Robotic Mapping & Localization

Kaveh Fathian
Assistant Professor
Computer Science Department
Colorado School of Mines

Lec10: Linear Algebra

Reference 1:

Linear Algebra Review and Reference by Zico Kolter

https://www.cs.cmu.edu/~zkolter/course/15-884/linalg-review.pdf

Linear Algebra Review and Reference

Zico Kolter (updated by Chuong Do)

September 30, 2015

Contents

1 Basic Concepts and Notation

Linear algebra provides a way of compactly representing and operating on sets of linear equations. For example, consider the following system of equations:

$$\begin{aligned} 4x_1 - 5x_2 &= -13 \\ -2x_1 + 3x_2 &= 9. \end{aligned}$$

This is two equations and two variables, so as you know from high school algebra, you can find a unique solution for x_1 and x_2 (unless the equations are somehow degenerate, for example if the second equation is simply a multiple of the first, but in the case above there is in fact a unique solution). In matrix notation, we can write the system more compactly as

$$Ax = b$$

with

$$A = \begin{bmatrix} 4 & -5 \\ -2 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} -13 \\ 9 \end{bmatrix}.$$

As we will see shortly, there are many advantages (including the obvious space savings) to analyzing linear equations in this form.

1.1 Basic Notation

We use the following notation:

- By $A \in \mathbb{R}^{m \times n}$ we denote a matrix with m rows and n columns, where the entries of A are real numbers.
- By $x \in \mathbb{R}^n$, we denote a vector with n entries. By convention, an n-dimensional vector is often thought of as a matrix with n rows and 1 column, known as a ***column vector***. If we want to explicitly represent a ***row vector*** — a matrix with 1 row and n columns — we typically write x^T (here x^T denotes the transpose of x, which we will define shortly).
- The ith element of a vector x is denoted x_i:

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

- We use the notation a_{ij} (or A_{ij}, $A_{i,j}$, etc) to denote the entry of A in the ith row and jth column:

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

- We denote the jth column of A by a_j or $A_{:,j}$:

$$A = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix}.$$

- We denote the ith row of A by a_i^T or $A_{i,:}$:

$$A = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix}.$$

- Note that these definitions are ambiguous (for example, the a_1 and a_1^T in the previous two definitions are *not* the same vector). Usually the meaning of the notation should be obvious from its use.

2 Matrix Multiplication

The product of two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$ is the matrix

$$C = AB \in \mathbb{R}^{m \times p},$$

where

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.$$

Note that in order for the matrix product to exist, the number of columns in A must equal the number of rows in B. There are many ways of looking at matrix multiplication, and we'll start by examining a few special cases.

2.1 Vector-Vector Products

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

$$x^T y \in \mathbb{R} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i.$$

Observe that inner products are really just special case of matrix multiplication. Note that it is always the case that $x^T y = y^T x$.

Given vectors $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$ (not necessarily of the same size), $xy^T \in \mathbb{R}^{m \times n}$ is called the ***outer product*** of the vectors. It is a matrix whose entries are given by $(xy^T)_{ij} = x_i y_j$, i.e.,

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

As an example of how the outer product can be useful, let $\mathbf{1} \in \mathbb{R}^n$ denote an n-dimensional vector whose entries are all equal to 1. Furthermore, consider the matrix $A \in \mathbb{R}^{m \times n}$ whose columns are all equal to some vector $x \in \mathbb{R}^m$. Using outer products, we can represent A compactly as,

$$A = \begin{bmatrix} | & | & & | \\ x & x & \cdots & x \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x_1 & x_1 & \cdots & x_1 \\ x_2 & x_2 & \cdots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_m & x_m & \cdots & x_m \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} = x\mathbf{1}^T.$$

2.2 Matrix-Vector Products

Given a matrix $A \in \mathbb{R}^{m \times n}$ and a vector $x \in \mathbb{R}^n$, their product is a vector $y = Ax \in \mathbb{R}^m$. There are a couple ways of looking at matrix-vector multiplication, and we will look at each of them in turn.

