Linear Algebra Refresher

What is Linear Algebra

Linear Algebra

Linear algebra is the branch of mathematics concerning vector spaces and linear mappings between such spaces. It includes the study of lines, planes, and subspaces, but is also concerned with properties common to all vector spaces.

Why do we study Linear Algebra?

- Provides a way to compactly represent & operate on sets of linear equations.
- In machine learning, we represent data as matrices and hence it is natural to use notions and formalisms developed in Linear Algebra.

Introduction to LinAl

- Consider the following system of equations:

$$4x_1 - 5x_2 = -13$$
$$-2x_1 + 3x_2 = 9$$

- In matrix notation, the system is more compactly represented as:

$$Ax = b$$

$$A = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}$$

$$b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}$$

Vector Space

Definition

A set V with two operations $+$ and $\cdot$ is said to be a **vector space** if it is closed under both these operations and satisfies the following eight axioms.

1. Commutative Law

$$x + y = y + x, \quad \forall x, y \in V$$

2. Associative Law

$$(x + y) + z = x + (y + z), \quad \forall x, y, z \in V$$

3. Additive identity

$$\exists 0 \in V \quad s.t \quad x + 0 = x, \quad \forall x \in V$$

4. Additive inverse

$$\forall x \in V, \quad \exists \tilde{x} \quad s.t \quad x + \tilde{x} = 0$$

Vector Space (Contd..)

5. Distributive Law

$$\alpha \cdot (x + y) = \alpha \cdot x + \alpha \cdot y, \quad \forall \alpha \in \mathbb{R}, x, y \in V$$

6. Distributive Law

$$(\alpha + \beta) \cdot x = \alpha \cdot x + \beta \cdot x, \quad \forall \alpha, \beta \in \mathbb{R}, x \in V$$

7. Associative Law

$$(\alpha\beta) \cdot x = \alpha \cdot (\beta \cdot x), \quad \forall \alpha, \beta \in \mathbb{R}, x \in V$$

8. Unitary Law

$$1 \cdot x = x, \quad \forall x \in V$$

Subspace

Definition

Let W be a subset of a vector space V. Then W is called a **subspace** of V if W is a vector space.

- Do we have to verify all 8 conditions to check whether a given subset of a vector space is a subspace?
- **Theorem**: Let W be a subset of a vector space V. Then W is a subspace of V if and only if W is non-empty and
$x + \alpha y \in W, \quad \forall x, y \in W, \alpha \in \mathbb{R}$

Norm

Definition

Norm is any function $f : \mathbb{R}^n \to \mathbb{R}$ satisfying:

1. $\forall x \in \mathbb{R}^n, \quad f(x) \geq 0$ (non-negativity)
2. $f(x) = 0$ iff $x = 0$ (definiteness)
3. $\forall x \in \mathbb{R}^n, \quad f(tx) = |t| f(x)$ (homogeneity)
4. $\forall x, y \in \mathbb{R}^n, \quad f(x + y) \leq f(x) + f(y)$ (triangle inequality)

- Example - l_p norm

$$||x||_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$$

- Matrices can have norms too - e.g., Frobenius norm

$$||A||_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij}^2} = \sqrt{tr(A^T A)} \tag{1}$$

Range Of A Matrix

- The **span** of a set of vectors $X = \{x_1, x_2, \cdots x_n\}$ is the set of all vectors that can be expressed as a linear combination of the vectors in X.
  In other words, set of all vectors v such that $v = \sum_{i=1}^{i=|X|} \alpha_i x_i, \alpha_i \in R$
- The **range** or **columnspace** of a matrix A, denoted by $R(A)$ is the span of its columns. In other words, it contains all linear combinations of the columns of A. For instance, the columnspace of $A = \begin{bmatrix} 1 & 0 \\ 5 & 4 \\ 2 & 4 \end{bmatrix}$ is the plane spanned by the vectors $\begin{bmatrix} 1 \\ 5 \\ 2 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 4 \\ 4 \end{bmatrix}$

Nullspace Of A Matrix

Definition

The nullspace $N(A)$ of a matrix $A \in R^{m \times n}$ is the set of all vectors that equal 0 when multiplied by A. The dimensionality of the nullspace is also referred to as the **nullity** of A.

$$N(A) = \{x \in \mathbb{R}^n : Ax = 0\}$$

- Note that vectors in $N(A)$ are of dimension n, while those in $R(A)$ are of size m, so vectors in $R(A^T)$ and $N(A)$ are both of dimension n.

