IFT 6760A - Lecture 5
Matrix Norms, Low Rank Approximation and Min-max Theorem

Scribe(s): Erik-Olivier Riendeau, William Duguay **Instructor:** Guillaume Rabusseau

1 Summary

In the previous lecture we continued reviewing the fundamentals of linear algebra. We started by covering diagonalizability. Then, we looked at the difference between the geometric and algebraic multiplicity. In particular, for a matrix to be diagonalizable, both these multiplicities must be equal. Also, we defined the definiteness of a matrix and the Spectral theorem. Moreover, we covered the Shur decomposition. We finished by proving the singular-value decomposition.

In this lecture we reviewed matrix norms, in particular the p-norm and the Frobenius norm. Then, we covered low rank approximation, specifically the Ecart-Young-Mirsky theorem. We looked at the Rayleigh-Ritz theorem and the Courant-Fischer (Min-max) theorem. Finally, we showed trace maximization and minimization.

2 Matrix Norms

2.1 Matrix p-norm

We start by defining the basic building block of the matrix p-norm:

Definition 1 (Vector p-norm)**.** *The* vector p-norm*, where $p \in \mathbb{R}$ is greater than* 1*, is defined as:*

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |\mathbf{x}_i|^p \right)^{1/p}.$$

See [1] for more information on vector norms.

Now, any norm on vectors induces a norm on matrices. The matrix p-norm of an arbitrary matrix $\mathbf{A}$, denoted $\|\mathbf{A}\|_p$, is defined as:

$$\|\mathbf{A}\|_p = \sup_{\mathbf{x} \neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}.$$

Remember that the difference between the supremum and the maximum is that the maximum must be an element of the set while the supremum need not to be. More specifically, if $\mathcal{X}$ is an ordered set, and $\mathcal{S}$ is a subset, then s_0 is the *supremum* of $\mathcal{S}$ iff:

1. $s \leq s_0, \forall s \in \mathcal{S}$
2. if $x \in \mathcal{X}$ and $s \leq x, \forall s \in \mathcal{S}$, then $s_0 \leq x$

On the other hand, an element m is the *maximum* of $\mathcal{S}$ iff:

1. $s \leq m, \forall s \in \mathcal{S}$

2. $m \in \mathcal{S}$

Considering that a property of a vector norm is $\|c\mathbf{x}\|_p = |c|\|\mathbf{x}\|_p$, for any scalar c, we choose c such that $\|\mathbf{x}\|_p = 1$. So the following equivalent statement defines the matrix p-norm.

Definition 2 (Matrix p-norm)**.** *The* matrix p-norm*, where $p \in \mathbb{R}$, is defined as:*

$$\|\mathbf{A}\|_p = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|_p$$

Some geometric intuition about the matrix p-norm can be seen in [2].

As a side note, the main difference between a norm and a distance is that one can consider the norm of only one element, while a distance needs at least two elements.

2.2 Matrix Frobenius Norm

Definition 3 (Frobenius Norm)**.** *The* Frobenius norm *is the 2-norm of the vector obtained by concatenating the rows (or equivalently the columns) of the matrix $\mathbf{A}$:*

$$\|\mathbf{A}\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} |a_{ij}|^2 \right)^{1/2}$$

From the previous definition, the Frobenius norm can also be obtained by rearranging the square of the norm in the following way:

$$\|\mathbf{A}\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} |a_{ij}|^2 = \mathrm{Tr}(\mathbf{A}^T\mathbf{A}) = \mathrm{Tr}(\mathbf{A}\mathbf{A}^T)$$

Where we used the following property of traces: Tr($\mathbf{ABC}$) = Tr($\mathbf{CAB}$) = Tr($\mathbf{BCA}$)

Property 4. *Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{x} \in \mathbb{R}^n$ and let $\mathbf{P} \in \mathbb{R}^{m\times m}$, $\mathbf{Q} \in \mathbb{R}^{n\times n}$ be two orthogonal matrices. Then, the following holds:*

