1. (a) Prove the cyclic property of the trace:

$$\mathsf{trace}(ABC) = \mathsf{trace}(BCA) = \mathsf{trace}(CAB) \tag{1}$$

for all A, B, C such that their product is defined and is a square matrix.

Solution: We prove first the analogous statement for two matrices,

$$\mathsf{trace}(AB) = \mathsf{trace}(BA).$$

Suppose A is an $m \times n$ matrix, in which case in order for both the products AB and BA to be well-defined, B must be an $n \times m$ matrix. Further, note that

$$(AB)_{ii} = \sum_{j=1}^{n} a_{ij} b_{ji} \qquad\qquad (BA)_{jj} = \sum_{i=1}^{m} b_{ji} a_{ij}$$

and hence the result follows since $a_{ij} b_{ji} = b_{ji} a_{ij}$ and by interchanging the sums,

$$\mathsf{trace}(AB) = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ji} = \sum_{i=1}^{m} \sum_{j=1}^{n} b_{ji} a_{ij} = \sum_{j=1}^{n} \sum_{i=1}^{m} b_{ji} a_{ij} = \mathsf{trace}(BA).$$

Hence, we have proven the cyclic property of the trace for two matrices.

Now let A, B, and C be matrices such that the products ABC, BCA, and CAB are well-defined and suppose A is an $m \times n$ matrix. Then it must be that B is an $n \times p$ matrix, and C is a $p \times m$ matrix. In particular, this means that BC is an $n \times m$ matrix, and so

$$\mathsf{trace}(ABC) = \mathsf{trace}(A(BC)) = \mathsf{trace}((BC)A) = \mathsf{trace}(BCA).$$

Similarly, since CA is a $p \times n$ matrix,

$$\mathsf{trace}(BCA) = \mathsf{trace}(B(CA)) = \mathsf{trace}((CA)B) = \mathsf{trace}(CAB),$$

and so we have the following cyclic property of the trace

$$\mathsf{trace}(ABC) = \mathsf{trace}(BCA) = \mathsf{trace}(CAB).$$

(b) Prove that

$$\|A\|_F^2 = \sum_{i=1}^{d} \sigma_i^2. \tag{2}$$

Hint: use the full SVD of A and *the cyclic property of trace.*

Solution: We claim first that $\|A\|_F^2 = \mathsf{trace}(A^\top A)$. By definition of the Frobenius norm, we have that

$$\|A\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{d} |a_{ij}|^2.$$

Further, notice that

$$(A^\top A)_{jj} = \sum_{i=1}^{n} |a_{ij}|^2$$

and hence interchanging the sums,

$$\|A\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{d} |a_{ij}|^2 = \sum_{j=1}^{d}\sum_{i=1}^{n} |a_{ij}|^2 = \sum_{j=1}^{d} (A^\top A)_{jj} = \mathsf{trace}(A^\top A),$$

so $\|A\|_F^2 = \mathsf{trace}(A^\top A)$. Now by the SVD, we can write $A = U\Sigma V^\top$, in which case

$$A^\top A = (U\Sigma V^\top)^\top (U\Sigma V^\top) = V\Sigma^2 V^\top$$

and so via the cyclic property of the trace proven in item (a),

$$\mathsf{trace}(A^\top A) = \mathsf{trace}(V\Sigma^2 V^\top) = \mathsf{trace}(\Sigma^2 V^\top V) = \mathsf{trace}(\Sigma^2) = \sum_{i=1}^{d} \sigma_i^2$$

meaning $\|A\|_F^2 = \sum_{i=1}^{d} \sigma_i^2$ as desired.

(c) Prove that

$$\|A + B\|_F^2 = \|A\|_F^2 + \|B\|_F^2 + 2\langle A, B\rangle_F \tag{3}$$

where $\langle A, B\rangle_F$ is the Frobenius inner product. The Frobenius inner product is defined as

$$\langle A, B\rangle_F \coloneqq \sum_{i,j} a_{ij} b_{ij} = \mathsf{trace}(A^\top B) = \mathsf{trace}(B^\top A). \tag{4}$$

Solution: By definition of the Frobenius norm, we have that

$$\|A + B\|_F^2 = \sum_{i,j} |a_{ij} + b_{ij}|^2 = \sum_{i,j} (a_{ij} + b_{ij})^2 = \sum_{i,j} (a_{ij}^2 + b_{ij}^2 + 2a_{ij}b_{ij})$$

and further, splitting the sum, we find

$$\begin{aligned}\|A + B\|_F^2 &= \sum_{i,j} (a_{ij}^2 + b_{ij}^2 + 2a_{ij}b_{ij}) = \sum_{i,j} a_{ij}^2 + \sum b_{ij}^2 + 2\sum_{i,j} a_{ij}b_{ij} \\ &= \sum_{i,j} |a_{ij}|^2 + \sum_{i,j} |b_{ij}|^2 + 2\sum_{i,j} a_{ij}b_{ij} = \|A\|_F^2 + \|B\|_F^2 + 2\langle A, B\rangle_F\end{aligned}$$

as desired.