Example

Consider the matrix

$$A = \begin{bmatrix} 1 & 0 \\ 5 & 4 \\ 2 & 4 \end{bmatrix}$$

The nullspace of A is made up of vectors x of the form $\begin{bmatrix} u \\ v \end{bmatrix}$, such that

$$\begin{bmatrix} 1 & 0 \\ 5 & 4 \\ 2 & 4 \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The nullspace here only contains the vector (0,0).

Another Example

Now, consider the matrix

$$B = \begin{bmatrix} 1 & 0 & 1 \\ 5 & 4 & 9 \\ 2 & 4 & 6 \end{bmatrix}$$

Here, the third column is a linear combination of the first two columns.
Here, the nullspace is the line of all points $x = c, y = c, z = -c$.

Linear Independence and Rank

Definition

A set of vectors $\{x_1, x_2, \cdots x_n\} \in \mathbb{R}^n$ is said to be **(linearly) independent** if no vector can be represented as a linear combination of the remaining vectors.

- i.e., if $x_n = \sum_{i=1}^{n-1} \alpha_i x_i$ for some scalar values $\alpha_1, \cdots, \alpha_{n-1} \in \mathbb{R}$, then we say that the vectors $\{x_1, x_2, \cdots x_n\}$ are linearly dependent; otherwise, the vectors are linearly independent
- The **column rank** of a matrix $A \in \mathbb{R}^{m \times n}$ is the size of the largest subset of columns of A that constitute a linearly independent set
- Similarly, **row rank** of a matrix is the largest number of rows of A that constitute a linearly independent set

Properties Of Ranks

- For any matrix $A \in \mathbb{R}^{m \times n}$, it turns out that the column rank of A is equal to the row rank of A, collectively as the rank of A, denoted as **rank(A)**
- Some basic properties of the rank:
  1. For $A \in \mathbb{R}^{m \times n}$, $rank(A) \leq min(m, n)$.
     If $rank(A) = min(m, n)$, A is said to be **full rank**
  2. For $A \in \mathbb{R}^{m \times n}$, $rank(A) = rank(A^T)$
  3. For $A \in \mathbb{R}^{m \times n}, B \in \mathbb{R}^{n \times p}$, $rank(AB) \leq min(rank(A), rank(B))$
  4. For $A, B \in \mathbb{R}^{m \times n}$, $rank(A + B) \leq rank(A) + rank(B)$

Orthogonal Matrices

- A square matrix $U \in R^{n \times n}$ is **orthogonal** iff
  - All columns are mutually orthogonal $v_i^T v_j = 0, \forall i \neq j$
  - All columns are normalized $v_i^T v_i = 1, \forall i$
- If U is orthogonal, $UU^T = U^T U = I$. This also implies that the inverse of U happens to be its transpose.
- Another salient property of orthogonal matrices is that **they do not change** the Euclidean norm of a vector when they operate on it, i.e $||Ux||_2 = ||x||_2$.
  Multiplication by an orthogonal matrix can be thought of as a pure rotation,i.e., it does not change the magnitude of the vector, but changes the direction.

Quadratic Form of Matrices

- Given a square matrix $A \in \mathbb{R}^{n \times n}$ and a vector $x \in \mathbb{R}^n$, the scalar value $x^T Ax$ is called a **quadratic form**
- A symmetric matrix $A \in \mathbb{S}^n$ is positive definite (PD) if for all non-zero vectors $x \in \mathbb{R}^n$, $x^T Ax > 0$
- Similarly, positive semidefinite if $x^T Ax \geq 0$, negative definite if $x^T Ax < 0$ and negative semidefinite if $x^T Ax \leq 0$
- One important property of positive definite and negative definite matrices is that they are always full rank, and hence, invertible.
- **Gram matrix**: Given any matrix $A \in \mathbb{R}^{m \times n}$, matrix $G = A^T A$ is always positive semidefinite.
  Further if $m \geq n$, then G is positive definite.

