LEMMA 7.2. Conditioned on a particular realization of the latent features associated to rows u and v, i.e. $\mathbf{x}_1(u) = x_1(u), \mathbf{x}_1(v) = x_1(v)$, the expected mean between the difference in their values $\mu_{x_1(u)x_1(v)}$ is a constant. Recall that the empirical mean m_{uv} is defined as

$$(7.5) \qquad m_{uv} = \frac{1}{|\mathcal{O}^{uv}|} \left(\sum_{j \in \mathcal{O}^{uv}} Z(u,j) - Z(v,j) \right).$$

We would like to show that m_{uv} concentrates to $\mu_{x_1(u)x_1(v)}$. For any column j, recall that the column latent feature variable $\mathbf{x}_2(j)$ is sampled according to $P_{\mathcal{X}_2}$, independently from the row features $\mathbf{x}_1(u)$ and $\mathbf{x}_1(v)$. The additive noise terms associated to each observation is an independent and zero-mean random variable. For each pair (u,v), let us define a new independent random variable,

$$W_{uv}(j) = Z(u,j) - Z(v,j) \text{ for } j \in \mathcal{O}^{uv},$$

which has mean $\mu_{x_1(u)x_1(v)}$ when conditioned on $\mathbf{x}_1(u) = x_1(u)$ and $\mathbf{x}_1(v) = x_1(v)$. It follows then that $\tilde{W}_{uv}(j) = W_{uv}(j) - \mu_{x_1(u)x_1(v)}$ for $j \in \mathcal{O}^{uv}$ are zero-mean independent random variables conditioned on the row latent features.

Therefore, conditioned on $\mathbf{x}_1(u) = x_1(u), \mathbf{x}_1(v) = x_1(v)$ and the cardinality of overlap $|\mathcal{O}^{uv}|$, we can rewrite the difference $\mu_{x_1(u)x_1(v)} - m_{uv}$ to be the average of $|\mathcal{O}^{uv}|$ independent, zero mean random variables $\tilde{W}_{uv}(j)$. By the definition of B_0 in (2.7), $|W_{uv}(j)|$ is upper bounded by B_0 for every pair of rows (u,v) and column j. It follows that $\left| \mu_{x_1(u)x_1(v)} \right| \leq B_0$, such that

$$\left| \tilde{W}_{uv}(j) \right| = \left| W_{uv}(j) - \mu_{x_1(u)x_1(v)} \right| \leq |W_{uv}(j)| + \left| \mu_{x_1(u)x_1(v)} \right| \leq 2B_0.$$

In addition, conditioned on $\mathbf{x}_1(u) = x_1(u), \mathbf{x}_1(v) = x_1(v)$, $\mu_{x_1(u)x_1(v)}$ is a constant so that

$$Var\left(\tilde{W}_{uv}(j) \right) = Var\left(W_{uv}(j) \right) = \mathbb{E}\left[W_{uv}(j)^2 \right] - \mathbb{E}\left[W_{uv}(j) \right]^2$$
$$\leq \mathbb{E}\left[W_{uv}(j)^2 \right] \leq B_0^2.$$

Therefore, an application of Bernstein's inequality (see Lemma A.2) for a sum of bounded random variables implies that

$$(7.6) \qquad \mathbb{P}\left(\left| \mu_{\mathbf{x}_1(u)\mathbf{x}_1(v)} - m_{uv} \right| > \nu \, \middle| \, \mathbf{x}_1(u) = x_1(u), \mathbf{x}_1(v) = x_1(v), |\mathcal{O}^{uv}| \right)$$
$$\leq \exp\left(-\frac{3|\mathcal{O}^{uv}|\nu^2}{6B_0^2 + 4B_0\nu} \right).$$

When $v \in \mathcal{S}_u^\beta(i)$, $|\mathcal{O}^{uv}| \geq \beta$. Further, since the above holds inequalities for all possibilities of $x_1(u), x_1(v)$, we conclude that

$$\mathbb{P}\left(\left|\mu_{x_1(u)x_1(v)} - m_{uv}\right| > \nu \,\Big|\, v \in \mathcal{S}_u^\beta(i)\right) \leq \exp\left(-\frac{3\beta\nu^2}{6B_0^2 + 4B_0\nu}\right).$$

$\square$

Next we prove that the sample variance s_{uv}^2 converges to its expected value $\sigma_{x_1(u),x_1(v)}^2$ conditioned on the row latent features and a large enough overlap $\mathcal{O}^{uv}$.

PROOF OF LEMMA 7.3. Recall $\sigma_{ab}^2 \triangleq \mathrm{Var}[f(a, \mathbf{x}_2) - f(b, \mathbf{x}_2)]$ for $a, b \in \mathcal{X}_1$, $\mathbf{x}_2 \sim P_{\mathcal{X}_2}$, and the sample variance between rows u, v is defined as

$$s_{uv}^2 = \frac{1}{2|\mathcal{O}^{uv}|(|\mathcal{O}^{uv}| - 1)} \sum_{j_1, j_2 \in \mathcal{O}^{uv}} ((Z(u, j_1) - Z(v, j_1)) - (Z(u, j_2) - Z(v, j_2)))^2$$

$$= \frac{1}{|\mathcal{O}^{uv}| - 1} \sum_{j \in \mathcal{O}^{uv}} (Z(u, j) - Z(v, j) - m_{uv})^2.$$

Conditioned on $\mathbf{x}_1(u) = x_1(u), \mathbf{x}_1(v) = x_1(v)$, it follows from our model of independent additive noise that

$$\mathbb{E}\left[s_{uv}^2 \mid x_1(u), x_1(v)\right] = \sigma_{x_1(u)x_1(v)}^2 + 2\gamma^2,$$

with respect to the randomness of the sampled latent column features induced by $P_{\mathcal{X}_2}$ and the independent noise terms. Recall as well that, $W_{uv}(j) = Z(u, j) - Z(v, j)$ are independent random variables conditioned on $\mathbf{x}_1(u) = x_1(u), \mathbf{x}_1(v) = x_1(v)$, where $|W_{uv}(j)| \leq B_0$, by the model assumptions and the definition of B_0 in (2.7).

Therefore, by an application of Maurer-Pontil inequality for the sample variance of a set of bounded random variables (see Lemma A.3 in Appendix A), it follows that

$$\mathbb{P}\left(|s_{uv}^2 - (\sigma_{\mathbf{x}_1(u)\mathbf{x}_1(v)}^2 + 2\gamma^2)| > \tau \,\big|\, v \in \mathcal{S}_u^\beta(i), \mathbf{x}_1(u) = x_1(u), \mathbf{x}_1(v) = x_1(v)\right)$$

$$(7.7) \qquad \leq 2\exp\left(-\frac{(\beta - 1)\tau^2}{4B_0^2(2(\sigma_{x_1(u)x_1(v)}^2 + 2\gamma^2) + \tau)}\right),$$

where we used the property that $v \in \mathcal{S}_u^\beta(i)$ implies $|\mathcal{O}^{uv}| \geq \beta$. Since we assumed $\beta \geq 2$, $\beta - 1 \geq \beta/2$. By the same argument used in the proof of Lemma 7.2, we

can bound $\sigma^2_{x_1(u)x_1(v)} + 2\gamma^2 = Var\left(W_{uv}(j)\right) \leq \mathbb{E}\left[W_{uv}(j)^2\right] \leq B_0^2$. Therefore, the right hand side of (7.7) can be bounded by

$$(7.8) \qquad\qquad \leq 2\exp\left(-\frac{\beta\tau^2}{8B_0^2(2B_0^2 + \tau)}\right).$$

Given that this bound is indepedent of $\mathbf{x}(u), \mathbf{x}(v)$, we can conclude the desired result. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Sufficiently good k-nearest neighbors. We will next prove that with high probability, the k rows in set $\mathcal{S}_u^\beta(i)$ with the smallest values of $\sigma^2_{x_1(u)x_1(v)}$ are indeed "close", i.e. the variance is small. This result depends on the local geometry of the probability measure on the latent space, according to the function $\phi\left(\cdot\right)$.

DEFINITION 7.1. *Given a compact metric space $(\mathcal{X}, d)$ equipped with a Borel probability measure $P_\mathcal{X}$, we define the measure underestimator function $\phi\left(\cdot\right)$ as the essential infimum of the measure of an r-ball. For $r > 0$,*

$$\phi\left(r\right) := ess \inf_{x \in \mathcal{X}} P_\mathcal{X}\left(B(x, r)\right),$$

where $B(x, r) := \{x' \in \mathcal{X} : d(x, x') \leq r\}$.

We recall the notion of essential infimum. For a function f measurable on $(\mathcal{X}, P_\mathcal{X})$, its essential infimum on $\mathcal{X}$ with respect to measure $P_\mathcal{X}$ is defined as

$$ess \inf_{x \in \mathcal{X}} f(x) = \sup\left\{a : P_\mathcal{X}\left(\{x : f(x) < a\}\right) = 0\right\}.$$

For each $r > 0$, we can define $f_r(x) := P_\mathcal{X}\left(B(x, r)\right)$ so that f_r is measurable. Because $P_\mathcal{X}$ is a probability measure, $0 \leq f_r(x) \leq 1$ for all $x \in \mathcal{X}$ and hence $\phi(r) := ess \inf_{x \in \mathcal{X}} f_r(x)$ is well-defined and takes value in $[0, 1]$.

The function $\phi\left(\cdot\right)$ behaves as an underestimator of the cumulative distribution function of $P_\mathcal{X}$, and it always exists under our assumptions that $\mathcal{X}$ is compact (see Appendix B for a proof of its existence). It is used to ensure that for any $u \in [m]$, with high probability, there exists another row $v \in \mathcal{S}_u^\beta(i)$ such that $d_\mathcal{X}(x_1(v), x_1(v))$ is small, implied by the Lipschitz condition that we can use the values of row v to approximate the values of row u well.

PROOF OF LEMMA 7.4. Define the random variable $Y_{uv} = \mathbb{I}(\sigma^2_{x_1(u)x_1(v)} \leq \zeta)$. Then the following two events are equivalent.

$$\left\{\sum_{v \in \mathcal{S}_u^\beta(i)} Y_{uv} < k\right\} \equiv \left\{\left\{\sigma^2_{x_1(u)x_1(v)}\right\}^{(k)}_{v \in \mathcal{S}_u^\beta(i)} > \zeta\right\}.$$

Observe that $\sum_{v \in \mathcal{S}_u^\beta(i)} Y_{uv}$ is simply a Binomial random variable with parameters $\left|\mathcal{S}_u^\beta(i)\right|$ and $\mathbb{P}\left(\sigma_{x_1(u)x_1(v)}^2 \leq \zeta\right)$. By the Lipschitz property of f, for any $a, b \in \mathcal{X}_1$, and $y \in \mathcal{Y}$,

$$(7.9) \qquad |f(a, y) - f(b, y)| \leq L d_{\mathcal{X}_1}(a, b).$$

Therefore, it follows that

$$\sigma_{ab}^2 = \text{Var}[f(a, \mathbf{y}) - f(b, \mathbf{y})]$$
$$\leq \mathbb{E}[(f(a, \mathbf{y}) - f(b, \mathbf{y}))^2]$$
$$(7.10) \qquad \leq L^2 d_{\mathcal{X}_1}(a, b)^2.$$

From (7.10), it follows that $d_{\mathcal{X}_1}(a, b) \leq \sqrt{\zeta/L^2}$ is a sufficient condition to ensure $\sigma_{ab}^2 < \zeta$. Therefore, using the definition of $\phi_1(\cdot)$, it follows that

$$\mathbb{P}\left(\sigma_{x_1(u)x_1(v)}^2 \leq \zeta\right) \geq \mathbb{P}\left(d_{\mathcal{X}_1}(x_1(u), x_1(v)) \leq \sqrt{\frac{\zeta}{L^2}}\right)$$
$$\geq \phi_1\left(\sqrt{\frac{\zeta}{L^2}}\right).$$

Since we have also conditioned on the fact that $|\mathcal{S}_u^\beta(i)| \geq \frac{1}{2}(m-1)p$, if U is a Binomial random variable with parameters $\frac{1}{2}(m-1)p$ and $\phi_1\left(\sqrt{\frac{\zeta}{L^2}}\right)$, then $\sum_{v \in \mathcal{S}_u^\beta(i)} Y_{uv}$ dominates U, in the sense that

$$\mathbb{P}\left(\sum_{v \in \mathcal{S}_u^\beta(i)} Y_{uv} < k\right) \leq \mathbb{P}(U < k).$$

We let N_0 denote the expectation of U, i.e., $N_0 := \mathbb{E}[U] = \frac{1}{2}(m-1)p\phi_1\left(\sqrt{\frac{\zeta}{L^2}}\right)$. Then it follows from Chernoff's bound for the lower tail that

$$\mathbb{P}(U < k) \leq \exp\left(-\frac{(N_0 - k)^2}{2N_0}\right).$$

Therefore,

$$\mathbb{P}\left(\left\{\sigma_{x_1(u)x_1(v)}^2\right\}_{v \in \mathcal{S}_u^\beta(i)}^{(k)} > \zeta \,\middle|\, |\mathcal{S}_u^\beta(i)| \in \left[\frac{1}{2}(m-1)p, \frac{3}{2}(m-1)p\right]\right)$$
$$\leq \exp\left(-\frac{(N_0 - k)^2}{2N_0}\right).$$

$\square$

Bound on tail probability of error conditioned on good events. Conditioned on the above events that the overlaps are sufficiently large, mean and variance concentrate, and there are sufficiently many good nearest neighbors, we can upper bound the estimation error using Chebyshev's inequality.

PROOF OF LEMMA 7.5. Recall that $\mathcal{S}_u^{\beta,k}(i)$ denotes the set of the k best row indices v in $\mathcal{S}_u^{\beta}(i)$ which have minimum sample variance s_{uv}^2. We are interested in the probabilistic tail bound of the error $Err^k(u,i)$, which can be expressed as

$$
\begin{aligned}
Err^k(u,i) &\triangleq A(u,i) - \hat{A}^k(u,i) \\
&= \frac{1}{|\mathcal{S}_u^{\beta,k}(i)|} \sum_{v \in \mathcal{S}_u^{\beta,k}(i)} (A(u,i) - \hat{A}_v(u,i)) \\
&= \frac{1}{|\mathcal{S}_u^{\beta,k}(i)|} \sum_{v \in \mathcal{S}_u^{\beta,k}(i)} (A(u,i) - Z(v,i) - m_{uv}) \\
&= \frac{1}{|\mathcal{S}_u^{\beta,k}(i)|} \sum_{v \in \mathcal{S}_u^{\beta,k}(i)} \left(\left(A(u,i) - A(v,i) - \eta(v,i) - \mu_{x_1(u)x_1(v)} \right) \right) \\
&\quad - \frac{1}{|\mathcal{S}_u^{\beta,k}(i)|} \sum_{v \in \mathcal{S}_u^{\beta,k}(i)} \left(\left(m_{uv} - \mu_{x_1(u)x_1(v)} \right) \right)
\end{aligned}
$$

Its absolute value can be bounded by

$$
\begin{aligned}
&\left| A(u,i) - \hat{A}^k(u,i) \right| \\
&\leq \left| \frac{1}{|\mathcal{S}_u^{\beta,k}(i)|} \sum_{v \in \mathcal{S}_u^{\beta,k}(i)} \left(A(u,i) - A(v,i) - \eta(v,i) - \mu_{x_1(u)x_1(v)} \right) \right| \\
&\quad + \left| \frac{1}{|\mathcal{S}_u^{\beta,k}(i)|} \sum_{v \in \mathcal{S}_u^{\beta,k}(i)} \left(m_{uv} - \mu_{x_1(u)x_1(v)} \right) \right|
\end{aligned}
$$

By the same argument as in the proof of Lemma 7.1, $\left| A(u,i) - \hat{A}^k(u,i) \right| > \varepsilon$ implies either

$$
(7.11) \quad \left| \frac{1}{|\mathcal{S}_u^{\beta,k}(i)|} \sum_{v \in \mathcal{S}_u^{\beta,k}(i)} \left(A(u,i) - A(v,i) - \eta(v,i) - \mu_{x_1(u)x_1(v)} \right) \right| > \varepsilon - \varepsilon_1
$$

or

$$(7.12) \qquad \left| \frac{1}{|\mathcal{S}_u^{\beta,k}(i)|} \sum_{v \in \mathcal{S}_u^{\beta,k}(i)} m_{uv} - \mu_{x_1(u)x_1(v)} \right| > \varepsilon_1$$

for any choice of ε_1. Recall the definitions of events E_1, E_2, E_3, E_4 from the statement of Lemma 7.5. Conditioned on E_2, $\left| \mu_{x_1(u)x_1(v)} - m_{uv} \right| \leq \nu$, $\forall\ v \in \mathcal{S}_u^{\beta}(i)$, the second event (7.12) never happens for $\varepsilon_1 \geq \nu$. By choosing $\varepsilon_1 = \nu$, it follows that the error probability conditioned on E_1, E_2, E_3, and E_4 can be bounded by the probability of the first event (7.11). We denote $E := E_1 \cap E_2 \cap E_3 \cap E_4$, and we use "expr" to denote the left hand expression in (7.11).

(7.13)

$$\mathbb{P}\left(\left| A(u,i) - \hat{A}^k(u,i) \right| > \varepsilon \mid E \right)$$

$$\leq \mathbb{P}\left(\left| \frac{1}{|\mathcal{S}_u^{\beta,k}(i)|} \sum_{v \in \mathcal{S}_u^{\beta,k}(i)} \left(A(u,i) - A(v,i) - \eta(v,i) - \mu_{x_1(u)x_1(v)} \right) \right| > \varepsilon - \nu \mid E \right)$$

$$= \int_{\vec{y} \in \mathcal{X}_1^m} \sum_{\mathcal{S}_0 \subset [n]:|\mathcal{S}_0|=k} \mathbb{P}\left(|\text{"expr"}| > \varepsilon - \nu \mid (x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0, E \right)$$

$$\mathbb{P}\left((x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0 \mid E \right) d\vec{y}.$$

We can verify that the event E and the set $\mathcal{S}_u^{\beta,k}(i)$ do not depend upon any information from column i, therefore $\eta(v,i)$ is independent from E for all v. For the same reason, $x_2(i)$ is independent from E and $\mathcal{S}_u^{\beta,k}(i)$. Therefore, we can show that each term in the summation $\left(A(u,i) - A(v,i) - \eta(v,i) - \mu_{x_1(u)x_1(v)} \right)$ has zero mean by definition of $\mu_{x_1(u)x_1(v)}$. We use $\mathbf{x}_2$ to denote an independent random variable sampled as per $P_{\mathcal{X}_2}$.

$$\mathbb{E}\left[A(u,i) - A(v,i) - \eta(v,i) - \mu_{x_1(u)x_1(v)} \mid (x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0, E \right]$$

$$= \mathbb{E}\left[f(x_1(u), x_2(i)) - f(x_1(v), x_2(i)) \mid (x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0, E \right] - \mu_{y_u y_v}$$

$$= \mathbb{E}_{\mathbf{x}_2}\left[f(y_u, \mathbf{x}_2) - f(y_v, \mathbf{x}_2) \mid (x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0, E \right] - \mu_{y_u y_v}$$

$$= 0.$$

Therefore an application of Chebyshev inequality yields,

(7.14)

$$
\mathbb{P}\left(\left|\text{``expr''}\right| > \varepsilon - \nu \mid (x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0, E\right)
$$

$$
\leq \frac{\mathrm{Var}\left[\frac{1}{|\mathcal{S}_u^{\beta,k}(i)|} \sum_{v \in \mathcal{S}_u^{\beta,k}(i)} (A(u,i) - A(v,i) - \eta(v,i)) \mid (x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0, E\right]}{(\varepsilon - \nu)^2}.
$$

Recall that $\eta(v,i)$ is independent from E and $\mathcal{S}_u^{\beta,k}(i)$ and has variance γ^2. In addition, $|\mathcal{S}_u^{\beta,k}(i)| = |\mathcal{S}_0| = k$. Therefore, by independence of the noise term and using Cauchy-Schwarz inequality, it follows that

$$
\mathrm{Var}\left[\frac{1}{|\mathcal{S}_u^{\beta,k}(i)|} \sum_{v \in \mathcal{S}_u^{\beta,k}(i)} (A(u,i) - A(v,i) - \eta(v,i)) \,\Big|\, (x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0, E\right]
$$

$$
= \mathrm{Var}\left[\frac{1}{k} \sum_{v \in \mathcal{S}_u^{\beta,k}(i)} (A(u,i) - A(v,i)) \,\Big|\, (x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0, E\right]
$$

$$
+ \mathrm{Var}\left[\frac{1}{k} \sum_{v \in \mathcal{S}_u^{\beta,k}(i)} \eta(v,i) \,\Big|\, (x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0, E\right]
$$

(7.15)

$$
\leq \left(\frac{1}{k} \sum_{v \in \mathcal{S}_0} \sqrt{\mathrm{Var}\left[A(u,i) - A(v,i) \mid (x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0, E\right]}\right)^2 + \frac{\gamma^2}{k}.
$$

Next we will bound $\mathrm{Var}\left[A(u,i) - A(v,i) \mid (x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0, E\right]$ for any $v \in \mathcal{S}_0$. Recall that the event E and set $\mathcal{S}_u^{\beta,k}(i)$ is completely independent from $x_2(i)$ because the event does not depend on any data from column i. Therefore using the independence of $x_2(i)$, if we let $\mathbf{x}_2$ denote a random variable which is sampled independently from $P_{\mathcal{X}_2}$, then

$$
\mathrm{Var}\left[A(u,i) - A(v,i) \mid (x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0, E\right]
$$

$$
= \mathrm{Var}\left[f(x_1(u), x_2(i)) - f(x_1(v), x_2(i)) \mid (x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0, E\right]
$$

$$
= \mathrm{Var}\left[f(y_u, \mathbf{x}_2) - f(y_v, \mathbf{x}_2) \mid (x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0, E\right]
$$

$$
= \sigma^2_{y_u y_v}.
$$

Let $\tilde{\mathcal{V}}$ denote the subset of rows $v \in \mathcal{S}_u^\beta(i)$ such that $\sigma^2_{x_1(u)x_1(v)} \leq \zeta$. Conditioned on E_4, the size of set $\tilde{\mathcal{V}}$ must be at least k. Conditioned on E_3, for every $v \in \tilde{\mathcal{V}} \subset \mathcal{S}_u^\beta(i)$,

$$s_{uv}^2 \leq \left(\sigma^2_{x_1(u)x_1(v)} + 2\gamma^2\right) + \tau \leq \zeta + 2\gamma^2 + \tau.$$

Therefore, it follows by the definition of $\mathcal{S}_u^{\beta,k}(i)$ as the set of k rows with minimum sample variance, that for all $v \in \mathcal{S}_u^{\beta,k}(i)$, $s_{uv}^2 \leq \zeta + \tau + 2\gamma^2$. Due to event E_3 and $\mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0$, this implies that for all $v \in \mathcal{S}_0$, $\sigma^2_{x_1(u)x_1(v)} \leq \left(s_{uv}^2 + \tau\right) - 2\gamma^2 \leq \zeta + 2\tau$. Therefore for all $v \in \mathcal{S}_0$,

$$\mathrm{Var}\left[A(u,i) - A(v,i) \mid (x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0, E\right] \leq \zeta + 2\tau.$$

Finally we can plug the variance bound into (7.15) and (7.14) to show that

$$\mathbb{P}\left(|\text{``expr''}| > \varepsilon - \nu \mid (x_1(v))_{v \in [m]} = \vec{y}, \mathcal{S}_u^{\beta,k}(i) = \mathcal{S}_0, E\right)$$

$$\leq \frac{1}{(\varepsilon - \nu)^2}\left(\left(\frac{1}{k}\sum_{v \in \mathcal{S}_0}\sqrt{\zeta + 2\tau}\right)^2 + \frac{\gamma^2}{k}\right)$$

$$\leq \frac{1}{(\varepsilon - \nu)^2}\left(\zeta + 2\tau + \frac{\gamma^2}{k}\right).$$

Since this above bound does not depend on the particular choice of $\vec{y}$ and $\mathcal{S}_0$, by plugging into (7.13), it follows that

$$(7.16) \qquad \mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon \mid E\right)$$

$$(7.17) \qquad \leq \frac{1}{(\varepsilon - \nu)^2}\left(\zeta + 2\tau + \frac{\gamma^2}{k}\right).$$

$\square$

7.3. *Proof of Theorem 7.6.* Lemmas 7.1, 7.2, 7.3, and 7.4 provide probability bounds on the following events defined in Lemma 7.5:

$$E_1 := \left\{|\mathcal{S}_u^\beta(i)| \in \left[\frac{1}{2}(m-1)p, \frac{3}{2}(m-1)p\right]\right\},$$

$$E_2 := \left\{\left|\mu_{x_1(u)x_1(v)} - m_{uv}\right| \leq \nu, \ \forall \ v \in \mathcal{S}_u^\beta(i)\right\},$$

$$E_3 := \left\{\left|s_{uv}^2 - (\sigma^2_{x_1(u)x_1(v)} + 2\gamma^2)\right| \leq \tau, \ \forall \ v \in \mathcal{S}_u^\beta(i)\right\},$$

$$E_4 := \left\{\left\{\sigma^2_{x_1(u)x_1(v)}\right\}_{v \in \mathcal{S}_u^\beta(i)}^{(k)} \leq \zeta\right\}.$$

Lemma 7.5 provides a bound on the tail of the probability of error conditioned on the events E_1, E_2, E_3, E_4. Theorem 7.6 simply combines the lemmas with the proper conditioning to prove an upper bound on the tail of the error probability.

