We show that given an estimate \( \hat{A} \) that is close to a general high-rank positive semi-definite (PSD) matrix \( A \) in spectral norm (i.e., \( \| \hat{A} - A \|_2 \leq \delta \)), the simple truncated Singular Value Decomposition of \( \hat{A} \) produces a multiplicative approximation of \( A \) in Frobenius norm. This observation leads to many interesting results on general high-rank matrix estimation problems:

1. **High-rank matrix completion**: we show that it is possible to recover a general high-rank matrix \( A \) up to \((1 + \varepsilon)\) relative error in Frobenius norm from partial observations, with sample complexity independent of the spectral gap of \( A \).

2. **High-rank matrix denoising**: we design an algorithm that recovers a matrix \( A \) with error in Frobenius norm from its noise-perturbed observations, without assuming \( A \) is exactly low-rank.

3. **Low-dimensional approximation of high-dimensional covariance**: given \( N \) i.i.d. samples of dimension \( n \) from \( \mathcal{N}_n(0, A) \), we show that it is possible to approximate the covariance matrix \( A \) with relative error in Frobenius norm with \( N \approx n \), improving over classical covariance estimation results which requires \( N \approx n^2 \).

## 1 Introduction

Let \( A \) be an unknown general high-rank \( n \times n \) PSD data matrix that one wishes to estimate. In many machine learning applications, though \( A \) is unknown, it is relatively easy to obtain a crude estimate \( \hat{A} \) that is close to \( A \) in spectral norm (i.e., \( \| \hat{A} - A \|_2 \leq \delta \)). For example, in matrix completion a simple procedure that fills all unobserved entries with 0 and re-scales observed entries produces an estimate that is consistent in spectral norm (assuming the matrix satisfies a spikeness condition, standard assumption in matrix completion literature). In matrix de-noising, an observation that is corrupted by Gaussian noise is close to the underlying signal, because Gaussian noise is isotropic and has small spectral norm. In covariance estimation, the sample covariance in low-dimensional settings is close to the population covariance in spectral norm under mild conditions [Bunea and Xiao, 2015].

However, in most such applications it is not sufficient to settle for a spectral norm approximation. For example, in recommendation systems (an application of matrix completion) the zero-filled re-scaled rating matrix is close to the ground truth in spectral norm, but it is an absurd estimator because most of the estimated ratings are zero. It is hence mandatory to require a more stringent measure of performance. One commonly used measure is the Frobenius norm of the estimation error \( \| \hat{A} - A \|_F \), which ensures that (on average) the estimate is close to the ground truth in an element-wise sense. A spectral norm approximation \( \hat{A} \) is in general not a good estimate under Frobenius norm, because in high-rank scenarios \( \| \hat{A} - A \|_F \) can be \( \sqrt{n} \) times larger than \( \| \hat{A} - A \|_2 \).
In this paper, we show that in many cases a powerful multiplicative low-rank approximation in Frobenius norm can be obtained by applying a simple truncated SVD procedure on a crude, easy-to-find spectral norm approximate. In particular, given the spectral norm approximation condition $\|\hat{A} - A\|_2 \leq \delta$, the top-$k$ SVD of $\hat{A}_k$ of $\hat{A}$ multiplicatively approximates $\hat{A}$ in Frobenius norm; that is, $\|\hat{A}_k - A\|_F \leq C(k, \delta, \sigma_{k+1}(A))\|A - \hat{A}_k\|_F$, where $\hat{A}_k$ is the best rank-$k$ approximation of $A$ in Frobenius and spectral norm. To our knowledge, the best existing result under the assumption $\|\hat{A} - A\|_2 \leq \delta$ is due to Achlioptas and McSherry [2007], who showed that $\|\hat{A}_k - A\|_F \leq \|A - A_k\|_F + \sqrt{k}\delta + 2k^{1/4}\sqrt{\delta}\|A_k\|_F$, which depends on $\|A_k\|_F$ and is not multiplicative in $\|A - A_k\|_F$.

Below we summarize applications in several matrix estimation problems.

**High-rank matrix completion** Matrix completion is the problem of (approximately) recovering a data matrix from very few observed entries. It has wide applications in machine learning, especially in online recommendation systems. Most existing work on matrix completion assumes the data matrix is exactly low-rank [Candes and Recht, 2012, Sun and Luo, 2016, Jain et al., 2013]. Candes and Plan [2010], Keshavan et al. [2010] studied the problem of recovering a low-rank matrix corrupted by stochastic noise; Chen et al. [2016] considered sparse column corruption. All of the aforementioned work assumes that the ground-truth data matrix is exactly low-rank, which is rarely true in practice.

