Abstract  This lecture deals with the Frequent Directions algorithm (1) for computing an estimation to a matrix of the type $A^TA$. The usefulness of this method is shown below.

Matrices

Matrices can be used to represent many things. For example:

- Graphs: $A(i,j) = 1 \iff$ edge $i \rightarrow j$.
- A collection of data with features, e.g. a corpus of documents over a fixed set of words. $A(i,j) = 1 \iff$ word $j$ appears in document $i$.
- A collection of bipartite relations, e.g. $A(i,j) = 1 \iff$ viewer $i$ watched movie $j$.

Let’s look at the second example. If we take, for example, columns $u$ and $v$ (representing words $u$ and $v$) we can measure the similarity between the words with $|u^Tv|$ (if they generally appear in the same documents we assume they are similar). Thus, we see that the matrix $A^TA$ is a matrix of word similarity. Likewise, the matrix $AA^T$ is a matrix of document similarity. We are interested in fast computation of these matrices.

Let’s say we receive the documents one by one and we would like to compute $A^TA$. We compute some other matrix $B$ and would like to minimize the norm $\|A^TA - B\|$. What are some important matrix norms to know?

1. Frobenius - $\|M\|_F^2 = \sum M_{i,j}^2 = \text{trace} (M^TM) = \text{trace} (MM^T)$.

2. Spectral - $\|M\|^2 = \max_{x \in \mathbb{R}^m, x \neq 0} \frac{\|Mx\|}{\|x\|} = \max_{y \in \mathbb{R}^n} \|y^TM\|$. This is the most the matrix stretches.

3. $\|M\|_{p\rightarrow q} = \max_{x \in \mathbb{R}^m, x \neq 0} \frac{\|Mx\|_p}{\|x\|_q}$ where $\|x\|_p = (\sum |x_i|^p)^{\frac{1}{p}}$. Also $\|x\|_\infty = \max_{i=1..m} |x_i|$.

$\text{rank} (M) = \text{dim} (\text{span} (\text{cols} (M))) = \text{dim} (\text{span} (\text{rows} (M)))$.

To continue, we must first understand the Singular Value Decomposition, or SVD.
Singular Value Decomposition (SVD):

A singular value is related to an eigenvalue. We define it iteratively. The top singular value is exactly the spectral norm of the matrix:

$$\sigma_1(M) = \|M\|$$

As mentioned above, by definition of the norm, the norm can be achieved by multiplication on the right and also on the left. So by definition:

$$\exists \begin{align*} x \in \mathbb{R}^m, \quad y \in \mathbb{R}^n, \quad &s.t. \quad \|Mx\| = \sigma_1 = \|y^TM\| \end{align*}$$

x is a right singular vector and y is a left singular vector. We will assume x and y are unique. We will also assume $m < n$. In fact $Mx = \sigma_1y$, $M^Ty = \sigma_1x$.

This means we can decompose M into:

$$M = \begin{pmatrix} n & \begin{array}{c} \vdots \\ \vdots \\ \vdots \end{array} \\ \vdots \\ \vdots \\ \vdots \end{pmatrix}_{m \times n} \begin{pmatrix} \begin{array}{cccc} \sigma_1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \end{array} \end{pmatrix}_{m \times m} \begin{pmatrix} \begin{array}{c} 0 \\ \vdots \\ \vdots \\ \vdots \end{array} \\ \vdots \\ \vdots \end{pmatrix}_{m \times n}$$

Orthogonal to $y$

Orthogonal to $x$

Now we can continue by induction and get the following decomposition:

$$M = U \Sigma V^T$$

where $U$ and $V$ are orthogonal matrices and $\sigma_1 > \sigma_2 > \ldots > \sigma_m$. Note that $\sigma_1 = \text{spectral norm } \|M\|_F^2 = \sum \sigma_i^2$ and rank $(M) = \# \text{non-zeros among } \sigma_1, \ldots, \sigma_m$. The kernel of M is the space spanned by the singular vectors corresponding to the singular values that are equal to 0.

If we have a matrix with singular values close to 0, then we can say it is approximately a lower rank matrix for which the corresponding singular values are exactly 0. This is how we define the numerical rank of a matrix, though we will not go into that.

Singular values and eigenvalues:

Now, let’s connect singular values to eigenvalues. We will look at $M^TM$.

$$M^TM = V\Sigma^TUU^T\Sigma V^T = V\Sigma^T\Sigma V^T$$

where $\Sigma$ is the diagonal matrix of singular values.
where

\[
\Sigma^T \Sigma = \begin{pmatrix}
\sigma_1^2 & & \\
& \ddots & \\
& & \sigma_r^2
\end{pmatrix}
\]

\(\sigma_1^2...\sigma_r^2\) are both singular values and eigenvalues of \(M^T M\).

Any real matrix has a complete set of singular values and they are all at least 0.

**Approximating matrices using the SVD:**

We want to approximate a matrix by “cleaning out the noise”. Basically we would like to replace the 0.001 (very small) singular values with 0. Then we want to ask if this is a good approximation and is it the best possible? Let’s say we have a matrix \(M = U\Sigma V^T\) which we suspect is “close” to low rank (this is common since most matrices in nature are technically of full rank, e.g. the derivative of an image in not exactly sparse but is approximately sparse).

