Residual-Based Sampling for Online Outlier-Robust PCA

Tianhao Zhu [1] **Jie Shen** [1]

Abstract

Outlier-robust principal component analysis (ORPCA) has been broadly applied in scientific discovery in the last decades. In this paper, we study online ORPCA, an important variant that addresses the practical challenge that the data points arrive in a sequential manner and the goal is to recover the underlying subspace of the clean data with one pass of the data. Our main contribution is the first provable algorithm that enjoys comparable recovery guarantee to the best known batch algorithm, while significantly improving upon the state-of-the-art online ORPCA algorithms. The core technique is a robust version of the residual norm which, informally speaking, leverages not only the importance of a data point, but also how likely it behaves as an outlier.

1. Introduction

Principal Component Analysis (PCA) is a fundamental tool for analyzing high-dimensional data. The key idea is to get the optimal subspace approximation. In the absence of outliers, such subspace can be computed by the top-k left singular vectors, or Principal Components, of the sample covariance matrix. This is one of the most important problems in machine learning that has been studied for a long time. The PCA problem is well understood when the data matrix is fully observed, and there is no noise. Thus, a large body of recent works focus on the online setting or when the data are corrupted.

Robust PCA. Many robust PCA algorithms have been proposed because the traditional PCA breaks down immediately in the presence of outliers. Some of them focused on the robust PCA in the low-dimension regime. They either applied the standard PCA on the robust estimate of the covariance matrix (Xu & Yuille, 1995; Yang & Wang, 1999; la Torre & Black, 2001; Brubaker, 2009; Klivans et al., 2009), or tried to find a low-dimension that has the maximum robust estimation of the projected distance, called projection pursuit (Croux et al., 2007). A more challenging work lies in the high-dimension regime where the ambient dimension is higher than the number of observed data. In (Candès et al., 2011; Chandrasekaran et al., 2011), the exact recovery of the low-rank and sparse component via convex optimization was well established by minimizing a weighted combination of a nuclear-norm of one component and an ℓ_1-norm of another component. Specifically, Xu et al. (2012) showed that matrix decomposition using nuclear norm minimization can recover the column space and identify the outliers. The nearly optimal recovery guarantee was also achieved by Cherapanamjeri et al. (2017) using the threshold-based approaches, which reduced the computational cost significantly. However, these algorithms require not only observing all the samples, but also inliers being incoherent (thus inliers are not thresholded). Meanwhile, another widely used convex surrogate is the max-norm regularization (Srebro et al., 2004), where the max-norm promoted the a low-rank structure. Srebro & Shraibman (2005) studied collaborative filtering and showed a tighter generalization bound than the nuclear norm theoretically, and Lee et al. (2010); Jalali & Srebro (2012) showed its appealing performance in some practical applications empirically. The very recent work Deshpande & Pratap (2021) extended the outlier low-rank approximation problem under ℓ_p-metric, and extended the analysis over M-estimator loss functions and affine subspace approximation. Readers may refer to Lerman & Maunu (2018) for a comprehensive survey of the works on robust subspace recovery.

Online PCA. The online setting is more restricted than offline in that we can only observe the samples in a sequential manner. Note that this is different from stochastic optimization where the algorithms can access an arbitrary sample in each iteration (Ozawa et al., 2004; Nie et al., 2013; Arora et al., 2013; Garber & Hazan, 2015; Hallgren & Northrop, 2018). The recent work Garber (2019) cast the online PCA into the regret minimization framework. It proposed the regularized Online Gradient Ascent model, and enjoyed the poly-logarithmic regret bound, requiring only linear memory and run-time per iteration. Another way is to progressively expand the subspace of inliers. Boutsidis et al. (2015)

[1] Department of Computer Science, Stevens Institute of Technology, Hoboken, New Jersey, USA. Correspondence to: Tianhao Zhu <tzhu12@stevens.edu>, Jie Shen <jie.shen@stevens.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

proposed the first online algorithm for computing the PCA embedding under such setting, resulting in an additive error guarantee. In a recent work, Bhaskara et al. (2019) gave an online algorithm for Column Subsection Selection(CSS) as well as PCA, and achieved the multiplicative approximation by residual-based sampling method. Meanwhile, based on Krasulina's method Krasulina (1969) and Oja's rule Oja & Karhunen (1985), some works achieved the non-asymptotic convergence guarantee for the streaming PCA problem (Jain et al., 2016; Li et al., 2016; Allen-Zhu & Li, 2017; Tang, 2019; Amid & Warmuth, 2020). However, it turns out that all of these algorithms fail in the presence of outliers.

Online Robust PCA. As a combination of robust and online settings, there are two variants of Online Robust PCA. One assumes that the features of each newly observed sample are sparsely corrupted; see e.g. (Feng et al., 2013b; Shen et al., 2014). Another class assumes that some samples are arbitrary or even adversarial while the rest form a low-dimensional space, which is the regime of our interest. In this spectrum, The most related work to this paper is Feng et al. (2013a) in the sense that they considered online optimization for outlier-robust PCA. Unfortunately, their theoretical guarantees were less favorable for practical problems for two reasons: 1) they required strong conditions on the initial iterate, which is hard to satisfy; and 2) their results held only in an asymptotic sense while we present finite convergence guarantee. The reader may refer to Table 1 for a comparison.

1.1. Main results

The main algorithmic contribution of the paper is a novel online outlier-robust PCA (ORPCA) algorithm that uses adaptive sampling technique which shows promising results in the non-robust setting (Bhaskara et al., 2019).

Let matrix $A \in \mathbb{R}^{d\times n}$ be the observed matrix with n samples in d dimensions, with A_i being the i-th column. Let $k < \min\{d, n\}$ be the target rank of the subspace of inliers.

We assume the following conditions.

Assumption 1. The point arrives one after another, and we can only make a one-time pass over it.

Assumption 2. The matrix consists inliers A_{in} and outliers A_{out}, where the entries of A_{out} can be arbitrary and shown in columns in arbitrary order. There exists an upper bound z on the number of outliers. Namely, $|A_{\text{out}}| \leq z$.

Let ξ be a quantity such that $\left\|A_{\text{in}} - \text{SVD}_k(A_{\text{in}})\right\|_F^2 \leq \xi$. Let M be the points marked as outliers, and $A_{\text{in}}\backslash M$ be the inliers not marked as outliers. We present our first theorem below.

Theorem 1. *If Assumption 1 and Assumption 2 are satisfied, then there exists an efficient algorithm that upon seeing each A_i, decides to add it to the outlier set M, or outputs an embedding $Y_i \in \mathbb{R}^r$. In the end we have, the error bound on the inliers not marked as outliers is $\min_{\Phi\in\mathbb{R}^{d\times r},\Phi^T\Phi=I} \left\|A_{in}\backslash M - \Phi Y\right\|_F^2 \leq O(\xi \log \frac{\|A_{in}\|_F^2}{\xi})$, with the output dimension bounded by $r \leq O(k \cdot (\log \frac{\|A_{in}\|_F^2}{\xi})^2)$. The number of marked outliers satisfies $|M| \leq O(z \cdot (\log \frac{\|A_{in}\|_F^2}{\xi})^2)$.*

Firstly, let us compare the bound above with that of Bhaskara et al. (2019) which is the state-of-the-art online PCA algorithm. For direct comparison, let us think of the approximation error ξ as $\frac{1}{\gamma}\left\|A_{\text{in}}\right\|_F^2$, where $\gamma \geq 1$ is a constant value. Then we can obtain the error guarantee $O(\frac{\log\gamma}{\gamma}\left\|A_{\text{in}}\right\|_F^2)$ using an embedding dimension $O(k \cdot (\log\gamma)^2)$ with high probability. It is worth noticing that they assume every point is an inlier, thus their algorithm only works for noise-less setting; while our algorithm is robust to at most z outliers, and enjoys the same error and embedding dimension guarantee over inliers not marked outliers, by discarding $O(z \cdot (\log\gamma)^2)$ points as outliers.

We note that error is bounded by roughly $O(\xi \cdot \log \frac{\|A_{\text{in}}\|_{\text{F}}^2}{\xi})$, where ξ is parameter.

It is an important estimation, because it is not only a parameter to run Algorithm 1 and 2, but also appear in their theoretical guarantees. It is unclear whether this is an artifact of our analysis, or is fundamental for the problem. We note that if we relax the restriction of Assumption 1, and permit a second pass over the data, then we can combine our algorithm with the additive approximation algorithm in incremental fashion, and enjoy an error guarantee with a constant approximation factor. We show our second result as follows.

Theorem 2. *If Assumption 2 is satisfied, and we allow a second-time pass over the point, then there exists an efficient algorithm that upon seeing each A_i, decide to add it to the outlier set M, or outputs an embedding $Y_i \in \mathbb{R}^r$. In the end we have, the error bound on the inliers not marked as outliers is $\min_{\Phi\in\mathbb{R}^{d\times r},\Phi^T\Phi=I} \left\|A_{in}\backslash M - \Phi Y\right\|_F^2 \leq \left\|A_{in} - \text{SVD}_k(A_{in})\right\|_F^2 + \epsilon\xi$, with the output dimension bounded by $r \leq O(\frac{k}{\epsilon^2}(\log \frac{\|A_{in}\|_F^2}{\xi})^4)$.*

On the positive side, the above theorem significantly improves upon the approximation error. On the other hand, there is a trade-off between the approximation error and the output dimension. Indeed, we show that by sacrificing additional $O(\frac{1}{\epsilon^2} \cdot (\log \frac{\|A_{\text{in}}\|_F^2}{\xi})^2)$ factor of embedding dimension, we can improve our approximation ratio from data-dependent to a constant.

Table 1. A comparison with prior algorithms. In the "Online" column, the "≈" of Algorithm 2 means that the algorithm requires the two-passes over data. It turns out that with such one additional pass, the algorithm can make a correction to its prior prediction and hence improves the performance guarantee. In the "Error" column, the "unknown" of Feng et al. (2013a) means that their algorithm does not converge to vanishing approximation error provably. Notably, they only show computational convergence to a stationary point with the unknown statistical property.

Work	Online?	Outlier-robust?	Error
Xu et al. (2012)	✗	✓	$\left\|A_{\text{in}} - \text{SVD}_k(A_{\text{in}})\right\|_F^2$
Cherapanamjeri et al. (2017)	✗	✓	$\left\|A_{\text{in}} - \text{SVD}_k(A_{\text{in}})\right\|_F^2$
Boutsidis et al. (2015)	✓	✗	$\left\|A_{\text{in}} - \text{SVD}_k(A_{\text{in}})\right\|_F^2 + \epsilon \left\|A_{\text{in}}\right\|_F^2$
Bhaskara et al. (2019)	✓	✗	$O((\log \frac{\|A_{\text{in}}\|_F^2}{\xi})^2 \cdot \xi)$
Feng et al. (2013a)	✓	✓	unknown
This work (Algorithm 1)	✓	✓	$O((\log \frac{\|A_{\text{in}}\|_F^2}{\xi})^2 \cdot \xi)$
This work (Algorithm 2)	≈	✓	$\left\|A_{\text{in}} - \text{SVD}_k(A_{\text{in}})\right\|_F^2 + \epsilon\xi$

1.2. Overview of main techniques

Our algorithm is inspired by Bhaskara et al. (2019) in part, but has crucial differences. We present an overview of the techniques below, and highlight our novelty.

