Streaming Robust PCA

Yang Shi shiy4@uci.edu
University of California, Irvine
U.N. Niranjan Niranjan.Uma@microsoft.com
Microsoft Corporation

Abstract

In this paper, we consider the problem of robust PCA in the streaming setting with space constraints. The problem can be stated as follows: at time t, we are given a n-dimensional data vector $x_t = uz_t + s_t$ where u is a fixed vector, z_t is a Gaussian random variable and s_t is an arbitrary sparse perturbation. Without storing samples, we wish to recover u and subsequently also separate the sparse perturbation s_t from each sample. Essentially, our algorithm performs simple iterative hard-thresholding followed by stochastic block power method. Our algorithm also has the optimal space complexity of $O(n)$ and a sample complexity of $O(n \log n)$.

1 Introduction

The robust PCA problem addresses the following question: suppose we are given a data matrix which is the sum of an unknown low-rank matrix and an unknown sparse matrix, can we recover each of the component matrices? Despite the inherent non-convexity of the problem, recent advances have provided algorithms with near-optimal convergence guarantees. However, these bounds hold only in the batch setting, ie, when the entire data matrix is known. In the present work, we analyze robust PCA in the streaming setting, focusing on the rank-1 case where we would like to recover the top eigenvector of the true covariance without the perturbation effect due to sparse corruptions.

1.1 Our Contribution

To the best of our knowledge, we obtain the first-known convergence guarantees from streaming robust PCA while having finite sample complexity of $O(n \log n)$ and also having optimal space complexity of $O(n)$ where n is the dimension; the precise result is stated in Theorem 3.1. The assumptions that we use are natural identifiability assumptions used in the batch case as well, the details of which are presented in Section 3. At a high level, our algorithm performs alternating hard-thresholding followed by stochastic block power method. Two specific improvements from earlier works are: (1) we have the weaker deterministic assumption for the sparse perturbation (2) We do not need incoherence of the intermediate updates in our analysis.

1.2 Related work

PCA: *Principal Component Analysis (PCA)* is an ubiquitous unsupervised learning algorithm and has a rich history. Oja's algorithm is a classical method for streaming PCA Oja and Karhunen [1985]. Though the convergence and empirical performance were known, the asymptotic convergence rate was first provided in Balsubramani et al. [2013]. Improving on the analysis of Balsubramani et al. [2013], linear convergence is presented in Shamir [2015] but with the requirement that the initialization vector must have a constant correlation with the true eigenvector. The convergence of block stochastic power method is considered in Mitliagkas et al. [2013] for PCA in the streaming setting. Recently, a tighter analysis for Oja's algorithm is provided in Jain et al. [2016]. Also, Alecton is a SGD algorithm for low-rank matrix problems presented in De Sa et al. [2014]. Their analysis is based on control of martingales to achieve $O\left(\frac{1}{\epsilon}\right)$ convergence rate where ϵ is the desired numerical error.

Robust PCA: The convergence of the non-convex alternating projections based method was analyzed in Netrapalli et al. [2014] in the batch setting. Recently, a projected gradient method on factorized matrices was presented in Yi et al. [2016]. They also match the time complexity lower bound of $\tilde{O}(rn^2)$ in the fully observed setting and also provide guarantees under the partially observed setting, however, only in the batch setting. For the online setting, the work by He et al He et al. [2011] presented an algorithm based on online ℓ_1-minimization which also had good empirical performance.

OPTML 2017: 10th NIPS Workshop on *Optimization for Machine Learning* (NIPS 2017).

2 Problem Setup

2.1 Model

We consider the popular spiked-covariance model with sparse perturbations in n-dimensions, ie, $x_t = Az_t + s_t$ where A is an unknown $n \times r$ matrix of rank-r and s_t is deterministic sparse perturbation with unknown support and magnitude. Given a sample x_t at time t, we wish to recover s_t and with finite such samples, we wish to find the space spanned by the columns of A upto a fixed numerical accuracy ϵ. In other words, if $A = U\Sigma V^\top$ is the SVD, we wish to find the eigenvectors U. In this paper, we will focus only on the rank-1 case, ie, at time instance t, the data vector is given by $x_t = uz_t + s_t$ where $\|u\|_2 = 1$.

2.2 Notations and Assumptions

We introduce natural (standard) conditions, similar to Netrapalli et al. [2014] under which the problem is identifiable:

1. Low-rank part: For simplicity we have assumed that the eigenvalue corresponding to the top eigenvector is 1. u is μ-incoherent, ie, $\|u\|_\infty \leq \frac{\mu}{\sqrt{n}}$ and $z_t \overset{iid}{\sim} \mathcal{N}(0,1)$. We note that we may relax this generative assumption on z_t to a random variable such that $\mathbb{E}[z_t] = 0, \mathbb{E}[z_t^2] = 1, |z_t| \leq Z_{\max}$ almost surely with a little care.
2. Sparse part: we have a deterministic sparsity condition, ie, $\|s_t\|_0 \leq d_h$ and also without loss of generality, we assume $\|s_t\|_\infty \leq \frac{s_{\max}}{\sqrt{n}}$.

