Chapter7 Presentation

ZhouXuRui

HITSZ

December 9, 2022

Contents

Contents

Notation

Vector

A vector $\boldsymbol{x} \in \mathbb{R}^n$ is a list of n numbers, usually written as a column vector

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

The unit vector $\boldsymbol{e}_i$ is a vector of all 0 's, except entry i, which has value 1:

$$\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$$

This is also called a one-hot vector.

Notation

Matrix

A matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ with m rows and n columns is a 2 d array of numbers, arranged as follows:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

If $m = n$, the matrix is said to be square.

We use the notation A_{ij} or $A_{i,j}$ to denote the entry of $\mathbf{A}$ in the i th row and j th column. We use the notation $\mathbf{A}_{i,:}$ to denote the i 'th row and $\mathbf{A}_{:,j}$ to denote the j 'th column. We treat all vectors as column vectors by default (so $\mathbf{A}_{i,:}$ is viewed as a column vector with n entries). We use bold upper case letters to denote matrices, bold lower case letters to denote vectors, and non-bold letters to denote scalars.

Notation

We can view a matrix as a set of columns stacked along the horizontal axis:

$$\mathbf{A} = \begin{bmatrix} | & | & & | \\ \mathbf{A}_{:,1} & \mathbf{A}_{:,2} & \cdots & \mathbf{A}_{:,n} \\ | & | & & | \end{bmatrix}$$

For brevity, we will denote this by

$$\mathbf{A} = \left[\mathbf{A}_{:,1}, \mathbf{A}_{:,2}, \ldots, \mathbf{A}_{:,n}\right]$$

We can also view a matrix as a set of rows stacked along the vertical axis:

$$\mathbf{A} = \begin{bmatrix} - & \mathbf{A}_{1,:}^\top & - \\ - & \mathbf{A}_{2,:}^\top & - \\ & \vdots & \\ - & \mathbf{A}_{m,:}^\top & - \end{bmatrix}$$

Notation

For brevity, we will denote this by

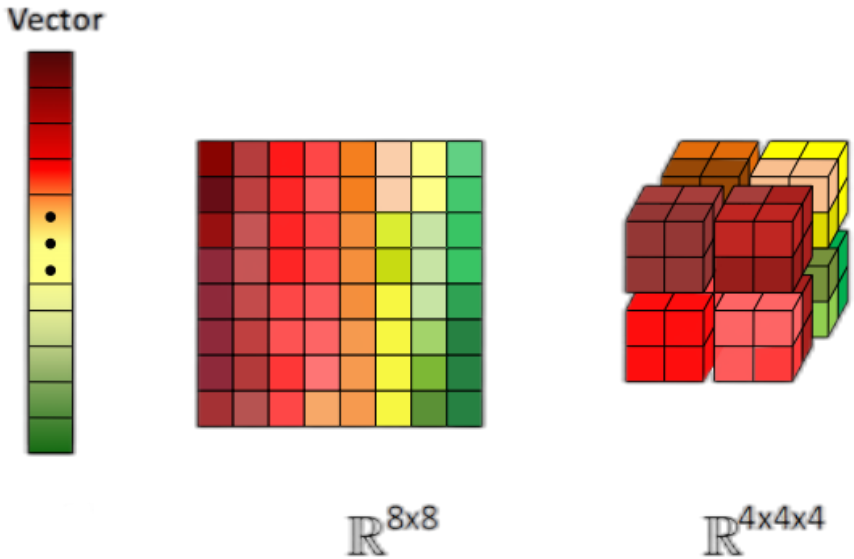
$$\mathbf{A} = \left[\mathbf{A}_{1,:}; \mathbf{A}_{2,:}; \ldots; \mathbf{A}_{m,:}\right]$$

Notation

Tensor

A tensor (in machine learning terminology) is just a generalization of a 2d array to more than 2 dimensions.The number of dimensions is known as the order or rank of the tensor.

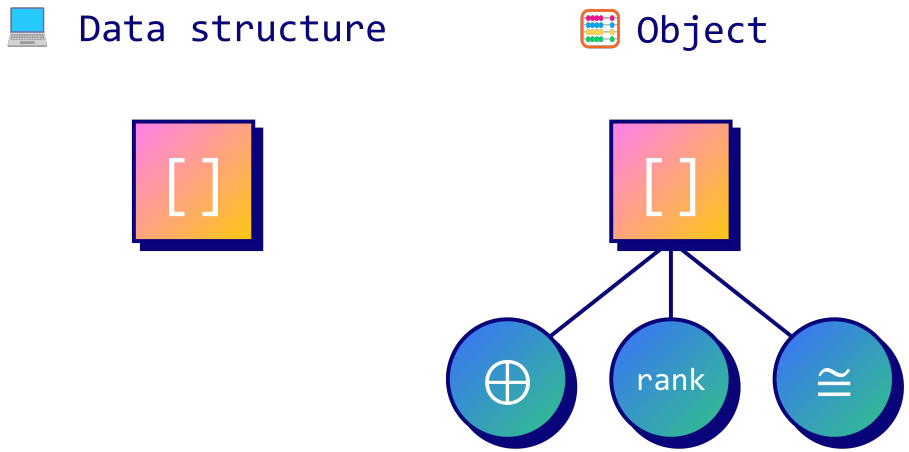
In mathematics, tensors can be viewed as a way to define multilinear maps, just as matrices can be used to define linear functions, although we will not need to use this interpretation.



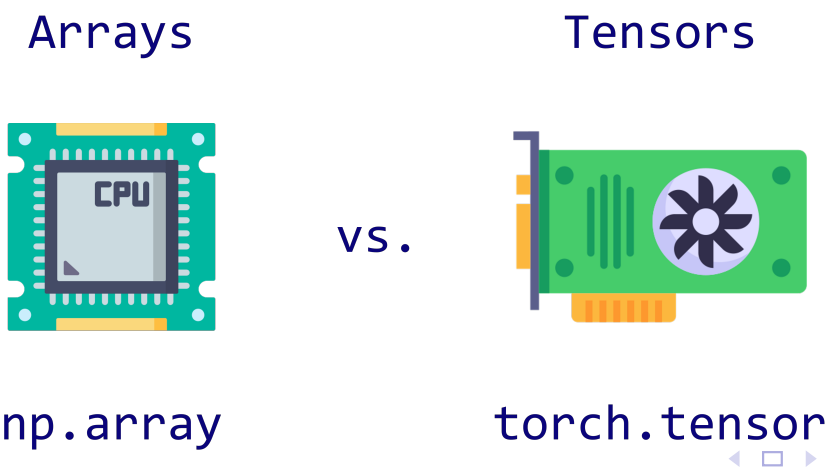
Notation

More introduction about Tensor:

Tensors in machine learning and Math/Physics:



Tensors in deep learning:



Notation

To put it more clearly:

Machine learning Tensor

In machine learning, the term "rank r tensor" is just another way to say "an r dimensional array of numbers." It just a way to refer to a data structure. There is no mathematical content to the usage of the term.