PROOF. First of all, note that the error probability can be decomposed into bounding the probability of violating each event, according to

$$
\begin{aligned}
&\mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon\right) \\
&= \mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon \,\Big|\, E_1 \cap E_2 \cap E_3 \cap E_4\right) \mathbb{P}\left(E_1 \cap E_2 \cap E_3 \cap E_4\right) \\
&\quad + \mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon \,\Big|\, E_1^c \cup E_2^c \cup E_3^c \cup E_4^c\right) \mathbb{P}\left(E_1^c \cup E_2^c \cup E_3^c \cup E_4^c\right) \\
&\leq \mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon \,\Big|\, E_1 \cap E_2 \cap E_3 \cap E_4\right) \\
&\quad + \mathbb{P}\left(E_1^c \cup E_2^c \cup E_3^c \cup E_4^c\right) \\
&\leq \mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon \,\Big|\, E_1 \cap E_2 \cap E_3 \cap E_4\right) \\
&\quad + \mathbb{P}\left(E_1^c\right) + \mathbb{P}\left(E_2^c | E_1\right) + \mathbb{P}\left(E_3^c | E_1\right) + \mathbb{P}\left(E_4^c | E_1\right).
\end{aligned}
$$

The first inequality uses the fact that probabilities cannot exceed 1. The last inequality is derived from the additivity of the measure $\mathbb{P}$ and the union bound. Because $A \cup B = A \cup (B \cap A^c)$,

$$
\begin{aligned}
\mathbb{P}\left(A \cup B\right) &= \mathbb{P}\left(A\right) + \mathbb{P}\left(B \cap A^c\right) \\
&= \mathbb{P}\left(A\right) + \mathbb{P}\left(B | A^c\right) \mathbb{P}\left(A^c\right) \\
&\leq \mathbb{P}\left(A\right) + \mathbb{P}\left(B | A^c\right) \qquad \because \mathbb{P}\left(A^c\right) \leq 1.
\end{aligned}
$$

Then we can obtain the last inequality by applying the union bound after taking $A = E_1^c$ and $B = E_2^c \cup E_3^c \cup E_4^c$. By Lemma 7.5,

$$
\mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon \,\Big|\, E_1, E_2, E_3, E_4\right) \leq \frac{1}{(\varepsilon - \nu)^2}\left(\zeta + 2\tau + \frac{\gamma^2}{k}\right).
$$

Using Lemma 7.1, we have

$$
\begin{aligned}
\mathbb{P}\left(E_1^c\right) &= \mathbb{P}\left(|\mathcal{S}_u^\beta(i)| \notin \left[\frac{1}{2}(m-1)p, \frac{3}{2}(m-1)p\right]\right) \\
&\leq (m-1)\exp\left(-\frac{(np^2 - \beta)^2}{2np^2}\right) + 2\exp\left(-\frac{(m-1)p}{12}\right).
\end{aligned}
$$

Similarly, by using Lemma 7.2 with union bound,

$$\mathbb{P}\left(E_2^c|E_1\right) = \mathbb{P}\left(\bigcup_{v\in\mathcal{S}_u^\beta(i)}\left\{|\mu_{x_1(u)x_1(v)} - m_{uv}| > \nu\right\}\right)$$

$$\leq \frac{3}{2}(m-1)p\exp\left(-\frac{3\beta\nu^2}{6B_0^2 + 4B_0\nu}\right),$$

where we recall that $B_0 = LD_{\mathcal{X}_1} + 2B_e$. By using Lemma 7.3 with union bound again,

$$\mathbb{P}\left(E_3^c|E_1\right) = \mathbb{P}\left(\bigcup_{v\in\mathcal{S}_u^\beta(i)}\left\{\left|s_{uv}^2 - \left(\sigma_{x_1(u)x_1(v)}^2 + 2\gamma_{ui}^2\right)\right| > \tau\right\}\right)$$

$$\leq 3(m-1)p\exp\left(-\frac{\beta\tau^2}{8B_0^2\left(2B_0^2 + \tau\right)}\right).$$

By Lemma 7.4 with $N_0 := \frac{1}{2}(m-1)p\phi_1\left(\sqrt{\frac{\zeta}{L^2}}\right)$ as previously defined,

$$\mathbb{P}\left(E_4^c|E_1\right) \leq \exp\left(-\frac{(N_0 - k)^2}{2N_0}\right).$$

Putting everything together, we obtain the following inequality,

$$\mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon\right)$$

$$\leq \frac{1}{(\varepsilon - \nu)^2}\left(\zeta + 2\tau + \frac{\gamma^2}{k}\right)$$

$$(7.18) \qquad + (m-1)\exp\left(-\frac{\left(np^2 - \beta\right)^2}{2np^2}\right) + 2\exp\left(-\frac{(m-1)p}{12}\right)$$

$$(7.19) \qquad + \frac{3}{2}(m-1)p\exp\left(-\frac{3\beta\nu^2}{6B_0^2 + 4B_0\nu}\right)$$

$$(7.20) \qquad + 3(m-1)p\exp\left(-\frac{\beta\tau^2}{8B_0^2\left(2B_0^2 + \tau\right)}\right)$$

$$(7.21) \qquad + \exp\left(-\frac{(N_0 - k)^2}{2N_0}\right).$$

Note that ζ, ν, τ are parameters which are introduced purely for the purpose of analysis. Requiring all exponential terms decay to 0 as $m, n \to \infty$ restricts the

range of values ζ, ν, τ can take. We will let $\nu = \tau = \beta^{-1/3}$. Also, we enforce ζ satisfies $\phi_1\left(\sqrt{\frac{\zeta}{L^2}}\right) \geq c_\phi (mp)^{-2/3}$ so that $N_0 \geq \frac{c_\phi}{4}(mp)^{1/3}$ as described in the theorem statement. Additionally, we require $\beta \to \infty$ as $m, n \to \infty$ and $np^2 - \beta \geq cnp^2$ for some $c > 0$. We will show these are sufficient conditions for the convergence of exponential error terms.

Given a sequence of problems of size (m, n), suppose that $p = \omega(m^{-1})$ and $p = \omega(n^{-1/2})$. Then $mp, np^2 \to \infty$ as $m, n \to \infty$. We may assume without loss of generality that m and n are large enough such that $mp, np^2, \beta \geq 2$ for simplicity.

Assuming $\beta \to \infty$ as $m, n \to \infty$, we can observe that $\nu, \tau \to 0$ as $m, n \to \infty$. In addition, we have $\nu \leq 1$, $\tau \leq 1$, since we assumed $\beta \geq 2$ (hence, ≥ 1). Therefore, the exponential terms Eq. (7.19) and Eq. (7.20) reduce as follows:

$$\frac{3}{2}(m-1)p\exp\left(-\frac{3\beta\nu^2}{6B_0^2 + 4B_0\nu}\right) \leq \frac{3}{2}(m-1)p\exp\left(-\frac{3}{6B_0^2 + 4B_0}\beta^{1/3}\right),$$

$$3(m-1)p\exp\left(-\frac{\beta\tau^2}{8B_0^2\left(2B_0^2 + \tau\right)}\right) \leq 3(m-1)p\exp\left(-\frac{1}{8B_0^2\left(2B_0^2 + 1\right)}\beta^{1/3}\right).$$

In the theorem statement, we assumed that $k \leq c_k N_0$ for some $c_k \in (0, 1)$, and the constraints on the parameter ζ, which depend on the latent geometry via ϕ_1, allow us to reduce the last exponential error term Eq. (7.21) to the following bound:

$$\exp\left(-\frac{(N_0 - k)^2}{2N_0}\right) \leq \exp\left(-\frac{c_\phi(1 - c_k)^2}{8}(mp)^{1/3}\right).$$

From the assumption $np^2 - \beta \geq cnp^2$ for some $c \in (0, 1)$, it follows that $-\frac{(np^2 - \beta)^2}{2np^2} \leq -\frac{c^2}{2}np^2$. Since we assumed $\beta < np^2$,

$$(m-1)\exp\left(-\frac{(np^2 - \beta)^2}{2np^2}\right) \leq (m-1)\exp\left(-\frac{c^2}{2}np^2\right) \leq (m-1)\exp\left(-\frac{c^2}{2}\beta\right).$$

Finally, we can obtain the following upper bound on Eq. (7.18) using our assumption that $\beta > 1$ and $mp \geq 2$:

$$(m-1)\exp\left(-\frac{(np^2 - \beta)^2}{2np^2}\right) + 2\exp\left(-\frac{(m-1)p}{12}\right)$$
$$\leq (m-1)\exp\left(-\frac{c^2}{2}\beta^{1/3}\right) + 2\exp\left(-\frac{1}{24}(mp)^{1/3}\right).$$

By substituting in the above inequalities, the error probability bound reduces to

$$\mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon\right)$$

$$\leq \frac{1}{\left(\varepsilon - \beta^{-1/3}\right)^2}\left(\zeta + 2\beta^{-1/3} + \frac{\gamma^2}{k}\right)$$

$$+ 3\exp\left(-c_1\left(mp\right)^{1/3}\right) + \left(m + \frac{9}{2}mp\right)\exp\left(-c_2\beta^{1/3}\right)$$

where $c_1 := \min\left\{\frac{1}{24}, \frac{c_\phi(1-c_k)^2}{8}\right\}$ and $c_2 := \min\left\{\frac{c^2}{2}, \frac{3}{6B_0^2 + 4B_0}, \frac{1}{8B_0^2(2B_0^2 + 1)}\right\}$ are absolute constants, which may depend only on the geometry of the latent spaces. $\qquad\square$

7.4. *Proof of Theorem 4.1 (Main Result) in the Main Article.* Given the error probability bound in Theorem 7.6, we integrate the probability to provide a bound on the mean squared error.

PROOF. For a fixed (u,i),

$$Err^k(u,i) \triangleq A(u,i) - \hat{A}^k(u,i)$$

$$= \frac{1}{|\mathcal{S}_u^{\beta,k}(i)|}\sum_{v\in\mathcal{S}_u^{\beta,k}(i)}\left(A(u,i) - \hat{A}_v(u,i)\right)$$

$$= \frac{1}{|\mathcal{S}_u^{\beta,k}(i)|}\sum_{v\in\mathcal{S}_u^{\beta,k}(i)}\left(A(u,i) - Z(v,i) - m_{uv}\right).$$

By our model assumptions and the definition of B_0, it follows that $|m_{uv}| \leq B_0$, and $|A(u,i) - Z(v,i)| \leq LD_{\mathcal{X}_1} + B_e \leq B_0$. Therefore $|Err^k(u,i)| \leq 2B_0$. Since $Err^k(u,i)^2 \geq 0$, the mean squared error can be written as the following integral:

$$MSE(\hat{A}^k) := \frac{1}{mn}\sum_{u,i}\mathbb{E}\left[Err^k(u,i)^2\right]$$

$$= \frac{1}{mn}\sum_{u,i}\int_0^\infty \mathbb{P}\left(Err^k(u,i)^2 \geq t\right)dt$$

$$= \frac{1}{mn}\sum_{u,i}\int_0^\infty \mathbb{P}\left(Err^k(u,i) \geq \sqrt{t}\right)dt$$

$$= \frac{1}{mn}\sum_{u,i}\int_0^{4B_0^2}\mathbb{P}\left(Err^k(u,i) \geq \sqrt{t}\right)dt.$$

A probability cannot exceed 1, hence it follows from Theorem 7.6 that for all (u, i),

$$\mathbb{P}\left(Err^k(u, i) \geq \sqrt{t}\right) \leq \left[\frac{F_1}{(t^{1/2} - F_2)^2} + F_3\right] \wedge 1,$$

where $F_1 = \zeta + 2\beta^{-1/3} + \frac{\gamma^2}{k}$, $F_2 = \beta^{-1/3}$, and $F_3 = 3\exp\left(-c_1 (mp)^{1/3}\right) + \left(m + \frac{9}{2}mp\right)\exp\left(-c_2\beta^{1/3}\right)$, with the parameters ζ, k, β and constants c_1, c_2 as described in the theorem statement.

Since the function $\frac{F_1}{(t^{1/2}-F_2)^2} + F_3$ is monotone decreasing for $t > F_2^2$, for any partitioning parameter $t^* > F_2^2$, it follows that

$$\begin{aligned}
MSE(\hat{A}^k) &\leq \int_0^{t^*} 1 dt + \int_{t^*}^{4B_0^2} \left(\frac{F_1}{(t^{1/2} - F_2)^2} + F_3\right) dt \\
&= t^* + 2F_1 \left(-\frac{F_2}{t^{1/2} - F_2} + \ln(t^{1/2} - F_2)\right)\Bigg|_{t^*}^{4B_0^2} + F_3(4B_0^2 - t^*) \\
&= t^* - \frac{2F_1 F_2}{2B_0 - F_2} + 2F_1 \ln(2B_0 - F_2) \\
&\quad + \frac{2F_1 F_2}{(t^*)^{1/2} - F_2} - 2F_1 \ln((t^*)^{1/2} - F_2) + F_3(4B_0^2 - t^*) \\
&\leq t^* + 2F_1 F_2 \left(\frac{1}{(t^*)^{1/2} - F_2} - \frac{1}{2B_0 - F_2}\right) \\
&\quad + 2F_1 \ln\left(\frac{2B_0 - F_2}{(t^*)^{1/2} - F_2}\right) + 4B_0^2 F_3.
\end{aligned}$$

Let's choose $t^* = (F_1 + F_2)^2$, which satisfies $t^* > F_2^2$ because $F_1 > 0$. We can verify that that the definitions of F_1 and F_2, along with the constraints on the parameters specified in the theorem statement, guarantee that $F_1 \to 0$ and $F_2 \to 0$ as $n, m \to \infty$, whereas B_0 is a constant. Therefore we can assume n, m are large enough such that this choice of $t^* \leq 4B_0^2$. By substituting in our choice of t^*, it follows that

$$MSE(\hat{A}^k) \leq (F_1 + F_2)^2 + 2F_1 F_2 \left(\frac{1}{F_1} - \frac{1}{2B_0 - F_2}\right) + 2F_1 \ln\left(\frac{2B_0 - F_2}{F_1}\right) + 4B_0^2 F_3.$$

We assume without loss of generality that m, n are large enough such that $F_2 < 2B_0$, which implies

$$\frac{1}{F_1} - \frac{1}{2B_0 - F_2} \leq \frac{1}{F_1} \quad \text{and} \quad \frac{2B_0 - F_2}{F_1} \leq \frac{2B_0}{F_1}.$$

Therefore by substitution,

$$MSE(\hat{A}^k) \le (F_1 + F_2)^2 + 2F_2 + 2F_1 \ln\left(\frac{2B_0}{F_1}\right) + 4B_0^2 F_3$$

$$(7.22) \qquad = 2F_1 \ln\left(\frac{2B_0}{F_1}\right) + (F_1 + F_2)^2 + 2F_2 + 4B_0^2 F_3.$$

This MSE upper bound converges to 0 as $F_1, F_2, F_3 \to 0$. □

7.5. *Results for Specific Probability Measures.* In this section, we simplify the results from Theorem 4.1 (main article) for specific choices of probability measures, namely the uniform measure over a d dimensional Euclidean cube, or a measure which is only supported on finitely many points. Each of these cases leads to a specific form of the underestimator function $\phi(\cdot)$, which gives concrete bounds. We then choose specific expressions for ζ, and k to ensure that the mean squared error of our user-user k-nearest neighbor algorithm converges to zero. The parameter ζ in Theorem statement is introduced purely for the purpose of analysis, and is not used within the implementation of the the algorithm. Recall that it is used to define event E_4, which holds when the k rows in $\mathcal{S}_u^{\bar\beta}(i)$ with minimum variance all satisfy $\sigma^2_{x_1(u)x_1(v)} \le \zeta$. Intuitively, ζ is the thresholding parameter for the membership of similar neighbors.

PROOF OF COROLLARY 4.3 IN THE MAIN ARTICLE. When the latent space is a cube in $\mathbb{R}^d$ equipped with the uniform probability measure,

$$\phi_1\left(\sqrt{\frac{\zeta}{L^2}}\right) = C(2L)^{-d}\zeta^{d/2},$$

where C is a normalization constant to ensure $\phi_1(\mathcal{X}_1) = 1$. We need to choose ζ so that $\zeta \ge \left(\frac{c_\phi}{C}\right)^{2/d}(2L)^2(mp)^{-4/3d}$ to satisfy the constraint on ζ in the statement of Theorem 4.1 in the main article:

$$(7.23) \qquad \phi_1\left(\sqrt{\frac{\zeta}{L^2}}\right) \ge c_\phi(mp)^{-2/3} \text{ for some } c_\phi \ge 0.$$

Therefore, let $c_\phi = 1$, and set $\zeta = \frac{(2L)^2}{C^{2/d}}(mp)^{-4/3d}$. We can plug it into the expression F_1 in Theorem 4.1 (main article) to get

$$F_1 = \zeta + 2\beta^{-1/3} + \frac{\gamma^2}{k}$$

$$(7.24) \qquad \le \left[\frac{(2L)^2}{C^{2/d}} + 2\right]\max\left\{(mp)^{-4/3d}, \beta^{-1/3}\right\} + \frac{\gamma^2}{k}.$$

The assumption that $p \geq \max\left\{m^{-1+\delta}, n^{-\frac{1}{2}+\delta}\right\}$ for some $\delta > 0$ guarantees that mp and np^2 diverge to ∞ as $m, n \to \infty$. As a result, $F_2 = (np^2)^{-1/3}$ converges to 0.

To ensure $F_3 \to 0$ exponentially fast, we additionally require that $\log m < (np^2)^{\delta'/3}$ for some $\delta' > 0$. Additionally recall that $\beta = np^2/2$. The second term of F_3 is thus upper bounded by

$$
\begin{aligned}
\left(m + \frac{9}{2}mp\right) \exp\left(-c_2 \beta^{1/3}\right) &\leq \frac{11}{2} m \exp\left(-c_2 (np^2/2)^{1/3}\right) \\
&\leq \exp\left(\log \frac{11}{2} + (np^2)^{\delta'/3} - c_2(np^2/2)^{1/3}\right) \\
&\leq \exp\left(\log \frac{11}{2} - (np^2)^{1/3}\left(2^{-1/3}c_2 - (np^2)^{-(1-\delta')/3}\right)\right).
\end{aligned}
$$

Therefore, the exponent diverges to $-\infty$ polynomially in np^2 as $n \to \infty$ and hence, F_3 decays exponentially to 0 as $n \to \infty$.

We observe in Eq. (7.24) that the rate of convergence of F_1 is determined by the slowest among $\left\{(mp)^{-4/3d}, \beta^{-1/3}, \frac{1}{k}\right\}$. Since $F_2 = \beta^{-1/3}$ and F_3 decays exponentially fast, we can see that F_1 is the critical error term, which converges to 0 most slowly among F_1, F_2, F_3. By definition, $F_1 \geq 2F_2$. Hence, the convergence rate of MSE in Eq. (7.22) is dominated by the first term $2F_1 \ln\left(\frac{2B_0}{F_1}\right)$. To be more concrete, since $F_1 + F_2 = \zeta + 3\beta^{-1/3} + \frac{\gamma^2}{k} \leq \frac{3}{2}F_1$, $(F_1 + F_2)^2 \leq \frac{9}{4}F_1^2 \leq \frac{9}{4}F_1$. Given $F_1 \leq \frac{1}{2}$, we have $F_1 \leq \ln\left(\frac{2B_0}{F_1}\right)$ because $B_0 \geq 1$ and $\ln\left(\frac{2}{x}\right) \geq 1$ for $0 < x \leq \frac{1}{2}$. Therefore,

$$
\begin{aligned}
MSE(\hat{A}) &\leq 2F_1 \ln\left(\frac{2B_0}{F_1}\right) + (F_1 + F_2)^2 + 2F_2 + 4B_0^2 F_3 \\
&\leq 2F_1 \ln\left(\frac{2B_0}{F_1}\right) + \frac{9}{4}F_1^2 + F_1 + 4B_0^2 F_3 \\
&\leq \frac{21}{4}F_1 \ln\left(\frac{2B_0}{F_1}\right) + 4B_0^2 F_3
\end{aligned}
$$

(7.25)

given $F_1 \leq \frac{1}{2}$, which is satisfied for large enough $np^2 \gg 1$ and $k \gg 1$.

Now suppose that we choose $k = \frac{1}{8}(mp)^{1/3}$. Our choice of $\zeta = \frac{(2L)^2}{C^{2/d}}(mp)^{-4/3d}$ guarantees $\phi_1\left(\sqrt{\frac{\zeta}{L^2}}\right) \geq (mp)^{-2/3}$ in Eq. (7.23), and hence, $k \leq \frac{c_k}{2}(m - $

$1) p \phi_1 \left(\sqrt{\frac{\zeta}{L^2}} \right)$ for $c_k = \frac{1}{2}$ (we assumed $\frac{m}{2} \leq m - 1$ here). Then

$$F_1 = \zeta + 2\beta^{-1/3} + \frac{\gamma^2}{k}$$
$$\leq \left[\frac{(2L)^2}{C^{2/d}} + 2 + 8\gamma^2 \right] \max \left\{ (mp)^{-4/3d}, \beta^{-1/3}, (mp)^{-1/3} \right\}.$$

Plugging this into Eq. (7.25) gives a simplified bound in Corollary 4.3.

$\square$

A slight modification of the above proof yields a proof for the MSE bound when the latent row feature variable distribution $P_{\mathcal{X}_1}$ is only supported on a finite number of atoms.

PROOF OF COROLLARY 4.4 IN THE MAIN ARTICLE. When the latent space consists of a finite number of points, in other words, $P_{\mathcal{X}_1}$ is supported only on a finitely many points, any choice of $\zeta > 0$ satisfies $\phi_1 \left(\sqrt{\frac{\zeta}{L^2}} \right) \geq \min_{x \in supp(P_{\mathcal{X}_1})} \{ \phi_1(x) \}$, and hence, one can easily find $c_\phi \geq 0$ for which $\phi_1 \left(\sqrt{\frac{\zeta}{L^2}} \right) \geq c_\phi (mp)^{-2/3}$. We will simply let $\zeta = \beta^{-1/3}$ in such a case. Again, suppose that we choose $k = \frac{1}{8} (mp)^{1/3}$. This leads to

$$F_1 = \zeta + 2\beta^{-1/3} + \frac{\gamma^2}{k}$$
$$\leq \left[3 + 8\gamma^2 \right] \max \left\{ \beta^{-1/3}, (mp)^{-1/3} \right\}.$$

The remaining arguments in the proof of Corollary 4.3 with regards to terms F_2 and F_3 also follow. We plug in this upper bound for F_1 again into Eq. (7.25) to prove Corollary 4.4. $\square$

8. Proofs: Tensor Completion. Recall that our tensor completion algorithm followed from flattening the tensor to a matrix and applying our user-user k-nearest neighbor method to the resulting matrix. Therefore, the proof follows a similar outline to the proofs of the matrix results. However, it requires a separate (and little more involved) proof since in the flattened matrix obtained from tensor, the row and column latent features are likely to be correlated, not independent from each other. This subtlety requires careful handling of various arguments which we present here.

Recall that in Sections 5.1 and 5.2, we presented the latent variable model for a tensor, and we discussed the procedure for flattening a tensor to a matrix, and the model for the resulting flattened matrix.

Tensor Model. We summarize the model and relevant notation for a t-order tensor $T_A \in \mathbb{R}^{n_1 \times n_2 \times \ldots n_t}$. For each dimension $q \in [t]$, each index $\alpha_q \in [n_q]$ is associated to a latent feature $x_q(\alpha_q)$ drawn i.i.d. from the compact metric space $\mathcal{X}_q$ according to probability measure $P_{\mathcal{X}_q}$. The space $\mathcal{X}_q$ is endowed with metric $d_{\mathcal{X}_q}$ and has diameter $D_{\mathcal{X}_q}$. The underestimator function ϕ_q is defined in (5.1).

An entry of the tensor indexed by $\vec{\alpha} = (\alpha_1, \ldots \alpha_t) \in [n_1] \times \cdots \times [n_t]$ is described by the L-Lipschitz function f applied to the latent features according to (5.2). $T_Z \in \mathbb{R}^{n_1 \times n_2 \times \ldots n_t}$ is the noisy tensor derived from T_A by adding independent noise terms $\eta(\vec{\alpha})$ to each entry $T_A(\vec{\alpha})$ according to (5.4). The additive noise terms $\eta(\vec{\alpha})$ are assumed to be independent, zero mean, bounded between $[-B_e, B_e]$, and have uniform variance equal to γ^2. $\mathcal{D}$ denotes the index set of observed entries, such that $\vec{\alpha} \in \mathcal{D}$ with probability p and $\vec{\alpha} \notin \mathcal{D}$ with probability $1-p$ for any index $\vec{\alpha}$ independently of all other entries.

Flattened Matrix. There may be many ways to flatten a tensor into a matrix, each corresponding to a specific bi-partition of $[t]$ denoted by $(\mathcal{I}_1, \mathcal{I}_2)$ where $\mathcal{I}_1 = \{\pi(1), \ldots, \pi(t_1)\}$ and $\mathcal{I}_2 = \{\pi(t_1 + 1), \ldots, \pi(t)\}$ for some $1 \leq t_1 \leq t - 1$ and some permutation $\pi : [t] \to [t]$. Recall $t_2 = t - t_1$. Given a partitioning of the dimensions, the rows of the flattened matrix would correspond to taking the cartesian product of all dimensions in $\mathcal{I}_1$, and the columns of the matrix would correspond to taking the cartesian product of all dimensions in $\mathcal{I}_2$. If the dimensions of the matrix are denoted by $m \times n$, then $m = \times_{q \in \mathcal{I}_1}[n_q]$ and $n = \times_{q \in \mathcal{I}_2}[n_q]$.