- *Matrix norms induced by vector norms:* $\|\mathbf{A}\mathbf{x}\|_p \leq \|\mathbf{A}\|_p\|\mathbf{x}\|_p$
- *Orthogonal matrices preserve the Frobenius norm:* $\|\mathbf{PAQ}\|_F = \|\mathbf{A}\|_F$
- *Orthogonal matrices preserve the 2-norm:* $\|\mathbf{PAQ}\|_2 = \|\mathbf{A}\|_2$

Property 5. *For any $\mathbf{A} \in \mathbb{R}^{m\times n}$, we have $\|\mathbf{A}\|_2 = \sigma_1$, where σ_1 is the largest singular value of $\mathbf{A}$.*

Proof.

$$\begin{aligned}
\|\mathbf{A}\|_2 &= \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|_2 && \text{(definition of the 2-norm)} \\
&= \max_{\|\mathbf{x}\|=1} \|\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{x}\|_2 && \text{(replacing } \mathbf{A} \text{ by its SVD)} \\
&= \max_{\|\mathbf{x}\|=1} \|\mathbf{\Sigma}\mathbf{V}^T\mathbf{x}\|_2 && (\mathbf{U} \text{ is orthogonal and does not change the norm}) \\
&= \max_{\|\mathbf{y}\|=1} \|\mathbf{\Sigma}\mathbf{y}\|_2 && \text{(replacing } \mathbf{V}^T\mathbf{x} \text{ by } \mathbf{y}) \\
\mathbf{y}^T &= \begin{bmatrix} 1 & 0 & 0 & \cdots \end{bmatrix} && \text{(since we want to maximize, and } \mathbf{\Sigma} \text{ is an ordered diagonal matrix, we set } \mathbf{y}_1 = 1) \\
\|\mathbf{A}\|_2 &= \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix} = \sigma_1
\end{aligned}$$

□

3 Low Rank Approximation

Low rank approximation is a minimization problem with a cost function that measures the difference between a given matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and an approximating matrix with reduced rank. This minimization problem has an analytical solution in terms of the singular value decomposition.

Theorem 6 (Eckart-Young-Mirsky). *Let $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathbf{T}} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$ be the SVD of $\mathbf{A}$, where*

- $\sigma_i \in \mathbb{R}, \mathbf{u}_i \in \mathbb{R}^m, \mathbf{v}_i \in \mathbb{R}^n$
- $r = rank(\mathbf{A})$ *and* $\sigma_1 > \sigma_2 > ... > \sigma_r > 0$ *are the singular values of* $\mathbf{A}$
- $\mathbf{U} \in \mathbb{R}^{m \times m}$, $\mathbf{V} \in \mathbb{R}^{n \times n}$ *are orthogonal matrices.*

Now, let $\mathbf{A}_k = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T$, where only the first k terms are kept from the sum defining $\mathbf{A}$. Then

$$\min_{\mathbf{X} \text{ s.t. } \operatorname{rank}(\mathbf{X}) \leq k} \|\mathbf{A} - \mathbf{X}\|_F = \|\mathbf{A} - \mathbf{A}_k\|_F$$

Note that this is a convex optimization problem over a non convex set. Also, the same result holds for the 2-norm:

$$\min_{\mathbf{X} \text{ s.t. } \operatorname{rank}(\mathbf{X}) \leq k} \|\mathbf{A} - \mathbf{X}\|_2 = \|\mathbf{A} - \mathbf{A}_k\|_2$$

Proof. We show the result for the 2-norm. We start by showing that $\|\mathbf{A}\| = \sigma_{k+1}$. Let $\mathbf{D}_k$ be the diagonal matrix with diagonal entries $0, \cdots, 0, \sigma_{k+1}, \cdots, \sigma_r, 0, \cdots, 0$.

$$\mathbf{D}_k = \begin{bmatrix} 0 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 0 & & & & & & \\ & & & \sigma_{k+1} & & & & & \\ & & & & \ddots & & & & \\ & & & & & \sigma_r & & & \\ & & & & & & 0 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 0 \end{bmatrix}$$