Mathematics Tensor

The tensors of mathematics are the linear representations of multilinear maps, and vector bundles constructed from them. The ideas make sense for arbitrary modules over arbitrary rings.

Notation

Pyhsics Tensor

The tensors of physics are sections of vector bundles on manifolds modeling configuration spaces/spacetime which are themselves tensor products of tangent and cotangent spaces. This means that relative to a choice of coordinate systems on the manifold, they are functions taking values in multilinear maps whose arguments are either vectors like the partial derivative operators in each of the coordinate directions, or forms like the differentials of the coordinate functions. Evaluating them on these basis elements yields a multidimensional array of functions.

Contents

Vector Space

Space in Mathematics

A space consists of selected mathematical objects that are treated as points, and selected relationships(structure) between these points.

Vector Space

Definition: A vector space consists of a set V (elements of V are called vectors), a field $\mathbb{F}$ (elements of $\mathbb{F}$ are called scalars), and two operations

- An operation called vector addition that takes two vectors $v, w \in V$, and produces a third vector, written $v + w \in V$.
- An operation called scalar multiplication that takes a scalar $c \in \mathbb{F}$ and a vector $v \in V$, and produces a new vector, written $cv \in V$

which satisfy the following conditions (called axioms).

Vector Space

Axioms

1. Associativity of vector addition: $(u+v)+w=u+(v+w)$ for all $u, v, w \in V$.

2. Existence of a zero vector: There is a vector in V, written 0 and called the zero vector, which has the property that $u+0=u$ for all $u \in V$

3. Existence of negatives: For every $u \in V$, there is a vector in V, written $-u$ and called the negative of u, which has the property that $u+(-u)=0$.

4. Associativity of multiplication: $(ab)u=a(bu)$ for any $a, b \in \mathbb{F}$ and $u \in V$.

5. Distributivity: $(a+b)u=au+bu$ and $a(u+v)=au+av$ for all $a, b \in \mathbb{F}$ and $u, v \in V$.

6. Unitarity: $1u=u$ for all $u \in V$.

Vector Space

Linear Independence

A set of vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n\}$ is said to be (linearly) independent if no vector can be represented as a linear combination of the remaining vectors. Conversely, a vector which can be represented as a linear combination of the remaining vectors is said to be (linearly) dependent.

Span

The span of a set of vectors $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ is the set of all vectors that can be expressed as a linear combination of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. That is,

$$\text{span}\left(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}\right) \triangleq \left\{ \boldsymbol{v} : \boldsymbol{v} = \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i, \quad \alpha_i \in \mathbb{R} \right\}$$

Vector Space

Basis

A basis $\mathcal{B}$ is a set of linearly independent vectors that spans the whole space, meaning that $\text{span}(\mathcal{B}) = \mathbb{R}^n$.

There are often multiple bases to choose from, the standard basis uses the coordinate vectors $\boldsymbol{e}_1 = (1, 0, \ldots, 0)$, up to $\boldsymbol{e}_n = (0, 0, \ldots, 0, 1)$. This lets us translate back and forth between viewing a vector in $\mathbb{R}^2$ as an either an "arrow in the plane", rooted at the origin, or as an ordered list of numbers (corresponding to the coefficients for each basis vector).

Linear Transformation

A linear map or linear transformation is any function $f : \mathcal{V} \to \mathcal{W}$ such that $f(\boldsymbol{v} + \boldsymbol{w}) = f(\boldsymbol{v}) + f(\boldsymbol{w})$ and $f(a\boldsymbol{v}) = af(\boldsymbol{v})$ for all $\boldsymbol{v}, \boldsymbol{w} \in \mathcal{V}$.

Once the basis of $\mathcal{V}$ is chosen, a linear map $f : \mathcal{V} \to \mathcal{W}$ is completely determined by specifying the images of the basis vectors, because any element of $\mathcal{V}$ can be expressed uniquely as a linear combination of them.

Suppose $\mathcal{V} = \mathbb{R}^n$ and $\mathcal{W} = \mathbb{R}^m$. We can compute $f(\boldsymbol{v}_i) \in \mathbb{R}^m$ for each basis vector in $\mathcal{V}$, and store these along the columns of an $m \times n$ matrix $\mathbf{A}$. We can then compute $\boldsymbol{y} = f(\boldsymbol{x}) \in \mathbb{R}^m$ for any $\boldsymbol{x} \in \mathbb{R}^n$ as follows:

$$\boldsymbol{y} = \left(\sum_{j=1}^{n} a_{1j} x_j, \ldots, \sum_{j=1}^{n} a_{mj} x_j \right)$$

Vector Space

This corresponds to multiplying the vector $\boldsymbol{x}$ by the matrix $\mathbf{A}$:

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$$

That is, Matrix can be viewed as Linear Transformation.

Vector Space

Span

Suppose we view a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ as a set of n vectors in $\mathbb{R}^m$. The range (sometimes also called the column space) of this matrix is the span of the columns of $\mathbf{A}$. In other words,