Let b_i denote the i^{th} basis vector in n dimensions. Define the entry-wise hard-thresholding operation of a vector v, denoted as $\mathrm{Thresh}_a(v)$ as follows: for every i,

$$\mathrm{Thresh}_a(b_i^\top v) = \begin{cases} b_i^\top v, & \text{if } \left|b_i^\top v\right| > a \\ 0, & \text{else} \end{cases}$$

We use $h, t, and \tau$ to denote outermost,middle, and innermost loop indices in Algorithm 1. For all τ, $e_t^\tau = s_t - \hat{s}_t^\tau$ where $\hat{s}_t^\tau$ is our estimate of s_t resulting from the τ-th thresholding step. Let $\hat{s}_t = \hat{s}_t^T$, $e_t = e_t^T$ for every t. u_h is the estimated u after h outermost loops. Note that we can decompose u_h as $u_h = \pm\left(\sqrt{1-\alpha_h}u + \sqrt{\alpha_h}v_h\right)$, where $u \perp v_h$ and $\alpha_h \in (0,1)$, for every $h \geq 1$. Let $\Sigma_h = \sum_{t=B(h-1)+1}^{Bh} \frac{1}{B}(x_t - \hat{s}_t)(x_t - \hat{s}_t)^\top$ denote our estimate of the true covariance at epoch h, $\Sigma = uu^T$, and define $\Delta_h = \Sigma - \Sigma_h$. We define quantities $Z_{h-1} = Z_{\max}\left(\alpha_{h-1}\frac{\mu}{\sqrt{n}} + \sqrt{\alpha_{h-1}(1-\alpha_{h-1})}\right)$. We assume $d_h < \min\left(\frac{1}{100}\frac{n}{\mu^2 + 3\mu n\alpha_{h-1}}, \frac{1}{250 Z_{\max}^2\sqrt{\alpha_{h-1}}}\right)$ where d_h is the number of non-zeros in s_t that appears at epoch h.

$\|.\|_2$ denotes the two-norm of a vector or the spectral norm of a matrix, $\|.\|_\infty$ denotes the maximum of the absolute values of the entries of a vector or a matrix, $\|.\|_1$ is the sum of absolute values of entries of a vector, $\|.\|_0$ is the number of non-zeros in a vector.

2.3 Algorithm

We present our algorithm for the rank-1 case in Algorithm 1. There are three key loops in Algorithm 1 namely, (1) (innermost) τ-loop which we call alternations, (2) (middle) t-loop which we call iterations, and (3) the (outermost) h-loop which we call epochs. Our algorithm uses random initialization for our eigenvector estimate, which is also very easy in practice. Intuitively, the τ-loop is performs denoising via *iterative hard-thresholding*, ie, it solves the optimization problem:

$$\{\hat{z}_t, \hat{s}_t\} = \arg\min_{a\in\mathbb{R}, b\in\mathbb{R}^n} \|x_t - (ua+b)\|_2 \quad \text{s.t.} \quad \|b\|_0 \leq d_h$$

From this, we obtain an estimate of the sparse perturbation vector and consequently, the scaling factor associated with u. By subtracting this out, we obtain vector which is close to our desired subspace. Using a block of B such vectors, the t-loop accumulates the sample covariance matrix. Finally, the h-loop performs a noisy power method update on the accumulated covariance matrix until our estimate reach the desired numerical accuracy ϵ with respect to the true eigenvector. Note that this is effectively a block version of the usual power method but the key challenge is to control the perturbation in the sample covariance estimate due to (1) the error induced by thresholding (ie, running only a finite number of alternations), and (2) the error in our estimate of the top eigenvector in the current epoch. Additionally, we note that samples are never revisited and hence this is a one-pass algorithm. As described in Section 3, note that $s_{\max}$ is a constant which may be assumed to be known, without loss of generality.

Remark 2.1. *Note that we don't know α_{h-1} in practice and hence don't know the exact bound on Z_{h-1} but we will see that from Theorem 3.4 that a simple rule is to set $Z_{h-1} = C_1\sqrt{n}C_2^{-(h-1)/2}$ where $C_1, C_2 > 0$ are constants.*

Algorithm 1 Block Stochastic Power Method with Hard Thresholding

Input: Samples $\{x_1, \ldots, x_T\} \in \mathbb{R}^n$ such that $x_t = uz_t + s_t$
Output: Leading eigenvector of the denoised samples u_H
$u'_0 \sim \mathcal{N}(0, I_{n\times n})$
$u_0 \leftarrow \frac{u'_0}{\|u'_0\|_2}$
for $h = 1, \ldots, H = \frac{T}{B}$ **do**
  $u'_h \leftarrow 0$
  for $t = B(h-1)+1, \ldots, Bh$ **do**
    $\hat{s}_t^0 \leftarrow 0$
    for $\tau = 1, \ldots, \mathcal{T}$ **do**
      $\zeta_t^\tau \leftarrow 2Z_{h-1} + \frac{1}{5}\left(\frac{1}{10}\right)^\tau \frac{s_{\max}}{\sqrt{n}}$
      $\hat{z}_t^\tau \leftarrow u_{h-1}^\top (x_t - \hat{s}_t^{\tau-1})$
      $\hat{s}_t^\tau \leftarrow \mathrm{Thresh}_{\zeta_t^\tau}(x_t - u_{h-1}\hat{z}_t^\tau)$
    end for
    $\hat{s}_t \leftarrow \hat{s}_t^{\mathcal{T}}$
    $u'_h \leftarrow u'_h + \frac{1}{B}(x_t - \hat{s}_t)(x_t - \hat{s}_t)^\top u_{h-1}$
  end for
  $u_h \leftarrow \frac{u'_h}{\|u'_h\|_2}$
end for
return $\hat{z}_1, \ldots, \hat{z}_T, \hat{s}_1, \ldots, \hat{s}_T, u_{T/B}$

3 Analysis

For simplicity and concreteness, we now present the main result for the rank-1 case and present the proof details in appendix.