A row u in the matrix is associated to a vector of indices $\vec{u} = (u_1, \ldots, u_{t_1}) \in [n_{\pi(1)}] \times \cdots \times [n_{\pi(t_1)}]$, and similarly a column i in the matrix is associated to a vector of indices $\vec{i} = (i_1, \ldots, i_{t_2}) \in [n_{\pi(t_1+1)}] \times \cdots \times [n_{\pi(t)}]$. The corresponding row and column features are denoted by

$$\vec{x}_1^\pi(\vec{u}) = \left(x_{1,1}^\pi(u_1), \ldots, x_{1,t_1}^\pi(u_{t_1})\right) = \left(x_{\pi(1)}(u_1), \ldots, x_{\pi(t_1)}(u_{t_1})\right),$$
$$\vec{x}_2^\pi(\vec{i}) = \left(x_{2,1}^\pi(i_1), \ldots, x_{2,t_2}^\pi(i_{t_2})\right) = \left(x_{\pi(t_1+1)}(i_1), \ldots, x_{\pi(t_1+t_2)}(i_{t_2})\right).$$

The latent row features belong to space $\mathcal{X}_1^\pi = \times_{q \in \mathcal{I}_1} \mathcal{X}_q$, and the latent column features belong to $\mathcal{X}_2^\pi = \times_{q \in \mathcal{I}_2} \mathcal{X}_q$. We let $\mathcal{X}_{1,k}^\pi$ denote $\mathcal{X}_{\pi(k)}$ for $k \in [t_1]$, and we let $\mathcal{X}_{2,k}^\pi$ denote $\mathcal{X}_{\pi(t_1+k)}$ for $k \in [t_2]$. We define the metric for the product spaces according to the max over the distance in the individual latent space associated to each dimension,

$$(8.1) \qquad\qquad d_{\mathcal{X}_1^\pi}(\vec{x}, \vec{x}') = \max_{q \in [t_1]}(d_{\mathcal{X}_{\pi(q)}}(x_q, x_q'))$$

$$(8.2) \qquad\qquad d_{\mathcal{X}_2^\pi}(\vec{x}, \vec{x}') = \max_{q \in [t_2]}(d_{\mathcal{X}_{\pi(t_1+q)}}(x_q, x_q')).$$

Therefore the diameter of $\mathcal{X}_1^\pi$ is $D_{\mathcal{X}_1^\pi} := \max(D_{\mathcal{X}_{1,1}^\pi}, \ldots D_{\mathcal{X}_{1,t_1}^\pi})$. The measure associated to the row latent space is the product measure corresponding to $P_{\mathcal{X}_{1,1}^\pi} \ldots P_{\mathcal{X}_{1,t_1}^\pi}$,

and its underestimator function is denoted by

$$\phi_1^\pi(r) := \prod_{q \in \mathcal{I}_1} \phi_q(r).$$

If the tensors T_A and T_Z are drawn from the tensor latent variable model, then the entries of the associated matrices A and Z, obtained by flatting the tensors T_A and T_Z, can be described as:

$$(8.3) \qquad\qquad A(u,i) = f(\vec{x}_1^\pi(\vec{u}), \vec{x}_2^\pi(\vec{i}))$$

$$(8.4) \qquad\qquad Z(u,i) = A(u,i) + \eta(u,i).$$

This flattened matrix satisfies some of the conditions required in the matrix setting in Section 2, namely f is Lipschitz, the latent variables are drawn from a bounded metric space, and the data is observed with additive bounded zero-mean noise. We define

$$(8.5) \qquad\qquad B_0 \triangleq L D_{\mathcal{X}_1^\pi} + 2B_e,$$

such that for any $u, v \in [m]$ and any $i \in [n]$,

$$|Z(u,i) - Z(v,i)| = \left| f(\vec{x}_1(\vec{u}), \vec{x}_2(\vec{i})) + \eta(u,i) - f(\vec{x}_1(\vec{v}), \vec{x}_2(\vec{i})) - \eta(v,i) \right|$$
$$\leq L D_{\mathcal{X}_1^\pi} + 2B_e =: B_0.$$

However the flattened matrix does not satisfy the condition that the row and column latent variables are fully independent. Although the latent variables for each coordinate of each dimension of the tensor are independently sampled, if two rows $\vec{u}$ and $\vec{u}'$ in the flattened matrix correspond to tensor indices that coincide, i.e. if $u_h = u_h'$ for any $h \in [t_1]$, then the corresponding component of the latent variable $x_{1,h}^\pi(u_h)$ and $x_{1,h}^\pi(u_h')$ will be correlated, in fact they must be equal. The correlation structure amongst the rows and columns of the flattened matrix is very specific and follows from the partitioning and flattening of the tensor to a matrix. For some column i, recall that $\mathcal{N}_i \subset [n]$ denote the set of columns which do not share a tensor coordinate with i, and thus whose latent features are uncorrelated with $\vec{x}_2^\pi(\vec{i})$,

$$\mathcal{N}_i := \{j \in [n] \ s.t. \ j_k \neq i_k \text{ for all } k \in [t_2]\}.$$

The number of remaining columns in the matrix after removing one coordinate from each tesnor dimension is denoted by

$$|\mathcal{N}_i| = n' := \prod_{q \in \mathcal{I}_2} (n_q - 1).$$

Proof Outline.   The proof follows similar steps to the proof of the matrix results. The algorithm corresponds to first computing the empirical means and variances of the differences between pairs of rows in the flattened matrix. These computed variances are used to choose the k-nearest neighbors which are included in the final estimate. The differences we made to the algorithm is that the set $\mathcal{S}_u^{\beta_l,\beta_h}(i)$ both imposes and upper and lower bound on the overlap between rows. Additionally, since the matrix results from flattening a tensor, the latent variables for rows and columns are correlated according to the original tensor structure. Therefore, we modified the computation of empirical means and variances such that when we are estimating an entry at (u, i), we only consider columns j which do not share any coordinates in the original tensor representation, i.e. $i_k \neq j_k$ for all $k \in [t_2]$. In this way, the computation of $m_{uv}(i)$, $s_{uv}^2(i)$, and $\mathcal{S}_u^{\beta_l,\beta_h}(i)$ become fully independent from the latent variables associated to i denoted by $\vec{x}_2^\pi(\vec{i})$.

The key steps of the proof are stated within a few lemmas that mirror the same format as the matrix proof. However, we will need to modify the proof of these corresponding tensor lemmas to account for the fact that the row and column features are no longer drawn independently from the product spaces $\mathcal{X}_1^\pi, \mathcal{X}_2^\pi$. We first prove that the overlaps between pairs of rows is sufficiently large, which is presented in Lemma 8.1 and is equivalent to the corresponding matrix Lemma 7.1. Next we prove concentration of the empirical means and variances in Lemmas 8.2 and 8.3, which correspond to the matrix Lemmas 7.2 and 7.3. Then we prove that each row has sufficiently many "good neighbors", i.e. other rows whose latent features are close to the target row's latent features. This is presented in Lemma 8.4, which corresponds to 7.4. The proofs of Lemmas 8.2, 8.3, and 8.4 will need to account for the correlations across latent variables. Finally, conditioned on these above good events, we combine them together to prove a bound on the probability of error in Lemma 8.5.

We first prove that every pair of rows has sufficient, but not too huge overlap with high probability. Specifically, Lemma 8.1 proves that for any (u, i), the number of the candidate rows, $|\mathcal{S}_u^{\beta_l,\beta_h}(i)|$, concentrates around $(m - 1)p$. This relies on concentration of Binomial random variables via Chernoff's bound, using the fact that every entry is independently observed uniformly at random. The proof is essentially identical to the proof of Lemma 7.1, and the result upper bounds the probability of the complement of the following good event

$$(8.6) \qquad E_1' := \left\{ \left| \mathcal{S}_u^{\beta_l,\beta_h}(i) \right| \in \left[\frac{1}{2}(m - 1)p, \frac{3}{2}(m - 1)p \right] \right\}.$$

Because $|\mathcal{O}_i^{uv}|$ only considers columns that are do not share any tensor coordinates with i, it is distributed according to a Binomial with parameters n' and p^2, where $n' = \prod_{q \in \mathcal{I}_2}(n_q - 1)$.

LEMMA 8.1. *Given $p > 0$, $2 \le \beta_l < n'p^2$, $n'p^2 \le \beta_h < n'$ and $\theta \in (0, 1)$, for any $(u, i) \in [m] \times [n]$,*

$$\mathbb{P}\left(|\mathcal{S}_u^{\beta_l, \beta_h}(i) - (m-1)p| > \theta(m-1)p\right)$$

$$\le (m-1)\exp\left(-\frac{(n'p^2 - \beta_l)^2}{2n'p^2}\right) + (m-1)\exp\left(-\frac{(\beta_h - n'p^2)^2}{3n'p^2}\right)$$

$$+ 2\exp\left(-\frac{\theta^2}{3}(m-1)p\right).$$

Next, assuming the overlap between two rows u and v is sufficiently large (but not too large), Lemmas 8.2 and 8.3 prove that the sample mean and variance, $m_{uv}(i)$ and $s_{uv}^2(i)$ of the difference $Z(u, i) - Z(v, i)$ concentrate around their expectations $\mu_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}$ and $\sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}^2$ with high probability. For any pair of rows $(u, v) \in [m] \times [n]$, we define two levels of empirical means and variances. As defined in the algorithm, $m_{uv}(i)$ and $s_{uv}^2(i)$ are computed from the observed entries in the overlap of u and v, where we recall that $j \in \mathcal{O}_i^{uv}$ if $(u, j) \in \mathcal{D}$, $(v, j) \in \mathcal{D}$, and $j_k \ne i_k$ for all $k \in [t_2]$.

$$m_{uv}(i) := \frac{1}{|\mathcal{O}_i^{uv}|}\left[\sum_{j \in \mathcal{O}_i^{uv}} Z(u, j) - Z(v, j)\right],$$

$$s_{uv}^2(i) := \frac{1}{|\mathcal{O}_i^{uv}| - 1}\sum_{j \in \mathcal{O}_i^{uv}} [Z(u, j) - Z(v, j) - m_{uv}(i)]^2.$$

We define $\tilde{m}_{uv}(i)$ and $\tilde{s}_{uv}^2(i)$ to be the mean and variances of the differences between rows u and v in the full matrix Z including the unobserved entries, but not including all columns j which share a tensor coordinate with i. Recall that the set $\mathcal{N}_i$ denotes columns which have no shared tensor coordinate with i,

$$\mathcal{N}_i := \{j \in [n] \; s.t. \; j_k \ne i_k \text{ for all } k \in [t_2]\}.$$

The remaining number of columns would be $|\mathcal{N}_i| = n' = \prod_{q \in \mathcal{I}_2}(n_q - 1)$ instead of $n = \prod_{q \in \mathcal{I}_2} n_q$.

$$\tilde{m}_{uv}(i) := \frac{1}{n'}\sum_{j \in \mathcal{N}_i} [Z(u, j) - Z(v, j)],$$

$$\tilde{s}_{uv}^2(i) := \frac{1}{n' - 1}\sum_{j \in \mathcal{N}_i} [Z(u, j) - Z(v, j) - \tilde{m}_{uv}(i)]^2.$$

While Z may be fully instantiated from the latent row and column variables along with entrywise noise terms, only the entries denoted by $\mathcal{D}$ are actually observed, such that we are not actually able to compute $\tilde{m}_{uv}(i)$ and $\tilde{s}^2_{uv}(i)$ from the observed dataset. Whereas $\tilde{m}_{uv}(i)$ and $\tilde{s}^2_{uv}(i)$ denote the sample mean and variance between a pair of rows for the given sampled columns, $m_{uv}(i)$ and $s^2_{uv}(i)$ denote the sample mean and variance between a pair of rows for the observed datapoints, which can be thought of as drawn uniformly at random without replacement from the columns. We define μ_{ab} and σ^2_{ab} to be the mean and variance for a fixed pair of rows with latent features $a, b \in \mathcal{X}^\pi_1$ and for a randomly sampled column feature $\vec{\mathbf{x}}^\pi_2$ according to the product measure over $P_{\mathcal{X}^\pi_{2,1}} \ldots P_{\mathcal{X}^\pi_{2,t_2}}$:

$$
\begin{aligned}
\mu_{ab} &:= \mathbb{E}_{\vec{\mathbf{x}}^\pi_2} \left[f(a, \vec{\mathbf{x}}^\pi_2) - f(b, \vec{\mathbf{x}}^\pi_2) \right], \\
\sigma^2_{ab} &:= Var_{\vec{\mathbf{x}}^\pi_2} \left[f(a, \vec{\mathbf{x}}^\pi_2) - f(b, \vec{\mathbf{x}}^\pi_2) \right].
\end{aligned}
$$

Due to the fact that the latent variables across columns in the flattened matrix are correlated according to the original tensor structure, the simple proof from Lemmas 7.2 and 7.3 are no longer sufficient. Instead we use a 2-step analysis, which arises from the equivalent perspective of the model as generated from a 2-stage sampling procedure. In step one, the latent variables and individual noise terms are sampled according to the distributions specified in the model, i.e. each feature of a coordinate in the original tensor representation is drawn i.i.d. and the noise terms are independent, zero-mean, and bounded. The matrices A and Z are fully determined from the latent variables and individual noise terms. In step two, the index set $\mathcal{D}$ of observed entries is sampled such that for any $(u, i) \in [m] \times [n]$, with probability p, $(u, i) \in \mathcal{D}$ independently of other entries. The algorithm only observes the entries in Z which correspond to indices specified in the set $\mathcal{D}$.

To prove that $m_{uv}(i)$ concentrates around $\mu_{\vec{x}^\pi_1(\vec{u})\vec{x}^\pi_1(\vec{v})}$, we first show using Mc-Diarmid's inequality that $\tilde{m}_{uv}(i)$ concentrates around $\mu_{\vec{x}^\pi_1(\vec{u})\vec{x}^\pi_1(\vec{v})}$ due to the sampling of the latent variables and independent noise terms. Second we show that conditioned on $|\mathcal{O}^{uv}_i|$, $m_{uv}(i)$ concentrates around $\tilde{m}_{uv}(i)$ by the Kontorovich-Ramanan inequality, since the observed datapoints are sampled without replacement from a finite set of columns. We formally state the result in Lemma 8.2, which upper bounds the probability of the complement of the following good event

$$
(8.7) \qquad E'_2 := \left\{ \left| m_{uv}(i) - \mu_{\vec{x}^\pi_1(\vec{u})\vec{x}^\pi_1(\vec{v})} \right| \le \nu, \ \forall \ v \in \mathcal{S}^{\beta_l, \beta_h}_u(i) \right\}.
$$

LEMMA 8.2. *Given $(u, i) \in [m] \times [n]$, for any $\nu > 0$,*

$$\mathbb{P}\left(\left| m_{uv}(i) - \mu_{\tilde{x}_1^\pi(\tilde{u})\tilde{x}_1^\pi(\tilde{v})} \right| > \nu \,\middle|\, v \in \mathcal{S}_u^{\beta_l,\beta_h}(i) \right)$$

$$\leq 2\exp\left(\frac{-\nu^2}{8(LD_{\mathcal{X}_1^\pi})^2 \sum_{q=1}^{t_2} \frac{1}{n_q-1} + \frac{16B_e^2}{n'}} \right) + 2\exp\left(\frac{-\nu^2}{32B_0^2}\Delta \right),$$

where $\Delta = \min\left\{ \frac{n'^2\beta_l}{\left(n'+\beta_l^2\right)^2}, \frac{n'^2\beta_h}{\left(n'+\beta_h^2\right)^2} \right\}$.

Similarly, to prove that $s_{uv}^2(i)$ concentrates around $\sigma_{\tilde{x}_1^\pi(\tilde{u})\tilde{x}_1^\pi(\tilde{v})}^2 + 2\gamma^2$, we first show using McDiarmid's inequality that $\tilde{s}_{uv}^2(i)$ concentrates around $\sigma_{\tilde{x}_1^\pi(\tilde{u})\tilde{x}_1^\pi(\tilde{v})}^2 + 2\gamma^2$ due to the sampling of the latent variables and independent noise terms. Second we show that conditioned on $|\mathcal{O}_i^{uv}|$, $s_{uv}^2(i)$ concentrates around $\tilde{s}_{uv}^2(i)$ by the Kontorovich-Ramanan inequality, since the observed datapoints are sampled without replacement from a finite set of columns. We formally state the result in Lemma 8.3, which upper bounds the probability of the complement of the following good event

(8.8)
$$E_3' := \left\{ s_{uv}^2(i) \in \left[(1-\theta)\sigma_{\tilde{x}_1^\pi(\tilde{u})\tilde{x}_1^\pi(\tilde{v})}^2 - \tau, (1+\theta)\sigma_{\tilde{x}_1^\pi(\tilde{u})\tilde{x}_1^\pi(\tilde{v})}^2 + \tau \right], \forall\, v \in \mathcal{S}_u^{\beta_l,\beta_h}(i) \right\}.$$

LEMMA 8.3. *Given $u \in [m], i \in [n]$, for any $\tau > 0$,*

$$\mathbb{P}\left(s_{uv}^2(i) - 2\gamma^2 \in \left[(1-\theta)\sigma_{\tilde{x}_1^\pi(\tilde{u})\tilde{x}_1^\pi(\tilde{v})}^2 - \tau, (1+\theta)\sigma_{\tilde{x}_1^\pi(\tilde{u})\tilde{x}_1^\pi(\tilde{v})}^2 + \tau \right] \,\middle|\, v \in \mathcal{S}_u^{\beta_l,\beta_h}(i) \right)$$

$$\leq 2\exp\left(\frac{-\tau^2}{32L^2 D_{\mathcal{X}_1^\pi}^2 \left(3LD_{\mathcal{X}_1^\pi} + 4B_e\right)^2 \sum_{q=1}^{k_2} \frac{1}{n_q-1} + \frac{64B_e^2\left(2LD_{\mathcal{X}_1^\pi} + 5B_e\right)^2}{n'}} \right)$$

$$+ 2\exp\left(\frac{-\tau^2}{128B_0^4}\Delta \right),$$

where $\theta = \sum_{q \in \mathcal{I}_2} \frac{1}{n_q-1}$ is a quantity which depends only on the shape of the given tensor and vanishes to 0 as $n_q \to \infty, \forall q \in \mathcal{I}_2$. The overlap-dependent factor in the exponent $\Delta = \min\left\{ \frac{n'^2\beta_l}{\left(n'+\beta_l^2\right)^2}, \frac{n'^2\beta_h}{\left(n'+\beta_h^2\right)^2} \right\}$.

From Lemmas 8.2 and 8.3, we observe that the error bound vanishes as $n \to \infty$ if and only if $\Delta \to \infty$, which occurs when $\beta_l = \omega(1)$ and $\beta_h = o\left(n^{2/3}\right)$, since $\beta_l \leq \beta_h$ and $n' = \Theta(n)$ by definition.

In Lemma 8.4, we prove that for any index pair (u, i), there exists k "good" neighboring rows in $\mathcal{S}_u^{\beta_l, \beta_h}(i)$, whose variance $\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}$ is sufficiently small. Unlike Lemma 7.4, we cannot assume our row features are drawn independently the product latent space. However, we recall that the row features come from the original tensor structure, in which for each dimension, the latent feature associated to each coordinate is drawn independently and identically distributed. Consider a row u in the flattened matrix, with a vector of latent features $\vec{x}_1^\pi(\vec{u}) = (x_{1,1}^\pi(u_1), \ldots, x_{1,t_1}^\pi(u_{t_1}))$. For each dimension of the original tensor $q \in [t_1]$, by Chernoff's bound with high probability there are sufficiently many coordinates whose latent feature is close to $x_{1,q}^\pi(u_q)$. Then by combining these events by a simple union bound, it follows that there are sufficiently many rows $v \in [m]$ in the flattened matrix whose latent feature vector $\vec{x}_1^\pi(\vec{v})$ is close to $\vec{x}_1^\pi(\vec{u})$, implying that there are at least k good neighbors. Let $\left\{\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}\right\}_{v \in \mathcal{S}_u^{\beta_l, \beta_h}(i)}^{(k)}$ denote the value of the k-th minimum element in the set $\left\{\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}\right\}_{v \in \mathcal{S}_u^{\beta_l, \beta_h}(i)}$. The following Lemma 8.4 upper bounds the probability of the complement of the following good event

$$(8.9) \qquad E_4' := \left\{\left\{\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}\right\}_{v \in \mathcal{S}_u^{\beta_l, \beta_h}(i)}^{(k)} \leq \zeta\right\}.$$

LEMMA 8.4.   *Given $u \in [m]$, $i \in [n]$, for any $\zeta > 0$ and for any positive integer $k \leq \frac{1}{8} m p \phi_2^\pi\left(\sqrt{\frac{\zeta}{L^2}}\right) = \frac{p}{8} \prod_{q \in [t_1]} n_{\pi(q)} \phi_{\pi(q)}\left(\sqrt{\frac{\zeta}{L^2}}\right)$,*

$$\mathbb{P}\left(\left\{\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}\right\}_{v \in \mathcal{S}_u^{\beta_l, \beta_h}(i)}^{(k)} > \zeta \ \middle| \ |\mathcal{S}_u^{\beta_l, \beta_h}(i)| \in \left[\frac{1}{2}(m-1)p, \frac{3}{2}(m-1)p\right]\right)$$

$$\leq \sum_{q \in \mathcal{I}_1} \exp\left(-\frac{(1 - 2^{-1/t_1})^2}{2}\mu_q\right) + \exp\left(-\frac{k}{8}\right)$$

$$+ (m-1)\exp\left(-\frac{(n'p^2 - \beta_l)^2}{2n'p^2}\right) + (m-1)\exp\left(-\frac{(\beta_h - n'p^2)^2}{3n'p^2}\right),$$

where $\mu_q := n_{\pi(q)} \phi_{\pi(q)}\left(\sqrt{\frac{\zeta}{L^2}}\right)$ and $\phi_{\pi(q)}(r) := ess\inf_{x' \in \mathcal{X}_{\pi(q)}} \mathbb{P}_{\mathcal{X}_{\pi(q)}}\left(d_{\mathcal{X}_{\pi(q)}}(\mathbf{x}, x') \leq r\right)$.

Given that there exist k good neighbors in $\mathcal{S}_u^{\beta_l, \beta_h}(i)$ whose variance is small, and conditioned on the event that all the sample variances concentrate, it follows that the true variance between u and its k nearest neighbors are small with high probability. Each of these good events were denoted as E_1', E_2', E_3' and E_4', defined in (8.6), (8.7), (8.8), and (8.9) respectively, and we let E' denote the intersection of these good events,

$$E' := E_1' \cap E_2' \cap E_3' \cap E_4'.$$

We can provide a bound on the tail probability of the estimation error conditioned on E' by using Chebyshev's inequality, similar to that in Lemma 7.5 with minor changes.

LEMMA 8.5. *Given $\nu > 0$, $\tau > 0$, for any $\varepsilon > \nu$, $\zeta \geq 0$ and any positive integer $k \leq \frac{1}{8}mp\phi_1^\pi\left(\sqrt{\frac{\zeta}{L^2}}\right)$, the tail probability of the estimation error is bounded by*

$$\mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon \mid E'\right) \leq \frac{1}{(\varepsilon - \nu)^2}\left(\frac{(1+\theta)\zeta + 2\tau}{1-\theta} + \frac{\gamma^2}{k}\right).$$

where $\theta = \sum_{q \in \mathcal{I}_2} \frac{1}{n_q - 1}$ is a quantity which depends only on the shape of the given tensor and vanishes to 0 as $n_q \to \infty, \forall q \in \mathcal{I}_2$.

Then we bound the "bad" event, denoted by the complement of E' according to the above definition.

LEMMA 8.6. *Suppose that*

$$\max\left\{m^{-1+\delta}, n'^{-\frac{1}{2}+\delta}\right\} \leq p \leq n'^{-\frac{1}{6}-\delta} \text{ for some } \delta > 0$$

$$\forall q \in \mathcal{I}_1, \zeta \text{ satisfies } \phi_q\left(\sqrt{\frac{\zeta}{L^2}}\right) \geq c_q n_q^{-\frac{\log mp}{2\log m}} \text{ for some } c_q > 0,$$

$$2 \leq \beta_l \leq c_l \min\left\{n'p^2, n'^{1/2}\right\} \text{ for some } c_l \in (0,1),$$

$$c_h \max\left\{n'^{1/2}, n'p^2\right\} \leq \beta_h \leq n'^{\frac{2}{3}-\delta} \text{ for some } c_h > 1 \text{ and}$$

$$k \leq \frac{1}{8}mp\phi_1^\pi\left(\sqrt{\frac{\zeta}{L^2}}\right) = \frac{p}{8}\prod_{q \in \mathcal{I}_1} n_q \phi_q\left(\sqrt{\frac{\zeta}{L^2}}\right).$$

Then

$$\mathbb{P}\left(E'^c\right) \leq 4(m-1)\exp\left(-C_1 n'p^2\right) + 2\exp\left(-\frac{1}{24}mp\right)$$

$$+ 6(m-1)p\exp\left(-C_2(n_{q^*}-1)^{1/3}\right) + 6(m-1)p\exp\left(-C_3 \min\left\{\frac{n'^{2/3}}{4\beta_h}, \frac{\beta_l^{1/3}}{4}\right\}\right)$$

$$+ t_1 \exp\left(-C_4 n_{q^*}^{1/2}\right) + \exp\left(-\frac{k}{8}\right).$$

where $q^* := \arg\min_{q \in \mathcal{I}_1} n_q$ and $\nu = \tau = \max\left\{(n_{q^*}-1)^{-1/3}, \left(\frac{n'^2}{\beta_h^3}\right)^{-1/3}, \beta_l^{-1/3}\right\}$.

Here, $\theta = \sum_{q \in \mathcal{I}_2} \frac{1}{n_q - 1}$ is a quantity which depends only on the shape of the given

tensor and vanishes to 0 as $n_q \to \infty, \forall q \in \mathcal{I}_2$. Note that

$$C_1 := \min\left\{\frac{(1 - c_l)^2}{2}, \frac{(c_h - 1)^2}{3}\right\},$$

$$C_2 := \min\left\{\frac{1}{8L^2 D^2 t_2 + 16B_e^2}, \frac{1}{32L^2 D^2 (3LD + 4B_e)^2 t_2 + 64B_e^2 (2LD + 5B_e)^2}\right\},$$

$$C_3 := \min\left\{\frac{1}{32(LD + 2B_e)^2}, \frac{1}{128(LD + 2B_e)^4}\right\},$$

$$C_4 := \min_{q \in \mathcal{I}_q}\left\{\frac{(1 - 2^{-1/t_1})^2}{2} c_q\right\}$$

are some absolute constants, which may depend only on the geometry of the latent spaces.

Finally we combine the lemmas which bound each of the deviating events to get a final bound on the tail probability of the error.