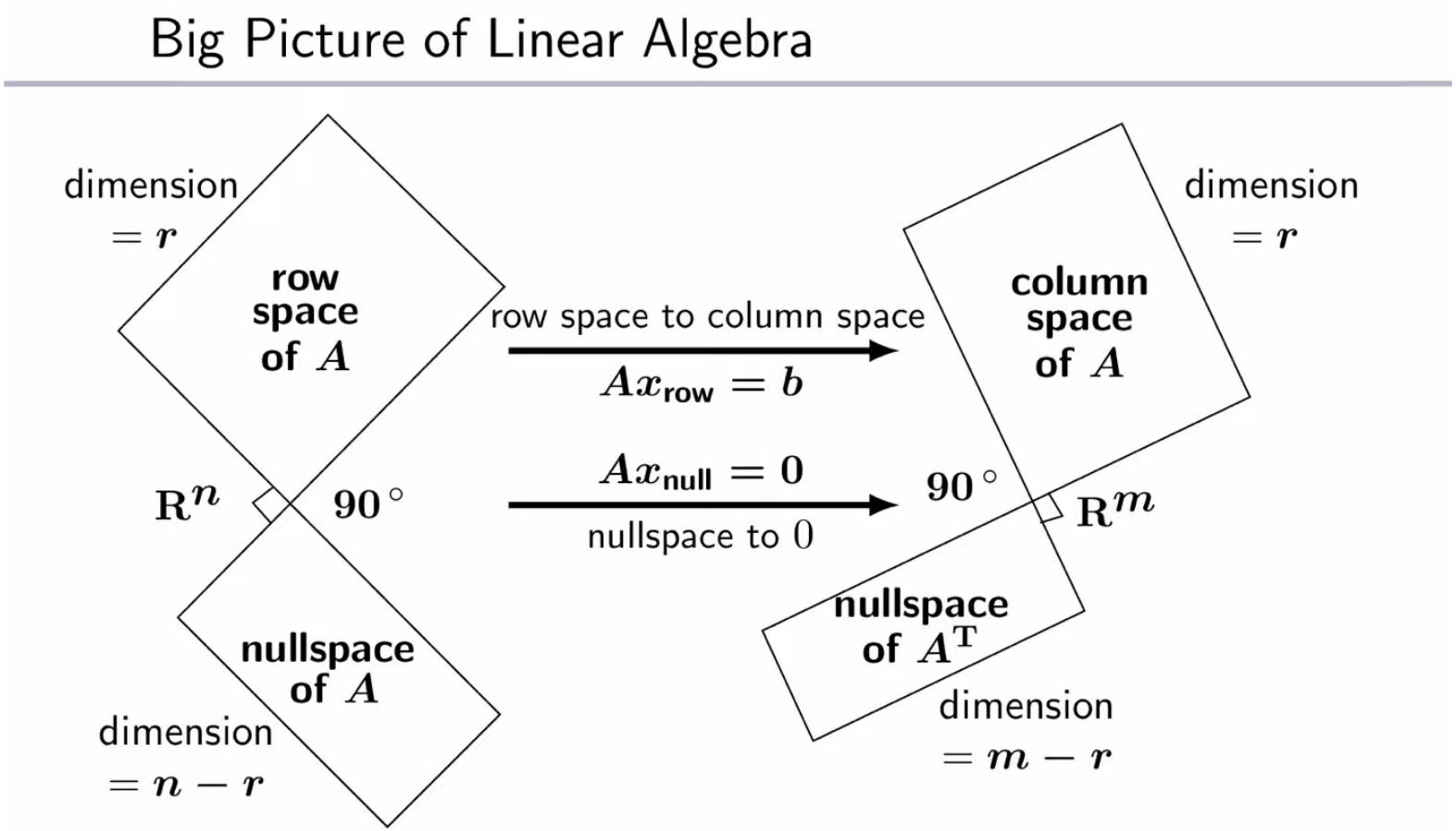
$$\text{range}(\mathbf{A}) \triangleq \{\boldsymbol{v} \in \mathbb{R}^m : \boldsymbol{v} = \mathbf{A}\boldsymbol{x}, \boldsymbol{x} \in \mathbb{R}^n\}.$$

This can be thought of as the set of vectors that can be "reached" or "generated" by $\mathbf{A}$; it is a subspace of $\mathbb{R}^m$ whose dimensionality is given by the rank of $\mathbf{A}$. The nullspace of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ is the set of all vectors that get mapped to the null vector when multiplied by $\mathbf{A}$, i.e.,

$$\text{nullspace}(\mathbf{A}) \triangleq \{\boldsymbol{x} \in \mathbb{R}^n : \mathbf{A}\boldsymbol{x} = \mathbf{0}\}.$$

Vector Space

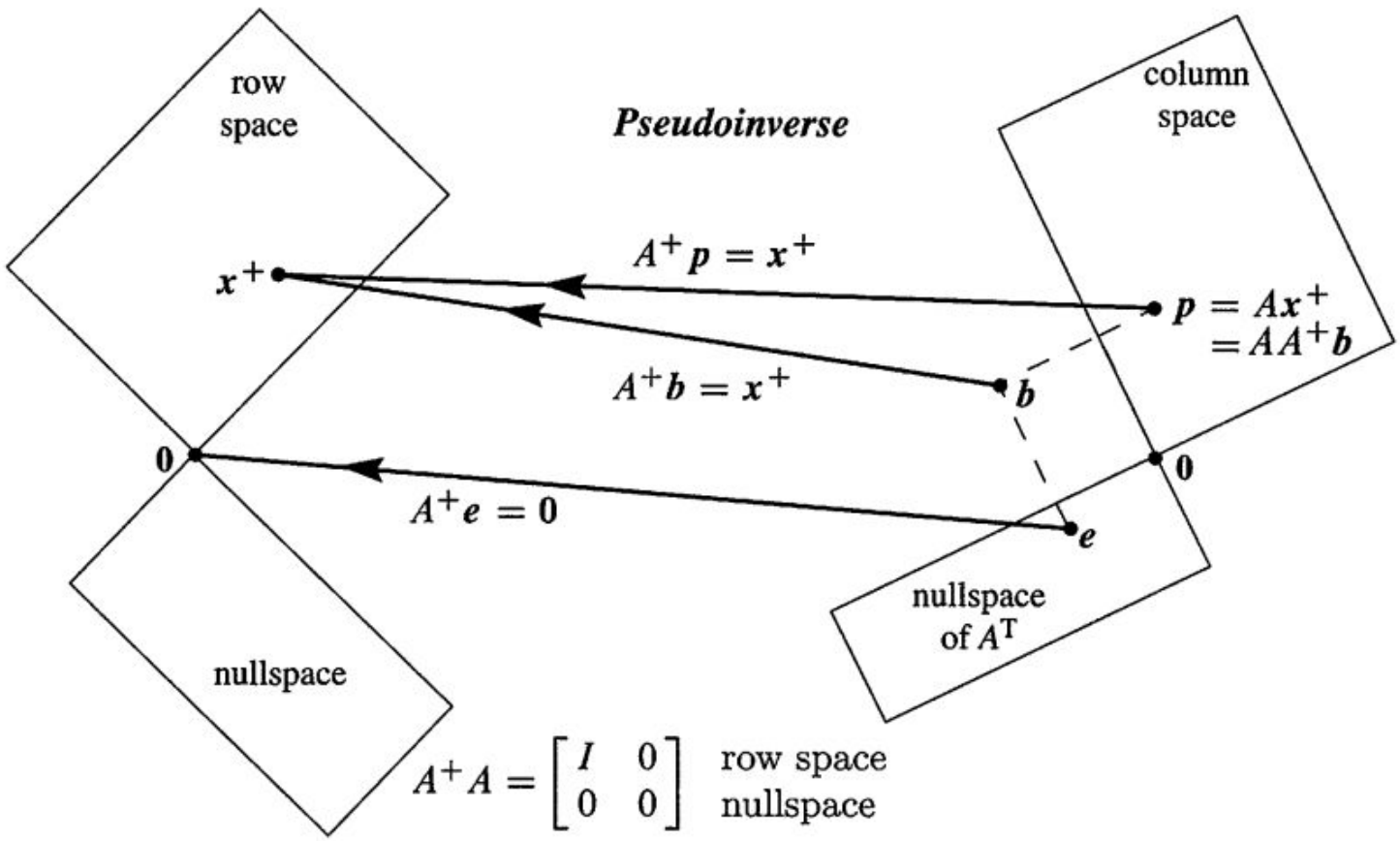
Here is the big picture of linear algebra in MIT 18.06:



This is the Big Picture—two subspaces in $\mathbf{R}^n$ and two subspaces in $\mathbf{R}^m$.