We have

$$\begin{aligned}
\|\mathbf{A} - \mathbf{A}_k\|_2 &= \|\sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T - \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T\|_2 && \text{(Definition of } \mathbf{A}_k\text{)} \\
&= \|\sum_{i=k+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T\|_2 \\
&= \|\mathbf{U}\mathbf{D}_k\mathbf{V}^T\|_2 && \text{(Definition of } \mathbf{D}_k\text{)} \\
&= \|\mathbf{D}_k\|_2 && \text{(By } \textbf{Property 4}\text{, since } \mathbf{U} \text{ and } \mathbf{V} \text{ are orthogonal)} \\
&= \sigma_{k+1} && \text{(By } \textbf{Property 5}\text{)}
\end{aligned}$$

We want to show that for any matrix $\mathbf{B} = \mathbf{X}\mathbf{Y}$, where $r = rank(\mathbf{B})$, $\mathbf{X} \in \mathbb{R}^{m \times r}, \mathbf{Y} \in \mathbb{R}^{k \times n}$, $\mathbf{A}_k$ will always be closer to $\mathbf{A}$ than $\mathbf{B}$ with respect to the matrix 2-norm.

$$\|\mathbf{A} - \mathbf{B}\|_2 \geq \|\mathbf{A} - \mathbf{A}_k\|_2 \qquad \text{(Statement to prove)}$$

Let $\mathbf{V}_{k+1} = \begin{bmatrix} | & & | \\ \mathbf{v}_1 & \cdots & \mathbf{v}_{k+1} \\ | & & | \end{bmatrix}$, Where $\mathbf{v}_1, \cdots, \mathbf{v}_{k+1}$ are the eigenvectors associated with the top $k+1$ singular values.
By the Rank-Nullity Theorem:

$$\dim \mathcal{N}(\mathbf{B}) = n - \operatorname{rank}(\mathbf{B})$$

hence,

$$\dim \mathcal{N}(\mathbf{B}) + \dim \mathcal{R}(\mathbf{V_{k+1}}) > n$$

Which implies that $\mathcal{N}(\mathbf{B}) \cap \mathcal{R}(\mathbf{V_{k+1}}) \neq \{\mathbf{0}\}$. Then, we can take a unit vector $\mathbf{x} \in \mathcal{N}(\mathbf{B}) \cap \mathcal{R}(\mathbf{V_{k+1}})$. We have

$$\begin{aligned}
\|\mathbf{A}-\mathbf{B}\|_2^2 &\geq \|(\mathbf{A}-\mathbf{B})\mathbf{x}\|_2^2 \\
&= \|\mathbf{A}\mathbf{x}\|_2^2 && \text{(Since } \mathbf{x} \text{ is in } \mathcal{N}(\mathbf{B}), \text{ then } \mathbf{B}\mathbf{x} = \mathbf{0}) \\
&= \|\mathbf{U}\mathbf{D}\mathbf{V}^T\mathbf{x}\|_2^2 && \text{(SVD of } \mathbf{A}.) \\
&= \|\mathbf{D}\mathbf{V}^T\mathbf{x}\|_2^2 && \text{(By } \textbf{Property 4}, \text{ since } \mathbf{U} \text{ is orthogonal)} \\
&= \sum_{i=1}^{r} \sigma_i^2 \langle \mathbf{v}_i, \mathbf{x}\rangle^2 \\
&= \sum_{i=k+1}^{r} \sigma_i^2 \langle \mathbf{v}_i, \mathbf{x}\rangle^2 && (\mathbf{x} \in \mathcal{R}(\mathbf{V_{k+1}}), \text{ which implies that } \mathbf{x} \text{ is orthogonal to every } \mathbf{v}_i \text{ for } i < k+1) \\
&\geq \sigma_{k+1}^2 \sum_{i=k+1}^{r} \langle \mathbf{v}_i, \mathbf{x}\rangle^2 && (\sigma_{k+1} \text{ is the largest singular value of } \mathbf{D}) \\
&= \sigma_{k+1}^2 \|\mathbf{V}_{k+1}^T\mathbf{x}\|_2^2 && \text{(Definition of 2-norm)} \\
&= \sigma_{k+1}^2 \|\mathbf{V}^T\mathbf{x}\|_2^2 && (\mathbf{V}_{k+1}^T\mathbf{x} = \mathbf{V}^T\mathbf{x} \text{ because } \mathbf{x} \in \mathcal{R}(\mathbf{V_{k+1}})) \\
&= \sigma_{k+1}^2 \|\mathbf{x}\|_2^2 && \text{(By } \textbf{Property 4}, \text{ since } \mathbf{V} \text{ is orthogonal)} \\
&= \sigma_{k+1}^2 && (\mathbf{x} \text{ is of unit length}) \\
&= \|\mathbf{A}-\mathbf{A}_k\|_2^2 && \text{(As shown in the first part of the proof)}
\end{aligned}$$