1) Online PCA using adaptive sampling. At a high level, the online PCA algorithm processes in phases. In each phase, a new direction would be added to the subspace when it is deemed "significant". By adaptive sampling method, the "significance" of every point is proportional to its residual norm (to be defined) over current subspace. Therefore, each point is either informative (thus directly adding its direction to the subspace) or "not-informative" (thus creating a running sketch to sum these vectors until the sum of their residual norms is informative, then adding its direction to the subspace). Classic adaptive sampling is related to column subset selection. For example, Deshpande & Rademacher (2010); Paul et al. (2015) show that there exists a sub-matrix that projects full points onto its span and enjoy a favorable ratio to the best rank-k approximation. Specifically, we show that adaptive sampling can be slightly modified to sample informative residual norms, such that it can solve online PCA. Moreover, the processes of committing subspace and outputting embedding can be done in a one-shot manner, so that the algorithm enjoys favorable computational complexity. The guarantee of the bound on the number of phases is formally shown Lemma 7.

2) Threshold-based outlier removal. Suppose the data is corrupted by $z \gg k$ outliers. Intuitively, this is challenging in an online model, because if we encounter a point far from the current subspace V, we are not sure if it represents a new direction or is simply an outlier. We show each category of the whole phases by introducing the inlier and outlier phases, which are identified by the residual norm of points under or over the threshold ξ/z. This outlier threshold ensures that once z/k of the distant points are seen, there is a sufficient probability of picking the direction. We show that our adaptive sampling design can choose $O(z \cdot \log \frac{\|A_{\text{in}}\|_F^2}{\xi})$ points being outliers; Putting all these ideas leads to an $O(\xi \cdot \log \frac{\|A_{\text{in}}\|_F^2}{\xi})$ error bound for online PCA; see Section 2.1

3) Incremental additive approximation algorithm to improve error. We define the algorithm in an incremental fashion, when it maintains and incrementally adjusts the objective in each step. For example, our algorithm updates the subspace V and a covariance matrix U incrementally when getting new points. It is commonly applied in online fashion, where the memory and computation cost are limited. Intuitively, if two algorithms have incremental property, we can combine them asynchronously (if the second algorithm requires the results from the first algorithm) or synchronously (if two algorithms are independent). By observing that the error of the algorithm 1 is a new residual, the main idea is to process it to an additive approximation algorithm (for example, the first algorithm of Boutsidis et al. (2015)). However, it require a second pass over the data. We show that we can process the residual of the marked inliers and reduce the error over inliers not marked as outliers; see Section 2.2.

4) Removing the dependence on ξ. Furthermore, we note that the two theorems assumed that the parameter ξ is known and satisfies the bound $\xi \geq \left\|A_{\text{in}} - \text{SVD}_k(A_{\text{in}})\right\|_F^2$, which can only be estimated in hindsight. Indeed, we show that we can remove this assumption by assigning an arbitrary small approximation error, and double its value when sufficient phases are met; see Section 2.3.

1.3. Notations

We use capital letters, for example, M, to denote a matrix, and M_i denotes its ith column. The capital letter A is reserved for the data matrix. The ℓ_2-norm of a vector M_i is denoted by $\|M_i\|$. The Frobenius norm of a matrix M is denoted by $\|M\|_F$. The size of the set is denoted by $|\cdot|$. In the ith phase, the algorithm observes the sample A_i and maintains the subspace V^i. We denote the output embedding $Y_i = (V^i)^T A_i$, where the subspace V^i is a set of normalized residuals over previous phases. The projection of A_i to the space orthogonal to V^i is denoted by $\Pi_{V^i}^{\perp} A_i = A_i - V^i(V^i)^T A_i$. We also use $\Pi_i^{\perp} = I - V^i(V^i)^T$.

Recall that each column of A is a sample. By Assumption 2, the set of all columns of matrix A can be partitioned into inliers and the outliers. That is, $A = A_{\text{in}} \cup A_{\text{out}}$. We assume there exists an upper bound z on the number of outliers: $|A_{\text{out}}| \le z$. Let $\text{SVD}_k(A_{\text{in}})$ be the optimal rank-k approximation of A_{in}, then the optimum error of the k-dimension embedding is

$$\text{OPT}_k = \left\| A_{\text{in}} - \text{SVD}_k(A_{\text{in}}) \right\|_F^2 . \tag{1}$$

1.4. Roadmap

In Section 2, we describe our main algorithms, and in Section 3 we show the guarantees. We conclude our work in Section 4, and defer all the proof details and numerical experiments to the appendix.

2. Main Algorithms

In this section, we present our main algorithms.

2.1. Logarithmic approximation

We present our online robust logarithmic approximation algorithm in Algorithm 1. Intuitively, our algorithm uses adaptive sampling based on residual vectors to form the subspace, and process the outlier removal.

Definition 3. In the iteration i, we assign the residual norm in approximating the new point A_i using the current subspace V^i by

$$\left\| \Pi_{V^i}^{\perp} A_i \right\| = \left\| A_i - V^i (V^i)^T A_i \right\| \tag{2}$$

A collection of A_i would be combined as a "proxy" vector until the sum of these residual square norm values are $\ge O(\xi/k)$. Then the phase ends, and we add the direction of "proxy" vector in that phase to V^i. At a high level, the "proxy" vector with the sum of residual square norm $\ge O(\xi/k)$ is deemed significant, thus adding its direction to V^i will occur a smaller error in the future.

In Bhaskara et al. (2019), the author uses one threshold of residual square norm $O(\xi/k)$ to identify the informative and non-informative points. On top of that, our algorithm adds an additional threshold $O(\xi/z)$ to identify the inliers and outliers. By our assumption, we have $z \gg k$. Thus, we can always get $O(\xi/k) \gg O(\xi/z)$, and classify points into three categories.

1. When the point residual square norm is $\le O(\xi/z)$, we name it non-special inlier, and accumulate them until the sum of their residual norms $\ge O(\xi/k)$. Then the span of the "proxy" vector can be treated as an important direction in the subspace, and we add it to the subspace. If the phase terminates with a non-special inlier, it is called an non-special inlier phase; see Steps 4 – 10.

2. When the point residual square norm is between $O(\xi/k)$ and $O(\xi/z)$, the point could be an non-informative outlier, or an non-informative inlier. Such threshold can be thought as the cross-field of inliers and outliers. We name such point non-special outlier. Then we linearly combine it to the "proxy" vector with probability $O(k \left\| \Pi_{V^i}^{\perp} A_i \right\| / \xi)$. When z/k number of such points are observed, the "proxy" vector is deemed important, and we add it to the subspace. If the phase terminates with a non-special outlier, it is called an non-special outlier phase; see Steps 13 – 18.

3. When the point residual square norm is $\ge O(\xi/k)$, the point could be an informative outlier, or an informative inlier lying in the new directions. Intuitively, there should be a decent number of inliers lying in the new directions, while outliers are separated from other points. Therefore, a special point is an inlier near the new direction with high probability. We name such point special outlier, and add it to the subspace with probability $O(k/z)$. If the phase terminates with a special point, it is called a special phase; see Steps 19 – 21.

We note that when outliers are arbitrarily far from authentic data, rendering distance-based sampling is prone to pick outliers, so that we can mitigate their effect. The special outliers and non-special outliers are distant from the current subspace, so it is reasonable to mark them as an outlier. Yet, it is possible that inliers can also be classified as outliers, for example, when a new direction starts to form. We claim that if one direction has more than z/k points, then once z/k of them are observed, there is a sufficiently large probability of adding this direction to the subspace. Meanwhile, for the directions having $< z/k$ points, the total number of picked points near these directions is $< z$. Thus classifying the points as outliers would only increase a factor 2 in the number of marked outliers.

2.2. Constant approximation

We observe that for each vector A_i, Algorithm 1 outputs an embedding of A_i on the current subspace spanned by V^i

Algorithm 1 Online ORPCA with Logarithmic Approximation Error

```
Require: Matrix $A \in \mathbb{R}^{d\times n}$ whose columns $\{A_i\}$ arrive one by one, parameter $\xi$ that upper bounds the optimal approximation error over the inlier points, an upper bound $z$ on the number of outliers, parameter $k > 1$.
Ensure: The online low-dimensional embedding $y_i$; the subspace $V$ at the end.
$V \leftarrow \emptyset$, $U \leftarrow 0^{d\times d}$, $w \leftarrow 0$, $t \leftarrow 0$, $\alpha \leftarrow 0$, $\beta \leftarrow 0$, $r \leftarrow 2k \log \frac{\|A_{\text{in}}\|_F^2}{\xi}$.
while columns $A_i$ arrive do
    $\Pi_V^\perp A_i \leftarrow A_i - UA_i$, $p_{A_i} \leftarrow \frac{k\left\|\Pi_V^\perp A_i\right\|^2}{512\xi}$.
    if $p_{A_i} < \frac{k}{z}$ then
        $\alpha \leftarrow \alpha + p_{A_i}$, $w \leftarrow w + \mathcal{X}A_i$, where $\mathcal{X}$ is $\pm 1$ uniformly at random.
        if $\alpha \geq 1$ then
            $w' \leftarrow \frac{\Pi_V^\perp w}{\left\|\Pi_V^\perp w\right\|}$, add $w'$ to $V$, $U \leftarrow U + w'w'^T$.
            Reset $w$, $\alpha$, $t$ and $\beta$ to 0.
        end if
    else
        Mark $A_i$ as outliers.
        if $p_{A_i} < 1$ then
            With probability $p_{A_i}$: $t \leftarrow t + \mathcal{X}\frac{A_i}{\sqrt{p_{A_i}}}$, where $\mathcal{X}$ is $\pm 1$ uniformly at random, $\beta \leftarrow \beta + \frac{k}{z}$.
            if $\beta \geq 1$ then
                $t' \leftarrow \frac{\Pi_V^\perp t}{\left\|\Pi_V^\perp t\right\|}$, add $t'$ to $V$, $U \leftarrow U + t't'^T$.
                Reset $w$, $\alpha$, $t$ and $\beta$ to 0.
            end if
        else if $p_{A_i} \geq 1$ then
            With probability $k/z$: Set $A'_i \leftarrow \frac{\Pi_V^\perp A_i}{\left\|\Pi_V^\perp A_i\right\|}$, $w' \leftarrow \frac{\Pi_V^\perp w}{\left\|\Pi_V^\perp w\right\|}$, add $A'_i$ and $w'$ to $V$, $U \leftarrow U + A'_i A'^T_i + w'w'^T$.
            Reset $w$, $\alpha$, $t$ and $\beta$ to 0.
        end if
    end if
    Return the embedding $y_i \leftarrow V^T A_i$, resized to dimension $r$ by adding zeros, so that the dimension of all embedding is fixed.
end while
return $V$.
```

Algorithm 2 Online ORPCA with Constant Approximation

```
Require: Matrix $A \in \mathbb{R}^{d\times n}$ whose columns arrive one by one, parameters $\xi$, $k$ and $\epsilon$.
Ensure: The online low-dimensional embedding $y_i$; the subspace $W$ at the end.
Initialize $V', V'' \leftarrow \emptyset$.
Set output dimension $l \leftarrow O(k'/\epsilon'^2) + O(k \cdot \log \frac{\|A_{\text{in}}\|_F^2}{\xi})$, where the first term is from Theorem 4, and the second from Theorem 5.
while columns $A_i$ arrive do
    Execute Algorithm 1 with input $A_i$; this updates $V'$ and updates $y'_i \leftarrow (V')^T A_i$.
    if $\left\|\Pi_V^\perp A_i\right\|^2 < \frac{\xi}{z}$ then
        Execute a step of OPCA-ADD($\Pi_{V'}^\perp A_i$, $k'$, $\epsilon'$, $\Gamma$), where $k'$, $\epsilon'$, $\Gamma$ are defined in (3); this updates $V''$, and outputs $y''_i \leftarrow (V'')^T(\Pi_{V'}^\perp A_i)$
        Let $W$ be an orthogonal basis for span($V' \cup V''$)
    end if
    Return the embedding $y_i \leftarrow W^T A_i$
end while
return $W$.
```

with the guarantee of the residual square norm. We can pass the residual and its guarantee to the additive algorithm of Theorem 4 only for marked inliers. The final output is the joint embedding of the outputs from the two algorithms.

