Deep Learning Book Notes
Chapter 2: Linear Algebra

Compiled By: Abhinaba Bala, Dakshit Agrawal, Mohit Jain

Section 2.1: Scalars, Vectors, Matrices and Tensors

- Scalar
    - Single Number
    - Lowercase names in *italic* typeface
- Vector
    - Array of numbers
    - Lowercase names in **bold** typeface
    - Each element denoted in *italic* typeface with corresponding subscript
    - Can be visualized as point, with each element giving the coordinate along a different axis
- Matrices
    - 2-D array of numbers
    - Uppercase names in **bold** typeface
    - Each element denoted by name of matrix in *italic* typeface with corresponding indices
    - Can be visualized as a collection of vectors. Vectors taken row-wise constitute the **row space**, whereas vectors taken column-wise constitute the **column space**.
- Tensors
    - 3 or more dimensional array of numbers
- Transpose of a matrix is the mirror image of the matrix across the main diagonal
- Addition is done by adding corresponding elements.
- While implementing matrix operations, addition of a matrix and a vector can take place by adding the vector to each row or column of the matrix, and is known as **broadcasting**.

Section 2.2: Multiplying Matrices and Vectors

- Multiplication of a matrix by a vector can be visualized as transforming the vector according to the matrix (i.e. the matrix denotes a transformation). See the following 3Blue1Brown video for more details:

  https://www.youtube.com/watch?v=kYB8IZa5AuE&index=4&list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab

  Also, for visualization and intuition one may refer to "Matrices as Transformation" by KhanAcademy:

  https://www.khanacademy.org/math/precalculus/precalc-matrices/matrices-as-transformations/a/matrices-as-transformations

- Element-wise product of corresponding elements of two matrices is the **Hadamard Product**, denoted by $\boldsymbol{A} \odot \boldsymbol{B}$.
- A **system of linear equations** can be denoted by matrices as $\boldsymbol{Ax} = \mathbf{b}$. Here $\boldsymbol{A} \in \mathbb{R}^{mxn}$ is a known matrix, $\mathbf{b} \in \mathbb{R}^{m}$ is a known vector, and $\mathbf{x} \in \mathbb{R}^{n}$ is a vector of unknown variables.

Section 2.3: Identity and Inverse Matrices

- **Identity Matrix** is a matrix that does not change the vector when multiplied to it.
- **Inverse Matrix** is a matrix that reverses the transformation applied by a matrix $\boldsymbol{A}$, and is denoted by $\boldsymbol{A}^{-1}$. See the following 3Blue1Brown video for more details:

  https://www.youtube.com/watch?v=uQhTuRlWMxw&index=8&list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab

- Given the inverse matrix $\boldsymbol{A}^{-1}$ we can find solutions for different $\mathbf{b}$. However, the inverse matrix should not be used in software applications as there is a loss of precision. Hence, $\boldsymbol{A}^{-1}$ is calculated every time for a new $\mathbf{b}$.

Section 2.4: Linear Dependence and Span

- Linear combination of vectors means a summation of vectors scaled by their respective scalars.

$$\boldsymbol{Ax} = \sum x_i \boldsymbol{A}_{:,i}$$

  If a vector $\mathbf{v}^{(n)}$ from a set of vectors $\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(n)}\}$ can be represented as a linear combination of the other vectors, then the set of vectors is **linearly dependent**. Elsewise, the set of vectors is **linearly independent**.

- The **span** of a set of vectors is the set of all points obtainable by a linear combination of the original vectors. For more details about span and linear combinations, refer to the following video by 3Blue1Brown:

  https://www.youtube.com/watch?v=k7RM-ot2NWY&index=3&list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab

- Adding a vector to a set that is a linear combination of other vectors in the set does not expand the span of the set of vectors.
- The system of linear equations can be thought of as a set of column vectors of $\boldsymbol{A}$. Our objective is to find scalars corresponding to each vector so that they can reach a point specified by the vector **b**.
- A **necessary** condition so that each point **b** in the space $\mathbb{R}^m$ has a solution is for $\mathbf{n \geq m}$. However, even if we have 'n' column vectors to cover 'm' dimensions, some of the vectors may be linearly dependent, causing the corresponding span to cover space less than $\mathbb{R}^m$. Hence a **sufficient** condition is for the 'n' column vectors to have a set of exactly 'm' linearly independent vectors.
- A square matrix having linearly dependent columns is known as **singular**.

