A Notation

Here we state the notation that will be used throughout this appendix. For $n \in \mathbb{Z}^+$, $[n]$ denotes the set $\{1, \ldots, n\}$. If X is a matrix, then $\|X\|_2$, $\|X\|_F$, and $\|X\|_*$ are respectively the operator, frobenius, and nuclear norms of X. $\sigma_i(X)$ is the i^{th} largest singular value of X. For a matrix $U \in \mathbb{R}^{m\times r}$ with orthonormal columns, we will refer to U as a *matrix* and *subspace* interchangeably, where the subspace is the space in $\mathbb{R}^m$ spanned by the columns of U; $P_U = UU^\top$ is the projection operator onto the subspace U. We use $d(U, \hat{U}) = \left\|P_U - P_{\hat{U}}\right\|_F$ as a metric for subspaces.

B Comparison of slice rank to existing tensor ranks

There are already many definitions of rank for tensors that have been studied. The two most common, on which the tensor recovery literature has focused, are referred to here as CP rank and Tucker rank. We review the canonical definitions of these ranks here. See Kolda and Bader (2009) for a more thorough treatment of these concepts.

CP rank The CP rank of a tensor relates to its orthogonal decompositions. A rank-one tensor is any tensor $M \in \mathbb{R}^{m\times m\times n}$ that is the tensor product of three vectors, i.e. $M = u \otimes v \otimes w$ for some $u \in \mathbb{R}^m$, $v \in \mathbb{R}^m$, and $w \in \mathbb{R}^n$, or equivalently, $M^j_{i_1,i_2} = u_{i_1} v_{i_2} w_j$. For any tensor M, we denote its CP rank as $\text{CP}(M)$, which is the minimum number r such that M can be expressed as the sum of r rank-one tensors.

Tucker rank The Tucker rank of a tensor M, denoted $\text{Tucker}(M)$, is the vector (r_1, r_2, r_3), where r_d is the rank of its mode-d unfolding. This relates to its higher order singular value decomposition: given a tensor of Tucker rank (r_1, r_2, r_3), there exist vectors $u^1, \ldots, u^{r_1} \in \mathbb{R}^m$, $v^1, \ldots, v^{r_2} \in \mathbb{R}^m$, and $w^1, \ldots, w^{r_3} \in \mathbb{R}^n$, and a smaller tensor $S \in \mathbb{R}^{r_1\times r_2\times r_3}$, such that $M = \sum_{k_1}^{r_1} \sum_{k_2}^{r_2} \sum_{k_3}^{r_3} S^{k_3}_{k_1,k_2} u^{k_1} \otimes v^{k_2} \otimes w^{k_3}$.

The following Proposition establishes the equivalent definitions for the CP rank and Tucker rank that were referred to in §2.1 in the main text.

Proposition 2. *(a) $M \in \mathbb{R}^{m\times m\times n}$ has CP rank at most r if and only if there exist matrices $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{m\times r}$ such that each slice of M, M^j can be decomposed as $M^j = US^jV^\top$, where each S^j is a* diagonal *matrix.*

(b) $M \in \mathbb{R}^{m\times m\times n}$ has Tucker rank component-wise at most (r, r, l) if and only if there exist matrices $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{m\times r}$ such that each slice of M, M^j can be decomposed as $M^j = US^jV^\top$, wherein $\text{rank}\left(\text{span}\left(s^1, s^2, \ldots, s^n\right)\right) = l$ *where $s^j \in \mathbb{R}^{r^2}$ is a vectorization of S^j.*

Proof. We first prove part (a). By definition, M has CP rank at most r if and only if there exist vectors $u^1, \ldots, u^r \in \mathbb{R}^m$, $v^1, \ldots, v^r \in \mathbb{R}^m$, and $w^1, \ldots, w^r \in \mathbb{R}^n$, such that each entry of M can be expressed as

$$M^j_{i_1,i_2} = \sum_{k=1}^{r} u^k_{i_1} v^k_{i_2} w^k_j.$$

Let U and V be the matrices with columns $u^1, \ldots, u^r$ and $v^1, \ldots, v^r$, respectively. Then we can equivalently write the above expression for slices as

$$M^j = \sum_{k=1}^{r} w^k_j (u^k v^{k\top}) = US^jV^\top,$$

where S^j is the diagonal matrix, whose diagonal elements of $w^1_j, \ldots, w^r_j$.