From row space to column space, A is invertible.

Vector Space



Projection

The projection of a vector $\boldsymbol{y} \in \mathbb{R}^m$ onto the span of $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ (here we assume $\boldsymbol{x}_i \in \mathbb{R}^m$) is the vector $\boldsymbol{v} \in \operatorname{span}(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\})$, such that $\boldsymbol{v}$ is as close as possible to $\boldsymbol{y}$, as measured by the Euclidean norm $\|\boldsymbol{v} - \boldsymbol{y}\|_2$. We denote the projection as $\operatorname{Proj}(\boldsymbol{y}; \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\})$ and can define it formally as

$$\operatorname{Proj}(\boldsymbol{y}; \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}) = \operatorname{argmin}_{\boldsymbol{v} \in \operatorname{span}(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\})} \|\boldsymbol{y} - \boldsymbol{v}\|_2$$

Given a (full rank) matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ with $m \geq n$, we can define the projection of a vector $\boldsymbol{y} \in \mathbb{R}^m$ onto the range of $\mathbf{A}$ as follows:

$$\operatorname{Proj}(\boldsymbol{y}; \mathbf{A}) = \operatorname{argmin}_{\boldsymbol{v} \in \mathcal{R}(\mathbf{A})} \|\boldsymbol{v} - \boldsymbol{y}\|_2 = \mathbf{A}\left(\mathbf{A}^\top \mathbf{A}\right)^{-1} \mathbf{A}^\top \boldsymbol{y}.$$

Proof

When you want to find the projection p of b onto $C(A)$, then you can start with a "parameterization" of p that guarantees that your result is in $C(A)$. So you set $p = A\bar{x}$ and $\bar{x} \in V$. Now you want $p - b$ to be orthogonal to $C(A)$, which means it's in the null space of A^T, so we can get $A^T(p - b) = 0$ or $A^T(A\bar{x} - b) = 0$ or

$$\bar{x} = \left(A^T A\right)^{-1} A^T b$$

Now you have the particular $\bar{x} \in V$ that provides the correct parameters for your projection p, and you just have to apply your initial choice $p = A\bar{x}$ for the parametrization to obtain the projection p.

Vector Space

When $Ax = b$ is inconsistent, its least-squares solution minimizes $\|Ax - b\|^2$:

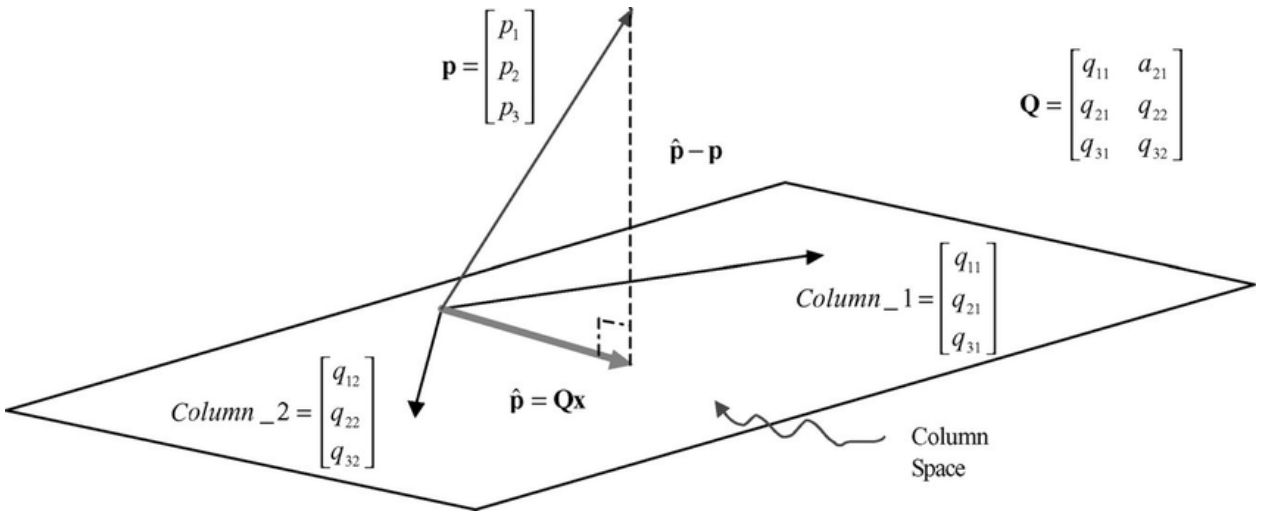
- Normal equations $A^{\mathrm{T}}A\hat{x} = A^{\mathrm{T}}b$.

$A^{\mathrm{T}}A$ is invertible exactly when the columns of A are linearly independent! Then,

- Best estimate $\hat{x}$ $\hat{x} = \left(A^{\mathrm{T}}A\right)^{-1}A^{\mathrm{T}}b$.

The projection of b onto the column space is the nearest point $A\hat{x}$:

- Projection $p = A\hat{x} = A\left(A^{\mathrm{T}}A\right)^{-1}A^{\mathrm{T}}b$.



Contents

Norms

Vector Norms

A norm of a vector $\|\boldsymbol{x}\|$ is, informally, a measure of the "length" of the vector. More formally, a norm is any function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies 4 properties:

- For all $\boldsymbol{x} \in \mathbb{R}^n$, $f(\boldsymbol{x}) \geq 0$ (non-negativity).
- For all $\boldsymbol{x} \in \mathbb{R}^n, t \in \mathbb{R}$, $f(t\boldsymbol{x}) = |t|f(x)$ (absolute value homogeneity).
- For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, $f(\boldsymbol{x} + \boldsymbol{y}) \leq f(\boldsymbol{x}) + f(\boldsymbol{y})$ (triangle inequality).

