School of Computer Science and Technology
University of Science and Technology of China

Exercise Sheet 1 for Algorithms for Big Data 2023 Spring

Solution

Note: In the following, we will refer the book "Foundations of Data Science" by Blum, Hopcroft and Kannan as **[BHK]**.

Note: In the following, for a vector $\boldsymbol{x}$, $\|\boldsymbol{x}\| = \|\boldsymbol{x}\|_2$.

Exercise 1

(Exercise 3.12 in [BHK]) Let $\sum_i \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ be the singular value decomposition of a rank r matrix A. Let $A_k = \sum_{i=1}^k \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ be a rank k approximation to A for some $k < r$. Express the following quantities in terms of the singular values $\{\sigma_i, 1 \le i \le r\}$.

(a) $\|A_k\|_F^2$

(b) $\|A_k\|_2^2$

(c) $\|A - A_k\|_F^2$

(d) $\|A - A_k\|_2^2$

Solution.

(a) The definition that $A_k = \sum_{i=1}^k \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ is also the SVD of A_k. As a result, the singular values of A_k are $\sigma_1, \sigma_2, \dots, \sigma_k$. By the connection between Frobenius norm and singular values of a matrix, we can see that $\|A_k\|_F^2 = \sum_{i=1}^k \sigma_i^2$.

(b) Since the singular values of A_k are $\sigma_1, \sigma_2, \dots, \sigma_k$ and $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_k$, we can see that the largest singular value of A_k is σ_1. Since the 2-norm of a matrix is equal to its largest singular value, it is obvious that $\|A_k\|_2 = \sigma_1$ and thus $\|A_k\|_2^2 = \sigma_1^2$.

(c) Since $A = \sum_{i=1}^r \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ and $A_k = \sum_{i=1}^k \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, we can see that $A - A_k = \sum_{i=k+1}^r \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, which is the SVD of $A - A_k$. As a result, the singular values of $A - A_k$ are $\sigma_{k+1}, \dots, \sigma_r$ and thus $\|A - A_k\|_F^2 = \sum_{i=k+1}^r \sigma_i^2$.

(d) Since the singular values of $A - A_k$ are $\sigma_{k+1}, \dots, \sigma_r$ and $\sigma_{k+1} \ge \cdots \ge \sigma_r$, we can see that the largest singular value of $A - A_k$ is σ_{k+1} and thus $\|A - A_k\|_2^2 = \sigma_{k+1}^2$.

Exercise 2

(Exercise 3.18 in [BHK]) A matrix A is positive semi-definite if for all $\boldsymbol{x}$, $\boldsymbol{x}^\top A \boldsymbol{x} \ge 0$.

(a) Let A be a real valued matrix. Prove that $B = AA^\top$ is positive semi-definite.

(b) Let A be the adjacency matrix of a graph. The Laplacian of A is $L = D - A$ where D is a diagonal matrix whose diagonal entries are the row sums of A. Prove that L is positive semi-definite by showing that $L = B^\top B$ where B is an m-by-n matrix with a row for each edge in the graph, a column for each vertex, and we define

$$b_{ei} = \begin{cases} -1 & \text{if } i \text{ is the endpoint of } e \text{ with lesser index} \\ 1 & \text{if } i \text{ is the endpoint of } e \text{ with greater index} \\ 0 & \text{if } i \text{ is not an endpoint of } e \end{cases}$$

Proof.

(a) For all $\boldsymbol{x}$, we have $\boldsymbol{x}^\top B\boldsymbol{x} = \boldsymbol{x}^\top AA^\top \boldsymbol{x} = (A^\top \boldsymbol{x})^\top (A^\top \boldsymbol{x}) = \left\|A^\top \boldsymbol{x}\right\|_2^2 \geq 0$. Therefore, B is positive semi-definite.

(b) Let $G = (V, E, \psi)$ denote the graph, where V denotes the set of vertices, E denotes the set of edges, and $\psi : E \to \mathcal{P}(V)$ is a function. Here $\mathcal{P}(V)$ denotes the power set of V. Note that the vertices in $\psi(e)$ are the endpoints of the edge e.

