4 Frequent Directions

Edo Liberty[3] discovered a strong connection between *"matrix sketching"* and *"frequent items"* problems.

In FREQUENTITEMS problem, we are given a stream $S = \langle s_1, s_2, \ldots, s_n \rangle$ of n items from domain $[u] = \{1, 2, \ldots, u\}$, and the goal is to find the frequency of each item $j \in [u]$ as $f_j = |\{s_i \in S \mid s_i = j\}|$. One obvious solution to this problem would be to maintain u counters, one for each member of domain, and upon receiving an item, we increment the corresponding counter. This solution gives exact frequencies but it is not applicable when domain $[u]$ is large, or stream is infinite. Therefore we seek an approximate solution that outputs estimates $\hat{f}_j$ (for *all* $j \in [u]$) so that $|f_j - \hat{f}_j| \leq \varepsilon n$.

One of approximate solutions to this problem is *Misra-Gries (MG)* [4] sketch, which uses $O(1/\varepsilon)$ space to keep $\ell - 1 = 1/\varepsilon$ counters, each labeled by a value $j \in [u]$: it increments a counter if the new item matches the associated label or for an empty counter, and it decrements all counters if there is no empty counter and none match the stream element. At any time, the approximated frequency of an item $j \in [u]$, i.e. $\hat{f}_j$, is the associated counter value, or 0 if there is no associated counter.

To see the connection between matrix sketching and frequent items problems, consider a matrix A as a stream of its rows. For now, let us constrain the rows of A to be indicator vectors. In other words, we have $A_{i,:} \in \{e_1, ..., e_d\}$, where e_j is the jth standard basis vector. Note that such a matrix can encode a stream of items (as above). If ith element in the stream is j, then the i-th row of the matrix is set to $A_{i,:} = e_j$. The frequency f_j can be expressed as $f_j = \|Ae_j\|^2$. Assume that we construct a matrix $B \in \mathbb{R}^{\ell \times d}$ as follows. First, we run FREQUENTITEMS on the input. Then, for every item j for which $\hat{f}_j > 0$ we generate one row in B equal to $\hat{f}_j^{1/2} \cdot e_j$. The result is a low rank approximation of A. Note that $\mathsf{rank}(B) = \ell$ and that $\|Be_j\|^2 = \hat{f}_j$. Also notice that $\|A\|_F^2 = n$ and that $A^T A = \mathsf{diag}(f_1, \ldots, f_d)$ and that $B^T B = \mathsf{diag}(\hat{f}_1, \ldots, \hat{f}_d)$. Porting the results we obtained from FREQUENTITEMS, we get that $\|A^T A - B^T B\|_2 = \max_j |f_j - \hat{f}_j| \leq \|A\|_F^2/(\ell - k)$.

This connection is the idea behind FREQUENTDIRECTIONS, below we describe this algorithm in detail.

4.1 Frequent Directions

FREQUENTDIRECTIONS algorithm, described in 4.1.1, receives and processes the input matrix $A \in \mathbb{R}^{n \times d}$ in streaming fasion, row by row. It maintains a $\ell \times d$ sketch matrix B that is updated every time a new row from A is received. It maintains the invariant that the last row of sketch B is always all-zero valued. During the execution of the algorithm, rows from A simply replace the all-zero valued row in B. Then, the last row is nullified by a two-stage process. First, the sketch is rotated (from left) using its SVD such that its rows are orthogonal and in descending magnitude order. Then, the norm of sketch rows are "shrunk" so that at least one of them is set to zero.

Setting $\ell = k + 1/\varepsilon$, FREQUENTDIRECTIONS achieves the spectral error bound[1]

$$\forall x \in \mathbb{R}^d, \|x\| = 1 \quad 0 \leq \|Ax\|^2 - \|Bx\|^2 \leq \varepsilon \|A - A_k\|_F^2 \tag{4.1}$$

And using $\ell = \lceil k + k/\varepsilon \rceil$, FREQUENTDIRECTIONS obtains Frobenius error bound[1]

$$\|A - \pi_B^k(A)\|_F \leq (1 + \varepsilon)\|A - A_k\|_F \tag{4.2}$$

Where $\pi_B^k(A)$ is the projection of rows of A onto space spanned by top k right singular vectors of B.