2. Prove the Eckart-Young-Mirsky theorem for any Ky-Fan norm.

Theorem 1. Let $A = U\Sigma V^\top$ be an SVD of A and M be any matrix of the size of A such that $\mathsf{rank}(M) \leq k$. Then

$$\|A - M\| \geq \|A - U_k \Sigma_k V_k^\top\| \text{ for any Ky-Fan norm } \|\cdot\|,$$

where U_k and V_k consist of the first k columns of U and V, respectively, and $\Sigma_k = \mathsf{diag}\{\sigma_1, \ldots, \sigma_k\}$.

You can use Lemma 1 in Section 4.3 in `LinearAlgebra.pdf`. In this case, write it in your proof and explain every nontrivial equality in it. This will help you understand this Lemma.

Solution: Let A be an $n \times d$ matrix. If $\|\cdot\|$ is a Ky-Fan norm, then for some p, it corresponds to the ℓ^p norm of the vectors of singular values. Then truncating the sum, we see that

$$\|A - M\|^p = \sum_{i=1}^{d} \sigma_i(A - M)^p \geq \sum_{i=1}^{d-k} \sigma_i(A - M)^p.$$

Now taking $B \coloneqq A - M$, $C \coloneqq M$, and $j \coloneqq k + 1$ in Lemma 1, we have that

$$\sigma_{i+k}(A) \leq \sigma_i(A - M) + \sigma_{k+1}(M)$$

and further note that $\sigma_{k+1}(M) = 0$ since $\mathsf{rank}(M) \leq k$, hence $\sigma_i(A - M) \geq \sigma_{i+k}(A)$ and so

$$\sum_{i=1}^{d-k} \sigma_i(A - M)^p \geq \sum_{i=1}^{d-k} \sigma_{i+k}(A)^p = \sum_{i=k+1}^{d} \sigma_i(A)^p = \|A - U_k \Sigma_k V_k^\top\|^p.$$

Finally, taking the pth root on both sides, we obtain

$$\|A - M\| \geq \|A - U_k \Sigma_k V_k^\top\|$$

as desired.

3. The dataset for this problem is downloaded from this webpage: TechTC – Technion Repository of Text Categorization. The particular file that I used is Preprocessed feature vectors: techtc300 preprocessed.zip (117,951,459 bytes; approx. 301Mb uncompressed). Its description is available here.

   I extracted two data files from it: `vectors.txt` and `words_idx.txt`. Each line of `words_idx.txt` is of the form

   ```
   <word index><word>
   ```

   A total of 18446 words. The file `vectors.txt` contains 278 lines. Lines 1, 3, 5, etc, i.e. all odd lines, contain a single number, the index of a document. Lines 2, 4, 6, etc, i.e. all even lines, contain information about the content of the document whose index is in the line above. The information is given as follows. The first number is 1 or -1, a label of the document. Label -1 attributes the document to Florida, while label 1 corresponds to Indiana. The rest of the numbers are the word indices and the counts of the corresponding words in the document. For example, line 2 starts with

   ```
   -1 1   54 2   11 3   53 4   22 5   44 6
   ```

   This means that document 1 belong to the category -1. Furthermore, word 1 is encountered 54 times, word 2 is encountered 11 times, word 3 is encountered 53 times, word 4 is encountered 22 times, etc.

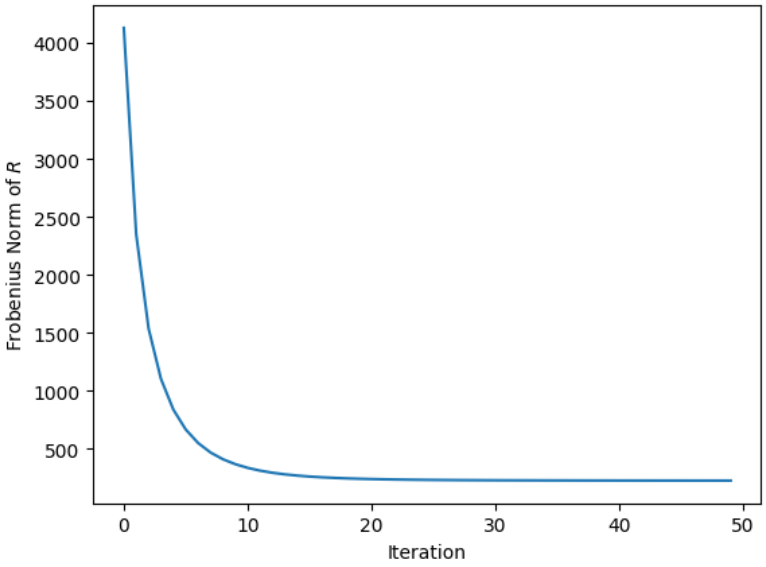
   My code `DocsLeeSeung.ipynb` reads the data from files, creates an $N_{\text{words}} \times N_{\text{docs}}$ matrix A