Negahban and Wainwright [2012] derived minimax rates of estimation error when the spectrum of the data matrix lies in an $\ell_p$ ball. Zhang et al. [2015], Koltchinskii et al. [2011] derived oracle inequalities for general matrix completion; however their error bound has an additional $O(\sqrt{n})$ multiplicative factor. These results also require solving computationally expensive nuclear-norm penalized optimization problems whereas our method only requires solving a single truncated singular value decomposition. Chatterjee et al. [2015] also used the truncated SVD estimator for matrix completion. However, his bound depends on the nuclear norm of the underlying matrix which may be $\sqrt{n}$ times larger than our result. Hardt and Wootters [2014] used a “soft-deflation” technique to remove condition number dependency in the sample complexity; however, their error bound for general high-rank matrix completion is additive and depends on the “consecutive” spectral gap $(\sigma_k(A) - \sigma_{k+1}(A))$, which can be small in practical settings [Balcan et al., 2016, Anderson et al., 2015]. Eriksson et al. [2012] considered high-rank matrix completion with additional union-of-subspaces structures.

In this paper, we show that if the $n \times n$ data matrix $A$ satisfies a $\mu_0$-spikiness condition, \(^1\) then for any $\epsilon \in (0, 1)$, the truncated SVD of zero-filled matrix $\hat{A}_k$ satisfies $\|\hat{A}_k - A\|_F \leq (1 + O(\epsilon))\|A - A_k\|_F$ if the sample complexity is lower bounded by $\Omega(\frac{n \max \{\epsilon^{-2}, k^2\} r_k(A) \log n}{\sigma_{k+1}(A)^2})$, which can be further simplified to $\Omega(\mu_0^2 \max \{\epsilon^{-2}, k^2\} \gamma_k(A)^2 \sqrt{n r_s(A) \log n})$, where $\gamma_k(A) = \sigma_k(A)/\sigma_{k+1}(A)$ is the $k$th-order condition number and $r_s(A) = \|A\|_F^2/\|A\|_2^2 \leq \text{rank}(A)$ is the stable rank of $A$. Compared to existing work, our bound is multiplicative, gap-free, and the estimator is computationally efficient. \(^2\)

**High-rank matrix de-noising** Let $\hat{A} = A + E$ be a noisy observation of $A$, where $E$ is a PSD Gaussian noise matrix with zero mean and $\nu^2/n$ variance on each entry. By simple concentration results we have $\|\hat{A} - A\|_2 = \nu$ with high probability; however, $\hat{A}$ is in general not a good estimator of $A$ in Frobenius norm when $A$ is high-rank. Specifically, $\|\hat{A} - A\|_F$ can be as large as $\sqrt{n\nu}$.

Applying our main result, we show that if $\nu < \sigma_{k+1}(A)$ for some $k \ll n$, then the top-$k$ SVD $\hat{A}_k$ of $\hat{A}$ satisfies $\|\hat{A}_k - A\|_F \leq (1 + O(\sqrt{n}/\sigma_{k+1}(A)))\|A - A_k\|_F + \sqrt{\nu}$. This suggests a form of bias-variance decomposition as larger rank threshold $k$ induces smaller bias $\|A - \hat{A}_k\|_F$ but larger variance $k\nu^2$. Our results generalize existing work on matrix de-noising [Donoho and Gavish, 2014, Donoho et al., 2013, Gavish and Donoho, 2014], which focus primarily on exact low-rank $A$.

\(^1\) $n\|A\|_{\text{max}} \leq \mu_0\|A\|_F$; see also Definition 2.1.