If we write A by rows, then we can express Ax as,

$$y = Ax = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} x = \begin{bmatrix} a_1^T x \\ a_2^T x \\ \vdots \\ a_m^T x \end{bmatrix}.$$

In other words, the ith entry of y is equal to the inner product of the ith *row* of A and x, $y_i = a_i^T x$.

Alternatively, let's write A in column form. In this case we see that,

$$y = Ax = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \\ a_1 \\ \\ \end{bmatrix} x_1 + \begin{bmatrix} \\ a_2 \\ \\ \end{bmatrix} x_2 + \ldots + \begin{bmatrix} \\ a_n \\ \\ \end{bmatrix} x_n \ .$$

In other words, y is a ***linear combination*** of the *columns* of A, where the coefficients of the linear combination are given by the entries of x.

So far we have been multiplying on the right by a column vector, but it is also possible to multiply on the left by a row vector. This is written, $y^T = x^T A$ for $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^m$, and $y \in \mathbb{R}^n$. As before, we can express y^T in two obvious ways, depending on whether we express A in terms on its rows or columns. In the first case we express A in terms of its columns, which gives

$$y^T = x^T A = x^T \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} = \begin{bmatrix} x^T a_1 & x^T a_2 & \cdots & x^T a_n \end{bmatrix}$$

which demonstrates that the ith entry of y^T is equal to the inner product of x and the ith *column* of A.

Finally, expressing A in terms of rows we get the final representation of the vector-matrix product,

$$\begin{aligned} y^T &= x^T A \\ &= \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} \\ &= x_1 \begin{bmatrix} — & a_1^T & — \end{bmatrix} + x_2 \begin{bmatrix} — & a_2^T & — \end{bmatrix} + \ldots + x_n \begin{bmatrix} — & a_n^T & — \end{bmatrix} \end{aligned}$$

so we see that y^T is a linear combination of the *rows* of A, where the coefficients for the linear combination are given by the entries of x.

2.3 Matrix-Matrix Products

Armed with this knowledge, we can now look at four different (but, of course, equivalent) ways of viewing the matrix-matrix multiplication $C = AB$ as defined at the beginning of this section.

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)th

entry of C is equal to the inner product of the ith row of A and the jth column of B. Symbolically, this looks like the following,

$$C = AB = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} a_1^T b_1 & a_1^T b_2 & \cdots & a_1^T b_p \\ a_2^T b_1 & a_2^T b_2 & \cdots & a_2^T b_p \\ \vdots & \vdots & \ddots & \vdots \\ a_m^T b_1 & a_m^T b_2 & \cdots & a_m^T b_p \end{bmatrix}.$$

Remember that since $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$, $a_i \in \mathbb{R}^n$ and $b_j \in \mathbb{R}^n$, so these inner products all make sense. This is the most "natural" representation when we represent A by rows and B by columns. Alternatively, we can represent A by columns, and B by rows. This representation leads to a much trickier interpretation of AB as a sum of outer products. Symbolically,

$$C = AB = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & & | \end{bmatrix} \begin{bmatrix} — & b_1^T & — \\ — & b_2^T & — \\ & \vdots & \\ — & b_n^T & — \end{bmatrix} = \sum_{i=1}^{n} a_i b_i^T \ .$$

Put another way, AB is equal to the sum, over all i, of the outer product of the ith column of A and the ith row of B. Since, in this case, $a_i \in \mathbb{R}^m$ and $b_i \in \mathbb{R}^p$, the dimension of the outer product $a_i b_i^T$ is $m \times p$, which coincides with the dimension of C. Chances are, the last equality above may appear confusing to you. If so, take the time to check it for yourself!

