Streaming Anomaly Detection Using Randomized Matrix Sketching

Hao Huang
General Electric Global Research
haohuangcssbu@gmail.com

Shiva Kasiviswanathan
General Electric Global Research
kasivisw@gmail.com

Abstract

Data is continuously being generated from sources such as machines, network traffic, application logs, etc. Timely and accurate detection of anomalies in massive data streams have important applications such as in preventing machine failures, intrusion detection, and dynamic load balancing. In this paper, we introduce a novel anomaly detection algorithm, which can detect anomalies in a streaming fashion by making only one pass over the data while utilizing limited storage. The algorithm uses ideas from matrix sketching and randomized low-rank matrix approximations to maintain, in a streaming model, a set of few orthogonal vectors that form a good approximate basis for the data. Using this constructed orthogonal basis, anomalies in new incoming data are detected based on a simple reconstruction error test. We theoretically prove that our algorithm compares favorably with an offline approach based on global singular value decomposition updates. The experimental results show the effectiveness and efficiency of our approach over other popular fast anomaly detection methods.

1 Introduction

Detecting anomalies in huge volumes of data has many important real-life applications in areas such as machine health monitoring, intrusion detection systems, financial fraud detection, and medical diagnosis [8, 1]. However, it is also a challenging problem because in many modern applications the data arrives in a streaming fashion. The streaming data could be infinite, so offline algorithms that attempt to store the entire stream for analysis will not scale. Also in these situations, there is usually a lack of a complete (labeled) training set as new anomalous and non-anomalous patterns arise over time (this is sometimes referred to as *concept drift*). A common requirement in many mission-critical applications is to detect anomalies in near real-time, as new data values are encountered.

Although a lot of recent research has been focused on streaming anomaly detection [8, 1], there is still lack of theoretically sound and practically effective algorithms that operate efficiently in a streaming model by making just one pass over the data. In practice, however, because of inherent correlations in the data, it is possible to reduce a large sized numerical stream into just a handful of hidden bases that can compactly describe the key patterns [31], and therefore dramatically reduce the complexity of further analysis. This general idea has been referred to as the *subspace method* in the literature [22, 20, 19, 11]. We exploit this observation in our proposed algorithms by maintaining a set of few orthogonal vectors that conceptually constitute up-to-date normal patterns.

In this paper, we introduce a novel approach to anomaly detection in an unsupervised setting based on ideas from *matrix sketching*.[1] We use matrix sketching to maintain (over time) a low-rank matrix with orthogonal columns that can linearly represent well all the identified non-anomalous datapoints. We utilize this for anomaly detection as follows: let U be a low-rank matrix representing all non-anomalous datapoints till time $t-1$, for a new datapoint $\mathbf{y}$ arriving at time t, if there does not exist a good representation of $\mathbf{y}$ using U, then $\mathbf{y}$ does not lie close to the space of non-anomalous datapoints, and $\mathbf{y}$ could be an anomaly. At the end of timestep t, the low-rank matrix is updated to capture all the non-anomalous points introduced at t.

For efficient sketching, we adapt a recent deterministic sketching algorithm called *Frequent Directions* (proposed by Liberty [26]) and combine it with ideas from the theory of randomized low-rank matrix approximations. The

[1] Informally, a sketch of a matrix Z is another matrix Z' that is of smaller size than Z, but still approximates it well.

Frequent Directions algorithm operates in a streaming model and constructs a sketch matrix using a (surprisingly) simple idea of "shrinking" a few orthogonal vectors.

Our Contributions. We propose a novel streaming anomaly detection algorithm for the unsupervised setting. Our theoretical analysis generalizes the analysis of *Frequent Directions* by [26, 12, 13] and combines it with recent results in matrix perturbation theory [9] to prove that our streaming algorithm has a similar performance to that of a global algorithm based on costly singular value decomposition updates. Our proposed algorithm has the following *salient features*:

(1) It can identify anomalies in near-real time, ensuring that the detection keeps up with the rate of data collection.

(2) It is pass-efficient, in that only one pass is required for each datapoint.

(3) It is space-efficient and require only a small amount of bookkeeping space.

(4) It easily adapts to unseen normal patterns and provide timely updated identification of anomalies.

Our experimental results corroborate the performance and scalability of our approach on datasets drawn from diverse domains such as biomedical, network security, and broadcast news.

2 Related Work

Anomaly detection is a well-studied topic and we refer the reader to the excellent surveys by Chandola *et al.* [8] and Aggarwal *et al.* [1] for an introduction. We only mention a few relevant results here.

Many anomaly detection approaches have been suggested based on approximating the *sample density*. This includes distance-based methods [4] and manifold based methods [18]. However, these methods don't work well on large datasets since they require either computing all pair-wise distances or the complete affinity matrix, both of which are time and space consuming. Recently, inlier-based outlier detection methods were proposed in [17]. However, their training and computational complexity requirements render them unsuitable for real-time streaming applications.

Some more efficient techniques such as IForest [27] and Mass [32] are based on attribute-wise analysis. But they tend to fail when the distribution for anomalous points becomes less discriminative, e.g., if the anomalous and non-anomalous points share similar attribute range/distribution [18].

Subspace methods are popular for *distributed anomaly detection* [22, 20, 25, 19, 11, 10]. But the previously proposed methods here, either make decisions using a sliding time window [20] or utilize a random subspace embedding to project each point to a lower dimension space [11]. However, our proposed algorithm at time t makes decisions utilizing a sketch of the entire data history till t, and also it operates in the original feature space thereby avoiding the instability issues that arise from random subspace embedding.

In a streaming setup the training set is usually never perfect, and the detection model needs to be updated as new data comes in. The ideal scenario is to detect the arrival of a new normal pattern, and then improve the model suitably. Some methods achieve this by relying on probabilistic modeling of the data distributions and monitoring the likelihood for new-coming observations; see the survey by Markou and Singh [29]. But they usually rely on accessing the whole of the past historical data at each timestep. Hence, these methods cannot efficiently deal with very large data sets that arise in streaming applications.

Kernel-based online anomaly detection algorithm proposed by Ahmed *et al.* [2] uses a dictionary learnt over normal data to detect anomalies. The dictionary atoms are updated using a simple heuristic rule, but [2] provide no theoretical guarantees on the performance of the algorithm.

3 Preliminaries

Notation. We denote $[n] = 1 : n$. Vectors are always in column-wise fashion and are denoted by boldface letters. For a vector $\mathbf{v}$, $\mathbf{v}^\top$ denotes its transpose and $\|\mathbf{v}\|$ denotes its Euclidean norm. For a matrix $Z \in \mathbf{R}^{m\times n} = \{z_{i,j}\}$, its Frobenius norm $\|Z\|_F^2 = \sum_{ij} z_{ij}^2$, and its spectral norm $\|Z\| = \sup\{\|Z\mathbf{v}\| \,:\, \|\mathbf{v}\| = 1\}$. We use $\mathrm{r}ank(Z)$ to denote the rank of Z. We use $Z \succeq 0$ to denote that if Z is a positive semidefinite matrix, and if $Z - Y \succeq 0$, then we write $Z \succeq Y$. For a vector $(z_1, \dots, z_m) \in \mathbf{R}^m$, $\mathrm{d}iag(z_1, \dots, z_m) \in \mathbf{R}^{m\times m}$ denotes a diagonal matrix with $z_1, \dots, z_m$ as its diagonal entries. Given a matrix Z, we abuse notation and use $\mathbf{y} \in Z$ to represent that $\mathbf{y}$ is a column in Z. Let $\mathbf{I}_m$ denote an identity matrix of dimension $m \times m$. Given a set of matrices, $Z_1, \dots, Z_t$, we use the notation $Z_{[t]}$ to denote the matrix obtained by horizontally concatenating $Z_1, \dots, Z_t$, i.e., $Z_{[t]} = [Z_1|\dots|Z_t]$.

We use $\textsc{Svd}(Z)$ to denote the singular value decomposition of Z, i.e., $\textsc{Svd}(Z) = U\Sigma V^\top$. Here U is an $m \times m$ orthogonal matrix, Σ is an $m \times n$ diagonal matrix, and V is an $n \times n$ orthogonal matrix. The diagonal entries of Σ are known as the singular values of Z. We follow the common convention to list the singular values in non-increasing order. For a symmetric matrix $S \in \mathbf{R}^{m\times m}$, we use $\textsc{Eig}(S)$ to denote its eigenvalue decomposition, i.e., $U\Lambda U^\top = \textsc{Eig}(S)$. Here U is an $m \times m$ orthogonal matrix and Λ is an $m \times m$ diagonal matrix whose (real) entries are $\lambda_1, \dots, \lambda_m$ are known as the eigenvalues of S (again listed in non-increasing order).

The best rank-k approximation (in both the spectral and Frobenius norm) to a matrix $Z \in \mathbf{R}^{m\times n}$ is $Z_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$ are the top-k singular values of Z, with associated left and right singular vectors $\mathbf{u}_i \in \mathbf{R}^m$ and $\mathbf{v}_i \in \mathbf{R}^n$, respectively. We use $\textsc{Svd}_k(Z)$ to denote the the singular value decomposition of Z_k, i.e., $Z_k = \textsc{Svd}_k(Z) = U_k \Sigma_k V_k^\top$. Here $\Sigma_k = \mathrm{d}iag(\sigma_1, \dots, \sigma_k) \in \mathbf{R}^{k\times k}$, $U_k = [\mathbf{u}_1, \dots, \mathbf{u}_k] \in \mathbf{R}^{m\times k}$, and $V_k = [\mathbf{v}_1, \dots, \mathbf{v}_k] \in \mathbf{R}^{n\times k}$. The following celebrated theorem bounds the approximation error.

Theorem 1. *[14] Let $Z \in \mathbf{R}^{m\times n}$ with $n > m$, and let $\sigma_1 \geq \cdots \geq \sigma_m$ be the singular values of Z. Let $U_k \Sigma_k V_k^\top = \textsc{Svd}_k(Z)$. Then*

$$\min_{\mathrm{r}ank(X)\leq k} \|Z - X\|_2 = \|Z - U_k\Sigma_k V_k^\top\|_2 = \sigma_{k+1}\,,$$

$$\min_{\mathrm{r}ank(X)\leq k} \|Z - X\|_F = \|Z - U_k\Sigma_k V_k^\top\|_F = \sqrt{\sum_{j=k+1}^{m} \sigma_{k+1}^2}.$$

Definition 1. *Define the k-condition number of a matrix $Z \in \mathbf{R}^{m\times n}$ with $n > m$ as $\kappa_k(Z) = \sigma_1/\sigma_k \geq 1$ where $\sigma_1 \geq \cdots \geq \sigma_m$ are the singular values of Z.*

The following claim is quite standard.

Claim 2. *Let $Z \in \mathbf{R}^{m\times n}$, and let Z_k be a rank-k approximation of Z according to Theorem 1. For any vector $\mathbf{x} \in \mathbf{R}^m$, $\kappa_k(Z)\|Z_k^\top \mathbf{x}\| \geq \|Z^\top \mathbf{x}\|$.*

Proof. Follows from the fact that

$$\|Z^\top \mathbf{x}\| \leq \sigma_1(Z)\|\mathbf{x}\| \leq \sigma_1(Z)\frac{\|Z_k^\top \mathbf{x}\|}{\sigma_k(Z)} = \kappa_k(Z)\|Z_k^\top \mathbf{x}\|,$$

where the last inequality comes from: $\|Z_k^\top \mathbf{x}\| \geq \sigma_k(Z_k)\|\mathbf{x}\| = \sigma_k(Z)\|\mathbf{x}\|$. □

4 Streaming Anomaly Detection

In this section, we propose an anomaly detection scheme for streaming data based on matrix sketching, and also provide theoretical guarantees for its efficacy. We start by describing the problem of streaming anomaly detection.