\(^2\) We remark that our relative-error analysis does not, however, apply to exact rank-$k$ matrix where $\sigma_{k+1} = 0$. This is because for exact rank-$k$ matrix a bound of the form $(1 + O(\epsilon))\|A - A_k\|_F$ requires exact recovery of $A$, which truncated SVD cannot achieve. On the other hand, in the case of $\sigma_{k+1} = 0$ a weaker additive-error bound is always applicable, as we show in Theorem 2.3.
We first state our main result, which shows that truncated SVD on a weak estimator with small approximation error in spectral norm leads to a strong estimator with multiplicative error bound. We remark that truncated SVD in general has time complexity $\Theta(nk^2)$, whereas our analysis is based on $\Omega(k)$ i.i.d. samples. Using similar arguments in the proof, we believe similar results for general asymmetric matrices can be obtained as well.
To our knowledge, the best existing bound for $\|\hat{A}_k - A\|_F$ assuming $\|\hat{A} - A\|_2 \leq \delta$ is due to Achlioptas and McSherry [2007], who showed that
\[
\|\hat{A}_k - A\|_F \leq \|A - A_k\|_F + (\hat{A} - A)_k\|_F + 2\sqrt{(\hat{A} - A)_k\|_F \|A_k\|_F}
\leq \|A - A_k\|_F + \sqrt{k}\delta \|A - A_k\|_2 + 2k^{1/4}\sqrt{\delta \|A_k\|_F}.
\]
(3)

Compared to Theorem 2.1, Eq. (3) is not relative because the third term $2k^{1/4}\sqrt{\delta \|A_k\|_F}$ depends on the $k$ largest eigenvalues of $A$, which could be much larger than the remainder term $\|A - A_k\|_F$. In contrast, Theorem 2.1, together with Remark 2.1, shows that $\|\hat{A}_k - A\|_F$ could be upper bounded by a small factor multiplied with the remainder term $\|A - A_k\|_F$.

We also provide a gap-dependent version.

Theorem 2.2. Suppose $A$ is an $n \times n$ PSD matrix with eigenvalues $\sigma_1(A) \geq \cdots \geq \sigma_n(A) \geq 0$, and a symmetric matrix $\hat{A}$ satisfies $\|\hat{A} - A\|_2 \leq \delta = \epsilon (\sigma_k(A) - \sigma_{k+1}(A))$ for some $\epsilon \in (0, 1/4]$. Let $A_k$ and $\hat{A}_k$ be the best rank-$k$ approximations of $A$ and $\hat{A}$. Then
\[
\|\hat{A}_k - A\|_F \leq \|A - A_k\|_F + 102\sqrt{2k}\epsilon (\sigma_k(A) - \sigma_{k+1}(A)).
\]
(4)

If $A$ is an exact rank-$k$ matrix, Theorem 2.2 implies that truncated SVD gives an $\epsilon \sqrt{2k}\sigma_k$ error approximation in Frobenius norm, which has been established by many previous works [Yi et al., 2016, Tu et al., 2015, Wang et al., 2016].

Before we proceed to the applications and proof of Theorem 2.1, we list several examples of $A$ with classical distribution of eigenvalues and discuss how Theorem 2.1 could be applied to obtain good Frobenius-norm approximations of $A$. We begin with the case where eigenvalues of $A$ have a polynomial decay rate (i.e., power law). Such matrices are ubiquitous in practice [Liu et al., 2015].

Corollary 2.1 (Power-law spectral decay). Suppose $\|\hat{A} - A\|_2 \leq \delta$ for some $\delta \in (0, 1/2]$ and $\sigma_j(A) = j^{-\beta}$ for some $\beta > 1/2$. Set $k = \lfloor \min\{C_1\delta^{-1/\beta}, n\} - 1 \rfloor$. If $k \geq 1$ then
\[
\|\hat{A}_k - A\|_F \leq C' \max \left\{ \delta^{\frac{2^{q+1}}{2^{q+1}}}, n^{-\frac{2^{q+1}}{2^{q+1}}} \right\},
\]
where $C_1, C_2' > 0$ are constants that only depend on $\beta$.

We remark that the assumption $\sigma_j(A) = j^{-\beta}$ implies that the eigenvalues lie in an $\ell_q$ ball for $q = 1/\beta$; that is, $\sum_{j=1}^{n} \sigma_j(A)^q = O(1)$. The error bound in Corollary 2.1 matches the minimax rate (derived by Negahban and Wainwright [2012]) for matrix completion when the spectrum is constrained in an $\ell_q$ ball, by replacing $\delta$ with $\sqrt{n/N}$ where $N$ is the number of observed entries.

Next, we consider the case where eigenvalues satisfy a faster decay rate.

Corollary 2.2 (Exponential spectral decay). Suppose $\|\hat{A} - A\|_2 \leq \delta$ for some $\delta \in (0, e^{-16})$ and $\sigma_j(A) = \exp\{-cj\}$ for some $c > 0$. Set $k = \lfloor \min\{c^{-1}\log(1/\delta) - c^{-1}\log(1/\delta), n\} - 1 \rfloor$. If $k \geq 1$ then
\[
\|\hat{A}_k - A\|_F \leq C' \max \left\{ \delta \sqrt{\log(1/\delta)^3}, n^{1/2} \exp(-cn) \right\},
\]
where $C_2' > 0$ is a constant that only depends on $c$.