Now we prove part (b). By definition, M has Tucker rank component-wise at most (r, r, l) if and only if there exist vectors $u^1, \ldots, u^r \in \mathbb{R}^m$, $v^1, \ldots, v^r \in \mathbb{R}^m$, and $w^1, \ldots, w^l \in \mathbb{R}^n$, and a tensor $S \in \mathbb{R}^{r\times r\times l}$, such that $M = \sum_{k_1}^{r} \sum_{k_2}^{r} \sum_{k_3}^{l} S^{k_3}_{k_1,k_2} u^{k_1} \otimes v^{k_2} \otimes w^{k_3}$. Equivalently, each entry of M can be expressed as

$$M^j_{i_1,i_2} = \sum_{k_1=1}^{r} \sum_{k_2=1}^{r} \sum_{k_3=3}^{l} S^{k_3}_{k_1,k_2} u^{k_1}_{i_1} v^{k_2}_{i_2} w^{k_3}_j.$$

Let U and V be the matrices with columns $u^1, \ldots, u^r$ and $v^1, \ldots, v^r$, respectively. Then we can equivalently write the above expression for slices as

$$M^j = \sum_{k_3=1}^{l} w_j^{k_3} \sum_{k_1=1}^{r} \sum_{k_2=1}^{r} S_{k_1,k_2}^{k_3} (u^{k_1} v^{k_2 \top}) = \sum_{k_3=1}^{l} w_j^{k_3} (U S^{k_3} V^\top) = U \left(\sum_{k_3=1}^{l} w_j^{k_3} S^{k_3} \right) V^\top.$$

Let $W^j = \sum_{k_3=1}^{l} w_j^{k_3} S^{k_3}$ for each j. Then each W^j is a linear combination of some matrices $S^1, \ldots, S^l$, or equivalently, $W^1, \ldots, W^n$, when viewed as vectors in $\mathbb{R}^{r^2}$, span an l-dimensional subspace. □

C Proof of Proposition 1

Our proof follows a standard Bayesian argument for minimax lower bounds; for example, see the proof of Theorem 1.2 in Chatterjee (2014). We will separately show that $\mathrm{MSE}(\hat{M}) \geq C(r^2/m^2)$ and $\mathrm{MSE}(\hat{M}) \geq C(r/mn)$. We first give a detailed proof that $\mathrm{MSE}(\hat{M}) \geq C(r^2/m^2)$. For each ground truth slice M^k, let the elements sitting in the first r rows and first r columns be drawn independently from a uniform distribution, and the remaining elements set equal to 0:

$$M_{ij}^k \sim \begin{cases} \mathrm{Uniform}[0,1] & \text{if } i \leq r \text{ and } j \leq r \\ 0 & \text{if } i > r \text{ or } j > r \end{cases}. \tag{5}$$

Note that all slices of M share the same column and row spaces, both with dimension at most r. Finally, conditional on M, each entry $X_{i,j}^k$ of X is drawn from the following two point distribution:

$$X_{ij}^k \sim \mathrm{Ber}(M_{ij}^k). \tag{6}$$

Then for each $i \leq r$ and $j \leq r$, we have

$$\begin{aligned} \mathsf{E}\left[\mathrm{Var}\left(M_{ij}^k | X\right)\right] &= \mathrm{Var}\left(M_{ij}^k\right) - \mathrm{Var}\left(\mathsf{E}\left[M_{ij}^k | X\right]\right) \\ &= \mathrm{Var}\left(M_{ij}^k\right) - \mathrm{Var}\left(\mathsf{E}\left[M_{ij}^k | X_{ij}^k\right]\right) \\ &= \mathrm{Var}\left(M_{ij}^k\right) - \mathrm{Var}\left(\frac{1 + X_{ij}^k}{3}\right) \\ &= \frac{1}{12} - \frac{1}{36} = \frac{1}{18} \end{aligned} \tag{7}$$

The first equality is the law of total variance. For the second equality, observe that M_{ij}^k is independent of all entries of X except for its corresponding entry X_{ij}^k. The third equality comes from the fact that, having defined M_{ij}^k to be distributed as Uniform$[0,1]$ (or equivalently Beta$(1,1)$), its distribution conditional on X_{ij}^k is Beta$(1 + X_{ij}^k, 2 - X_{ij}^k)$.

Then for any estimator $\hat{M}$, the definition of variance implies that

$$\mathsf{E}\left[\left(\hat{M}_{ij}^k - M_{ij}^k\right)^2 \middle| X\right] \geq \mathrm{Var}\left(M_{ij}^k | X\right).$$

Taking expectations of both sides, and applying (7), we have

$$\mathsf{E}\left[\left(\hat{M}_{ij}^k - M_{ij}^k\right)^2\right] \geq \frac{1}{18}.$$

The proof concludes by summing both sides over all entries of M in the first r rows and first r columns (i.e. nr^2 entries in total) and dividing by $m^2 n$.