**Definition** A rank-r approximation of M is a matrix \(\tilde{M}\) of rank \(\leq r\) such that \(\|M - \tilde{M}\|?\) is “small” (for some norm ?).

**Theorem** If \(? \in \{F, \text{spectral}\}\) then the best rank-r approximation of M is given by:

\[
\tilde{M} = U \begin{pmatrix}
\sigma_1 & & \\
& \ddots & \\
& & \sigma_r
\end{pmatrix} V^T
\]

This means \(\tilde{M} = \arg\min_{M' \text{ of rank } r} \|M - M'\|?\)

Why is this surprising? Any rank r matrix can be written as \(\tilde{M} = (L_{i,j}) (R_{k,l})\) where the first matrix is n*r and the second is r*m. As we said we want ot minimize \(\|M - M'\|?\). The norm is a convex function. The problem is that \(\tilde{M}\) is not a convex function of the L’s and R’s. So this is a non convex optimization problem in a very strong sense. But nevertheless it has a closed form solution which is very simple once you know the SVD of M and the SVD of any matrix can be computed efficiently.
We call the error $\epsilon_r$. So

$$\epsilon_r = M - \tilde{M} = U \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{r+1} \\ \sum & \cdots & \sigma_m \end{pmatrix} V^T$$

$$\|\epsilon_r\| = \sigma_{r+1}$$

$$\|\epsilon_r\|_F^2 = \sum_{i=r+1}^{m} \sigma_i^2$$

**Frequent Directions (1):**

The following algorithm is used to approximate $A^T A$, when the rows of $A$ appear in a streaming fashion, using a matrix $B$ such that $\|A^T A - B^T B\|_M$ is small and $B$ has low rank.

**Algorithm 1** Frequent Directions (Edo Liberty)

```
Input: $A \in \mathbb{R}^{n \times m}$ ($n \gg m$) (given row by row)
1 $B^0 \leftarrow 0^{l \times m}$;
2 for $i = 1..n$ (obtain $A_i$) do
3       insert $A_i$ into a zero row of $B$ (always exists);
4       if $B$ now has no zero row then
5           $[U, \Sigma, V] = \text{SVD}(B)$ (in other words $U \Sigma V^T = B$);
6           $C \leftarrow \Sigma V^T$;
7           $\hat{\Sigma} := \sqrt{\max(\Sigma^2 - I_d\sigma_{med}^2, 0)}$;
8           $B = \hat{\Sigma} V^T$
9       else
10          $C = \Sigma V^T$
11     end
12 end
```

Notice that in line 6 of the algorithm we ignore $U$. This is because we are computing $B^T B = V\Sigma \tilde{U}^T U \Sigma V^T$. $\Sigma$ is an $l \times l$ matrix which has in its diagonal $\sigma_1...\sigma_l$, some of which might be 0. Our goal is to take $C$ and clear out some row there. So we are going to clear out the row with the lowest $\sigma$ values. We calculate $\hat{\Sigma}$ by clearing the rows (about half) whose $\sqrt{\sigma^2 - \sigma_{med}^2}$ falls below 0.
So it looks like this:

\[
\hat{\Sigma} = \begin{pmatrix}
\sqrt{\sigma_1^2 - \sigma_{\text{med}}^2} & & \\
& \ddots & \\
& & \sqrt{\sigma_{l/2}^2 - \sigma_{\text{med}}^2} \\ 0 & & \\
& & & \ddots & \\
& & & & \ 0
\end{pmatrix}
\]

This is almost like subtracting the median singular value from each one of the singular values, but in the squared domain. This frees up half of the singular values, freeing up half of the rows of \(B\). Note that this algorithm is similar in its concept the the “Heavy Hitters” algorithm (finding elements with high frequency in a stream) which we discussed in a previous lecture. Here we try to “track” the highest singular values.

**Theorem**  For \(B\) yielded by the above algorithm \(\|A^T A - B^T B\| \leq \frac{2\|A\|^2_F}{l} \).

**Proof**

Superscript \(i\) - after \(i\) iterations have completed.

Subscript \(i\) - row \(i\).

Note that for any symmetric matrix \(S\) we have \(\|S\| = \max_{\|x\|=1} \sqrt{x^T S x}\).

\[
\frac{\|A^T A - B^T B\|}{\|A^T A - B^T B\|}\text{ symmetric} = \max_{\|x\|=1} \|x^T (A^T T - B^T B) x\| = \|x^T A^T Ax - x^T B^T Bx\| \text{ telescopic sum}
\]

\[
= \sum_{i=1}^{n} (A_i x)^2 + x^T (B^{i-1})^T B^{i-1} x - \sum_{i=1}^{n} x^T (B^i)^T B^i x = \sum_{i=1}^{n} x^T \left( (C^i)^T C^i - (B^i)^T B^i \right) x \leq \sum_{i=1}^{n} \| (C^i)^T C^i - (B^i)^T B^i \| \leq \sum_{i=1}^{n} \left( \sigma^i_{\text{med}} \right)^2
\]

Multiplying by \(V\) from left and right doesn’t change the norm

The rest of the proof is to show the last part is small.
References