Consider the following common examples:

p-norm $\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$, for $p \geq 1$

2-norm $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$, also called Euclidean norm. Note that $\|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^\top \boldsymbol{x}$.

1-norm $\|\boldsymbol{x}\|_1 = \sum_{i=1}^n |x_i|$.

Max-norm $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$.

Matrix Norms

Suppose we think of a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ as defining a linear function $f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x}$. We define the induced norm of $\mathbf{A}$ as the maximum amount by which f can lengthen any unit-norm input:

$$\|\mathbf{A}\|_p = \max_{\boldsymbol{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p} = \max_{\|\boldsymbol{x}\|=1} \|\mathbf{A}\boldsymbol{x}\|_p$$

Typically $p = 2$, in which case

$$\|\mathbf{A}\|_2 = \sqrt{\lambda_{\max}(\mathbf{A}^\top \mathbf{A})} = \max_i \sigma_i$$

Contents

Properties

Trace

The trace of a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, denoted $\mathrm{tr}(\mathbf{A})$, is the sum of diagonal elements in the matrix:

$$\mathrm{tr}(\mathbf{A}) \triangleq \sum_{i=1}^{n} A_{ii}.$$

The trace has the following properties, where $c \in \mathbb{R}$ is a scalar, and $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$ are square matrices:

$$\mathrm{tr}(\mathbf{A}) = \mathrm{tr}\left(\mathbf{A}^{\top}\right)$$
$$\mathrm{tr}(\mathbf{A} + \mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$$
$$\mathrm{tr}(c\mathbf{A}) = c\,\mathrm{tr}(\mathbf{A})$$
$$\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$$

Properties

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i \text{ where } \lambda_i \text{ are the eigenvalues of } \mathbf{A}$$

We also have the following important cyclic permutation property: For $\mathbf{A}, \mathbf{B}, \mathbf{C}$ such that ABC is square,

$$\text{tr}(\mathbf{ABC}) = \text{tr}(\mathbf{BCA}) = \text{tr}(\mathbf{CAB})$$

From this, we can derive the trace trick, which rewrites the scalar inner product $\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x}$ as follows

$$\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x} = \text{tr}\left(\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x}\right) = \text{tr}\left(\boldsymbol{x}\boldsymbol{x}^\top \mathbf{A}\right)$$

Contents

Positive Definite Matrix

Given a square matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ and a vector $\boldsymbol{x} \in \mathbb{R}^n$, the scalar value $\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x}$ is called a quadratic form. Written explicitly, we see that

$$\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j.$$

Note that,

$$\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x} = \left(\boldsymbol{x}^\top \mathbf{A} \boldsymbol{x} \right)^\top = \boldsymbol{x}^\top \mathbf{A}^\top \boldsymbol{x} = \boldsymbol{x}^\top \left(\frac{1}{2} \mathbf{A} + \frac{1}{2} \mathbf{A}^\top \right) \boldsymbol{x}$$

For this reason, we often implicitly assume that the matrices appearing in a quadratic form are symmetric.

Special Types

Positive Definite Matrix

The following conditions are equivalent:
1) $\mathbf{x}^T A\mathbf{x} > 0$ for all $x \neq 0$.
2) All eigenvalues of A satisfy $\lambda_j > 0$.
3) All NW (upper left) minors of A are positive.
4) In the reduction of A to row echelon form, no exchanges are required, and all pivots are positive.
5) $A = R^T R$ for some non-singular R.

Positive Definite Matrix

For a symmetric matrix A, the following conditions are equivalent.
(1) $A \succeq 0$.
(2) $A = U^\top U$ for some matrix U.
(3) $x^\top Ax \geq 0$ for every $x \in \mathbb{R}^n$.
(4) All principal minors of A are nonnegative.

Special Types

AA^T is positive semidefinite.
Proof: Let $y = A^T x$.

$$x^T AA^T x = y^T y = \sum_{k=1}^{N} y_k^2 \geqslant 0$$

Contents

Vector–vector products

Given two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the quantity $\boldsymbol{x}^\top \boldsymbol{y}$, called the inner product, dot product or scalar product of the vectors, is a real number given by

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \triangleq \boldsymbol{x}^\top \boldsymbol{y} = \sum_{i=1}^{n} x_i y_i$$

Given vectors $\boldsymbol{x} \in \mathbb{R}^m, \boldsymbol{y} \in \mathbb{R}^n$ (they no longer have to be the same size), $\boldsymbol{x}\boldsymbol{y}^\top$ is called the outer product of the vectors. It is a matrix whose entries are given by $\left(\boldsymbol{x}\boldsymbol{y}^\top\right)_{ij} = x_i y_j$, i.e.,

$$\boldsymbol{x}\boldsymbol{y}^\top \in \mathbb{R}^{m \times n} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix}.$$

Contents

Matrix–vector products

Given a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and a vector $\boldsymbol{x} \in \mathbb{R}^n$, their product is a vector $\boldsymbol{y} = \mathbf{A}\boldsymbol{x} \in \mathbb{R}^m$. There are a couple ways of looking at matrix-vector multiplication, and we will look at them both.
If we write $\mathbf{A}$ by rows, then we can express $\boldsymbol{y} = \mathbf{A}\boldsymbol{x}$ as follows:

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{x} = \begin{bmatrix} - & \boldsymbol{a}_1^\top & - \\ - & \boldsymbol{a}_2^\top & - \\ & \vdots & \\ - & \boldsymbol{a}_m^\top & - \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \boldsymbol{a}_1^\top \boldsymbol{x} \\ \boldsymbol{a}_2^\top \boldsymbol{x} \\ \vdots \\ \boldsymbol{a}_m^\top \boldsymbol{x} \end{bmatrix}$$

Matrix–vector products

In other words, the i th entry of $\boldsymbol{y}$ is equal to the inner product of the i th row of $\mathbf{A}$ and $\boldsymbol{x}$, $y_i = \boldsymbol{a}_i^\top \boldsymbol{x}$. Alternatively, let's write $\mathbf{A}$ in column form. In this case we see that

$$\boldsymbol{y} = \mathbf{A}\boldsymbol{x} = \begin{bmatrix} | & | & | & \\ \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_n \\ | & | & | & \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} | \\ \boldsymbol{a}_1 \\ | \end{bmatrix} x_1 + \ldots + \begin{bmatrix} | \\ \boldsymbol{a}_n \\ | \end{bmatrix} x_n.$$

In other words, $\boldsymbol{y}$ is a linear combination of the columns of $\mathbf{A}$, where the coefficients of the linear combination are given by the entries of $\boldsymbol{x}$. We can view the columns of $\mathbf{A}$ as a set of basis vectors defining a linear subspace. We can contstruct vectors in this subspace by taking linear combinations of the basis vectors.