THEOREM 8.7.   *Suppose that*

$$\max\left\{m^{-1+\delta}, n'^{-\frac{1}{2}+\delta}\right\} \le p \le n'^{-\frac{1}{6}-\delta} \text{ for some } \delta > 0$$

$$\forall q \in \mathcal{I}_1, \zeta \text{ satisfies } \phi_q\left(\sqrt{\frac{\zeta}{L^2}}\right) \ge c_q n_q^{-\frac{\log mp}{2\log m}} \text{ for some } c_q > 0,$$

$$2 \le \beta_l \le c_l \min\left\{n'p^2, n'^{1/2}\right\} \text{ for some } c_l \in (0, 1),$$

$$c_h \max\left\{n'^{1/2}, n'p^2\right\} \le \beta_h \le n'^{\frac{2}{3}-\delta} \text{ for some } c_h > 1 \text{ and}$$

$$k \le \frac{1}{8}mp\phi_1^\pi\left(\sqrt{\frac{\zeta}{L^2}}\right) = \frac{p}{8}\prod_{q \in \mathcal{I}_1} n_q \phi_q\left(\sqrt{\frac{\zeta}{L^2}}\right).$$

For any given $\varepsilon > \max\left\{\max_{q \in \mathcal{I}_1}\left((n_q - 1)^{-1/3}\right), \left(\frac{n'^2}{\beta_h^3}\right)^{-1/3}, \beta_l^{-1/3}\right\}$, the tail probability of the error of the estimate produced by the user-user k-smoothed variant of our method with overlap parameters β_l, β_h is upper bounded by:

$$\mathbb{P}\left(\left|A(u, i) - \hat{A}^k(u, i)\right| > \varepsilon\right)$$

$$\le \frac{1}{(\varepsilon - \nu)^2}\left(\frac{(1 + \theta)\zeta + 2\tau}{1 - \theta} + \frac{\gamma^2}{k}\right) + F_3',$$

where

$$F_3' = 4(m-1)\exp\left(-C_1 n' p^2\right) + 2\exp\left(-\frac{1}{24}mp\right)$$

$$+ 6(m-1)p\exp\left(-C_2(n_{q^*}-1)^{1/3}\right) + 6(m-1)p\exp\left(-C_3 \min\left\{\frac{n'^{2/3}}{4\beta_h}, \frac{\beta_l^{1/3}}{4}\right\}\right)$$

$$+ t_1\exp\left(-C_4 n_{q^*}^{1/2}\right) + \exp\left(-\frac{k}{8}\right),$$

with $q^* := \arg\min_{q \in \mathcal{I}_1} n_q$ and $\nu = \tau = \max\left\{(n_{q^*}-1)^{-1/3}, \left(\frac{n'^2}{\beta_h^3}\right)^{-1/3}, \beta_l^{-1/3}\right\}$.
Here, $\theta = \sum_{q \in \mathcal{I}_2} \frac{1}{n_q - 1}$ is a quantity which depends only on the shape of the given
tensor and vanishes to 0 as $n_q \to \infty, \forall q \in \mathcal{I}_2$. Note that

$$C_1 := \min\left\{\frac{(1-c_l)^2}{2}, \frac{(c_h-1)^2}{3}\right\},$$

$$C_2 := \min\left\{\frac{1}{8L^2 D^2 t_2 + 16 B_e^2}, \frac{1}{32 L^2 D^2 (3LD + 4B_e)^2 t_2 + 64 B_e^2 (2LD + 5B_e)^2}\right\},$$

$$C_3 := \min\left\{\frac{1}{32(LD + 2B_e)^2}, \frac{1}{128(LD + 2B_e)^4}\right\},$$

$$C_4 := \min_{q \in \mathcal{I}_q}\left\{\frac{(1 - 2^{-1/t_1})^2}{2} c_q\right\}$$

are some absolute constants, which may depend only on the geometry of the latent
spaces.

By integrating the upper bound on the tail of the error probability, we obtain an upper bound on the MSE of the estimate, as stated in the final Theorem 5.1.

In this section we prove the five key lemmas introduced in the proof outline.

8.1. *Sufficiently many rows with good overlap.* Using the fact that every entry is independently observed uniformly at random, the size of the overlap between a pair of rows is distributed according to a Binomial random variable. Therefore Lemma 8.1 follows from a straightforward application of Chernoff's bound.

PROOF OF LEMMA 8.1. A slight modification of the proof for Lemma 7.1 yields the result. It suffices to redefine R_{uv} as

$$R_{uv} = \mathbb{I}\{|\mathcal{O}_i^{uv}| \geq \beta_l\}\mathbb{I}\{|\mathcal{O}_i^{uv}| \leq \beta_h\}.$$

Since the measurement at each entry still happens i.i.d. with probability p, using Chernoff's bound for deviation from mean above and below, we can bound $\mathbb{P}(R_{uv} = 0)$. By an application of union bound as in Lemma 7.1, we can obtain the desired result. □

8.2. *Concentration of mean and variance.* Assuming the overlap between two rows u and v is sufficiently large (but not too large), Lemmas 8.2 and 8.3 prove that the sample mean and variance, $m_{uv}(i)$ and $s^2_{uv}(i)$ of the difference $Z(u,i) - Z(v,i)$ concentrate around their expectations $\mu_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}$ and $\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}$ with high probability. For any pair of rows $(u, v) \in [m] \times [m]$, recall the following definitions

$$m_{uv}(i) := \frac{1}{|\mathcal{O}_i^{uv}|} \left(\sum_{j \in \mathcal{O}_i^{uv}} Z(u,j) - Z(v,j) \right),$$

$$s^2_{uv}(i) := \frac{1}{|\mathcal{O}_i^{uv}| - 1} \sum_{j \in \mathcal{O}_i^{uv}} (Z(u,j) - Z(v,j) - m_{uv}(i))^2,$$

$$:= \frac{1}{2|\mathcal{O}_i^{uv}|(|\mathcal{O}_i^{uv}| - 1)} \sum_{j,h \in \mathcal{O}_i^{uv} \times \mathcal{O}_i^{uv}} ((Z(u,j) - Z(v,j)) - (Z(u,h) - Z(v,h)))^2,$$

$$\tilde{m}_{uv}(i) := \frac{1}{n'} \sum_{j \in \mathcal{N}_i} (Z(u,j) - Z(v,j)),$$

$$\tilde{s}^2_{uv}(i) := \frac{1}{n' - 1} \sum_{j \in \mathcal{N}_i} (Z(u,j) - Z(v,j) - \tilde{m}_{uv}(i))^2,$$

$$= \frac{1}{2n'(n' - 1)} \sum_{j,h \in \mathcal{N}_i \times \mathcal{N}_i} ((Z(u,j) - Z(v,j)) - (Z(u,h) - Z(v,h)))^2,$$

$$\mu_{ab} := \mathbb{E}_{\vec{\mathbf{x}}_2^\pi} [f(a, \vec{\mathbf{x}}_2^\pi) - f(b, \vec{\mathbf{x}}_2^\pi)],$$

$$\sigma^2_{ab} := Var_{\vec{\mathbf{x}}_2^\pi} [f(a, \vec{\mathbf{x}}_2^\pi) - f(b, \vec{\mathbf{x}}_2^\pi)].$$

As $\mathcal{O}_i^{uv}$ only contains columns j for which both $Z(u,j)$ and $Z(v,j)$ are observed, i.e. $(u,j) \in \mathcal{D}$ and $(v,j) \in \mathcal{D}$, the quantities $m_{uv}(i)$ and $s^2_{uv}(i)$ can be computed from the data and are used in the algorithm. $\tilde{m}_{uv}(i)$ and $\tilde{s}^2_{uv}(i)$ is defined from the full matrix Z, which is fully determined by the latent row and column latent variables along with the individual noise terms. However, as many entries are unobserved, these quantities are not computable from the data and are defined for the purpose of the analysis.

PROOF OF LEMMA 8.2. Without loss of generality, to simplify the notation,

assume that $i = (n_{\pi(t_1+1)}, \ldots n_{\pi(t)})$. Therefore

$$\mathcal{N}_i = \{j \in [n] \ s.t. \ j_k \neq i_k \text{ for all } k \in [t_2]\}$$
$$= \{j \in [n] \ s.t. \ j_k \in [n_{\pi(t_1+k)} - 1] \text{ for all } k \in [t_2]\}$$
$$= \{j \in [n] \ s.t. \ \vec{j} \in \times_{k \in [t_2]} [n_{\pi(t_1+k)} - 1]\}.$$

The proof is equivalent for all choice of $i \in [n]$ by symmetry.

By the triangle inequality,

$$\left| \mu_{\vec{x}_1^\pi(\vec{u}) \vec{x}_1^\pi(\vec{v})} - m_{uv}(i) \right| \leq \left| \mu_{\vec{x}_1^\pi(\vec{u}) \vec{x}_1^\pi(\vec{v})} - \tilde{m}_{uv}(i) \right| + \left| \tilde{m}_{uv}(i) - m_{uv}(i) \right|.$$

Therefore, $\left| \mu_{\vec{x}_1^\pi(\vec{u}) \vec{x}_1^\pi(\vec{v})} - m_{uv}(i) \right| > \nu$ implies that either $\left| \mu_{\vec{x}_1^\pi(\vec{u}) \vec{x}_1^\pi(\vec{v})} - \tilde{m}_{uv}(i) \right|$ or $|\tilde{m}_{uv}(i) - m_{uv}(i)|$ are larger than $\nu/2$, such that a simple application of union bound results in

$$\mathbb{P}\left(\left| m_{uv}(i) - \mu_{\vec{x}_1^\pi(\vec{u}) \vec{x}_1^\pi(\vec{v})} \right| > \nu \ \middle| \ v \in \mathcal{S}_u^{\beta_l, \beta_h}(i) \right)$$
$$= \mathbb{P}\left(\left| \mu_{\vec{x}_1^\pi(\vec{u}) \vec{x}_1^\pi(\vec{v})} - \tilde{m}_{uv}(i) \right| + |\tilde{m}_{uv}(i) - m_{uv}(i)| > \nu \ \middle| \ v \in \mathcal{S}_u^{\beta_l, \beta_h}(i) \right)$$
$$\leq \mathbb{P}\left(\left| \mu_{\vec{x}_1^\pi(\vec{u}) \vec{x}_1^\pi(\vec{v})} - \tilde{m}_{uv}(i) \right| > \frac{\nu}{2} \ \middle| \ v \in \mathcal{S}_u^{\beta_l, \beta_h}(i) \right)$$
$$+ \mathbb{P}\left(|\tilde{m}_{uv}(i) - m_{uv}(i)| > \frac{\nu}{2} \ \middle| \ v \in \mathcal{S}_u^{\beta_l, \beta_h}(i) \right)$$

To bound the first term, we use McDiarmid's inequality (Theorem A.4) to show that $\tilde{m}_{uv}(i)$ concentrates around $\mu_{\vec{x}_1^\pi(\vec{u}) \vec{x}_1^\pi(\vec{v})}$ due to the sampling of the latent variables and independent noise terms. Conditioned on $|\mathcal{O}_i^{uv}|$, we bound the second term using the Kontorovich-Ramanan inequality (Theorem A.5) to show that $m_{uv}(i)$ concentrates around $\tilde{m}_{uv}(i)$ since the observed datapoints are sampled without replacement from a finite set of columns.

To apply McDiarmid's inequality for bounding $\left| \mu_{\vec{x}_1^\pi(\vec{u}) \vec{x}_1^\pi(\vec{v})} - \tilde{m}_{uv}(i) \right|$, we define a function ξ_{ab} which maps from the latent variables and noise terms to $\tilde{m}_{uv}(i)$ for $\vec{x}_1^\pi(\vec{u}) = a$ and $\vec{x}_1^\pi(\vec{v}) = b$. We show that this function satisfies the bounded difference condition. The entries in row u and v of matrix Z are fully determined by the row latent variables $\vec{x}_1^\pi(\vec{u})$ and $\vec{x}_1^\pi(\vec{v})$, the column latent variables $\{\vec{x}_2^\pi(\vec{j})\}_{j \in \mathcal{N}_i}$, and the independent noise terms $\{\eta(u,j), \eta(v,j)\}_{j \in \mathcal{N}_i}$. The n' column latent variables are fully determined by the $\sum_{q \in \mathcal{I}_2} (n_q - 1)$ latent variables for corresponding coordinates in the tensor. Recall that the number of relevant columns in the flattened matrix is $n' = \prod_{q \in \mathcal{I}_2} (n_q - 1)$, such that in fact the number of free parameters is significantly smaller than the number of total columns. Each column latent variable $\vec{x}_2^\pi(\vec{j})$ is associated to a vector of t_2 latent variables in the original tensor,

$$\vec{x}_2^\pi(\vec{j}) = \left(x_{\pi(t_1+1)}(j_1), \ldots, x_{\pi(t_1+t_2)}(j_{t_2}) \right).$$

Therefore, the set of column latent variables $\{\vec{x}_2^\pi(\vec{j})\}_{j \in \mathcal{N}_i}$ is fully determined by the corresponding latent variables in the t_2 dimensions of the original tensor, described by the sets $\{x_{\pi(t_1+1)}(j)\}_{j \in [n_{\pi(t_1+1)}-1]}, \ldots, \{x_{\pi(t_1+t_2)}(j)\}_{j \in [n_{\pi(t_1+t_2)}-1]}$.

For a pair of row latent variables $a, b \in \mathcal{X}_1^\pi$, we construct function ξ_{ab} to map from the column latent variables and individual noise terms to $\tilde{m}_{uv}(i)$ conditioned on $\vec{x}_1^\pi(\vec{u}) = a, \vec{x}_1^\pi(\vec{v}) = b$,

$$\xi_{ab} : \times_{q \in \mathcal{I}_2}(\mathcal{X}_q)^{n_q-1} \times \mathbb{R}^{2n'} \to \mathbb{R}.$$

The function ξ_{ab} is defined according to

(8.10)
$$\xi_{ab}\Big(\{x_q(l)\}_{q \in \mathcal{I}_2, l \in [n_q-1]}, \{\eta(u,j), \eta(v,j)\}_{j \in \mathcal{N}_i}\Big)$$
$$= \frac{1}{n'} \sum_{j \in \mathcal{N}_i} \Big(f\Big(a, \vec{x}_2^\pi(\vec{j})\Big) + \eta(u,j) - f\Big(b, \vec{x}_2^\pi(\vec{j})\Big) - \eta(v,j)\Big)$$

where $\vec{x}_2^\pi(\vec{j}) = \big(x_{2,1}^\pi(j_1), \ldots, x_{2,t_2}^\pi(j_{t_2})\big)$. We will overload the notation and also let ξ_{ab} denote the random variable which is evaluated with respect to the column latent variables and noise terms, leaving out the terms in the parentheses for readability. We can verify that by construction,

$$\xi_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} = \tilde{m}_{uv}(i).$$

Since the noise terms are zero-mean and $\vec{x}_2^\pi(\vec{j})$ are identically distributed for all $j \in [n]$,
$$\mathbb{E}\xi_{ab} = \mathbb{E}_{\mathbf{x}_2^\pi}\left[f(a, \mathbf{x}_2^\pi) - f(b, \mathbf{x}_2^\pi)\right] = \mu_{ab}.$$

For any tensor dimension $q \in \mathcal{I}_2$ and for any coordinate $l \in [n_q - 1]$, the associated latent variable $x_q(l)$ shows up in exactly $\frac{n'}{n_q-1}$ terms of the summation in (8.10). We assumed in our model that the function f is L-Lipschitz such that by for any $a, b \in \mathcal{X}_1^\pi$ and $x_2 \in \mathcal{X}_2^\pi$,

$$|f(a, x_2) - f(b, x_2)| \le LD_{\mathcal{X}_1^\pi},$$

where we recall that $D_{\mathcal{X}_1^\pi}$ is the diameter of $\mathcal{X}_1^\pi$. Since the above expression only takes values within $[-LD_{\mathcal{X}_1^\pi}, LD_{\mathcal{X}_1^\pi}]$, this implies that if we changed a single variable $x_q(l)$ arbitrarily, for each of the $\frac{n'}{n_q-1}$ terms which it participates in, the value will at most change by $2LD_{\mathcal{X}_1^\pi}$, which is then divided by n' from the term in front of the summation. Thus, the overall value of ξ_{ab} can change at most by $\frac{2LD_{\mathcal{X}_1^\pi}}{n_q-1}$ when a single latent variable is changed arbitrarily.

For any $j \in \mathcal{N}_i$, the noise variables $\eta(u,j)$ and $\eta(v,j)$ each only show up in a single term of the summation in (8.10). Furthermore, we assumed in our model

statement that the noise is bounded, i.e. $\eta(u, j) \in [-B_e, B_e]$. Thus, the overall value of ξ_{ab} can change at most by $\frac{2B_e}{n'}$ when a single additive noise variable is changed arbitrarily.

By McDiarmid's inequality, we can conclude that for any $\nu > 0$ and $u, v \in [m]$,

$$\mathbb{P}\left(\left|\tilde{m}_{uv}(i) - \mu_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}\right| \geq \tfrac{\nu}{2} \mid \vec{x}_1^\pi(\vec{u}) = a, \vec{x}_1^\pi(\vec{v}) = b\right)$$
$$= \mathbb{P}\left(|\xi_{ab} - \mathbb{E}\xi_{ab}| \geq \tfrac{\nu}{2}\right)$$
$$\leq 2 \exp\left(\frac{-\nu^2}{8(LD_{\mathcal{X}_1^\pi})^2 \sum_{q \in \mathcal{I}_2} \frac{1}{n_q - 1} + \frac{16B_e^2}{n'}}\right).$$

Observe that $v \in \mathcal{S}_u^{\beta_l, \beta_h}(i)$ only depends on the locations of observed samples, specified by the set of data indices $\mathcal{D}$, whereas $\tilde{m}_{uv}(i)$ and $\mu_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}$ are completely independent from $\mathcal{D}$. Additionally since the upper bound does not depend on a and b, it follows that

$$\mathbb{P}\left(\left|\tilde{m}_{uv}(i) - \mu_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}\right| \geq \tfrac{\nu}{2} \mid v \in \mathcal{S}_u^{\beta_l, \beta_h}(i)\right) \leq 2 \exp\left(\frac{-\nu^2}{8(LD_{\mathcal{X}_1^\pi})^2 \sum_{q \in \mathcal{I}_2} \frac{1}{n_q - 1} + \frac{16B_e^2}{n'}}\right).$$

Next, we bound $|\tilde{m}_{uv}(i) - m_{uv}(i)|$ using the Kontorovich-Ramanan inequality. Recall that the set $\mathcal{O}_i^{uv}$ denotes the columns in $\mathcal{N}_i$ for which samples have been observed from both rows u and v. Suppose that $|\mathcal{O}_i^{uv}| = k$, such that $\mathcal{O}_i^{uv}$ can be viewed as k randomly chosen columns out of n' total columns. This is equivalent to drawing k elements uniformly at random from a finite population of n' without replacement. We will define a function ψ_{uvk} which maps from the k sampled columns to the quantity $m_{uv}(i)$,

$$\psi_{uvk} : \mathcal{N}_i^k \to \mathbb{R}.$$

Let $J_1 \ldots J_k \in \mathcal{N}_i$ denote the randomly selected column indices, then

$$\psi_{uvk}(\{J_1, \ldots, J_k\}) = \frac{1}{k} \sum_{l=1}^{k} (Z(u, J_l) - Z(v, J_l)).$$

We can verify that by construction, conditioned on $|\mathcal{O}_i^{uv}| = k$,

$$\psi_{uvk}(\mathcal{O}_i^{uv}) = m_{uv}(i) \text{ and } \mathbb{E}[\psi_{uvk}(\mathcal{O}_i^{uv})] = \tilde{m}_{uv}(i).$$

For any two sequences $\{J_l\}, \{J'\} \in \mathcal{N}_i^k$, the function values differ by

$$
\psi_{uvk}(\{J_1, \ldots, J_k\}) - \psi_{uvk}(\{J'_1, \ldots, J'_k\})
$$

$$
= \frac{1}{k} \sum_{l=1}^{k} (Z(u, J_l) - Z(v, J_l) - Z(u, J'_l) + Z(v, J'_l))
$$

$$
= \frac{1}{k} \sum_{l:J_l \neq J'_l} (Z(u, J_l) - Z(v, J_l) - Z(u, J'_l) + Z(v, J'_l)).
$$

Recall that by definition of B_0 in (8.5), for any $u, v \in [m]$ and $i \in [n]$,

$$
|Z(u, i) - Z(v, i)| \leq B_0,
$$

since the latent space and noise terms are bounded. Therefore, if we consider the Hamming metric on $\mathcal{N}_i^k$, we can verify that $\psi_{uvk}(\cdot)$ is $(\frac{2B_0}{k})$-Lipschitz, since

$$
\psi_{uvk}(\{J_1, \ldots, J_k\}) - \psi_{uvk}(\{J'_1, \ldots, J'_k\}) \leq \frac{2B_0}{k} |\{l : J_l \neq J'_l\}|.
$$

Next we compute the mixing coefficient of the randomly chosen columns $\mathcal{O}_i^{uv} = \{J_1, \ldots J_k\}$, as defined in the setup of the Kontorovich-Ramanan inequality (see Theorem A.5 in Appendix A for details). The sequence of columns $(J_1, \ldots, J_k)$ is uniformly distributed amongst all $\frac{n'!}{(n'-k)!}$ possible sequences of choosing k columns out of n' total columns without replacement. The mixing coefficient Δ_k is defined according to $\Delta_k := \max_{s \in [k]} \left(1 + \sum_{s'=s+1}^{k} \bar{\lambda}_{ss'}\right)$, where

$$
\bar{\lambda}_{ss'} := \sup_{\substack{\vec{\tau} \in \mathcal{N}_i^{s-1}, \\ w, \hat{w} \in \mathcal{N}_i}} \frac{1}{2} \sum_{\vec{\tau}' \in \mathcal{N}_i^{k-s'+1}} \left| \mathbb{P}\left((J_{s'} \ldots J_k) = \vec{\tau}' \mid (J_1 \ldots J_s) = (\vec{\tau}, w)\right) \right.
$$

$$
\left. - \mathbb{P}\left((J_{s'} \ldots J_k) = \vec{\tau}' \mid (J_1 \ldots J_s) = (\vec{\tau}, \hat{w})\right) \right|,
$$

where the supremum is only taken over valid sequences $(\vec{\tau}, w)$ and $(\vec{\tau}, \hat{w})$ which do no repeat any elements.

Conditioned on $(J_1 \ldots J_s) = (\vec{\tau}, w)$, the sequence $(J_{s'} \ldots J_k)$ is equally likely amongst all sequences of length $(k - s' + 1)$ chosen from the n' columns without replacement as long as they also do not repeat any elements already chosen in the vector $(\vec{\tau}, w)$, i.e. all permutations of length $(k - s' + 1)$ out of $n' - s$ remaining columns. Similarly, conditioned on $(J_1 \ldots J_s) = (\vec{\tau}, \hat{w})$, the sequence $(J_{s'} \ldots J_k)$ is equally likely amongst all permutations of length $(k - s' + 1)$ out of the $n' - s$ remaining columns that do not contain any elements in $(\vec{\tau}, \hat{w})$. There are a total of $\frac{(n'-s)!}{((n'-s)-(k-s'+1))!}$ such valid permutations.

Therefore, if τ' does not contain symbols w or $\hat{w}$, then the probability of $(J_{s'} \ldots J_k)$ $= \vec{\tau}'$ is equal whether conditioned on $(J_1 \ldots J_s) = (\vec{\tau}, w)$ or $(J_1 \ldots J_s) = (\vec{\tau}, \hat{w})$. If τ' contains both w and $\hat{w}$, then it will have probability zero in both conditioned events. If τ' contains only one out of either w or $\hat{w}$, then the absolute value of the difference in the conditional probabilities of $(J_{s'} \ldots J_k) = \vec{\tau}'$ will be equal to $\frac{((n'-s)-(k-s'+1))!}{(n'-s)!}$. The total number of such sequences τ' which contain only one out of either w or $\hat{w}$ (and do not repeat elements in τ) is equal to $2(k - s' + 1)\frac{(n'-s-1)!}{((n'-s-1)-(k-s'))!} = \frac{2(k-s'+1)}{n'-s}\frac{(n'-s)!}{((n'-s)-(k-s'+1))!}$. Therefore,

$$\bar{\lambda}_{ss'} = \frac{(k - s' + 1)}{n' - s} \frac{(n' - s)!}{((n' - s) - (k - s' + 1))!} \frac{((n' - s) - (k - s' + 1))!}{(n' - s)!}$$

$$= \frac{(k - s' + 1)}{n' - s}.$$

We can then compute the mixing coefficient Δ_k,

$$(8.11) \qquad \Delta_k = \max_{s \in [k]} \left(1 + \sum_{s'=s+1}^{k} \frac{(k - s' + 1)}{n' - s} \right)$$

$$(8.12) \qquad = \max_{s \in [k]} \left(1 + \frac{(k - s)(k - s + 1)}{2(n' - s)} \right)$$

$$(8.13) \qquad = 1 + \frac{(k - 1)k}{2(n' - 1)}.$$

The last step follows from showing that the derivative with respect to s is negative, such that the expression is maximized for the smallest value of s, which is $s = 1$.

Consequently, it follows from the Kontorovich-Ramanan theorem (Theorem A.5) that for any $\nu > 0$,

$$\mathbb{P}\left(|m_{uv}(i) - \tilde{m}_{uv}(i)| \geq \tfrac{\nu}{2} \mid |\mathcal{O}_i^{uv}| = k\right) = \mathbb{P}\left(|\psi_{uvk} - \mathbb{E}\psi_{uvk}| \geq \tfrac{\nu}{2}\right)$$

$$\leq 2\exp\left(\frac{-\nu^2}{8k\left(\frac{2B_0}{k}\right)^2 \left(1 + \frac{(k-1)k}{2(n'-1)}\right)^2} \right)$$

$$= 2\exp\left(\frac{-k\nu^2}{32B_0^2 \left(1 + \frac{(k-1)k}{2(n'-1)}\right)^2} \right)$$

$$\leq 2\exp\left(\frac{-\nu^2}{32B_0^2} \frac{n'^2 k}{(n' + k^2)^2} \right).$$

The last inequality follows when we assume that $n' \geq 2$. Then $2('n - 1) \geq n'$ and $1 + \frac{(k-1)k}{2(n'-1)} \leq 1 + \frac{k^2}{n'}$, hence, $\frac{k}{\left(1 + \frac{(k-1)k}{2(n'-1)}\right)^2} \geq \frac{k}{\left(1 + \frac{k^2}{n'}\right)^2} = \frac{n'^2 k}{(n'+k^2)^2}$.

We would like to compute a lower bound that holds for all k. By taking the derivative with respect to k,

$$\frac{\partial}{\partial k} \frac{n'^2 k}{(n' + k^2)^2} = \frac{n'^2(n' + k^2)(n' - 3k^2)}{(n' + k^2)^4}.$$

Therefore, $\frac{n'^2 k}{(n'+k^2)^2}$ is monotone increasing for $k \leq \sqrt{\frac{n'}{3}}$, and monotone decreasing for $k > \sqrt{\frac{n'}{3}}$. In other words, for any $k \in [\beta_l, \beta_h]$,

$$\frac{n'^2 k}{(n' + k^2)^2} \geq \min\left\{\frac{n'^2 \beta_l}{\left(n' + \beta_l^2\right)^2}, \frac{n'^2 \beta_h}{\left(n' + \beta_h^2\right)^2}\right\} =: \Delta.$$

Therefore,

$$\mathbb{P}\left(|m_{uv}(i) - \tilde{m}_{uv}(i)| \geq \tfrac{\nu}{2} \;\middle|\; v \in \mathcal{S}_u^{\beta_l, \beta_h}(i)\right)$$
$$= \mathbb{P}\left(|m_{uv}(i) - \tilde{m}_{uv}(i)| \geq \tfrac{\nu}{2} \mid |\mathcal{O}_i^{uv}| \in [\beta_l, \beta_h]\right)$$
$$\leq 2\exp\left(\frac{-\nu^2}{32 B_0^2}\Delta\right).$$

Finally, we combine the two bounds to obtain the final result.