Additive approximation algorithm. To obtain the desired bound on error and output dimension, we need to use the previous work of Boutsidis et al. (2015). It proposed an algorithm for online PCA. The algorithm requires the knowledge of the Frobenius norm of the entire matrix $\|A\|_F^2$ and achieved an additive error $\epsilon \|A\|_F^2$, and maintains the subspace U by only appending the new direction if necessary when the new sample arrives. We denote it OPCA-ADD, and leverage it into Algorithm 2. We established the following theorem for OPCA-ADD algorithm and will invoke it to analyze Algorithm 2.

Theorem 4. *Given an input matrix $A \in \mathbb{R}^{d\times n}$, a parameter $\epsilon > 0$ and an upper bound Γ on $\|A\|_F^2$, there exist an algorithm for online PCA that, at every time i, upon seeing a vector A_i, outputs an embedding $y_i \in \mathbb{R}^l$, where $l = O(k/\epsilon^2)$, and maintains a matrix V''^i with d rows and orthonormal columns. V''^i is only incremented as the algorithm proceeds. The embedding y_i of the vector A_i is precisely $(V''^i)^T A_i$. One has the guarantee that*

$$\sum_i \left\|A_i - V''^i y_i\right\|_F^2 \leq \left\|A - \mathrm{SVD}_k(A_{in})\right\|_F^2 + \epsilon\Gamma.$$

We now formally present the constant approximation algorithm in Algorithm 2. The OPCA-ADD refers to the additive approximation of Boutsidis et al. (2015).

Define the following:

$$k' = 20k \cdot \log \frac{\|A_{\text{in}}\|_F^2}{\xi}, \tag{3}$$

$$\epsilon' = \frac{\epsilon}{\log \frac{\|A_{\text{in}}\|_F^2}{\xi}}, \tag{4}$$

$$\Gamma = \xi \cdot \log \frac{\|A_{\text{in}}\|_F^2}{\xi}. \tag{5}$$

In Algorithm 2, when a point A_i arrives, we first feed it to Algorithm 1 and update V' as necessary. If it is non-special inlier, we apply the residual projection $\Pi_{V'}^{\perp} A_i$ to OPCA-ADD and get the second set V''. It is worth noting that we only execute OPCA-ADD algorithm for non-special inliers. This is because we do not need to refine the cost we have incurred on marked outliers. The output W is the union of V' and V''.

2.3. Remove the dependence on ξ

The assumption on having a parameter ξ such that $\xi \geq \text{OPT}_k$ is important, because it not only appears in the theoretical guarantee but also is actually required to run the algorithm. In the Algorithm 3, we show how to apply a general way to remove this assumption, at the expense of an additional factor of $\log\left(\frac{\left\|A_{\text{in}} - \text{SVD}_k(A_{\text{in}})\right\|_F^2}{\xi}\right)$ in the embedding dimension and the number of marked outliers.

Let us denote $L_\delta = \log \frac{\|A_{\text{in}}\|_F^2}{\xi} + \log(1/\delta)$. We show the removing procedure in two steps. Firstly, we show an analog of Theorem 5 without the assumption on ξ. Then for the same incremental property, we can run the residuals through the instantiation of the algorithm in Boutsidis et al. (2015) and output the constant approximation.

First step. We start with the given value of ξ_0, and run Algorithm 1. If the total number of phases with the current ξ_0 exceeds kL_δ, we conclude that ξ_0 is too small, and double ξ_0. Once $\xi_0 \geq \left\|A_{\text{in}} - \text{SVD}_k(A_{\text{in}})\right\|_F^2$, we will no longer exceed the bound on the number of phases. The number of doubling steps needed is $\log \left\|A_{\text{in}} - \text{SVD}_k(A_{\text{in}})\right\|_F^2 / \xi$. Since this number is bounded by L_δ, with probability at least $1-\delta$, the output dimension becomes $O(kL_\delta^2)$ and the number of marked outliers becomes $O(zL_\delta^2)$, which establishes Theorem 1.

Second step. We observe that Algorithm 3 also has the incremental property. That is, we maintain the subset $V_{\text{old}} + V$ and output the projection on this space. In the end, we have an $O(L_\delta)$ approximation to the error. Thereby, we can combine it with Algorithm 1 from Boutsidis et al. (2015), with the following parameters:

$$k' = 20k \cdot L_\delta^2, \epsilon' = \epsilon \cdot \frac{1}{L_\delta}, \Gamma = \xi \cdot L_\delta. \tag{6}$$

Algorithm 3 Online ORPCA Logarithmic Approximation without Parameter ξ

```
Require: Matrix $A \in \mathbb{R}^{d \times n}$ whose columns $\{A_i\}$ arrive
    one by one, arbitrary $\xi$, and parameters $k$, $\epsilon$.
Ensure: The online low-dimensional embedding $y_i$; the
    subspace $V_{\text{old}}$ at the end.
Initialize $V_{\text{old}} \leftarrow \emptyset$, $V \leftarrow \emptyset$, $U_{\text{old}} \leftarrow \emptyset$, $U \leftarrow 0^{d \times d}$,
    $w \leftarrow 0$, $t \leftarrow 0$ and running sum $\alpha \leftarrow 0$, $\beta \leftarrow 0$.
while columns $A_i$ arrive do
    Invoke Algorithm 1 with input $\Pi_{V_{\text{old}}}^{\perp} A_i$; this updates
        updates $V$.
    Return the embedding $y_i \leftarrow V_{\text{old}}^T A_i \cup V^T A_i$.
    if number of phases (the dimension of $V$) exceeds
        $kL_\delta$ then
        $V_{\text{old}} \leftarrow V_{\text{old}} \cup V$; $U_{\text{old}} \leftarrow U_{\text{old}+U}$; $V \leftarrow \emptyset, \xi \leftarrow 2\xi$.
    end if
end while
return $V_{\text{old}}$.
```

Thus the number of columns used overall is $O(k'/\epsilon'^2) = O(\frac{kL_\delta^4}{\epsilon^2})$. This establishes Theorem 2. We defer the detailed proof to the appendix.

3. Performance Guarantee

We state the guarantee of our algorithms. Recall that the analysis of Algorithm 3 has been given in Section 2.3.

3.1. Logarithmic approximation

We start by showing the following theorem that characterizes the performance of Algorithm 1, which is almost our result of Theorem 1, except for the requirement of $\xi \geq \text{OPT}_k$.

Theorem 5. *If Assumption 1 and Assumption 2 are satisfied, and $\delta > 0$, then with probability $1-\delta$, Algorithm 1 satisfies: the number of phases, and the number of columns r of the subspace V, is $\leq O(k \cdot \log \frac{\|A_{in}\|_F^2}{\xi} + \log 1/\delta)$. The number of points marked as outliers is $O(z \cdot \log \frac{\|A_{in}\|_F^2}{\xi} + \frac{z}{k} \log 1/\delta)$. The objective cost for the inlier points not marked as outliers is $O(\xi \cdot \log \frac{\|A_{in}\|_F^2}{\xi} + \frac{\xi}{k} \log 1/\delta)$. The running time of each step is $O(d^2)$.*

To ease the discussion of the theorem, we need the following definition.

Definition 6. A phase is said to be a majority outlier phase if one of the following conditions hold:

1. The phase is non-special inlier and the following inequality holds (where $\{u_i\}_{i=1}^r$ are the non-special inliers in the phase): $\sum_{i \in [r] \wedge u_i \in A \setminus A_{in}} \frac{k\left\|\Pi_V^{\perp} u_i\right\|^2}{\xi} \geq \frac{1}{2}$.
2. The phase is non-special outlier and the following in-

equality holds (where $\{u_i\}_{i=1}^r$ are the non-special outliers in the phase): $\sum_{i\in[r]\wedge u_i\in A\setminus A_{in}} \frac{k}{z} \geq \frac{1}{2}$.

If a phase is not majority outlier, then it is said to be a majority inlier phase.

Observe that for a majority outlier phases, the number of outliers in the phase has the lower bound $\frac{z}{2k}$, whereas we have an upper bound for the total number of outliers z. Thereby, the number of a successful majority outliers phases must be at most $O(k)$, otherwise the number of outliers would exceed z, which contradict to our Assumption 2.

For majority inlier phases, we show that they are either special, non-special inlier or non-special outlier. To get the support of their bounds on number, we use a crucial geometric lemma proposed in Bhaskara et al. (2019). The key observation is that if each vector has a non-trivial orthogonal component to all of the preceding vectors, we can find upper bound the total number of such vectors in terms of k.

Lemma 7 (Bhaskara et al. (2019))**.** *Let $v_1, v_2, ..., v_r \in \mathbb{R}^d$ be a set of linearly independent vectors, $r \leq d$. Let $c > 0$ be any constant, and let Γ be a parameter satisfying $\Gamma \geq \frac{1}{c}\left\|V - \mathrm{SVD}_k(k)\right\|_F^2$. Suppose that v_i satisfy $\left\|\Pi_{i-1}^{\perp}\right\|^2 \geq \gamma$. Suppose additionally that $\gamma^2 \geq \frac{2c\Gamma}{k}$. Then the number of columns r satisfies the bound $r \leq 2k \cdot \log\left(\frac{\|V\|_F^2}{2c\Gamma}\right)$.*

By lemma 7, we immediately have the estimate of number of special phases, since the total number of the special points over inliers and outliers is $O(k \cdot \log\frac{\|A_{\text{in}}\|_F^2}{\xi} + z)$, and we accept them with probability k/z. Recall $k/z < 1$, by the law of large numbers, we have the number of special phases $O(k \cdot \log\frac{\|A_{\text{in}}\|_F^2}{\xi})$.

The following definition summarizes the sense that deemed "successful", in which we require w to be a "proxy" for the phase. We note that the analysis of non-special inlier and non-special outlier phases are similar, so we only show that of non-special inlier phase. For the full and detailed proof, readers can refer to the appendix.

Lemma 8. *(Non-special inlier phases with majority inliers) Let the phase be a majority inlier non-special inlier phase. Let $\{u_i\}_{i=1}^t$ be the non-special inliers in the phase. Then with probability at least $1/4$ we have:*

1. $\left\|\Pi_V^{\perp} w\right\|^2 \geq \frac{1}{8}\sum_{i=0}^t \left\|\Pi_V^{\perp} u_i\right\|^2$.

2. *If Π_k is the projection matrix orthogonal to the k-SVD space of the inliers A_{in}, then*

$$\|\Pi_k w\|^2 \leq 16\sum_{i=1}^t \|\Pi_k u_i\|^2$$

3. $\|w\|^2 \leq 16\sum_{i=1}^t \|u_i\|^2$

Definition 9. A non-special inlier phase is said to be successful if all the inequalities in Lemma 8 are satisfied. Otherwise, it is said unsuccessful.

Combining the lemma 7 and 8, we get that the number of successful majority inlier phases of non-special inlier and non-special outlier is $O(k \cdot \log\frac{\|A_{\text{in}}\|_F^2}{\xi})$.

