Linear Algebra Review
(Adapted from Punit Shah’s slides)

Introduction to Machine Learning (CSC 311)

University of Toronto

Fall 2021
Matrix Decomposition

- We can decompose an integer into its prime factors, e.g., $12 = 2 \times 2 \times 3$.

- Similarly, matrices can be decomposed into product of other matrices.

$$A = V \text{diag}(\lambda)V^{-1}$$

- Examples are Eigendecomposition, SVD, Schur decomposition, LU decomposition, ....
Eigenvectors

- An eigenvector of a square matrix \( A \) is a nonzero vector \( v \) such that multiplication by \( A \) only changes the scale of \( v \).

\[
Av = \lambda v
\]

- The scalar \( \lambda \) is known as the eigenvector.

- If \( v \) is an eigenvector of \( A \), so is any rescaled vector \( sv \). Moreover, \( sv \) still has the same eigenvalue. Thus, we constrain the eigenvector to be of unit length:

\[
\|v\|_2 = 1
\]
Characteristic Polynomial (1)

- Eigenvalue equation of matrix $A$.

$$Av = \lambda v$$

$$\lambda v - Av = 0$$

$$(\lambda I - A)v = 0$$

- If nonzero solution for $v$ exists, then it must be the case that:

$$\det(\lambda I - A) = 0$$

- Unpacking the determinant as a function of $\lambda$, we get:

$$P_A(\lambda) = \det(\lambda I - A) = 1 \times \lambda^n + c_{n-1} \times \lambda^{n-1} + \ldots + c_0$$

- This is called the characteristic polynomial of $A$. 
If $\lambda_1, \lambda_2, \ldots, \lambda_n$ are roots of the characteristic polynomial, they are eigenvalues of $A$ and we have $P_A(\lambda) = \prod_{i=1}^{n}(\lambda - \lambda_i)$.

$c_{n-1} = -\sum_{i=1}^{n} \lambda_i = -tr(A)$. This means that the sum of eigenvalues equals to the trace of the matrix.

$c_0 = (-1)^n \prod_{i=1}^{n} \lambda_i = (-1)^n det(A)$. The determinant is equal to the product of eigenvalues.

Roots might be complex. If a root has multiplicity of $r_j > 1$ (This is called the algebraic dimension of eigenvalue), then the geometric dimension of eigenspace for that eigenvalue might be less than $r_j$ (or equal but never more). But for every eigenvalue, one eigenvector is guaranteed.
Example

- Consider the matrix:
  \[ A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \]

- The characteristic polynomial is:
  \[
  \det(\lambda I - A) = \det \begin{bmatrix} \lambda - 2 & -1 \\ -1 & \lambda - 2 \end{bmatrix} = 3 - 4\lambda + \lambda^2 = 0
  \]

- It has roots \( \lambda = 1 \) and \( \lambda = 3 \) which are the two eigenvalues of \( A \).

- We can then solve for eigenvectors using \( A\mathbf{v} = \lambda\mathbf{v} \):
  \[ \mathbf{v}_{\lambda=1} = [1, -1]^\top \quad \text{and} \quad \mathbf{v}_{\lambda=3} = [1, 1]^\top \]
Eigendecomposition

Suppose that $n \times n$ matrix $A$ has $n$ linearly independent eigenvectors $\{v^{(1)}, \ldots, v^{(n)}\}$ with eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$.

- Concatenate eigenvectors (as columns) to form matrix $V$.
- Concatenate eigenvalues to form vector $\lambda = [\lambda_1, \ldots, \lambda_n]^T$.

The eigendecomposition of $A$ is given by:

$$AV = V \text{diag}(\lambda) \implies A = V \text{diag}(\lambda)V^{-1}$$
Symmetric Matrices

- Every symmetric (hermitian) matrix of dimension $n$ has a set of (not necessarily unique) $n$ orthogonal eigenvectors. Furthermore, all eigenvalues are real.