Theorem 3.1. *Under the assumptions in Section 2.2, if $B \geq \frac{32Cn(\log H)^2}{\epsilon^2}$, $H \geq C_5 \log\left(\frac{n}{\epsilon}\right)$, $\mathcal{T} > \log_{10}\left(\frac{C_1\sqrt{n}\log(H)s_{\max}}{\epsilon\sqrt{B}}\right)$, with probability at least $1 - 6C$, Algorithm 1 yields an ϵ-close solution in the sense that $\alpha_H \leq \epsilon$.*

Proof outline: At a high level, the proof of convergence involves analyzing the three loops in Algorithm 1, namely: (1) convergence of (innermost) τ-loop (alternations), (2) concentration properties in (middle) t-loop (iterations), and (3) convergence of (outermost) h-loop (epochs). We wish re-emphasize that the concentration arguments are different from Mitliagkas et al. [2013, 2014] since we do not have any randomness assumptions on the support of the sparse perturbation. Lemma 3.1 quantifies the property of our initialization that is proved in Lemma 6 of Mitliagkas et al. [2013] but we provide it here for completeness.

Lemma 3.1. *The initialization given by Steps 3 and 4 of Algorithm 1 yields a vector u_0 such that $\sqrt{1-\alpha_0} = O(1/\sqrt{n})$ with probability $1 - o(1)$.*

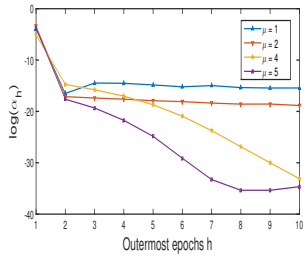
3.1 Convergence of the Innermost Loop

The main result of this section is the validity of the hard-thresholding operation, stated as:

Theorem 3.2. *For every t, after $\mathcal{T} > \log_{10}\left(\frac{s_{\max}}{\epsilon\sqrt{n}}\right)$ alternations we have $\|e_t\|_\infty \leq 4Z_{h-1} + \epsilon$.*

3.2 Concentration Properties in the Middle Loop

We analyze the concentration properties of many iterations within a single epoch, ie, with enough number of samples, the covariance within a block concentrates. In this step, it is essential to show that the covariance in a single epoch converges to the true covariance plus a perturbation that depends on the sparse perturbation and is decaying as epochs proceed, so that this estimate may then be used for block power method updates. We note that the index t in this section runs from $B(h-1)+1$ to Bh and to simplify notation, we will omit this range in the summations. Thus, the main result of this section is:



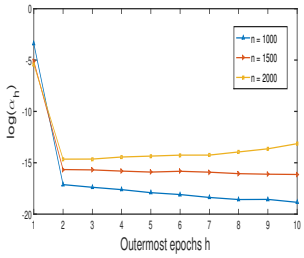


Figure 1: Log of α_h with $T = 1000$, $B = 100$. (Left) $n = 1000$, varying μ. (Right) $\mu = 2$, varying n.

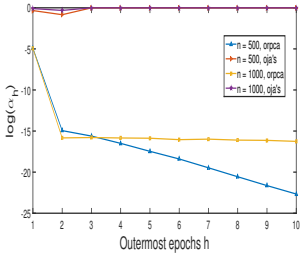


Figure 2: Log of α_h with $T = 1000$, $B = 100$, $\mu = 2$ using Algorithm 1 and Oja's algorithm.

Theorem 3.3. *Setting* $\mathcal{T} > \log_{10}\left(\frac{C_1\sqrt{n}\log(H)s_{\max}}{\epsilon\sqrt{B}}\right)$ *for every* t, *letting* $B \geq \frac{32Cn(\log H)^2}{\epsilon^2}$, *with probability* $1 - \frac{6C}{H}$, *we have* $\|\Delta_h\|_2 \leq \epsilon + 100d_h Z_{\max}^2 \alpha_{h-1}$.

3.3 Convergence of the Outermost Loop

The goal here is to show that $\alpha_h \to 0$ by quantifying the improvement (decrease) of α_h over α_{h-1}. The main result for this section is:

Theorem 3.4. *If* $H \geq C_5 \log\left(\frac{n}{\epsilon}\right)$ *with probability at least* $1 - 6C$, *we obtain* $\alpha_H \leq \epsilon$ *where* C_5 *and* C *are constants.*

4 Experiment

In this section, we show some synthetic results using Algorithm 1. We generate the samples in a batch, but only feed them in the algorithm one by one. $x_t = uz_t + s_t$. Entities in $u \in \mathbb{R}^n$ are i.i.d. samples generated from $\mathcal{N}(0, 1)$. z_t is also generated from $\mathcal{N}(0, 1)$. s_t is generated at the beginning of each outermost loop and then added to the low rank part before we use the noisy sample in middle and inner loops.

In Figure 1, we show the convergence of α_h varying incoherence parameter μ and sample dimension n. We also compare our algorithm with Oja's algorithm Oja and Karhunen [1985]. Since Oja's algorithm considered pure online PCA problem, we expect it to fail in our task. From Figure 2, we can see that the log distance using Oja's algorithm is close to 0, which means α_h is close to 1. In other word, the estimated u_h is far away from u.

5 Conclusion and Future Work

In this paper, we have presented the first convergence result for the robust PCA problem in the streaming setting under the most general assumptions compared to previous works mentioned in Section 1.2. Extending the results in this paper to the rank-r case should be possible along similar lines but we note two points: (1) a naïve analysis would lead to loose bounds and hence care must be taken in accounting for the r-eigenvectors while using the distance between subspaces to track the progress of the algorithm, and (2) we suspect that the convergence guarantee for the rank-r analogue of Algorithm 1 (ie, via block stochastic orthogonal iteration with hard thresholding) will have a sub-optimal dependence on the condition number and hence one plausible fix would be consider the streaming version of the stage-wise algorithm in Netrapalli et al. [2014]. We defer these analyses of the rank-r case to future work. Though the main focus of this paper was to obtain convergence guarantees, we wish to note that potential applications include real-time background-foreground separation in videos and real-time subspace tracking similar to He et al. [2011].