$\square$

PROOF OF LEMMA 8.3. Without loss of generality, to simplify the notation, assume that $i = (n_{\pi(t_1+1)}, \ldots n_{\pi(t)})$. Therefore

$$\mathcal{N}_i = \{j \in [n] \text{ s.t. } j_k \neq i_k \text{ for all } k \in [t_2]\}$$
$$= \{j \in [n] \text{ s.t. } j_k \in [n_{\pi(t_1+k)} - 1] \text{ for all } k \in [t_2]\}$$
$$= \{j \in [n] \text{ s.t. } \vec{j} \in \times_{k \in [t_2]}[n_{\pi(t_1+k)} - 1]\}.$$

The proof is equivalent for all choice of $i \in [n]$ by symmetry. Recall that $|\mathcal{N}_i| = n' = \prod_{q \in \mathcal{I}_2}(n_q - 1)$.

In this proof, we will show concentration of $s_{uv}^2(i)$ to $\mathbb{E}\tilde{s}_{uv}^2(i)$, and then show that $\mathbb{E}\tilde{s}_{uv}^2(i)$ conditioned on $\vec{x}_1^\pi(\vec{u})$ and $\vec{x}_1^\pi(\vec{v})$ is approximately equal to $\sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}^2 + 2\gamma^2$. By the triangle inequality,

$$\left|\mathbb{E}\tilde{s}_{uv}^2(i) - s_{uv}^2(i)\right| \leq \left|\mathbb{E}\tilde{s}_{uv}^2(i) - \tilde{s}_{uv}^2(i)\right| + \left|\tilde{s}_{uv}^2(i) - s_{uv}^2(i)\right|.$$

In a similar vein as in the proof of the Lemma 8.2, $\left|\mathbb{E}\tilde{s}_{uv}^2(i) - s_{uv}^2(i)\right| > \tau$ implies that either $\left|\mathbb{E}\tilde{s}_{uv}^2(i) - \tilde{s}_{uv}^2(i)\right| > \tau/2$ or $\left|\tilde{s}_{uv}^2(i) - s_{uv}^2(i)\right| > \tau/2$. Again, a simple

application of union bound results in

$$\mathbb{P}\left(\left|\mathbb{E}\tilde{s}_{uv}^2(i) - s_{uv}^2(i)\right| > \tau \;\Big|\; v \in \mathcal{S}_u^{\beta_l,\beta_h}(i)\right)$$

$$\leq \mathbb{P}\left(\left|\mathbb{E}\tilde{s}_{uv}^2(i) - \tilde{s}_{uv}^2(i)\right| > \frac{\tau}{2}\;\Big|\; v \in \mathcal{S}_u^{\beta_l,\beta_h}(i)\right) + \mathbb{P}\left(\left|\tilde{s}_{uv}^2(i) - s_{uv}^2(i)\right| > \frac{\tau}{2}\;\Big|\; v \in \mathcal{S}_u^{\beta_l,\beta_h}(i)\right)$$

To apply McDiarmid's inequality for bounding $\left|\mathbb{E}\tilde{s}_{uv}^2(i) - \tilde{s}_{uv}^2(i)\right|$, we define a function ξ_{ab} which maps from the latent variables and noise terms to $\tilde{s}_{uv}^2(i)$ for $\vec{x}_1^\pi(\vec{u}) = a$ and $\vec{x}_1^\pi(\vec{v}) = b$. We show that this function satisfies the bounded difference condition. The entries in row u and v of matrix Z are fully determined by the row latent variables $\vec{x}_1^\pi(\vec{u})$ and $\vec{x}_1^\pi(\vec{v})$, the column latent variables $\{\vec{x}_2^\pi(\vec{j})\}_{j \in \mathcal{N}_i}$, and the independent noise terms $\{\eta(u,j), \eta(v,j)\}_{j \in \mathcal{N}_i}$.

For a pair of row latent variables $a, b \in \mathcal{X}_1^\pi$, we construct function ξ_{ab} to map from the column latent variables and individual noise terms to $\tilde{s}_{uv}^2(i)$ conditioned on $\vec{x}_1^\pi(\vec{u}) = a, \vec{x}_1^\pi(\vec{v}) = b$,

$$\xi_{ab} : \times_{q \in \mathcal{I}_2}(\mathcal{X}_q)^{n_q-1} \times \mathbb{R}^{2n'} \to \mathbb{R}.$$

We define f_{ab} and η_{uv} according to

$$f_{ab}(x) = f(a,x) - f(b,x),$$
$$\eta_{uv}(j) = \eta(u,j) - \eta(v,j).$$

The function ξ_{ab} is defined according to

$$\xi_{ab}\Big(\{x_q(l)\}_{q \in \mathcal{I}_2, l \in [n_q-1]}, \{\eta(u,j), \eta(v,j)\}_{j \in \mathcal{N}_i}\Big)$$

(8.14)

$$= \frac{1}{2n'(n'-1)} \sum_{j,h \in \mathcal{N}_i \times \mathcal{N}_i} \left(f_{ab}\left(\vec{x}_2^\pi(\vec{j})\right) + \eta_{uv}(j) - f_{ab}\left(\vec{x}_2^\pi(\vec{h})\right) - \eta_{uv}(h)\right)^2$$

where $\vec{x}_2^\pi(\vec{j}) = \left(x_{2,1}^\pi(j_1), \ldots, x_{2,t_2}^\pi(j_{t_2})\right)$. We will overload the notation and also let ξ_{ab} denote the random variable which is evaluated with respect to the column latent variables and noise terms, leaving out the terms in the parentheses for readability. We can verify that by construction,

$$\xi_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} = \tilde{s}_{uv}^2(i) \text{ and } \mathbb{E}\xi_{ab} = \mathbb{E}\left[\tilde{s}_{uv}^2(i)\big|\, \vec{x}_1^\pi(\vec{u}) = a, \vec{x}_1^\pi(\vec{v}) = b\right].$$

For any tensor dimension $q \in \mathcal{I}_2$ and for any coordinate $l \in [n_q - 1]$, the associated latent variable $x_q(l)$ shows up in at most $\frac{2n'(n'-1)}{n_q-1}$ terms out of $n'(n'-1)$ of the summation in (8.14), used to evaluate ξ_{ab}. We assumed in our model that the function f is L-Lipschitz such that by for any $a, b \in \mathcal{X}_1^\pi$ and $x_2 \in \mathcal{X}_2^\pi$,

$$|f(a, x_2) - f(b, x_2)| \leq LD_{\mathcal{X}_1^\pi},$$

where we recall that $D_{\mathcal{X}_1^\pi}$ is the diameter of $\mathcal{X}_1^\pi$. Therefore,

$$\left| f_{ab}\left(\vec{x}_2^\pi(\vec{j})\right) - f_{ab}\left(\vec{x}_2^\pi(\vec{j'})\right) \right| \le 2LD_{\mathcal{X}_1^\pi}.$$

We assumed that the noise was bounded such that $|\eta_{uv}(j)| \le 2B_e$. Therefore,

$$\left| f_{ab}\left(\vec{x}_2^\pi(\vec{j})\right) + \eta_{uv}(j) - f_{ab}\left(\vec{x}_2^\pi(\vec{j'})\right) - \eta_{uv}(j') \right| \le 2LD_{\mathcal{X}_1^\pi} + 4B_e.$$

If we changed a single variable $x_q(l)$ arbitrarily, for each of the $\frac{2n'(n'-1)}{n_q-1}$ terms which it participates in, the value of the summand

$$\left[f_{ab}\left(\vec{x}_2^\pi(\vec{j})\right) + \eta_{uv}(j) - f_{ab}\left(\vec{x}_2^\pi(\vec{j'})\right) - \eta_{uv}(j') \right]^2$$

will at most change by $2\left(2LD_{\mathcal{X}_1^\pi} + 4B_e\right)\left(2LD_{\mathcal{X}_1^\pi}\right) + \left(2LD_{\mathcal{X}_1^\pi}\right)^2$ because if we consider a perturbation Δ, $(X + \Delta)^2 - X^2 = 2X\Delta + \Delta^2$, which is then divided by $2n'(n' - 1)$ from the term in front of the summation. Thus, the overall value of ξ_{ab} can change at most by

$$\frac{1}{n_q - 1}\left(2\left(2LD_{\mathcal{X}_1^\pi} + 4B_e\right)\left(2LD_{\mathcal{X}_1^\pi}\right) + \left(2LD_{\mathcal{X}_1^\pi}\right)^2\right) = \frac{4LD_{\mathcal{X}_1^\pi}}{n_q - 1}\left(3LD_{\mathcal{X}_1^\pi} + 4B_e\right)$$

when a single latent variable is changed arbitrarily.

For any $j \in \mathcal{N}_i$, the noise variables $\eta(u, j)$ and $\eta(v, j)$ each shows up in exactly $2(n' - 1)$ terms of the summation in (8.14). Furthermore, we assumed in our model statement that the noise is bounded, i.e. $\eta(u, j) \in [-B_e, B_e]$. Thus, by a similar argument as above, the overall value of ξ_{ab} can change at most by

$$\frac{1}{n'}\left[2\left(2LD_{\mathcal{X}_1^\pi} + 4B_e\right)(2B_e) + (2B_e)^2\right] = \frac{4B_e}{n'}\left(2LD_{\mathcal{X}_1^\pi} + 5B_e\right)$$

when a single additive noise variable is changed arbitrarily.

By McDiarmid's inequality, we can conclude that for any $\tau > 0$ and $u, v \in [m]$,

$$\mathbb{P}\left(\left| \tilde{s}_{uv}^2(i) - \mathbb{E}\tilde{s}_{uv}^2(i) \right| \ge \tfrac{\tau}{2} \mid \vec{x}_1^\pi(\vec{u}) = a, \vec{x}_1^\pi(\vec{v}) = b \right)$$
$$= \mathbb{P}\left(|\xi_{ab} - \mathbb{E}\xi_{ab}| \ge \tfrac{\tau}{2} \right)$$
$$\le 2\exp\left(\frac{-\tau^2}{32L^2 D_{\mathcal{X}_1^\pi}^2 \left(3LD_{\mathcal{X}_1^\pi} + 4B_e\right)^2 \sum_{q=1}^{t_2} \frac{1}{n_q - 1} + \frac{64B_e^2\left(2LD_{\mathcal{X}_1^\pi} + 5B_e\right)^2}{n'}} \right).$$

Observe that $v \in \mathcal{S}_u^{\beta_l, \beta_h}(i)$ only depends on the locations of observed samples, specified by the set of data indices $\mathcal{D}$, whereas $\tilde{s}_{uv}$ and $\sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}$ are completely

independent from $\mathcal{D}$. Additionally since the upper bound does not depend on a and b, it follows that

$$\mathbb{P}\left(\left|\tilde{s}_{uv}^2(i) - \mathbb{E}\tilde{s}_{uv}^2(i)\right| \geq \tfrac{\tau}{2} \mid v \in \mathcal{S}_u^{\beta_l,\beta_h}(i)\right)$$

$$\leq 2\exp\left(\frac{-\tau^2}{32L^2 D_{\mathcal{X}_1^\pi}^2 \left(3LD_{\mathcal{X}_1^\pi} + 4B_e\right)^2 \sum_{q=1}^{t_2} \frac{1}{n_q - 1} + \frac{64B_e^2\left(2LD_{\mathcal{X}_1^\pi} + 5B_e\right)^2}{n'}}\right).$$

Next, we bound $\left|\tilde{s}_{uv}^2(i) - s_{uv}^2(i)\right|$ using the Kontorovich-Ramanan inequality. Recall that the set $\mathcal{O}_i^{uv}$ denotes the columns for which samples have been observed from both rows u and v. Suppose that $|\mathcal{O}_i^{uv}| = k$, such that $\mathcal{O}_i^{uv}$ can be viewed as k randomly chosen columns out of n' total columns. This is equivalent to drawing k elements uniformly at random from a finite population of n' without replacement. We will define a function $\psi_{uvk} : \mathcal{N}_i^k \to \mathbb{R}$, which maps from the k sampled columns to the quantity $s_{uv}^2(i)$. Let $J_1 \ldots J_k \in \mathcal{N}_i$ denote the selected column indices, then

$$\begin{aligned}&\psi_{uvk}(\{J_1, \ldots, J_k\})\\ (8.15) \quad &= \frac{1}{2k(k-1)} \sum_{l=1}^{k} \sum_{l'=1}^{k} \left(Z(u, J_l) - Z(v, J_l) - Z(u, J_{l'}) + Z(v, J_{l'})\right)^2.\end{aligned}$$

We can verify that by construction, conditioned on $|\mathcal{O}_i^{uv}| = k$, $\psi_{uvk}(\mathcal{O}_i^{uv}) = s_{uv}^2(i)$. Next we show that its expectation conditioned on the matrix Z equals $\tilde{s}_{uv}^2(i)$. Since we condition on the matrix Z, the randomness only comes from the randomly sampled columns $\mathcal{O}_i^{uv} = \{J_l\}_{l \in [k]}$. Observe that the randomly chosen column indices J_l and $J_{l'}$ are identically distributed, and thus $Z(u, J_l) - Z(v, J_l)$ and $Z(u, J_{l'}) - Z(v, J_{l'})$ are also identically distributed. Recall from (8.16) that if two variables are identically distributed, then the expectation of the square of their difference is equal to two times the variance minus the covariance. Therefore if we let $X_l := Z(u, J_l) - Z(v, J_l)$,

$$\mathbb{E}\left[\psi_{uvk}(\mathcal{O}_i^{uv})\right] = \frac{1}{k(k-1)} \sum_{l=1}^{k} \sum_{l'=1}^{k} \left(\text{Var}[X_l] - \text{Cov}[X_l, X_{l'}]\right)$$

$$= \text{Var}[X_l] - \frac{1}{k(k-1)} \sum_{l=1}^{k} \sum_{l' \neq l} \text{Cov}[X_l, X_{l'}].$$

Conditioned on Z, for $l \neq l'$, the random variables X_l and $X_{l'}$ are determined by the sampled columns J_l and $J_{l'}$, which can be viewed as two samples drawn from

a finite population of size n' without replacement. Therefore, their covariance is equal to $-\frac{1}{n'-1}$ times the variance of the population. Therefore,

$$\mathbb{E}\left[\psi_{uvk}(\mathcal{O}_i^{uv})\right] = \text{Var}[X_l] + \frac{1}{k(k-1)(n'-1)} \sum_{l=1}^{k} \sum_{l' \neq l} \text{Var}[X_l] = \frac{n'}{(n'-1)} \text{Var}[X_l].$$

By construction, conditioned on Z, the variance of X_l with respect to the randomly chosen column index J_l is equal to $\frac{n'-1}{n'} \tilde{s}_{uv}^2(i)$, and the mean of X_l with respect to the randomly chosen column index J_l is equal to $\tilde{m}_{uv}$. Therefore $\mathbb{E}\left[\psi_{uvk}(\mathcal{O}_i^{uv})\right] = \tilde{s}_{uv}^2(i)$.

Next we show that the function ψ_{uvk} is Lipschitz with respect to the Hamming metric on $\mathcal{N}_i^k$, and we compute the associated Lipschitz constant. The definition of ψ_{uvk} in (8.15) involves a double summation, and a particular index variable J_l shows up in exactly $2(k-1)$ of these terms in the summation. Consider two sets of column indices $\{J_l \in \mathcal{N}_i\}_{l \in [k]}$ and $\{J_l' \in \mathcal{N}_i\}_{l \in [k]}$. For each position l_0 at which they differ, i.e. $J_{l_0} \neq J_{l_0}'$, we can bound the maximum resulting difference in the function value of ψ_{uvk}. By definition of B_0 in (8.5), for any $u, v \in [m]$ and $i \in [n]$, $|Z(u,i) - Z(v,i)| \leq B_0$, since the latent space and noise terms are bounded. For any $J_l, J_l' \in [n]$, it follows that

$$\left(Z(u, J_l) - Z(v, J_l) - Z(u, J_l') + Z(v, J_l')\right)^2 \leq 4B_0^2.$$

The difference in the value of ψ_{uvl} evaluated with column indices $\{J_l\}_{l \in [k]}$ versus $\{J_l'\}_{l \in [k]}$ is bounded above by

$$\psi_{uvk}(\{J_l\}_{l \in [k]}) - \psi_{uvk}(\{J_l'\}_{l \in [k]}) \leq \frac{1}{2k(k-1)} 2(k-1) |\{l : J_l \neq J_l'\}| (4B_0^2)$$
$$= \frac{4B_0^2}{k} |\{l : J_l \neq J_l'\}|.$$

Therefore, if we consider the Hamming metric on $\mathcal{N}_i^k$, the function $\psi_{uvk}(\cdot)$ is $\left(\frac{4B_0^2}{k}\right)$-Lipschitz.

Next we compute the mixing coefficient of the randomly chosen columns $\mathcal{O}_i^{uv} = \{J_1, \ldots J_k\}$, as defined in the setup of the Kontorovich-Ramanan inequality (see Theorem A.5 in Appendix A for details). This is exactly the same with the argument in the proof of Lemma 8.2. We just restate the mixing coefficient Δ_k obtained in (8.12):

$$\Delta_k = 1 + \frac{(k-1)k}{2(n'-1)}.$$

Consequently, it follows from the Kontorovich-Ramanan theorem (Theorem A.5) that for any $\nu > 0$,

$$\mathbb{P}\left(\left|s_{uv}^2(i) - \tilde{s}_{uv}^2(i)\right| \geq \tfrac{\tau}{2} \mid |\mathcal{O}_i^{uv}| = k\right) = \mathbb{P}\left(\left|\psi_{uvk} - \mathbb{E}\psi_{uvk}\right| \geq \tfrac{\tau}{2}\right)$$

$$\leq 2\exp\left(\frac{-\tau^2}{8k\left(\frac{4B_0^2}{k}\right)^2\left(1 + \frac{(k-1)k}{2(n'-1)}\right)^2}\right)$$

$$= 2\exp\left(\frac{-k\tau^2}{128B_0^4\left(1 + \frac{(k-1)k}{2(n'-1)}\right)^2}\right)$$

$$\leq 2\exp\left(\frac{-\tau^2}{128B_0^4}\frac{n'^2 k}{(n' + k^2)^2}\right).$$

The last inequality follows when we assume that $n' \geq 2$. Then $2(n' - 1) \geq n'$ and $1 + \frac{(k-1)k}{2(n'-1)} \leq 1 + \frac{k^2}{n'}$, hence, $\frac{k}{\left(1 + \frac{(k-1)k}{2(n'-1)}\right)^2} \geq \frac{k}{\left(1 + \frac{k^2}{n'}\right)^2} = \frac{n'^2 k}{(n'+k^2)^2}$.

We would like to compute a lower bound that holds for all k. By taking the derivative with respect to k,

$$\frac{\partial}{\partial k}\frac{n'^2 k}{(n' + k^2)^2} = \frac{n'^2(n' + k^2)(n' - 3k^2)}{(n' + k^2)^4}.$$

Therefore, $\frac{n'^2 k}{(n'+k^2)^2}$ is monotone increasing for $k \leq \sqrt{\frac{n'}{3}}$, and monotone decreasing for $k > \sqrt{\frac{n'}{3}}$. In other words, for any $k \in [\beta_l, \beta_h]$,

$$\frac{n'^2 k}{(n' + k^2)^2} \geq \min\left\{\frac{n'^2\beta_l}{\left(n' + \beta_l^2\right)^2}, \frac{n'^2\beta_h}{\left(n' + \beta_h^2\right)^2}\right\} =: \Delta.$$

Therefore,

$$\mathbb{P}\left(\left|s_{uv}^2(i) - \tilde{s}_{uv}^2(i)\right| \geq \tfrac{\tau}{2} \;\middle|\; v \in \mathcal{S}_u^{\beta_l,\beta_h}(i)\right) = \mathbb{P}\left(\left|s_{uv}^2(i) - \tilde{s}_{uv}^2(i)\right| \geq \tfrac{\tau}{2} \mid |\mathcal{O}_i^{uv}| \in [\beta_l, \beta_h]\right)$$

$$\leq 2\exp\left(\frac{-\tau^2}{128B_0^4}\Delta\right).$$

Combining the two bounds on $\left|s_{uv}^2(i) - \tilde{s}_{uv}^2(i)\right|$ and $\left|\tilde{s}_{uv}^2(i) - \mathbb{E}\tilde{s}_{uv}^2(i)\right|$ leads to a bound on the probability that $\left|s_{uv}^2(i) - \mathbb{E}\tilde{s}_{uv}^2(i)\right| \geq \tau$.

Lastly, we bound the difference between $\mathbb{E}\tilde{s}_{uv}^2(i)$ and $\sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}^2 + 2\gamma^2$. By definition, we can write $\tilde{s}_{uv}^2(i)$, conditioned on $\vec{x}_1^\pi(\vec{u}) = a$ and $\vec{x}_1^\pi(\vec{v}) = b$, according

to

$$\tilde{s}_{uv}^2(i) = \frac{1}{2n'(n'-1)} \sum_{j,h \in \mathcal{N}_i \times \mathcal{N}_i} (X_j - X_h)^2,$$

where we define the variables

$$X_j = Z(u,j) - Z(v,j) = f_{ab}\left(\vec{x}_2^\pi(\vec{j})\right) + \eta_{uv}(\vec{j})$$

for all $j \in [n]$. Since the noise terms are i.i.d, and since $\vec{x}_2^\pi(\vec{j})$ are identically distributed for all $j \in \mathcal{N}_i$, $\{X_j\}_{j \in \mathcal{N}_i}$ are also identically distributed. When random variables X_j and X_h are identically distributed,

$$
\begin{aligned}
\mathbb{E}\left[(X_j - X_h)^2\right] &= \mathbb{E}\left[X_j^2 - 2X_jX_h + X_h^2\right] \\
&= 2\mathbb{E}\left[X_j^2\right] - 2\mathbb{E}\left[X_jX_h\right] \\
&= 2\mathrm{Var}[X_j] - 2\mathrm{Cov}[X_j, X_h].
\end{aligned}
$$

(8.16)

If $h = j$, then $\mathrm{Cov}[X_j, X_h] = \mathrm{Var}[X_j]$. Therefore, by substitution and using the fact that $\{X_j\}_{j \in \mathcal{N}_i}$ are identically distributed, it follows that

$$
\begin{aligned}
&\mathbb{E}\left[\tilde{s}_{uv}^2(i)\,\middle|\, \vec{x}_1^\pi(\vec{u}) = a, \vec{x}_1^\pi(\vec{v}) = b\right] \\
&= \frac{1}{n'(n'-1)} \sum_{j \in \mathcal{N}_i} \sum_{h \neq j, h \in \mathcal{N}_i} (\mathrm{Var}[X_j] - \mathrm{Cov}[X_j, X_h]) \\
&= \mathrm{Var}[X_j] - \frac{1}{n'(n'-1)} \sum_{j \in \mathcal{N}_i} \sum_{h \neq j, h \in \mathcal{N}_i} \mathrm{Cov}[X_j, X_h].
\end{aligned}
$$

First we can show that because the noise terms are independent, and by definition of σ_{ab}^2, the variance is equal to

$$
\begin{aligned}
\mathrm{Var}[X_j] &= \mathrm{Var}[f_{ab}\left(\vec{x}_2^\pi(\vec{j})\right) + \eta_{uv}(\vec{j})] \\
&= \mathrm{Var}[f_{ab}\left(\vec{x}_2^\pi(\vec{j})\right)] + \mathrm{Var}[\eta_{uv}(\vec{j})] \\
&= \sigma_{ab}^2 + 2\gamma^2.
\end{aligned}
$$

Next we bound the contribution from the covariance terms. By definition,

$$
\begin{aligned}
\mathrm{Cov}(X_j, X_h) &= \mathrm{Cov}\left(f_{ab}\left(\vec{x}_2^\pi(\vec{j})\right) + \eta_{uv}(\vec{j}), f_{ab}\left(\vec{x}_2^\pi(\vec{h})\right) + \eta_{uv}(\vec{h})\right) \\
&= \mathrm{Cov}\left(f_{ab}\left(\vec{x}_2^\pi(\vec{j})\right), f_{ab}\left(\vec{x}_2^\pi(\vec{h})\right)\right) + \mathrm{Cov}\left(\eta_{uv}(\vec{j}), \eta_{uv}(\vec{h})\right).
\end{aligned}
$$

Since the noise terms are independent across entries, $\text{Cov}\left(\eta_{uv}(\vec{j}), \eta_{uv}(\vec{h})\right) = 0$ if $j \neq h$. If the Hamming distance between $\vec{j}$ and $\vec{h}$ is t_2, it means that the entry corresponding to j and h in the original tensor do not share any coordinates, and therefore their associated latent variables $\vec{x}_2^{\pi}(\vec{j})$ and $\vec{x}_2^{\pi}(\vec{h})$ are independent, such that $Cov\left(f_{ab}\left(\vec{x}_2^{\pi}(\vec{j})\right), f_{ab}\left(\vec{x}_2^{\pi}(\vec{h})\right)\right)$ is equal to zero. For each j, there can exist at most $\sum_{q \in \mathcal{I}_2} \frac{n'}{n_q - 1}$ indices $h \in \mathcal{N}_i$ (including j itself) for which $d_H(\vec{j}, \vec{h}) < t_2$, where $d_H(\vec{j}, \vec{h}) = |\{l : j_l \neq h_l\}|$ denotes the Hamming distance between $\vec{j}$ and $\vec{h}$. Furthermore, we know that the absolute value of the covariance of two identically distributed random variables is always less than the variance. We assume without loss of generality that $n_q \geq 2$, otherwise $\mathcal{N}_i$ would be an empty set. It follows that

$$\left| \frac{1}{n'(n'-1)} \sum_{j \in \mathcal{N}_i} \sum_{h \neq j, h \in \mathcal{N}_i} \text{Cov}(X_j, X_h) \right|$$

$$= \left| \frac{1}{n'(n'-1)} \sum_{j \in \mathcal{N}_i} \sum_{\substack{h : h \neq j, \\ d_H(\vec{j}, \vec{h}) < t_2}} \text{Cov}\left(f_{ab}\left(\vec{x}_2^{\pi}(\vec{j})\right), f_{ab}\left(\vec{x}_2^{\pi}(\vec{h})\right)\right) \right|$$

$$\leq \frac{1}{n'(n'-1)} \sum_{j \in \mathcal{N}_i} \sum_{\substack{h : h \neq j, \\ d_H(\vec{j}, \vec{h}) < t_2}} \text{Var}\left(f_{ab}\left(\vec{x}_2^{\pi}(\vec{j})\right)\right)$$

$$\leq \frac{\sigma_{ab}^2}{n'-1} \left(\sum_{q \in \mathcal{I}_2} \left(\frac{n'}{n_q - 1} - 1 \right) \right)$$

$$\leq \frac{\sigma_{ab}^2}{n'-1} \left(\sum_{q \in \mathcal{I}_2} \left(\frac{n'-1}{n_q - 1} \right) \right) \qquad \because n_q \geq 2, \forall q \in \mathcal{I}_2$$

$$= \sigma_{ab}^2 \left(\sum_{q \in \mathcal{I}_2} \frac{1}{n_q - 1} \right)$$

Therefore, it follows that

$$(1 - \theta)\sigma_{\vec{x}_1^{\pi}(\vec{u})\vec{x}_1^{\pi}(\vec{v})}^2 + 2\gamma^2 \leq \mathbb{E}\tilde{s}_{uv}^2(i) \leq (1 + \theta)\sigma_{\vec{x}_1^{\pi}(\vec{u})\vec{x}_1^{\pi}(\vec{v})}^2 + 2\gamma^2,$$

where $\theta = \sum_{q \in \mathcal{I}_2} \frac{1}{n_q - 1}$, which converges to 0 as $n_q \to \infty$ for all $q \in \mathcal{I}_2$.