Eigenvalues & Eigenvectors

- Given a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, λ is said to be an eigenvalue of $\mathbf{A}$ and vector $\vec{x}$ the corresponding eigenvector if

$$A\vec{x} = \lambda\vec{x}$$

- **Geometrical interpretation**
  We can think of the eigenvectors of a matrix A as those vectors which upon being operated by A are only scaled but not rotated.
- **Example**

$$A = \begin{bmatrix} 6 & 5 \\ 1 & 2 \end{bmatrix}, \vec{x} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$$

$$A\vec{x} = \begin{bmatrix} 35 \\ 7 \end{bmatrix} = 7\vec{x}$$

Characteristic Equation

- Trivially, the $\vec{0}$ vector would always be an eigenvector of any matrix. Hence, we only refer only to non-zero vectors as eigenvectors.
- Given a matrix A, how do we find all eigenvalue-eigenvector pairs?

$$A\vec{x} = \lambda\vec{x}$$
$$A\vec{x} - \lambda I\vec{x} = 0$$
$$(A - \lambda I)\vec{x} = 0$$

The above will hold iff

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

This equation is also referred to as the characteristic equation of A. Solving the equation gives us all the eigenvalues λ of A. Note that these eigenvalues can be **complex**.

Properties

1. The trace tr(A) of a matrix A also equals the sum of its n eigenvalues.

$$tr(A) = \sum_{i=1}^{n} \lambda_i$$

2. The determinant $|A|$ is equal to the product of the eigenvalues.

$$|A| = \prod_{i=1}^{n} \lambda_i$$

3. The rank of a matrix is equal to the number of non zero eigenvalues of A.
4. If A is invertible, then the eigenvalues of A^{-1} are of form $\frac{1}{\lambda_i}$, where λ_i are the eigenvalues of A.

Theorem

- If a matrix has all its eigen values distinct and non-zero, all its eigenvectors are linearly independent

Diagonalization

Given a matrix A, we consider the matrix S with each column being an eigenvector of A

$$S = \begin{bmatrix} \vdots & \vdots & \dots & \vdots \\ \vec{v_1} & \vec{v_2} & \dots & \vec{v_n} \\ \vdots & \vdots & \dots & \vdots \end{bmatrix}$$

$$AS = \begin{bmatrix} \vdots & \vdots & \dots & \vdots \\ \lambda_1\vec{v_1} & \lambda_2\vec{v_2} & \dots & \lambda_n\vec{v_n} \\ \vdots & \vdots & \dots & \vdots \end{bmatrix}$$

$$AS = \begin{bmatrix} \vdots & \vdots & \dots & \vdots \\ \vec{v_1} & \vec{v_2} & \dots & \vec{v_n} \\ \vdots & \vdots & \dots & \vdots \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \dots \\ \vdots & \ddots & \dots \\ 0 & \dots & \lambda_n \end{bmatrix}$$

Diagonalization

$$AS = S\Lambda$$
$$A = S\Lambda S^{-1}$$

- $S^{-1}AS$ is diagonal
- Note that the above result is dependent on S being invertible. In the case where the eigenvalues are distinct, this will be true since the eigenvectors will be linearly independent

Properties of Diagonalization

1. A square matrix A is said to be **diagonalizable** if $\exists S$ such that $A = S\Lambda S^{-1}$.
2. Diagonalization can be used to simplify computation of the higher powers of a matrix A, if the diagonalized form is available

$$A^n = (S\Lambda S^{-1})(S\Lambda S^{-1})\ldots(S\Lambda S^{-1})$$
$$A^n = S\Lambda^n S^{-1}$$

Λ^n is simple to compute since it is a diagonal matrix.

Eigenvalues & Eigenvectors of Symmetric Matrices

- Two important properties for a symmetric matrix A:
  1. All the eigenvalues of A are real
  2. The eigenvectors of A are orthonormal, i.e., matrix S is orthogonal. Thus, $A = S\Lambda S^T$.
- Definiteness of a symmetric matrix depends entirely on the sign of its eigenvalues. Suppose $A = S\Lambda S^T$, then

$$x^T Ax = x^T S\Lambda S^T x = y^T \Lambda y = \sum_{i=1}^{n} \lambda_i y_i^2$$

- Since $y_i^2 \geq 0$, sign of expression depends entirely on the λ_i's. For example, if all $\lambda_i > 0$, then matrix A is positive definite.

Eigenvalues of a PSD Matrix

Consider a positive semi definite matrix A. Then, $\forall \vec{x}$ which are eigenvectors of A.

$$\vec{x}^T A \vec{x} \geq 0$$
$$\lambda \vec{x}^T \vec{x} \geq 0$$
$$\lambda ||\vec{x}||^2 \geq 0$$

Hence, all eigenvalues of a PSD matrix are non-negative.