It suffices to prove that each entry of $B^\top B$ is equal to that of L, i.e., $\ell_{ij} = (B^\top B)_{ij}$, $\forall i, j \in V$. We consider the following two cases:

(i) $i \neq j$

Since D is a diagonal matrix, $d_{ij} = 0$. Since A is the adjacency matrix of G, a_{ij} is equal to the number of edges between vertices i and j, i.e., $a_{ij} = |\{e \in E \mid \psi(e) = \{i, j\}\}|$. Therefore, $\ell_{ij} = d_{ij} - a_{ij} = -|\{e \in E \mid \psi(e) = \{i, j\}\}|$ because $L = D - A$.

Note that $(B^\top B)_{ij} = \sum_{e \in E} b_{ei} b_{ej}$. For each $e \in E$, if $b_{ei} \neq 0$ and $b_{ej} \neq 0$ hold, either (1) $b_{ei} = 1$ and $b_{ej} = -1$ or (2) $b_{ei} = -1$ and $b_{ej} = 1$ must hold, because 1 and -1 are the only two possible non-zero values in the e-th row of B. As a result, $b_{ei} b_{ej} = -1$ or $b_{ei} b_{ej} = 0$ holds for all $e \in E$ and $i, j \in V$.

By the definition of B, $b_{ei} \neq 0$ iff (if and only if) i is the endpoint of e. Therefore, $b_{ei} b_{ej} = -1$ iff i and j are both the endpoints of e, i.e., $\psi(e) = \{i, j\}$. As a result,

$$\begin{aligned}
(B^\top B)_{ij} = \sum_{e \in E} b_{ei} b_{ej} &= \sum_{e \in E, b_{ei} b_{ej} = -1} b_{ei} b_{ej} + \sum_{e \in E, b_{ei} b_{ej} = 0} b_{ei} b_{ej} \\
&= \sum_{e \in E, b_{ei} b_{ej} = -1} -1 + \sum_{e \in E, b_{ei} b_{ej} = 0} 0 \\
&= -\left|\{e \in E \mid b_{ei} b_{ej} = -1\}\right| \\
&= -\left|\{e \in E \mid \psi(e) = \{i, j\}\}\right| \\
&= \ell_{ij}
\end{aligned}$$

(ii) $i = j$

In this case, we only need to prove that $\forall i \in V$, $\ell_{ii} = (B^\top B)_{ii}$. Note that $\left(B^\top B\right)_{ii} = \sum_{e \in E} b_{ei}^2$. Since $b_{ei}^2 \in \{0, 1\}$, $b_{ei}^2 = 1$ iff i is an endpoint of e, i.e., $i \in \psi(e)$. It is easy to obtain that

$$\begin{aligned}
(B^\top B)_{ii} = \sum_{e \in E} b_{ei}^2 &= \sum_{e \in E, b_{ei}^2 = 1} b_{ei}^2 + \sum_{e \in E, b_{ei}^2 = 0} b_{ei}^2 \\
&= \sum_{e \in E, b_{ei}^2 = 1} 1 + \sum_{e \in E, b_{ei}^2 = 0} 0 \\
&= \left|\left\{e \in E \mid b_{ei}^2 = 1\right\}\right| \\
&= \left|\{e \in E \mid i \in \psi(e)\}\right| \\
&= \sum_{j \in V \setminus \{i\}} \left|\{e \in E \mid \psi(e) = \{i, j\}\}\right| \\
&= \sum_{j \in V \setminus \{i\}} a_{ij} \\
&= \sum_{j \in V} a_{ij} - a_{ii} \\
&= d_{ii} - a_{ii} \\
&= \ell_{ii}
\end{aligned}$$

$\square$

Exercise 3
(Exercise 3.22 in [BHK])

(a) For any matrix A, show that $\sigma_k \leq \frac{\|A\|_F}{\sqrt{k}}$.

(b) Prove that there exists a matrix B of rank at most k such that $\|A - B\|_2 \leq \frac{\|A\|_F}{\sqrt{k}}$.

(c) Can the 2-norm on the left-hand side in (b) be replaced by Frobenius norm?

Proof.

(a) Let r denote the rank of A and $\sigma_1, \sigma_2, \ldots, \sigma_r$ denote the singular values of A. By the connection between singular values and Frobenius norm of a matrix, and the fact that singular values are non-negative, we can see that

$$\|A\|_F^2 = \sum_{i=1}^{r} \sigma_i^2 \geq \sum_{i=1}^{k} \sigma_i^2 \geq \sum_{i=1}^{k} \sigma_k^2 = k\sigma_k^2$$

Therefore,

$$\sigma_k^2 \leq \frac{\|A\|_F^2}{k}$$
$$\sigma_k \leq \frac{\|A\|_F}{\sqrt{k}}$$

(b) We choose $B = A_k$, i.e., the projection of A onto the best-fitting k-dimensional subspace, and thus $\|A - B\|_2 = \|A - A_k\|_2 = \sigma_{k+1} \leq \sigma_k \leq \frac{\|A\|_F}{\sqrt{k}}$.