Section 2.5: Norms

- L^2 norm is known as the **Euclidean norm**. It's common to use the squared L^2 norm as well.
- L^1 norm is known as the **Manhattan distance**, whose name comes from the way distance is calculated on the grid-like streets of Manhattan, New York.
- If sparsity (some values be exactly 0) or feature extraction is necessary, then L^1 norm is used. If the values are needed to be distributed among the features, then L^2 norm is used.
- Norms of $n < 1$ results in concave surfaces and hence are avoided. Norm of $n = \infty$ is the max function. A beautiful visualization of norm functions can be found in the following video:

  https://www.youtube.com/watch?v=SXEYIGqXSxk

- The **Frobenius norm** is used to measure the size of the matrix and is given by taking the L^2 norm of all the elements in the matrix.

Section 2.6: Special Kinds of Matrices and Vectors

- Diagonal Matrices
  - A matrix $\boldsymbol{D}$ is diagonal if and only if $D_{i,j} = 0$, for all $i \neq j$. It is not necessarily square.
  - To compute $(\text{diag}(\mathbf{v}))\mathbf{x}$, we only need to scale each element x_i by v_i, i.e. $(\text{diag}(\mathbf{v}))\mathbf{x} = \mathbf{v} \circ \mathbf{x}$.
  - The inverse exists only if all the elements are non-zero. To compute the inverse, we only need to invert the individual elements, i.e. $\text{diag}(\mathbf{v})^{-1} = \text{diag}([1/v_1, \ldots, 1/v_n]^T)$.
- A **symmetric matrix** is any matrix that is equal to its own transpose, i.e $\boldsymbol{A} = \boldsymbol{A}^T$.
- Orthogonal Matrix
  - A vector $\mathbf{x}$ and a vector $\mathbf{y}$ are **orthogonal** to each other if $\mathbf{x}^T\mathbf{y} = 0$.
  - A vector that is orthogonal as well as having unit norm are called **orthonormal**.

- **Orthogonal Matrix** is a square matrix whose rows are mutually orthonormal, and whose columns are mutually orthonormal.

$$\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}$$

$$\boldsymbol{A}^{-1} = \boldsymbol{A}^T$$

Therefore orthogonal matrices are of interest because their inverse is very cheap to calculate.

Section 2.7: Eigendecomposition

- A vector **v** that does not change its direction after transformation by a matrix ***A*** is known as an **eigenvector** of matrix ***A***. The factor λ by which the vector is scaled by the transformation is called the **eigenvalue** corresponding to that **eigenvector**.

$$\boldsymbol{A}\mathbf{v} = \lambda\mathbf{v}$$

If **v** is an eigenvector of ***A***, then so is any rescaled vector $s\mathbf{v}$ for $s \in \mathbb{R}$, $s \neq 0$. Moreover, $s\mathbf{v}$ has the same eigenvalue. For this reason, we usually look for unit eigenvectors.

- **Eigendecomposition** of ***A*** is given by $\boldsymbol{A} = \boldsymbol{V}\text{diag}(\boldsymbol{\lambda})\boldsymbol{V}^{-1}$. It's visualization can be interpreted as taking the basis vectors (the vectors relative to which coordinates of a vector are given) to be the eigenvectors. Since during transformation, the eigenvectors scale by corresponding factor λ, any vector represented with the eigenvectors as its basis is scaled by a fixed number according to each basis vector (hence the diagonal matrix). $\boldsymbol{V}^{-1}$ helps to bring the vector back to the original basis vectors (usually $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$). For more details and good visualization, refer to the following video by 3Blue1Brown:

https://www.youtube.com/watch?v=PFDu9oVAE-g&index=14&list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab

- Not every matrix can be decomposed into eigenvalues and eigenvectors. Sometimes the decomposition exists but contains complex numbers. Every real symmetric

matrix can be decomposed into an expression using only real-valued eigenvectors and eigenvalues:

$$\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^{T}$$

Where $\boldsymbol{Q}$ is an orthogonal matrix composed of eigenvectors of $\boldsymbol{A}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix. The eigenvalue $\Lambda_{i,\,i}$ is associated with the eigenvector in column *i* of $\boldsymbol{Q}$, denoted as $\boldsymbol{Q}_{:,\,i}$.