Algorithm 4.1.1 FREQUENTDIRECTIONS

Input: $\ell, A \in R^{n \times d}$
$B \leftarrow 0^{\ell \times d}$
for $i \in 1, \ldots, n$ **do**
  $B_{\ell,:} \leftarrow A_{i,:}$ # ith row of A replaced (all-zeros) ℓth row of B
  $[U, \Sigma, V] \leftarrow \mathsf{svd}(B)$
  $C \leftarrow \Sigma V^T$ # Not computed, only needed for proof notation
  $\delta \leftarrow \Sigma^2_{\ell,\ell}$
  $B \leftarrow \sqrt{\Sigma^2 - \delta I_\ell} \cdot V^T$ # The last row of B is again zero
return B

4.1.1 Error Analysis

In this section, we first show FREQUENTDIRECTIONS satisfies three properties, and then we prove *any algorithm* that produces an approximate matrix B which satisfies the properties (for any choice of Δ) will achieve the error bounds 4.1 and 4.2.

In what follows, we denote by δ_i, $B_{[i]}$, $C_{[i]}$ the values of δ, B and C respectively *after* the ith row of A was processed. Let $\Delta = \sum_{i=1}^n \delta_i$, be the total mass we subtract from the stream during the algorithm. We first prove FREQUENTDIRECTIONS satisfies three auxiliary properties[1].

Property 1. *For any vector x we have $\|Ax\|^2 - \|Bx\|^2 \geq 0$.*

Proof. Use the observations that $\langle A_{i,:}, x\rangle^2 + \|B_{[i-1]}x\|^2 = \|C_{[i]}x\|^2$.

$$\|Ax\|^2 - \|Bx\|^2 = \sum_{i=1}^n [\langle A_{i,:}, x\rangle^2 + \|B_{[i-1]}x\|^2 - \|B_{[i]}x\|^2] \geq \sum_{i=1}^n [\|C_{[i]}x\|^2 - \|B_{[i]}x\|^2] \geq 0. \qquad \square$$

Property 2. *For any unit vector $x \in \mathbb{R}^d$ we have $\|Ax\|^2 - \|Bx\|^2 \leq \Delta$.*

Proof. To see this, first note that $\|C_{[i]}x\|^2 - \|B_{[i]}x\|^2 \leq \|C_{[i]}^T C_{[i]} - B_{[i]}^T B_{[i]}\| \leq \delta_i$. Now, consider the fact that $\|C_{[i]}x\|^2 = \|B_{[i-1]}x\|^2 + \|A_{i,:}x\|^2$. Substituting for $\|C_{[i]}x\|^2$ above and taking the sum yields

$$\begin{aligned}\sum_i \|C_{[i]}x\|^2 - \|B_{[i]}x\|^2 &= \sum_i (\|B_{[i-1]}x\|^2 + \|A_{i,:}x\|^2) - \|B_{[i]}x\|^2 \\ &= \|Ax\|^2 + \|B_{[0]}x\|^2 - \|B_{[n]}x\|^2 = \|Ax\|^2 - \|Bx\|^2.\end{aligned}$$

Combining this with $\sum_i \|C_{[i]}x\|^2 - \|B_{[i]}x\|^2 \leq \sum_i \delta_i = \Delta$ yields that $\|Ax\|^2 - \|Bx\|^2 \leq \Delta$. $\square$

Property 3. $\Delta\ell \leq \|A\|_F^2 - \|B\|_F^2$.

Proof. In the ith round of the algorithm $\|C_{[i]}\|_F^2 \geq \|B_{[i]}\|_F^2 + \ell\delta_i$ and $\|C_{[i]}\|_F^2 = \|B_{[i-1]}\|_F^2 + \|A_{i,:}\|^2$. By solving for $\|A_{i,:}\|^2$ and summing over i we get

$$\|A\|_F^2 = \sum_{i=1}^n \|A_{i,:}\|^2 \leq \sum_{i=1}^n \|B_{[i]}\|_F^2 - \|B_{[i-1]}\|_F^2 + \ell\delta_i = \|B\|_F^2 + \ell\Delta. \qquad \square$$

Additive Error Bound.