Singular Value Decomposition

1. We saw that diagonalization is applicable only to square matrices. We need some analogue for rectangular matrices too, since we often encounter them, e.g the Document-Term matrix. For a rectangular matrix, we consider left singular and right singular vectors as two bases instead of a single base of eigenvectors for square matrices.
2. The Singular Value Decomposition is given by $A = U\Sigma V^T$ where $U \in R^{m \times m}$, $\Sigma \in R^{m \times n}$ and $V \in R^{n \times n}$.

Singular Value Decomposition

1. U is such that the m columns of U are the eigenvectors of AA^T, also known as the left singular vectors of A.
2. V is such that the n columns of V are the eigenvectors of A^TA, also known as the right singular vectors of A.
3. Σ is a rectangular diagonal matrix with each element being the square root of an eigenvalue of AA^T or A^TA

Significance: SVD allows us to construct a lower rank approximation of a rectangular matrix. We choose only the top r singular values in Σ, and the corresponding columns in U and rows in V^T

Matrix Calculus

1. **The Gradient**
   Consider a function $f : \Re^{m \times n} \rightarrow \Re$. The gradient $\nabla_A f(A)$ denotes the matrix of partial derivatives with respect to every element of the matrix A. Each element is given by $(\nabla_A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}}$
2. **The Hessian**
   Suppose a function $f : \Re^n \rightarrow \Re$ takes in vectors and returns real numbers. The Hessian, denoted as $\nabla_x^2 f(x)$ or H is the $n \times n$ matrix of partial derivatives. $(\nabla_x^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$. Note that the Hessian is always symmetric.
3. Note that the Hessian is not the gradient of the gradient, since the gradient is a vector, and we cannot take the gradient of the vector. However, if we do take elementwise gradients of every element of the gradient, then we can construct the Hessian.

Differentiating Linear and Quadratic Functions

If $f(x) = b^T x$, for some constant $b \in \Re^n$. Let us find the gradient of f.

$$f(x) = \sum_{i=1}^{i=n} b_i x_i$$

$$\frac{\partial f(x)}{x_k} = b_k$$

We can see that $\frac{\partial b^T x}{\partial x} = b$. We can intuitively see how this relates to differentiating $f(x) = ax$ with respect to x when a and x are real scalars.

Differentiating Linear and Quadratic Functions

Consider the function $f(x) = x^T Ax$ where $x \in R^n$ and $A \in \Re^{n \times n}$ is a known symmetric matrix.

$$f(x) = \sum_{i=1}^{i=n} \sum_{j=1}^{j=n} A_{ij} x_i x_j$$

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \left[\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} {x_k}^2 \right]$$

$$\frac{\partial f(x)}{\partial x_k} = \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2 A_{kk} x_k$$

$$\frac{\partial f(x)}{\partial x_k} = \sum_{i=1}^{n} A_{ik} x_i + \sum_{j=1}^{n} A_{kj} x_j$$

$$\frac{\partial f(x)}{\partial x_k} = 2 \sum_{i=1}^{n} A_{ki} x_i$$

Differentiating Linear and Quadratic Functions

Thus $\nabla_x(x^T Ax) = 2Ax$. Now, let us find the Hessian H.

$$\frac{\partial}{\partial x_k}\frac{\partial f(x)}{\partial x_l} = \frac{\partial}{\partial x_k}(2\sum_{i=1}^{i=n} A_{li}x_i) \qquad = 2A_{kl}$$

Hence, $\nabla_x^2(x^T Ax) = 2A$.

Definition 5.6 (Jacobian). The collection of all first-order partial derivatives of a vector-valued function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ is called the *Jacobian*. The Jacobian $\boldsymbol{J}$ is an $m \times n$ matrix, which we define and arrange as follows:

$$\boldsymbol{J} = \nabla_{\boldsymbol{x}} \boldsymbol{f} = \frac{\mathrm{d}\boldsymbol{f}(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \tag{5.57}$$

$$= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix}, \tag{5.58}$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad J(i,j) = \frac{\partial f_i}{\partial x_j}. \tag{5.59}$$

Example 5.9 (Gradient of a Vector-Valued Function)

We are given

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}\,, \qquad \boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^M, \quad \boldsymbol{A} \in \mathbb{R}^{M\times N}, \quad \boldsymbol{x} \in \mathbb{R}^N\,.$$