References

A. Balsubramani, S. Dasgupta, and Y. Freund. The fast convergence of incremental pca. In *Advances in Neural Information Processing Systems*, pages 3174–3182, 2013.

C. De Sa, K. Olukotun, and C. Ré. Global convergence of stochastic gradient descent for some non-convex matrix problems. *arXiv preprint arXiv:1411.1134*, 2014.

J. He, L. Balzano, and J. Lui. Online robust subspace tracking from partial information. *arXiv preprint arXiv:1109.3827*, 2011.

P. Jain, C. Jin, S. M. Kakade, P. Netrapalli, and A. Sidford. Streaming pca: Matching matrix bernstein and near-optimal finite sample guarantees for oja's algorithm. In *29th Annual Conference on Learning Theory*, pages 1147–1164, 2016.

I. Mitliagkas, C. Caramanis, and P. Jain. Memory limited, streaming pca. In *Advances in Neural Information Processing Systems*, pages 2886–2894, 2013.

I. Mitliagkas, C. Caramanis, and P. Jain. Streaming pca with many missing entries. *Preprint*, 2014.

P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain. Non-convex robust pca. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

E. Oja and J. Karhunen. 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.

O. Shamir. A stochastic pca and svd algorithm with an exponential convergence rate. In *Proc. of the 32st Int. Conf. Machine Learning (ICML 2015)*, pages 144–152, 2015.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

X. Yi, D. Park, Y. Chen, and C. Caramanis. Fast algorithms for robust pca via gradient descent. *arXiv preprint arXiv:1605.07784*, 2016.

6 Appendix

6.1 Proof of Theorem 3.2

Proof. This follows from Lemmas 6.1 and 6.2. □

Lemma 6.1. *We have the following useful short results.*

1. $\left\|\left(I - u_{h-1}u_{h-1}^\top\right)u\right\|_\infty \le \alpha_{h-1}\frac{\mu}{\sqrt{n}} + \sqrt{\alpha_{h-1}\left(1-\alpha_{h-1}\right)}$.
2. $Z_{h-1} \le 2Z_{\max}\sqrt{\alpha_{h-1}}$.
3. *When* $\alpha_{h-1} > \frac{1}{n}$, $\|u_{h-1}\|_\infty^2 \le \frac{\mu^2+3\mu n\alpha_{h-1}}{n}$. *Else,* $\|u_{h-1}\|_\infty^2 \le \frac{4\mu^2}{n}$.

Proof. 1. $\left|u^\top u_{h-1}\right| = \sqrt{1-\alpha_{h-1}}$ and $\operatorname{sign}(u^\top u_{h-1}).\operatorname{sign}(b_i^\top u_{h-1}) = \operatorname{sign}(b_i^\top u)$. Thus,

$$\begin{aligned}\left\|u - (u_{h-1}^\top u)u_{h-1}\right\|_\infty &= \left\|u - \sqrt{1-\alpha_{h-1}}\left(\sqrt{1-\alpha_{h-1}}u + \sqrt{\alpha_{h-1}}v_{h-1}\right)\right\|_\infty \\ &\le \alpha_{h-1}\|u\|_\infty + \sqrt{\alpha_{h-1}\left(1-\alpha_{h-1}\right)}\|v_{h-1}\|_\infty \\ &\le \alpha_{h-1}\frac{\mu}{\sqrt{n}} + \sqrt{\alpha_{h-1}\left(1-\alpha_{h-1}\right)}\end{aligned}$$

2. We can obtain an upper bound on Z_{h-1} as follows:

$$\begin{aligned}Z_{h-1} &= Z_{\max}\left(\alpha_{h-1}\frac{\mu}{\sqrt{n}} + \sqrt{\alpha_{h-1}\left(1-\alpha_{h-1}\right)}\right) \\ &\le Z_{\max}\sqrt{\alpha_{h-1}}\left(\sqrt{\alpha_{h-1}}\frac{\mu}{\sqrt{n}} + \sqrt{1-\alpha_{h-1}}\right)\end{aligned}$$

   By noting that $\alpha_{h-1} \in (0,1)$ and $\mu \le \sqrt{n}$, we obtain $Z_{h-1} \le 2Z_{\max}\sqrt{\alpha_{h-1}}$.
3. Since $\alpha_{h-1} \in (0,1)$, we have $1-\alpha_{h-1} < 1$. Using this in ξ_1,

$$\begin{aligned}\|u_{h-1}\|_\infty^2 &= \left\|\left(\sqrt{1-\alpha_{h-1}}u + \sqrt{\alpha_{h-1}}v_{h-1}\right)\right\|_\infty^2 \\ &\le \left(1-\alpha_{h-1}\right)\|u\|_\infty^2 + \alpha_{h-1}\|v_{h-1}\|_\infty^2 \\ &\qquad + 2\sqrt{\alpha_{h-1}\left(1-\alpha_{h-1}\right)}\|u\|_\infty\|v_{h-1}\|_\infty \\ &\le \left(1-\alpha_{h-1}\right)\frac{\mu^2}{n} + \alpha_{h-1} + 2\sqrt{\alpha_{h-1}\left(1-\alpha_{h-1}\right)}\frac{\mu}{\sqrt{n}} \\ &\overset{\xi_1}{\le} \frac{\mu^2}{n} + \alpha_{h-1} + 2\sqrt{\alpha_{h-1}}\frac{\mu}{\sqrt{n}} = \frac{\mu^2 + n\alpha_{h-1} + 2\mu\sqrt{n\alpha_{h-1}}}{n}\end{aligned}$$

   When $\alpha_{h-1} > \frac{1}{n}$, we have $n\alpha_{h-1} > \sqrt{n\alpha_{h-1}}$. So, $\|u_{h-1}\|_\infty^2 \le \frac{\mu^2+3\mu n\alpha_{h-1}}{n}$ by noting $\mu \ge 1$.

