CSE 521: Design and Analysis of Algorithms I **Fall 2021**

Lecture 10: Linear Algebra Background

Lecturer: Shayan Oveis Gharan *11/01/2021*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

Now that we have finished our lecture series on randomized algorithms, we start with a bit of linear algebra review so that we can use these tools in the algorithms we learn next. The book 'Matrix Analysis' by Horn and Johnson is an excellent reference for all the concepts reviewed here.

10.1 Eigenvalues

For a matrix $A \in \mathbb{R}^{n\times n}$, the eigenvalue-eigenvector pair is defined as (λ, x), where

$$Ax = \lambda x.$$

For an indeterminant (variable) x the polynomial $\det(xI - A)$ is called the characteristic polynomial of A. It turns out that the roots of this polynomial are exactly the eigenvalues of A.

Let us justify this fact. If λ is a root of this polynomial it means that $\det(\lambda I - A) = 0$. But that means that columns of the matrix $\lambda I - A$, say $v_1, \ldots, v_n \in \mathbb{R}^n$ are not linearly independent, i.e., there exists coefficients $c_1, \ldots, c_n$ such that

$$c_1 v_1 + c_2 v_2 + \cdots + c_n v_n = 0.$$

Now, the vector $c = (c_1, \ldots, c_n) \in \mathbb{R}^n$ is an eigenvector of $\lambda I - A$ with eigenvalue 0, i.e., $(\lambda I - A)c = 0$, or equivalently,

$$\lambda c = \lambda I c = Ac.$$

So, λ is an eigenvalue of A. Since any degree n polynomial has n roots any square matrix A has exactly n eigenvalues.

Many of our algorithms will deal with the family of symmetric matrices (which we denote by $\mathcal{S}_n$), with special properties of eigenvalues. We start with the fact that a symmetric matrix has real eigenvalues. This means we can order them and talk about the largest/smallest eigenvalues.

10.1.1 Spectral Theorem

Theorem 10.1 (Spectral Theorem)**.** *For any symmetric matrix, there are eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, with corresponding eigenvectors $v_1, v_2, \ldots, v_n$ which are orthonormal (that is, they have unit length measured in the ℓ_2 norm and $\langle v_i, v_j\rangle = 0$ for all i and j). We can then write*

$$M = \sum_{i=1}^{n} \lambda_i v_i v_i^T = V\Lambda V^T. \tag{10.1}$$

where V is the matrix with v_i's arranged as column vectors and Λ is the diagonal matrix of eigenvalues.

The v_i's in the above theorem form a basis for all vectors in $\mathbb{R}^n$. This means that for any vector x we can uniquely write it as

$$x = \sum_{i=1}^{n} \langle v_i, x \rangle v_i.$$

An application of this is being able to write complicated functions of a symmetric matrix in terms of functions of the eigenvalues that is, $f(M) = \sum_{i=1}^{n} f(\lambda_i) v_i v_i^T$ for $M \in \mathcal{S}_n$. For example:

- $M^2 = \sum_{i=1}^{n} \lambda_i^2 v_i v_i^T$.
- $\exp(M) = \sum_{i=1}^{\infty} \frac{A^k}{k!} = \sum_{i=1}^{n} \exp(\lambda_i) v_i v_i^T$
- For an invertible matrix, $M^{-1} = \sum_{i=1}^{n} (\frac{1}{\lambda_i}) v_i v_i^T$.

We say a symmetric matrix M is positive semidefinite (PSD) if all eigenvalues of M are nonnegative. For a positive semidefinite M we can write

$$\sqrt{M} = M^{1/2} = \sum_{i=1}^{n} \sqrt{\lambda_i} v_i v_i^{\intercal}.$$

We usually use the notation $M \succeq 0$ to denote that M is PSD. In particular, any PSD matrix M can be written as AA^T for some matrix A defined above. later we see the converse of this statement is also true.

Two special functions of eigenvalues are the *trace* and *determinant*, described in the next subsection.