To compute the gradient $\mathrm{d}\boldsymbol{f}/\mathrm{d}\boldsymbol{x}$ we first determine the dimension of $\mathrm{d}\boldsymbol{f}/\mathrm{d}\boldsymbol{x}$: Since $\boldsymbol{f} : \mathbb{R}^N \to \mathbb{R}^M$, it follows that $\mathrm{d}\boldsymbol{f}/\mathrm{d}\boldsymbol{x} \in \mathbb{R}^{M\times N}$. Second, to compute the gradient we determine the partial derivatives of f with respect to every x_j:

$$f_i(\boldsymbol{x}) = \sum_{j=1}^{N} A_{ij}x_j \implies \frac{\partial f_i}{\partial x_j} = A_{ij} \tag{5.67}$$

We collect the partial derivatives in the Jacobian and obtain the gradient

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix} = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & & \vdots \\ A_{M1} & \cdots & A_{MN} \end{bmatrix} = \boldsymbol{A} \in \mathbb{R}^{M\times N}\,. \tag{5.68}$$

Example 5.11 (Gradient of a Least-Squares Loss in a Linear Model)
Let us consider the linear model

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\theta}\,, \tag{5.75}$$

where $\boldsymbol{\theta} \in \mathbb{R}^D$ is a parameter vector, $\boldsymbol{\Phi} \in \mathbb{R}^{N\times D}$ are input features and $\boldsymbol{y} \in \mathbb{R}^N$ are the corresponding observations. We define the functions

$$L(\boldsymbol{e}) := \|\boldsymbol{e}\|^2\,, \tag{5.76}$$
$$\boldsymbol{e}(\boldsymbol{\theta}) := \boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}\,. \tag{5.77}$$

We seek $\frac{\partial L}{\partial \boldsymbol{\theta}}$, and we will use the chain rule for this purpose. L is called a *least-squares loss* function.

Before we start our calculation, we determine the dimensionality of the gradient as

$$\frac{\partial L}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{1\times D}\,. \tag{5.78}$$

The chain rule allows us to compute the gradient as

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \frac{\partial L}{\partial \boldsymbol{e}}\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}\,, \tag{5.79}$$

where the dth element is given by

$$\frac{\partial L}{\partial \boldsymbol{\theta}}[1,d] = \sum_{n=1}^{N}\frac{\partial L}{\partial \boldsymbol{e}}[n]\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}[n,d]\,. \tag{5.80}$$

We know that $\|\boldsymbol{e}\|^2 = \boldsymbol{e}^\top\boldsymbol{e}$ (see Section 3.2) and determine

$$\frac{\partial L}{\partial \boldsymbol{e}} = 2\boldsymbol{e}^\top \in \mathbb{R}^{1\times N}\,. \tag{5.81}$$

Furthermore, we obtain

$$\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}} = -\boldsymbol{\Phi} \in \mathbb{R}^{N\times D}\,, \tag{5.82}$$

such that our desired derivative is

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = -2\boldsymbol{e}^\top\boldsymbol{\Phi} \overset{(5.77)}{=} -\underbrace{2(\boldsymbol{y}^\top - \boldsymbol{\theta}^\top\boldsymbol{\Phi}^\top)}_{1\times N}\underbrace{\boldsymbol{\Phi}}_{N\times D} \in \mathbb{R}^{1\times D}\,. \tag{5.83}$$

5.5 Gradients of Matrices

Tensor is a multidimensional array.

Example 5.12 (Gradient of Vectors with Respect to Matrices)
Let us consider the following example, where

$$\boldsymbol{f} = \boldsymbol{A}\boldsymbol{x}\,, \quad \boldsymbol{f} \in \mathbb{R}^{M}\,, \quad \boldsymbol{A} \in \mathbb{R}^{M\times N}\,, \quad \boldsymbol{x} \in \mathbb{R}^{N} \tag{5.85}$$

and where we seek the gradient $\mathrm{d}\boldsymbol{f}/\mathrm{d}\boldsymbol{A}$. Let us start again by determining the dimension of the gradient as

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{A}} \in \mathbb{R}^{M\times(M\times N)}\,. \tag{5.86}$$

By definition, the gradient is the collection of the partial derivatives:

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{A}} = \begin{bmatrix} \frac{\partial f_1}{\partial \boldsymbol{A}} \\ \vdots \\ \frac{\partial f_M}{\partial \boldsymbol{A}} \end{bmatrix}\,, \quad \frac{\partial f_i}{\partial \boldsymbol{A}} \in \mathbb{R}^{1\times(M\times N)}\,. \tag{5.87}$$