A nearly identical argument shows that $\mathrm{MSE}(\hat{M}) \geq C(r/mn)$. For the first slice M^1, let the elements in the first r rows be drawn independently from a uniform distribution, and the remaining elements set equal to 0:

$$M_{ij}^1 \sim \begin{cases} \mathrm{Uniform}[0,1] & \text{if } i \leq r \\ 0 & \text{if } i > r \end{cases}. \tag{8}$$

Set the entries of the remaining slices equal to the corresponding entries in the first slice, i.e. $M^k_{ij} = M^1_{ij}$ for all k. Once again, conditional on M, each entry X^k_{ij} is drawn independently from the distribution in (6), so while the slices of M are copies of each other, the slices of X are not. Then for each $i \le r$, we have

$$\begin{aligned}\mathsf{E}\left[\mathrm{Var}\left(M^k_{ij}|X\right)\right] &= \mathrm{Var}\left(M^k_{ij}\right) - \mathrm{Var}\left(\mathsf{E}\left[M^k_{ij}|X\right]\right)\\ &= \mathrm{Var}\left(M^k_{ij}\right) - \mathrm{Var}\left(\mathsf{E}\left[M^k_{ij}|X^1_{ij},\dots X^n_{ij}\right]\right)\\ &= \mathrm{Var}\left(M^k_{ij}\right) - \mathrm{Var}\left(\frac{1+X^1_{ij}+\cdots+X^n_{ij}}{n+2}\right)\\ &= \frac{1}{12} - \frac{n}{12(n+2)} = \frac{1}{6(n+2)} \end{aligned} \tag{9}$$

Again, the first equality is the law of total variance, and for the second equality, observe that M^k_{ij} is independent of all entries of X except for the $(i,j)^{\text{th}}$ entry of each slice. For the third equality, the distribution of M^k_{ij} conditional on $X^1_{ij},\dots,X^n_{ij}$ is $\mathrm{Beta}(1+X^1_{ij}+\cdots+X^n_{ij}, n+1-(X^1_{ij}+\cdots+X^n_{ij}))$.

Therefore, for any estimator $\hat{M}$ we have

$$\mathsf{E}\left[\left(\hat{M}^k_{ij} - M^k_{ij}\right)^2\right] \ge \frac{1}{6(n+2)}.$$

The proof concludes by summing both sides over all entries of M in the first r rows (i.e. nmr entries in total) and dividing by $m^2 n$. □

D Proof of Theorem 1

We will present a more general version of Theorem 1 that relaxes the balanced noise assumption and reflects the recovery error caused by 'unbalanced' noise.

To proceed, we need to precisely quantify the concept of 'unbalanced' noise. Recall that if v is the tensor whose entries are the variances of the corresponding entries of ϵ, i.e. $v^k_{ij} = \mathsf{E}[(\epsilon^k_{ij})^2]$, then ϵ is balanced if the row-sums of $v_{(1)}$ are equal and the row-sums of $v_{(2)}$ are equal. An equivalent way to state this assumption is $\mathsf{E}[\epsilon_{(1)}\epsilon_{(1)}^\top] = \rho_1 I_m$ and $\mathsf{E}[\epsilon_{(2)}\epsilon_{(2)}^\top] = \rho_2 I_m$ for some constants ρ_1 and ρ_2, where I_m is the $m \times m$ identity matrix. To see this, note that the off-diagonal elements of $\mathsf{E}[\epsilon_{(1)}\epsilon_{(1)}^\top]$ and $\mathsf{E}[\epsilon_{(2)}\epsilon_{(2)}^\top]$ are always equal to zero when the noise terms are independent, and the diagonal elements are exactly the row sums of $v_{(1)}$ and $v_{(2)}$, so the balanced noise assumption states that $\mathsf{E}[\epsilon_{(1)}\epsilon_{(1)}^\top]$ and $\mathsf{E}[\epsilon_{(2)}\epsilon_{(2)}^\top]$ are multiples of the identity matrix.

For general, possibly unbalanced noise, it turns out that the appropriate quantities to measure the level of 'unbalance' in the noise are

$$\min_\rho \frac{1}{m}\left\|\mathsf{E}[\epsilon_{(1)}\epsilon_{(1)}^\top] - \rho I_m\right\|_F^2 \quad \text{and} \quad \min_\rho \frac{1}{m}\left\|\mathsf{E}[\epsilon_{(2)}\epsilon_{(2)}^\top] - \rho I_m\right\|_F^2.$$

These quantities measure how far $\mathsf{E}[\epsilon_{(1)}\epsilon_{(1)}^\top]$ and $\mathsf{E}[\epsilon_{(2)}\epsilon_{(2)}^\top]$ are from a multiple of the identity matrix. One nice interpretation is that the quantities correspond to population variances, one each for the row-sums of $v_{(1)}$ and the row-sums of $v_{(2)}$. We denote the maximum of these two quantities as δ^2, and can now state our more general result. We will in fact prove Theorem 2 below. Theorem 1 is just the special case where $\delta = 0$.