Both corollaries are proved in the appendix. The error bounds in both Corollaries 2.1 and 2.2 are significantly better than the trivial estimate $\hat{A}$, which satisfies $\|\hat{A} - A\|_F \leq n^{1/2}\delta$. We also remark that the bound in Corollary 2.1 cannot be obtained by a direct application of the weaker bound Eq. (3), which yields a $\delta^{\frac{1}{\beta-1}}$ bound.

We next state results that are consequences of Theorem 2.1 in several matrix estimation problems.

2.1 High-Rank Matrix Completion

Suppose $A$ is a high-rank $n \times n$ PSD matrix that satisfies $\mu_0$-spikeness condition defined as follows:
Furthermore, for fixed \( \epsilon \in \mathbb{R} \) we have

\[
\| \hat{A} - A \|_F \leq O(\sqrt{k}) \cdot \| A - A_k \|_F \quad \text{if} \quad p = \Omega \left( \frac{\mu_0^2 \| A \|_F^2 \log(n/t)}{n^2 \sigma_k(A)^2} \right),
\]

**Theorem 2.2.** Fix \( t \in (0, 1) \). Then with probability \( 1 - t \) we have

\[
\| \hat{A}_k - A \|_F \leq \sqrt{\frac{k}{n}} \| A - A_k \|_F \quad \text{if} \quad p = \Omega \left( \frac{\mu_0^2 \max\{\epsilon^{-4}, k^2\} \| A \|_F^2 \log(n/t)}{n \sigma_k(A)^2} \right).
\]

Furthermore, for fixed \( \epsilon \in (0, 1/4] \), with probability \( 1 - t \) we have

\[
\| \hat{A}_k - A \|_F \leq \| A - A_k \|_F + \epsilon (\sigma_k(A) - \sigma_{k+1}(A)) \quad \text{if} \quad p = \Omega \left( \frac{\mu_0^2 \max\{\epsilon^{-4}, k^2\} \| A \|_F^2 \log(n/t)}{n \sigma_k(A)^2} \right).
\]

As a remark, because \( \mu_0 \geq 1 \) and \( \| A \|_F^2 / \sigma_{k+1}(A) \geq \sqrt{k} \) always hold, the sample complexity is lower bounded by \( \Omega(nk \log n) \), the typical sample complexity in noiseless matrix completion. In the case of high rank \( A \), the results in Theorem 2.2 are the strongest when \( A \) has small stable rank \( r_s(A) = \| A \|_F^2 / \| A \|_2^2 \) and the top-\( k \) condition number \( \gamma_k(A) = \sigma_1(A) / \sigma_{k+1}(A) \) is not too large. For example, if \( A \) has stable rank \( r_s(A) = r \) then \( \| \hat{A}_k - A \|_F^2 \) has an \( O(\sqrt{k}) \) multiplicative error bound with sample complexity \( \Omega(\mu_0^2 \gamma_k(A)^2 \cdot nk \log n) \); or an \( (1 + O(\epsilon)) \) relative error bound with sample complexity \( \Omega(\mu_0^2 \max\{\epsilon^{-4}, k^2\} \gamma_k(A)^2 \cdot nk \log n) \). Finally, when \( \sigma_{k+1}(A) \) is very small and the “gap” \( \sigma_k(A) - \sigma_{k+1}(A) \) is large, a weaker additive-error bound is applicable with sample complexity independent of \( \sigma_{k+1}(A)^{-1} \).

Comparing with previous works, if the gap \( 1 - \sigma_{k+1}(A) / \sigma_k(A) \) is of order \( \epsilon \), then sample complexity of [Hardt, 2014] Theorem 1.2 and [Hardt and Wooters, 2014] Theorem 1 scale with \( 1/\epsilon^7 \). Our result improves their results to the scaling of \( 1/\epsilon^4 \) with a much simpler algorithm (truncated SVD).
2.2 High-rank matrix de-noising

Let $A$ be an $n \times n$ PSD signal matrix and $E$ a symmetric random Gaussian matrix with zero mean and $\nu^2/n$ variance. That is, $E_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, \nu^2/n)$ for $1 \leq i \leq j \leq n$ and $E_{ij} = E_{ji}$. Define $\hat{A} = A + E$. The matrix de-noising problem is then to recover the signal matrix $A$ from noisy observations $\hat{A}$. We refer the readers to [Gavish and Donoho, 2014] for a list of references that shows the ubiquitous application of matrix de-noising in scientific fields.