$\square$

8.3. *Sufficiently Many Good Neighbors.*

PROOF OF LEMMA 8.4. According to the Bayes' rule, the conditional probability of interest can be written as

$$
\mathbb{P}\left(\left\{\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}\right\}^{(k)}_{v\in\mathcal{S}_u^{\beta_l,\beta_h}(i)} > \zeta \;\middle|\; |\mathcal{S}_u^{\beta_l,\beta_h}(i)| \in \left[\frac{1}{2}(m-1)p, \frac{3}{2}(m-1)p\right]\right)
$$

$$
= \frac{\mathbb{P}\left(\left\{\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}\right\}^{(k)}_{v\in\mathcal{S}_u^{\beta_l,\beta_h}(i)} > \zeta \;\bigcap\; |\mathcal{S}_u^{\beta_l,\beta_h}(i)| \in \left[\frac{1}{2}(m-1)p, \frac{3}{2}(m-1)p\right]\right)}{\mathbb{P}\left(|\mathcal{S}_u^{\beta_l,\beta_h}(i)| \in \left[\frac{1}{2}(m-1)p, \frac{3}{2}(m-1)p\right]\right)}
$$

$$
\le 2\mathbb{P}\left(\left\{\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}\right\}^{(k)}_{v\in\mathcal{S}_u^{\beta_l,\beta_h}(i)} > \zeta\right),
$$

assuming that $\mathbb{P}\left(|\mathcal{S}_u^{\beta_l,\beta_h}(i)| \in \left[\frac{1}{2}(m-1)p, \frac{3}{2}(m-1)p\right]\right) \ge \frac{1}{2}$, which is true when m and n are sufficiently large from Lemma 8.1.

We can observe that this probability can be represented equivalently as

$$
\mathbb{P}\left(\left\{\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}\right\}^{(k)}_{v\in\mathcal{S}_u^{\beta_l,\beta_h}(i)} > \zeta\right) = \mathbb{P}\left(\sum_{v\in\mathcal{S}_u^{\beta_l,\beta_h}(i)} \mathbb{I}\left\{\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} \le \zeta\right\} < k\right)
$$

$$
= \mathbb{P}\left(\sum_v \mathbb{I}\left\{v \in \mathcal{S}_u^{\beta_l,\beta_h}(i)\right\} \mathbb{I}\left\{\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} \le \zeta\right\} < k\right).
$$

For any N, define event $Q := \left\{\sum_v \mathbb{I}\left\{\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} \le \zeta\right\} < N\right\}$. Therefore we can condition on event Q to show that

$$
\mathbb{P}\left(\sum_v \mathbb{I}\left\{v \in \mathcal{S}_u^{\beta_l,\beta_h}(i)\right\} \mathbb{I}\left\{\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} \le \zeta\right\} < k\right)
$$

$$
= \mathbb{P}\left(\sum_v \mathbb{I}\left\{v \in \mathcal{S}_u^{\beta_l,\beta_h}(i)\right\} \mathbb{I}\left\{\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} \le \zeta\right\} < k \;\middle|\; Q\right) \mathbb{P}(Q) +
$$

$$
+ \mathbb{P}\left(\sum_v \mathbb{I}\left\{v \in \mathcal{S}_u^{\beta_l,\beta_h}(i)\right\} \mathbb{I}\left\{\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} \le \zeta\right\} < k \;\middle|\; Q^c\right) \mathbb{P}(Q^c)
$$

$$
\le \mathbb{P}(Q) + \mathbb{P}\left(\sum_v \mathbb{I}\left\{v \in \mathcal{S}_u^{\beta_l,\beta_h}(i)\right\} \mathbb{I}\left\{\sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} \le \zeta\right\} < k \;\middle|\; Q^c\right).
$$

First we provide a bound for $\mathbb{P}(Q)$. For any $a, b \in \mathcal{X}_1^\pi$, and random variable

$\mathbf{x}_2 \in \mathcal{X}_2^\pi$, by the L-Lipschitz property of function f,

$$
\begin{aligned}
\sigma_{ab}^2 &= \mathrm{Var}_{\mathbf{x}_2}[f(a, \mathbf{x}_2) - f(b, \mathbf{x}_2)] \\
&\leq \mathbb{E}_{\mathbf{x}_2}[(f(a, \mathbf{x}_2) - f(b, \mathbf{x}_2))^2] \\
&\leq L^2 d_{\mathcal{X}_1^\pi}(a, b)^2,
\end{aligned}
$$

(8.17)

where we recall that the metric $d_{\mathcal{X}_1^\pi}$ is defined as the maximum over the distance in each of the t_1 dimensions. Therefore,

$$
d_{\mathcal{X}_{\pi(q)}}(x_{\pi(q)}(u_q), x_{\pi(q)}(v_q)) \leq \zeta^{1/2} L^{-1} \text{ for all } q \in [t_1]
$$

implies that $d_{\mathcal{X}_1^\pi}(\vec{x}_1^\pi(\vec{u}), \vec{x}_1^\pi(\vec{v}))$ is less than $\zeta^{1/2} L^{-1}$, which implies $\sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}^2 \leq \zeta$ by substituting into (8.17). For any t_1-sequence of positive integers $(N_1, N_2, \ldots, N_{t_1})$ such that $\prod_{q \in [t_1]} N_q \geq N$,

$$
\bigcap_{q \in [t_1]} \left\{ \sum_{l \in [n_{\pi(q)}]} \mathbb{I}\left\{ d_{\mathcal{X}_{\pi(q)}}\left(x_{\pi(q)}(u_q), x_{\pi(q)}(l)\right) \leq \zeta^{1/2} L^{-1} \right\} \geq N_q \right\}
$$
$$
\implies \left\{ \sum_v \mathbb{I}\left\{ \sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}^2 \leq \zeta \right\} \geq N \right\}.
$$

Therefore, we have the following inequality from applying the union bound:

$$
\mathbb{P}(Q) = \mathbb{P}\left(\sum_v \mathbb{I}\left\{ \sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}^2 \leq \zeta \right\} < N \right)
$$
$$
\leq \sum_{q \in [t_1]} \mathbb{P}\left(\sum_{l \in [n_{\pi(q)}]} \mathbb{I}\left\{ d_{\mathcal{X}_{\pi(q)}}\left(\mathbf{x}_{\pi(q)}(u_q), \mathbf{x}_{\pi(q)}(l)\right) \leq \zeta^{1/2} L^{-1} \right\} < N_q \right).
$$

Because $\{\mathbf{x}_{\pi(q)}(l)\}_{l \in [n_{\pi(q)}]}$ are drawn i.i.d. from $\mathcal{X}_{\pi(q)}$, the indicator variables are i.i.d. Bernoulli random variables whose success parameter is at least $\phi_{\pi(q)}\left(\zeta^{1/2} L^{-1}\right)$ by definition of the underestimator function $\phi_{\pi(q)}$. Therefore, by Chernoff's bound, it follows that

$$
\mathbb{P}\left(\sum_{l \in [n_{\pi(q)}]} \mathbb{I}\left\{ d_{\mathcal{X}_{\pi(q)}}\left(\mathbf{x}_{\pi(q)}(u_q), \mathbf{x}_{\pi(q)}(l)\right) \leq \zeta^{1/2} L^{-1} \right\} < N_q \right)
$$
$$
\leq \mathbb{P}\left(\mathrm{Binomial}\left(n_{\pi(q)}, \phi_{\pi(q)}\left(\zeta^{1/2} L^{-1}\right)\right) \leq N_q \right)
$$
$$
\leq \exp\left(-\frac{(\mu_q - N_q)^2}{2\mu_q} \right),
$$

(8.18)

for $N_q < \mu_q$ where $\mu_q := n_{\pi(q)} \phi_{\pi(q)} \left(\zeta^{1/2} L^{-1} \right)$. Given parameter N and information on the underestimator functions $\{\phi_{\pi(q)}\}_{q \in [t_1]}$, we can optimize the choice of $\{N_q\}_{q \in [t_1]}$ by balancing $\frac{(\mu_q - N_q)^2}{2\mu_q}$.

Next we provide a bound for $\mathbb{P}\left(\sum_{v \in \mathcal{S}_u^{\beta_l, \beta_h}(i)} \mathbb{I}\left\{ \sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} \leq \zeta \right\} < k \,\Big|\, Q^c \right)$. The event $\left\{ v \in \mathcal{S}_u^{\beta_l, \beta_h}(i) \right\}$ is solely determined by the locations of the sampled datapoints represented in index set $\mathcal{D}$, thus it is completely independent from the latent variables and the entry values. Thus it is independent from $\left\{ \sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} \leq \zeta \right\}$ as long as $v \neq u$. Therefore, the probability of interest

$$\mathbb{P}\left(\sum_v \mathbb{I}\left\{ v \in \mathcal{S}_u^{\beta_l, \beta_h}(i) \right\} \mathbb{I}\left\{ \sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} \leq \zeta \right\} < k \,\Bigg|\, Q^c \right)$$

$$= \mathbb{P}\left(\sum_{v : \sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} \leq \zeta} \mathbb{I}\left\{ v \in \mathcal{S}_u^{\beta_l, \beta_h}(i) \right\} < k \,\Bigg|\, Q^c \right)$$

$$\leq \mathbb{P}\left(\sum_{v=1, v \neq u}^N \mathbb{I}\left\{ v \in \mathcal{S}_u^{\beta_l, \beta_h}(i) \right\} < k \right)$$

$$= \mathbb{P}\left(\left| \mathcal{S}_u^{\beta_l, \beta_h}(i) \cap ([N] \setminus \{u\}) \right| < k \right),$$

where we used the fact that Q^c implies by definition that the set of rows v such that $\mathbb{I}\left\{ \sigma^2_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} \leq \zeta \right\}$ is larger than N and the probability of event of interest is monotonically decreasing in the number of such vs. Recall that Lemma 8.1 proved concentration of $|\mathcal{S}_u^{\beta_l, \beta_h}(i)|$. By the same proof as Lemma 8.1, replacing m with N, it follows that for $k < Np$,

(8.19)
$$\mathbb{P}\left(\left| \mathcal{S}_u^{\beta_l, \beta_h}(i) \cap ([N] \setminus \{u\}) \right| < k \right) \leq \exp\left(-\frac{((N-1)p - k)^2}{2(N-1)p} \right)$$

(8.20)
$$+ (N-1) \exp\left(-\frac{(n'p^2 - \beta_l)^2}{2n'p^2} \right) + (N-1) \exp\left(-\frac{(\beta_h - n'p^2)^2}{3n'p^2} \right).$$

We choose $N_q = 2^{-1/t_1} \mu_q = 2^{-1/t_1} n_{\pi(q)} \phi_{\pi(q)} \left(\zeta^{1/2} L^{-1} \right)$, such that it also follows that

$$N = \prod_{q \in [t_1]} N_q = \frac{1}{2} \prod_{q \in [t_1]} n_{\pi(q)} \phi_{\pi(q)} \left(\zeta^{1/2} L^{-1} \right) = \frac{1}{2} m \phi_1^\pi \left(\zeta^{1/2} L^{-1} \right).$$

Therefore, given the constraint on k such that $k \leq \frac{1}{8}mp\phi_1^\pi \left(\zeta^{1/2}L^{-1}\right)$, we can verify that $k < Np$.

With this choice of N_q, the expression in (8.18) is upper bounded by

$$\exp\left(-\frac{(\mu_q - N_q)^2}{2\mu_q}\right) = \exp\left(-\frac{(1 - 2^{-1/t_1})^2}{2}\mu_q\right).$$

Since we may assume $N \geq 2$ (because $N = 1$ gives a meaningless event E and we are considering the large tensor limit as $n_q \to \infty, \forall q$), it holds that $N - 1 \geq \frac{N}{2}$. Therefore, the first expression in (8.19) is upper bounded by

$$\exp\left(-\frac{((N-1)p - k)^2}{2(N-1)p}\right) \leq \exp\left(-\frac{\left(\frac{1}{4}Np\right)^2}{2Np}\right) = \exp\left(-\frac{Np}{32}\right) \leq \exp\left(-\frac{k}{8}\right).$$

Using $N \leq \prod_{q \in \mathcal{I}_1} n_q = m$ in (8.19) and combining everything together, we obtain the desired result. $\square$

8.4. *Tail Probability Bound Conditioned on Good Events.*

PROOF OF LEMMA 8.5. We redefine S^k so that it denotes the set of the k best row indices v in $\mathcal{S}_u^{\beta_l, \beta_h}(i)$ which have minimum sample variance $s_{uv}^2(i)$. We define $Err^k(u,i) := A(u,i) - \hat{A}^k(u,i)$ and are interested in its probabilistic tail bound. By the same argument as in the proof of Lemma 7.5, its absolute value can be bounded as below

$$\left|A(u,i) - \hat{A}^k(u,i)\right|$$

$$\leq \left|\frac{1}{|S^k|}\sum_{v \in S^k}\left(A(u,i) - A(v,i) - \eta(v,i) - \mu_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}\right)\right|$$

$$+ \left|\frac{1}{|S^k|}\sum_{v \in S^k} m_{uv}(i) - \mu_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}\right|.$$

Recall that we denote $E' := E_1' \cap E_2' \cap E_3' \cap E_4'$. Following the same lines of

argument as in the proof of Lemma 7.5, we can show that

$$\mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon \mid E'\right)$$

$$\leq \mathbb{P}\left(\left|\frac{1}{|S^k|}\sum_{v\in S^k}\left(A(u,i) - A(v,i) - \eta(v,i) - \mu_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}\right)\right| > \varepsilon - \nu \;\middle|\; E'\right)$$

$$= \int_{\vec{y}\in(\mathcal{X}_1^\pi)^m}\sum_{\mathcal{S}_0\subset[n]:|\mathcal{S}_0|=k}\mathbb{P}\left(\left|\text{"expr"}\right| > \varepsilon - \nu \mid (\vec{x}_1^\pi(\vec{v}))_{v\in[m]} = \vec{y}, S^k = \mathcal{S}_0, E'\right)$$

$$\mathbb{P}\left((\vec{x}_1^\pi(\vec{v}))_{v\in[m]} = \vec{y}, S^k = \mathcal{S}_0 \mid E'\right)d\vec{y},$$

where "expr" denotes the expression $\frac{1}{|S^k|}\sum_{v\in S^k}\left(A(u,i) - A(v,i) - \eta(v,i) - \mu_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}\right)$.

Recall that we modified the tensor algorithm such that the overlap $\mathcal{O}_i^{uv}$, sample variance $s_{uv}^2(i)$, row base $\mathcal{S}_u^{\beta_l,\beta_h}(i)$, and thus the set S^k are computed from the data matrix after removing all columns j which share any of the original tensor coordinates, i.e. $j_k = i_k$ for any $k \in [t_2]$. Therefore, the event E' and the set S^k are independent from the i-th column latent feature $\vec{x}_2(\vec{i})$ and the noise terms $\eta(v,i)$ for any $v \in [m]$. Let $\vec{\mathbf{x}}_2$ be a random variable sampled independently from the product space $P_{\mathcal{X}_2^\pi}$. Then we can verify that

$$\mathbb{E}[A(u,i) - A(v,i) - \eta(v,i) - \mu_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} \mid (\vec{x}_1^\pi(\vec{v}))_{v\in[m]} = \vec{y}, S^k = \mathcal{S}_0, E']$$

$$= \mathbb{E}[f(\vec{x}_1^\pi(\vec{u}), \vec{x}_2^\pi(\vec{i})) - f(\vec{x}_1^\pi(\vec{v}), \vec{x}_2^\pi(\vec{i})) - \eta(v,i) \mid (\vec{x}_1^\pi(\vec{v}))_{v\in[m]} = \vec{y}, S^k = \mathcal{S}_0, E'] - \mu_{y_u y_v}$$

$$= \mathbb{E}[f(y_u, \vec{\mathbf{x}}_2) - f(y_v, \vec{\mathbf{x}}_2) - \eta(v,i) \mid (\vec{x}_1^\pi(\vec{v}))_{v\in[m]} = \vec{y}, S^k = \mathcal{S}_0, E'] - \mu_{y_u y_v}$$

$$= 0.$$

Next we apply Chebyshev's inequality and Cauchy-Schwarz inequality along with the fact that $\eta(v,i)$ is independent from E' because the events in E' do not depend on the data observed from column i.

$$\mathbb{P}\left(\left|\text{"expr"}\right| > \varepsilon - \nu \mid (\vec{x}_1^\pi(\vec{v})))_{v\in[m]} = \vec{y}, S^k = \mathcal{S}_0, E\right)$$

$$\leq \frac{\text{Var}\left[\frac{1}{|S^k|}\sum_{v\in S^k}(A(u,i) - A(v,i) - \eta(v,i)) \mid (\vec{x}_1^\pi(\vec{v}))_{v\in[m]} = \vec{y}, S^k = \mathcal{S}_0, E'\right]}{(\varepsilon - \nu)^2}$$

$$\leq \frac{1}{(\varepsilon - \nu)^2}\left(\left(\frac{1}{k}\sum_{v\in\mathcal{S}_0}\sqrt{\text{Var}\left[A(u,i) - A(v,i) \mid (\vec{x}_1^\pi(\vec{v}))_{v\in[m]} = \vec{y}, S^k = \mathcal{S}_0, E'\right]}\right)^2 + \frac{\gamma^2}{k}\right).$$

Next we will bound $\text{Var}\left[A(u,i) - A(v,i) \mid (\vec{x}_1^\pi(\vec{v}))_{v\in[m]} = \vec{y}, S^k = \mathcal{S}_0, E'\right]$

for any $v \in \mathcal{S}_0$. Therefore,

$$
\begin{aligned}
&\mathrm{Var}\left[A(u,i) - A(v,i) \mid (\vec{x}_1^\pi(\vec{v}))_{v\in[m]} = \vec{y}, S^k = \mathcal{S}_0, E'\right] \\
&= \mathrm{Var}\left[f(y_u, \vec{x}_2(\vec{i})) - f(y_v, \vec{x}_2(\vec{i})) \mid (\vec{x}_1^\pi(\vec{v}))_{v\in[m]} = \vec{y}, S^k = \mathcal{S}_0, E'\right] \\
&= \mathrm{Var}\left[f(y_u, \mathbf{x}_2) - f(y_v, \mathbf{x}_2) \mid (\vec{x}_1^\pi(\vec{v}))_{v\in[m]} = \vec{y}, S^k = \mathcal{S}_0, E'\right] \\
&= \sigma_{y_u y_v}^2.
\end{aligned}
$$

Let $\tilde{\mathcal{V}}$ denote the subset of rows $v \in \mathcal{S}_u^{\beta_l, \beta_h}(i)$ such that $\sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}^2 \leq \zeta$. Conditioned on E_4, the size of set $\tilde{\mathcal{V}}$ must be at least k. Conditioned on E_3', for every $v \in \tilde{\mathcal{V}} \subset \mathcal{S}_u^{\beta_l, \beta_h}(i)$,

$$
s_{uv}^2(i) \leq (1+\theta)\, \sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}^2 + 2\gamma^2 + \tau \leq (1+\theta)\zeta + 2\gamma^2 + \tau.
$$

Therefore, it follows by definition of S^k as the set of k rows with minimum sample variance, that for all $v \in S^k$, $s_{uv}^2(i) \leq (1+\theta)\zeta + \tau + 2\gamma^2$. Again due to event E_3', this implies that for all $v \in S^k$, $\sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}^2 \leq \frac{(1+\theta)\zeta + 2\tau}{1-\theta}$ because $(1-\theta)\sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}^2 - \tau \leq s_{uv}^2(i) - 2\gamma^2$. Therefore for all $v \in S^k$,

$$
\mathrm{Var}\left[A(u,i) - A(v,i) \mid (\vec{x}_1^\pi(\vec{v}))_{v\in[m]} = \vec{y}, S^k = \mathcal{S}_0, E'\right] \leq \frac{(1+\theta)\zeta + 2\tau}{1-\theta}.
$$

Finally we can show that

$$
\begin{aligned}
&\mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon \mid (\vec{x}_1^\pi(\vec{v}))_{v\in[m]} = \vec{y}, S^k = \mathcal{S}_0, E'\right) \\
&\qquad \leq \frac{1}{(\varepsilon - \nu)^2}\left(\frac{(1+\theta)\zeta + 2\tau}{1-\theta} + \frac{\gamma^2}{k}\right).
\end{aligned}
$$

Since this bound does not depend on the choice of $\vec{y}$ and $\mathcal{S}_0$, when we integrate over all choices of $\vec{y}$ and $\mathcal{S}_0$, we obtain the same bound. $\qquad\square$

PROOF OF LEMMA 8.6. We restate the following conditioning events (defined in Lemma 8.5) for readability:

$$
\begin{aligned}
E' &:= E_1' \cap E_2' \cap E_3' \cap E_4', \\
E_1' &:= \left\{\left|\mathcal{S}_u^{\beta_l, \beta_h}(i)\right| \in \left[\frac{1}{2}(m-1)p, \frac{3}{2}(m-1)p\right]\right\}, \\
E_2' &:= \left\{\left|m_{uv} - \mu_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}\right| \leq \nu, \; \forall\, v \in \mathcal{S}_u^{\beta_l, \beta_h}(i)\right\}, \\
E_3' &:= \left\{s_{uv}^2(i) \in \left[(1-\theta)\sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}^2 - \tau, (1+\theta)\sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}^2 + \tau\right], \; \forall\, v \in \mathcal{S}_u^{\beta_l, \beta_h}(i)\right\}, \\
E_4' &:= \left\{\left\{\sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}^2\right\}_{v\in\mathcal{S}_u^{\beta_l, \beta_h}(i)}^{(k)} \leq \zeta\right\},
\end{aligned}
$$

First of all, note that (see the proof of Theorem 7.6 for more detail)

$$\mathbb{P}\left(E'^c\right) \le \mathbb{P}\left(E_1'^c\right) + \mathbb{P}\left(E_2'^c|E_1'\right) + \mathbb{P}\left(E_3'^c|E_1'\right) + \mathbb{P}\left(E_4'^c|E_1'\right).$$

Using Lemma 8.1, we have

$$\mathbb{P}\left(E_1'^c\right) = \mathbb{P}\left(|\mathcal{S}_u^{\beta_l,\beta_h}(i)| \notin \left[\frac{1}{2}(m-1)p, \frac{3}{2}(m-1)p\right]\right)$$

$$\le (m-1)\exp\left(-\frac{(n'p^2 - \beta_l)^2}{2n'p^2}\right) + (m-1)\exp\left(-\frac{(\beta_h - n'p^2)^2}{3n'p^2}\right)$$

$$+ 2\exp\left(-\frac{(m-1)p}{12}\right).$$

Similarly, by using Lemma 8.2 with union bound,

$$\mathbb{P}\left(E_2'^c|E_1'\right) = \mathbb{P}\left(\bigcup_{v \in \mathcal{S}_u^{\beta_l,\beta_h}(i)} \left\{|\mu_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} - m_{uv}| > \nu\right\}\right)$$

$$\le 3(m-1)p\exp\left(\frac{-\nu^2}{8L^2D^2\theta + \frac{16B_e^2}{n'}}\right)$$

$$+ 3(m-1)p\exp\left(\frac{-\nu^2}{32(LD + 2B_e)^2}\Delta\right),$$

where we recall that $\Delta = \min\left\{\frac{n'^2\beta_l}{(n'+\beta_l^2)^2}, \frac{n'^2\beta_h}{(n'+\beta_h^2)^2}\right\}$ and $\theta = \sum_{q \in \mathcal{I}_2} \frac{1}{n_q - 1}$ is a quantity which depends only on the shape of the given tensor instance, and it vanishes to 0 as $n_q \to \infty, \forall q \in \mathcal{I}_2$.