10.1.2 Trace, Determinant and Rank

Definition 10.2. *The trace of a square matrix is the sum of its diagonal entries.*

Alternatively, we can say the following:

Lemma 10.3. *The trace of a symmetric matrix $A \in \mathbb{R}^{n \times n}$ is equal to the sum of its eigenvalues.*

Proof 1. By definition of trace,

$$\mathrm{Tr}(A) = \sum_{i=1}^{n} \mathbf{1}_i^T A \mathbf{1}_i,$$

where $\mathbf{1}_i$ is the indicator vector of i, i.e., it is a vector which is equal to 1 in the i-th coordinate and it is 0 everwhere else. Using (10.1) we can write,

$$\begin{aligned}
\mathrm{Tr}(A) &= \sum_{i=1}^{n} \mathbf{1}_i^T \left(\sum_{j=1}^{n} \lambda_j v_j v_j^T \right) \mathbf{1}_i \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_j \mathbf{1}_i^T v_j v_j^T \mathbf{1}_i \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_j \langle \mathbf{1}_i, v_j \rangle^2 \\
&= \sum_{j=1}^{n} \lambda_j \sum_{i=1}^{n} \langle \mathbf{1}_i, v_j \rangle^2 = \sum_{j=1}^{n} \lambda_j.
\end{aligned}$$

The last identity uses the fact that for any vector v_j, $\sum_{i=1}^n \langle \mathbf{1}_i, v_j\rangle^2 = \|v_j\|^2 = 1$, as $\mathbf{1}_1, \dots, \mathbf{1}_n$ form another orthonormal basis of $\mathbb{R}^n$. □

Proof 2. Recall that

$$\det(xI - A) = (x - \lambda_1) \dots (x - \lambda_n)$$

Observe that the coefficient of x^{n-1} in the RHS is equal to $-(\lambda_1 + \dots - \lambda_n)$. So, to prove the claim it is enough to show that the coefficient of x^{n-1} is the negative of the trace of A.

Let us expand $\det(xI - A)$

$$\det(xI - A) = \sum_\sigma \prod_{i=1}^n \operatorname{sgn}(\sigma)(xI - A)_{i,\sigma_i}$$

Observe that for every permutation σ in the RHS either $\sigma_i = i$ for all i or there exists at least two indices i, j such that $\sigma_i \neq i$ and $\sigma_j \neq j$. But the latter case does not give any monomial of degree $n-1$ in x. It can only give monomials of degree at most $n-2$.

Now, consider the terms coming from the identity permutation σ as the coefficient of x^{n-1} comes from this permutation. It follows that such a permutaiton has sign $+1$. So we just need to figure out the coefficient of x^{n-1} coming from the product of diagonal entries of the matrix $xI - A$,

$$\prod_{i=1}^n (xI - A)_{i,i} = \prod_{i=1}^n (x - A_{i,i})$$

but that is exactly the negative of the sum of diagonal entries of A.

Note that we did not use eigenvectors of A in this proof. So, unlike proof 1 and proof 3, this proof works out even if A is *not a symmetric matrix.* □

Proof 3*:* Recall the *cyclic permutation* property of trace is that

$$\operatorname{Tr}(ABC) = \operatorname{Tr}(BCA) = \operatorname{Tr}(CAB)$$

This is derived simply from definition. Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of A with corresponding eigenvalues $v_1, \dots, v_n$. We have

$$\begin{aligned}
\operatorname{Tr}(A) &= \operatorname{Tr}\left(\sum_{i=1}^n \lambda_i v_i v_i^T\right) \\
&= \sum_{i=1}^n \operatorname{Tr}(\lambda_i v_i v_i^T) \\
&= \sum_{i=1}^n \lambda_i \operatorname{Tr}(\langle v_i, v_i^T\rangle) \\
&= \sum_{i=1}^n \lambda_i.
\end{aligned}$$

In the last identity we used that $\|v_i\| = 1$ for all i. □

Lemma 10.4. *The determinant of a matrix is the product of its eigenvalues.*

To prove the lemma once again we use the characteristic polynomial $\det(xI - A) = (x - \lambda_1) \dots (x - \lambda_n)$. So, if we plug in $x = 0$ we obtain, $\det(-A) = \prod_{i=1}^n -\lambda_i$ or equivalently that $\det(A) = \prod_{i=1}^n \lambda_i$.