□

Lemma 6.2. *If* $\left\|e_t^{\tau-1}\right\|_\infty \le 4Z_{h-1} + \left(\frac{1}{10}\right)^{\tau-1}\frac{s_{\max}}{\sqrt{n}}$, *then we have:*

1. $\left|b_i^\top(uz_t - u_{h-1}\hat{z}_t^\tau)\right| \le \frac{26}{25}Z_{h-1} + \left(\frac{1}{10}\right)^{\tau+1}\frac{s_{\max}}{\sqrt{n}}$
2. $\left\|e_t^\tau\right\|_\infty \le 4Z_{h-1} + \left(\frac{1}{10}\right)^{\tau}\frac{s_{\max}}{\sqrt{n}}$.
3. *Moreover,* $\operatorname{Supp}\left(e_t^\tau\right) \subseteq \operatorname{Supp}\left(e_t^{\tau-1}\right) \subseteq \operatorname{Supp}\left(s_t\right)$.

Proof. 1. $x_t - u_{h-1}\hat{z}_t^\tau = (uz_t + s_t) - u_{h-1}\hat{z}_t^\tau$
$\implies \left|b_i^\top (x_t - u_{h-1}\hat{z}_t^\tau - s_t)\right| = \left|b_i^\top (uz_t - u_{h-1}\hat{z}_t^\tau)\right|$.

$$\begin{aligned}
\left|b_i^\top (uz_t - u_{h-1}\hat{z}_t^\tau)\right| &\overset{\xi_1}{=} \left|b_i^\top (uz_t - u_{h-1}u_{h-1}^\top (x_t - \hat{s}_t^{\tau-1}))\right| \\
&\overset{\xi_2}{=} \left|b_i^\top (uz_t - u_{h-1}u_{h-1}^\top (uz_t + e_t^{\tau-1}))\right| \\
&\overset{\xi_3}{\leq} \left|b_i^\top (uz_t - u_{h-1}u_{h-1}^\top uz_t)\right| + \left|b_i^\top u_{h-1}u_{h-1}^\top e_t^{\tau-1}\right| \\
&\leq Z_{\max} \max_i \left|b_i^\top (u - (u_{h-1}^\top u)u_{h-1})\right| + \max_i \left|b_i^\top u_{h-1}u_{h-1}^\top e_t^{\tau-1}\right| \\
&\overset{\xi_4}{\leq} Z_{\max} \left\|u - (u_{h-1}^\top u)u_{h-1}\right\|_\infty + \|u_{h-1}\|_\infty^2 \left\|e_t^{\tau-1}\right\|_1 \\
&\overset{\xi_5}{\leq} Z_{\max} \left(\alpha_{h-1}\frac{\mu}{\sqrt{n}} + \sqrt{\alpha_{h-1}(1-\alpha_{h-1})}\right) + d_h \|u_{h-1}\|_\infty^2 \left\|e_t^{\tau-1}\right\|_\infty \\
&\overset{\xi_6}{\leq} Z_{h-1} + \frac{1}{100}\left\|e_t^{\tau-1}\right\|_\infty \overset{\xi_7}{\leq} Z_{h-1} + \frac{1}{100}\left(4Z_{h-1} + \left(\frac{1}{10}\right)^{\tau-1}\frac{s_{\max}}{\sqrt{n}}\right) \\
&\leq \frac{26}{25}Z_{h-1} + \left(\frac{1}{10}\right)^{\tau+1}\frac{s_{\max}}{\sqrt{n}}
\end{aligned}$$

where ξ_1 is by substituting $\hat{z}_t^\tau = u_{h-1}^\top (x_t - \hat{s}_t^{\tau-1})$, ξ_2 by recalling the definition that $e_t^{\tau-1} = s_t - \hat{s}_t^{\tau-1}$, ξ_3 by triangle inequality, ξ_4 by using $|\langle a, b\rangle| \leq \|a\|_\infty \|b\|_1$, ξ_5 by Lemma 6.1-(1) and noting that $\left\|e_t^{\tau-1}\right\|_1 \leq d_h \left\|e_t^{\tau-1}\right\|_\infty$, ξ_6 by using the definition of Z_{h-1} and the assumption on d_h, ξ_7 by inductive hypothesis that $\left\|e_t^{\tau-1}\right\|_\infty \leq 4Z_{h-1} + \left(\frac{1}{10}\right)^{\tau-1}\frac{s_{\max}}{\sqrt{n}}$.