(c) No. We choose $r = 5$, $k = 2$ and $A = I_r$ (the $r \times r$ identity matrix). It is obvious that all the singular values of A are 1. By the properties of SVD, we can see that $A_k = \arg\min_{\mathrm{rank}(B)=k} \|A - B\|_F$. Therefore,

$$\|A - B\|_F \geq \|A - A_k\|_F = \sqrt{\sum_{i=k+1}^{r} \sigma_i^2} = \sqrt{3} > \frac{\sqrt{5}}{\sqrt{2}} = \frac{\|A\|_F}{\sqrt{k}}$$

□

....

(Exercise 3.23 in [BHK]) Suppose an $n \times d$ matrix A is given and you are allowed to preprocess A. Then you are given a number of d-dimensional vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ and for each of these vectors you must find the vector $A\boldsymbol{x}_j$ approximately, in the sense that you must find a vector $\boldsymbol{y}_j$ satisfying $\left\|\boldsymbol{y}_j - A\boldsymbol{x}_j\right\| \leq \varepsilon \|A\|_F \|\boldsymbol{x}_j\|$. Here $\varepsilon > 0$ is a given error bound. Describe an algorithm that accomplishes this in time $O\left(\frac{d+n}{\varepsilon^2}\right)$ per $\boldsymbol{x}_j$ not counting the preprocessing time.

Hint: use Exercise 3.22 in [BHK].

Solution. When preprocessing, we calculate the SVD of A, i.e., $\{(\sigma_i, \boldsymbol{u}_i, \boldsymbol{v}_i)\}_{i=1}^r$ s.t. $A = \sum_{i=1}^r \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$ and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r$. When answering a query, we use the following algorithm:

Algorithm 1: Query

Data: Singular value decomposed matrix $A = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$
Input: Input vector $\boldsymbol{x}$ and error bound ε
Output: $\boldsymbol{y} \approx A\boldsymbol{x}$

1 Calculate $k = \lceil \frac{1}{\varepsilon^2} \rceil$.
2 For $i = 1, 2, \ldots, k$, iteratively calculate $a_i \triangleq \sigma_i \boldsymbol{v}_i^\top \boldsymbol{x}$.
3 Calculate $\boldsymbol{y} \triangleq \sum_{i=1}^{k} a_i \boldsymbol{u}_i$ and return.

Analysis of the algorithm:

(a) Running time: The above algorithm calculates $\boldsymbol{y} = A_k\boldsymbol{x}$ in $O(k(d+n)) = O(\frac{d+n}{\varepsilon^2})$ time, as calculating k inner products of two d-dimensional vectors (Step 2) takes $O(kd)$ time and accumulating k vectors of n dimensions (Step 3) takes $O(kn)$ time.

(b) Correctness: Note that $\sqrt{k} \geq \frac{1}{\varepsilon}$ (since $k = \lceil \frac{1}{\varepsilon^2} \rceil \geq \frac{1}{\varepsilon^2}$), and $\|A\boldsymbol{x}\|_2 \leq \|A\|_2 \|\boldsymbol{x}\|_2$ (since $\|A\|_2 = \sup_{\|\boldsymbol{x}\|_2=1} \|A\boldsymbol{x}\|_2 = \sup_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \geq \frac{\|A\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2}$ when $\boldsymbol{x} \neq \boldsymbol{0}$), we have

$$\|\boldsymbol{y} - A\boldsymbol{x}\|_2 = \|(A_k - A)\boldsymbol{x}\|_2 \leq \|A_k - A\|_2 \|\boldsymbol{x}\|_2 \leq \frac{\|A\|_F}{\sqrt{k}} \|\boldsymbol{x}\|_2 \leq \varepsilon \|A\|_F \|\boldsymbol{x}\|_2$$

Exercise 4
Let $A \in \mathbb{R}^{n\times d}$ be a data matrix with the SVD: $A = UDV^\top = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, where $r \leq d$. Suppose that $\sigma_2 < (1-\varepsilon)\sigma_1$ for some $\varepsilon > 0$. Let $\boldsymbol{x}$ be a unit vector such that $\boldsymbol{x}^\top \boldsymbol{v}_1 \geq \frac{1}{2}$. For each integer $k \geq 1$, define a vector $\boldsymbol{b}_k = (A^\top A)^k \boldsymbol{x}$.
Find the smallest possible k such that

$$\left|\boldsymbol{b}_k^\top \cdot \boldsymbol{v}_1\right| \geq \left(1 - \varepsilon^{10}\right) \|\boldsymbol{b}_k\|,$$

and explain why.