10.2 Rayleigh Quotient

Let A be a symmetric matrix. The Rayleigh coefficient gives a characterization of all eigenvalues (and eigenvectors of A) in terms of the solution to optimization problems. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of A. Then,

$$\lambda_1(A) = \max_{\|x\|_2=1} x^T A x = \max_x \frac{x^T A x}{x^T x} \tag{10.2}$$

Let x_1 be the optimum vector in the above. It follows that x_1 is the eigenvector of A corresponding to λ_1. Then,

$$\lambda_2(A) = \max_{x:\langle x, x_1\rangle=0, \|x\|=1} x^T A x$$

And so on, the third eigenvector is the vector maximizing the quadratic form $x^T A x$ over all vectors that orthogonal to the first two eigenvectors. Similarly, we can write

$$\lambda_n(A) = \min_{\|x\|_2=1} x^T A x$$

Let us derive, Equation (10.2). Note that $f(x) = x^T A x$ is a continuous function and $\{x \mid \|x\|_2 = 1\}$ is a compact set. So by Weierstrass Theorem, the maximum is attained. Now we diagonalize A using Equation (10.1) as $A = \sum_{i=1}^n \lambda_i v_i v_i^T$ and multiply on either side by x to get the following chain of equalities:

$$\begin{aligned} x^T A x &= x^T \left(\sum_{i=1}^n \lambda_i v_i v_i^T \right) x \\ &= \sum_{i=1}^n \lambda_i x^T v_i v_i^T x \\ &= \sum_{i=1}^n \lambda_i \langle x, v_i \rangle^2. \end{aligned} \tag{10.3}$$

Since $\|x\| = 1$ and $v_1, \ldots, v_n$ form an orthonormal basis of $\mathbb{R}^n$, $\sum_{i=1}^n \langle v_i, x\rangle^2 = \|x\|^2 = 1$. Therefore, (10.3) is maximized when $\langle x, v_1\rangle = 1$ and the rest are 0. This means the vector x for which this optimum value is attained is v_1 as desired.

In the same way, we can also get the characterization for the minimum eigenvalue.

Positive (Semi) Definite Matrices An equivalent definition for a symmetric matrix $A \in \mathbb{R}^{n\times n}$ to be PSD is that

$$x^T A x \geq 0$$

for all $x \in \mathbb{R}^n$. It follows by the Rayleigh quotient that $x^\intercal A x \geq 0$ for all vectors $x \in \mathbb{R}^n$ if and only if all eigenvalues of A are nonnegative.

10.3 Singular Value Decomposition

Of course not every matrix is unitarily diagonalizable. In fact non-symmetric matrices may not have real eigenvalues the space of eigenvectors is not necessarily orthonormal.

Instead, when dealing with a non-symmetric matrix, first we turn it into a symmetric matrix and then we apply the spectral theorem to that matrix. This idea is called the Singular Value Decomposition (SVD). For any matrix $A \in \mathbb{R}^{m\times n}$ (with $m \leq n$) can be written as

$$A = U\Sigma V^T = \sum_{i=1}^{m} \sigma_i u_i v_i^T \tag{10.4}$$

where $\sigma_1 \geq \cdots \geq \sigma_m \geq 0$ are the singular values of A, $u_1, \ldots, u_m$ are orthonormal and are called the left singular vectors of A and $v_1, \ldots, vm \in \mathbb{R}^n$ are orthonormal and are call the right singular vectors of A. To construct this decomposition we need to apply the spectral theorem to the matrix A^TA. Observe that if the above identity holds then

$$A^TA = \sum_{i=1}^{m} \sigma_i v_i u_i^T \sum_{j=1}^{n} \sigma_j u_j v_j^T = \sum_{i=1}^{n} \sigma_i^T v_i v_i^T$$

where we used that $\langle u_i, u_j\rangle$ is 1 if $i = j$ and it is zero otherwise. Therefore, $v_1, \ldots, v_m$ are in fact the eigenvectors of A^TA and $\sigma_1^2, \ldots, \sigma_m^2$ are the eigenvalues of A^TA. By a similar argument it follows that $u_1, \ldots, u_m$ are eigenvectors of AA^T and $\sigma_1^2, \ldots, \sigma_m^2$ are its eigenvalues.

Note that both matrices AA^T and A^TA are symmetric PSD matrices. In the matrix form the above identities can be written as

$$A^TA = V\Sigma U^T U\Sigma V^T = V\Sigma^2 V^T = [V \;\; \tilde{V}] \begin{bmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{bmatrix} [V \;\; \tilde{V}]^T \tag{10.5}$$

$$AA^T = U\Sigma V^T V\Sigma U^T = U\Sigma^2 U^T = [U \;\; \tilde{U}] \begin{bmatrix} \Sigma^2 & 0 \\ 0 & 0 \end{bmatrix} [U \;\; \tilde{U}]^T \tag{10.6}$$

where $\tilde{V}, \tilde{U}$ are any matrices for which $[V \;\; \tilde{V}]$ and $[U \;\; \tilde{U}]$ are orthonormal. The righthand expressions are eigenvalue decompositions of A^TA and AA^T.