2. Next, to complete the induction over τ, let us calculate $\|e_t^\tau\|_\infty$. We have two cases
   (a) *Case 1* $\left(\left|b_i^\top (x_t - u_{h-1}\hat{z}_t^\tau)\right| > \zeta_t^\tau\right)$: $\left|b_i^\top e_t^\tau\right| = \left|b_i^\top (s_t - \hat{s}_t^\tau)\right| = \left|b_i^\top (s_t - (x_t - u_{h-1}\hat{z}_t^\tau))\right| = \left|b_i^\top (uz_t - u_{h-1}\hat{z}_t^\tau)\right| \leq \frac{26}{25}Z_{h-1} + \left(\frac{1}{10}\right)^{\tau+1}\frac{s_{\max}}{\sqrt{n}}$.
   (b) *Case 2* $\left(\left|b_i^\top (x_t - u_{h-1}\hat{z}_t^\tau)\right| \leq \zeta_t^\tau\right)$: $b_i^\top \hat{s}_t^\tau = 0 \implies b_i^\top e_t^\tau = b_i^\top s_t$ and $\left|b_i^\top (x_t - u_{h-1}\hat{z}_t^\tau)\right| = \left|b_i^\top (uz_t + s_t - u_{h-1}\hat{z}_t^\tau)\right| \leq \zeta_t^\tau$. So, we have

$$\begin{aligned}
\left|b_i^\top e_t^\tau\right| = \left|b_i^\top s_t\right| &\leq \zeta_t^\tau + \left|b_i^\top (uz_t - u_{h-1}\hat{z}_t^\tau)\right| \\
&\leq \left(2Z_{h-1} + \frac{1}{5}\left(\frac{1}{10}\right)^\tau \frac{s_{\max}}{\sqrt{n}}\right) + \left(\frac{26}{25}Z_{h-1} + \left(\frac{1}{10}\right)^{\tau+1}\frac{s_{\max}}{\sqrt{n}}\right) \\
&= \frac{76}{25}Z_{h-1} + \left(\frac{1}{5} + \frac{1}{10}\right)\left(\frac{1}{10}\right)^\tau \frac{s_{\max}}{\sqrt{n}} \leq 4Z_{h-1} + \left(\frac{1}{10}\right)^\tau \frac{s_{\max}}{\sqrt{n}}
\end{aligned}$$

3. If $b_i^\top s_t = 0$, then we have $b_i^\top e_t^\tau = \mathbf{I}_{\{|b_i^\top (uz_t - u_{h-1}\hat{z}_t^\tau)|\} > \zeta_t^\tau}\left(b_i^\top (uz_t - u_{h-1}\hat{z}_t^\tau)\right)$ but note that

$$\left|b_i^\top (uz_t - u_{h-1}\hat{z}_t^\tau)\right| \leq \frac{26}{25}Z_{h-1} + \left(\frac{1}{10}\right)^{\tau+1}\frac{s_{\max}}{\sqrt{n}} < 2Z_{h-1} + \frac{1}{5}\left(\frac{1}{10}\right)^\tau \frac{s_{\max}}{\sqrt{n}} = \zeta_t^\tau$$

This is a contradiction since the indicator is inactive at location i, so $b_i^\top e_t^\tau = 0$.

□

6.2 Proof of Theorem 3.3

Proof. We have $\Delta_h = \Sigma_h - \Sigma$.

$$\begin{aligned}
\|\Sigma_h - \Sigma\|_2 &= \left\|\sum_{t=1}^B \frac{1}{B}(x_t - \hat{s}_t)(x_t - \hat{s}_t)^\top - uu^\top\right\|_2 = \left\|\frac{1}{B}\sum_t (uz_t + e_t)(uz_t + e_t)^\top - uu^\top\right\|_2 \\
&\leq \underbrace{\frac{1}{B}\left\|uu^\top \sum_t (z_t^2 - 1)\right\|_2}_{Term-1} + \underbrace{\frac{1}{B}\left\|u\sum_t z_t e_t^\top\right\|_2}_{Term-2} + \underbrace{\frac{1}{B}\left\|\sum_t z_t e_t u^\top\right\|_2}_{Term-3} + \underbrace{\frac{1}{B}\left\|\sum_t e_t e_t^\top\right\|_2}_{Term-4}
\end{aligned}$$

The second step is by triangle inequality on the spectral norm of the perturbed matrix. Now, we bound each of the terms using similar techniques as Mitliagkas et al. ((2013)).

Term-1: Using tail bounds for sub-Gaussian random variables from Vershynin ((2010)), with probability at least $1 - \frac{2C}{H}$

$$\frac{1}{B}\left\|uu^\top \sum_t \left(z_t^2 - 1\right)\right\|_2 \leq \left|\frac{1}{B}\sum_t \left(z_t^2 - 1\right)\right| \left\|uu^\top\right\|_2 \leq \sqrt{\frac{C\log H}{B}}$$

Term-2: As spectral norm is sub-multiplicative, $\left\|\sum_t uz_t e_t^\top\right\|_2 = \left\|u(E_h\mathbf{z})^\top\right\|_2 \leq \|u\|_2 \|E_h\mathbf{z}\|_2$. Now, for every i, since $z_t \overset{iid}{\sim} \mathcal{N}(0,1)$, we have $b_i^\top E_h\mathbf{z} \sim \mathcal{N}(0,\sigma_e^2)$ where $\sigma_e^2 \leq B\|E_h\|_\infty^2$; this is because $var\left(b_i^\top E_h\mathbf{z}\right) = var\left(\sum_{j=1}^B b_i^\top E_h b_j b_j^\top \mathbf{z}\right) \leq B\|E_h\|_\infty^2$. Hence, with probability $1 - \frac{2C}{H}$

$$\|E_h\mathbf{z}\|_2 = \sqrt{\sum_{i=1}^n \left(b_i^\top E_h\mathbf{z}\right)^2} \leq \sqrt{n\left(b_1^\top E_h\mathbf{z}\right)^2} \leq \sqrt{n}\sqrt{2}\sigma_e \log\left(\frac{H}{C}\right)$$

where the last line was obtained by using the Hoeffding bound, ie, tail bound for $X \sim \mathcal{N}\left(0,\sigma_e^2\right)$ is given by $Pr\left(-t \leq X \leq t\right) \leq 1 - 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$ and noting that $\sqrt{X^2}$ is half-normal distribution satisfying this bound. Further simplifying by substituting $\|E_h\|_\infty$, using Theorem 3.2 and Lemma 6.1-(2), we get

$$\|E_h\mathbf{z}\|_2 \leq \sqrt{2nB}\|E_h\|_\infty \log\left(\frac{H}{C}\right) \leq \sqrt{2nB}\log\left(\frac{H}{C}\right)\left(8Z_{\max}\sqrt{\alpha_{h-1}} + \epsilon\right)$$

Dividing both sides by B, we obtain $\frac{1}{B}\|E_h\mathbf{z}\|_2 \leq \sqrt{\frac{2n}{B}}\log\left(\frac{H}{C}\right)\left(8Z_{\max}\sqrt{\alpha_{h-1}} + \epsilon\right)$.