Consider each phase as a coin toss, then the successful phase can be thought as the head result of the coin. Assume that we know the number of successful phases, by the Chernoff bound, we can see that it is very unlikely that the number of unsuccessful phase is much larger.

Now we are ready to present the proof sketch of our main theorem.

Proof Sketch of Theorem 5. Combined the conclusion above, we can get that with probability at least $1-\delta$, the total number of phases is $O(k \cdot \log\frac{\|A_{\text{in}}\|_F^2}{\xi} + \log(1/\delta))$. We remark that for either special, non-special inlier, non-special outlier phases, the cumulative probability over non-special inliers should be bounded by 2, which shows that the cost of inliers not marked as outliers in each phase is $O(\xi/k)$. Thereby, the total cost over inliers not marked as outliers is $O(\xi \cdot \log\frac{\|A_{\text{in}}\|_F^2}{\xi} + \frac{\xi}{k}\log(1/\delta))$. Similarly, we observe that for either phase, the number of points marked as outliers should be bounded by $\frac{2z}{k}$. Multiply it with the total number of phases we have the total number of marked outliers is $O(z \cdot \log\frac{\|A_{\text{in}}\|_F^2}{\xi} + \frac{z}{k}\log 1/\delta)$. For the running time, we note that in each iteration, we only need to maintain the covariance matrix $U = VV^T$, where $V \in \mathbb{R}^{d\times 1}$. Thus, the running time is only $O(d^2)$.

3.2. Constant approximation

Then we present the guarantee of Algorithm 2. Again, this result is almost our result of Theorem 2, except for the requirement of $\xi \geq \mathrm{OPT}_k$.

Theorem 10. *If Assumption 2 is satisfied, and we allow a second-time pass over the point, and $\delta > 0$, then we have that with probability at least $1-\delta$, Algorithm 2 satisfies: the number of phase and the number of columns r to be $\leq O(\frac{k}{\epsilon^2}(\log\frac{\|A_{in}\|_F^2}{\xi} + \log 1/\delta)^3)$. The objective cost for the output embedding Y, over the points not marked as outliers $\leq OPT_k + \epsilon\xi$.*

Proof Sketch of Theorem 10. We show that the residual squared norm in Algorithm 1 is bounded by $O(\xi \cdot \log\frac{\|A_{\text{in}}\|_F^2}{\xi})$, so we can set $\epsilon' = \epsilon/(\log\frac{\|A_{\text{in}}\|_F^2}{\xi})$ and satisfy the desired additive error. By Theorem 4, we have $\sum_i \left\|A_i - W^T y_i\right\|_F^2 \leq \mathrm{OPT}_k + \epsilon\xi$ (it is satisfied because we set $k' \geq k$, so $\mathrm{OPT}_{k'} \leq \mathrm{OPT}_k$). Thus we get the desired error bound. We plugged the value from Eq. (3) to the embedding dimension l, and the dominant part in the dimension is $O((\frac{k}{\epsilon^2})(\log \frac{\|A_{\text{in}}\|_F^2}{\xi} + \log 1/\delta)^3)$.

4. Conclusion and Future Work

In this paper, we have presented a robust PCA algorithm in the online manner that builds up a low-rank embedding by adding the new directions to the subspace in each iteration, when the data is corrupted by outliers. Prior to this work, existing PCA algorithms either only considered partial assumption, or failed to present the convergence analysis within finite data. To our best knowledge, this paper is the first to provide a provable algorithm using sampling method.

We raise three open questions for future work. Firstly, our work assumes that the inlier data are noiseless. It would always be interesting to examine whether our model can preserve the guarantee when the inliers are with additive noise. Secondly, the number of outliers marked by our algorithm slightly violates the bounds on the actual number of outliers z and the dimension of the subspace k by a logarithmic factor. It would be interesting to examine whether we can get a better dependence on these parameters. For example, Bhaskara & Kumar (2018) proposes an offline robust sampling algorithm that violates the number of outliers by only a factor $(1 + \delta)$, where δ is an additional input, which improves the dependence on z from a logarithmic factor to a constant ratio. Finally, our work focuses on finding low-rank approximation under Frobenius error, and it would be interesting to study whether our residual-based algorithm can solve the problem under general ℓ_p error.

Acknowledgements

We thank the anonymous reviewers and meta-reviewer for valuable comments on improving the presentation. This work is supported by NSF-IIS-1948133 and the startup funding from Stevens Institute of Technology.

References

Allen-Zhu, Z. and Li, Y. First efficient convergence for streaming k-PCA: A global, gap-free, and near-optimal rate. In *Proceedings of the 58th IEEE Annual Symposium on Foundations of Computer Science*, pp. 487–492, 2017.

Amid, E. and Warmuth, M. K. An implicit form of krasulina's k-PCA update without the orthonormality constraint. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pp. 3179–3186, 2020.

Arora, R., Cotter, A., and Srebro, N. Stochastic optimization of PCA with capped MSG. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pp. 1815–1823, 2013.

Bhaskara, A. and Kumar, S. Low rank approximation in the presence of outliers. In *Proceedings of Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, volume 116, pp. 4:1–4:16, 2018.

Bhaskara, A., Lattanzi, S., Vassilvitskii, S., and Zadimoghaddam, M. Residual based sampling for online low rank approximation. In *Proceedings of the 60th IEEE Annual Symposium on Foundations of Computer Science*, pp. 1596–1614, 2019.

Boutsidis, C., Garber, D., Karnin, Z. S., and Liberty, E. Online principal components analysis. In *Proceedings of the 2015 Annual ACM-SIAM Symposium on Discrete Algorithms*, 2015.

Brubaker, S. Robust PCA and clustering in noisy mixtures. pp. 1078–1087, 01 2009.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM*, 58(3):11:1–11:37, 2011.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

Cherapanamjeri, Y., Jain, P., and Netrapalli, P. Thresholding based outlier robust PCA. In *Proceedings of the 30th Conference on Learning Theory*, pp. 593–628, 2017.

Croux, C., Filzmoser, P., and Oliveira, M. R. Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 87: 218–225, 06 2007.

Deshpande, A. and Pratap, R. Sampling-based dimension reduction for subspace approximation with outliers. *Theoretical Computer Science*, 858:100–113, 2021.

Deshpande, A. and Rademacher, L. Efficient volume sampling for row/column subset selection. In *Proceedings of the 51th Annual IEEE Symposium on Foundations of Computer Science*, pp. 329–338, 2010.

Feng, J., Xu, H., Mannor, S., and Yan, S. Online PCA for contaminated data. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pp. 764–772, 2013a.

Feng, J., Xu, H., and Yan, S. Online robust PCA via stochastic optimization. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pp. 404–412, 2013b.

Garber, D. On the regret minimization of nonconvex online gradient ascent for online PCA. In *Proceedings of the 32th Conference on Learning Theory*, pp. 1349–1373, 2019.

Garber, D. and Hazan, E. Fast and simple PCA via convex optimization. *CoRR*, abs/1509.05647, 2015.

Hallgren, F. and Northrop, P. Incremental kernel PCA and the nyström method. *CoRR*, abs/1802.00043, 2018.

Jain, P., Jin, C., Kakade, S. M., Netrapalli, P., and Sidford, A. Streaming PCA: matching matrix bernstein and near-optimal finite sample guarantees for Oja's algorithm. In *Proceedings of the 29th Conference on Learning Theory*, pp. 1147–1164, 2016.

Jalali, A. and Srebro, N. Clustering using max-norm constrained optimization. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Klivans, A. R., Long, P. M., and Servedio, R. A. Learning halfspaces with malicious noise. In *Proceedings of the 36th International Colloquium, Automata, Languages and Programming*, pp. 609–621, 2009.

Krasulina, T. P. The method of stochastic approximation for the determination of the least eigenvalue of a symmetrical matrix. *Ussr Computational Mathematics and Mathematical Physics*, 9:189–195, 1969.

la Torre, F. D. and Black, M. J. Robust principal component analysis for computer vision. In *Proceedings of the 8th International Conference On Computer Vision*, pp. 362–369, 2001.

Lee, J. D., Recht, B., Salakhutdinov, R., Srebro, N., and Tropp, J. A. Practical large-scale optimization for max-norm regularization. In *Proceedings of the 24th Annual Conference on Neural Information Processing Systems*, pp. 1297–1305, 2010.

Lerman, G. and Maunu, T. An overview of robust subspace recovery. *Proceedings of the IEEE*, 106(8):1380–1410, 2018.

Li, C.-L., Lin, H.-T., and Lu, C.-J. Rivalry of two families of algorithms for memory-restricted streaming PCA. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 473–481, 2016.

Nie, J., Kotlowski, W., and Warmuth, M. K. Online PCA with optimal regrets. In *Proceedings of 24th International Conference on Algorithmic Learning Theory*, pp. 98–112, 2013.

Oja, E. and Karhunen, J. On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix. *Journal of Mathematical Analysis and Applications*, 106(1):69–84, 1985.

Ozawa, S., Pang, S., and Kasabov, N. A modified incremental principal component analysis for on-line learning of feature space and classifier. In *Proceedings of the 8th Pacific Rim International Conference on Artificial Intelligence*, 2004.

Paul, S., Magdon-Ismail, M., and Drineas, P. Column selection via adaptive sampling. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems*, pp. 406–414, 2015.

Shen, J., Xu, H., and Li, P. Online optimization for max-norm regularization. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*, pp. 1718–1726, 2014.

Srebro, N. and Shraibman, A. Rank, trace-norm and max-norm. In *Proceedings of the 18th Annual Conference on Learning Theory*, pp. 545–560, 2005.

Srebro, N., Rennie, J. D. M., and Jaakkola, T. S. Maximum-margin matrix factorization. In *Proceedings of the 17th Annual Conference on Neural Information Processing Systems*, pp. 1329–1336, 2004.

Tang, C. Exponentially convergent stochastic k-PCA without variance reduction. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems*, pp. 12393–12404, 2019.

Xu, H., Caramanis, C., and Sanghavi, S. Robust PCA via outlier pursuit. *IEEE Transactions on Information Theory*, 58(5):3047–3064, 2012.

Xu, L. and Yuille, A. L. Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Transactions on Neural Networks*, 6(1): 131–143, 1995.

Yang, T. and Wang, S. Robust algorithms for principal component analysis. *Pattern Recognition Letters*, 20(9): 927–933, 1999.

A. Restatement of Useful Notations

Recall we have a set of n points with d features, represented by a matrix $A \in \mathbb{R}^{d\times n}$, and a low-rank embedding dimension $k < \min\{d, n\}$. A can be partitioned into inliers and the outliers. That is, $A = A_{\text{in}} \cup A_{\text{out}}$. We assume there exists an upper bound z on the number of outliers: $|A_{\text{out}}| \le z$. We also set ξ be the approximation error over A_{in}, and $\mathrm{SVD}_k(A_{\text{in}})$ be the optimal rank-k approximation of A_{in}, $\xi \ge \left\|A_{\text{in}} - \mathrm{SVD}_k(A_{\text{in}})\right\|_F^2$.

B. Omitted Proofs for Tightness

In this section, we show an example to explain why the logarithmic term $\log \frac{\|A_{\text{in}}\|_F^2}{\xi}$ is a unavailable for residual-based sampling algorithm.