Now equipped with the above observations, and *no additional requirements about the construction of* B, we can prove FREQUENTDIRECTIONS achieves bound 4.1. We use Property 2 verbatim and bootstrap Property 3 to prove a tighter bound on Δ. In the following, y_is correspond to singular vectors of A ordered with respect to a decreasing corresponding singular values.

$$\begin{aligned}
\Delta\ell &\leq \|A\|_F^2 - \|B\|_F^2 && \text{via Property 3} \\
&= \sum_{i=1}^{k} \|Ay_i\|^2 + \sum_{i=k+1}^{d} \|Ay_i\|^2 - \|B\|_F^2 && \|A\|_F^2 = \sum_{i=1}^{d} \|Ay_i\|^2 \\
&= \sum_{i=1}^{k} \|Ay_i\|^2 + \|A - A_k\|_F^2 - \|B\|_F^2 \\
&\leq \|A - A_k\|_F^2 + \sum_{i=1}^{k} \left(\|Ay_i\|^2 - \|By_i\|^2\right) && \sum_{i=1}^{k} \|By_i\|^2 < \|B\|_F^2 \\
&\leq \|A - A_k\|_F^2 + k\Delta. && \text{via Property 2}
\end{aligned}$$

Solving $\Delta\ell \leq \|A - A_k\|_F^2 + k\Delta$ for Δ to obtain $\Delta \leq \|A - A_k\|_F^2/(\ell - k)$, which combined with Property 1 and Property 2 proves that for any unit vector x we have

$$0 \leq \|Ax\|^2 - \|Bx\|^2 \leq \Delta \leq \|A - A_k\|_F^2/(\ell - k)$$

Setting $\ell = k + 1/\varepsilon$ completes the proof of additive error bound.

Relative Error Bound.

Now we can show that projecting A onto B_k provides a relative error approximation[1]. Here, y_is correspond to the singular vectors of A as above and v_is correspond to the singular vectors of B in a similar fashion.

$$\begin{aligned}
\|A - \pi_B^k(A)\|_F^2 = \|A\|_F^2 - \|\pi_B^k(A)\|_F^2 &= \|A\|_F^2 - \sum_{i=1}^{k} \|Av_i\|^2 && \text{Pythagorean theorem} \\
&\leq \|A\|_F^2 - \sum_{i=1}^{k} \|Bv_i\|^2 && \text{via Property 1} \\
&\leq \|A\|_F^2 - \sum_{i=1}^{k} \|By_i\|^2 && \text{since } \sum_{i=1}^{j} \|Bv_i\|^2 \geq \sum_{i=1}^{j} \|By_i\|^2 \\
&\leq \|A\|_F^2 - \sum_{i=1}^{k} (\|Ay_i\|^2 - \Delta) && \text{via Property 2} \\
&= \|A\|_F^2 - \|A_k\|_F^2 + k\Delta \\
&\leq \|A - A_k\|_F^2 + \frac{k}{\ell - k}\|A - A_k\|_F^2 && \text{by } \Delta \leq \|A - A_k\|_F^2/(\ell - k) \\
&= \frac{\ell}{\ell - k}\|A - A_k\|_F^2.
\end{aligned}$$

Setting $\ell = \lceil k + k/\varepsilon \rceil$ results in the standard bound form $\|A - \pi_B^k(A)\|_F^2 \leq (1 + \varepsilon)\|A - A_k\|_F^2$.

4.1.2 Running Time Analysis

Each iteration of Algorithm 4.1.1 is dominated by the computation of the $\mathsf{svd}(B)$. The standard running time of this operation is $O(d\ell^2)$ [2]. Since this loop is executed once per row in A, the total running time would naïvely be $O(nd\ell^2)$. However, we show below that at the expense of doubling the space (i.e. considering 2ℓ rows for B), we can achieve lower running time $O(nd\ell)$[3]. Algorithm 4.1.2 gives the details.