Theorem 2. *Assume the entries of M lie in $[-1,1]$. Suppose the entries of ϵ are independent, mean-zero, and $\mathsf{E}[\epsilon^6_{ij}] \le K^6$. Then there exists a universal constant c such that for the slice learning algorithm (without trimming),*

$$\mathrm{SMSE}(\hat{M}) \le c\left[\frac{K^2 r^2}{m^2} + \frac{K^2(K^4+1)r^2}{\gamma_M^2 mn} + \frac{(K^2+1)r^2\delta^2}{\gamma_M^2 m^3 n^2}\right]$$

The proof of Theorem 2 involves two steps, corresponding to the two stages of the algorithm: learning subspaces and projection. In the first step, we show that we are able to closely estimate the column and row spaces, and in the second step, we show that if our estimates of the 'true' column and row spaces are close, then our estimate of each slice is close.

D.1 Step 1: Column and Row Space Estimation

To estimate the column space (and similarly the row space), we take the top column singular vectors of $X_{(1)} = M_{(1)} + \epsilon_{(1)}$, so it is important to understand the extent to which $\epsilon_{(1)}$ changes the singular vectors of $M_{(1)}$. Lemma 1 bounds the error of this step. The first result in Lemma 1 is an upper bound on $\mathsf{E}\left[d(U,\hat{U})^2\right]$, which is the expected error of our subspace estimate. Because of the decomposition we make later on, we also need to bound $\mathsf{E}\left[\left\|\epsilon^k\right\|_F^2 d(U,\hat{U})^2\right]$ for any slice of the noise tensor ϵ^k. It would be tempting to say that, since $\mathsf{E}\left[\left\|\epsilon^k\right\|_F^2\right] \le K^2m^2$, we can multiply the first result by K^2m^2, but unfortunately ϵ^k and $d(U,\hat{U})$ are not independent. The second result in Lemma 1 states that this bound still holds.

Lemma 1. *Let $M \in \mathbb{R}^{m\times mn}$ be a matrix with column space $U \in \mathbb{R}^{m\times r}$. Suppose $\epsilon \in \mathbb{R}^{m\times mn}$ is a random matrix with independent elements, where each element ϵ_{ij} is mean-zero and $\mathsf{E}[\epsilon_{ij}^6] \le K^6$. Let $X = M + \epsilon$, and let $\hat{U} \in \mathbb{R}^{m\times r}$ be the column singular vectors of X corresponding to its largest r singular values. Then taking expectation over ϵ, we have*

$$\mathsf{E}\left[d(U,\hat{U})^2\right] \le 24\frac{4K^2m\left\|M\right\|_F^2 + K^4m^3n + \min_\rho\left\|\mathsf{E}[\epsilon\epsilon^\top] - \rho I_m\right\|_F^2}{\sigma_r^4(M)}, \quad \textit{and}$$

$$\mathsf{E}\left[\left\|\epsilon^1\right\|_F^2 d(U,\hat{U})^2\right] \le 24K^2m^2\frac{4K^2m\left\|M\right\|_F^2 + K^4m^3n + \min_\rho\left\|\mathsf{E}[\epsilon\epsilon^\top] - \rho I_m\right\|_F^2}{\sigma_r^4(M)},$$

where ϵ^1 is any $m\times m$ submatrix of ϵ.

The proof of Lemma 1 relies on the Davis-Kahan Theorem (Davis and Kahan (1970)), via a recent extension by Yu et al. (2015), which we reproduce as Lemma 2. Note that Lemma 2 is a statement about symmetric matrices, which we adapt to our setting where the matrices are not symmetric or even square; Yu et al. (2015) also show a version of Lemma 2 for rectangular matrices that is a similar modification to Wedin's Theorem (Wedin (1972)), but applying that directly would not yield as strong a bound as Lemma 1. However, this stronger bound requires the noise to be balanced.