Contents

Matrix–matrix products

Below we look at four different (but, of course, equivalent) ways of viewing the matrix-matrix multiplication $\mathbf{C} = \mathbf{AB}$.

First we can view matrix-matrix multiplication as a set of vector-vector products. The most obvious viewpoint, which follows immediately from the definition, is that the i, j entry of $\mathbf{C}$ is equal to the inner product of the i th row of $\mathbf{A}$ and the j th column of $\mathbf{B}$. Symbolically, this looks like the following:

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} - & \boldsymbol{a}_1^\top & - \\ - & \boldsymbol{a}_2^\top & - \\ & \vdots & \\ - & \boldsymbol{a}_m^\top & - \end{bmatrix} \begin{bmatrix} | & | & & | \\ \boldsymbol{b}_1 & \boldsymbol{b}_2 & \cdots & \boldsymbol{b}_p \\ | & | & & | \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{a}_1^\top \boldsymbol{b}_1 & \boldsymbol{a}_1^\top \boldsymbol{b}_2 & \cdots & \boldsymbol{a}_1^\top \boldsymbol{b}_p \\ \boldsymbol{a}_2^\top \boldsymbol{b}_1 & \boldsymbol{a}_2^\top \boldsymbol{b}_2 & \cdots & \boldsymbol{a}_2^\top \boldsymbol{b}_p \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{a}_m^\top \boldsymbol{b}_1 & \boldsymbol{a}_m^\top \boldsymbol{b}_2 & \cdots & \boldsymbol{a}_m^\top \boldsymbol{b}_p \end{bmatrix}$$

Matrix–matrix products

Alternatively, we can represent **A** by columns, and **B** by rows, which leads to the interpretation of AB as a sum of outer products. Symbolically,

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} | & | & & | \\ \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} - & \boldsymbol{b}_1^\top & - \\ - & \boldsymbol{b}_2^\top & - \\ & \vdots & \\ - & \boldsymbol{b}_n^\top & - \end{bmatrix} = \sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{b}_i^\top$$

Put another way, **AB** is equal to the sum, over all i, of the outer product of the i th column of **A** and the i th row of **B**. Since, in this case, $\boldsymbol{a}_i \in \mathbb{R}^m$ and $\boldsymbol{b}_i \in \mathbb{R}^p$, the dimension of the outer product $\boldsymbol{a}_i \boldsymbol{b}_i^\top$ is $m \times p$, which coincides with the dimension of **C**.

Matrix–matrix products

We can also view matrix-matrix multiplication as a set of matrix-vector products. Specifically, if we represent **B** by columns, we can view the columns of **C** as matrix-vector products between **A** and the columns of B. Symbolically,

$$\mathbf{C} = \mathbf{AB} = \mathbf{A}\begin{bmatrix} | & | & & | \\ \boldsymbol{b}_1 & \boldsymbol{b}_2 & \cdots & \boldsymbol{b}_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \mathbf{A}\boldsymbol{b}_1 & \mathbf{A}\boldsymbol{b}_2 & \cdots & \mathbf{A}\boldsymbol{b}_p \\ | & | & & | \end{bmatrix}.$$

Here the i th column of **C** is given by the matrix-vector product with the vector on the right, $\boldsymbol{c}_i = \mathbf{A}\boldsymbol{b}_i$. These matrix-vector products can in turn be interpreted using both viewpoints given in the previous subsection.

Matrix–matrix products

Finally, we have the analogous viewpoint, where we represent **A** by rows, and view the rows of **C** as the matrix-vector product between the rows of **A** and the matrix **B**. Symbolically,

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} - & \boldsymbol{a}_1^\top & - \\ - & \boldsymbol{a}_2^\top & - \\ & \vdots & \\ - & \boldsymbol{a}_m^\top & - \end{bmatrix} \mathbf{B} = \begin{bmatrix} - & \boldsymbol{a}_1^\top \mathbf{B} & - \\ - & \boldsymbol{a}_2^\top \mathbf{B} & - \\ & \vdots & \\ - & \boldsymbol{a}_m^\top \mathbf{B} & - \end{bmatrix}$$

It may seem like overkill to dissect matrix multiplication to such a large degree, especially when all these viewpoints follow immediately from the initial definition we gave (in about a line of math) at the beginning of this section. However, virtually all of linear algebra deals with matrix multiplications of some kind, and it is worthwhile to spend some time trying to develop an intuitive understanding of the viewpoints presented here.

Contents

Manipulating data matrices

Suppose $\mathbf{X}$ is an $N \times D$ matrix. We can sum across the rows by premultiplying by a $1 \times N$ matrix of ones to create a $1 \times D$ matrix:

$$\mathbf{1}_N^\top \mathbf{X} = \begin{pmatrix} \sum_n x_{n1} & \cdots & \sum_n x_{nD} \end{pmatrix}$$

Hence the mean of the data vectors is given by

$$\overline{\boldsymbol{x}}^\top = \frac{1}{N} \mathbf{1}_N^\top \mathbf{X}$$

We can sum across the columns by postmultiplying by a $D \times 1$ matrix of ones to create a $N \times 1$ matrix:

$$\mathbf{X}\mathbf{1}_D = \begin{pmatrix} \sum_d x_{1d} \\ \vdots \\ \sum_d x_{Nd} \end{pmatrix}$$

Manipulating data matrices

We can sum all entries in a matrix by pre and post multiplying by a vector of 1s:

$$\mathbf{1}_N^\top \mathbf{X} \mathbf{1}_D = \sum_{ij} X_{ij}$$

Hence the overall mean is given by

$$\bar{x} = \frac{1}{ND} \mathbf{1}_N^\top \mathbf{X} \mathbf{1}_D$$

We often want to scale rows or columns of a data matrix (e.g., to standardize them). We now show how to write this in matrix notation.