Streaming Anomaly Detection Task. We assume that the data arrives in streams. Let $\{Y_t \in \mathbf{R}^{m \times n_t},\ t = 1, 2, \dots\}$ denote a sequence of streaming data matrices, where Y_t represents the datapoints introduced at timestep t. Here m is the size of the feature space, and n_t is the number of datapoints arriving at time t. We typically assume that there are more datapoints than number of features ($n_t > m$).[2] We normalize Y_t such that each column (point) in Y_t has a unit L_2-norm. Under this setup, the goal of streaming anomaly detection is to **identify "anomalous datapoints" in Y_t at every timestep t.**

Our Anomaly Detection Framework. Our idea is based on maintaining, at every timestep t, a *low-rank matrix with orthogonal columns* that can reconstruct "well" all the prior (till time $t-1$) non-anomalous datapoints that the algorithm has identified. At time t, a new point $\mathbf{y}_i \in Y_t$ is marked as *anomaly* if it can not be well (linearly) reconstructed using these basis vectors (intuitively, if $\mathbf{y}_i$ does not lie "close" to the space of non-anomalous points). Similar ideas based on using the projection of the data onto a residual subspace as means for detecting anomalies are quite popular and are also known to empirically work well [22, 20, 19, 10, 25].[3]

More formally, let $Y_{[t-1]_{good}}$ be the set of all non-anomalous datapoints that the algorithm has identified till time $t-1$ (i.e., $Y_{[t-1]_{good}}$ is $Y_{[t-1]}$ restricted to non-anomalous datapoints).[4] Consider the rank-k approximation of $Y_{[t-1]_{good}}$ for an appropriately chosen parameter k:[5]

$$Y_{[t-1]_{\mathrm{good}_k}} = \mathrm{SVD}_k(Y_{[t-1]_{\mathrm{good}}}) = U_{t-1_k} \Sigma_{t-1_k} V_{t-1_k}^\top.$$

First observation is that U_{t-1_k} is a "good" rank-k matrix to linearly represent all the points in $Y_{[t-1]_{\mathrm{good}}}$.[6] This follows from the observation that by setting $X = \Sigma_{t-1_k} V_{t-1_k}^\top$:

$$\min_X \|Y_{[t-1]_{\mathrm{good}}} - U_{t-1_k} X\|_F^2 \leq \|Y_{[t-1]_{\mathrm{good}}} - Y_{[t-1]_{\mathrm{good}_k}}\|_F^2.$$

The bound on $\|Y_{[t-1]_{\mathrm{good}}} - Y_{[t-1]_{\mathrm{good}_k}}\|_F^2$ follows from Theorem 1. In many practical scenarios, most of the mass from $Y_{[t]_{\mathrm{good}}}$ would be in its top-k singular values, resulting in $\|Y_{[t-1]_{good}} - Y_{[t-1]_{good_k}}\|_F^2$ being small.

We can now use U_{t-1_k} to detect anomalies in Y_t by following a simple approach. Since U_{t-1_k} is a good basis to linearly reconstruct all the observed non-anomalous points in $Y_{[t-1]}$, we can use it to test whether a point $\mathbf{y}_i \in Y_t$ is "close" to space of non-anomalous points or not. This can be easily achieved by solving the following simple least-squares problem:

$$\min_{\mathbf{x}} \|\mathbf{y}_i - U_{t-1_k}\mathbf{x}\|. \tag{1}$$

As the columns of U_{t-1_k} are orthogonal to each other, this least-squares problem has a simple closed-form solution $\mathbf{x}^* = (U_{t-1_k}^\top U_{t-1_k})^{-1} U_{t-1_k}^\top \mathbf{y}_i = U_{t-1_k}^\top \mathbf{y}_i$. The objective value of (1) at $\mathbf{x}^*$ is used as the anomaly score to decide if $\mathbf{y}_i$ is anomalous or not, with larger objective values denoting anomalies. In other words, the anomaly score for $\mathbf{y}_i$ is $\|(\mathbb{I}_m - U_{t-1_k} U_{t-1_k}^\top)\mathbf{y}_i\|$. Note that this anomaly score is exactly the length of the orthogonal projection of $\mathbf{y}_i$ onto the orthogonal complement U_{t-1_k}.

In Algorithm ANOMDETECT, we present a simple prototype procedure for anomaly detection based on maintaining the left singular vectors (corresponding to the top-k singular values) of the streaming data. Since we have normalized all input points ($\mathbf{y}_i$'s) to have unit L_2-length, the objective values in (1) for all points are in the same scale. The Algorithm ANOMDETECT alternates between an anomaly detection and singular vector updating step. In the anomaly detection step, we use the past singular vectors to detect anomalies among the new incoming points by

[2] The algorithm and the analysis can be easily reworked if this assumption does not hold.

[3] These prior works typically use residual of the principal component representation.

[4] At time $t = 1$, a training set consisting of only non-anomalous data is used to bootstrap the process. As we discuss in Section 5, this training set could be very small.

[5] We defer the discussion on setting of k to later. Readers could think of k as a small number, much smaller than m or n_t.

[6] It is possible to use other (non-SVD) matrix factorization approaches to construct a basis matrix that can linearly represent $Y_{[t-1]_{good}}$, however, using a low-rank SVD is attractive becomes it naturally comes with guarantees of Theorem 1.

Algorithm 1: AnomDetect (prototype algorithm for detecting anomalies at time t)

```
Input: $Y_t \in \mathbf{R}^{m \times n_t}$ (new observance), $U_{t-1_k} \in \mathbf{R}^{m \times k}$ (low-rank matrix with orthogonal columns), and $\zeta \in \mathbf{R}$ (threshold parameter).
Anomaly score construction step:
$Y_{t_{\text{good}}} \leftarrow [\,]$ , $Y_{t_{\text{bad}}} \leftarrow [\,]$
for each point (column) $\mathbf{y}_i \in Y_t$ do
    Solve the following least-squares problem:
    $\mathbf{x}_i^* = \text{argmin}_{\mathbf{x}} \|\mathbf{y}_i - U_{t-1_k}\mathbf{x}\| \quad (\Longrightarrow \mathbf{x}_i^* \leftarrow U_{t-1_k}^\top \mathbf{y}_i)$
    Define anomaly score: $a_i \leftarrow \|(\mathbb{I}_m - U_{t-1_k}U_{t-1_k}^\top)\mathbf{y}_i\|$
    if $a_i \leq \zeta$ then
        $Y_{t_{\text{good}}} \leftarrow [Y_{t_{\text{good}}}|\mathbf{y}_i]$
    else
        $Y_{t_{\text{bad}}} \leftarrow [Y_{t_{\text{bad}}}|\mathbf{y}_i]$   ($\mathbf{y}_i$ is marked as anomaly)
    end if
end for
Updating the singular vectors:
Generate $U_{t_k} \in \mathbf{R}^{m \times k}$, a matrix with orthogonal columns which is (or approximates) the left singular vectors corresponding to top-$k$ singular values of $Y_{[t]_{\text{good}}}$ (could use either Algorithms 2, 3, or 4 here)
Return: $Y_{t_{\text{good}}}$, $Y_{t_{\text{bad}}}$, and $U_{t_k}$
```

thresholding on the objective value of the least-squares problem (1). It is important to note that the thresholding also highly depends on the setting of ζ, which we will analyze in Section 5. We note here that our above framework is reminiscent to that used in *dictionary learning* where the goal is to estimate a collection of basis vectors over which a given data collection can be accurately reconstructed [28, 21]. In that context, U_{t-1_k} is referred to as the dictionary matrix.

The main challenge comes in updating the singular vectors. To start off, we first present an inefficient (but simple) approach based on global SVD updates, and later show how ideas from matrix sketching and randomized low-rank matrix approximations could be used to speedup the updating without any *significant loss in quality*.

4.1 Global Algorithm

The simplest way of updating the singular vectors (without any errors) is to simply (re)generate them from the globally collected sample set $Y_{[t]_{\text{good}}}$. A more mature approach for incrementally and correctly generating the singular vectors of a matrix (with addition of new columns) [5] is outlined in Algorithm GlobalUpdate.

Algorithm 2: GlobalUpdate (global update of the singular vectors at time t)

```
Input: $\hat{U}_{t-1}$, $\hat{\Sigma}_{t-1}$, and $Y_{t_{\text{good}}} \in \mathbf{R}^{m \times n_t}$
$F \leftarrow [\hat{\Sigma}_{t-1}|\hat{U}_{t-1}^\top Y_{t_{\text{good}}}]$
$U_F \Sigma_F V_F^\top \leftarrow \textsc{Svd}(F)$
$\hat{U}_t \leftarrow \hat{U}_{t-1} U_F$
$\hat{\Sigma}_t \leftarrow \Sigma_F$
$\hat{U}_{t_k} \leftarrow [\mathbf{u}_1, \ldots \mathbf{u}_k]$   (where $\hat{U}_t = [\mathbf{u}_1, \ldots, \mathbf{u}_m]$)
Return: $\hat{U}_t$, $\hat{\Sigma}_t$, and $\hat{U}_{t_k}$
```

We mention here that there is one additional line of work, referred to as *Incremental Principal Component Analysis* [16, 23, 7, 3], that attempts to maintain a low-rank approximation of a matrix Z (using SVD and a small amount of bookkeeping) as rows/columns of Z arrive in a stream. The low-rank approximation is updated after addition each new set of rows/columns. However, as noted before in [12], these approaches can have arbitrarily bad matrix approximation error on adversarial data. In Section 5, we also present experimental evidence demonstrating that, for anomaly detection, our approach outperforms a recent incremental PCA technique proposed by [3].

At timestep t, Algorithm GLOBALUPDATE takes $O(m^2 \sum_{j=1}^{t} n_{j_{\text{good}}})$ time, where $n_{j_{\text{good}}}$ denotes the number of columns in the matrix $Y_{j_{\text{good}}}$ (thus $\sum_{j=1}^{t} n_{j_{\text{good}}}$ is the number of columns in $Y_{[t]_{\text{good}}}$). It is obvious that a significant disadvantage of Algorithm GLOBALUPDATE is that both its computational and memory requirement increases with time. We overcome this problem by using an efficient matrix sketching approach outlined next. Importantly, we show that while gaining in computational efficiency, the sketching approach still produces anomaly scores which are similar to that of the Algorithm GLOBALUPDATE.

4.2 Sketching Algorithm

In his recent paper Liberty [26] showed that by adapting the Misra and Gries [30] approach for approximating frequency counts in a stream, one could obtain additive error bounds for matrix sketching. More formally, in the setting of [26], the input is a matrix $Z \in \mathbf{R}^{p\times d}$. In each step, one row of Z is processed by the algorithm (called *Frequent Directions* in [26]) in a streaming fashion, and the algorithm iteratively updates a matrix $Q \in \mathbf{R}^{q\times d}$ ($q \ll p$) such that for any unit vector $\mathbf{x} \in \mathbf{R}^d$, $\|Z\mathbf{x}\|^2 - \|Q\mathbf{x}\|^2 \leq 2\|Z\|_F^2/q$. Recently Ghashami and Philips [12] reanalyzed the *Frequent Directions* algorithm to show that it provides relative error bounds for low-rank matrix approximation. Instead of Q, their algorithm return Q_k (a rank-k approximation of Q) and their main result shows that $\|Z\|_F^2 - \|Q_k\|_F^2 \leq q/(q-k) \cdot \|Z - Z_k\|_F^2$.

Our approach for updating the singular vectors (outlined in Algorithm RANDSKETCHUPDATE) is based on extending the *Frequent Directions* algorithm of [26] to a more general setting. In contrast to [26, 12], where one new row (or column) is added at every timestep t, we add $n_t \gg 1$ new columns. With this generality, for computational efficiency, at each timestep, we perform a low-rank SVD, instead of the full SVD as in [26, 12, 13]. This could be accomplished by computing either the actual low-rank SVD or approximating it using a computationally faster randomized approach. We focus on the later here because computational efficiency is paramount in a streaming setting, and as our experimental results suggest, for anomaly detection, the performance of the randomized approach is almost identical to performing an exact low-rank SVD (see the experimental results in Section 5).[7]

Randomized low-rank matrix approximation has been a subject of lot of recent research with approaches based on sparsification, column selection, dimensionality reduction, etc., been devised for solving many matrix problems (see [15] and references therein). Here we use a simple technique suggested by [15] that is based on combining a randomized pre-processing step (multiplying by a random matrix and QR decomposition) along with a simple post-processing step (eigenvalue decomposition of a small matrix).

At timestep t, Algorithm RANDSKETCHUPDATE takes $O(\ell T_{\text{mult}} + (m + n_t)\ell^2)$ time, where T_{mult} denotes the cost of a matrix-vector multiplication with the input matrix M_t. The matrix-vector multiplication is a well-studied topic with numerous known efficient sequential/parallel algorithms. Between iterations the algorithm only needs to store the E_t matrices (the up-to-date randomized matrix sketch) which take $O(m\ell)$ storage. We discuss the setting of ℓ in the next section.

4.3 Analysis of Algorithm RANDSKETCHUPDATE

In this section, we show that the anomaly detection results obtained by using $\breve{U}_{t_k}$ (output of Algorithm RANDSKETCHUPDATE) in Algorithm ANOMDETECT is similar to using the (true) singular vectors based on a global update (output of Algorithm GLOBALUPDATE).