Lemma 2 (Davis-Kahan Variant; Yu et al. (2015), Theorem 2)**.** *Suppose S and $\hat{S}$ are symmetric matrices, and let U and $\hat{U}$ be the eigenvectors corresponding to the r largest eigenvalues of S and $\hat{S}$, respectively. Let $\lambda_r(S)$ and $\lambda_{r+1}(S)$ be the r^{th} and $r+1^{th}$ largest eigenvalues of S. Then assuming $\lambda_r(S) \ne \lambda_{r+1}(S)$, we have*

$$d(U,\hat{U}) \le \frac{2\sqrt{2}\left\|S - \hat{S}\right\|_F}{\lambda_r(S) - \lambda_{r+1}(S)}.$$

Proof. First note that the column singular vectors of M and X are identical to the eigenvectors of $MM^\top$ and $XX^\top$, respectively, and further, the eigenvectors of $XX^\top - \rho I_m$ are the same for any $\rho \in \mathbb{R}$. Thus, Lemma 2 can be applied directly with $S = MM^\top$, and $\hat{S} = XX^\top - \rho I_m$ for any $\rho \in \mathbb{R}$, and $\lambda_r(MM^\top) - \lambda_{r+1}(MM^\top) = \sigma_r(M)^2 - \sigma_{r+1}(M)^2 = \sigma_r(M)^2$:

$$d(U,\hat{U})^2 \le \frac{8\min_\rho\left\|MM^\top - (XX^\top - \rho I_m)\right\|_F^2}{\sigma_r^4(M)}. \tag{10}$$

To upper bound the numerator, we make the following decomposition:

$$\begin{aligned}\min_\rho\left\|MM^\top - (XX^\top - \rho I_m)\right\|_F &\le 2\left\|M\epsilon^\top\right\|_F + \min_\rho\left\|\epsilon\epsilon^\top - \rho I_m\right\|_F \\ &\le 2\left\|M\epsilon^\top\right\|_F + \left\|\epsilon\epsilon^\top - \mathsf{E}[\epsilon\epsilon^\top]\right\|_F + \min_\rho\left\|\mathsf{E}[\epsilon\epsilon^\top] - \rho I_m\right\|_F,\end{aligned}$$

and since $(a+b+c)^2 \le 3(a^2+b^2+c^2)$ for any $a,b,c \in \mathbb{R}$, we have

$$\min_\rho\left\|MM^\top - (XX^\top - \rho I_m)\right\|_F^2 \le 3\left(4\left\|M\epsilon^\top\right\|_F^2 + \left\|\epsilon\epsilon^\top - \mathsf{E}[\epsilon\epsilon^\top]\right\|_F^2 + \min_\rho\left\|\mathsf{E}[\epsilon\epsilon^\top] - \rho I_m\right\|_F^2\right).$$

We have decomposed the numerator of (10) into three terms. The last term is a deterministic quantity. The proof concludes by bounding the expectation of the first two terms. All of the following calculations proceed in the same manner: the first equality is a rewriting of expressions in expanded form, the second equality comes from setting any summands with a lone $\mathsf{E}[\epsilon_{ij}]$ to zero, and the inequality applies the fact that $\mathsf{E}[\epsilon_{ij}^6] \leq K^6$.

$$\mathsf{E}\left[\left\|M\epsilon^\top\right\|_F^2\right] = \sum_{i_1\in[m],i_2\in[m]} \mathsf{E}\left[\left(\sum_{j\in[mn]} M_{i_1j}\epsilon_{i_2j}\right)^2\right] = \sum_{i_1\in[m],i_2\in[m]} \sum_{j\in[mn]} M_{i_1j}^2 \mathsf{E}\left[\epsilon_{i_2j}^2\right] \leq K^2 m \left\|M\right\|_F^2 \tag{11}$$

$$\begin{aligned}\mathsf{E}\left[\left\|\epsilon^1\right\|_F^2 \left\|M\epsilon^\top\right\|_F^2\right] &= \mathsf{E}\left[\left\|\epsilon^1\right\|_F^2 \sum_{i_1,i_2\in[m]} \left(\sum_{j\in[mn]} M_{i_1j}\epsilon_{i_2j}\right)^2\right] \\ &= \sum_{i_1,i_2\in[m]} \sum_{j\in[mn]} M_{i_1j}^2 \mathsf{E}\left[\epsilon_{i_2j}^2 \left\|\epsilon^1\right\|_F^2\right] \\ &\leq K^4 m^3 \left\|M\right\|_F^2\end{aligned} \tag{12}$$