Manipulating data matrices

If we pre-multiply $\mathbf{X}$ by a diagonal matrix $\mathbf{S} = \text{diag}(\boldsymbol{s})$, where $\boldsymbol{s}$ is an N-vector, then we just scale each row of $\mathbf{X}$ by the corresponding scale factor in s :

$$\text{diag}(\boldsymbol{s})\mathbf{X} = \begin{pmatrix} s_1 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & s_N \end{pmatrix} \begin{pmatrix} x_{1,1} & \cdots & x_{1,D} \\ & \ddots & \\ x_{N,1} & \cdots & x_{N,D} \end{pmatrix} =$$

$$= \begin{pmatrix} s_1 x_{1,1} & \cdots & s_1 x_{1,D} \\ & \ddots & \\ s_N x_{N,1} & \cdots & s_N x_{N,D} \end{pmatrix}$$

Manipulating data matrices

If we post-multiply $\mathbf{X}$ by a diagonal matrix $\mathbf{S} = \text{diag}(\boldsymbol{s})$, where $\boldsymbol{s}$ is a D-vector, then we just scale each column of $\mathbf{X}$ by the corresponding element in s.

$$\mathbf{X}\,\text{diag}(\boldsymbol{s}) = \begin{pmatrix} x_{1,1} & \cdots & x_{1,D} \\ & \ddots & \\ x_{N,1} & \cdots & x_{N,D} \end{pmatrix} \begin{pmatrix} s_1 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & s_D \end{pmatrix} =$$

$$\begin{pmatrix} s_1 x_{1,1} & \cdots & s_D x_{1,D} \\ & \ddots & \\ s_1 x_{N,1} & \cdots & s_D x_{N,D} \end{pmatrix}$$

Manipulating data matrices

Thus we can rewrite the standardization operation in matrix form as follows:

$$\text{standardize}(\mathbf{X}) = \left(\mathbf{X} - \mathbf{1}_N \boldsymbol{\mu}^T\right) \text{diag}(\boldsymbol{\sigma})^{-1}$$

where $\boldsymbol{\mu} = \overline{\boldsymbol{x}}$ is the empirical mean, and $\boldsymbol{\sigma}$ is a vector of the empirical standard deviations.

Manipulating data matrices

Sum Square Matrix

The sum of squares matrix is $D \times D$ matrix defined by

$$\mathbf{S}_0 \triangleq \mathbf{X}^\top \mathbf{X} = \sum_{n=1}^{N_{\mathcal{D}}} \boldsymbol{x}_n \boldsymbol{x}_n^\top = \sum_{n=1}^{N_{\mathcal{D}}} \begin{pmatrix} x_{n,1}^2 & \cdots & x_{n,1} x_{n,D} \\ & \ddots & \\ x_{n,D} x_{n,1} & \cdots & x_{n,D}^2 \end{pmatrix}$$

The scatter matrix is a $D \times D$ matrix defined by

$$\mathbf{S}_{\overline{\boldsymbol{x}}} \triangleq \sum_{n=1}^{N} (\boldsymbol{x}_n - \overline{\boldsymbol{x}}) (\boldsymbol{x}_n - \overline{\boldsymbol{x}})^\top = \left(\sum_n \boldsymbol{x}_n \boldsymbol{x}_n^\top \right) - N \overline{\boldsymbol{x}} \overline{\boldsymbol{x}}^\top$$

Manipulating data matrices

We see that this is the sum of squares matrix applied to the mean-centered data. More precisely, define $\tilde{\mathbf{X}}$ to be a version of $\mathbf{X}$ where we subtract the mean $\overline{\boldsymbol{x}} = \frac{1}{N}\mathbf{X}^\top \mathbf{1}_N$ off every row. Hence we can compute the centered data matrix using

$$\tilde{\mathbf{X}} = \mathbf{X} - \mathbf{1}_N \overline{\boldsymbol{x}}^\top = \mathbf{X} - \frac{1}{N}\mathbf{1}_N \mathbf{1}_N^\top \mathbf{X} = \mathbf{C}_N \mathbf{X}$$

where

$$\mathbf{C}_N \triangleq \mathbf{I}_N - \frac{1}{N}\mathbf{J}_N$$

is the centering matrix, and $\mathbf{J}_N = \mathbf{1}_N \mathbf{1}_N^\top$ is a matrix of all 1s. The scatter matrix can now be computed as follows:

$$\mathbf{S}_{\overline{\boldsymbol{x}}} = \tilde{\mathbf{X}}^\top \tilde{\mathbf{X}} = \mathbf{X}^\top \mathbf{C}_N^\top \mathbf{C}_N \mathbf{X} = \mathbf{X}^\top \mathbf{C}_N \mathbf{X}$$

Manipulating data matrices

Gram Matrix

The $N \times N$ matrix $\mathbf{X}\mathbf{X}^\top$ is a matrix of inner products called the Gram matrix:

$$\mathbf{K} \triangleq \mathbf{X}\mathbf{X}^\top = \begin{pmatrix} \boldsymbol{x}_1^\top \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_1^\top \boldsymbol{x}_N \\ & \ddots & \\ \boldsymbol{x}_n^\top \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_N^\top \boldsymbol{x}_N \end{pmatrix}$$

Sometimes we want to compute the inner products of the mean-centered data vectors, $\tilde{\mathbf{K}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top$. However, if we are working with a feature similarity matrix instead of raw features, we will only have access to K, not $\mathbf{X}$. Fortunately, we can compute $\tilde{\mathbf{K}}$ from $\mathbf{K}$ using the double centering trick:

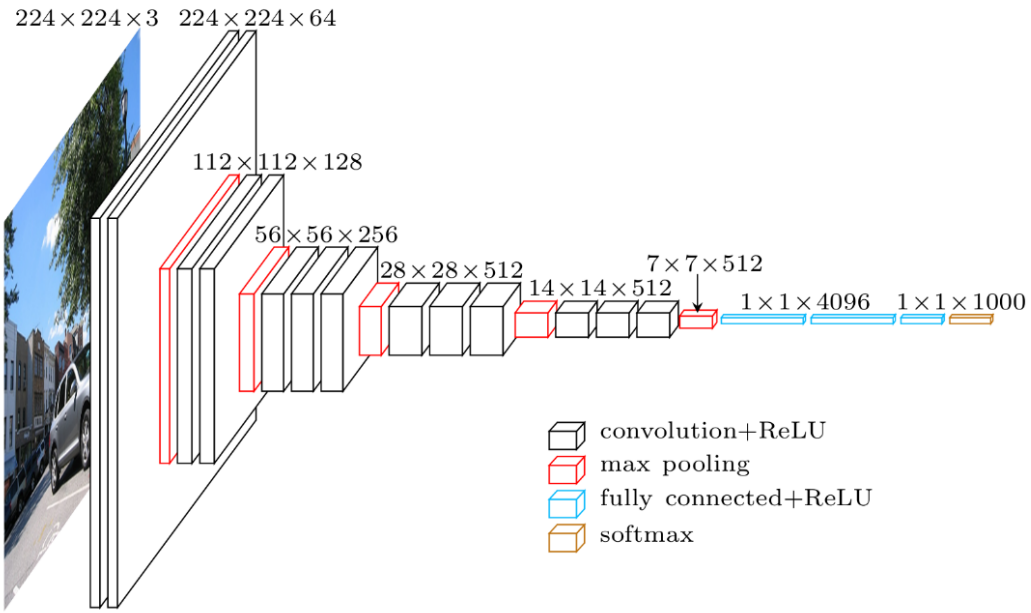
$$\tilde{\mathbf{K}} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top = \mathbf{C}_N \mathbf{K} \mathbf{C}_N = \mathbf{K} - \frac{1}{N}\mathbf{J}\mathbf{K} - \frac{1}{N}\mathbf{K}\mathbf{J} + \frac{1}{N^2}\mathbf{J}\mathbf{K}\mathbf{J}$$

Manipulating data matrices

This subtracts the row means and column means from $\mathbf{K}$, and adds back the global mean that gets subtracted twice, so that both row means and column means of $\tilde{\mathbf{K}}$ are equal to zero.