By using Lemma 8.3 with union bound again,

$$\mathbb{P}\left(E_3'^c|E_1'\right) = \mathbb{P}\left(\bigcup_{v \in \mathcal{S}_u^\beta(i)} \left\{s_{uv}^2(i) \in \left[(1-\theta)\sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}^2 - \tau, (1+\theta)\sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})}^2 + \tau\right]\right\}\right)$$

$$\le 3(m-1)p\exp\left(\frac{-\tau^2}{32L^2D^2\left(3LD + 4B_e\right)^2\theta + \frac{64B_e^2(2LD+5B_e)^2}{n'}}\right)$$

$$+ 3(m-1)p\exp\left(\frac{-\tau^2}{128(LD + 2B_e)^4}\Delta\right).$$

By Lemma 8.4 with $\mu_q := n_q\phi_q\left(\sqrt{\frac{\zeta}{L^2}}\right)$ and $\phi_q(r) := \operatorname{ess\,inf}_{x' \in \mathcal{X}_q} \mathbb{P}_{\mathcal{X}_q}\left(d_{\mathcal{X}_q}(\mathbf{x}, x') \le r\right)$

as previously defined,

$$\mathbb{P}\left(E_4'^c|E_1'\right) \le \sum_{q\in\mathcal{I}_1} \exp\left(-\frac{(1-2^{-1/t_1})^2}{2}\mu_q\right) + \exp\left(-\frac{k}{8}\right)$$

$$+ (m-1)\exp\left(-\frac{(n'p^2-\beta_l)^2}{2n'p^2}\right) + (m-1)\exp\left(-\frac{(\beta_h-n'p^2)^2}{3n'p^2}\right).$$

Putting everything together, we obtain the following bound

(8.21)
$$\mathbb{P}\left(E'^c\right) \le 2(m-1)\exp\left(-\frac{\left(n'p^2-\beta_l\right)^2}{2n'p^2}\right) + 2(m-1)\exp\left(-\frac{\left(\beta_h-n'p^2\right)^2}{3n'p^2}\right)$$

(8.22)
$$+ 2\exp\left(-\frac{(m-1)p}{12}\right)$$

(8.23)
$$+ 3(m-1)p\exp\left(\frac{-\nu^2}{8L^2D^2\theta + \frac{16B_e^2}{n'}}\right)$$

(8.24)
$$+ 3(m-1)p\exp\left(\frac{-\nu^2}{32(LD+2B_e)^2}\Delta\right)$$

(8.25)
$$+ 3(m-1)p\exp\left(\frac{-\tau^2}{32L^2D^2\left(3LD+4B_e\right)^2\theta + \frac{64B_e^2(2LD+5B_e)^2}{n'}}\right)$$

(8.26)
$$+ 3(m-1)p\exp\left(\frac{-\tau^2}{128(LD+2B_e)^4}\Delta\right)$$

(8.27)
$$+ \sum_{q\in\mathcal{I}_1} \exp\left(-\frac{(1-2^{-1/t_1})^2}{2}\mu_q\right) + \exp\left(-\frac{k}{8}\right).$$

Note that ζ, ν, τ are parameters which are introduced purely for the purpose of analysis. Requiring all exponential terms decay to 0 as $m, n \to \infty$ restricts the range of values ζ, ν, τ can take. Define $q^* := \arg\min_{q\in\mathcal{I}_1} n_q$. Then we choose0 $\nu = \tau = \max\left\{(n_{q^*}-1)^{-1/3}, \left(\frac{n'^2}{\beta_h^3}\right)^{-1/3}, \beta_l^{-1/3}\right\}$. Also, we enforce ζ satisfies

$\phi_q\left(\sqrt{\frac{\zeta}{L^2}}\right) \geq c_q n_q^{-\frac{\log mp}{2\log m}}$ so that $\mu_q \geq c_q n_q^{\delta/2}$ even for the worst choice of p, and $mp\phi_1^\pi\left(\sqrt{\frac{\zeta}{L^2}}\right) \geq \left(\prod_{q\in\mathcal{I}_1} c_q\right)\sqrt{mp}$ as described in the theorem statement.

Additionally, we require $\beta_l, \beta_h \to \infty$ as $m, n \to \infty$ while $n'p^2 - \beta_l, \beta_h - n'p^2 \to \infty$ as well. We will assume $\beta_h \leq n'^{2/3-\delta}$ for convenience in analysis. Lastly, we want $k \to \infty$, which is possible while satisfying the constraint $k \leq \frac{1}{8}mp\phi_1^\pi\left(\sqrt{\frac{\zeta}{L^2}}\right) = \frac{\prod_{q\in\mathcal{I}_1} c_q}{8}\sqrt{mp}$, which originated from Lemma 8.5. Let's suppose that we take $k = \frac{\prod_{q\in\mathcal{I}_1} c_q}{8}\sqrt{mp}$ by default. We will show these are sufficient conditions for the convergence of exponential error terms.

Given a sequence of problems of size (m, n), suppose that $p = \omega(m^{-1})$ and $p = \omega(n'^{-1/2})$. Then $mp, n'p^2 \to \infty$ as $m, n \to \infty$, since $n' = \Theta(n)$. We may assume without loss of generality that m and n' are large enough such that $mp, n'p^2 \geq 2$. As $n_q \to \infty, \forall q$, we can observe that $\nu, \tau \to 0$. In addition, we have $\nu \leq 1, \tau \leq 1$, since we assumed $\beta_h \leq n'^{2/3-\delta}$.

Now we will clean up the exponential error terms in Eq. (8.21)-(8.27) to obtain simpler yet more enlightening forms.

Eq. (8.21): Since $\beta_l \leq c_l n'p^2$, we have $(n'p^2 - \beta_l)^2 \geq \left[(1-c_l)n'p^2\right]^2$. Similarly, $\beta_h \geq c_h n'p^2$ implies that $(\beta_h - n'p^2)^2 \geq \left[(c_h-1)n'p^2\right]^2$. Therefore, the terms in Eq. (8.21) are upper bounded by

$$2(m-1)\exp\left(-\frac{(1-c_l)^2}{2}n'p^2\right) + 2(m-1)\exp\left(-\frac{(c_h-1)^2}{3}n'p^2\right).$$

Eq. (8.22): The term in Eq. (8.22) is upper bounded by $2\exp\left(-\frac{1}{24}mp\right)$, using the fact that $m - 1 \geq \frac{m}{2}$ for $m \geq 2$, which is implied by our assumption that $mp \geq 2$.

Eq. (8.23) and (8.25): Due to our choice of ν (and τ), we have $\nu \geq (n_{q^*}-1)^{-1/3}$ where $q^* = \arg\max_{q\in\mathcal{I}_1}\left((n_q-1)^{-1/3}\right)$. For any constants $A, B > 0$, because $(n_q-1)^{-1} \leq (n_{q^*}-1)^{-1}, \forall q \in \mathcal{I}_2$, and $n' = \prod_{q\in\mathcal{I}_2}(n_q-1) \geq n_{q^*}-1$, it follows that

$$\frac{\nu^2}{A\sum_{q\in\mathcal{I}_2}\frac{1}{n_q-1} + B\frac{1}{n'}} \geq \frac{(n_{q^*}-1)^{-2/3}}{(At_2+B)(n_{q^*}-1)^{-1}} = \frac{(n_{q^*}-1)^{1/3}}{At_2+B}.$$

Eq. (8.24) and (8.26): Again from our choice of ν (and τ), the following inequality is true: $\nu \geq \max\left\{\left(\frac{n'^2}{\beta_h^3}\right)^{-1/3}, \beta_l^{-1/3}\right\}$. Meanwhile, $\Delta := \min\{\Delta_l, \Delta_h\}$ where $\Delta_l := \frac{n'^2\beta_l}{(n'+\beta_l^2)^2}$ and $\Delta_h := \frac{n'^2\beta_h}{(n'+\beta_h^2)^2}$. Therefore, $\nu^2\Delta \geq \min\left\{\left(\frac{n'^2}{\beta_h^3}\right)^{-2/3}\Delta_h, \beta_l^{-2/3}\Delta_l\right\}$.

From the assumption $\beta_h \geq \sqrt{n'}$,

$$\left(\frac{n'^2}{\beta_h^3}\right)^{-2/3} \Delta_h = \left(\frac{n'^2}{\beta_h^3}\right)^{-2/3} \frac{n'^2 \beta_h}{\left(n' + \beta_h^2\right)^2}$$

$$\geq \left(\frac{n'^2}{\beta_h^3}\right)^{-2/3} \frac{n'^2 \beta_h}{\left(2\beta_h^2\right)^2}$$

$$= \frac{n'^{2/3}}{4\beta_h}.$$

Because we assumed $\beta_h \leq n'^{2/3-\delta}$, this value diverges no slower than $n'^\delta/4$ as $n' = \Theta(n) \to \infty$.

Similarly, from the assumption $\beta_l \leq \sqrt{n'}$,

$$\beta_l^{-2/3} \Delta_l = \beta_l^{-2/3} \frac{n'^2 \beta_l}{\left(n' + \beta_l^2\right)^2}$$

$$\geq \beta_l^{-2/3} \frac{n'^2 \beta_l}{\left(2n'\right)^2}$$

$$= \frac{\beta_l^{1/3}}{4}.$$

Eq. (8.27): $\mu_q = n_q \phi_q\left(\sqrt{\frac{\zeta}{L^2}}\right) \geq c_q n_q^{1-\frac{\log mp}{2\log m}} \geq c_q n_q^{1-\frac{\log m}{2\log m}} = c_q n_q^{1/2}$ because $p \leq 1$. We will leave k as a free parameter, but our default choice for k, which is $\frac{\prod_{q\in\mathcal{I}_1} c_q}{8}\sqrt{mp}$, diverges as $m \to \infty$.

Consequently, we have

$$\mathbb{P}\left(E'^c\right) \leq 4(m-1)\exp\left(-c_1 n' p^2\right) + 2\exp\left(-\frac{1}{24}mp\right)$$

$$+ 6(m-1)p\exp\left(-c_2(n_{q^*}-1)^{1/3}\right) + 6(m-1)p\exp\left(-c_3\min\left\{\frac{n'^{2/3}}{4\beta_h}, \frac{\beta_l^{1/3}}{4}\right\}\right)$$

$$+ t_1\exp\left(-c_4 n_{q^*}^{1/2}\right) + \exp\left(-\frac{k}{8}\right),$$

where

$$q^* := \arg\min_{q \in \mathcal{I}_1} n_q,$$

$$c_1 := \min\left\{\frac{(1-c_l)^2}{2}, \frac{(c_h-1)^2}{3}\right\},$$

$$c_2 := \min\left\{\frac{1}{8L^2D^2t_2 + 16B_e^2}, \frac{1}{32L^2D^2(3LD+4B_e)^2t_2 + 64B_e^2(2LD+5B_e)^2}\right\},$$

$$c_3 := \min\left\{\frac{1}{32(LD+2B_e)^2}, \frac{1}{128(LD+2B_e)^4}\right\},$$

$$c_4 := \min_{q \in \mathcal{I}_q}\left\{\frac{(1-2^{-1/t_1})^2}{2}c_q\right\}.$$

Note that c_1, c_2, c_3, c_4 are absolute constants, which may depend only on the geometry of the latent spaces.

$\square$

8.5. *Proof of Main Results for Tensor Completion.*

PROOF OF THEOREM 8.7. The proof follows from simply combining the results from Lemmas 8.5 and 8.6. By conditioning on event E', it follows that

$$\mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon\right) \leq \mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon \,\Big|\, E'\right) + \mathbb{P}\left(E'^c\right).$$

By Lemma 8.5,

$$\mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon \,\Big|\, E'\right) \leq \frac{1}{(\varepsilon - \nu)^2}\left(\frac{(1+\theta)\zeta + 2\tau}{1-\theta} + \frac{\gamma^2}{k}\right).$$

By Lemma 8.6, $\mathbb{P}\left(E'^c\right) \leq F_3'$ for

$$F_3' = 4(m-1)\exp\left(-c_1 n'p^2\right) + 2\exp\left(-\frac{1}{24}mp\right)$$

$$+ 6(m-1)p\exp\left(-c_2(n_{q^*}-1)^{1/3}\right) + 6(m-1)p\exp\left(-c_3\min\left\{\frac{n'^{2/3}}{4\beta_h}, \frac{\beta_l^{1/3}}{4}\right\}\right)$$

$$+ t_1\exp\left(-c_4 n_{q^*}^{1/2}\right) + \exp\left(-\frac{k}{8}\right),$$

where $q^* := \arg\min_{q \in \mathcal{I}_1} n_q$, and c_1, c_2, c_3, c_4 are some absolute constants derived in Lemma 8.6, which may depend only on the geometry of the latent spaces.

Therefore,

$$\mathbb{P}\left(\left|A(u,i) - \hat{A}^k(u,i)\right| > \varepsilon\right) \leq \frac{1}{(\varepsilon - \nu)^2}\left(\frac{(1+\theta)\zeta + 2\tau}{1-\theta} + \frac{\gamma^2}{k}\right) + F_3'.$$

$\square$

PROOF OF THEOREM 5.1 IN THE MAIN ARTICLE. The proof is the same with that of Theorem 4.1 (main article), and follows by integrating the tail of the probability bound presented in Theorem 8.7. $\square$

8.6. *Results Given an Optimal Flattening and Uniform Probability Measure.* In this section, we simplify the results from Theorem 5.1 (main article) when the latent probability measure is a uniform measure over a d dimensional Euclidean cube. We thus compute the specific form of the underestimator function $\phi_1^\pi(\cdot)$, which leads to concrete bounds. We then choose specific expressions for ζ, and k to ensure that the mean squared error of our user-user k-nearest neighbor algorithm converges to zero. The parameter ζ in Theorem 5.1 (main article) is introduced purely for the purpose of analysis, and is not used within the implementation of the the algorithm. Recall that it is used to define event E_4', which holds when the k rows in $\mathcal{S}_u^{\beta_l,\beta_h}(i)$ with minimum variance all satisfy $\sigma_{\vec{x}_1^\pi(\vec{u})\vec{x}_1^\pi(\vec{v})} \leq \zeta$. Intuitively, ζ is the thresholding parameter for the membership of similar neighbors.

PROOF OF COROLLARY 5.3 IN THE MAIN ARTICLE. Recall that we assumed an equilateral tensor with $n_q = l$ for all $q \in [t]$, and we assumed that the algorithm is applied to a user-optimally flattened tensor, such that $|\mathcal{I}_1| = t - \lfloor\frac{2t}{3}\rfloor$, and $|\mathcal{I}_2| = \lfloor\frac{2t}{3}\rfloor$. Therefore,

$$l^{\frac{t}{3}} \leq m = l^{t-\lfloor\frac{2t}{3}\rfloor} \leq l^{\lceil\frac{t}{3}\rceil},$$
$$l^{\lfloor\frac{2t}{3}\rfloor} \leq n = l^{\lfloor\frac{2t}{3}\rfloor} \leq l^{\frac{2t}{3}},$$
$$(l-1)^{\lfloor\frac{2t}{3}\rfloor} \leq n = (l-1)^{\lfloor\frac{2t}{3}\rfloor} \leq (l-1)^{\frac{2t}{3}}.$$

When the latent space for each coordinate $q \in \mathcal{I}_1$ is a cube in $\mathbb{R}^d$ equipped with the uniform probability measure, then for $q \in \mathcal{I}_1$,

$$\phi_q\left(\sqrt{\frac{\zeta}{L^2}}\right) = C(2L)^{-d}\zeta^{d/2},$$

where C is a normalization constant to ensure $\phi_q(\mathcal{X}_q) = 1$. Therefore, the corre-

sponding underestimator function of the corresponding product space is

$$\phi_1^\pi\left(\sqrt{\frac{\zeta}{L^2}}\right) = \prod_{q\in\mathcal{I}_1}\phi_q\left(\sqrt{\frac{\zeta}{L^2}}\right)$$
$$= \left(C(2L)^{-d}\zeta^{d/2}\right)^{t_1}.$$

As specified in the conditions of Theorem 5.1 (main article), we need to choose ζ so that for all q,

(8.28) $$\phi_q\left(\sqrt{\frac{\zeta}{L^2}}\right) = C(2L)^{-d}\zeta^{d/2} \geq c_q n_q^{-\frac{\log mp}{2\log m}} \text{ for some } c_q \geq 0.$$

Therefore we need

(8.29) $$\zeta \geq \left(\frac{c_q}{C}\right)^{2/d}(2L)^2 n_q^{-\frac{\log mp}{d\log m}}$$

for all q. Since $m = l^{|\mathcal{I}_1|}$ and $n_q = l \,\forall\, q$, it follows that

$$n_q^{-\frac{\log mp}{\log m}} = l^{-\frac{\log mp}{|\mathcal{I}_1|\log l}} = (mp)^{-\frac{1}{|\mathcal{I}_1|}}.$$

Therefore, we will let $c_q = 1$, and choose $\zeta = \frac{(2L)^2}{C^{\frac{2}{d}}}(mp)^{-\frac{1}{d|\mathcal{I}_1|}}$, which satisfies the above conditions.

Since we have $\phi_q\left(\sqrt{\frac{\zeta}{L^2}}\right) \geq (mp)^{-\frac{1}{2|\mathcal{I}_1|}}$ for each $q \in \mathcal{I}_1$, taking $k = \frac{1}{8}\sqrt{mp}$ satisfies the required condition that

$$k = \frac{mp}{8}\left((mp)^{-\frac{1}{2|\mathcal{I}_1|}}\right)^{|\mathcal{I}_1|} \leq \frac{mp}{8}\phi_q^\pi\left(\sqrt{\frac{\zeta}{L^2}}\right).$$

The assumption that $p \geq \max\left\{m^{-1+\delta}, n'^{-\frac{1}{2}+\delta}\right\}$ for some $\delta > 0$ guarantees that mp and $n'p^2$ diverge to ∞ as $m, n \to \infty$, which occurs when $l \to \infty$. For the choice of $\beta_l = \frac{1}{2}\min\left\{n'p^2, \sqrt{n'}\right\}$ and $\beta_h = 2\max\left\{n'p^2, \sqrt{n'}\right\}$, we can verify that $F_2' \to 0$ and $l \to \infty$. By definition

$$F_2' = \max\left\{(n_{q*} - 1)^{-1/3}, \left(\frac{n'^2}{\beta_h^3}\right)^{-1/3}, \beta_l^{-1/3}\right\}.$$

Each of these terms shrink to 0 as $l \to \infty$, using the fact that $p \geq n'^{-\frac{1}{2}+\delta}$, and $p \leq n'^{-\frac{1}{6}-\delta}$ for some $\delta > 0$,

$$(n_{q^*} - 1)^{-1/3} = (l-1)^{-1/3},$$

$$\left(\frac{n'^2}{\beta_h^3}\right)^{-1/3} = \frac{2 \max\left\{n'p^2, \sqrt{n'}\right\}}{n'^{2/3}} \leq 2 \max\left\{n'^{-2\delta}, n'^{-1/6}\right\},$$

$$\beta_l^{-1/3} = \left(\frac{1}{2} \min\left\{n'p^2, \sqrt{n'}\right\}\right)^{-1/3} \leq \left(\frac{1}{2} \min\left\{n'^{2\delta}, \sqrt{n'}\right\}\right)^{-1/3}.$$

Next, we will show $F_1' \to 0$ as $l \to \infty$. First of all, $\theta = \sum_{q \in \mathcal{I}_1} \frac{1}{n_q - 1} = \frac{t - \lfloor \frac{2t}{3} \rfloor}{l-1} \leq \frac{t}{l-1} \to 0$ as $l \to \infty$. To be more specific, we can assume that $\theta \leq \frac{1}{2}$ whenever $l \geq 2t + 1$. Therefore, by plugging in the conditions on θ and the expressions for ζ, k, n', and F_2', it follows that

$$F_1' = \frac{(1+\theta)\zeta + 2F_2'}{1-\theta} + \frac{\gamma^2}{k}$$

$$\leq 3\zeta + 4F_2' + \frac{\gamma^2}{k}$$

$$\leq \left[3\frac{(2L)^2}{C^{\frac{2}{d}}} + 4 + 8\gamma^2\right] \max\left\{(mp)^{-\frac{1}{d|\mathcal{I}_1|}}, (l-1)^{-\frac{1}{3}}, \frac{\beta_h}{(l-1)^{\frac{2}{3}|\mathcal{I}_2|}}, \beta_l^{-\frac{1}{3}}, (mp)^{-\frac{1}{2}}\right\}$$

$$\to 0 \quad as \ l \to \infty.$$

To show that F_3' decays exponentially fast as l grows, we need to show that the exponents in each of the terms in F_3' (see the expression in Theorem 5.1 in the main article) grow faster than the multiplicative coefficients in front. The multiplicative factors $4(m-1)p, 6(m-1)p, t_1$ contribute at most only a logarithmic increase in l when included into the exponents. By showing that the exponents grow polynomially with l, it follows that F_3' decays exponentially with l. We can verify this by simply plugging in the choices for n', m, n_q, k, β_l and β_h, and additionally using the conditions that $\max\left\{m^{-1+\delta}, n'^{-\frac{1}{2}+\delta}\right\} \leq p \leq n'^{-\frac{1}{6}-\delta}$ for some $\delta > 0$. We list the simplified expressions for each of the terms in the exponents below, showing

that they indeed increase polynomially in l.

$$n'p^2 \geq (l-1)^{\frac{2t}{3}\delta},$$

$$mp \geq l^{\frac{t}{3}\delta},$$

$$(n_{q^*}-1)^{1/3} = (l-1)^{1/3},$$

$$n_{q^*}^{1/2} = l^{1/2},$$

$$k = \frac{1}{8}\sqrt{m} \geq \frac{1}{8}l^{\frac{t}{6}}$$

$$\frac{n'^{\frac{2}{3}}}{\beta_h} = \frac{1}{2}\min\left\{n'^{\frac{1}{6}}, \frac{n'^{\frac{2}{3}}}{n'p^2}\right\} \geq \frac{1}{2}\min\left\{n'^{\frac{1}{6}}, n'^{2\delta}\right\}$$

$$\frac{\beta_l^{\frac{1}{3}}}{4} = \frac{1}{4\sqrt[3]{2}}\min\left\{(n'p^2)^{\frac{1}{3}}, n'^{\frac{1}{6}}\right\} \geq \frac{1}{4\sqrt[3]{2}}\min\left\{n'^{\frac{2}{3}\delta}, n'^{\frac{1}{6}}\right\}.$$

From the definition of expression F_1' in Theorem 5.1 (main article), it follows that $F_1' \geq \frac{2F_2'}{1-\theta} \geq 2F_2'$, because $\theta = \sum_{q \in \mathcal{I}_2} \frac{1}{n_q-1} = \frac{\lfloor\frac{2t}{3}\rfloor}{l-1} \in (0, 1]$. Therefore, if we can certify that F_3' decays exponentially fast as l (and thus n and m) grows, and that $F_1' \leq \frac{1}{2}$ when l is sufficiently large compared to t, then by the same lines of argument as in the proof of Corollary 4.3 (main article), we can argue that F_1' dominates the mean-squared-error thereby achieving a similar form of diminishing error bound.

We then simplify the remaining terms assuming $F_1' \leq \frac{1}{2}$. This can be assumed once we show that $F_1' \to 0$ as $l \to \infty$, which can be deduced from observing that $F_2' \to 0$, $\zeta \to 0$, $\theta \to 0$, and $k \to \infty$ as $l \to \infty$ as $l \to \infty$. If $F_1' \leq \frac{1}{2}$ is assumed, we can apply the same idea as in the proof of Corollary 4.3 (main article), where we upper bound $F_2' \leq \frac{1}{2}F_1'$, and then simplify the terms in the MSE bound presented in Theorem 5.1 (main article) using $\log\left(\frac{2B_0}{F_1'}\right) \geq 1$ as long as $F_1' \leq \frac{1}{2}$ and $B_0 \geq 1$:

$$MSE(\hat{A}) \leq 2F_1'\ln\left(\frac{2B_0}{F_1'}\right) + \left(F_1' + F_2'\right)^2 + 2F_2' + 4B_0^2F_3'$$

$$\leq 2F_1'\ln\left(\frac{2B_0}{F_1'}\right) + \frac{9}{4}F_1' + F_1' + 4B_0^2F_3'$$

(8.30)
$$\leq \frac{21}{4}F_1'\ln\left(\frac{2B_0}{F_1'}\right) + 4B_0^2F_3'.$$

Plugging the bounds for F_1' and F_3' into Eq. (8.30) gives the simplified bound presented in Corollary 5.3 (main article). □

References.

AIROLDI, E. M., COSTA, T. B. and CHAN, S. H. (2013). Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Advances in Neural Information Processing Systems* 692–700.

ALDOUS, D. J. (1981). Representations for partially eschangeable arrays of random variables. *J. Multivariate Anal.* **11** 581 - 598.

ANANDKUMAR, A., GE, R., HSU, D., KAKADE, S. M. and TELGARSKY, M. (2014). Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research* **15** 2773–2832.

ARORA, S., GE, R. and MOITRA, A. (2012). Learning topic models–going beyond SVD. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on* 1–10. IEEE.

ARORA, S., GE, R., KANNAN, R. and MOITRA, A. (2012). Computing a nonnegative matrix factorization–provably. In *Proceedings of the 44th annual ACM symposium on Theory of computing* 145–162. ACM.

AUSTIN, T. (2012). Exchangeable random arrays. *Technical Report, Notes for IAS workshop.*

BELL, R. M. and KOREN, Y. (2007). Scalable Collaborative Filtering with Jointly Derived Neighborhood Interpolation Weights. In *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*. ICDM '07 43–52. IEEE Computer Society, Washington, DC, USA.

BRESLER, G., CHEN, G. H. and SHAH, D. (2014). A latent source model for online collaborative filtering. In *Advances in Neural Information Processing Systems* 3347–3355.

BRESLER, G., SHAH, D. and VOLOCH, L. F. (2015). Collaborative Filtering with Low Regret. *arXiv preprint arXiv:1507.05371.*

CAI, D., HE, X., WU, X. and HAN, J. (2008). Non-negative matrix factorization on manifold. In *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on* 63–72. IEEE.

CANDÈS, E. J. and RECHT, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational mathematics* **9** 717–772.

CHATTERJEE, S. (2015). Matrix estimation by universal singular value thresholding. *The Annals of Statistics* **43** 177–214.

DE SILVA, V. and LIM, L. H. (2008). Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications* **30** 1084 – 1127.

FAZEL, M., HINDI, H. and BOYD, S. P. (2003). Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices. In *Proceedings of ACC* **3** 2156–2162. IEEE.

GANDY, S., RECHT, B. and YAMADA, I. (2011). Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems* **27** 025010.

GANTI, R. S., BALZANO, L. and WILLETT, R. (2015). Matrix Completion Under Monotonic Single Index Models. In *Advances in Neural Information Processing Systems* 1864–1872.

GAO, C., LU, Y. and ZHOU, H. H. (2015). Rate-optimal graphon estimation. *The Annals of Statistics* **43** 2624–2652.

GOLDBERG, D., NICHOLS, D., OKI, B. M. and TERRY, D. (1992). Using Collaborative Filtering to Weave an Information Tapestry. *Commun. ACM*.