$$\begin{aligned}\mathsf{E}\left[\left\|\epsilon\epsilon^\top - \mathsf{E}[\epsilon\epsilon^\top]\right\|_F^2\right] &= \sum_{i\in[m]} \mathrm{Var}\left(\sum_{j\in[mn]} \epsilon_{ij}^2\right) + \sum_{i_1\in[m],i_2\in[m],i_1\neq i_2} \mathsf{E}\left[\left(\sum_{j\in[mn]} \epsilon_{i_1j}\epsilon_{i_2j}\right)^2\right] \\ &= \sum_{i\in[m]} \sum_{j\in[mn]} \mathrm{Var}\left[\epsilon_{ij}^2\right] + \sum_{i_1\in[m],i_2\in[m],i_1\neq i_2} \sum_{j\in[mn]} \mathsf{E}\left[\epsilon_{i_1j}^2\right] \mathsf{E}\left[\epsilon_{i_2j}^2\right] \\ &\leq K^4 m^2 n + K^4 m^2 (m-1) n = K^4 m^3 n\end{aligned} \tag{13}$$

(14)

$$\begin{aligned}\mathsf{E}\left[\left\|\epsilon^1\right\|_F^2 \left\|\epsilon\epsilon^\top - \mathsf{E}[\epsilon\epsilon^\top]\right\|_F^2\right] &= \mathsf{E}\left[\left\|\epsilon^1\right\|_F^2 \sum_{i\in[m]} \left(\sum_{j\in[mn]} \epsilon_{ij}^2 - \mathsf{E}\left[\epsilon_{ij}^2\right]\right)^2 + \left\|\epsilon^1\right\|_F^2 \sum_{i_1,i_2\in[m],i_1\neq i_2} \left(\sum_{j\in[mn]} \epsilon_{i_1j}\epsilon_{i_2j}\right)^2\right] \\ &= \mathsf{E}\left[\left\|\epsilon^1\right\|_F^2 \sum_{i\in[m]} \sum_{j\in[mn]} \left(\epsilon_{ij}^2 - \mathsf{E}\left[\epsilon_{ij}^2\right]\right)^2 + \left\|\epsilon^1\right\|_F^2 \sum_{i_1,i_2\in[m],i_1\neq i_2} \sum_{j\in[mn]} \epsilon_{i_1j}^2\epsilon_{i_2j}^2\right] \\ &\leq K^6 m^4 n + K^6 m^4 (m-1) n = K^6 m^5 n\end{aligned}$$

Combining (11) and (13) completes the first result, and combining (12) and (14), along with the fact that $\mathsf{E}\left[\left\|\epsilon^1\right\|_F^2\right] \leq K^2 m^2$, completes the second.

□

D.2 Step 2: Projection onto Estimated Spaces

Lemma 3 decomposes the error of the projection step in terms of the error of our column and row space estimates. For any slice M^k, our estimate of this slice is the projection of X^k onto the estimated subspaces $\hat{U}$ and $\hat{V}$, i.e. $P_{\hat{U}} M^k P_{\hat{V}} + P_{\hat{U}} \epsilon^k P_{\hat{V}}$. If $\hat{U}$ and $\hat{V}$ are close to U and V, then $P_{\hat{U}} M^k P_{\hat{V}} \approx P_U M^k P_V = M^k$. Furthermore, since $\hat{U}$ and $\hat{V}$ are low-dimensional subspaces, $P_{\hat{U}} \epsilon^k P_{\hat{V}}$ will be small (this argument needs to be made carefully as $\hat{U}$ and $\hat{V}$ depend on ϵ^k).

Lemma 3. *Let $M^1 \in \mathbb{R}^{m\times m}$ be a matrix with column and row spaces $U, V \in \mathbb{R}^{m\times r}$. Let $\epsilon^1 \in \mathbb{R}^{m\times m}$ be a random matrix, and let $\hat{U}, \hat{V} \in \mathbb{R}^{m\times r}$ be random subspaces, where none of these variables are required to be independent. If $\hat{M}^1 = P_{\hat{U}}(M^1 + \epsilon^1)P_{\hat{V}}$, then taking expectation over $\epsilon^1, \hat{U}$, and $\hat{V}$:*

$$\mathsf{E}\left[\left\|\hat{M}^1 - M^1\right\|_F^2\right] \leq 9\mathsf{E}\left[\left\|P_U\epsilon^1P_V\right\|_F^2\right] + 3\left\|M^1\right\|_F^2 \mathsf{E}\left[4d(U,\hat{U})^2 + d(V,\hat{V})^2\right] + 9\mathsf{E}\left[\left\|\epsilon^1\right\|_F^2 \left(4d(U,\hat{U})^2 + d(V,\hat{V})^2\right)\right].$$