Second, 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,

$$C = AB = A \begin{bmatrix} | & | & & | \\ b_1 & b_2 & \cdots & b_p \\ | & | & & | \end{bmatrix} = \begin{bmatrix} | & | & & | \\ Ab_1 & Ab_2 & \cdots & Ab_p \\ | & | & & | \end{bmatrix}.$$

Here the ith column of C is given by the matrix-vector product with the vector on the right, $c_i = Ab_i$. These matrix-vector products can in turn be interpreted using both viewpoints given in the previous subsection. 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 C. Symbolically,

$$C = AB = \begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} B = \begin{bmatrix} — & a_1^T B & — \\ — & a_2^T B & — \\ & \vdots & \\ — & a_m^T B & — \end{bmatrix}.$$

Here the ith row of C is given by the matrix-vector product with the vector on the left, $c_i^T = a_i^T B$.

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.

In addition to this, it is useful to know a few basic properties of matrix multiplication at a higher level:

- Matrix multiplication is associative: $(AB)C = A(BC)$.
- Matrix multiplication is distributive: $A(B + C) = AB + AC$.
- Matrix multiplication is, in general, *not* commutative; that is, it can be the case that $AB \neq BA$. (For example, if $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times q}$, the matrix product BA does not even exist if m and q are not equal!)

If you are not familiar with these properties, take the time to verify them for yourself. For example, to check the associativity of matrix multiplication, suppose that $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times p}$, and $C \in \mathbb{R}^{p\times q}$. Note that $AB \in \mathbb{R}^{m\times p}$, so $(AB)C \in \mathbb{R}^{m\times q}$. Similarly, $BC \in \mathbb{R}^{n\times q}$, so $A(BC) \in \mathbb{R}^{m\times q}$. Thus, the dimensions of the resulting matrices agree. To show that matrix multiplication is associative, it suffices to check that the (i, j)th entry of $(AB)C$ is equal to the (i, j)th entry of $A(BC)$. We can verify this directly using the definition of matrix multiplication:

$$\begin{aligned}
((AB)C)_{ij} &= \sum_{k=1}^{p}(AB)_{ik}C_{kj} = \sum_{k=1}^{p}\left(\sum_{l=1}^{n}A_{il}B_{lk}\right)C_{kj} \\
&= \sum_{k=1}^{p}\left(\sum_{l=1}^{n}A_{il}B_{lk}C_{kj}\right) = \sum_{l=1}^{n}\left(\sum_{k=1}^{p}A_{il}B_{lk}C_{kj}\right) \\
&= \sum_{l=1}^{n}A_{il}\left(\sum_{k=1}^{p}B_{lk}C_{kj}\right) = \sum_{l=1}^{n}A_{il}(BC)_{lj} = (A(BC))_{ij}.
\end{aligned}$$

Here, the first and last two equalities simply use the definition of matrix multiplication, the third and fifth equalities use the distributive property for *scalar multiplication over addition*, and the fourth equality uses the *commutative and associativity of scalar addition*. This technique for proving matrix properties by reduction to simple scalar properties will come up often, so make sure you're familiar with it.

3 Operations and Properties

In this section we present several operations and properties of matrices and vectors. Hopefully a great deal of this will be review for you, so the notes can just serve as a reference for these topics.

3.1 The Identity Matrix and Diagonal Matrices

The ***identity matrix***, denoted $I \in \mathbb{R}^{n \times n}$, is a square matrix with ones on the diagonal and zeros everywhere else. That is,

$$I_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

It has the property that for all $A \in \mathbb{R}^{m \times n}$,

$$AI = A = IA.$$

Note that in some sense, the notation for the identity matrix is ambiguous, since it does not specify the dimension of I. Generally, the dimensions of I are inferred from context so as to make matrix multiplication possible. For example, in the equation above, the I in $AI = A$ is an $n \times n$ matrix, whereas the I in $A = IA$ is an $m \times m$ matrix.