such that $A_{i,j} = 1$ if word i is present in document j, and computes its factorization $A \approx WH$ where $W \in \mathbb{R}_+^{N_{\text{words}} \times k}$ and $H \in \mathbb{R}_+^{k \times N_{\text{docs}}}$ using the Lee-Seung algorithm. I set $k = 10$.

You can choose to do this problem in Matlab or in Python.

(a) Implement the projected gradient descent to factorize $A \approx WH$ where $W \in \mathbb{R}_+^{N_{\text{words}} \times k}$ and $H \in \mathbb{R}_+^{k \times N_{\text{docs}}}$, $k = 10$. Plot the Frobenius norm of $R \coloneqq A - WH$ versus the number of iteration. Determine (approximately) the number of iterations sufficient to make the residual $\|R\|_F$ stop changing visibly. Check if the eventual norm of the residual depends on the initial approximation. Check which words correspond to relatively high numbers in the columns of W and, looking at them, hypothesize what is the common theme of this set of documents.

Solution: We produce the following plot for projected gradient descent via this code.

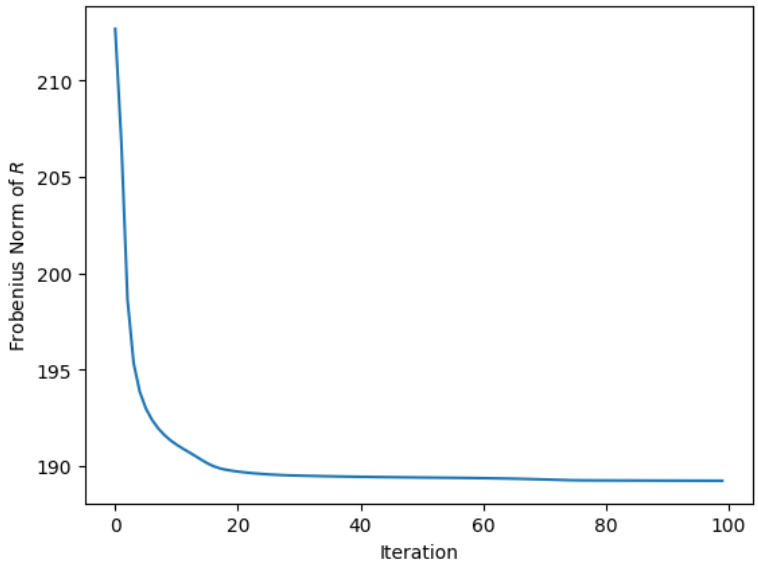


We see that after around 20 iterations, the norm of the residual $\|R\|_F$ stops changing visibly. Running the code several times, the eventual norm does not appear to depend on the initial approximation.

I hypothesize the common theme of these documents is that they are letters, hence words such as "you" and "your" corresponding to high numbers in the columns of W.

(b) Do the same task for the HALS algorithm (Section 5.3.2 in `LinearAlgebra.pdf`).

Solution: We produce the following plot for the HALS algorithm via this code.



We see that after around 30 iterations, the norm of the residual $\|R\|_F$ stops changing visibly. Running the code several times, the eventual norm does not appear to depend on the initial approximation.

(c) Find the best approximation A_{10} of A by a matrix of rank ≤ 10 using SVD. Compute $\|A - A_{10}\|_F$. Compare it with the residuals.

Solution: After 50 iterations of projected gradient descent with a learning rate of $\alpha = 10^{-5}$, we achieve the final residual norm $\|R\|_F \approx 227.85$. Likewise, after 100 iterations of the HALS algorithm, we achieve the final residual norm $\|R\|_F \approx 189.22$.

In contrast, taking the truncated SVD A_{10} of A, we find that $\|A - A_{10}\|_F \approx 187.72$. Hence, we see that the HALS algorithm comes much closer to the norm of the best approximation of A with rank ≤ 10.