To summarize,

- The singular values σ_i are the squareroots of eigenvalues of A^TA and AA^T, that is, $\sigma_i(A) = \sqrt{\lambda_i(A^TA)} = \sqrt{\lambda_i(AA^T)} \quad (\lambda_i(A^TA) = \lambda_i(AA^T) = 0$ for $i > r)$.
- The left singular vectors $u_1, \ldots, u_r$ are the eigenvectors of AA^T the right singular vectors $V = [v_1, \ldots, v_m]$ are the eigenvectors of A^TA.

In general, computing the singular value decomposition can take $\mathcal{O}(n^3)$ time.

10.4 Matrix Norms

Any matrix $A \in \mathbb{R}^{n\times n}$ can be thought of as a vector of n^2 dimensions. Therefore, we can measure the 'size' of a matrix using matrix norms. For a function $\|.\| : \mathbb{R}^{n\times n} \to \mathbb{R}$ to be a matrix norm, it must satisfy the properties of non-negativity (and zero only when the argument is zero), homogeneity, triangle inequality and submultiplicativity. We list below a few important matrix norms that we'll repeatedly encounter:

Frobenius norm:

$$\|A\|_F = |\operatorname{Tr}(AA^T)|^{1/2} = (\sum_{i,j=1}^{n} a_{ij}^2)^{1/2}. \tag{10.7}$$

The Frobenius norm is just the Euclidean norm of matrix A thought of as a vector. As we just saw in Section 10.3,

$$\mathrm{Tr}(AA^T) = \sum_{i=1}^{n} \lambda_i(AA^T) = \sum_{i=1}^{n} \sigma_i(A)^2,$$

therefore this gives us an important alternative characterization of Frobenius norm:

$$\|A\|_F = (\sum_{i=1}^{n} \sigma_i(A)^2)^{1/2}. \tag{10.8}$$

Operator norm: The operator norm $\|.\|_2$ is defined as

$$\|A\|_2 = \max_{\|x\|=1} \|Ax\| = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} \tag{10.9}$$

It follows by the Rayleigh-Ritz characterization that

$$\max_x \frac{\|Ax\|}{\|x\|} = \sqrt{\max_x \frac{\|Ax\|^2}{\|x\|^2}} = \sqrt{\max_x \frac{x^T A^T Ax}{x^T x}} = \sqrt{\lambda_{max}(A^T A)} = \sigma_{\max}(A).$$

10.5 A Geometric Intuition of a Matrix/Operator

Let $M = \sum_{i=1}^n \lambda_i v_i v_i^\intercal$ be a symmetric matrix. We can represent M geometrically by a ellipse defined with the following equation:

$$x \in \mathbb{R}^n : x^\intercal M^{-2} x = 1.$$

The axis of this ellipse correspond to eigenvectors of M and length of the i-th axis is equal to the i-th largest eigenvalue in absolute value. This is because if we let $x = \lambda_i v_i$, then

$$x^\intercal M^{-2} x = (\lambda_i v_i)^\intercal M^{-2} (\lambda_i v_i) = (\lambda_i v_i)^\intercal \frac{1}{\lambda_i} v_i = 1,$$

where we used the v_i is an eigenvector of M^{-2} with corresponding eigenvalue of $1/\lambda_i^2$. So, this says that along the v_i direction the farthest point is exactly $|\lambda_i|$ away from the origin. In particular, the farthest point of the ellipse is $\max_i |\lambda_i|$ away from the origin.

We can understand this ellipse differently: It can be seen as the image of the unit sphere around the origin with respect to operator M. Recall that the set points on the unit sphere are all $x \in \mathbb{R}^n$ such that $\|x\| = 1$. The image with respect to M, is Mx. We claim that for any x such that $\|x\| = 1$, Mx is on the ellipse. It is enough to see

$$(Mx)^\intercal M^{-2} (Mx) = x^\intercal M^\intercal M^{-2} Mx = x^\intercal x \|x\|^2 = 1.$$

So, in particular, the volume of ellipse defined above is equal to the volume of the ball of radius 1 times $\det(M)$.