A ***diagonal matrix*** is a matrix where all non-diagonal elements are 0. This is typically denoted $D = \text{diag}(d_1, d_2, \ldots, d_n)$, with

$$D_{ij} = \begin{cases} d_i & i = j \\ 0 & i \neq j \end{cases}$$

Clearly, $I = \text{diag}(1, 1, \ldots, 1)$.

3.2 The Transpose

The ***transpose*** of a matrix results from "flipping" the rows and columns. Given a matrix $A \in \mathbb{R}^{m \times n}$, its transpose, written $A^T \in \mathbb{R}^{n \times m}$, is the $n \times m$ matrix whose entries are given by

$$(A^T)_{ij} = A_{ji}.$$

We have in fact already been using the transpose when describing row vectors, since the transpose of a column vector is naturally a row vector.

The following properties of transposes are easily verified:

- $(A^T)^T = A$
- $(AB)^T = B^T A^T$
- $(A + B)^T = A^T + B^T$

3.3 Symmetric Matrices

A square matrix $A \in \mathbb{R}^{n \times n}$ is ***symmetric*** if $A = A^T$. It is ***anti-symmetric*** if $A = -A^T$. It is easy to show that for any matrix $A \in \mathbb{R}^{n \times n}$, the matrix $A + A^T$ is symmetric and the

matrix $A - A^T$ is anti-symmetric. From this it follows that any square matrix $A \in \mathbb{R}^{n \times n}$ can be represented as a sum of a symmetric matrix and an anti-symmetric matrix, since

$$A = \frac{1}{2}(A + A^T) + \frac{1}{2}(A - A^T)$$

and the first matrix on the right is symmetric, while the second is anti-symmetric. It turns out that symmetric matrices occur a great deal in practice, and they have many nice properties which we will look at shortly. It is common to denote the set of all symmetric matrices of size n as $\mathbb{S}^n$, so that $A \in \mathbb{S}^n$ means that A is a symmetric $n \times n$ matrix;

3.4 The Trace

The ***trace*** of a square matrix $A \in \mathbb{R}^{n \times n}$, denoted $\mathrm{tr}(A)$ (or just $\mathrm{tr}A$ if the parentheses are obviously implied), is the sum of diagonal elements in the matrix:

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

As described in the CS229 lecture notes, the trace has the following properties (included here for the sake of completeness):

- For $A \in \mathbb{R}^{n \times n}$, $\mathrm{tr}A = \mathrm{tr}A^T$.
- For $A, B \in \mathbb{R}^{n \times n}$, $\mathrm{tr}(A + B) = \mathrm{tr}A + \mathrm{tr}B$.
- For $A \in \mathbb{R}^{n \times n}$, $t \in \mathbb{R}$, $\mathrm{tr}(tA) = t\ \mathrm{tr}A$.
- For A, B such that AB is square, $\mathrm{tr}AB = \mathrm{tr}BA$.
- For A, B, C such that ABC is square, $\mathrm{tr}ABC = \mathrm{tr}BCA = \mathrm{tr}CAB$, and so on for the product of more matrices.

As an example of how these properties can be proven, we'll consider the fourth property given above. Suppose that $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times m}$ (so that $AB \in \mathbb{R}^{m \times m}$ is a square matrix). Observe that $BA \in \mathbb{R}^{n \times n}$ is also a square matrix, so it makes sense to apply the trace operator to it. To verify that $\mathrm{tr}AB = \mathrm{tr}BA$, note that

$$\begin{aligned}
\mathrm{tr}AB &= \sum_{i=1}^{m} (AB)_{ii} = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} A_{ij} B_{ji} \right) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ji} = \sum_{j=1}^{n} \sum_{i=1}^{m} B_{ji} A_{ij} \\
&= \sum_{j=1}^{n} \left(\sum_{i=1}^{m} B_{ji} A_{ij} \right) = \sum_{j=1}^{n} (BA)_{jj} = \mathrm{tr}BA.
\end{aligned}$$

Here, the first and last two equalities use the definition of the trace operator and matrix multiplication. The fourth equality, where the main work occurs, uses the commutativity of scalar multiplication in order to reverse the order of the terms in each product, and the commutativity and associativity of scalar addition in order to rearrange the order of the summation.