- Every real symmetric matrix $A$ can be decomposed into real-valued eigenvectors and eigenvalues:

$$ A = Q\Lambda Q^\top $$

- $Q$ is an orthogonal matrix of the eigenvectors of $A$, and $\Lambda$ is a diagonal matrix of eigenvalues.

- We can think of $A$ as scaling space by $\lambda_i$ in direction $v^{(i)}$. 

![Plot of unit vectors $u \in \mathbb{R}^2$ (circle) and Plot of vectors $Au$ (ellipse)](image)
Eigendecomposition is not Unique

- Decomposition is not unique when two eigenvalues are the same.

- By convention, order entries of $\Lambda$ in descending order. Then, eigendecomposition is unique if all eigenvalues have multiplicity equal to one.

- If any eigenvalue is zero, then the matrix is singular. Because if $v$ is the corresponding eigenvector we have: $Av = 0v = 0$. 

Positive Definite Matrix

- If a symmetric matrix $A$ has the property:

  $x^\top A x > 0$  for any nonzero vector $x$

  Then $A$ is called **positive definite**.

- If the above inequality is not strict then $A$ is called **positive semidefinite**.

- For positive (semi)definite matrices all eigenvalues are positive (nonnegative).
Singular Value Decomposition (SVD)

- If $A$ is not square, eigendecomposition is undefined.
- SVD is a decomposition of the form $A = UDV^\top$.
- SVD is more general than eigendecomposition.
- Every real matrix has a SVD.
Write $A$ as a product of three matrices: $A = UDV^\top$.

If $A$ is $m \times n$, then $U$ is $m \times m$, $D$ is $m \times n$, and $V$ is $n \times n$.

$U$ and $V$ are orthogonal matrices, and $D$ is a diagonal matrix (not necessarily square).

Diagonal entries of $D$ are called singular values of $A$.

Columns of $U$ are the left singular vectors, and columns of $V$ are the right singular vectors.
SVD can be interpreted in terms of eigendecomposition.

Left singular vectors of $A$ are the eigenvectors of $AA^\top$.

Right singular vectors of $A$ are the eigenvectors of $A^\top A$.

Nonzero singular values of $A$ are square roots of eigenvalues of $A^\top A$ and $AA^\top$.

Numbers on the diagonal of $D$ are sorted largest to smallest and are non-negative ($A^\top A$ and $AA^\top$ are semipositive definite.).
Matrix norms

- We may define norms for matrices too. We can either treat a matrix as a vector, and define a norm based on an entrywise norm (example: Frobenius norm). Or we may use a vector norm to “induce” a norm on matrices.

- Frobenius norm:
  \[ \|A\|_F = \sqrt{\sum_{i,j} a_{i,j}^2}. \]

- Vector-induced (or operator, or spectral) norm:
  \[ \|A\|_2 = \sup_{\|x\|_2=1} \|Ax\|_2. \]
SVD Optimality

- Given a matrix $A$, SVD allows us to find its “best” (to be defined) rank-$r$ approximation $A_r$.
- We can write $A = U D V^\top$ as $A = \sum_{i=1}^{n} d_i u_i v_i^\top$.
- For $r \leq n$, construct $A_r = \sum_{i=1}^{r} d_i u_i v_i^\top$.
- The matrix $A_r$ is a rank-$r$ approximation of $A$. Moreover, it is the best approximation of rank $r$ by many norms:
  - When considering the operator (or spectral) norm, it is optimal. This means that $\|A - A_r\|_2 \leq \|A - B\|_2$ for any rank $r$ matrix $B$.
  - When considering Frobenius norm, it is optimal. This means that $\|A - A_r\|_F \leq \|A - B\|_F$ for any rank $r$ matrix $B$. One way to interpret this inequality is that rows (or columns) of $A_r$ are the projection of rows (or columns) of $A$ on the best $r$ dimensional subspace, in the sense that this projection minimizes the sum of squared distances.