By taking the square root on each side of the inequality, we obtain that for any matrix $\mathbf{B} = \mathbf{X}\mathbf{Y}$:

$$\|\mathbf{A}-\mathbf{B}\|_2 \geq \|\mathbf{A}-\mathbf{A}_k\|_2$$

□

4 Variational Characterization of Eigenvalues of Symmetric Matrices

Definition 7 (Rayleigh Quotient)**.** *Let* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *be symmetric, then the* Rayleigh Quotient *is the ratio*

$$\frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$$

The quotient is independent of the scale of $\mathbf{x}$ *since the denominator is the squared norm of* $\mathbf{x}$*.*

Theorem 8 (Rayleigh-Ritz). *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be symmetric. The solution to maximizing (resp. minimizing) the Rayleigh-Ritz Quotient for $\mathbf{x} \neq \mathbf{0}$ is given by the largest (resp. smallest) eigenvalue of $\mathbf{A}$:*

$$\max_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \max_{\|\mathbf{x}\|_2=1} \mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda_{max}(\mathbf{A}) \tag{1}$$

$$\min_{\mathbf{x}\neq\mathbf{0}} \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} = \min_{\|\mathbf{x}\|_2=1} \mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda_{min}(\mathbf{A}) \tag{2}$$

Moreover, if $\mathbf{v}_1, \cdots, \mathbf{v}_k$ are the eigenvectors corresponding to the top k eigenvalues $\lambda_1, \cdots, \lambda_k$ of $\mathbf{A}$, then

$$\max_{\substack{\|\mathbf{x}\|_2=1 \\ \mathbf{x}\in \mathrm{span}(\mathbf{v}_1,\cdots,\mathbf{v}_k)^\perp}} \mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda_{k+1}(\mathbf{A}) \tag{3}$$

where the maximum is obtained by letting $\mathbf{x} = \mathbf{v}_{k+1}$. The constraint $\mathbf{x} \in \mathrm{span}(\mathbf{v}_1, \cdots, \mathbf{v}_k)^\perp$ means that $\mathbf{x}$ must be orthogonal to the first k eigenvectors of $\mathbf{A}$. Since $\mathbf{A}$ is assumed to be symmetric, all its eigenvectors are orthogonal.

Proof. (1) Let $\mathbf{A} = \mathbf{V}\mathbf{D}\mathbf{V}^T$ be the eigendecomposition of $\mathbf{A}$, where $\mathbf{v}_1, \cdots, \mathbf{v}_n$ are the eigenvectors of $\mathbf{A}$ and $\mathbf{V}$ is an orthogonal matrix constructed as:

$$\mathbf{V} = \begin{bmatrix} | & & | \\ \mathbf{v}_1 & \cdots & \mathbf{v}_n \\ | & & | \end{bmatrix} \in \mathbb{R}^{n\times n}$$