Term-3: Same as Term-2.

Term-4: Let $\epsilon \leq Z_{h-1} \leq 2Z_{\max}\sqrt{\alpha_{h-1}}$. Using triangle inequality, sub-multiplicative property, Theorem 3.2 and Lemma 6.1-(4), we have

$$\begin{aligned}\frac{1}{B}\left\|\sum_t e_t e_t^\top\right\|_2 &\leq \frac{1}{B}\sum_t \left\|e_t e_t^\top\right\|_2 \leq \frac{1}{B}\sum_t \|e_t\|_2^2 \leq \frac{1}{B}.B\left(\sqrt{d_h}\|E_h\|_\infty\right)^2 \leq d_h\|E_h\|_\infty^2 \\ &\leq d_h\left(4Z_{h-1} + \epsilon\right)^2 \leq d_h\left(25Z_{h-1}^2\right) \leq 100d_h Z_{\max}^2\alpha_{h-1}\end{aligned}$$

Note that by setting $\mathcal{T} > \log_{10}\left(\frac{C_1\sqrt{n}\log(H)s_{\max}}{\epsilon\sqrt{B}}\right)$ we have $\|e_t\|_\infty \leq 4Z_{h-1} + \frac{\epsilon\sqrt{B}}{C_1 n\log H}$ where C_1 is a constant. Using this, combining all the terms, letting $B \geq \frac{32Cn(\log H)^2}{\epsilon^2}$ and assuming $\alpha_{h-1} \leq \frac{1}{256Z_{\max}^2}$, we obtain, with probability $1 - \frac{6C}{H}$:

$$\begin{aligned}\|\Delta_h\|_2 &= \|\Sigma_h - \Sigma\|_2 \\ &\leq \sqrt{\frac{C\log H}{B}} + 2\sqrt{\frac{2n}{B}}\log\left(\frac{H}{C}\right)\left(8Z_{\max}\sqrt{\alpha_{h-1}} + \frac{\epsilon\sqrt{B}}{C_1 n\log H}\right) + 100d_h Z_{\max}^2\alpha_{h-1} \\ &\leq \epsilon + 100d_h Z_{\max}^2\alpha_{h-1}\end{aligned}$$

□

6.3 Proof of Theorem 3.4

Proof. Noting $u_h = \frac{u_h'}{\|u_h'\|_2}$, decomposing u_h' as $u_h' = \langle u_h', u\rangle u + \langle u_h', v_h\rangle v_h$ and similarly for u_h, using Lemma 6.3 and assuming $10\epsilon < \sqrt{\alpha_{h-1}}$, we have,

$$\begin{aligned}\alpha_h = \langle u_h, v_h\rangle^2 &= \left\langle \frac{u_h'}{\|u_h'\|_2}, v_h\right\rangle^2 = \frac{\langle u_h', v_h\rangle^2}{\|u_h'\|_2^2} = \frac{\langle u_h', v_h\rangle^2}{\langle u_h', u\rangle^2 + \langle u_h', v_h\rangle^2} \\ &\leq \frac{\left(\epsilon + 100d_h Z_{\max}^2\alpha_{h-1}\right)^2}{\left(\sqrt{1-\alpha_{h-1}}\left(1 - \epsilon - \sqrt{\frac{\alpha_{h-1}}{1-\alpha_{h-1}}}100d_h Z_{\max}^2\alpha_{h-1}\right)\right)^2 + \left(\epsilon + 100d_h Z_{\max}^2\alpha_{h-1}\right)^2} \\ &\leq \frac{\alpha_{h-1}\left(0.1 + 100d_h Z_{\max}^2\sqrt{\alpha_{h-1}}\right)^2}{(1-\alpha_{h-1})\left(1 - \left(0.1 + \sqrt{\frac{\alpha_{h-1}}{1-\alpha_{h-1}}}100d_h Z_{\max}^2\sqrt{\alpha_{h-1}}\right)\right)^2 + \alpha_{h-1}\left(0.1 + 100d_h Z_{\max}^2\sqrt{\alpha_{h-1}}\right)^2}\end{aligned}$$

The second inequality above is obtained by noting that $\frac{x}{c+x}$ is an increasing function in x for positive x and c. Let $C_3 = \left(\left(0.1 + 100 d_h Z_{\max}^2 \sqrt{\alpha_{h-1}}\right) / \left(0.9 - \sqrt{\frac{\alpha_{h-1}}{1-\alpha_{h-1}}} 100 d_h Z_{\max}^2 \sqrt{\alpha_{h-1}}\right)\right)^2$. Note that if $d_h < 1/250 Z_{\max}^2 \sqrt{\alpha_{h-1}}$, we note that the constant $C_3 < 1$. Using this and also applying Lemmas 2, 6 of Mitliagkas et al. ((2013)), we get $\alpha_h \overset{\xi_1}{=} \frac{C_3 \alpha_{h-1}}{1-\alpha_{h-1}+C_3\alpha_{h-1}} \overset{\xi_2}{\leq} \frac{C_3^h \alpha_0}{1-\left(1-C_3^h\right)\alpha_0} \leq C_4 C_3^h n$ where ξ_1 holds with probability at least $1 - \frac{6C}{H}$ and where ξ_2 holds with probability at least $1 - \frac{6hC}{H}$ where the factor of h comes by accounting for the failure of atleast one epoch followed by applying the union bound. Hence, if $H \geq \log_{1/C_3}\left(\frac{C_4 n}{\epsilon}\right)$ with probability atleast $1 - 6C$, we obtain $\alpha_H \leq \epsilon$. □