[7] For completeness, in Appendix A, we present Algorithm SKETCHUPDATE for singular value updation based on the exact low-rank SVD and its analysis.

Algorithm 3: RANDSKETCHUPDATE (randomized streaming update of the singular vectors at time t)

```
Input: $Y_{t_{\text{good}}} \in \mathbf{R}^{m \times n_t}$, and $E_{t-1} \in \mathbf{R}^{m \times \ell}$ the randomized matrix sketch computed at time $t-1$
$M_t \leftarrow [E_{t-1} | Y_{t_{\text{good}}}]$
$r \leftarrow 100\ell$
Generate an $m \times r$ random Gaussian matrix $\Omega$
$Y \leftarrow M_t M_t^\top \Omega$
$QR \leftarrow \textsc{Qr}(Y)$ (QR factorization for $Y$)
$A_t \breve{\Sigma}_t^2 A_t^\top \leftarrow \textsc{Eig}(Q^\top M_t M_t^\top Q)$    (with $\breve{\Sigma}_t^2 = \mathrm{d}iag(\breve{\sigma}_{t_1}^2, \dots, \breve{\sigma}_{t_r}^2)$)
$\breve{U}_t \leftarrow Q A_t$    ($QQ^\top M_t M_t^\top Q Q^\top$ approximates $M_t M_t^\top$)
$\breve{U}_{t_\ell} \leftarrow [\mathbf{u}_1, \dots, \mathbf{u}_\ell]$    (where $\breve{U}_{t_r} = [\mathbf{u}_1, \dots, \mathbf{u}_r]$ and $\ell \leq r$)
$\breve{\Sigma}_{t_\ell}^{(trunc)} \leftarrow \mathrm{d}iag\left(\sqrt{\breve{\sigma}_{t_1}^2 - \breve{\sigma}_{t_\ell}^2}, \sqrt{\breve{\sigma}_{t_2}^2 - \breve{\sigma}_{t_\ell}^2}, \dots, \sqrt{\breve{\sigma}_{t_{\ell-1}}^2 - \breve{\sigma}_{t_\ell}^2}, 0\right)$
$E_t \leftarrow \breve{U}_{t_\ell} \breve{\Sigma}_{t_\ell}^{(trunc)}$
Return: $E_t$ and $\breve{U}_{t_k}$
```

Our proof generalizes the analysis of *Frequent Directions* by [12], and then carefully combines it with known matrix perturbation results. Our first aim will be to bound the Frobenius norm of the difference between $Y_{[t]_{\text{good}_k}} Y_{[t]_{\text{good}_k}}^\top$ and $E_{t_k} E_{t_k}^\top$, for which we will use the following result from [15] that bounds the error due to the randomized SVD.

Theorem 3. *(Restated from Corollary 10.9 of [15]) Consider Algorithm* RANDSKETCHUPDATE *at timestep t. Let $\mathrm{d}iag(\bar{\sigma}_{t_1}, \dots, \bar{\sigma}_{t_m})$ be the eigenvalues of $M_t M_t^\top$, then with probability at least $1 - 6e^{-99\ell}$, $\|M_t M_t^\top - \breve{U}_t \breve{\Sigma}_t^2 \breve{U}_t\| \leq 38\bar{\sigma}_{t_{\ell+1}} + 2(\sum_{i=\ell+1}^m \bar{\sigma}_{t_i}^2)^{1/2}/\sqrt{\ell}$.*

We will need a few additional notations:

1. $E_{t_k} = \breve{U}_{t_k} \breve{\Sigma}_{t_k}^{(trunc)}$ (rank-k approximation of E_t),
2. $\breve{\Delta}_t = \sum_{j=1}^t \breve{\sigma}_{j_\ell}^2$,
3. $\upsilon_j = 38\bar{\sigma}_{j_{\ell+1}} + 2\left(\sum_{i=\ell+1}^m \bar{\sigma}_{j_i}^2\right)^{1/2} / \sqrt{\ell}$ (error bound from Theorem 3, at timestep j),
4. $\kappa = \sigma_1(Y_{[t]_{\text{good}}})/\sigma_k(Y_{[t]_{\text{good}}})$, where $\sigma_1(Y_{[t]_{\text{good}}}) \geq \cdots \geq \sigma_m(Y_{[t]_{\text{good}}})$ are the singular values of $Y_{[t]_{\text{good}}}$,
5. $N_t = QQ^\top M_t$,
6. $P_t = QA_t\breve{\Sigma}_t = \breve{U}_t \breve{\Sigma}_t$ (by construction in Algorithm RANDSKETCHUPDATE, $N_t N_t^\top = P_t P_t^\top$).

As columns of Q are orthogonal to each other, $QQ^\top$ is a projection matrix, and therefore by standard properties of projection matrices and noting that $(QQ^\top)^\top = QQ^\top$,

$$\|M_t\|_F^2 \geq \|QQ^\top M_t\|_F^2 = \|N_t\|_F^2 = \|P_t\|_F^2. \tag{2}$$

Similarly for all unit vectors $\mathbf{x} \in \mathbf{R}^m$,

$$\|M_t^\top \mathbf{x}\|^2 \geq \|(QQ^\top M_t)^\top \mathbf{x}\|^2 = \|N_t^\top \mathbf{x}\|^2 = \|P_t^\top \mathbf{x}\|^2. \tag{3}$$

For ease of presentation, in the following, we are going to assume, that $t \cdot 6e^{-99\ell} \ll 1$.[8] Note that $e^{-99\ell}$ is a very tiny number (as we set $\ell \approx \sqrt{m}$).

Lemma 4. *At timestep t, Algorithm* RANDSKETCHUPDATE *maintains that: $\|Y_{[t]_{\text{good}}}\|_F^2 - \|E_t\|_F^2 \geq \ell\breve{\Delta}_t$.*

[8] If t gets extremely large such that this inequality is violated, then one could use a slightly larger r in Algorithm RANDSKETCHUPDATE.

Proof. At timestep t, $\|M_t\|_F^2 = \|E_{t-1}\|_F^2 + \|Y_{t_{\text{good}}}\|_F^2$. We also have $\|P_t\|_F^2 \geq \|E_t\|_F^2 + \ell\breve{\sigma}^2_{t_\ell}$. Since, $\|M_t\|_F^2 \geq \|P_t\|_F^2$ (from (2)), we have $\|M_t\|_F^2 \geq \|E_t\|_F^2 + \ell\breve{\sigma}^2_{t_\ell}$. Solving for $\|Y_{t_{\text{good}}}\|_F^2$ from these inequalities and summing over all $j \leq t$, we get,

$$\|Y_{[t]_{\text{good}}}\|_F^2 = \sum_{j=1}^{t} \|Y_{j_{\text{good}}}\|_F^2 \geq \sum_{j=1}^{t} (\|E_j\|_F^2 - \|E_{j-1}\|_F^2 + \ell\breve{\sigma}^2_{j_\ell}) \geq \|E_t\|_F^2 + \ell\breve{\Delta}_t.$$

The last line follows as E_0 is all zeros matrix. □

The following lemma shows that for any direction $\mathbf{x}$, $Y_{[t]_{\text{good}}}$ and E_t are with high probability not too far apart. Define

$$\Upsilon_t = \sum_{j=1}^{t} \upsilon_j. \tag{4}$$

Lemma 5. *For any unit vector $\mathbf{x} \in \mathbf{R}^m$, at any timestep t, $0 \leq \|Y_{[t]_{\text{good}}}^\top \mathbf{x}\|^2 - \|E_t^\top \mathbf{x}\|^2$, and with probability at least $1 - t \cdot 6e^{-99\ell}$,*

$$\|Y_{[t]_{\text{good}}}^\top \mathbf{x}\|^2 - \|E_t^\top \mathbf{x}\|^2 \leq \breve{\Delta}_t + \Upsilon_t.$$

Proof. To show $|Y_{[t]_{\text{good}}}^\top \mathbf{x}\|^2 - \|E_t^\top \mathbf{x}\|^2 > 0$, observe that $\|E_{t-1}^\top \mathbf{x}\|^2 + \|Y_{t_{\text{good}}}^\top \mathbf{x}\|^2 = \|M_t^\top \mathbf{x}\|^2$. Since $\|P_t^\top \mathbf{x}\|^2 \geq \|E_t^\top \mathbf{x}\|^2$ (by construction) and $\|M_t^\top \mathbf{x}\|^2 \geq \|P_t^\top \mathbf{x}\|^2$ (from (3)), we have,

$$\|Y_{[t]_{\text{good}}}^\top \mathbf{x}\|^2 = \sum_{j=1}^{t} \|Y_{j_{\text{good}}}^\top \mathbf{x}\|^2 \geq \sum_{j=1}^{t} (\|E_j^\top \mathbf{x}\|^2 - \|E_{j-1}^\top \mathbf{x}\|^2) \geq \|E_t^\top \mathbf{x}\|^2.$$

Here we used that E_0 is an all zeros matrix. Now let us concentrate on showing

$$\|Y_{[t]_{\text{good}}}^\top \mathbf{x}\|^2 - \|E_t^\top \mathbf{x}\|^2 \leq \Upsilon_t + \breve{\Delta}_t.$$

Let $\mathbf{u}_i$ be the ith column in $\breve{U}_t$. $\breve{\sigma}^2_{t_i} - \breve{\sigma}^2_{t_\ell}$ is the ith singular value of E_t. Let $R_p = \mathrm{r}ank(P_t)$.

$$\begin{aligned}
\|P_t^\top \mathbf{x}\|^2 &= \sum_{i=1}^{R_p} \breve{\sigma}^2_{t_i} \langle \mathbf{u}_i, \mathbf{x}\rangle^2 = \sum_{i=1}^{R_p} (\breve{\sigma}^2_{t_i} + \breve{\sigma}^2_{t_\ell} - \breve{\sigma}^2_{t_\ell}) \langle \mathbf{u}_i, \mathbf{x}\rangle^2 \\
&= \sum_{i=1}^{R_p} (\breve{\sigma}^2_{t_i} - \breve{\sigma}^2_{t_\ell}) \langle \mathbf{u}_i, \mathbf{x}\rangle^2 + \sum_{i=1}^{R_p} \breve{\sigma}^2_{t_\ell} \langle \mathbf{u}_i, \mathbf{x}\rangle^2 \\
&\leq \sum_{i=1}^{\ell} (\breve{\sigma}^2_{t_i} - \breve{\sigma}^2_{t_\ell}) \langle \mathbf{u}_i, \mathbf{x}\rangle^2 + \breve{\sigma}^2_{t_\ell} \sum_{i=1}^{R_p} \langle \mathbf{u}_i, \mathbf{x}\rangle^2 \leq \|E_t^\top \mathbf{x}\|^2 + \breve{\sigma}^2_{t_\ell}.
\end{aligned}$$

For the first inequality we used that for $i > \ell$, $\breve{\sigma}^2_{t_i} \leq \breve{\sigma}^2_{t_\ell}$. For the second inequality, we use that $\sum_{i=1}^{R_p} \langle \mathbf{u}_i, \mathbf{x}\rangle^2 \leq \|\mathbf{x}\|^2 = 1$ (as $\mathbf{x}$ is a unit vector).