Let $\mathbf{x} \in \mathbb{R}^n$ be of unit norm ($\|\mathbf{x}\|_2 = 1$) and let $\mathbf{y}$ be a linear combination of the eigenbasis of $\mathbf{A}$, such that $\mathbf{y} = \mathbf{V}^T\mathbf{x} \in \mathbb{R}^n$. Then we can derive the following inequality:

$$\begin{aligned}
\mathbf{x}^T\mathbf{A}\mathbf{x} &= \mathbf{x}^T\mathbf{V}\mathbf{D}\mathbf{V}^T\mathbf{x} && \text{(Eigendecomposition of } \mathbf{A}\text{)} \\
&= \mathbf{y}^T\mathbf{D}\mathbf{y} && \text{(Definition of } \mathbf{y}\text{)} \\
&= \sum_{i=1}^n \lambda_i \mathbf{y}_i^2 && \text{(Where } \lambda_i \text{ are the diagonal elements of } \mathbf{D}\text{)} \\
&\leq \lambda_{max}(\mathbf{A}) \sum_{i=1}^n \mathbf{y}_i^2 && (\ \lambda_{max}(\mathbf{A}) \geq \lambda_i\ \forall i\ ) \\
&= \lambda_{max}(\mathbf{A})\|\mathbf{y}\|_2^2 \\
&= \lambda_{max}(\mathbf{A}) && (\ \|\mathbf{x}\|_2 = 1, \text{ then } \|\mathbf{y}\|_2 = \|\mathbf{V}^T\mathbf{x}\|_2 = \|\mathbf{x}\|_2 = 1, \\
& && \text{by } \textbf{Property 4}\text{, since } \mathbf{V} \text{ is orthogonal)}
\end{aligned}$$

□

Proof. (2) We are using the same matrices $\mathbf{A}$ and $\mathbf{V}$, vector $\mathbf{x}$ and $\mathbf{y}$ as in (1). Then we can derive the following inequality:

$$\begin{aligned}
\mathbf{x}^T\mathbf{A}\mathbf{x} &= \mathbf{x}^T\mathbf{V}\mathbf{D}\mathbf{V}^T\mathbf{x} && \text{(Eigendecomposition of } \mathbf{A}\text{)} \\
&= \mathbf{y}^T\mathbf{D}\mathbf{y} && \text{(Definition of } \mathbf{y}\text{)} \\
&= \sum_{i=1}^n \lambda_i \mathbf{y}_i^2 && \text{(Where } \lambda_i \text{ are the diagonal elements of } \mathbf{D}\text{)} \\
&\geq \lambda_{min}(\mathbf{A}) \sum_{i=1}^n \mathbf{y}_i^2 && (\ \lambda_{min}(\mathbf{A}) \leq \lambda_i\ \forall i\ ) \\
&= \lambda_{min}(\mathbf{A})\|\mathbf{y}\|_2^2 \\
&= \lambda_{min}(\mathbf{A}) && (\ \|\mathbf{x}\|_2 = 1, \text{ then } \|\mathbf{y}\|_2 = \|\mathbf{V}^T\mathbf{x}\|_2 = \|\mathbf{x}\|_2 = 1, \\
& && \text{by } \textbf{Property 4}\text{, since V is orthogonal)}
\end{aligned}$$

□

Proof. (3) We will split the matrix $\mathbf{V}$ in two partitions such that $\mathbf{V} = \begin{bmatrix}\mathbf{V}_1 & \mathbf{V}_2\end{bmatrix}$, where $\mathbf{V}_1 \in \mathbb{R}^{n\times k}$ represents the top k eigenvectors of $\mathbf{A}$ and $\mathbf{V}_2 \in \mathbb{R}^{n\times n-k}$ represents the last $n-k$ eigenvectors of $\mathbf{A}$. If $\mathbf{x} \in \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_k)^\perp$, then $\mathbf{x}$ is orthogonal to all vectors in $\mathbf{V}_1$ and, most importantly, $\mathbf{x}$ is in the range of $\mathbf{V}_2$ ($\mathbf{x} \in \mathcal{R}(\mathbf{V}_2)$). Then, we can write