Lemma 6.3. *We have the following upper and lower bounds.*

1. $\langle u_h', v_h \rangle \leq \epsilon + 100 d_h Z_{\max}^2 \alpha_{h-1}$.
2. $\langle u_h', u \rangle \geq \sqrt{1-\alpha_{h-1}} \left(1 - \epsilon - \sqrt{\frac{\alpha_{h-1}}{1-\alpha_{h-1}}} 100 d_h Z_{\max}^2 \alpha_{h-1}\right)$.

Proof. 1. Recall that $\langle u, v_h \rangle = 0$. Now,

$$\begin{aligned}\langle u_h', v_h \rangle &= v_h^\top \Sigma_h u_{h-1} = v_h^\top \left(\Sigma + \Delta_h\right) u_{h-1} = v_h^\top u u^\top u_{h-1} + v_h^\top \Delta_h u_{h-1} \\ &\leq 0 + \|v_h\|_2 \|\Delta_h\|_2 \|u_{h-1}\|_2 = \epsilon + 100 d_h Z_{\max}^2 \alpha_{h-1}\end{aligned}$$

2. Next, we have lower bound the following term since it would appear in the denominator.

$$\begin{aligned}\langle u_h', u \rangle &= u^\top \Sigma_h u_{h-1} = u^\top \frac{1}{B} \sum_t \left(x_t - \hat{s}_t\right)\left(x_t - \hat{s}_t\right)^\top u_{h-1} \\ &= \frac{1}{B} \sum_t \left(u^\top \left(u z_t + e_t\right)\right) \left(\left(u z_t + e_t\right)^\top \left(\sqrt{1-\alpha_{h-1}} u + \sqrt{\alpha_{h-1}} v_{h-1}\right)\right) \\ &= \frac{1}{B} \sum_t \left(z_t + u^\top e_t\right) \left(\sqrt{1-\alpha_{h-1}} z_t + \sqrt{1-\alpha_{h-1}} e_t^\top u + \sqrt{\alpha_{h-1}} e_t^\top v_{h-1}\right) \\ &= \underbrace{\frac{\sqrt{1-\alpha_{h-1}}}{B} \sum_t \left(z_t + u^\top e_t\right)^2}_{Term-5} + \underbrace{\frac{\sqrt{\alpha_{h-1}}}{B} \sum_t \left(z_t + u^\top e_t\right)\left(e_t^\top v_{h-1}\right)}_{Term-6}\end{aligned}$$

Term-5: With probability $1 - \frac{2C}{H}$, using the settings for B and $\mathcal{T}$ from Section 3.2, and the upper bound for Term-2 with a negative sign (since this is an absolute value of scalar),

$$\begin{aligned}\frac{1}{B} \sum_t \left(z_t + u^\top e_t\right)^2 &= \frac{1}{B} \sum_t z_t^2 + \frac{1}{B} \sum_t \left(u^\top e_t\right)^2 + \frac{2}{B} \sum_t z_t u^\top e_t \\ &\geq 1 - \sqrt{\frac{C \log H}{B}} + 0 + \frac{2}{B} u^\top E_h \mathbf{z} \geq 1 - \frac{\epsilon}{4} - \frac{2}{B} \|u\|_2 \|E_h \mathbf{z}\|_2 \geq 1 - \frac{\epsilon}{2}\end{aligned}$$

Term-6: This is similar to spectral norm upper bounds in Step-2 but with a negative sign, ie,

$$\frac{1}{B} \sum_t \left(z_t + u^\top e_t\right)\left(e_t^\top v_{h-1}\right) \leq \frac{1}{B} \left|v_{h-1}^\top E_h \mathbf{z}\right| + \frac{1}{B} \left|u^\top E_h E_h^\top v_{h-1}\right| \leq \frac{\epsilon}{2} + 100 d_h Z_{\max}^2 \alpha_{h-1}$$

Thus, from Terms-5 and 6, and noting $\alpha_{h-1} \leq \frac{1}{4}$, we have,

$$\begin{aligned}\langle u_h', u \rangle &\geq \sqrt{1-\alpha_{h-1}} \left(1 - \frac{\epsilon}{2}\right) - \sqrt{\alpha_{h-1}} \left(\frac{\epsilon}{2} + 100 d_h Z_{\max}^2 \alpha_{h-1}\right) \\ &= \sqrt{1-\alpha_{h-1}} \left(1 - \left(1 + \sqrt{\frac{\alpha_{h-1}}{1-\alpha_{h-1}}}\right) \frac{\epsilon}{2} - \sqrt{\frac{\alpha_{h-1}}{1-\alpha_{h-1}}} 100 d_h Z_{\max}^2 \alpha_{h-1}\right) \\ &\geq \sqrt{1-\alpha_{h-1}} \left(1 - \epsilon - \sqrt{\frac{\alpha_{h-1}}{1-\alpha_{h-1}}} 100 d_h Z_{\max}^2 \alpha_{h-1}\right)\end{aligned}$$

□