Lemma 11 (Bhaskara et al. (2019))**.** *Let $k = 1$, and let $t, z > 2$ be parameters that will be fixed shortly. There exists a matrix A of dimensions $t \times t$, such that $\|A\|_F^2 = z^{2t}$, the rank-1 approximation error $\left\|A - \mathit{SVD}\text{-}1(A)\right\|_F^2 \le 2t^2/z^2$, and further, such that every column of A has a squared projection at least 1 orthogonal to the previous columns.*

Proof. We choose the matrix in the following, where $z > 2$. It is easy to see that each column has a length 1 orthogonal to the previous columns. The bound on the Frobenius norm is also easy to check.

$$M = \begin{pmatrix} 1 & z^2 & z^3 & \cdots & z^{t-1} \\ 0 & 1 & z^2 & \cdots & z^{t-2} \\ & & \cdots & & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Let $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_t \ge 0$ be the singular values of M. We demonstrate an explicit subspace S of dimension $t-1$ such that

$$\max_{x\in S, \|x\|=1} \|Mx\|^2 \le \frac{2t}{z^2}$$

By the min-max characterization of singular values, this implies the desired claim. Now define

$$S = \{x \in \mathbb{R}^t : \frac{x_1}{z^{t-1}} + \frac{x_2}{z^{t-2}} + \cdots + x_t = 0\}$$

Take any unit vector $x \in S$. Consider the ith coordinate of Mx. This is precisely

$$(Mx)_i = x_i + \frac{x_{i+1}}{z} + \cdots + \frac{x_t}{z^{t-1}} - (\frac{x_{t-1}}{z} + \frac{x_{t-2}}{z^2} + \cdots + \frac{x_1}{z^{t-1}})$$

where we used the definition of $x \in S$. Thus by Cauchy-Schwartz, we have that

$$(Mx)_i^2 \le t(\frac{x_{t-1}^2}{z^2} + \frac{x_{t-2}^2}{z^4} + \cdots + \frac{x_1^2}{z^{2t-2}})$$

Since $z > 2$, we have $\sum_i (Mx)_i^2 \le \frac{2t}{z} \sum_i x_i^2$, thus proving $\sigma_2^2 \le \frac{2t}{z^2}$. This immediately implies that $\left\|A - \mathrm{SVD}_k(A)\right\|_F^2 \le \frac{2t^2}{z^2}$. □

Assume $z = 2t$, and $\xi = 1$. Now we know that $\left\|A - \mathrm{SVD}_k(A)\right\|_F^2 \le \xi$. If we run an algorithm that samples the columns (as they arrive) with probability $\min(1, \left\|\Pi^\perp A_i\right\|_2^2/\xi)$ (as in our algorithms, $\Pi^\perp$ is the projection orthogonal to the chosen columns), this algorithm will indeed pick all the columns. Thus the algorithm is choosing $\Omega(k \log \frac{\|A_{\text{in}}\|_F^2}{\xi} / \log\log \frac{\|A_{\text{in}}\|_F^2}{\xi})$ columns. This matches the upper bound up to a $\log \frac{\|A_{\text{in}}\|_F^2}{\xi}$ factor.

C. Omitted Proofs for Logarithmic Approximation Algorithm

We can get the number of majority outlier phases immediately from the Definition 6.

Lemma 12. *The total number of majority outlier phases is $2k$.*

Proof. Suppose we have t majority outliers phases.

Let us consider one of these phases.

- If the phase is special or non-special outlier, and all points in that phase are $v_1, v_2, ..., v_r$, by the definition we would have $\sum_{i\in[r]\wedge v_i\in V\setminus V_{in}} \frac{k}{z} \geq \frac{1}{2}$.
- If the phase is non-special inlier, and all points in that phase are $v_1, v_2, ..., v_r$, by definition we would have $\sum_{i\in[r]\wedge v_i\in V\setminus V_{\text{in}}} \frac{k\left\|\Pi_V^\perp v\right\|^2}{\xi} \geq \frac{1}{2}$, and since $\frac{k}{z} \geq \frac{k\left\|\Pi_V^\perp v\right\|^2}{\xi}$, we can also obtain that $\sum_{i\in[r]\wedge v_i\in V\setminus V_{\text{in}}} \frac{k}{z} \geq \frac{1}{2}$.

Therefore we can see that in either case, the number of outlier points in a phase is at lease $\frac{z}{2k}$.

Suppose that $t > 2k$, we would have the number of outliers $> z$, which contradicts to the setting of the upper bound z on the number of outliers. This implies $t \leq 2k$. That is, the number of majority outliers phases is $2k$. □

Before talking about the majority inlier phase, we state the non-trivial Geometric Lemma, which would be used in the majority inlier phase analysis.

Lemma 13 (Restatement of Lemma 7)**.** *. Let $v_1, v_2, ..., v_r \in \mathbb{R}^d$ be a set of linearly independent vectors, $r \leq d$. Let $c > 0$ be any constant, and let Γ be a parameter satisfying $\Gamma \geq \frac{1}{c}\left\|V - V^{(k)}\right\|_F^2$. Suppose that v_i satisfy $\left\|\Pi_{i-1}^\perp\right\|^2 \geq \gamma$. Suppose additionally that $\gamma^2 \geq \frac{2c\Gamma}{k}$. Then the number of columns r satisfies the bound $r \leq 2k \cdot \log\left(\frac{\|V\|_F^2}{2c\Gamma}\right)$.*

Proof. Let K be the parallelopiped formed by the columns of V. It is well-known that $\text{vol}(K) = \sqrt{\det(V^TV)}$. The volume of the parallelopiped can also be computed iteratively using the "base times height" formula. If ℓ_i is the length of the projection of v_i orthogonal to span$\{v_1, ..., v_{i-1}\}$, then the volume is precisely $\Pi_{i=1}^r \ell_i$. In our case, this is at least γ^r by hypothesis.

Let $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$ be the singular values of the matrix V. Now, if $\left\|V - V^{(k)}\right\|_F^2 \leq c\Gamma$, then $\sigma_{2k+1}^2 \leq \frac{c\Gamma}{k}$. For suppose not, then:

$$\sigma_{k+1}^2 + \sigma_{k+2}^2 + ... + \sigma_{2k+1}^2 \geq (k+1)\sigma_{2k+1}^2 \geq c\Gamma,$$

contradicting the bound on $\left\|V - V^{(k)}\right\|_F^2$. Next, using the formula for the volume, we have

$$\left(\Pi_{i=1}^{2k}\sigma_i\right)\left(\frac{c\Gamma}{k}\right)^{(r-2k)/2} \geq \gamma^r.$$

Now, using standard convexity, we have

$$\Pi_{i=1}^{2k}\sigma_i \leq \left(\frac{\sum_{i=1}^{2k}\sigma_i}{2k}\right)^{2k} \leq \left(\frac{\sum_{i=1}^{2k}\sigma_i^2}{2k}\right)^{k} \leq \left(\frac{\|V\|_F^2}{2k}\right)^{k}.$$

Combine the two equations above, we have

$$\left(\frac{\|V\|_F^2}{2c\Gamma}\right)^{k} \geq \left(\frac{\gamma^2 k}{c\Gamma}\right)^{r/2}.$$

Take logarithms on both sides we have

$$k \cdot \log\left(\frac{\|V\|_F^2}{2c\Gamma}\right) \geq \frac{r}{2} \cdot \log\left(\frac{\gamma^2 k}{c\Gamma}\right).$$

Thereby we get

$$r \leq 2k \cdot \frac{\log\big(\frac{\|V\|_F^2}{2c\Gamma}\big)}{\log\big(\frac{\gamma^2 k}{c\Gamma}\big)},$$

where we assume $\gamma^2 \geq \frac{2c\Gamma}{k}$, so $\log\big(\frac{\gamma^2 k}{c\Gamma}\big) \geq \log 2$, which gives the desired bound $r \leq 2k \cdot \log\big(\frac{\|V\|_F^2}{2c\Gamma}\big)$. □

Applying the Geometric Lemma, we can get the number of mjaority inliers phase.

Lemma 14. *The total number of majority inliers special phase is* $2k \log \frac{\|A_{in}^2\|}{2\xi}$.

Proof. We claim that for the arriving point A_i, $p_{A_i} = \frac{k\left\|\Pi_V^\perp A_i\right\|^2}{\xi}$, where $n \geq 2$ is a constant number.

Let T be the matrix consisting of all the columns that made the phases special. Note that any column of T must have squared projection $\geq n\xi/k$ orthogonal to the span of all previous columns of T.

T is the subset of A, so we have $\|T\|_F^2 \leq \|A_{\text{in}}\|_F^2$, and $\left\|T - T^{(k)}\right\|_F^2 \leq \left\|A_{\text{in}} - \text{SVD}_k(A_{\text{in}})\right\|_F^2 \leq \xi$. Thus the hypothesis of lemma 7 holds. This implies that

$$\#\text{cols}(T) \leq 2k \log \frac{\|T\|_F^2}{n\xi} \leq 2k \log \frac{\|T\|_F^2}{2\xi} \leq 2k \log \frac{\left\|A_{\text{in}}^2\right\|}{2\xi}.$$

□

Lemma 15 (Restatement of Lemma 8)**.** *(Non-special inlier phases with majority inliers) Let the phase be a majority inlier non-special inlier phase. Let* $\{u_i\}_{i=1}^t$ *be the non-special inliers in the phase. Then with probability at least* $1/4$ *we have:*

1. $\left\|\Pi_V^\perp w\right\|^2 \geq \frac{1}{8} \sum_{i=0}^t \left\|\Pi_V^\perp u_i\right\|^2$.
2. *If* Π_k *is the projection matrix orthogonal to the* k*-SVD space of the inliers* A_{in}*, then*

$$\|\Pi_k w\|^2 \leq 16 \sum_{i=1}^t \|\Pi_k u_i\|^2$$

3. $\|w\|^2 \leq 16 \sum_{i=1}^t \|u_i\|^2$

Proof. By definition, $w = \sum_{i=1}^t \mathcal{X}_i A_i$ where $\mathcal{X}_i$ is the uniformly at random sign for vector A_i. In the end of the phase, we have $\Pi_V^\perp w = \sum_{i=1}^t \mathcal{X}_i \Pi_V^\perp A_i$. Denote the random variable Z to be $\left\|\Pi_V^\perp w\right\|^2$, we can obtain $Z = (\sum_{i=1}^t \Pi_V^\perp \mathcal{X}_i A_i)^2 = \sum_{i=1}^t \left\|\Pi_V^\perp A_i\right\|^2 + 2\sum_{i<j} \mathcal{X}_i \mathcal{X}_j \left\|\Pi_V^\perp A_i\right\| \left\|\Pi_V^\perp A_j\right\|$. By the linearity of expectation, we have:

$$\mathbb{E}[Z] = \sum_{1 \leq i,j \leq t} \mathbb{E}\left[\mathcal{X}_i \mathcal{X}_j\right] \left\|\Pi_V^\perp A_i\right\| \cdot \left\|\Pi_V^\perp A_j\right\| = \sum_{i=1}^t \left\|\Pi_V^\perp A_i\right\|^2,$$

where the second equality holds because $\mathbb{E}\left[\mathcal{X}_i \mathcal{X}_j\right]$ is 0 for any $i \neq j$ and 1 for $i = j$. We can apply the Paley-Zygmund inequality to lower bound the probability of this event:

$$\Pr\left[Z \geq \frac{2-\sqrt{3}}{2}\mathbb{E}[Z]\right] \geq \frac{3\mathbb{E}[Z]^2}{4\mathbb{E}[Z^2]} \tag{7}$$