$$\begin{aligned}\mathbf{x}^T\mathbf{A}\mathbf{x} &= \mathbf{x}^T\mathbf{V}_2\mathbf{V}_2^T\mathbf{A}\mathbf{V}_2\mathbf{V}_2^T\mathbf{x} && \text{(Replace each } \mathbf{x} \text{ by its projection onto } \mathcal{R}(\mathbf{V}_2), \text{ which}\\ &&& \text{does not change } \mathbf{x} \text{ since it is initially assumed that } \mathbf{x} \in \mathcal{R}(\mathbf{V}_2)\text{)} \quad (4)\\ &= \mathbf{x}^T\mathbf{V}_2\mathbf{V}_2^T\mathbf{V}\mathbf{D}\mathbf{V}^T\mathbf{V}_2\mathbf{V}_2^T\mathbf{x} && \text{(Eigendecomposition of } \mathbf{A})\end{aligned}$$

Now, using the fact that

$$\mathbf{V}_2^T\mathbf{V} = \mathbf{V}_2^T\begin{bmatrix}\mathbf{V}_1 & \mathbf{V}_2\end{bmatrix} = \begin{bmatrix}\mathbf{V}_2^T\mathbf{V}_1 & \mathbf{V}_2^T\mathbf{V}_2\end{bmatrix} = \begin{bmatrix}\mathbf{0} & \mathbf{I}\end{bmatrix}$$

it follows that

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\mathbf{V}\underbrace{\begin{bmatrix}0 & & & & & & \\ & \ddots & & & & & \\ & & 0 & & & & \\ & & & \lambda_{k+1} & & & \\ & & & & \ddots & & \\ & & & & & \lambda_n & \end{bmatrix}}_{\widetilde{\mathbf{A}}}\mathbf{V}^T\mathbf{x}$$

Hence,

$$\begin{aligned}\max_{\substack{\|\mathbf{x}\|_2=1\\ \mathbf{x}\in\text{span}(\mathbf{v}_1,\cdots,\mathbf{v}_k)^\perp}} \mathbf{x}^T\mathbf{A}\mathbf{x} &= \max_{\|\mathbf{x}\|_2=1}\mathbf{x}^T\widetilde{\mathbf{A}}\mathbf{x} && \text{(The constraint } \mathbf{x} \in \text{span}(\mathbf{v}_1, \cdots, \mathbf{v}_k)^\perp \text{ is now incorporated into the } \widetilde{\mathbf{A}}.\\ &&& \text{We now have the same maximization problem as in } \textbf{(1)},\\ &&& \text{except that } \widetilde{\mathbf{A}} \text{ replaces } \mathbf{A})\\ &= \lambda_{k+1}(\mathbf{A}) && \text{(By (1), since the largest eigenvalue of } \widetilde{\mathbf{A}} \text{ is } \lambda_{k+1})\end{aligned}$$

□

Theorem 9 (Courant-Fischer / Min-max)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be symmetric, and let $\mathcal{G}\text{r}(\text{k})$ denote the set of all k-dimensional subspaces of $\mathbb{R}^n$. Then,*

$$\min_{\mathcal{U}\in\mathcal{G}\text{r}(\text{n}-\text{k})}\max_{\substack{\mathbf{x}\in\mathcal{U}\\ \|\mathbf{x}\|=1}}\mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda_{k+1}(\mathbf{A})$$

$$\max_{\mathcal{U}\in\mathcal{G}\text{r}(\text{k})}\min_{\substack{\mathbf{x}\in\mathcal{U}\\ \|\mathbf{x}\|=1}}\mathbf{x}^T\mathbf{A}\mathbf{x} = \lambda_{k}(\mathbf{A})$$

5 Trace Maximization

Theorem 10 (Trace maximization)**.** *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a positive semi-definite matrix. We have*

$$\max_{\substack{\mathbf{U}\in\mathbb{R}^{n\times k}\\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}} \operatorname{Tr}(\mathbf{U}^T\mathbf{A}\mathbf{U}) = \sum_{i=1}^{k}\lambda_i(\mathbf{A})$$