Algorithm 4.1.2 FAST-FREQUENTDIRECTIONS

Input: $\ell, A \in R^{n \times d}$
$B \leftarrow$ all zeros matrix $\in R^{2\ell \times d}$
for $i \in 1, \ldots, n$ **do**
  Insert $A_{i,:}$ into a zero valued row of B
  if B has no zero valued rows **then**
    $[U, \Sigma, V] \leftarrow \mathsf{svd}(B)$
    $C = \Sigma V^T$ # Only needed for proof notation
    $\delta \leftarrow \Sigma^2_{\ell,\ell}$
    $B \leftarrow \sqrt{\max(\Sigma^2 - I_\ell \delta, 0)} \cdot V^T$ # The last $\ell + 1$ rows of B are zero valued.
return B

Note that in Algorithm 4.1.2 the svd of B is computed only $n/(\ell+1)$ times because the "if" statement is only triggered once every $\ell + 1$ iterations. Thereby exhibiting a total running time of $O((n/\ell)d\ell^2) = O(nd\ell)$. The reader should revisit the proofs in Section 4.1.1 and observe that they still hold. Consider the values of i for which the "if" statement is triggered. It still holds that $0 \preceq C_{[i]}^T C_{[i]} - B_{[i]}^T B_{[i]} \preceq \delta I_d$ and that $\|C_{[i]}\|_F^2 - \|B_{[i]}\|_F^2 \geq \ell\delta$. For the other values of i, the sketch simply aggregates the input rows and there is clearly no incurred error in doing that. This is sufficient for the same analysis to go through and complete the discussion on correctness of Algorithm 4.1.2.

4.1.3 Parallelization and Merging Sketches

In extremely large datasets, the processing is often distributed among several machines. Each machine receives a disjoint input of raw data and is tasked with creating a small space summary. Then to get a global summary of the entire data, these summaries need to be combined. The core problem is illustrated for case of just two machines, each processes a data set A_1 and A_2, where $A = [A_1; A_2]$, and create two summaries B_1 and B_2, respectively. Then the goal is to create a single summary B which approximates A using only B_1 and B_2. If B can achieve the same formal space/error tradeoff as each B_1 to A_1 in a streaming algorithm, then the summary is called a *mergeable summary*.

Here we show that the FREQUENTDIRECTIONS sketch is indeed mergeable[3] under the following procedure. Consider $B' = [B_1; B_2]$ which has 2ℓ rows; then run FREQUENTDIRECTIONS (in particular Algorithm 4.1.2) on B' to create sketch B with ℓ rows. Given that B_1 and B_2 satisfy properties 1, 2, and 3 with parameters Δ_1 and Δ_2, respectively, we will show that B satisfies the same properties with $\Delta = \Delta_1 + \Delta_2 + \delta$, where δ is taken from the single shrink operation used in Algorithm 4.1.2. This implies B automatically inherits the bounds in Theorem 4.1 and Theorem 4.2 as well.

First note that B' satisfies all properties with $\Delta' = \Delta_1 + \Delta_2$, by additivity of squared spectral norm along any direction x (e.g. $\|B_1 x\|^2 + \|B_2 x\|^2 = \|B' x\|^2$) and squared Frobenious norms (e.g. $\|B_1\|_F^2 + \|B_2\|_F^2 = \|B'\|_F^2$), but has space twice as large as desired. Property 1 holds since B only shrinks all directions in relation to B'. Property 2 follows by considering any unit vector x and expanding $\|Bx\|^2$ as

$$\|Bx\|^2 \geq \|B'x\|^2 - \delta \geq \|Ax\|^2 - (\Delta_1 + \Delta_2) - \delta = \|Ax\|^2 - \Delta.$$

Similarly, Property 3 can be seen as

$$\|B\|_F^2 \leq \|B'\|_F^2 - \delta\ell \leq \|A\|_F^2 - (\Delta_1 + \Delta_2)\ell - \delta\ell = \|A\|_F^2 - \Delta\ell.$$

This property trivially generalizes to any number of partitions of A. It is especially useful when the matrix (or data) is distributed across many machines. In this setting, each machine can independently compute a local sketch. These sketches can then be combined in an arbitrary order using FREQUENTDIRECTIONS.

Bibliography

[1] Mina Ghashami, Edo Liberty, Jeff M Phillips, and David P Woodruff. Frequent directions: Simple and deterministic matrix sketching. *arXiv preprint arXiv:1501.01711*, 2015.

[2] Gene H. Golub and Charles F. van Loan. *Matrix computations (3. ed.)*. Johns Hopkins University Press, 1996.

[3] Edo Liberty. Simple and deterministic matrix sketching. In *KDD*, pages 581–588, 2013.

[4] Jayadev Misra and David Gries. Finding repeated elements. *Sci. Comput. Program.*, 2(2):143–152, 1982.