3.5 Norms

A ***norm*** of a vector $\|x\|$ is informally a measure of the "length" of the vector. For example, we have the commonly-used Euclidean or ℓ_2 norm,

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}.$$

Note that $\|x\|_2^2 = x^T x$.

More formally, a norm is any function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies 4 properties:

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

Other examples of norms are the ℓ_1 norm,

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

and the ℓ_∞ norm,

$$\|x\|_\infty = \max_i |x_i|.$$

In fact, all three norms presented so far are examples of the family of ℓ_p norms, which are parameterized by a real number $p \geq 1$, and defined as

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

Norms can also be defined for matrices, such as the Frobenius norm,

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

Many other norms exist, but they are beyond the scope of this review.

3.6 Linear Independence and Rank

A set of vectors $\{x_1, x_2, \ldots x_n\} \subset \mathbb{R}^m$ is said to be ***(linearly) independent*** if no vector can be represented as a linear combination of the remaining vectors. Conversely, if one vector belonging to the set *can* be represented as a linear combination of the remaining vectors, then the vectors are said to be ***(linearly) dependent***. That is, if

$$x_n = \sum_{i=1}^{n-1} \alpha_i x_i$$

for some scalar values $\alpha_1, \ldots, \alpha_{n-1} \in \mathbb{R}$, then we say that the vectors $x_1, \ldots, x_n$ are linearly dependent; otherwise, the vectors are linearly independent. For example, the vectors

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \quad x_2 = \begin{bmatrix} 4 \\ 1 \\ 5 \end{bmatrix} \quad x_3 = \begin{bmatrix} 2 \\ -3 \\ -1 \end{bmatrix}$$

are linearly dependent because $x_3 = -2x_1 + x_2$.

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. With some abuse of terminology, this is often referred to simply as the number of linearly independent columns of A. In the same way, the ***row rank*** is the largest number of rows of A that constitute a linearly independent set.

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 (though we will not prove this), and so both quantities are referred to collectively as the ***rank*** of A, denoted as $\text{rank}(A)$. The following are some basic properties of the rank:

- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) \leq \min(m, n)$. If $\text{rank}(A) = \min(m, n)$, then A is said to be ***full rank***.
- For $A \in \mathbb{R}^{m \times n}$, $\text{rank}(A) = \text{rank}(A^T)$.
- For $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{n \times p}$, $\text{rank}(AB) \leq \min(\text{rank}(A), \text{rank}(B))$.
- For $A, B \in \mathbb{R}^{m \times n}$, $\text{rank}(A + B) \leq \text{rank}(A) + \text{rank}(B)$.

3.7 The Inverse

The ***inverse*** of a square matrix $A \in \mathbb{R}^{n \times n}$ is denoted A^{-1}, and is the unique matrix such that

$$A^{-1}A = I = AA^{-1}.$$

Note that not all matrices have inverses. Non-square matrices, for example, do not have inverses by definition. However, for some square matrices A, it may still be the case that

A^{-1} may not exist. In particular, we say that A is ***invertible*** or ***non-singular*** if A^{-1} exists and ***non-invertible*** or ***singular*** otherwise.[1]

In order for a square matrix A to have an inverse A^{-1}, then A must be full rank. We will soon see that there are many alternative sufficient and necessary conditions, in addition to full rank, for invertibility.

The following are properties of the inverse; all assume that $A, B \in \mathbb{R}^{n \times n}$ are non-singular:

- $(A^{-1})^{-1} = A$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(A^{-1})^T = (A^T)^{-1}$. For this reason this matrix is often denoted A^{-T}.