To compute the partial derivatives, it will be helpful to explicitly write out the matrix vector multiplication:

$$f_i = \sum_{j=1}^{N} A_{ij}x_j\,, \quad i = 1,\ldots,M\,, \tag{5.88}$$

and the partial derivatives are then given as

$$\frac{\partial f_i}{\partial A_{iq}} = x_q\,. \tag{5.89}$$

This allows us to compute the partial derivatives of f_i with respect to a row of $\boldsymbol{A}$, which is given as

$$\frac{\partial f_i}{\partial \boldsymbol{A}_{i,:}} = \boldsymbol{x}^{\top} \in \mathbb{R}^{1\times 1\times N}\,, \tag{5.90}$$

$$\frac{\partial f_i}{\partial \boldsymbol{A}_{k\neq i,:}} = \boldsymbol{0}^{\top} \in \mathbb{R}^{1\times 1\times N} \tag{5.91}$$

where we have to pay attention to the correct dimensionality. Since f_i maps onto $\mathbb{R}$ and each row of $\boldsymbol{A}$ is of size $1\times N$, we obtain a $1\times 1\times N$-sized tensor as the partial derivative of f_i with respect to a row of $\boldsymbol{A}$.

We stack the partial derivatives (5.91) and get the desired gradient in (5.87) via

$$\frac{\partial f_i}{\partial \boldsymbol{A}} = \begin{bmatrix} \boldsymbol{0}^{\top} \\ \vdots \\ \boldsymbol{0}^{\top} \\ \boldsymbol{x}^{\top} \\ \boldsymbol{0}^{\top} \\ \vdots \\ \boldsymbol{0}^{\top} \end{bmatrix} \in \mathbb{R}^{1\times(M\times N)}\,. \tag{5.92}$$

5.5 Gradients of Matrices

Tensor is a multidimensional array.

Example 5.13 (Gradient of Matrices with Respect to Matrices)
Consider a matrix $\boldsymbol{R} \in \mathbb{R}^{M \times N}$ and $\boldsymbol{f} : \mathbb{R}^{M \times N} \to \mathbb{R}^{N \times N}$ with

$$\boldsymbol{f}(\boldsymbol{R}) = \boldsymbol{R}^\top \boldsymbol{R} =: \boldsymbol{K} \in \mathbb{R}^{N \times N}, \tag{5.93}$$

where we seek the gradient $\mathrm{d}\boldsymbol{K}/\mathrm{d}\boldsymbol{R}$.

To solve this hard problem, let us first write down what we already know: The gradient has the dimensions

$$\frac{\mathrm{d}\boldsymbol{K}}{\mathrm{d}\boldsymbol{R}} \in \mathbb{R}^{(N \times N) \times (M \times N)}, \tag{5.94}$$

which is a tensor. Moreover,

$$\frac{\mathrm{d}K_{pq}}{\mathrm{d}\boldsymbol{R}} \in \mathbb{R}^{1 \times M \times N} \tag{5.95}$$

for $p, q = 1, \ldots, N$, where K_{pq} is the (p, q)th entry of $\boldsymbol{K} = \boldsymbol{f}(\boldsymbol{R})$. Denoting the ith column of $\boldsymbol{R}$ by $\boldsymbol{r}_i$, every entry of $\boldsymbol{K}$ is given by the dot product of two columns of $\boldsymbol{R}$, i.e.,

$$K_{pq} = \boldsymbol{r}_p^\top \boldsymbol{r}_q = \sum_{m=1}^{M} R_{mp} R_{mq}. \tag{5.96}$$

When we now compute the partial derivative $\frac{\partial K_{pq}}{\partial R_{ij}}$ we obtain

$$\frac{\partial K_{pq}}{\partial R_{ij}} = \sum_{m=1}^{M} \frac{\partial}{\partial R_{ij}} R_{mp} R_{mq} = \partial_{pqij}, \tag{5.97}$$

$$\partial_{pqij} = \begin{cases} R_{iq} & \text{if } j = p,\ p \neq q \\ R_{ip} & \text{if } j = q,\ p \neq q \\ 2R_{iq} & \text{if } j = p,\ p = q \\ 0 & \text{otherwise} \end{cases}. \tag{5.98}$$

From (5.94), we know that the desired gradient has the dimension $(N \times N) \times (M \times N)$, and every single entry of this tensor is given by ∂_{pqij} in (5.98), where $p, q, j = 1, \ldots, N$ and $i = 1, \ldots, M$.