Proof. We begin by making the following decomposition, where the first two inequalities rely on the sub-multiplicative and sub-additive properties of the frobenius norm, and the first inequality also relies on the fact that $\left\|P_{\hat{V}} - P_V\right\|_2 \leq 1$. The final inequality comes from $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$.

$$\begin{aligned}\left\|\hat{M}^1 - M^1\right\|_F^2 &= \left\|P_{\hat{U}}(M^1+\epsilon^1)P_{\hat{V}} - M^1\right\|_F^2 \\ &= \left\|[P_U + (P_{\hat{U}} - P_U)]M^1[P_V + (P_{\hat{V}} - P_V)] - M^1 + P_{\hat{U}}\epsilon^1 P_{\hat{V}}\right\|_F^2 \\ &= \left\|M^1(P_{\hat{V}} - P_V) + (P_{\hat{U}} - P_U)\left[M^1 + M^1(P_{\hat{V}} - P_V)\right] + P_{\hat{U}}\epsilon^1 P_{\hat{V}}\right\|_F^2 \\ &\leq \left(\left\|M^1(P_{\hat{V}} - P_V)\right\|_F + 2\left\|(P_{\hat{U}} - P_U)M^1\right\|_F + \left\|P_{\hat{U}}\epsilon^1 P_{\hat{V}}\right\|_F\right)^2 \\ &\leq \left(\left\|M^1\right\|_F\left(2d(U,\hat{U}) + d(V,\hat{V})\right) + \left\|P_{\hat{U}}\epsilon^1 P_{\hat{V}}\right\|_F\right)^2 \\ &\leq 3\left\|M^1\right\|_F^2\left(4d(U,\hat{U})^2 + d(V,\hat{V})^2\right) + 3\left\|P_{\hat{U}}\epsilon^1 P_{\hat{V}}\right\|_F^2.\end{aligned}$$

The last term is decomposed further in a similar way:

$$\begin{aligned}\left\|P_{\hat{U}}\epsilon^1 P_{\hat{V}}\right\|_F^2 &= \left\|[P_U + (P_{\hat{U}} - P_U)]\epsilon^1[P_V + (P_{\hat{V}} - P_V)]\right\|_F^2 \\ &\leq \left(\left\|P_U\epsilon^1 P_V\right\|_F + \left\|\epsilon^1(P_{\hat{V}} - P_V)\right\|_F + 2\left\|(P_{\hat{U}} - P_U)\epsilon^1\right\|_F\right)^2 \\ &\leq \left(\left\|P_U\epsilon^1 P_V\right\|_F + \left\|\epsilon^1\right\|_F\left(2d(U,\hat{U}) + d(V,\hat{V})\right)\right)^2 \\ &\leq 3\left\|P_U\epsilon^1 P_V\right\|_F^2 + 3\left\|\epsilon^1\right\|_F^2\left(4d(U,\hat{U})^2 + d(V,\hat{V})^2\right).\end{aligned}$$

We conclude the proof by taking expectations. □

D.3 Final Steps

We are now ready to conclude the proof. Fix any $k \in [n]$. Lemma 3 first gives us:

$$\mathsf{E}\left[\left\|\hat{M}^k - M^k\right\|_F^2\right] \leq 9\mathsf{E}\left[\left\|P_U\epsilon^k P_V\right\|_F^2\right] + 3\left\|M^k\right\|_F^2\mathsf{E}\left[4d(U,\hat{U})^2 + d(V,\hat{V})^2\right] + 9\mathsf{E}\left[\left\|\epsilon^k\right\|_F^2\left(4d(U,\hat{U})^2 + d(V,\hat{V})^2\right)\right].$$

The first term is bounded as follows:

$$\mathsf{E}\left[\left\|P_U\epsilon^k P_V\right\|_F^2\right] = \mathsf{E}\left[\left\|UU^\top\epsilon^k VV^\top\right\|_F^2\right] = \mathsf{E}\left[\left\|U^\top\epsilon^k V\right\|_F^2\right] = \sum_{i_1\in[r],i_2\in[r]}\mathsf{E}\left[\left(U_{i_1}^\top\epsilon^k V_{i_2}\right)^2\right] \leq r^2K^2.$$

The remaining terms are bounded by applying Lemma 1 directly:

$$\mathsf{E}\left[d(U,\hat{U})^2\right] \leq 24\frac{4K^2m\left\|M_{(1)}\right\|_F^2 + K^4m^3n + m\delta^2}{\sigma_r^4(M_{(1)})},$$

$$\mathsf{E}\left[\left\|\epsilon^k\right\|_F^2 d(U,\hat{U})^2\right] \leq 24K^2m^2\frac{4K^2m\left\|M_{(1)}\right\|_F^2 + K^4m^3n + m\delta^2}{\sigma_r^4(M_{(1)})},$$

and similarly for $\mathsf{E}\left[d(V,\hat{V})^2\right]$ and $\mathsf{E}\left[\left\|\epsilon^k\right\|_F^2 d(V,\hat{V})^2\right]$.