We expand $\mathbb{E}[Z]^2$ as:

$$\mathbb{E}[Z]^2 = \sum_{i=1}^t \left\|\Pi_V^\perp A_i\right\|^4 + \sum_{i<j} 2\left\|\Pi_V^\perp A_i\right\|^2 \left\|\Pi_V^\perp A_j\right\|^2$$

Then we need to figure out the value of

$$\mathbb{E}[Z^2] = \mathbb{E}\left[\left(\sum_i \left\|\Pi_V^\perp A_i\right\|^2 + 2\sum_{i<j} \mathcal{X}_i\mathcal{X}_j\langle \Pi_V^\perp A_i, \Pi_V^\perp A_j\rangle\right)^2\right]$$

We note any term with an odd power of $\mathcal{X}_i$ has expectation of 0. The remaining terms has the coefficients $\mathcal{X}_i^4$ or $\mathcal{X}_i^2\mathcal{X}_j^2$, which equals to 1. Therefore we have:

$$\mathbb{E}[Z^2] = \sum_{i=1}^t \left\|\Pi_V^\perp A_i\right\|^4 + 4\sum_{i<j}\langle \Pi_V^\perp A_i, \Pi_V^\perp A_j\rangle^2$$

Applying Cauchy-Schwartz inequality to the inner products above and then we can obtain $\mathbb{E}[Z^2] \leq 2\mathbb{E}[Z]^2$. Applying this into Equation 7 we obtain:

$$\Pr\left[Z \geq \frac{2-\sqrt{3}}{2}\mathbb{E}[Z]\right] \geq \frac{3}{8}$$

We also observe that $\frac{2-\sqrt{3}}{2}$ is at least $1/8$. Thus we have w.p. $3/8$, $\left\|\Pi_V^\perp w\right\|^2 \geq \frac{1}{8}\sum_{i=0}^t \left\|\Pi_V^\perp A_i\right\|^2$. This ensure the first statement holds.

Assume $Z = \|\Pi_k w\|^2$ where Π_k is the projection matrix orthogonal to the k-SVD space of the matrix A_{in}. Applying the Markov inequality we have:

$$\Pr\left(Z \geq 16\mathbb{E}[Z]\right) \leq \frac{1}{16}$$

Therefore we have w.p. $15/16$, $\|\Pi_k w\|^2 \leq 16\mathbb{E}\left[\|\Pi_k w\|^2\right]$. From the proof of the first statement we have $\mathbb{E}\left[\|\Pi_k w\|^2\right] = \sum_{i=1}^t \|\Pi_k A_i\|^2$. The second statement holds.

The proof of the third statement is identical to the second statement (Assuming $Z = \|w\|^2$, and applying it to Markov inequality). So it also holds w.p. $15/16$. Combine the results above , we conclude that with probability at least $1 - 5/8 - 1/16 - 1/16 = 1/4$, both of the inequalities hold. □

Similar to successful non-special inlier phases, we show the analysis of the successful non-special outlier phases.

Lemma 16. *(Majority inliers non-special outlier) Let the phase be a majority inlier non-special outlier phase. Let $\{A_i\}_{i=1}^r$ be the non-special outliers in the phase, V be the basis subspace at beginning of the phase, and t be the linear combination of the residuals from picked points at the end of phase. The phase is said to be successful with probability $1/80$:*

- $\left\|\Pi_V^\perp t\right\|^2 \geq \frac{1}{2}\sum_{i=1}^r \left\|\Pi_V^\perp A_i\right\|^2.$
- *If Π_k is the projection matrix orthogonal to the k-SVD space of the matrix A_{in}, then*

$$\|\Pi_k t\|^2 \leq 40\sum_{i=1}^r \|\Pi_k A_i\|^2$$

- $\|t\|^2 \leq 40\sum_{i=1}^r \|A_i\|^2$

Proof. Let $A_1, ..., A_r$ be the columns in a phase. Let Y_i be a indicator showing whether A_i is picked or not. By definition, we have $\Pr[y_i = 1] = p_{A_i}$.

Using above definition, we have

$$t = \sum_{i=1}^r \mathcal{X}_i Y_i \frac{A_i}{\sqrt{p_{A_i}}}.$$

Firstly, let's compute the expectation of residuals of A_i, we show

$$\mathbb{E}\left[\left\|\Pi_V^\perp t\right\|^2\right] = \sum_{i=1}^{r} \mathbb{E}[Y_i^2] \frac{\left\|\Pi_V^\perp A_i\right\|^2}{p_{A_i}} + 2\sum_{i<j} \mathbb{E}\left[\mathcal{X}_i \mathcal{X}_j\right] \mathbb{E}[Y_i Y_j] \frac{\langle \Pi_V^\perp A_i, \Pi_V^\perp A_j \rangle}{\sqrt{p_{A_i} p_{A_j}}} = \sum_{i=1}^{r} \left\|\Pi_V^\perp A_i\right\|^2$$

This holds because $\mathbb{E}\left[\mathcal{X}_i \mathcal{X}_j\right]$ is zero for $i \neq j$, and $\mathbb{E}[Y_i^2] = \mathbb{E}[Y_i] = p_{A_i}$.

We denote the random variable Z to be $\left\|\Pi_V^\perp t\right\|^2$ and apply it to Paley-Zygmund inequality shown in 7. We can easily obtain that $E[Z]^2 = \sum_{i=1}^r \left\|\Pi_V^\perp A_i\right\|^4 + \sum_{i<j} 2 \left\|\Pi_V^\perp A_i\right\|^2 \left\|\Pi_V^\perp A_j\right\|^2$ .Then we need to compute $\mathbb{E}[Z^2]$.

$$\mathbb{E}[Z^2] = \sum_{i=1}^{r} p_{A_i} \left\|\Pi_V^\perp \frac{A_i}{\sqrt{p_{A_i}}}\right\|^4 + 6\sum_{i<j} p_{A_i} p_{A_j} \langle \Pi_V^\perp \frac{A_i}{\sqrt{p_{A_i}}}, \Pi_V^\perp \frac{A_j}{\sqrt{p_{A_j}}} \rangle^2$$

Applying Cauchy-Schwartz inequality to the second part we obtain:

$$\mathbb{E}[Z^2] \leq \sum_{i=1}^{r} \frac{\left\|\Pi_V^\perp A_i\right\|^4}{p_{A_i}} + 6\sum_{i<j} \left\|\Pi_V^\perp A_i\right\|^2 \left\|\Pi_V^\perp A_j\right\|^2.$$

For the first term, since the phase is non-special, we have $p_{A_i} = \frac{k}{\xi} \left\|\Pi_V^\perp A_i\right\|^2$, so we can obtain

$$\sum_{i=1}^{r} \frac{\left\|\Pi_V^\perp A_i\right\|^4}{p_{A_i}} = \frac{\xi}{k} \sum_{i=1}^{r} \left\|\Pi_V^\perp A_i\right\|^2$$

The phase ends when $\sum_{i=1}^r p_{A_i} \geq 1$, so equivalently we have $\sum_{i=1}^r \left\|\Pi_V^\perp A_i\right\|^2 \geq \frac{\xi}{k}$. Combine this with the inequality above we have:

$$\frac{\xi}{k} \sum_{i=1}^{r} \left\|\Pi_V^\perp A_i\right\|^4 \leq \left(\sum_{i=1}^{r} \left\|\Pi_V^\perp A_i\right\|^2\right)^2$$

The second term ($6\sum_{i<j} \left\|\Pi_V^\perp A_i\right\|^2 \left\|\Pi_V^\perp A_j\right\|^2$) is not more than $3\mathbb{E}[Z^2]$. Thus we conclude that $\mathbb{E}[Z^2] \geq 4\mathbb{E}[Z]^2$. Applying it to 7 we have Pr $\left[Z \geq 1/2\mathbb{E}[Z]\right] \geq \frac{1}{16}$. This ensures that the first statement holds with probability 1/16.

Assume $Z = \|\Pi_k w\|^2$ where Π_k is the projection matrix orthogonal to the k-SVD space of the matrix A_{in}. Applying this to Markov inequality we have:

$$\text{Pr}\left(Z \geq 40\mathbb{E}[Z]\right) \leq \frac{1}{40}$$

Therefore we have w.p. $39/40$, $\|\Pi_k w\|^2 \leq 40\mathbb{E}\left[\|\Pi_k w\|^2\right]$. From the proof of the first statement we have $\mathbb{E}\left[\|\Pi_k w\|^2\right] = \sum_{i=1}^t \|\Pi_k A_i\|^2$. The second statement holds.

The proof of the third statement is identical to the second statement (Assuming $Z = \|w\|^2$, and applying it to Markov inequality). So it also holds w.p. $39/40$. Combine the results above, we conclude that with probability at least $1 - 15/16 - 1/40 - 1/40 = 1/80$, both of the inequalities hold together. □

Lemma 17. *The number of successful majority inliers non-special inlier phases is* $2k \log \frac{\|A_{in}^2\|}{2\xi}$.

Proof. We claim that for the arriving point A_i, $p_{A_i} = \frac{k\left\|\Pi_V^\perp A_i\right\|}{n\xi}$, where $n \geq 512$ is a constant number.

By the definition of a successful majority inlier non-special inlier phase, we have: $\sum_{i\in[r]\wedge A_i \in V_{in}} \left\|\Pi_V^\perp A_i\right\|^2 \geq \frac{n\xi}{2k}$. By the first inequality in the definition of a successful majority inliner non-special inlier phases, we have $\left\|\Pi_V^\perp w\right\|^2 \geq \frac{n\xi}{16k}$.

Let T be the matrix whose columns are all the vectors w at the end of successful majority inlier non-special inlier phases. The span of the columns of T is a subspace of the span of the columns of V. By the third inequality in the definition of a successful phase, we have $\|T\|^2 \leq 16 \|A_{\text{in}}\|^2$. Also, by the second inequality, we have $\left\|T - T^{(k)}\right\|_F^2 \leq 16 \left\|A_{\text{in}} - \text{SVD}_k(A_{\text{in}})\right\|_F^2$. Thus ξ satisfies that $\xi \geq \frac{1}{16} \left\|T - T^{(k)}\right\|_F^2$.

Above inequality implies that

$$\#\text{cols}(T) \leq 2k \log \frac{16 \|T\|^2}{n\xi} \leq 2k \log \frac{\|T\|^2}{32\xi} \leq 2k \log \frac{\left\|A_{\text{in}}^2\right\|}{2\xi}.$$

□

Lemma 18. *The number of successful majority inliers non-special outliers phases is* $2k \log \frac{\left\|A_{in}^2\right\|}{2\xi}$.

Proof. We claim that for the arriving point A_i, $p_{A_i} = \frac{k\left\|\Pi_V^\perp A_i\right\|}{n\xi}$, where $n \geq 320$ is a constant number.

By the definition of a successful majority inlier non-special outlier phase, we have:$\sum_{i\in[r]\wedge A_i\in V_{\text{in}}} \frac{k}{z} \geq \frac{1}{2}$, and since $\frac{k\left\|\Pi_V^\perp A_i\right\|^2}{\xi} \geq \frac{k}{z}$, we can also obtain $\sum_{i\in[r]\wedge A_i\in V_{\text{in}}} \frac{k\left\|\Pi_V^\perp A_i\right\|^2}{n\xi} \geq \sum_{i\in[r]\wedge A_i\in V_{\text{in}}} \frac{k}{z} \geq \frac{1}{2}$. Thus, $\sum_{i\in[r]\wedge A_i\in V_{in}} \left\|\Pi_V^\perp A_i\right\|^2 \geq \frac{n\xi}{2k}$. By the first inequality in the definition of a successful majority inliner non-special outlier phases, we have $\left\|\Pi_V^\perp w\right\|^2 \geq \frac{n\xi}{4k}$.