Since for all unit vectors $\mathbf{x} \in \mathbf{R}^m$, $\|M_t^\top \mathbf{x}\|^2 - \|P_t^\top \mathbf{x}\|^2 \leq \|M_t M_t^\top - P_t P_t^\top\|$, we get with probability at least $1 - 6e^{-99\ell}$,

$$\|M_t^\top \mathbf{x}\|^2 \leq \|P_t^\top \mathbf{x}\|^2 + \|M_t M_t^\top - P_t P_t^\top\| = \|P_t^\top \mathbf{x}\|^2 + \upsilon_t.$$

Since $\|M_t^\top \mathbf{x}\|^2 = \|E_{t-1}^\top \mathbf{x}\|^2 + \|Y_{t_{\text{good}}}^\top \mathbf{x}\|^2$, we get with probability at least $1 - 6e^{-99\ell}$,

$$\|E_{t-1}^\top \mathbf{x}\|^2 + \|Y_{t_{\text{good}}}^\top \mathbf{x}\|^2 \leq \upsilon_t + \|E_t^\top \mathbf{x}\|^2 + \breve{\sigma}^2_{t_\ell}.$$

Subtracting $\|E_{t-1}^\top \mathbf{x}\|^2$ from both sides and summing over $j \leq t$ with a union bound for probabilities, we get that with probability at least $1 - t \cdot 6e^{-99\ell}$,

$$\begin{aligned}\|Y_{[t]_{\text{good}}}^\top \mathbf{x}\|^2 &= \sum_{j=1}^{t} \|Y_{j_{\text{good}}}^\top \mathbf{x}\|^2 \\ &\leq \sum_{j=1}^{t} (\|E_j^\top \mathbf{x}\|^2 - \|E_{j-1}^\top \mathbf{x}\|^2 + \breve{\sigma}_{j_\ell}^2 + \upsilon_j) = \|E_t^\top \mathbf{x}\|^2 + \breve{\Delta}_t + \Upsilon_t.\end{aligned}$$

Again we used that E_0 is an all zeros matrix. □

Since for all unit vectors $\mathbf{x} \in \mathbf{R}^m$,

$$\|Y_{[t]_{\text{good}}}^\top \mathbf{x}\|^2 - \|E_t^\top \mathbf{x}\|^2 \geq 0 \Longrightarrow Y_{[t]_{\text{good}}} Y_{[t]_{\text{good}}}^\top \succeq E_t E_t^\top.$$

From Claim 2, for all unit vectors $\mathbf{x} \in \mathbf{R}^m$, $\kappa \|Y_{[t]_{\text{good}_k}}^\top \mathbf{x}\| \geq \|Y_{[t]_{\text{good}}}^\top \mathbf{x}\|$. Therefore,

$$\kappa^2 Y_{[t]_{\text{good}_k}} Y_{[t]_{\text{good}_k}}^\top \succeq Y_{[t]_{\text{good}}} Y_{[t]_{\text{good}}}^\top \succeq E_t E_t^\top \succeq E_{t_k} E_{t_k}^\top.$$

Lemma 6. *Let $Y_{[t]_{\text{good}_k}}$ be the best rank-k approximation to $Y_{[t]_{\text{good}}}$. Then with probability at least $1 - t \cdot 6e^{-99\ell}$, $\breve{\Delta}_t \leq (\|Y_{[t]_{\text{good}}} - Y_{[t]_{\text{good}_k}}\|_F^2 + k\Upsilon_t)/(\ell - k)$.*

Proof. From Lemma 4, $\|Y_{[t]_{\text{good}}}\|_F^2 - \|E_t\|_F^2 \geq \ell \breve{\Delta}_t$. Let $R_y = \mathrm{r}ank(Y_{[t]_{\text{good}}})$ and $\mathbf{v}_1, \ldots, \mathbf{v}_{R_y}$ be the left singular vectors of $Y_{[t]_{\text{good}}}$ corresponding to the non-zero singular values of $Y_{[t]_{\text{good}}}$, we have with probability at least $1 - t \cdot 6e^{-99\ell}$,

$$\begin{aligned}\ell \breve{\Delta}_t &\leq \|Y_{[t]_{\text{good}}}\|_F^2 - \|E_t\|_F^2 \\ &= \sum_{i=1}^{k} \|Y_{[t]_{\text{good}}}^\top \mathbf{v}_i\|^2 + \sum_{i=k+1}^{R_y} \|Y_{[t]_{\text{good}}}^\top \mathbf{v}_i\|^2 - \|E_t\|_F^2 \\ &= \sum_{i=1}^{k} \|Y_{[t]_{\text{good}}}^\top \mathbf{v}_i\|^2 + \|Y_{[t]_{\text{good}}} - Y_{[t]_{\text{good}_k}}\|_F^2 - \|E_t\|_F^2 \\ &\leq \sum_{i=1}^{k} \|Y_{[t]_{\text{good}}}^\top \mathbf{v}_i\|^2 + \|Y_{[t]_{\text{good}}} - Y_{[t]_{\text{good}_k}}\|_F^2 - \sum_{i=1}^{k} \|E_t^\top \mathbf{v}_i\|^2 \\ &= \|Y_{[t]_{\text{good}}} - Y_{[t]_{\text{good}_k}}\|_F^2 + \sum_{i=1}^{k} (\|Y_{[t]_{\text{good}}}^\top \mathbf{v}_i\|^2 - \|E_t^\top \mathbf{v}_i\|^2) \leq \|Y_{[t]_{\text{good}}} - Y_{[t]_{\text{good}_k}}\|_F^2 + k(\Upsilon_t + \breve{\Delta}_t).\end{aligned}$$

First inequality uses that $\sum_{i=1}^k \|E_t^\top \mathbf{v}_i\|^2 \leq \|E_t\|_F^2$, and the last inequality is based on Lemma 5. Solving for $\breve{\Delta}_t$ in the above inequality gives the claimed result. □

Using Lemma 6, we can relate $\|Y_{[t]_{\text{good}_k}}\|_F^2$ to $\|E_{t_k}\|\|_F^2$.

Lemma 7. *At any timestep t, $0 \leq \|Y_{[t]_{\text{good}_k}}\|_F^2 - \|E_{t_k}\|_F^2$, and with probability at least $1 - t \cdot 6e^{-99\ell}$,*

$$\|Y_{[t]_{\text{good}_k}}\|_F^2 - \|E_{t_k}\|_F^2 \leq k\Upsilon_t + k(\|Y_{[t]_{\text{good}}} - Y_{[t]_{\text{good}_k}}\|_F^2 + k\Upsilon_t)/(\ell - k).$$

Proof. As in Lemma 6, let $\mathbf{v}_1, \dots, \mathbf{v}_k$ be the left singular vectors of $Y_{[t]_{good}}$ corresponding to its top-k singular values. Let $\mathbf{u}_1, \dots, \mathbf{u}_k$ be the left singular vectors of E_t corresponding to its top-k singular values. We have

$$\|Y_{[t]_{\text{good}_k}}\|_F^2 = \sum_{i=1}^{k} \|Y_{[t]_{\text{good}}}^\top \mathbf{v}_i\|^2 \geq \sum_{i=1}^{k} \|Y_{[t]_{\text{good}}}^\top \mathbf{u}_i\|^2 \geq \sum_{i=1}^{k} \|E_t^\top \mathbf{u}_i\|^2 = \|E_{t_k}\|_F^2.$$

This proves that $0 \leq \|Y_{[t]_{\text{good}_k}}\|_F^2 - \|E_{t_k}\|_F^2$. The upper bound can be established by noticing that with probability at least $1 - t \cdot 6e^{-99\ell}$,

$$\|E_{t_k}\|_F^2 \geq \sum_{i=1}^{k} \|E_{t_k}^\top \mathbf{v}_i\|^2 \geq \sum_{i=1}^{k} (\|Y_{[t]_{\text{good}}}^\top \mathbf{v}_i\|^2 - \Upsilon_t - \breve{\Delta}_t) = \|Y_{[t]_{\text{good}_k}}\|_F^2 - k\Upsilon_t - k\breve{\Delta}_t,$$

where the second inequality follows from Lemma 5. Now substituting for $\breve{\Delta}_t$ from Lemma 6 gives the result. □

Using this above lemma and the fact that $\kappa^2 Y_{[t]_{\text{good}_k}} Y_{[t]_{\text{good}_k}}^\top \succeq E_{t_k} E_{t_k}^\top$, we can prove the following proposition.

Proposition 8. *At timestep t, E_{t_k} generated by Algorithm* RANDSKETCHUPDATE *satisfies,*

$$\|\kappa^2 Y_{[t]_{\text{good}_k}} Y_{[t]_{\text{good}_k}}^\top - E_{t_k} E_{t_k}^\top\|_F \leq \kappa^2 \|Y_{[t]_{\text{good}_k}}\|_F^2 - \|E_{t_k}\|_F^2.$$

Proof. For a positive semidefinite matrix, the trace is greater than or equal to the Frobenius norm. Since, we have established that $\kappa^2 Y_{[t]_{\text{good}_k}} Y_{[t]_{\text{good}_k}}^\top - E_{t_k} E_{t_k}^\top$ is a positive semidefinite matrix.

$$\begin{aligned}
&\|\kappa^2 Y_{[t]_{\text{good}_k}} Y_{[t]_{\text{good}_k}}^\top - E_{t_k} E_{t_k}^\top\|_F \\
&\leq \mathrm{tr}(\kappa^2 Y_{[t]_{\text{good}_k}} Y_{[t]_{\text{good}}}^\top - E_{t_k} E_{t_k}^\top) \\
&= \kappa^2 tr \left(Y_{[t]_{\text{good}_k}} Y_{[t]_{\text{good}_k}}^\top \right) - \mathrm{tr}(E_{t_k} E_{t_k}^\top) = \kappa^2 \|Y_{[t]_{\text{good}_k}}\|_F^2 - \|E_{t_k}\|_F^2.
\end{aligned}$$

The first inequality follows from the trace-Frobenius inequality of positive semidefinite matrices. □

We need couple of more definitions. Define Φ_a as,

$$\Phi_a = \frac{\kappa^2 \|Y_{[t]_{\text{good}_k}}\|_F^2 - \|E_{t_k}\|_F^2}{\|Y_{[t]_{\text{good}_k}}\|_F^2 - \|E_{t_k}\|_F^2}. \tag{5}$$

Note that $\Phi_a \geq 1$ as $\|Y_{[t]_{\text{good}_k}}\|_F^2 \geq \|E_{t_k}\|_F^2$ (from Lemma 7). In fact, for small k's (as in our setting), typically κ (the ratio between the largest and kth largest singular value of $Y_{[t]_{\text{good}}}$) is bounded, yielding $\Phi_a = O(1)$.

Define Φ_b as,

$$\Phi_b = \frac{\|\kappa^2 Y_{[t]_{\text{good}}} Y_{[t]_{\text{good}}}^\top - E_t E_t^\top\|}{\|\kappa^2 Y_{[t]_{\text{good}_k}} Y_{[t]_{\text{good}_k}}^\top - E_{t_k} E_{t_k}^\top\|}. \tag{6}$$

Claim 9.

$$\Phi_b \leq 1 + 2/(\kappa^2 - \|E_t\|^2/\|Y_{[t]_{\text{good}}}\|^2).$$

Proof. A simple manipulation show that the numerator of Φ_b,

$$\|\kappa^2 Y_{[t]_{\text{good}}} Y_{[t]_{\text{good}}}^\top - E_t E_t^\top\| \leq \kappa^2 \|Y_{[t]_{\text{good}}} Y_{[t]_{\text{good}}}^\top - Y_{[t]_{\text{good}_k}} Y_{[t]_{\text{good}_k}}^\top\| + \|E_t E_t^\top - E_{t_k} E_{t_k}^\top\| + \|\kappa^2 Y_{[t]_{\text{good}_k}} Y_{[t]_{\text{good}_k}}^\top - E_{t_k} E_{t_k}^\top\|.$$

Note that using Theorem 1,

$$\|Y_{[t]_{\text{good}}} Y_{[t]_{\text{good}}}^\top - Y_{[t]_{\text{good}_k}} Y_{[t]_{\text{good}_k}}^\top\| = \sigma_{k+1}^2,$$

where σ_{k+1} is the $(k+1)$st singular value of $Y_{[t]_{\text{good}}}$. Similarly by using Theorem 1, $\|E_t E_t^\top - E_{t_k} E_{t_k}^\top\|$ is equal to the square of the $(k+1)$st singular value of E_t. Since we have already established $Y_{[t]_{\text{good}}} Y_{[t]_{\text{good}}}^\top - E_t E_t^\top \succeq 0$, this implies that $\|E_t E_t^\top - E_{t_k} E_{t_k}^\top\| \leq \sigma_{k+1}^2$.

Let $\|Y_{[t]_{\text{good}}}\| = \sigma_1$. Substituting these observations into Φ_b, we get,

$$\Phi_b \leq 1 + \frac{(\kappa^2+1)\sigma_{k+1}^2}{\|\kappa^2 Y_{[t]_{\text{good}_k}} Y_{[t]_{\text{good}_k}}^\top - E_{t_k} E_{t_k}^\top\|} \leq 1 + \frac{(\kappa^2+1)\sigma_{k+1}^2}{\kappa^2\sigma_1^2 - \|E_t\|^2}.$$

The last inequality follows as by Weyl's inequality [14] the largest eigenvalue of $\|\kappa^2 Y_{[t]_{\text{good}_k}} Y_{[t]_{\text{good}_k}}^\top - E_{t_k} E_{t_k}^\top\|$ is greater than equal to $\kappa^2\|Y_{[t]_{\text{good}_k}}\|^2 - \|E_{t_k}\|^2$. We also used that $\|Y_{[t]_{\text{good}_k}}\|^2 = \|Y_{[t]_{\text{good}}}\|^2$ and $\|E_t\|^2 = \|E_{t_k}\|^2$. Since, $\kappa \leq \sigma_1/\sigma_{k+1}$, bound on Φ_b can be re-expressed as,

$$\Phi_b \leq 1 + \frac{(\kappa^2+1)\frac{\sigma_1^2}{\kappa^2}}{\kappa^2\sigma_1^2 - \|E_t\|^2} \leq 1 + \frac{2}{\kappa^2 - \|E_t\|^2/\sigma_1^2}.$$

Here we used that $(\kappa^2+1)/\kappa^2 \leq 2$ as $\kappa \geq 1$. □

Note that $\|E_t\|^2 \leq \|Y_{[t]_{\text{good}}}\|^2$ (as $Y_{[t]_{\text{good}}} Y_{[t]_{\text{good}}}^\top \succeq E_t E_t^\top$) . Typically κ is also bounded away from 1, yielding $\Phi_b = O(1)$.