- The eigendecomposition of any real symmetric matrix $\boldsymbol{A}$ may not be unique. If any two or more eigenvectors share the same eigenvalue, then any set of orthogonal vectors lying in their span are also eigenvectors with that eigenvalue. Visualize this as a vector that can be broken down into components of eigenvectors having same eigenvalue, then that vector will be scaled by the same factor and remain at its place (e.g. let eigenvectors be $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$, with their eigenvalue being 2. Then the whole space or span covered by them will be scaled by a factor of 2, with no change in the direction of any vector if $\hat{\mathbf{i}}$ and $\hat{\mathbf{j}}$ are taken as the basis vectors).
- By convention, we usually sort the entries of $\boldsymbol{\Lambda}$ in descending order.
- A matrix is singular if any one of the eigenvalues is zero.
- A **symmetric n x n real matrix** $\boldsymbol{M}$ is said to be **positive definite** if all its eigenvalues are all positive. A matrix whose eigenvalues are all positive or zero values is called **positive semidefinite**. Similarly, the matrix is **negative definite** if all its eigenvalues are all negative, and if all its eigenvalues are negative or zero valued, it is **negative semidefinite**.
- Positive definite matrices guarantee that $\forall \mathbf{x}$, $\mathbf{x}^{T}\boldsymbol{A}\mathbf{x} > 0$. Positive definite matrices guarantee that $\mathbf{x}^{T}\boldsymbol{A}\mathbf{x} = 0 \Rightarrow \mathbf{x} = \mathbf{0}$.

Proof:

Consider a matrix $\boldsymbol{A}$*, which is Positive Definite. As a result,* $\boldsymbol{A}$ *is also a real symmetric matrix. This implies that all of its eigenvalues are real and its eigenvectors are orthonormal. As a result, the eigendecomposition can be written as:*

$$\boldsymbol{A} = \boldsymbol{Q}\Lambda\boldsymbol{Q}^{T}$$

Now,

$\boldsymbol{x^T A x}$

$\Rightarrow \boldsymbol{x^T Q \Lambda Q^T x}$

$\Rightarrow \boldsymbol{(Q^T x)^T \Lambda (Q^T x)}$

Now, $\boldsymbol{Q^T x} = [q_1{}^T \boldsymbol{x} \quad q_2{}^T \boldsymbol{x} \quad . \; . \; . \quad q_n{}^T \boldsymbol{x}]^T$*, where* q_i *are the eigenvectors. Note that* $\boldsymbol{Q^T x}$ *is a column vector of dimension* ***n x 1****. Similarly,* $\boldsymbol{(Q^T x)^T}$ *is a row vector of dimension* ***1 x n****. For simplicity of notation, let,* $\boldsymbol{Q^T x} = [z_1 \;\; z_2 \;\; \ldots \;\; z_n]^T = \boldsymbol{z}$*, where* $z_i = q_i{}^T \boldsymbol{x}$ *are scalars.*

Hence, the expression becomes:

$\boldsymbol{z^T \Lambda z}$

As, Λ *is a diagonal matrix of dimensions* ***n x n****, multiplication with it scales up each column of the preceding matrix by the corresponding diagonal element, i.e. col_1 becomes* λ_1 *col_1 . . . col_n becomes* λ_n *col_n.*

Thus, $\boldsymbol{z^T \Lambda} = [\lambda_1 z_1 \;\; \lambda_2 z_2 \;\; \ldots \;\; \lambda_n z_n] = \boldsymbol{y^T}$ *(let)*

The expression then becomes:

$\boldsymbol{y^T z} = \sum_{i=1}^{n} \lambda_i z_i^2 > 0$

as all $\lambda_i > 0$ *(because* ***A*** *is Positive Definite) and squares of real numbers are positive.*

Hence, $\boldsymbol{x^T A x} > 0$

Section 2.8: Singular Value Decomposition

- The SVD factorizes a matrix into **singular vectors** and **singular values**. Every real matrix has a SVD, but the same is not true for eigendecomposition. For example, if a matrix is not square, the eigendecomposition is not defined, but we can use a singular value decomposition instead.
- The singular value decomposition is defined as follows:

$$\boldsymbol{A = UDV^T}$$

  Suppose ***A*** is an m x n matrix.