Let T be the matrix whose columns are all the vectors t at the end of successful non-special majority inlier non-special outlier phases. The span of the columns of T is a subspace of the span of the columns of V. By the third inequality in the definition of a successful phase, we have $\|T\|^2 \leq 40 \|A_{\text{in}}\|^2$. Also, by the second inequality, we have $\left\|T - T^{(k)}\right\|_F^2 \leq 40 \left\|A_{\text{in}} - \text{SVD}_k(A_{\text{in}})\right\|_F^2$. Thus ξ satisfies that $\xi \geq \frac{1}{40} \left\|T - T^{(k)}\right\|_F^2$.

Above inequality implies that

$$\#\text{cols}(T) \leq 2k \log \frac{4 \|T\|^2}{n\xi} \leq 2k \log \frac{\|T\|^2}{80\xi} \leq 2k \log \frac{\|A_{\text{in}}\|^2}{2\xi}.$$

□

Given the number of successful phase of majority inliers non-special inlier and outlier phases, we want to get their number of total phases. Formally, we need the following result.

Theorem 19. *(Bhaskara et al., 2019) We toss a coin n times. The tosses are independent of each other and in each toss, the probability of seeing a head is at least p. Let H_m and T_m denote the number of heads and tails we observe in the first $m \leq n$ coin tosses. With probability $1 - \delta$, we have $H_m \geq \frac{pm}{4} - \lceil 8log(\frac{2}{\delta}/p)$ for any $1 \leq m \leq n$. We note that although the claim is about conjunction of all these n events, the probability does not rely on n.*

Proof. We denote μ the expected number of heads in the first m tosses, which is at least pm. The lower tail inequality of Chernoff Bounds implies

$$\Pr[H_m < (1 - \frac{1}{2})\mu] \leq e^{-\mu/8} \leq e^{-pm/8}$$

The error probability $e^{-pm/8}$ is at most $\delta/2$ for $m \geq m' = \lceil 8\log(2/\delta)/p \rceil$. Instead of summing up the error bound for all values of m, we focus on the smaller geometrically growing sequence $M = \{2^\ell m' | \ell \in \mathbb{Z}^{\geq 0} \text{ AND } 2^\ell m' \leq n\}$. Having the lower bound on H_m for every $m \in M$ helps us achieve a universal lower bound on any $1 \leq m \leq n$ as follows. For any $m \leq m'$, the bound $H_m \geq pm - m'$ holds trivially.

For any other $m \leq n$, there exists an $m'' \in M$ such that $m'' \leq m \leq 2m''$. By definition H_m is at least $H_{m''}$. Assuming $H_{m''} \geq pm''/2$ implies H_m is at least $pm/4$ which proves the claim of the lemma. So we focus on bounding the error probabilities for values in set M.

For m', the error probability is at most $\delta/2$. The next value in M is $2m'$, so given the exponential form of the error, it is at most $(\delta/2)^2$. Using union bound, the aggregate error probability for set M does not exceed

$$\frac{\delta}{2} + (\frac{\delta}{2})^2 + (\frac{\delta}{2})^4 + ... \le \frac{\delta/2}{1-\delta/2} \le \delta.$$

Therefore with probability at least $1-\delta$, we have for every $m \in M$, $H_m \ge pm/2$, and consequently for every $1 \le m \le n$, $H_m \ge \frac{pm}{4} - m'$ which finishes the proof. □

Using lemma 19, we see that if H_m is the number of head coins, then with probability at least $1-\delta$, the total number of coin tosses $m = O(H_m + \log(\frac{1}{\delta}))$.

Lemma 20. *For any $\delta > 0$, with probability at least $1-\delta$, the total number of majority inliers non-special phases is $O(k \cdot \log \frac{\|A_{in}\|_F^2}{\xi} + \log(1/\delta))$.*

Proof. We consider each phase as coin toss. A head in the coin toss is associated with the phase being a successful one. For majority inliers non-special inlier phases, each phase has a probability $\ge 1/4$ of being successful. Applying it to the Theorem 19, with probability $1-\delta/2$, the bound on the number of majority inliers non-special inlier phases is $O(k \cdot k \cdot \log \frac{\|A_{\text{in}}\|_F^2}{\xi} + \log(1/\delta))$. Similarly, for majority inliers non-special outlier phases, each phase has a probability $\ge 1/80$ of successful. We also have that with probability $1-\delta/2$, the bound on the number of majority inliers non-special outlier phases is $O(k \cdot k \cdot \log \frac{\|A_{\text{in}}\|_F^2}{\xi} + \log(1/\delta))$.

Then the probability that both bounds hold is $\ge (1-\delta/2)^2 \ge 1-\delta$, and the number of the majority inliers non-special phases is $O(k \cdot \log \frac{\|A_{\text{in}}\|_F^2}{\xi} + \log(1/\delta))$. □

Lemma 21. *For any $\delta > 0$, with probability at least $1-\delta$, the total number of phases is $O(k \cdot \log \frac{\|A_{in}\|_F^2}{\xi} + \log(1/\delta))$.*

Proof. Combining the Lemmas 12, 14 and 20, we can immediately prove this lemma. □

Then, we get the number of columns in subspace, which is also the embedding dimension, $|V|$.

Lemma 22. *Given any $\delta > 0$, with probability at least $1-\delta$, the number of columns in subspace and the embedding dimension is $\le O(k \cdot \log \frac{\|A_{in}\|_F^2}{\xi} + \log(1/\delta))$.*

Proof. Every special phases and non-special outlier phases output at most two columns, while the non-special inlier phases output one column, so we have additional $O(k \cdot \log \frac{\|A_{\text{in}}\|_F^2}{\xi})$ number of columns compared with the number of phases, which implies this lemma. □

Then we are ready to prove our error guarantee and the number of marked outliers.

Lemma 23. *Define the inlier points not marked as outliers as $A_{in}\backslash M$. Given any $\delta > 0$, in the end, with probability $1-\delta$,*

$$\min_{\Phi \in \mathbb{R}^{d\times r}, \Phi^T\Phi = I} \left\| A_{in}\backslash M - \Phi Y \right\|_F^2 \le \xi \cdot O(\log \frac{\|A_{in}\|_F^2}{\xi} + \log(1/\delta)/k).$$

Proof. Firstly, we show the bound on the cost in each phase. To this end, let $A_1, ..., A_r$ be the points in a phase, and let V_{pre} be the basis vector of points at the start of the phase, and V_{cur} be the basis vector of points after the vector A_r has been processed. We now consider two cases,

- If the phase is non-special inlier:

  $$\sum_{i\in[r]\wedge A_i \in V_{\text{in}}\backslash M} \left\| \Pi^{\perp}_{V_{\text{cur}}} u \right\|^2 \le \sum_{i\in[r]\wedge A_i \in V_{in}\backslash M} \left\| \Pi^{\perp}_{V_{pre}} u \right\|^2 \le \sum_{i\in[r]\wedge A_i \in V_{in}\backslash M} p_{A_i} \frac{\xi}{k} \le 2\frac{\xi}{k}.$$

  The last inequality follows from that the sum of the selection probabilities of non-special inliers in a non-special inlier phase is ≤ 2.

- If the phase is special or non-special outlier:

$$\sum_{i\in[r]\wedge A_i\in V_{\text{in}}\setminus M}\left\|\Pi^{\perp}_{V_{\text{cur}}}u\right\|^2 \le \sum_{i\in[r]\wedge A_i\in V_{in}\setminus M}\left\|\Pi^{\perp}_{V_{pre}}u\right\|^2 \le \sum_{i\in[r]\wedge A_i\in V_{in}\setminus M} p_{A_i}\frac{\xi}{k} \le \frac{\xi}{k}.$$

  The last inequality follows from that the sum of the selection probabilities of non-special inliers or outliers is < 1.

This implies that in either case $\sum_{i\in[r]\wedge A_i\in V_{\text{in}}\setminus M}\left\|\Pi^{\perp}_{V_{\text{cur}}}u\right\| \le 2\frac{\xi}{k}$. Combined with the bound on the number of phases, we have that with probability at least $1-\delta$, the total cost over the inlier points which are not marked as outliers is $\xi \cdot O(\log\frac{\|A_{\text{in}}\|_F^2}{\xi} + \log{(1/\delta)}/k)$. □

Lemma 24. *The number of marked outliers satisfies $|M| \le z \cdot O(\log\frac{\|A_{in}\|_F^2}{\xi} + \frac{\log{(1/\delta)}}{k})$.*

Proof. We will first show the bound on number of points marked as outliers in each phase. Let the set of points marked as outliers be M. We consider three cases.

- case 1: The phase is non-special inliers:

$$\sum_{i\in[r]\wedge u\in M}\frac{k}{z} < 1.$$

  The inequality follows from the definition that if the phase is non-special inlier, then $\beta < 1$.

- case 2: The phase is non-special outlier:

$$\sum_{i\in[r]\wedge u\in M}\frac{k}{z} < 2.$$

  The inequality follows from the definition that if the phase is non-special outlier, then $1 \le \beta < 2$.

- case 3: The phase is special:

$$\sum_{i\in[r-1]\wedge u\in M}\frac{k}{z} < 1.$$

Therefore, we can see that in case 1 and case 2, $|i \in [r] \wedge u \in M| \le \frac{2z}{k}$. In case 3, we have the bound on the number of phase being $1+\frac{z}{k}$, while by our setting $\frac{z}{k} > 1$, we also obtain $|i\in[r-1]\wedge u \in M| + 1 \le \frac{2z}{k}$. Combined with the bound on the number of phases, with probability at least $1-\delta$, $|M| \le \frac{2z}{k}O(kL + \log(1/\delta)) = z\cdot O(\log\frac{\|A_{\text{in}}\|_F^2}{\xi} + \frac{\log(1/\delta)}{k})$. So we have the desire bound on the second term. □

Theorem 25 (Restatement of Theorem 5)**.** *If Assumption 1 and Assumption 2 are satisfied, and $\delta > 0$, then with probability $1-\delta$, Algorithm 1 satisfies: the number of phases, and the number of columns r of the subspace V, is $\le O(k\cdot\log\frac{\|A_{in}\|_F^2}{\xi} + \log 1/\delta)$. The number of points marked as outliers is $O(z\cdot\log\frac{\|A_{in}\|_F^2}{\xi} + \frac{z}{k}\log 1/\delta)$. The objective cost for the inlier points not marked as outliers is $O(\xi\cdot\log\frac{\|A_{in}\|_F^2}{\xi} + \frac{\xi}{k}\log 1/\delta)$. The running time of each step is $O(d^2)$.*

Proof. This is an immediate result by combining Lemmas 22, 24 and 23. The proof for running time has been stated in the last part of Section 2.1. □

D. Omitted Proofs for Constant Approximation Algorithm

We remark that since we only process the marked inliers to Algorithm 2, the guarantee for the number of marked outliers is identical. Thus in the section, we only show the analysis of output embedding and the total cost over inliers not marked as outliers.

Lemma 26. *The cost for the output embedding Y, over the points not marked as outliers is $\le OPT_k + \epsilon\xi$.*

Proof. In the ith iteration, let V_i' denote the subspace that Algorithm 1 maintains, and r_i' denote the residual of the Algorithm 1, $r_i' = \Pi^{\perp}_{V_i'} A_i$, which is also the input of the Algorithm 2. Let V_i'' denote the subspace that Algorithm 2 maintains and r_i'' denote its residual, $r_i'' = \Pi^{\perp}_{V_i''} r_i'$. Let R' denote the matrix whose ith column is r_i', and R'' denote the matrix whose ith column is r_i''.