We now use the theory of matrix perturbation to relate $\breve{U}_{t_k}$ (from Algorithm RANDSKETCHUPDATE) with the (true) left singular vectors corresponding to top-k singular values of $Y_{[t]_{\text{good}}}$. There is lot of prior work in matrix perturbation theory that relates the eigenvalues, singular values, eigenspaces, and singular subspaces, etc., of the matrix $Z + Z'$ to the corresponding quantity in Z, under various conditions on the matrices Z and Z'. Here we use a recent result from Chen *et al.* [9] that studies behavior of the eigenvector matrix of a Hermitian (symmetric) matrix under a small perturbation.

Theorem 10. *(Restated from Theorem 2.1 [9]) Let $A \in \mathbf{R}^{m\times m}$ be a symmetric matrix with distinct eigenvalues with* $\textsc{Eig}(A) = U\Lambda U^\top$ *where* $\Lambda = diag(\lambda_1,\dots,\lambda_m)$*. Let* $A_{\text{per}} = A + \Phi$ *be a symmetric matrix. Let* $L = L(\Lambda) = \min_{i\neq j}|\lambda_i - \lambda_j| > 0$*,* $\beta = \|\Phi\|_F/L$*, and* $\alpha = 2\|A\|/L$*, with β satisfying:* $\beta \leq 1/(1+4\alpha)$*. Then* $\textsc{Eig}(A_{\text{per}}) = U_{\text{per}}\Lambda_{\text{per}}U_{\text{per}}^\top$ *such that* $\|U - U_{\text{per}}\|_F \leq \sqrt{2}\beta/(1+4\alpha^2)^{1/4}$.

We now can apply Proposition 8 and Theorem 10 to bound $\|\hat{U}_{t_k} - \breve{U}_{t_k}\|_F$. We do so by constructing matrices: $A = \kappa^2 Y_{[t]_{\text{good}}} Y_{[t]_{\text{good}}}^\top$ and $A_{\text{per}} = E_t E_t^\top$.

Let ℓ be such that:

$$\frac{\sqrt{m}\Phi_b\Phi_a k\Upsilon_t}{L} + \frac{\sqrt{m}\Phi_b\Phi_a k(\|Y_{[t]_{\text{good}}} - Y_{[t]_{\text{good}_k}}\|_F^2 + k\Upsilon_t)}{(\ell-k)L} \leq \frac{L}{L + 4\kappa^2\|Y_{[t]_{\text{good}}}\|^2}. \tag{7}$$

An important point to note in the above equation (7) is that both terms in the left-hand side are monotonically decreasing functions in ℓ (for the first term, Υ_t decreases with ℓ).

Claim 11. *Let λ_i be the ith eigenvalue of* $Y_{[t]_{\text{good}}} Y_{[t]_{\text{good}}}^\top$ *and* $L = \min_{i\neq j}|\lambda_i - \lambda_j| > 0$*. If ℓ satisfies (7) for* $\Upsilon_t, \Phi_a, \Phi_b$ *defined in (4), (5), (6) respectively, then with probability at least* $1 - t\cdot 6e^{-99\ell}$,

$$\|\hat{U}_{t_k} - \breve{U}_{t_k}\|_F \leq \sqrt{2}L/\left(\sqrt{L + 8\kappa^2\|Y_{[t]_{\text{good}}}\|^2}\sqrt[4]{L^2 + 16\kappa^4\|Y_{[t]_{\text{good}}}\|^4}\right).$$

Proof. Set $A = \kappa^2 Y_{[t]_{\text{good}}} Y_{[t]_{\text{good}}}^\top$ and $A_{\text{per}} = E_t E_t^\top$. Now $\alpha = 2\|A\|/L$. Concentrating on β, with probability at least $1 - t \cdot 6e^{-99\ell}$,

$$\begin{aligned}\beta = \frac{\|A - A_{\text{per}}\|_F}{L} &\leq \frac{\sqrt{m}\Phi_b \|\kappa^2 Y_{[t]_{\text{good}_k}} Y_{[t]_{\text{good}_k}}^\top - E_{t_k} E_{t_k}^\top\|_F}{L} \\ &\leq \frac{\sqrt{m}\Phi_b (\kappa^2 \|Y_{[t]_{\text{good}_k}}\|_F^2 - \|E_{t_k}\|_F^2)}{L} \\ &= \frac{\sqrt{m}\Phi_b \Phi_a (\|Y_{[t]_{\text{good}_k}}\|_F^2 - \|E_{t_k}\|_F^2)}{L} \\ &\leq \frac{\sqrt{m}\Phi_b\Phi_a k \Upsilon_t}{L} + \frac{\sqrt{m}\Phi_b\Phi_a k (\|Y_{[t]_{\text{good}}} - Y_{[t]_{\text{good}_k}}\|_F^2 + k\Upsilon_t)}{(\ell - k)L}.\end{aligned}$$

The last inequality follows from Lemma 7. To apply Theorem 10, we need to satisfy the condition of $\beta \leq 1/(1+4\alpha)$. This translates to setting ℓ to satisfy (assuming $k < \sqrt{m}$ and the Lemma 7 holds),

$$\frac{\sqrt{m}\Phi_b\Phi_a k \Upsilon_t}{L} + \frac{\sqrt{m}\Phi_b\Phi_a k (\|Y_{[t]_{\text{good}}} - Y_{[t]_{\text{good}_k}}\|_F^2 + k\Upsilon_t)}{(\ell - k)L} \leq \frac{L}{L + 4\kappa^2 \|Y_{[t]_{\text{good}}}\|^2}.$$

The eigendecomposition of $Y_{[t]_{\text{good}}} Y_{[t]_{\text{good}}}^\top$ is:

$$Y_{[t]_{\text{good}}} Y_{[t]_{\text{good}}}^\top = \hat{U}_t \hat{\Sigma}_t \hat{U}_t^\top.$$

Similarly the eigendecomposition of $E_t E_t^\top$ is:

$$E_t E_t^\top = [\breve{U}_t | \emptyset_{r+1}, \dots, \emptyset_m] \times \mathrm{d}iag(\breve{\sigma}_{t_1}^2 - \breve{\sigma}_{t_\ell}^2, \dots, \breve{\sigma}_{t_{\ell-1}}^2 - \breve{\sigma}_{t_\ell}^2, 0, \dots, 0) \times [\breve{U}_t | \emptyset_{r+1}, \dots, \emptyset_m]^\top,$$

where $[\breve{U}_t | \emptyset_{r+1}, \dots, \emptyset_m]$ is an $m \times m$ orthogonal matrix. Note that $\breve{U}_t$ is an $m \times r$ matrix. The actual choice of $\emptyset_{r+1}, \dots, \emptyset_m$ will not matter for our result.

Substituting the values of $\beta \leq 1/(1 + 4\alpha)$ and α, we have by the bound of Theorem 10,

$$\|\hat{U}_t - [\breve{U}_t | \emptyset_{r+1}, \dots, \emptyset_m]\|_F \leq \frac{\sqrt{2}L}{\sqrt{L + 8\kappa^2 \|Y_{[t]_{\text{good}}}\|^2} \sqrt[4]{L^2 + 16\kappa^4 \|Y_{[t]_{\text{good}}}\|^4}}.$$

Noting that $\|\hat{U}_{t_k} - \breve{U}_{t_k}\|_F \leq \|\hat{U}_t - [\breve{U}_t | \emptyset_{r+1}, \dots, \emptyset_m]\|_F$ (as $\hat{U}_{t_k} - \breve{U}_{t_k}$ is a submatrix of $\hat{U}_t - [\breve{U}_t | \emptyset_{r+1}, \dots, \emptyset_m]$) completes the proof. □

Neither the numerical constants nor the precise form of the bound on ℓ in (7) are optimal because of the slackness in Theorem 10. The bound on ℓ in (7) could be simplified a bit for some interesting cases, e.g., when k is small and $1 < \kappa \leq O(1)$ then $\Gamma_a = O(1)$ and $\Gamma_b = O(1)$.

Remark: The assumption of $L > 0$ is something that is commonly satisfied in practice, especially if m is reasonably smaller than the number of datapoints in $Y_{[t]_{\text{good}}}$. The bound on ℓ from (7) should be treated as an existential result, as setting ℓ using it is tricky. Practically, we noticed that setting $\ell \approx \sqrt{m}$ suffices to get good results. Another important point to remember is that the Algorithm RANDSKETCHUPDATE can be used *with any value* of ℓ, the above bound on ℓ is only to ensure that its performance is similar to using global singular value decomposition updates in Algorithm ANOMDETECT (Theorem 13).

We can now compare the anomaly scores generated by using either $\hat{U}_{t_k}$ or $\breve{U}_{t_k}$ in Algorithm ANOMDETECT.

Claim 12. *Let* $\mathbf{x}_g^* = argmin_{\mathbf{x}} \|\mathbf{y} - \hat{U}_{t_k}\mathbf{x}\|$ *and let* $\mathbf{x}_s^* = argmin_{\mathbf{x}} \|\mathbf{y} - \breve{U}_{t_k}\mathbf{x}\|$. *Then,*

$$\Big| \|\mathbf{y} - \hat{U}_{t_k}\mathbf{x}_g^*\| - \|\mathbf{y} - \breve{U}_{t_k}\mathbf{x}_s^*\| \Big| \leq \|\hat{U}_{t_k} - \breve{U}_{t_k}\|_F.$$

Dataset	# Datapoints	#Features	% of Anomalies
Cod-RNA	488,565	8	33.33%
Protein-homology	145,751	74	0.89%
User-activity	129,328	83	10.69%
RCV1AD	100,274	1000	18.12%

Table 1: Statistics of the experimental datasets.

Proof. For any fixed $\mathbf{y} \in \mathbf{R}^m$ and $\mathbf{x} \in \mathbf{R}^k$,

$$\left| \|\mathbf{y} - \hat{U}_{t_k}\mathbf{x}\| - \|\mathbf{y} - \breve{U}_{t_k}\mathbf{x}\| \right| \leq \|\hat{U}_{t_k}\mathbf{x} - \breve{U}_{t_k}\mathbf{x}\| \leq \|\hat{U}_{t_k} - \breve{U}_{t_k}\|\|\mathbf{x}\| \leq \|\hat{U}_{t_k} - \breve{U}_{t_k}\|_F\|\mathbf{x}\|. \tag{8}$$

As the above inequality holds for every $\mathbf{x}$,

$$\left| \|\mathbf{y} - \hat{U}_{t_k}\mathbf{x}_g^*\| - \|\mathbf{y} - \breve{U}_{t_k}\mathbf{x}_s^*\| \right| \leq \max_{\mathbf{x}_j \in \{\mathbf{x}_g^*, \mathbf{x}_s^*\}} \left| \|\mathbf{y} - \hat{U}_{t_k}\mathbf{x}_j\| - \|\mathbf{y} - \breve{U}_{t_k}\mathbf{x}_j\| \right| \leq \|\hat{U}_{t_k} - \breve{U}_{t_k}\|_F\|\mathbf{x}_j\|. \tag{9}$$

The last inequality follows from (8). From solution to least-squares problem, $\mathbf{x}_g^* = \hat{U}_{t_k}^\top\mathbf{y}$ (similarly, $\mathbf{x}_s^* = \breve{U}_{t_k}^\top\mathbf{y}$). Since the input vectors are normalized, we get, $\|\mathbf{x}_g^*\| = \|\hat{U}_{t_k}^\top\mathbf{y}\| \leq \|\mathbf{y}\| = 1$ (similarly, $\|\mathbf{x}_s^*\| \leq 1$). This implies $\|\mathbf{x}_j\| \leq 1$. Using this bound on $\|\mathbf{x}_j\|$ in (9) completes the proof. □

The theorem follows from Claims 11 and 12.