Similarly, the following holds for the trace minimization problem:

$$\min_{\substack{\mathbf{U}\in\mathbb{R}^{n\times k}\\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}} \operatorname{Tr}(\mathbf{U}^T\mathbf{A}\mathbf{U}) = \sum_{i=n-k+1}^{n}\lambda_i(\mathbf{A})$$

So, the maximum trace is obtained by the top k eigenvectors of $\mathbf{A}$ and the minimum trace is obtained by the last k eigenvectors of $\mathbf{A}$.

Proof. We want to solve

$$\max_{\substack{\mathbf{u}_i^T\mathbf{u}_i=1\forall i\\ \mathbf{u}_i\mathbf{v}_j=0\forall i\neq j}} \sum_{i=1}^{k}\mathbf{u}_i^T\mathbf{A}\mathbf{u}_i$$

In order to do so, we will use the method of Lagrange multipliers.

$$\begin{aligned}
\mathcal{L} &= \sum_{i=1}^{k}(\mathbf{u}_i^T\mathbf{A}\mathbf{u}_i + \lambda_i(1-\mathbf{u}_i^T\mathbf{u}_i) + \sum_{j\neq i}\gamma_{ij}\mathbf{u}_i^T\mathbf{u}_j) && \text{(we write the lagrangian)}\\
\frac{\partial\mathcal{L}}{\partial\lambda_i} &= 1-\mathbf{u}_i^T\mathbf{u}_i\\
1 &= \mathbf{u}_i^T\mathbf{u}_i && \text{(setting the partial derivative to 0)}\\
\frac{\partial\mathcal{L}}{\partial\gamma_{ij}} &= \mathbf{u}_i^T\mathbf{u}_j\\
0 &= \mathbf{u}_i^T\mathbf{u}_j && \text{(setting the partial derivative to 0)}
\end{aligned}$$

Now, we have showed that the vectors $\mathbf{u}$ are orthonormal.

$$\begin{aligned}
\nabla_{\mathbf{u}_i}(\mathcal{L}) &= \mathbf{A}\mathbf{u}_i - \lambda_i\mathbf{u}_i\\
\mathbf{A}\mathbf{u}_i &= \lambda_i\mathbf{u}_i && \text{(setting the gradient to 0)}
\end{aligned}$$

Hence, $\mathbf{u}_i$ is an eigenvector of $\mathbf{A}$ with eigenvalue λ_i. From the initial formula, we can use the information gained from the method of Lagrange multipliers such that

$$\begin{aligned}
\sum_{i=1}^{k}\mathbf{u}_i^T\mathbf{A}\mathbf{u}_i &= \sum_{i=1}^{k}\mathbf{u}_i^T\lambda_i\mathbf{u}_i && \text{(because } \mathbf{A}\mathbf{u}_i=\lambda_i\mathbf{u}_i)\\
&= \sum_{i=1}^{k}\lambda_i && \text{(because } \mathbf{u}_i^T\mathbf{u}_i=1)
\end{aligned}$$

To summarize, we have showed that one can take the top k eigenvector-eigenvalue pairs to maximize the trace, or the bottom k eigenvector-eigenvalue pairs to minimize the the trace. □

The following minimization offers an alternative proof to the Eckart-Young-Mirsky theorem. Starting from the goal of minimizing the Frobenius norm of the difference between a matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and its orthogonal projection on $\mathbf{U} \in \mathbb{R}^{n\times k}$, we show that the minimization problem can be solved by projecting the columns of $\mathbf{A}$ onto the space

spanned by its top k left singular vectors:

$$
\begin{aligned}
&\min_{\substack{\mathbf{U}\in\mathbb{R}^{n\times k}\\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}} \|\mathbf{A}-\mathbf{U}\mathbf{U}^T\mathbf{A}\|_F^2 \\
&= \min_{\substack{\mathbf{U}\in\mathbb{R}^{n\times k}\\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}} \mathrm{Tr}((\mathbf{A}-\mathbf{U}\mathbf{U}^T\mathbf{A})^T(\mathbf{A}-\mathbf{U}\mathbf{U}^T\mathbf{A})) && \text{(writing the Frobenius norm in terms of trace)} \\
&= \min_{\substack{\mathbf{U}\in\mathbb{R}^{n\times k}\\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}} \mathrm{Tr}((\mathbf{A}^T-\mathbf{A}^T\mathbf{U}\mathbf{U}^T)(\mathbf{A}-\mathbf{U}\mathbf{U}^T\mathbf{A})) \\
&= \min_{\substack{\mathbf{U}\in\mathbb{R}^{n\times k}\\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}} \mathrm{Tr}(\mathbf{A}^T\mathbf{A}-\mathbf{A}^T\mathbf{U}\mathbf{U}^T\mathbf{A}-\mathbf{A}^T\mathbf{U}\mathbf{U}^T\mathbf{A}+\mathbf{A}^T\mathbf{U}\mathbf{U}^T\mathbf{U}\mathbf{U}^T\mathbf{A}) \\
&= \min_{\substack{\mathbf{U}\in\mathbb{R}^{n\times k}\\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}} \mathrm{Tr}(\mathbf{A}^T\mathbf{A}-\mathbf{A}^T\mathbf{U}\mathbf{U}^T\mathbf{A}-\mathbf{A}^T\mathbf{U}\mathbf{U}^T\mathbf{A}+\mathbf{A}^T\mathbf{U}\mathbf{U}^T\mathbf{A}) && \text{(using } \mathbf{U}^T\mathbf{U}=\mathbf{I}\text{)} \\
&= \min_{\substack{\mathbf{U}\in\mathbb{R}^{n\times k}\\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}} \mathrm{Tr}(\mathbf{A}^T\mathbf{A})-\mathrm{Tr}(\mathbf{A}^T\mathbf{U}\mathbf{U}^T\mathbf{A}) && (\mathrm{Tr}(\mathbf{A}+\mathbf{B})=\mathrm{Tr}(\mathbf{A})+\mathrm{Tr}(\mathbf{B})) \\
&= \min_{\substack{\mathbf{U}\in\mathbb{R}^{n\times k}\\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}} \mathrm{Tr}(\mathbf{A}^T\mathbf{A})-\mathrm{Tr}(\mathbf{U}^T\mathbf{A}\mathbf{A}^T\mathbf{U}) && \text{(changing the order of the matrices in the second trace)} \\
&= \min_{\substack{\mathbf{U}\in\mathbb{R}^{n\times k}\\ \mathbf{U}^T\mathbf{U}=\mathbf{I}}} \mathrm{Tr}(\mathbf{A}^T\mathbf{A})-\sum_{i=1}^{k}\mathbf{u}_i^T\mathbf{A}\mathbf{A}^T\mathbf{u}_i && \text{(rewriting the second trace in terms of Rayleigh terms)}
\end{aligned}
$$

The first term $\mathrm{Tr}(\mathbf{A}^T\mathbf{A})$ does not depend on $\mathbf{U}$, so we focus on the second term. To minimize the whole equation, we need to maximize the second term. As seen in the previous proof, to maximize this term, we take the top k eigenvectors of $\mathbf{A}\mathbf{A}^T$. This way, the minimal difference between $\mathbf{A}$ and its orthogonal projection onto $\mathbf{U}$ in the k-dimensional subspace is given by the top k eigenvectors of $\mathbf{A}\mathbf{A}^T$.

References

[1] J. Lambers. Mat610, lecture 2, summer 2009. `http://www.math.usm.edu/lambers/mat610/sum10/lecture2.pdf`. Accessed: 2019-01-25.

[2] J. P. Reilly. Ece lecture notes: Matrix computations for signal processing. `http://www.ece.mcmaster.ca/faculty/reilly/ece712/ch2.pdf`. Accessed: 2019-01-25.