For our choice of Γ, we have that with probability at least $1-\delta$,

$$\sum_i \left\|r_i'\right\|^2 \le \Gamma.$$

Then with probability at least $1-\delta$, we can get the error guarantee from Theorem 4

$$\sum_i \left\|r_i''\right\|^2 \le \left\|R' - (R')^{(k')}\right\| + \epsilon'\Gamma. \tag{8}$$

Based on the notion of residual, we have that $r_i' = A_i - V_i' y_i'$, and according to Theorem 5, V_i' at any time contains at most $O(kL + k + \log(1/\delta))$. Thus the rank-$k'$ approximation to the matrix R' has error at most the rank-k approximation to the matrix A_i. Namely,

$$\left\|R' - R'^{(k')}\right\|^2 \le \left\|A_{\text{in}} - \text{SVD}_k(A_{\text{in}})\right\|^2 \tag{9}$$

Based on Equation 3, we have $\epsilon\xi = \epsilon'\Gamma$. Combining it with Equation 8 and 9, we get:

$$\sum_i \left\|r_i''\right\|^2 \le \left\|A_{\text{in}} - \text{SVD}_k(A_{\text{in}})\right\|_F^2 + \epsilon\xi \tag{10}$$

As a final step, we observe that

$$r_i'' = r_i' - V_i'' y_i'' = A_i - V_i' y_i' - V_i'' y_i''.$$

This is the difference of A_i and a liner combination of the space $V_i' \cup V_i''$. Thus the length of r_i'' is upper bounded by the distance of A_i to the span of the space $V_i'n \cup V_i''n$, which is $A_i - W_i W_i^T A_i$. Namely, we get:

$$\sum_i \left\|A_i - W_i W_i^T A_i\right\|^2 \le \sum_i \left\|r_i''\right\|^2. \tag{11}$$

Combining Equation 10 with 11, and since $Y = \sum_i W_i^T A_i$, we get

$$\|A - WY\|_F^2 \le \left\|A_{\text{in}} - \text{SVD}_k(A_{\text{in}})\right\|_F^2 + \epsilon\xi,$$

which completes the proof. □

Lemma 27. *For any $\delta > 0$, with probability at least $1-\delta$, the number of phase and the number of output dimension r is $\le O(\frac{k}{\epsilon^2}(\log\frac{\|A_{in}\|_F^2}{\xi} + \log 1/\delta)^3)$.*

Proof. Since the dominating term in output dimension l is $O(k'/\epsilon')$. Plugging in the values from Equation 3 completes the proof. □

Theorem 28 (Restatement of Theorem 10)**.** *If Assumption 1 and Assumption 2 are satisfied, and $\delta > 0$, then we have that w.p. at least $1-\delta$, Algorithm 2 satisfies: the number of phase and the number of output dimension r is $\le O(\frac{k}{\epsilon^2}(\log\frac{\|A_{in}\|_F^2}{\xi} + \log 1/\delta)^3)$. The objective cost for the output embedding Y, over the points not marked as outliers is $\le OPT_k + \epsilon\xi$.*

Proof. This is an immediate result by combining Lemmas 26 and 27. □

E. Omitted Proofs for Algorithm without the dependence on ξ

E.1. Proof of Theorem 1

In this section, we present the proof for guarantee of the Logarithmic Approximation algorithm without the assumption $\xi \geq \mathrm{OPT}_k$. The guarantee for Constant Approximation has been explained in Section 2.3.

Recall $L_\delta = \log \frac{\|A_{\text{in}}\|_F^2}{\xi} + \log(1/\delta)$, and $0 < \delta \leq 1$.

Lemma 29. *For any $\delta > 0$, with probability at least $1 - \delta$, the output dimension is $O(kL_\delta^2)$.*

Proof. Assume we have a set of vectors $A = \{A_1, A_2, ..., A_n\}$. For a fixed subspace T, let $B = \{B_1, B_2, ..., B_n\}$ denote the orthogonal projections from B to T, that is, $B_i = \Pi_T^\perp A_i$. By the fact that the rank-k error only reduces upon projection, we get $\left\|A - A^{(k)}\right\|_F^2 \geq \left\|B - B^{(k)}\right\|_F^2$. By Algorithm 3, the guessing approximation error ξ_j does not increase once $\xi_j \geq \left\|A - A^{(k)}\right\|_F^2$, so the total number of the doubling is $\log \frac{\left\|A-A^{(k)}\right\|_F^2}{\xi}$, which is trivially bounded by L_δ, so the output dimension is bounded by $L_\delta \cdot O(kL_\delta) = O(kL_\delta^2)$. □

Lemma 30. *Suppose we have the initial approximation error ξ_0, and $\delta > 0$, in the end, with probability at least $1 - \delta$, the cost over inliers not marked as outliers is at most $O(\xi_0 L_\delta)$.*

Proof. After doubling j times, the guessing approximation error is $\xi_j = 2^j \xi_0$. Since the number of phases of each ξ_j is bounded by kL_δ, and the cumulative residual errors over non-special inliers in each phase is $< \frac{2\xi_j}{k}$, we conclude that the total residual error over inliers not marked as outliers is at most

$$kL_\delta \left(\frac{2\xi_0}{k} + \frac{2^2\xi_0}{k} + ... + \frac{2^{j+1}\xi_0}{k} \right) \leq O(\xi L_\delta).$$

□

Lemma 31. *For any $\delta > 0$, with probability at least $1 - \delta$, the total number of marked outliers is $O(zL_\delta^2)$.*

Proof. Since the number of doublings is bounded by L_δ, the number of phases of each ξ_j is bounded by kL_δ, and the number of marked outliers in each phase is $< 2z/k$, we conclude that the total number of marked outliers is bounded by

$$kL_\delta \cdot (2z/k + 2z/k + ... + 2z/k) = O(zL_\delta^2)$$

This establishes Theorem 1. □

F. Numerical Results

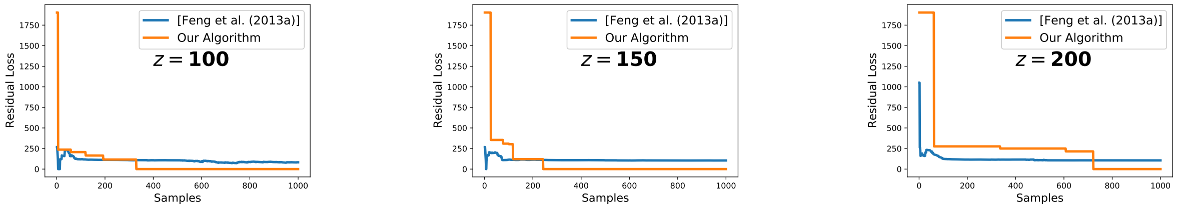
In this section, we report some numerical results on synthetic data. Our goal is to illustrate the properties of the online robust algorithm discussed in section 1, and compare our residual-based sampling for online robust PCA algorithm with the algorithm in Feng et al. (2013a), which uses matrix decomposition to update the subspace in each iteration.

To make a fair comparison, we simulate the contaminated data as follows. We randomly generate an $d \times k$ matrix A, and scale it to make its magnitudes of the leading eigenvalues $= 2$. Then we multiple A with another uniformly generated matrix $X = \mathbb{R}^{k \times n}$ to make $L = AX$. A fraction λ of outliers are generated with uniform distribution over $[-20, 20]$, where $z = \lambda n$ is the number of outliers.

The algorithm will return a subspace V after receiving every sample, and we will use it to compute the residual loss over inliers $\left\|A_{\text{in}} - VV^T A_{\text{in}}\right\|_F$ as the performance.

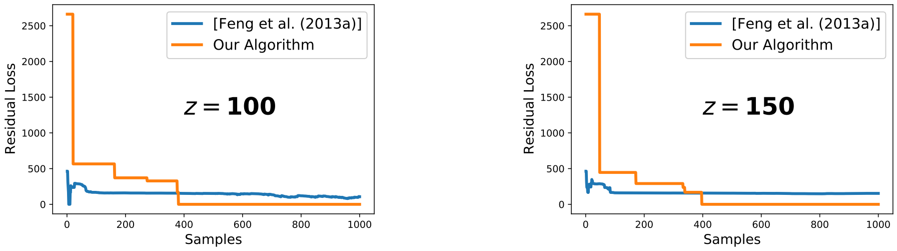
We firstly show the simulations with total $n = 1000$ samples and $d = 500$ dimensions. Simulation results for optimum low rank $k = 5$ with different number of outliers $z = 100, 150, 200$ have been shown in Figure 1.

While results show that our algorithm can recover the low-rank structure of inlier points, we also care about the dimension of the output embedding, the number of the marked outliers. By the bicriteria approximation, we do not expect these numbers



(a) (b) (c)

Figure 1. Performance comparison of our algorithm (red line) with Algorithm in Feng et al. (2013a) (blue line). Here $d = 500$, $n = 1000$, $k = 5$. We observe that Algorithm in Feng et al. (2013a) converges faster, but our algorithm can get a more accurate approximation. We notice that the red line is a broken line, while blue line is smooth. It makes sense because our algorithm only updates the subspace when the new direction is informative, while Feng et al. (2013a) updates subspace almost in every iteration.



(a) (b)

Figure 2. Performance comparison of our algorithm (red line) with Feng et al. (2013a) (blue line). Here $d = 1000$, $n = 1000$, $k = 5$.

too large compared with true values. Moreover, we assume our Algorithm enjoys better computation time than Feng et al. (2013a). Table 2, 3 and 4 support our assumption.

Figure 2 and Table 5, 6 show results of the similar numerical study for $n = 1000$ samples, and $d = 1000$ dimensions, where we observe similar trend.

Table 2. Comparison for the embedding dimension, marked outliers and execution time of our algorithm and Feng et al. (2013a) when $d = 500$, $k = 5$, $z = 100$. We observe that our algorithm sacrifices more embedding dimension and number of marked outliers to get the approximation. In contrast, Feng et al. (2013a) can keep the dimension, but mark too few points as outliers. We also show that out algorithm is faster than Feng et al. (2013a).

Algorithm	Embedding Dimension	#Marked Outliers	Running Time (s)
Our algorithm	12	132	7.37
Feng et al. (2013a)	5	29	41.52

Table 3. Comparison for the embedding dimension, marked outliers and execution time when $d = 500$, $k = 5$, $z = 150$.

Algorithm	Embedding Dimension	#Marked Outliers	Running Time (s)
Our algorithm	14	235	9.26
Feng et al. (2013a)	5	50	43.79

Table 4. Comparison for the embedding dimension, marked outliers and execution time when $d = 500$, $k = 5$, $z = 200$.

Algorithm	Embedding Dimension	#Marked Outliers	Running Time (s)
Our algorithm	6	564	8.42
Feng et al. (2013a)	5	58	50.96

Table 5. Comparison for the embedding dimension, marked outliers and execution time when $d = 1000$, $k = 5$, $z = 100$.

Algorithm	Embedding Dimension	#Marked Outliers	Running Time (s)
Our algorithm	10	358	35.14
Feng et al. (2013a)	5	20	325.01

Table 6. Comparison for the embedding dimension, marked outliers and running time when $d = 1000$, $k = 5$, $z = 150$.

Algorithm	Embedding Dimension	#Marked Outliers	Running Time (s)
Our algorithm	14	391	29.08
Feng et al. (2013a)	5	46	292.80