Theorem 13. *Let $Y_{1_{\text{good}}}, \dots, Y_{t_{\text{good}}}$ be a sequence of matrices with $Y_{[t]_{\text{good}}} = [Y_{1_{\text{good}}}|\dots|Y_{t_{\text{good}}}]$. Let $Y_{[t]_{\text{good}_k}} = \hat{U}_{t_k}\hat{\Sigma}_{t_k}\hat{V}_{t_k}^\top$ be the best rank-k approximation to $Y_{[t]_{\text{good}}}$. Let λ_i be the ith eigenvalue of $Y_{[t]_{\text{good}}}Y_{[t]_{\text{good}}}^\top$ and $L = \min_{i \neq j} |\lambda_i - \lambda_j| > 0$. Then for any unit vector $\mathbf{y} \in \mathbf{R}^m$, $\breve{U}_{t_k}$ (generated by the Algorithm* RandSketchUpdate*), under condition on ℓ from (7), with probability at least $1 - t \cdot 6e^{-99\ell}$, satisfies:*

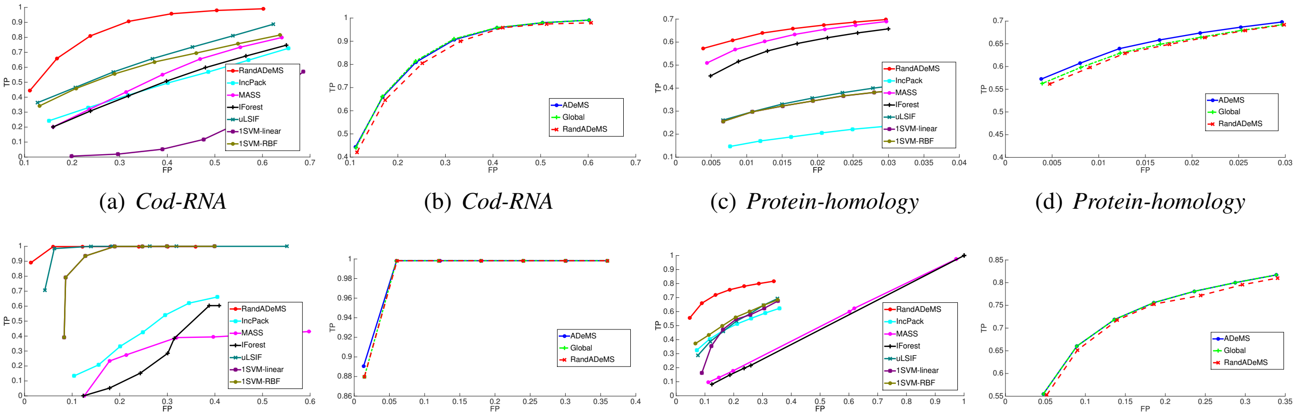
$$\left| \min_{\mathbf{x} \in \mathbf{R}^k} \|\mathbf{y} - \hat{U}_{t_k}\mathbf{x}\| - \min_{\mathbf{x} \in \mathbf{R}^k} \|\mathbf{y} - \breve{U}_{t_k}\mathbf{x}\| \right| \leq \sqrt{2}L/(\sqrt{L + 8\kappa^2\|Y_{[t]_{\text{good}}}\|^2}\sqrt[4]{L^2 + 16\kappa^4\|Y_{[t]_{\text{good}}}\|^4}).$$

The above bound on the difference in anomaly scores is an increasing function in L. Informally, the above theorem suggests that under some reasonable assumptions and settings of parameters, the anomaly scores in Algorithm AnomDetect obtained by using either Algorithms GlobalUpdate or RandSketchUpdate for singular value updation are "close". But as discussed before, Algorithm RandSketchUpdate is far more efficient both in space and time.

5 Experimental Analysis

In this section, we experimentally test our proposed approach in terms of effectiveness and efficiency. From now on, we refer to Algorithm AnomDetect with singular vectors updated using Algorithms GlobalUpdate, RandSketchUpdate, and SketchUpdate (presented in Appendix A) as **Global**, **RandADeMS**, and **ADeMS**, respectively.

Datasets. We use datasets drawn from a diverse set of domains to demonstrate the wide applicability of our anomaly detection approach (see Table 1). *Cod-RNA* dataset consists of sequenced genomes, and the task here is to detect *novel* non-coding RNAs (ncRNAs) [33], which are labeled as anomalies. *Protein-homology* dataset is from the protein homology prediction task of the KDD Cup 2004 [6], and the task here is to predict which proteins in the database are homologous to a native (query) sequence. Non-homologous sequences are labeled as anomalies.



(a) *Cod-RNA* (b) *Cod-RNA* (c) *Protein-homology* (d) *Protein-homology*

(e) *User-activity* (f) *User-activity* (g) *RCV1AD* (h) *RCV1AD*

Figure 1: ROC curves for compared approaches on various datasets.

The *User-activity* dataset is from an application of monitoring employee network activity log for an enterprise. The goal here is to identify malicious employee actions (anomalies) that result in loss of intellectual property for the enterprise. *RCV1* dataset consists of a corpus of newswire stories (documents), grouped into 103 categories [24]. In our evaluation, from these 103 categories, we used documents belonging to the 30 largest categories and documents belonging to the smallest 5 categories (labeled as anomalies). For features, we use a vocabulary of 1000 terms selected based on frequency. We refer to this modified *RCV1* dataset as *RCV1AD*.

Baselines. There are plenty of approaches for anomaly detection (see also the discussion in Section 2). We compare against six popular algorithms for anomaly detection. These algorithms were chosen taking into account their scalability on large datasets.

1SVM-linear and **1SVM-RBF** are one-class support vector machine classifiers with linear/radial-basis as kernel function. The output probability value of belonging to the anomalous class is treated as the anomaly score. We also compare against **IForest** [27], **Mass** [32], and Unconstrained Least-Squares Importance Fitting (**uLSIF**) [17] algorithms, which are all described in Section 2. As another streaming competitor, we implemented the incremental-PCA-based anomaly detection [3] to update the singular vectors in Algorithm ANOMDETECT. We call the resulting algorithm **IncPack**.

Parameter Settings. Except for IForest and Mass, all other competitors, including our proposed approach RANDADEMS, require a training set to bootstrap the process, and the training samples are required to be free of anomalies. We set the size of the training set as 2000, and we draw these training samples randomly from the set of non-anomalous datapoints. Note that the training set size is much smaller compared to the actual dataset size. We also observed that our results are stable to variations in the training set (see Section 5.1).

After training, for RANDADEMS, the number of data points given as input at each timestep is set to 5000, and $k = m/5$. The same parameter setting is also used for the ADEMS and IncPack streaming experiments. As mentioned in Section 4.3, we set $\ell = \sqrt{m}$. All other algorithms (1SVM-linear, 1SVM-RBF, IForest, Mass, and uLSIF) are considered to receive all the samples at once. The relevant parameters of these algorithms were tuned to obtain the best possible result. For evaluation, we do not use the threshold parameter ζ (as described in Algorithm ANOMDETECT), instead we assume that the percentage of anomalies among all samples is known *a priori* (say, this is $p\%$). At each timestep of the streaming process, we collect the anomaly scores of all the processed samples, and mark those datapoints as anomalies which occupy the top $p\%$ of these scores. All the experiments were run on a machine with Intel Xeon 2.67GHz processor and 124GB memory.

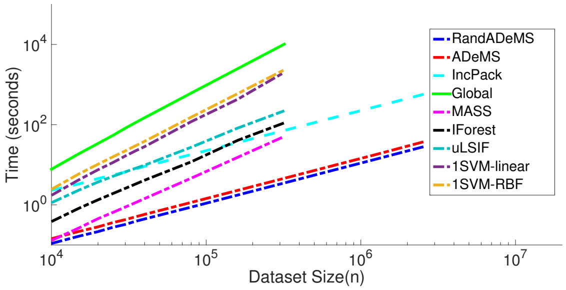


Figure 2: Scalability of various algorithms with increasing number of datapoints. The datasets for this test were created by down- and up-sampling the Cod-RNA dataset.

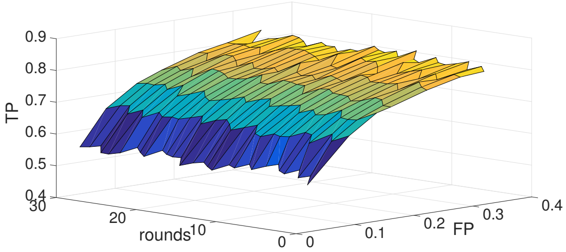


Figure 3: Stability of RANDADEMS under concept drift and different training set initializations.

Comparison between Different Algorithms. We use the standard evaluation metrics of True Positive rate (TP) and False Positive rate (FP). Figure 1 plots the ROC curve of the selected anomaly detection algorithms. To generate these curves, we use seven different threshold (ζ) numbers (selected based on the number of anomalies in each of the datasets). Each point represents the average (TP and FP) of a 30-fold cross-validation result, each time the training set is randomly selected from the normal samples and the order of samples is also randomly shuffled.

It is evident that RANDADEMS outperforms the other algorithms on all the datasets (except for the experiments on *User-activity* dataset, Figure 1(e), which shows a partial overlap between RANDADEMS, 1SVM-RBF, and uLSIF). Note that the performance of RANDADEMS is good, both when the fraction of anomalies is very high (such as in the *Cod-RNA* dataset, Figure 1(a)) or very small (such as in the *Protein-homology* dataset, Figure 1(c)). Due to the updates to the singular vectors, RANDADEMS can successfully deal with the concept shift problem in the normal data, i.e., new patterns of normal data appearing over time (more details in Section 5.1).

RANDADEMS has extremely similar performance to GLOBAL (Figures 1(b), 1(d), 1(f), and 1(h)). It confirms our theoretical analysis that the Algorithm RANDSKETCHUPDATE gives a desired approximation to Algorithm GLOBALUPDATE. Also these results suggest that using a randomized low-rank SVD (RANDADEMS) instead of the exact low-rank SVD (ADEMS) has little effect on the anomaly detection performance.

Figure 2 shows the scalability comparison (training + testing time) between the compared approaches. Among all the streaming competitors, RANDADEMS is 30% faster than ADEMS, and about 20 times faster than IncPack. Compared with other competitors, RANDADEMS is also even faster than the efficient IForest and Mass algorithms, and more than 20 times faster than the other methods. In particular, RANDADEMS and ADEMS, run in matter of couple of minutes, even when the dataset size increases to over 1 million.

5.1 Stability under Concept Drift and Training Set Initializations

We compare the performance of RANDADEMS in the order of timestamps in RCV1AD, with 30 different training sets randomly drawn from the set of non-anomalous datapoints. Each training set has 2000 samples. The ROC results of all the 30 test are plotted in Figure 3. For the test the size of each stream batch is set to be 1200, so that it corresponds to newswire stories arriving every 4 to 5 days on average. Therefore, as time goes on different topics arrive/fade in our experiment (concept drift). The points used for curves in Figure 3 are within 5% of the corresponding points in the (averaged) curve plotted in Figure 1(h). This demonstrates the stability of RANDADEMS under concept drift and different training set initializations.

6 Conclusion

We proposed a novel randomized sketching-based approach to efficiently and effectively detect anomalies in large data streams. The resulting algorithm consumes limited memory and requires just one pass over the data. Our theoretical results show this algorithm performs comparably with a global (batch) approach while being significantly faster. Empirical evaluation on a variety of datasets illustrate the effectiveness of the proposed approach.

Acknowledgments

We thank Abhishek Srivatsav and James Jobin for helpful initial discussions.

References

[1] C. Aggarwal. *Outlier Analysis*. Springer, 2013.

[2] Tarem Ahmed, Mark Coates, and Anukool Lakhina. Multivariate online anomaly detection using kernel recursive least squares. In *INFOCOM*, pages 625–633. IEEE, 2007.

[3] C. G. Baker, K. A. Gallivan, and P. Van Dooren. Low-rank incremental methods for computing dominant singular subspaces. *Linear Algebra and its Applications*, 2012.

[4] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander. Lof: Identifying density-based local outliers. *SIGMOD*, 29(2):93–104, 2000.

[5] P. Businger. Updating a singular value decomposition. *BIT*, 10(3):376–385, 1970.

[6] R. Caruana, T. Joachims, and L. Backstrom. Kdd-cup 2004: results and analysis. *JMLR*, 6(2):95–108, 2004.

[7] Y. Chahlaoui, K. Gallivan, and P. Van Dooren. Recursive calculation of dominant singular subspaces. *SIAM Journal on Matrix Analysis and Applications*, 25(2):445–463, 2003.