As an example of how the inverse is used, consider the linear system of equations, $Ax = b$ where $A \in \mathbb{R}^{n \times n}$, and $x, b \in \mathbb{R}^n$. If A is nonsingular (i.e., invertible), then $x = A^{-1}b$. (What if $A \in \mathbb{R}^{m \times n}$ is not a square matrix? Does this work?)

3.8 Orthogonal Matrices

Two vectors $x, y \in \mathbb{R}^n$ are ***orthogonal*** if $x^T y = 0$. A vector $x \in \mathbb{R}^n$ is ***normalized*** if $\|x\|_2 = 1$. A square matrix $U \in \mathbb{R}^{n \times n}$ is ***orthogonal*** (note the different meanings when talking about vectors versus matrices) if all its columns are orthogonal to each other and are normalized (the columns are then referred to as being ***orthonormal***).

It follows immediately from the definition of orthogonality and normality that

$$U^T U = I = U U^T.$$

In other words, the inverse of an orthogonal matrix is its transpose. Note that if U is not square — i.e., $U \in \mathbb{R}^{m \times n}$, $n < m$ — but its columns are still orthonormal, then $U^T U = I$, but $UU^T \neq I$. We generally only use the term orthogonal to describe the previous case, where U is square.

Another nice property of orthogonal matrices is that operating on a vector with an orthogonal matrix will not change its Euclidean norm, i.e.,

$$\|Ux\|_2 = \|x\|_2$$

for any $x \in \mathbb{R}^n$, $U \in \mathbb{R}^{n \times n}$ orthogonal.

3.9 Range and Nullspace of a Matrix

The ***span*** of a set of vectors $\{x_1, x_2, \ldots x_n\}$ is the set of all vectors that can be expressed as a linear combination of $\{x_1, \ldots, x_n\}$. That is,

$$\text{span}(\{x_1, \ldots x_n\}) = \left\{ v : v = \sum_{i=1}^{n} \alpha_i x_i, \quad \alpha_i \in \mathbb{R} \right\}.$$

[1] It's easy to get confused and think that non-singular means non-invertible. But in fact, it means the opposite! Watch out!

It can be shown that if $\{x_1, \ldots, x_n\}$ is a set of n linearly independent vectors, where each $x_i \in \mathbb{R}^n$, then $\text{span}(\{x_1, \ldots x_n\}) = \mathbb{R}^n$. In other words, *any* vector $v \in \mathbb{R}^n$ can be written as a linear combination of x_1 through x_n. The ***projection*** of a vector $y \in \mathbb{R}^m$ onto the span of $\{x_1, \ldots, x_n\}$ (here we assume $x_i \in \mathbb{R}^m$) is the vector $v \in \text{span}(\{x_1, \ldots x_n\})$, such that v is as close as possible to y, as measured by the Euclidean norm $\|v - y\|_2$. We denote the projection as $\text{Proj}(y; \{x_1, \ldots, x_n\})$ and can define it formally as,

$$\text{Proj}(y; \{x_1, \ldots x_n\}) = \text{argmin}_{v \in \text{span}(\{x_1, \ldots, x_n\})} \|y - v\|_2.$$

The ***range*** (sometimes also called the columnspace) of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $\mathcal{R}(A)$, is the the span of the columns of A. In other words,

$$\mathcal{R}(A) = \{v \in \mathbb{R}^m : v = Ax, x \in \mathbb{R}^n\}.$$

Making a few technical assumptions (namely that A is full rank and that $n < m$), the projection of a vector $y \in \mathbb{R}^m$ onto the range of A is given by,

$$\text{Proj}(y; A) = \text{argmin}_{v \in \mathcal{R}(A)} \|v - y\|_2 = A(A^T A)^{-1} A^T y \ .$$