It is well-known by concentration results of Gaussian random matrices, that $\|\hat{A} - A\|_F = \|E\|_F = O_P(\nu)$. Let $\hat{A}_k$ be the best rank-$k$ approximation of $\hat{A}$ in Frobenius/spectral norm. Applying Theorem 2.1, we immediately have the following result:

**Theorem 2.4.** There exists an absolute constant $c > 0$ such that, if $\nu < c \cdot \sigma_{k+1}(A)$ for some $1 \leq k < n$, then with probability at least 0.8 we have that

$$
\|\hat{A}_k - A\|_F \leq \left(1 + O\left(\frac{\nu}{\sigma_{k+1}(A)}\right)\right) \|A - A_k\|_F + O(\sqrt{k}\nu). 
$$

Eq. (6) can be understood from a classical bias-variance tradeoff perspective: the first $(1 + O(\nu/\sigma_{k+1}(A)))\|A - A_k\|_F$ acts as a bias term, which decreases as we increase cut-off rank $k$, corresponding to a more complicated model; on the other hand, the second $O(\sqrt{k}\nu)$ term acts as the (square root of) variance, which does not depend on the signal $A$ and increases with $k$.

2.3 Low-rank estimation of high-dimensional covariance

Suppose $A$ is an $n \times n$ PSD matrix and $X_1, \ldots, X_N$ are i.i.d. samples drawn from the multivariate Gaussian distribution $\mathcal{N}_n(0, A)$. The question is to estimate $A$ from samples $X_1, \ldots, X_N$. A common estimator is the sample covariance $\hat{A} = \frac{1}{N} \sum_{i=1}^N X_i X_i^\top$. While in low-dimensional regimes (i.e., $n$ fixed and $N \to \infty$) the asymptotic efficiency of $\hat{A}$ is obvious (cf. [Van der Vaart, 2000]), its statistical power in high-dimensional regimes where $n$ and $N$ are comparable is highly non-trivial. Below we cite results by Bunea and Xiao [2015] for estimation error $\|\hat{A} - A\|_F$, $\xi = 2/F$ when $n$ is not too large compared to $N$:

**Lemma 2.2** (Bunea and Xiao [2015]). Suppose $n = O(N^{\beta})$ for some $\beta \geq 0$ and let $r_e(A) = \text{tr}(A)/\|A\|_2$ denote the effective rank of the covariance $A$. Then the sample covariance $\hat{A} = \frac{1}{N} \sum_{i=1}^N X_i X_i^\top$ satisfies

$$
\|\hat{A} - A\|_F = O_P\left(\|A\|_2 r_e(A) \sqrt{\frac{\log N}{N}}\right) 
$$

and

$$
\|\hat{A} - A\|_2 = O_P\left(\|A\|_2 \max\left\{ \sqrt{\frac{r_e(A) \log(Nn)}{N}}, \frac{r_e(A) \log(Nn)}{N} \right\}\right). 
$$

Let $\hat{A}_k$ be the best rank-$k$ approximation of $\hat{A}$ in Frobenius/spectral norm. Applying Theorem 2.1 and 2.2 together with Eq. (8), we immediately arrive at the following theorem.

**Theorem 2.5.** Fix $\epsilon \in (0, 1/4]$ and $1 \leq k < n$. Recall that $r_e(A) = \text{tr}(A)/\|A\|_2$ and $\gamma_k(A) = \sigma_1(A)/\sigma_{k+1}(A)$. There exists a universal constant $c > 0$ such that, if

$$
r_e(A) \max\{\epsilon^{-4}, k^2\gamma_k(A)^2 \log(N)\} \leq c
$$

then with probability at least 0.8,

$$
\|\hat{A}_k - A\|_F \leq (1 + O(\epsilon)) \|A - A_k\|_F 
$$

and if

$$
\frac{r_e(A)k\|A\|_2^2 \log(N)}{N \epsilon^2 (\sigma_k(A) - \sigma_{k+1}(A))^2} \leq c
$$
then with probability at least 0.8,

$$\| \hat{A}_k - A \|_F \leq \| A - A_k \|_F + \epsilon (\sigma_k (A) - \sigma_{k+1} (A)) .$$

Theorem 2.5 shows that it is possible to obtain a reasonable Frobenius-norm approximation of $\hat{A}$ by truncated SVD in the asymptotic regime of $N = \Omega (r_e (A) \log k)$, which is much more flexible than Eq. (7) that requires $N = \Omega (r_e (A)^2 \log N)$.