[8] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–72, 2009.

[9] X. Chen, W. Li, and W. Xu. Perturbation analysis of the eigenvector matrix and singular vector matrices. *Taiwanese Journal of Mathematics*, 16(1):pp–179, 2012.

[10] Qi Ding and Eric D Kolaczyk. A compressed pca subspace method for anomaly detection in high-dimensional data. *Information Theory, IEEE Transactions on*, 59(11):7419–7433, 2013.

[11] Moshe Gabel, Assaf Schuster, and Daniel Keren. Communication-efficient distributed variance monitoring and outlier detection for multivariate time series. In *IPDPS*, 2014.

[12] M. Ghashami and J. M. Phillips. Relative errors for deterministic low-rank matrix approximations. In *SODA*, pages 707–717, 2014.

[13] Mina Ghashami, Edo Liberty, Jeff M. Phillips, and David P. Woodruff. Frequent directions : Simple and deterministic matrix sketching. *CoRR*, abs/1501.01711, 2015.

[14] G. H. Golub and C. F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[15] N. Halko, P. G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[16] P. M Hall, A. D. Marshall, and R. R. Martin. Incremental eigenanalysis for classification. In *BMVC*, volume 98, pages 286–295, 1998.

[17] S. Hido, Y. Tsuboi, H. Kashima, M. Sugiyama, and T. Kanamori. Statistical outlier detection using direct density ratio estimation. *KAIS*, 26(2):309–336, 2011.

[18] Hao Huang, Hong Qin, Shinjae Yoo, and Dantong Yu. Physics-based anomaly detection defined on manifold space. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 9(2):14, 2014.

[19] Ling Huang, XuanLong Nguyen, Minos Garofalakis, Joseph M Hellerstein, Michael I Jordan, Anthony D Joseph, and Nina Taft. Communication-efficient online detection of network-wide anomalies. In *INFOCOM*, pages 134–142. IEEE, 2007.

[20] Ling Huang, XuanLong Nguyen, Minos Garofalakis, Michael I Jordan, Anthony Joseph, and Nina Taft. In-network pca and anomaly detection. In *NIPS*, pages 617–624, 2006.

[21] S. Kasiviswanathan, H. Wang, A. Banerjee, and P. Melville. Online l_1-dictionary learning with application to novel document detection. *NIPS*, pages 2258–2266, 2012.

[22] Anukool Lakhina, Mark Crovella, and Christiphe Diot. Characterization of network-wide anomalies in traffic flows. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 201–206. ACM, 2004.

[23] A Levey and M. Lindenbaum. Sequential karhunen-loeve basis extraction and its application to images. *IEEE TIP*, 9(8):1371–1374, 2000.

[24] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *ACM SIGKDD Explorations Newsletter*, 5:361–397, 2004.

[25] Xin Li, Fang Bian, Mark Crovella, Christophe Diot, Ramesh Govindan, Gianluca Iannaccone, and Anukool Lakhina. Detection and identification of network anomalies using sketch subspaces. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, pages 147–152. ACM, 2006.

[26] E. Liberty. Simple and deterministic matrix sketching. In *ACM SIGKDD*, pages 581–588, 2013.

[27] F. T. Liu, K. M. Ting, and Z. H. Zhou. Isolation forest. *IEEE ICDM*, pages 413–422, 2008.

[28] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *JMLR*, 11:19–60, 2010.

[29] M. Markou and S. Singh. Novelty detection: a review–part 1: statistical approaches. *Signal processing*, 83(12):2481–2497, 2003.

[30] J. Misra and D. Gries. Finding repeated elements. *Science of computer programming*, 2(2):143–152, 1982.

[31] S. Papadimitriou, J. Sun, and C. Faloutsos. Streaming pattern discovery in multiple time-series. In *VLDB*, 2005.

[32] K. M. Ting, G. T. Zhou, F. T. Liu, and J. S. Tan. Mass estimation and its applications. *ACM SIGKDD*, 2010.

[33] A. V. Uzilov, J. M. Keegan, and D. H. Mathews. Detection of non-coding rnas on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics*, 2006.

Algorithm 4: SKETCHUPDATE (streaming update of the singular vectors at time t)

```
    **Input:** $Y_{t_{\text{good}}} \in \mathbf{R}^{m \times n_t}$ and $B_{t-1} \in \mathbf{R}^{m \times \ell}$
    $D_t \leftarrow [B_{t-1} | Y_{t_{\text{good}}}]$
    $\tilde{U}_{t_\ell} \tilde{\Sigma}_{t_\ell} \tilde{V}_{t_\ell}^\top \leftarrow \textsc{Svd}_\ell(D_t)$   (with $\tilde{\Sigma}_{t_\ell} = diag(\tilde{\sigma}_{t_1}, \dots, \tilde{\sigma}_{t_\ell})$)
    $\tilde{\Sigma}_{t_\ell}^{(trunc)} \leftarrow diag\left(\sqrt{\tilde{\sigma}_{t_1}^2 - \tilde{\sigma}_{t_\ell}^2}, \sqrt{\tilde{\sigma}_{t_2}^2 - \tilde{\sigma}_{t_\ell}^2}, \dots, \sqrt{\tilde{\sigma}_{t_{\ell-1}}^2 - \tilde{\sigma}_{t_\ell}^2}, 0\right)$
    $B_t \leftarrow \tilde{U}_{t_\ell} \tilde{\Sigma}_{t_\ell}^{(trunc)}$
    **Return:** $B_t$ and $\tilde{U}_{t_k}$
```

A Deterministic Matrix Sketching

As mentioned in Section 4.2, at each timestep t, instead of using a randomized low-rank matrix approximation, we could also compute an actual low-rank SVD. The resulting deterministic approach for singular value updation is presented in Algorithm SKETCHUPDATE. At timestep t, Algorithm SKETCHUPDATE takes $O(mn_t\ell)$ time (assuming $\ell \leq n_t$) by using power-iteration or rank-revealing QR decomposition for SVD of D_t [14]. Note that this running time is bigger than that of Algorithm RANDSKETCHUPDATE, which at timestep t takes $O(\ell T_{\text{mult}} + (m + n_t)\ell^2)$ time. Algorithm SKETCHUPDATE has a memory overhead of $O(m\ell)$.

A.1 Analysis of Algorithm SKETCHUPDATE

The analysis is similar to that of Algorithm RANDSKETCHUPDATE. Our first aim will be to bound the Frobenius norm of the difference between $Y_{[t]_{\text{good}_k}} Y_{[t]_{\text{good}_k}}^\top$ and $B_{t_k} B_{t_k}^\top$. We start off by defining some additional notation:

$$\begin{aligned} B_{t_k} &= \tilde{U}_{t_k} \tilde{\Sigma}_{t_k}^{(trunc)} \text{ (rank-}k\text{ approximation of } B_t\text{)}, \\ \Delta_t &= \textstyle\sum_{j=1}^t \tilde{\sigma}_{j_\ell}^2, \\ \tilde{U}_t \tilde{\Sigma}_t \tilde{V}_t^\top &= \textsc{Svd}(D_t), \\ \kappa &= \kappa_k(Y_{[t]_{\text{good}}}). \end{aligned}$$

Lemma 14 (Modified from Ghashami *et al.* [12])**.** *At timestep t, Algorithm* SKETCHUPDATE *maintains that* $\|Y_{[t]_{\text{good}}}\|_F^2 - \|B_t\|_F^2 \geq \ell \Delta_t$.

Proof. At timestep t, $\|D_t\|_F^2 = \|B_{t-1}\|_F^2 + \|Y_{t_{\text{good}}}\|_F^2$ and $\|D_t\|_F^2 \geq \|B_t\|_F^2 + \ell \tilde{\sigma}_{t_\ell}^2$. Solving for $\|Y_{t_{\text{good}}}\|_F^2$ and summing over all $j \leq t$, we get

$$\|Y_{[t]_{\text{good}}}\|_F^2 = \sum_{j=1}^t \|Y_{j_{\text{good}}}\|_F^2 \geq \sum_{j=1}^t \|B_j\|_F^2 - \|B_{j-1}\|_F^2 + \ell \tilde{\sigma}_{j_\ell}^2 = \|B_t\|_F^2 - \|B_0\|_F^2 + \ell \Delta_t.$$

By setting B_0 as all zeros matrix, we get the result. □

The following lemma shows that for any direction $\mathbf{x}$, $Y_{[t]_{\text{good}}}$ and B_t are not too far apart.

Lemma 15 (Modified from Ghashami *et al.* [12])**.** *For any unit vector* $\mathbf{x} \in \mathbf{R}^m$*, at any timestep t,* $0 \leq \|Y_{[t]_{\text{good}}}^\top \mathbf{x}\|^2 - \|B_t^\top \mathbf{x}\|^2 \leq \Delta_t$.

Proof. To show $|Y_{[t]_{\text{good}}}^\top \mathbf{x}\|^2 - \|B_t^\top \mathbf{x}\|^2 > 0$, observe that $\|B_{t-1}^\top \mathbf{x}\|^2 + \|Y_{t_{\text{good}}}^\top \mathbf{x}\|^2 = \|D_t^\top \mathbf{x}\|^2 \geq \|B_t^\top \mathbf{x}\|^2$. We have

$$\|Y_{[t]_{\text{good}}}^\top \mathbf{x}\|^2 = \sum_{j=1}^t \|Y_{j_{\text{good}}}^\top \mathbf{x}\|^2 \geq \sum_{j=1}^t \|B_j^\top \mathbf{x}\|^2 - \|B_{j-1}^\top \mathbf{x}\|^2 = \|B_t^\top \mathbf{x}\|^2 \geq 0.$$

Here we used that B_0 is an all zeros matrix. Now let us concentrate on showing

$$\|Y_{[t]_{good}}^{\top}\mathbf{x}\|^2 - \|B_t^{\top}\mathbf{x}\|^2 \leq \Delta_t.$$

Let $\mathbf{u}_i$ be the ith column in $\tilde{U}_t$. $\tilde{\sigma}_{t_i}^2 - \tilde{\sigma}_{t_\ell}^2$ is the ith singular value of B_t. Let $r_d = \mathrm{r}ank(D_t)$.

$$\begin{aligned}
\|D_t^{\top}\mathbf{x}\|^2 &= \sum_{i=1}^{r_d} \tilde{\sigma}_{t_i}^2 \langle \mathbf{u}_i, \mathbf{x}\rangle^2 = \sum_{i=1}^{r_d} (\tilde{\sigma}_{t_i}^2 + \tilde{\sigma}_{t_\ell}^2 - \tilde{\sigma}_{t_\ell}^2)\langle \mathbf{u}_i, \mathbf{x}\rangle^2 \\
&= \sum_{i=1}^{r_d} (\tilde{\sigma}_{t_i}^2 - \tilde{\sigma}_{t_\ell}^2)\langle \mathbf{u}_i, \mathbf{x}\rangle^2 + \sum_{i=1}^{r_d} \tilde{\sigma}_{t_\ell}^2 \langle \mathbf{u}_i, \mathbf{x}\rangle^2 \\
&\leq \sum_{i=1}^{\ell} (\tilde{\sigma}_{t_i}^2 - \tilde{\sigma}_{t_\ell}^2)\langle \mathbf{u}_i, \mathbf{x}\rangle^2 + \tilde{\sigma}_{t_\ell}^2 \sum_{i=1}^{r_d} \langle \mathbf{u}_i, \mathbf{x}\rangle^2 \leq \|B_t^{\top}\mathbf{x}\|^2 + \tilde{\sigma}_{t_\ell}^2.
\end{aligned}$$

For the first inequality we used that for $i > \ell$, $\tilde{\sigma}_{t_i}^2 \leq \tilde{\sigma}_{t_\ell}^2$. For the second inequality, we use that $\sum_{i=1}^{r_d} \langle \mathbf{u}_i, \mathbf{x}\rangle^2 \leq \|\mathbf{x}\|^2 = 1$ (as $\mathbf{x}$ is a unit vector). Since $\|D_t^{\top}\mathbf{x}\|^2 = \|B_{t-1}^{\top}\mathbf{x}\|^2 + \|Y_{t_{good}}^{\top}\mathbf{x}\|^2$. Using this along with the above established inequality, $|D_t^{\top}\mathbf{x}\|^2 - \|B_t^{\top}\mathbf{x}\|^2 \leq \tilde{\sigma}_{t_\ell}^2$, gives

$$\|B_{t-1}^{\top}\mathbf{x}\|^2 + \|Y_{t_{good}}^{\top}\mathbf{x}\|^2 \leq \|B_t^{\top}\mathbf{x}\|^2 + \tilde{\sigma}_{t_\ell}^2.$$