  - ***U*** is m x m matrix and its column vectors are known as the **left-singular vectors**.
  - ***V*** is n x n matrix and its column vectors are known as the **right-singular vectors**.
  - ***D*** is m x n matrix and the elements along the diagonal are known as the **singular values** of matrix ***A***.
  - Matrix ***U*** and ***V*** are orthogonal matrices, and ***D*** is a diagonal matrix, not necessarily square.
- SVD visualization is similar to eigendecomposition, but applied to non-square matrices.
- The left-singular vectors of ***A*** are the eigenvectors of $\boldsymbol{AA^T}$.
- The right-singular vectors of ***A*** are the eigenvectors of $\boldsymbol{A^TA}$.
- The nonzero singular values of ***A*** are the square roots of the eigenvalues of $\boldsymbol{AA^T}$ or $\boldsymbol{A^TA}$.
- It is used to partially calculate the inverse matrix of non-square matrices (Moore-Penrose Pseudoinverse).
- **Miscellaneous Extra Thought:** What is the difference between eigendecomposition and singular value decomposition? Answer in the link below.

  https://math.stackexchange.com/questions/320220/intuitively-what-is-the-difference-between-eigendecomposition-and-singular-valu

Section 2.9: The Moore-Penrose Pseudoinverse

- Practical algorithms for computing the pseudoinverse are based on the following formula

$$\boldsymbol{A}^{+} = \boldsymbol{V}\boldsymbol{D}^{+}\boldsymbol{U}^{\mathrm{T}}$$

  - $\boldsymbol{A}^{+}$ is the pseudoinverse of matrix $\boldsymbol{A}$.
  - $\boldsymbol{U}$, $\boldsymbol{D}$ and $\boldsymbol{V}$ are the singular value decomposition of $\boldsymbol{A}$.
  - Pseudoinverse $\boldsymbol{D}^{+}$ of $\boldsymbol{D}$ is obtained by taking the reciprocal of its nonzero elements then taking the transpose of the resulting matrix.
- When $\boldsymbol{A}$ has more columns than rows, then solving a linear equation using the pseudoinverse provides one of the many possible solutions. Specifically, it provides the solution $\mathbf{x} = \boldsymbol{A}^{+}\mathbf{y}$ with minimal Euclidean norm $\|\mathbf{x}\|_2$ among all possible solutions.
- When $\boldsymbol{A}$ has more rows than columns, it is possible for there to be no solution (solution possible only if $\mathbf{b}$ is in span of column vectors of $\boldsymbol{A}$). In this case, using a pseudoinverse gives us the $\mathbf{x}$ for which $\boldsymbol{A}\mathbf{x}$ is as close as possible to $\mathbf{y}$ in terms of Euclidean norm $\|\boldsymbol{A}\mathbf{x} - \mathbf{y}\|_2$. Geometrically, this means projecting $\boldsymbol{b}$ onto the column space of $\boldsymbol{A}$ resulting in a vector $\boldsymbol{p}$ and finding $\boldsymbol{x}^*$, such that $\boldsymbol{A}\boldsymbol{x}^* = \boldsymbol{p}$. Here, $\boldsymbol{x}^*$ is the best possible solution to the equation.

Section 2.10: The Trace Operator

- The trace operator gives the sum of all the diagonal entries of a matrix.
- The Frobenius norm of a matrix can be given as follows:

$$\|A\|_F = \sqrt{\mathrm{Tr}(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})}$$

- The trace of a square matrix composed of many factors is also invariant to moving the last factor into the first position, if the shapes of the corresponding matrices allow the resulting product to be defined:

$$\mathrm{Tr}(\boldsymbol{ABC}) = \mathrm{Tr}(\boldsymbol{CAB}) = \mathrm{Tr}(\boldsymbol{BCA})$$

- A scalar is its own trace: a = Tr(a).
- If **A** is an n*n matrix, then the sum of the n eigenvalues of **A** is the trace of **A**.
- **Proof of why Trace operation is cyclic:** *https://math.stackexchange.com/questions/561012/prove-that-operatornametraceabc-operatornametracebca-operatornam*

Section 2.11: Determinant

- If **A** is an **n x n** matrix, then the product of the n eigenvalues of **A** is the determinant of **A**.
- The absolute value of the determinant gives the scale factor by which area or volume (or a higher-dimensional analogue) is multiplied under the associated linear transformation, while its sign indicates whether the transformation preserves orientation. For more details, refer to the following video by 3Blue1Brown:

  https://www.youtube.com/watch?v=Ip3X9LOh2dk&index=7&list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab

- If determinant is 0, then space is contracted completely along at least one dimension, causing it to lose all its volume.
- If determinant is 1, then the transformation preserves volume.

Section 2.12: Example: Principal Components Analysis

Read from the book. Beautiful explanation given with a concrete proof.