3 Proof Sketch of Theorem 2.1

In this section we give a proof sketch of Theorem 2.1. The proof of Theorem 2.2 is similar and less challenging so we defer it to Appendix. We defer proofs of technical lemmas to Section A.

Because both $\hat{A}_k$ and $A_k$ are low-rank, $\| \hat{A}_k - A_k \|_F$ is upper bounded by an $O (\sqrt{r_k})$ factor of $\| \hat{A}_k - A_k \|_2$. From the condition that $\| \hat{A} - A \|_2 \leq \delta$, a straightforward approach to upper bound $\| \hat{A}_k - A_k \|_2$ is to consider the decomposition $\| \hat{A}_k - A_k \|_2 \leq \| \hat{A} - A \|_2 + 2 \| \hat{A}_k \bar{U}_k - U_k \bar{U}_k^T \|_2 \| \hat{A}_k \|_2$, where $U_k \bar{U}_k^T$ and $\hat{U}_k \hat{U}_k^T$ are projection operators onto the top-$k$ eigenspaces of $A$ and $\hat{A}$, respectively. Such a naive approach, however, has two major disadvantages. First, the upper bound depends on $\| \hat{A}_k \|_2$, which is additive and may be much larger than $\| \hat{A} - A \|_2$. Perhaps more importantly, the quantity $\| U_k \bar{U}_k^T - \hat{U}_k \hat{U}_k^T \|_2$ depends on the “consecutive” spectral gap $(\sigma_k (A) - \sigma_{k+1} (A))$, which could be very small for large matrices.

The key idea in the proof of Theorem 2.1 is to find an “envelope” $m_1 \leq k \leq m_2$ in the spectrum of $A$ surrounding $k$, such that the eigenvalues within the envelope are relatively close. Define

$$m_1 = \arg \max_{0 \leq j \leq k} \{ \sigma_j (A) \geq (1 + 2 \epsilon) \sigma_{k+1} (A) \};$$

$$m_2 = \arg \max_{0 \leq j \leq n} \{ \sigma_j (A) \geq \sigma_k (A) - 2 \epsilon \sigma_{k+1} (A) \},$$

where we let $\sigma_0 (A) = \infty$ for convenience. Let $U_m, \hat{U}_m$ be basis of the top $m$-dimensional linear subspaces of $A$ and $\hat{A}$, respectively. Also denote $U_{n-m}$ and $\hat{U}_{n-m}$ as bases of the orthogonal complement of $U_m$ and $\hat{U}_m$. By asymmetric Davis-Kahan inequality (Lemma C.1) and Wely’s inequality we can obtain the following result.

Lemma 3.1. If $\| \hat{A} - A \|_2 \leq \epsilon^2 \sigma_{k+1} (A)$ for $\epsilon \in (0, 1)$ then $\| \hat{U}_{m-k}^T U_{m_1} \|_2, \| \hat{U}_k^T U_{m_2-n} \|_2 \leq \epsilon$.

Let $U_{m_1:m_2}$ be the linear subspace of $A$ associated with eigenvalues $\sigma_{m_1+1} (A), \ldots, \sigma_{m_2} (A)$. Intuitively, we choose a $(k - m_1)$-dimensional linear subspace in $U_{m_1:m_2}$ that is “most aligned” with the top-$k$ subspace $\hat{U}_k$ of $\hat{A}$. Formally, define

$$W = \arg \max_{\text{dim}(W) = k-m_1} \sigma_{k-m_1} (W^T \hat{U}_k) \cdot \cdot \cdot$$

$W$ is then a $d \times (k - m_1)$ matrix with orthonormal columns that corresponds to a basis of $W$. $W$ is carefully constructed so that it is closely aligned with $\hat{U}_k$, yet still lies in $U_k$. In particular, Lemma 3.2 shows that $\sin (W, \hat{U}_k) = \| \hat{U}_{m-k}^T W \|_2$ is upper bounded by $\epsilon$.

Lemma 3.2. If $\| \hat{A} - A \|_2 \leq \epsilon^2 \sigma_{k+1} (A)$ for $\epsilon \in (0, 1)$ then $\| \hat{U}_{m-k}^T W \|_2 \leq \epsilon$.