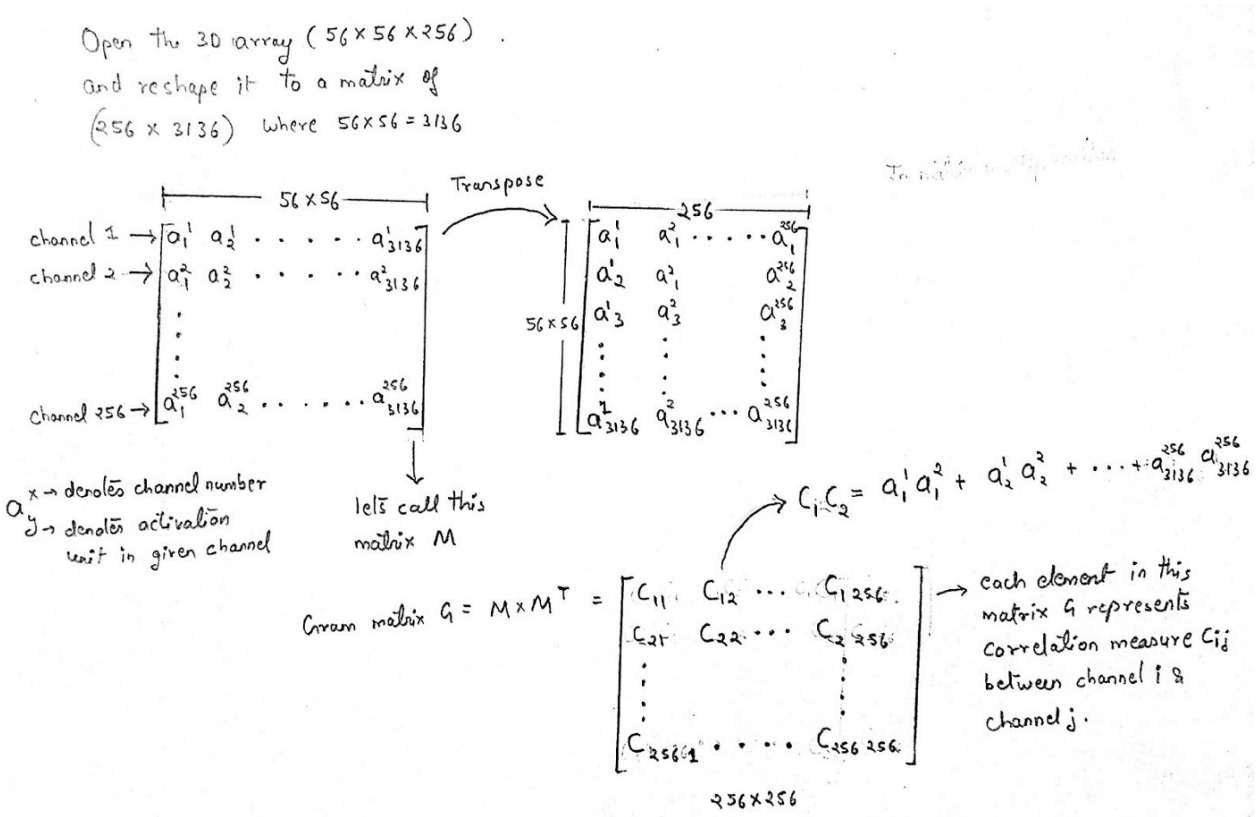
Manipulating data matrices

Gram matrix is often used in Deep learning, such as NN Transfer Style. The principle is simple: we define two distances, one for the content (DC) and one for the style (DS). DC measures how different the content is between two images while DS measures how different the style is between two images. Then, we take a third image, the input, and transform it to minimize both its content-distance with the content-image and its style-distance with the style-image. Actually we use Gram matrix to measure the DS part.



Manipulating data matrices

Calculation Process:



Manipulating data matrices

Distance Matrix

Let $\mathbf{X}$ be $N_x \times D$ datamatrix, and $\mathbf{Y}$ be another $N_y \times D$ datamatrix. We can compute the squared pairwise distances between these using

$$\mathbf{D}_{ij} = (\boldsymbol{x}_i - \boldsymbol{y}_j)^\top (\boldsymbol{x}_i - \boldsymbol{y}_j) = \|\boldsymbol{x}_i\|^2 - 2\boldsymbol{x}_i^\top \boldsymbol{y}_j + \|\boldsymbol{y}_j\|^2$$

Let us now write this in matrix form. Let $\hat{\boldsymbol{x}} = \left[\|\boldsymbol{x}_1\|^2; \cdots; \|\boldsymbol{x}_{N_x}\|^2\right] = \text{diag}\left(\mathbf{X}\mathbf{X}^\top\right)$ be a vector where each element is the squared norm of the examples in $\mathbf{X}$, and define $\hat{\boldsymbol{y}}$ similarly. Then we have

$$\mathbf{D} = \hat{\boldsymbol{x}}\mathbf{1}_{N_y}^\top - 2\mathbf{X}\mathbf{Y}^\top + \mathbf{1}_{N_x}\hat{\boldsymbol{y}}^\top$$

In the case that $\mathbf{X} = \mathbf{Y}$, we have

$$\mathbf{D} = \hat{\boldsymbol{x}}\mathbf{1}_{N}^\top - 2\mathbf{X}\mathbf{X}^\top + \mathbf{1}_{N}\hat{\boldsymbol{x}}^\top$$

Manipulating data matrices

Einstein Summation

Einstein summation, or einsum for short, is a notational shortcut for working with tensors. For example, instead of writing matrix multiplication as $C_{ij} = \sum_k A_{ik} B_{kj}$, we can just write it as $C_{ij} = A_{ik} B_{kj}$, where we drop the $\sum_k$.