Solution. Since the right singular vectors are orthonormal, we can add $\boldsymbol{v}_{r+1}, \boldsymbol{v}_{r+2}, \ldots, \boldsymbol{v}_d$ so that $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d\}$ forms an orthonormal basis of $\mathbb{R}^d$. Hence we can write $\boldsymbol{x}$ as the linear combination of $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_d$, i.e., $\boldsymbol{x} = \sum_{i=1}^{d} a_i \boldsymbol{v}_i$ where $a_i = \langle \boldsymbol{x}, \boldsymbol{v}_i \rangle$. Since $a_1 = \langle \boldsymbol{x}, \boldsymbol{v}_1 \rangle = \boldsymbol{x}^\top \boldsymbol{v}_1 \geq \frac{1}{2}$, i.e., $4a_1^2 \geq 1$, we can see that

$$\sum_{i=2}^{r} a_i^2 \leq \sum_{i=2}^{d} a_i^2 = 1 - a_1^2 \leq 3a_1^2$$

Notice that we can expand $\boldsymbol{b}_1$ in the following way:

$$\boldsymbol{b}_1 = A^\top A\boldsymbol{x} = \left(\sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top\right)^\top \left(\sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top\right) \sum_{i=1}^{d} a_i \boldsymbol{v}_i = \left(\sum_{i=1}^{r} \sigma_i \boldsymbol{v}_i \boldsymbol{u}_i^\top\right) \sum_{i=1}^{r} a_i \sigma_i \boldsymbol{u}_i = \sum_{i=1}^{r} a_i \sigma_i^2 \boldsymbol{v}_i$$

It is easy to prove that $\boldsymbol{b}_k = \sum_{i=1}^{d} a_i \sigma_i^{2k} \boldsymbol{v}_i$ by induction. Therefore,

$$\left|\boldsymbol{b}_k^\top \boldsymbol{v}_1\right| = \left|\left(\sum_{i=1}^{r} a_i \sigma_i^{2k} \boldsymbol{v}_i^\top\right) \boldsymbol{v}_1\right| = a_1 \sigma_1^{2k}$$

$$\|\boldsymbol{b}_k\| = \|\boldsymbol{b}_k\|_2 = \sqrt{\sum_{i=1}^{r} \left(a_i \sigma_i^{2k}\right)^2} \leq \sqrt{\left(a_1 \sigma_1^{2k}\right)^2 + \sum_{i=2}^{r} \left(a_i \sigma_2^{2k}\right)^2} < \sqrt{\left(a_1 \sigma_1^{2k}\right)^2 + (1-\varepsilon)^{4k} \sigma_1^{4k} \sum_{i=2}^{r} a_i^2}$$
$$\leq a_1 \sigma_1^{2k} \sqrt{1 + 3(1-\varepsilon)^{4k}}$$

In order to make $\left|\boldsymbol{b}_k^\top \boldsymbol{v}_1\right| \geq (1-\varepsilon^{10})\|\boldsymbol{b}_k\|$, we can just make sure that

$$
\begin{array}{crcl}
 & \left(1-\varepsilon^{10}\right)^2\left(1+3(1-\varepsilon)^{4k}\right) & \leq & 1 \\
\Longleftarrow & 1+3(1-\varepsilon)^{4k} & \leq & \left(1+\varepsilon^{10}\right)^2 \\
\Longleftarrow & 3(1-\varepsilon)^{4k} & \leq & 2\varepsilon^{10} \\
\Longleftarrow & 3e^{-4k\varepsilon} & \leq & 2\varepsilon^{10} \\
\Longleftarrow & k & \geq & \frac{\ln\frac{3}{2}-10\ln\varepsilon}{4\varepsilon} \\
\Longleftarrow & k & \geq & \frac{1}{9}\varepsilon^{-1}+\frac{5}{2}\varepsilon^{-1}\ln\varepsilon^{-1}
\end{array}
$$