Now define

$$\hat{A} = A_{m_1} + WW^T A WW^T .$$

We use $\hat{A}$ as the “reference matrix” because we can decompose $\| \hat{A}_k - A \|_F$ as

$$\| \hat{A}_k - A \|_F \leq \| A - \hat{A} \|_F + \| \hat{A}_k - \hat{A} \|_F \leq \| A - \hat{A} \|_F + \sqrt{2k} \| \hat{A}_k - \hat{A} \|_2$$

and bound each term on the right hand side separately. Here the last inequality holds because both $\hat{A}_k$ and $\hat{A}$ have rank at most $k$. The following lemma bounds the first term.

Lemma 3.3. If $\| \hat{A} - A \|_2 \leq \epsilon^2 \sigma_{k+1} (A)^2$ for $\epsilon \in (0, 1/4]$ then $\| A - \hat{A} \|_F \leq (1 + 32 \epsilon) \| A - A_k \|_F$. 

7
The proof of this lemma relies Pythagorean theorem and Poincaré separation theorem. Let $\mathcal{U}_{m_1:m_2}$ be the $(m_2 - m_1)$-dimensional linear subspace such that $\mathcal{U}_{m_2} = \mathcal{U}_{m_1} \oplus \mathcal{U}_{m_1:m_2}$. Define $A_{m_1:m_2} = U_{m_1:m_2} \Sigma_{m_1:m_2} U_{m_1:m_2}^T$, where $\Sigma_{m_1:m_2} = \text{diag}(\sigma_{m_1+1}(A), \ldots, \sigma_{m_2}(A))$ and $U_{m_1:m_2}$ is an orthonormal basis associated with $\mathcal{U}_{m_1:m_2}$. Applying Pythagorean theorem (Lemma C.2), we can decompose

$$\|A - \tilde{A}\|_F^2 = \|A - \tilde{A}_{m_2}\|_F^2 + \|A_{m_1:m_2}\|_F^2 - \|WW^TA_{m_1:m_2}WW^T\|_F^2.$$  

Applying Poincaré separation theorem (Lemma C.3) where $X = \Sigma_{m_1:m_2}$ and $P = U_{m_1:m_2}^T W$, we have $\|WW^TA_{m_1:m_2}W\|_F^2 \geq \sum_{j=m_1+1}^{m_2} \sigma_j(A_{m_1:m_2})^2 = \sum_{j=m_1+1}^{m_2} \sigma_j(A)^2$. With some routine algebra we can prove Lemma 3.3.

To bound the second term of Eq. (9) we use the following lemma.

**Lemma 3.4.** If $\|\hat{A} - A\|_2 \leq \epsilon^2 \sigma_{k+1}(A)$ for $\epsilon \in (0, 1/4]$ then $\|\hat{A}_k - \tilde{A}\|_2 \leq 102\epsilon^2\|A - A_k\|_2$.

The proof of Lemma 3.4 relies on the low-rankness of $\hat{A}_k$ and $\tilde{A}$. Recall the definition that $\tilde{U} = \text{Range}(\tilde{A})$ and $\tilde{U}_\perp = \text{Null}(\tilde{A})$. Consider $\|v\|_2 = 1$ such that $v^T(\hat{A}_k - \tilde{A})v = \|\tilde{A}_k - \tilde{A}\|_2$. Because $v$ maximizes $v^T(\hat{A}_k - \tilde{A})v$ over all unit-length vectors, it must lie in the range of $\{\hat{A}_k - \tilde{A}\}$ because otherwise the component outside the range will not contribute. Therefore, we can choose $v$ that $v = v_1 + v_2$ where $v_1 \in \text{Range}(\hat{A}_k) = \tilde{U}_k$ and $v_2 \in \text{Range}(\tilde{A}) = \tilde{U}_\perp$. Subsequently, we have that

$$v = \tilde{U}_k \hat{U}_k v + \tilde{U}_\perp \tilde{U}_{n-k} \hat{U}_{n-k} v$$

$$= \tilde{U} \hat{U} v + \tilde{U}_\perp \hat{U}_{n-k} v.$$  

(10)  

Consider the following decomposition:

$$\left|v^T(\hat{A}_k - \tilde{A})v\right| \leq |v^T(\hat{A} - A)v| + |v^T(\hat{A}_k - \tilde{A})v| + |v^T(A - \tilde{A})v|.$$  

The first term $|v^T(\hat{A} - A)v|$ is trivially upper bounded by $\|\hat{A} - A\|_2 \leq \epsilon^2 \sigma_{k+1}(A)$. The second and the third term can be bounded by Wely’s inequality (Lemma C.4) and basic properties of $\hat{A}$ (Lemma A.3). See Section A for details.