Putting all of this together, and replacing running constants with c, we have

$$\begin{aligned}\frac{1}{m^2}\mathsf{E}\left[\left\|\hat{M}^k - M^k\right\|_F^2\right] &\leq c\left[\frac{K^2r^2}{m^2} + \left(\frac{\left\|M^k\right\|_F^2}{m^2} + K^2\right)\left(\frac{K^2m\left\|M_{(1)}\right\|_F^2 + K^4m^3n + m\delta^2}{\min_{i=1,2}\{\sigma_r^4(M_{(i)})\}}\right)\right] \\ &\leq c\left[\frac{K^2r^2}{m^2} + \left(K^2+1\right)\left(\frac{K^2m^3n + K^4m^3n + m\delta^2}{\gamma_M^2 m^4n^2/r^2}\right)\right] \\ &\leq c\left[\frac{K^2r^2}{m^2} + \left(\frac{K^2(K^2+1)^2m^3n + (K^2+1)m\delta^2}{\gamma_M^2 m^4n^2/r^2}\right)\right] \\ &\leq c\left[\frac{K^2r^2}{m^2} + \frac{K^2(K^4+1)r^2}{\gamma_M^2 mn} + \frac{(K^2+1)r^2\delta^2}{\gamma_M^2 m^3n^2}\right].\end{aligned}$$

In the second step, we plug in our definition of γ_M, $\min_{i=1,2}\{\sigma_r^2(M_{(i)})\} \geq \gamma_M m^2 n/r$, and use the facts that $\left\|M^k\right\|_F^2 \leq m^2$ and $\left\|M_{(1)}\right\|_F^2 \leq m^2 n$. The last two steps are a rearrangement of terms. Note that this entire analysis holds for any $k \in [n]$, so

$$SMSE(\hat{M}) = \max_{k\in[n]} \frac{1}{m^2} \mathsf{E}\left[\left\|\hat{M}^k - M^k\right\|_F^2\right] \leq c\left[\frac{K^2 r^2}{m^2} + \frac{K^2(K^4+1)r^2}{\gamma_M^2 mn} + \frac{(K^2+1)r^2\delta^2}{\gamma_M^2 m^3 n^2}\right].$$

□

E Additional Numerical Results

Tables 2 and 3 summarize the results of the experiments using data from `Xiami.com`, in terms of recovering the Download and Listen slices. See Section 6.2 for a detailed description of the experiment.

Users	Songs	Sparsity	Naive	Matrix	Slice
2,412	1,541	9.6	0.84 (11)	0.87 (7)	0.91 (12)
4,951	2,049	7.9	0.83 (14)	0.85 (9)	0.91 (12)
27,411	3,472	3.2	0.83 (11)	0.86 (8)	0.91 (14)
23,300	10,106	14.2	0.94 (18)	0.93 (13)	0.94 (18)
53,713	10,199	8.2	0.93 (10)	0.93 (7)	0.94 (20)

Table 2: Summary of experiments on Xiami data for recovering the Download slice. Each row corresponds to an experiment on a subset of the data. Columns 'Users' and 'Songs' show the number of users and songs in each experiment, and 'Sparsity' gives the average number of downloads per user in the data. Results for the naive benchmark, the matrix-based benchmark, and the slice learning algorithm are shown in the last three columns. The average AUC over 10 replications is reported, along with the rank in parentheses.

Users	Songs	Sparsity	Naive	Matrix	Slice
2,412	1,541	14.8	0.88 (6)	0.88 (7)	0.91 (11)
4,951	2,049	12.6	0.88 (7)	0.87 (11)	0.91 (11)
27,411	3,472	7.5	0.87 (6)	0.87 (3)	0.90 (9)
23,300	10,106	21.3	0.94 (7)	0.92 (8)	0.94 (15)
53,713	10,199	14.1	0.92 (5)	0.92 (12)	0.93 (7)

Table 3: Summary of experiments on Xiami data for recovering the Listen slice. Each row corresponds to an experiment on a subset of the data. Columns 'Users' and 'Songs' show the number of users and songs in each experiment, and 'Sparsity' gives the average number of listens per user in the data. Results for the naive benchmark, the matrix-based benchmark, and the slice learning algorithm are shown in the last three columns. The average AUC over 10 replications is reported, along with the rank in parentheses.