Subtracting $\|B_{t-1}^{\top}\mathbf{x}\|^2$ from both sides and summing over $j \leq t$,

$$\|Y_{[t]_{good}}^{\top}\mathbf{x}\|^2 = \sum_{j=1}^{t} \|Y_{j_{good}}^{\top}\mathbf{x}\|^2 \leq \sum_{j=1}^{t} (\|B_j^{\top}\mathbf{x}\|^2 - \|B_{j-1}^{\top}\mathbf{x}\|^2 + \tilde{\sigma}_{j_\ell}^2) = \|B_t^{\top}\mathbf{x}\|^2 + \Delta_t.$$

Again we used that B_0 is an all zeros matrix. □

Since for all unit vectors $\mathbf{x} \in \mathbf{R}^m$,

$$\|Y_{[t]_{good}}^{\top}\mathbf{x}\|^2 - \|B_t^{\top}\mathbf{x}\|^2 \geq 0 \implies Y_{[t]_{good}} Y_{[t]_{good}}^{\top} \succeq B_t B_t^{\top}.$$

From Claim 2, we have for all vectors $\mathbf{x} \in \mathbf{R}^m$, $\kappa\|Y_{[t]_{good_k}}^{\top}\mathbf{x}\| \geq \|Y_{[t]_{good}}^{\top}\mathbf{x}\|$. Therefore,

$$\kappa^2 Y_{[t]_{good_k}} Y_{[t]_{good_k}}^{\top} \succeq Y_{[t]_{good}} Y_{[t]_{good}}^{\top} \succeq B_t B_t^{\top} \succeq B_{t_k} B_{t_k}^{\top}.$$

The following crucial lemma lower bounds the Frobenius norm between $Y_{[t]_{good}}$ and $Y_{[t]_{good_k}}$ (an upper bound for the same follows from Theorem 1).

Lemma 16 (Modified from Ghashami *et al.* [12]). *Let $Y_{[t]_{good_k}}$ be a rank-k approximation to $Y_{[t]_{good}}$ (as in Theorem 1). Then, $\Delta_t \leq \|Y_{[t]_{good}} - Y_{[t]_{good_k}}\|_F^2/(\ell - k)$.*

Proof. From Lemma 14, $\|Y_{[t]_{good}}\|_F^2 - \|B_t\|_F^2 \geq \ell\Delta_t$. Let $r = \mathrm{r}ank(Y_{[t]_{good}})$ and $\mathbf{v}_1, \ldots, \mathbf{v}_r$ be the left singular vectors of $Y_{[t]_{\mathrm{good}}}$ corresponding to the non-zero singular values of $Y_{[t]_{\mathrm{good}}}$, we have

$$\begin{aligned}
\ell\Delta_t &\le \|Y_{[t]_{\mathrm{good}}}\|_F^2 - \|B_t\|_F^2 \\
&= \sum_{i=1}^{k} \|Y_{[t]_{\mathrm{good}}}^\top \mathbf{v}_i\|^2 + \sum_{i=k+1}^{r} \|Y_{[t]_{\mathrm{good}}}^\top \mathbf{v}_i\|^2 - \|B_t\|_F^2 \\
&= \sum_{i=1}^{k} \|Y_{[t]_{\mathrm{good}}}^\top \mathbf{v}_i\|^2 + \|Y_{[t]_{\mathrm{good}}} - Y_{[t]_{\mathrm{good}_k}}\|_F^2 - \|B_t\|_F^2 \\
&\le \sum_{i=1}^{k} \|Y_{[t]_{\mathrm{good}}}^\top \mathbf{v}_i\|^2 + \|Y_{[t]_{\mathrm{good}}} - Y_{[t]_{\mathrm{good}_k}}\|_F^2 - \sum_{i=1}^{k} \|B_t^\top \mathbf{v}_i\|^2 \\
&= \|Y_{[t]_{\mathrm{good}}} - Y_{[t]_{\mathrm{good}_k}}\|_F^2 + \sum_{i=1}^{k}(\|Y_{[t]_{\mathrm{good}}}^\top \mathbf{v}_i\|^2 - \|B_t^\top \mathbf{v}_i\|^2) \le \|Y_{[t]_{\mathrm{good}}} - Y_{[t]_{\mathrm{good}_k}}\|_F^2 + k\Delta_t.
\end{aligned}$$

First inequality uses that $\sum_{i=1}^{k} \|B_t^\top \mathbf{v}_i\|^2 \le \|B_t\|_F^2$, and the last inequality is based on Lemma 15. Solving for Δ_t in the above inequality gives the claimed result. □

Now using Lemma 16, we can relate $\|Y_{[t]_{\mathrm{good}_k}}\|_F^2$ to $\|B_{t_k}|\|_F^2$.

Lemma 17 (Modified from Ghashami *et al.* [12])**.**

$$0 \le \|Y_{[t]_{\mathrm{good}_k}}\|_F^2 - \|B_{t_k}\|_F^2 \le \frac{k}{\ell - k} \cdot \|Y_{[t]_{\mathrm{good}}} - Y_{[t]_{\mathrm{good}_k}}\|_F^2.$$

Proof. As in Lemma 16, let $\mathbf{v}_1, \dots, \mathbf{v}_k$ be the left singular vectors of $Y_{[t]_{\mathrm{good}}}$ corresponding to its top-k singular values. Let $\mathbf{u}_1, \dots, \mathbf{u}_k$ be the left singular vectors of B_t corresponding to its top-k singular values. We have

$$\|Y_{[t]_{\mathrm{good}_k}}\|_F^2 = \sum_{i=1}^{k} \|Y_{[t]_{\mathrm{good}}}^\top \mathbf{v}_i\|^2 \ge \sum_{i=1}^{k} \|Y_{[t]_{\mathrm{good}}}^\top \mathbf{u}_i\|^2 \ge \sum_{i=1}^{k} \|B_t^\top \mathbf{u}_i\|^2 = \|B_{t_k}\|_F^2.$$

The second inequality ($\|Y_{[t]_{\mathrm{good}}}^\top \mathbf{u}_i\|^2 \ge \|B_t^\top \mathbf{u}_i\|^2$) follows from Lemma 15. This proves that $0 \le \|Y_{[t]_{\mathrm{good}_k}}\|_F^2 - \|B_{t_k}\|_F^2$. The upper bound can be established as follows.

$$\|B_{t_k}\|_F^2 \ge \sum_{i=1}^{k} \|B_{t_k}^\top \mathbf{v}_i\|^2 \ge \sum_{i=1}^{k}(\|Y_{[t]_{\mathrm{good}}}^\top \mathbf{v}_i\|^2 - \Delta_t) = \|Y_{[t]_{\mathrm{good}_k}}\|_F^2 - k\Delta_t,$$

where the second inequality follows from Lemma 15. Now substituting for Δ_t from Lemma 16 gives the result. □

Using this above lemma and the fact that $\kappa^2 Y_{[t]_{\mathrm{good}_k}} Y_{[t]_{\mathrm{good}_k}}^\top \succeq B_{t_k} B_{t_k}^\top$, we can derive a bound on $\|\kappa^2 Y_{[t]_{\mathrm{good}_k}} Y_{[t]_{\mathrm{good}_k}}^\top - B_{t_k} B_{t_k}^\top\|_F$ as in Proposition 8.

Proposition 18. *At timestep t, B_{t_k} generated by Algorithm* SKETCHUPDATE *satisfies,*

$$\|\kappa^2 Y_{[t]_{\mathrm{good}_k}} Y_{[t]_{\mathrm{good}_k}}^\top - B_{t_k} B_{t_k}^\top\|_F \le \kappa^2 \|Y_{[t]_{\mathrm{good}_k}}\|_F^2 - \|B_{t_k}\|_F^2.$$

We need couple more definitions.

1. Define Γ_a (which plays the same role as Φ_a defined in Section 4.3) as,

$$\Gamma_a = \frac{\kappa^2 \|Y_{[t]_{(k)}}\|_F^2 - \|B_{t_{(k)}}\|_F^2}{\|Y_{[t]_{(k)}}\|_F^2 - \|B_{t_{(k)}}\|_F^2}. \tag{10}$$

2. Define Γ_b (which plays the same role as Φ_b defined in Section 4.3) as,

$$\Gamma_b = 1 + \frac{2}{\kappa^2 - \|B_t\|^2/\|Y_{[t]}\|^2}. \tag{11}$$

The proof of following claim follows as Claim 9.

Claim 19. $\Gamma_b \leq 1 + \frac{2}{\kappa^2 - \|B_t\|^2/\|Y_{[t]_{\text{good}}}\|^2}$.

We now apply Proposition 18 and Theorem 10 to bound $\|\hat{U}_{t_k} - \tilde{U}_{t_k}\|_F$. To do so, we construct matrices

$$A = \kappa^2 Y_{[t]_{\text{good}}} Y_{[t]_{\text{good}}}^\top \quad \text{and} \quad A_{\text{per}} = B_t B_t^\top.$$

The proof of following claim follows as Claim 11.

Claim 20. *Let λ_i denote the ith eigenvalue of $Y_{[t]_{\text{good}}} Y_{[t]_{\text{good}}}^\top$ and $L = \min_{i \neq j} |\lambda_i - \lambda_j| > 0$. If*

$$\ell = \Omega\left(\frac{\sqrt{m}\kappa^2 \|Y_{[t]_{\text{good}}}\|^2 \Gamma_a \Gamma_b k \|Y_{[t]_{\text{good}}} - Y_{[t]_{\text{good}_k}}\|_F^2}{L^2}\right),$$

for Γ_a, Γ_b defined in (10), (11) respectively, then

$$\|\hat{U}_{t_k} - \tilde{U}_{t_k}\|_F \leq \|\hat{U}_t - \tilde{U}_t\|_F \leq \frac{\sqrt{2}L}{\sqrt{L + 8\kappa^2 \|Y_{[t]_{\text{good}}}\|^2 \sqrt[4]{L^2 + 16\kappa^4 \|Y_{[t]_{\text{good}}}\|^4}}}.$$

Now using this result, we are ready to compare the anomaly scores generated by using either $\hat{U}_{t_k}$ or $\tilde{U}_{t_k}$ in Algorithm AnomDetect. The theorem follows from Claims 20 and 12.

Theorem 21. *Let $Y_{1_{\text{good}}}, \dots, Y_{t_{\text{good}}}$ be a sequence of matrices with $Y_{[t]_{\text{good}}} = [Y_{1_{\text{good}}} | \dots | Y_{t_{\text{good}}}]$. Let $Y_{[t]_{\text{good}_k}} = \hat{U}_{t_k} \hat{\Sigma}_{t_k} \hat{V}_{t_k}^\top$ be the best rank-k approximation to $Y_{[t]_{\text{good}}}$. Let λ_i be the ith eigenvalue of $Y_{[t]_{\text{good}}} Y_{[t]_{\text{good}}}^\top$ and $L = \min_{i \neq j} |\lambda_i - \lambda_j| > 0$. Then for any unit vector $\mathbf{y} \in \mathbf{R}^m$, $\tilde{U}_{t_k}$ (generated by the Algorithm* SketchUpdate*), under condition on ℓ from Claim 20, satisfies:*

$$\left| \min_{\mathbf{x} \in \mathbf{R}^k} \|\mathbf{y} - \hat{U}_{t_k} \mathbf{x}\| - \min_{\mathbf{x} \in \mathbf{R}^k} \|\mathbf{y} - \tilde{U}_{t_k} \mathbf{x}\| \right| \leq \frac{\sqrt{2}L}{\sqrt{L + 8\kappa^2 \|Y_{[t]_{\text{good}}}\|^2 \sqrt[4]{L^2 + 16\kappa^4 \|Y_{[t]_{\text{good}}}\|^4}}}.$$

The above theorem has an interpretation similar to that of Theorem 13. However, notice that compared to Algorithm RandSketchUpdate, the requirement on ℓ is slightly weaker in Algorithm SketchUpdate.[9] This is because Algorithm SketchUpdate computes the exact low-rank matrices at each timestep.

[9] In fact, for small k's, and assuming $1 < \kappa \leq O(1)$ (implying $\Gamma_a = O(1)$ and $\Gamma_b = O(1)$), the bound on ℓ in Claim 20 could be simplified to,

$$\ell = \Omega\left(\frac{\sqrt{m}\|Y_{[t]_{\text{good}}}\|^2 \|Y_{[t]_{\text{good}}} - Y_{[t]_{\text{good}_k}}\|_F^2}{L^2}\right).$$