4 Discussion

We mention two potential directions to further extend results of this paper.

4.1 Model selection for general high-rank matrices

The validity of Theorem 2.1 depends on the condition $\|\hat{A} - A\|_2 \leq \epsilon^2 \sigma_{k+1}(A)$, which could be hard to verify if $\sigma_{k+1}(A)$ is unknown and difficult to estimate. Furthermore, for general high-rank matrices, the model selection problem of determining an appropriate (or even optimal) cut-off rank $k$ requires knowledge of the distribution of the entire spectrum of an unknown data matrix, which is even more challenging to obtain.

One potential approach is to impose a parametric pattern of decay of the eigenvalues (e.g., polynomial and exponential decay), and to estimate a small set of parameters (e.g., degree of polynomial) from the noisy observations $\hat{A}$. Afterwards, the optimal cut-off rank $k$ could be determined by a theoretical analysis, similar to the examples in Corollaries 2.1 and 2.2. Another possibility is to use repeated sampling techniques such as bootstrap in a stochastic problem (e.g., matrix de-noising) to estimate the “bias” term $\|A - A_k\|_F$ for different $k$, as the variance term $\sqrt{\kappa}$ is known or easy to estimate.

4.2 Minimax rates for polynomial spectral decay

Consider the class of PSD matrices whose eigenvalues follow a polynomial (power-law) decay: $\Theta(\beta, n) = \{A \in \mathbb{R}^{n \times n} : A > 0, \sigma_j(A) = j^{-\beta}\}$. We are interested in the following minimax rates for completing or de-noising matrices in $\Theta(\beta, n)$:
Question 1 (Completion of $\Theta(\beta, n)$). Fix $n \in \mathbb{N}$, $p \in (0, 1)$ and define $N = pn^2$. For $M \in \Theta(\beta, n)$, let $\tilde{A}_{ij} = M_{ij}/\sqrt{n}$ with probability $p$ and $\tilde{A}_{ij} = 0$ with probability $1 - p$. Also let $\Lambda(\mu_0, n) = \{M \in \mathbb{R}^{n \times n} : n \|M\|_{\text{max}} \leq \mu_0 \|M\|_F \}$ be the class of all non-spiky matrices. Determine

$$R_1(\mu_0, \beta, n, N) := \inf_{A \rightarrow M} \sup_{M \in \Theta(\beta, n) \cap \Lambda(\mu_0, n)} \mathbb{E}\|\tilde{M} - M\|_F^2.$$ 

Question 2 (De-noising of $\Theta(\beta, n)$). Fix $n \in \mathbb{N}$, $\nu > 0$ and let $\tilde{A} = M + \nu/\sqrt{n}Z$, where $Z$ is a symmetric matrices with i.i.d. standard Normal random variables on its upper triangle. Determine

$$R_2(\nu, \beta, n) := \inf_{A \rightarrow M} \sup_{M \in \Theta(\beta, n)} \mathbb{E}\|\tilde{M} - M\|_F^2.$$ 

Compared to existing settings on matrix completion and de-noising, we believe $\Theta(\beta, n)$ is a more natural matrix class which allows for general high-rank matrices, but also imposes sufficient spectral decay conditions so that spectrum truncation algorithms result in significant benefits. Based on Corollary 2.1 and its matching lower bounds for a larger $\mathcal{C}$ class [Negahban and Wainwright, 2012], we make the following conjecture:

**Conjecture 4.1.** For $\beta > 1/2$ and $\nu$ not too small, we conjecture that

$$R_1(\mu_0, \beta, n, N) \asymp C(\mu_0) \cdot \left[\frac{n}{N}\right]^{\frac{2(p-1)}{2p}} \quad \text{and} \quad R_2(\nu, \beta, n) \asymp [\nu^2]^{\frac{2(p-1)}{2p}},$$

where $C(\mu_0) > 0$ is a constant that depends only on $\mu_0$.

5 Acknowledgements

S.S.D. was supported by ARPA-E Terra program. Y.W. and A.S. were supported by the NSF CAREER grant IIS-1252412.

References