HOOVER, D. N. (1981). Row-column exchangeability and a generalized model for probability. In *Exchangeability in Probability and Statistics (Rome, 1981)* 281 - 291.

INDYK, P. (2001). Algorithmic applications of low-distortion geometric embeddings. In *focs* **1** 10–33.

INDYK, P. (2004). Nearest neighbors in high-dimensional spaces.

JAIN, P., NETRAPALLI, P. and SANGHAVI, S. (2013). Low-rank matrix completion using alternating minimization. In *Proceedings of the 45th annual ACM symposium on Theory of computing* 665–674. ACM.

JAIN, P. and OH, S. (2014). Provable tensor factorization with missing data. In *Advances in Neural Information Processing Systems* 1431–1439.

KESHAVAN, R., MONTANARI, A. and OH, S. (2009). Matrix completion from a few entries. *IEEE Trans. Inf. Theory* **56**.

KLOPP, O., TSYBAKOV, A. B. and VERZELEN, N. (2015). Oracle inequalities for network models and sparse graphon estimation. *To appear in Annals of Statistics*.

KOLDA, T. G. and BADER, B. W. (2009). Tensor decompositions and applications. *SIAM review* **51** 455–500.

KOLDA, T. G. and SUN, J. (2008). Scalable tensor decompositions for multi-aspect data mining. In *2008 Eighth IEEE International conference on data mining* 363 – 372.

KONTOROVICH, L. A. and RAMANAN, K. (2008). Concentration inequalities for dependent random variables via the martingale method. *Ann. Probab.* **36** 2126–2158.

KOREN, Y. (2008). Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. *KDD '08* 426–434. ACM, New York, NY, USA.

KOREN, Y. and BELL, R. (2011). Advances in Collaborative Filtering. In *Recommender Systems Handbook* 145-186. Springer US.

LEE, J., KIM, S., LEBANON, G., SINGER, Y. and BENGIO, S. (2016). LLORMA: Local Low-Rank Matrix Approximation. *Journal of Machine Learning Research* **17** 1-24.

LIN, Z., GANESH, A., WRIGHT, J., WU, L., CHEN, M. and MA, Y. (2009). Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. *CAMSAP* **61**.

LINDEN, G., SMITH, B. and YORK, J. (2003). Amazon.Com Recommendations: Item-to-Item Collaborative Filtering. *IEEE Internet Computing* **7** 76–80.

LIU, Z. and VANDENBERGHE, L. (2010). Interior-Point Method for Nuclear Norm Approximation with Application to System Identification. *SIAM Journal on Matrix Analysis and Applications* **31** 1235-1256.

LIU, J., MUSIALSKI, P., WONKA, P. and YE, J. (2013a). Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35** 208 – 220.

LIU, J., MUSIALSKI, P., WONKA, P. and YE, J. (2013b). Tensor completion for estimating missing values in visual data. *IEEE Trans. Pattern Analysis and Machine Intelligence* **35** 208–220.

MACK, Y. and SILVERMAN, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **61** 405–415.

MAURER, A. and PONTIL, M. (2009). Empirical Bernstein Bounds and Sample Variance Penalization. *ArXiv e-prints*.

MAZUMDER, R., HASTIE, T. and TIBSHIRANI, R. (2010a). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research* **11** 2287–2322.

MAZUMDER, R., HASTIE, T. and TIBSHIRANI, R. (2010b). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research* **11** 2287–2322.

MU, C., HUANG, B., WRIGHT, J. and GOLDFARB, D. (2014). Square Deal: Lower Bounds and Improved Relaxations for Tensor Recovery. In *ICML* 2014.

NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research* **13** 1665–1697.

NING, X., DESROSIERS, C. and KARYPIS, G. (2015). *Recommender Systems Handbook* A Comprehensive Survey of Neighborhood-Based Recommendation Methods, 37-76. Springer US.

OH, S. and SHAH, D. (2014). Learning mixed multinomial logit model from ordinal data. In *Advances in Neural Information Processing Systems* 595–603.

ORBANZ, P. and ROY, D. M. (2015). Bayesian models of graphs, arrays and other exchangeable random structures. *IEEE transactions on pattern analysis and machine intelligence* **37** 437 – 461.

RAVI, S., PASUPATHI, P., MUTHUKUMAR, S. and KRISHNAN, N. (2013). Image in-painting techniques-A survey and analysis. In *Innovations in Information Technology (IIT), 2013 9th International Conference on* 36–41. IEEE.

ROHDE, A., TSYBAKOV, A. B. et al. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics* **39** 887–930.

SHEN, B.-H., JI, S. and YE, J. (2009). Mining discrete patterns via binary matrix factorization. In *Proceedings of the 15th ACM SIGKDD international conference* 757–766. ACM.

SIGNORETTO, M., VAN DE PLAS, R., DE MOOR, B. and SUYKENS, J. A. (2011). Tensor versus matrix completion: a comparison with application to spectral data. *IEEE Signal Processing Letters* **18** 403 – 406.

SREBRO, N., ALON, N. and JAAKKOLA, T. S. (2004). Generalization error bounds for collaborative prediction with low-rank matrices. In *Advances In Neural Information Processing Systems* 1321–1328.

SUN, J., PAPADIMITRIOU, S., LIN, C. Y., CAO, N., LIU, S. and QIAN, W. (2009). MultiVis: Content-based social network exploration through multi-way visual analysis. In *Proc. SIAM Intl. Conf. on Data Mining* 1064 – 1075.

TOMIOKA, R., SUZUKI, T., HAYASHI, K. and KASHIMA, H. (2011). Statistical performance of convex tensor decomposition. In *Advances in Neural Information Processing Systems* 972–980.

WAND, M. P. and JONES, M. C. (1994). *Kernel smoothing.* Crc Press.

WANG, J., DE VRIES, A. P. and REINDERS, M. J. T. (2006). Unifying User-based and Item-based Collaborative Filtering Approaches by Similarity Fusion. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '06* 501–508. ACM, New York, NY, USA.

XU, J., MASSOULIÉ, L. and LELARGE, M. (2014). Edge Label Inference in Generalized Stochastic Block Models: from Spectral Theory to Impossibility Results. In *COLT* 903–920.

ZHANG, Y., LEVINA, E. and ZHU, J. (2015). Estimating network edge probabilities by neighborhood smoothing. *arXiv preprint arXiv:1509.08588.*

ZHANG, Z., ELY, G., AERON, S., HAO, N. and KILMER, M. (2014). Novel methods for multilinear data completion and denoising based on tensor-SVD. In *Proc. IEEE Conf. on CVPR* 3842 – 3849.

ZHAO, Q., ZHANG, L. and CICHOCKI, A. (2015). Bayesian CP factorization of incomplete tensors with automatic rank determination. *IEEE Trans. Pattern Analysis and Machine Intelligence* **37** 1751-1763.

APPENDIX A: AUXILIARY CONCENTRATION INEQUALITIES

Chernoff bound. There are various forms of Chernoff bounds, each of which are tuned to different assumptions. The following theorem gives the bound for a sum of independent Bernoulli trials.

THEOREM A.1. *Let* $X = \sum_{i=1}^{n} X_i$, *where* $X_i = 1$ *with probability* p_i, *and* $X_i = 0$ *with probability* $1 - p_i$, *and* X_i*'s are independent. Let* $\mu = \mathbb{E}[X] = \sum_{i=1}^{n} p_i$. *Then*

1. *Upper tail:* $\mathbb{P}(X \geq (1 + \delta)\mu) \leq \exp\left(-\frac{\delta^2}{2+\delta}\mu\right)$ *for all* $\delta > 0$.
2. *Lower tail:* $\mathbb{P}(X \leq (1 - \delta)\mu) \leq \exp\left(-\frac{\delta^2}{2}\mu\right)$ *for all* $0 < \delta < 1$.

For $\delta \in (0, 1)$, we can combine the upper and lower tails to obtain the following

simpler bound:

$$\mathbb{P}\left(|X - \mu| \geq \delta\mu\right) \leq 2\exp\left(-\frac{\mu\delta^2}{3}\right) \quad \textit{for all } 0 < \delta < 1.$$

Bernstein's inequality. We will present a form of Bernstein's inequality for bounded random variables.

THEOREM A.2. *Suppose that $X_1, \ldots, X_n$ are independent random variables with zero mean, and that there exists a constant M such that $|X_i| \leq M$ with probability 1 for each i. Let $\bar{S} := \frac{1}{n}\sum_{i=1}^{n} X_i$. Let variance of each random variable X_i be bounded above by V, i.e. $\mathrm{Var}(X_i) \leq V$ for $1 \leq i \leq n$. Then for any $t \geq 0$,*

$$\mathbb{P}\left(|\bar{S}| \geq t\right) \leq 2\exp\left(-\frac{3nt^2}{6V + 2Mt}\right).$$

Maurer-Pontil Inequality. This inequality provides bounds for the concentration of sample variance.

LEMMA A.3 (Maurer-Pontil Inequality Maurer and Pontil (2009) Theorem 7). *For $n \geq 2$, let $X_1, \ldots X_n$ be independent random variables such that $X_i \in [a, b]$. Let $V(X)$ denote their sample variance, i.e., $V(X) = \frac{1}{2n(n-1)}\sum_{i,j}(X_i - X_j)^2$. Let $\sigma^2 = \mathbb{E}[V(X)]$ denote the true variance. For any $s > 0$,*

$$\mathbb{P}\left(V(X) - \sigma^2 < -s\right) \leq \exp\left(-\frac{(n-1)s^2}{2(b-a)^2\sigma^2}\right),$$

and

$$\mathbb{P}\left(V(X) - \sigma^2 > s\right) \leq \exp\left(-\frac{(n-1)s^2}{(b-a)^2\left(2\sigma^2 + s\right)}\right).$$

McDiarmid's inequality. While the previous inequalities showed concentration for specific quantities like the mean and variance, McDiarmid's inequality provides concentration results for general functions that satisfy the bounded difference condition.

THEOREM A.4. *Let $X_1, \ldots, X_n$ be independent random variables such that for each $i \in [n]$, $X_i \in \mathcal{X}_i$. Let $\xi : \prod_{i=1}^{n} \mathcal{X}_i \to \mathbb{R}$ be a function of $(X_1, \ldots, X_n)$ that satisfies $\forall i, \forall x_1, \ldots, x_n, \forall x_i' \in \mathcal{X}_i$,*

$$\left|\xi\left(x_1, \ldots, x_i, \ldots, x_n\right) - \xi\left(x_1, \ldots, x_i', \ldots, x_n\right)\right| \leq c_i.$$

Then for all $\varepsilon > 0$,

$$\mathbb{P}\left(\xi - \mathbb{E}[\xi] \geq \varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^{n} c_i^2}\right).$$

By considering the negation of the function $-\xi$ in lieu of ξ, one can obtain the same tail bound for the opposite direction.

Kontorovich-Ramanan inequality.   While the previous concentration inequalities require the independence assumption, the Kontorovich-Ramanan inequality uses the martingale method to provide concentration results for a class of dependent random sequences on a countable state space. We setup some definitions and notations used in the theorem statement. The theorem and its proof is presented in section 1.2 of Kontorovich and Ramanan (2008).

Consider a sequence of random variables $(X_1, \ldots, X_n)$ each of which takes values in a countable space $\mathcal{X}$. Let the σ-algebra be the power set of $\mathcal{X}^n$, which we denote by $\mathcal{F}$. Let $\mathbb{P}$ denote the probability measure on $(\mathcal{X}^n, \mathcal{F})$. Let $\mathcal{X}^n$ be equipped with the Hamming metric $d : \mathcal{X}^n \times \mathcal{X}^n \to \mathbb{R}_+$, defined by $d(x, y) := \sum_{i=1}^{n} \mathbb{I}\{x_i \neq y_i\}$.

The total variation distance $\|P - Q\|_{TV}$ between probability measures P and Q defined on a countable space $\mathcal{X}$ with σ-algebra $\mathcal{F}$ is defined as

$$\|P - Q\|_{TV} := \sup_{A \in \mathcal{F}} |P(A) - Q(A)| = \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)|.$$

Given $1 \leq i < j \leq n$, let x_i^j denote the subsequence $(x_i, x_{i+1}, \ldots, x_j)$. When $i = 1$, x_1^j will be written as x^j for simplicity. For any $1 \leq i < j \leq n$, $x^{i-1} \in \mathcal{X}^{i-1}$ and $w, \hat{w} \in \mathcal{X}$, define λ_{ij} to be the total variation distance between the probability measures on X_j^n conditioned on the sequences of the first i letters and differing only at the i-th position. We denote the respective probability measures by $P_{X_j^n | X^i = x^{i-1} w}$ and $P_{X_j^n | X^i = x^{i-1} \hat{w}}$. Then

$$\lambda_{ij}\left(x^{i-1}, w, \hat{w}\right) := \left\| P_{X_j^n | X^i = x^{i-1} w} - P_{X_j^n | X^i = x^{i-1} \hat{w}} \right\|_{TV}.$$

Define $\bar{\lambda}_{ij}$ to be the supremum over possible choices of w, $\hat{w}$, and sequences x^{i-1} which have positive probability,

$$\bar{\lambda}_{ij} := \sup_{\substack{x^{i-1} \in \mathcal{X}^{i-1}, w, \hat{w} \in \mathcal{X} \\ \mathbb{P}(X^i = x^{i-1} w) > 0, \mathbb{P}(X^i = x^{i-1} \hat{w}) > 0}} \lambda_{ij}\left(x^{i-1}, w, \hat{w}\right).$$

Finally, we define the mixing coefficient Δ_n to be

$$\Delta_n := \max_{i \in [n]} \left(1 + \sum_{j=i+1}^{n} \bar{\lambda}_{ij} \right).$$

We use a slightly different notation from the original paper; they used Δ_n to denote a matrix, and consider its operator norm, which we denote directly by Δ_n.

THEOREM A.5.   *Suppose $\mathcal{X}$ is a countable space, $\mathcal{F}$ is the power set of $\mathcal{X}^n$, $\mathbb{P}$ is a probability measure on $(\mathcal{X}^n, \mathcal{F})$ and $\psi : \mathcal{X}^n \to \mathbb{R}$ is a L-Lipschitz function with respect to the Hamming metric on $\mathcal{X}^n$ and has mixing coefficient Δ_n. Then for any $\varepsilon > 0$,*

$$\mathbb{P}\left(|\psi - \mathbb{E}\psi| \geq \varepsilon\right) \leq 2\exp\left(\frac{-\varepsilon^2}{2nL^2\Delta_n^2}\right).$$

APPENDIX B: GEOMETRY OF LATENT PROBABILITY MEASURE

Existence of $\phi(\cdot)$.   We show that a compact metric space admits the underestimator function which we denote by ϕ throughout this paper. For that purpose, we define the notion of regular points. We let $B(x, r)$ denote the open ball of radius r centered at x.

DEFINITION B.1.   *Let (X, d) be a compact metric space and μ be a Borel probability measure on it. A point $x \in X$ is called regular if $\mu(B(x, r)) > 0$, for all $r > 0$.*

LEMMA B.1.   *Let $R := \{x \in X : x \text{ is regular}\}$. Then R is closed, i.e., $\bar{R} = R$.*

PROOF.   Suppose that $x \in \bar{R}$. For any $r > 0$, we can find a point $x_0 \in R$ such that $d(x, x_0) < \frac{r}{2}$. We have $\mu\left(B\left(x_0, \frac{r}{2}\right)\right) > 0$ because $x_0 \in R$. It follows from $B\left(x_0, \frac{r}{2}\right) \subset B(x, r)$ that $\mu(B(x, r)) \geq \mu\left(B\left(x_0, \frac{r}{2}\right)\right) > 0$. Therefore, $x \in R$ by definition. From this, we have $\bar{R} \subset R$, hence, $\bar{R} = R$.   □

LEMMA B.2.   *If $S \subset X$ satisfies $\mu(S) > 0$, then there exists at least one regular point in the clousre of S, i.e., $R \cap \bar{S} \neq \emptyset$.*

PROOF.   Assume there exists a set S for which $\mu(S) > 0$ and $R \cap \bar{S} = \emptyset$. We know that $\bar{S}$ is compact because it is a closed subset of X. For $n = 1, 2, \ldots$, we consider a sequence of open coverings, $\mathcal{G}_n := \left\{B(x, \frac{1}{n}) : x \in \bar{S}\right\}$, each of which is the family of open balls with radius $\frac{1}{n}$ centered at $x \in \bar{S}$. Due to the compactness of $\bar{S}$, we can find finite a subcover $\mathcal{G}'_n = \left\{B_1^{(n)}, \ldots, B_{N_n}^{(n)}\right\} \subset \mathcal{G}_n$ for each n. Here, $N_n := |\mathcal{G}'_n|$ is dependent on the choice of a subcover, and hence, not uniquely defined, however, we have $N_n < \infty$.

For all $n = 1, 2, \ldots, S \subset \bar{S} \subset \cup_{i=1}^{N_n} B_i^{(n)}$ since $\mathcal{G}'_n$ is still a cover of $\bar{S}$. Therefore, we have $\mu(S) \leq \mu\left(\cup_{i=1}^{N_n} B_i^{(n)}\right) \leq N_n \max_i \mu\left(B_i^{(n)}\right)$. Let z_n denote the center of $B_{i^*}^{(n)}$ where $i^* := \arg\max_i \mu\left(B_i^{(n)}\right)$. From the construction, $z_n \in \bar{S}$. Since X is a metric space, $\bar{S}$ is not only compact, but also sequentially compact. Therefore, there is a convergent subsequence of $\{z_n\}$, which converges to a point $z^* \in \bar{S}$.

This z^* is a regular point. For any $r > 0$, we can find $n(r) > \frac{2}{r}$ such that $d(z_{n(r)}, z^*) < \frac{r}{2}$, since $\{z_n\}$ has a convergent subseqeunce. Because

$$B\left(z_{n(r)}, \frac{1}{n(r)}\right) \subset B\left(z_{n(r)}, \frac{r}{2}\right) \subset B(z^*, r),$$

we can conclude that $\mu\left(B(z^*, r)\right) \geq \mu\left(B\left(z_{n(r)}, \frac{1}{n(r)}\right)\right) \geq \frac{\mu(S)}{N_{n(r)}} > 0$. Therefore, $z^* \in R$, and $R \cap \bar{S} \neq \emptyset$.

$\square$

The following lemma ascertains that the set of regular points has the full measure, i.e., for almost every point in compact X, we can find a neighborhood of strictly positive measure.

LEMMA B.3. *Let (X, d) be a compact metric space and μ be a Borel probability measure on it. Define $R := \{x \in X : x \text{ is regular}\}$. Then R has the full measure in X, i.e., $\mu(R) = 1$.*

PROOF. For $n = 1, 2, \ldots,$, let $\mathcal{F}_n := \left\{B(x, \frac{1}{n}) : x \in X\right\}$ be the family of open balls with radius $1/n$ centered at $x \in X$, each of which forms an open cover of X. Due to the compactness of X, we can find a finite subcover $\mathcal{F}'_n \subset \mathcal{F}_n$. Furthermore, we will filter out all zero-measure balls in $\mathcal{F}'_n$ and leave only those having positive measure by defining

$$\mathcal{F}''_n := \left\{B \in \mathcal{F}'_n : \mu(B) > 0\right\}.$$

Note that the choice of $\mathcal{F}'_n$ (and hence, that of $\mathcal{F}''_n$ also) is not unique. We suppose $\mathcal{F}''_n$ is fixed from now on, but the following arguments hold for any choice of subcovering $\mathcal{F}''_n$.

Next, we claim that $\cup_{B \in \mathcal{F}''_n} B \subset \cup_{x \in R} B\left(x, \frac{2}{n}\right) = R + B\left(0, \frac{2}{n}\right)$. From the construction above, $\mu(B) > 0$ for all $B \in \mathcal{F}''_n$. By Lemma B.2, there exists a regular point $x_B \in R \cap \bar{B}$. Since $\mathcal{F}''_n$ is a subfamily of $\mathcal{F}_n$, $B = B\left(x_0, \frac{1}{n}\right)$ for some $x_0 \in X$. Therefore, for every $z \in B$, $d(x_B, z) \leq d(x_B, x_0) + d(x_0, z) < \frac{2}{n}$ by triangle inequality. It follows that $B \subset B\left(x_B, \frac{2}{n}\right)$. Finally, $\cup_{B \in \mathcal{F}''_n} B \subset \cup_{B \in \mathcal{F}''_n} B\left(x_B, \frac{2}{n}\right) \subset \cup_{x \in R} B\left(x, \frac{2}{n}\right)$.

For each $n = 1, 2, \ldots,$ $\mu\left(\cup_{B \in \mathcal{F}''_n} B\right) = 1$. Because $\cup_{B \in \mathcal{F}'_n} B = \left(\cup_{B \in \mathcal{F}''_n} B\right) \cup \left(\cup_{B \in \mathcal{F}'_n \setminus \mathcal{F}''_n} B\right)$, we have $\mu\left(\cup_{B \in \mathcal{F}'_n} B\right) \leq \mu\left(\cup_{B \in \mathcal{F}''_n} B\right) + \mu\left(\cup_{B \in \mathcal{F}'_n \setminus \mathcal{F}''_n} B\right)$. Since $\mathcal{F}'_n \setminus \mathcal{F}''_n$ consists of a finite number of balls each of which has measure zero, $\mu\left(\cup_{B \in \mathcal{F}'_n \setminus \mathcal{F}''_n} B\right) = 0$. Therefore, $\mu\left(\cup_{B \in \mathcal{F}'_n} B\right) \leq \mu\left(\cup_{B \in \mathcal{F}''_n} B\right)$.

Combining these with the fact that $X \subset \cup_{B \in \mathcal{F}'_n} B$, we can conclude that

$$\mu\left(R + B\left(0, \frac{2}{n}\right)\right) \geq \mu\left(\cup_{B \in \mathcal{F}''_n} B\right) \geq \mu\left(\cup_{B \in \mathcal{F}'_n} B\right) \geq \mu(X) = 1.$$

Because $\bar{R} = \cap_{n=1}^{\infty} \left(R + B\left(0, \frac{2}{n}\right) \right)$ and $\left\{ R + B\left(0, \frac{2}{n_1}\right) \right\}_{n=1}^{\infty}$ consists of a decreasing nested set,

$$
\begin{aligned}
\mu\left(\bar{R}\right) &= \mu\left(\lim_{N\to\infty} \cap_{n=1}^{N} \left(R + B\left(0, \frac{2}{n}\right) \right) \right) \\
&= \lim_{N\to\infty} \mu \cap_{n=1}^{N} \left(R + B\left(0, \frac{2}{n}\right) \right) \\
&= \lim_{N\to\infty} \mu\left(R + B\left(0, \frac{2}{N}\right) \right) \\
&= 1.
\end{aligned}
$$

$\square$

LEMMA B.4. *Let (X, d) be a compact metric space, and μ be a Borel probability measure on it. Then there exists a function ϕ on (X, d, μ) which satisfies*

1. *$\phi : \mathbb{R}_{++} \to (0, 1]$,*
2. *ϕ is nondecreasing, and*
3. *$\mu\left(B(x, r)\right) \geq \phi(r)$, for all regular $x \in X$ and for all $r > 0$.*

PROOF. We define $\phi(r) := \inf_{x\in R} \mu\left(B(x, r)\right)$ and will show this ϕ satisfies the desired properties.

For any given $r > 0$, consider the open cover $\mathcal{G}_r := \left\{ B\left(x, \frac{r}{2}\right) : x \in R \right\}$ of R. We know that R is compact because it is a closed subset of X, which is compact. Therefore, we can find a finite subcover $\mathcal{G}_r' = \left\{ B\left(z_i, \frac{r}{2}\right) \right\}_{i=1}^{N(r)} \subset \mathcal{G}_r$ of R. For every $x \in R$, there should exist z such that $B\left(z, \frac{r}{2}\right) \in \mathcal{G}_r'$ and $d(x, z) < \frac{r}{2}$ because $\mathcal{G}_r'$ is a cover of R. Therefore, $\mu\left(B(x, r)\right) \geq \mu\left(B\left(z, \frac{r}{2}\right)\right) \geq \min_{i=1,\ldots,N(r)} \mu\left(B\left(z_i, \frac{r}{2}\right)\right) > 0$.

Suppose that $r_1 > r_2 > 0$. By definition $\phi(r_1) = \inf_x \mu\left(B(x, r_1)\right)$ and hence, for any $\varepsilon > 0$, there exists $x_0(\varepsilon) \in R$ such that $\mu\left(B\left(x_0(\varepsilon), r_1\right)\right) < \phi(r_1) + \varepsilon$. For every $x \in R$, we have $\mu\left(B(x, r_2)\right) \leq \mu\left(B(x, r_1)\right)$ because $B\left(x, r_2\right) \subset B\left(x, r_1\right)$. It follows that $\phi(r_2) \leq \mu\left(B\left(x_0(\varepsilon), r_2\right)\right) \leq \mu\left(B\left(x_0(\varepsilon), r_1\right)\right) < \phi(r_1) + \varepsilon$. By taking the limit $\varepsilon \to 0$, we can conclude that $\phi(r_2) \leq \phi(r_1)$. Therefore, ϕ is nondecreasing.

The last property is obvious from the definition of ϕ: for all $x \in R$, $\mu\left(B(x, r)\right) \geq \inf_{x\in R} \mu\left(B(x, r)\right) = \phi(r)$. $\square$

Let us consider two examples.

EXAMPLE 1. *Suppose $P_{\mathcal{X}_1}$ is a uniform measure over a unit cube in d dimen-*

sional Euclidean space. Then

$$\inf_{x_0 \in \mathcal{X}_1} P_{\mathcal{X}_1} \left(d_{\mathcal{X}_1}(\mathbf{x}, x_0) \le r \right) = \min \left\{ 1, \left(\frac{r}{2} \right)^d \right\}.$$

EXAMPLE 2. *Suppose $P_{\mathcal{X}_1}$ is supported only on a finite number of points, i.e., $|supp P_{\mathcal{X}_1}| = N$. Then for any $r \ge 0$,*

$$\inf_{x_0 \in \mathcal{X}_1} P_{\mathcal{X}_1} \left(d_{\mathcal{X}_1}(\mathbf{x}, x_0) \le r \right) \ge \min_{i=1,\dots,N} \phi_1 \left(x_i \right).$$

77 MASSACHUSETTS AVENUE, CAMBRIDGE, MA 02139,
E-MAIL: celee@mit.edu; liyihua@mit.edu; devavrat@mit.edu; dgsong@mit.edu