This last equation should look extremely familiar, since it is almost the same formula we derived in class (and which we will soon derive again) for the least squares estimation of parameters. Looking at the definition for the projection, it should not be too hard to convince yourself that this is in fact the same objective that we minimized in our least squares problem (except for a squaring of the norm, which doesn't affect the optimal point) and so these problems are naturally very connected. When A contains only a single column, $a \in \mathbb{R}^m$, this gives the special case for a projection of a vector on to a line:

$$\text{Proj}(y; a) = \frac{aa^T}{a^T a} y \ .$$

The ***nullspace*** of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $\mathcal{N}(A)$ is the set of all vectors that equal 0 when multiplied by A, i.e.,

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

Note that vectors in $\mathcal{R}(A)$ are of size m, while vectors in the $\mathcal{N}(A)$ are of size n, so vectors in $\mathcal{R}(A^T)$ and $\mathcal{N}(A)$ are both in $\mathbb{R}^n$. In fact, we can say much more. It turns out that

$$\left\{w : w = u + v, u \in \mathcal{R}(A^T), v \in \mathcal{N}(A)\right\} = \mathbb{R}^n \text{ and } \mathcal{R}(A^T) \cap \mathcal{N}(A) = \{\mathbf{0}\} \ .$$

In other words, $\mathcal{R}(A^T)$ and $\mathcal{N}(A)$ are disjoint subsets that together span the entire space of $\mathbb{R}^n$. Sets of this type are called ***orthogonal complements***, and we denote this $\mathcal{R}(A^T) = \mathcal{N}(A)^\perp$.

3.10 The Determinant

The ***determinant*** of a square matrix $A \in \mathbb{R}^{n \times n}$, is a function $\det : \mathbb{R}^{n \times n} \to \mathbb{R}$, and is denoted $|A|$ or $\det A$ (like the trace operator, we usually omit parentheses). Algebraically, one could write down an explicit formula for the determinant of A, but this unfortunately gives little intuition about its meaning. Instead, we'll start out by providing a geometric interpretation of the determinant and then visit some of its specific algebraic properties afterwards.

Given a matrix

$$\begin{bmatrix} — & a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_n^T & — \end{bmatrix},$$

consider the set of points $S \subset \mathbb{R}^n$ formed by taking all possible linear combinations of the row vectors $a_1, \ldots, a_n \in \mathbb{R}^n$ of A, where the coefficients of the linear combination are all between 0 and 1; that is, the set S is the restriction of $\text{span}(\{a_1, \ldots, a_n\})$ to only those linear combinations whose coefficients $\alpha_1, \ldots, \alpha_n$ satisfy $0 \leq \alpha_i \leq 1$, $i = 1, \ldots, n$. Formally,

$$S = \{v \in \mathbb{R}^n : v = \sum_{i=1}^{n} \alpha_i a_i \text{ where } 0 \leq \alpha_i \leq 1, i = 1, \ldots, n\}.$$

The absolute value of the determinant of A, it turns out, is a measure of the "volume" of the set S.[2]

For example, consider the 2×2 matrix,

$$A = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix}. \tag{1}$$

Here, the rows of the matrix are

$$a_1 = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \quad a_2 = \begin{bmatrix} 3 \\ 2 \end{bmatrix}.$$

The set S corresponding to these rows is shown in Figure 1. For two-dimensional matrices, S generally has the shape of a *parallelogram*. In our example, the value of the determinant is $|A| = -7$ (as can be computed using the formulas shown later in this section), so the area of the parallelogram is 7. (Verify this for yourself!)

In three dimensions, the set S corresponds to an object known as a *parallelepiped* (a three-dimensional box with skewed sides, such that every face has the shape of a parallelogram). The absolute value of the determinant of the 3×3 matrix whose rows define S give the three-dimensional volume of the parallelepiped. In even higher dimensions, the set S is an object known as an n-dimensional *parallelotope*.

[2]Admittedly, we have not actually defined what we mean by "volume" here, but hopefully the intuition should be clear enough. When $n = 2$, our notion of "volume" corresponds to the area of S in the Cartesian plane. When $n = 3$, "volume" corresponds with our usual notion of volume for a three-dimensional object.

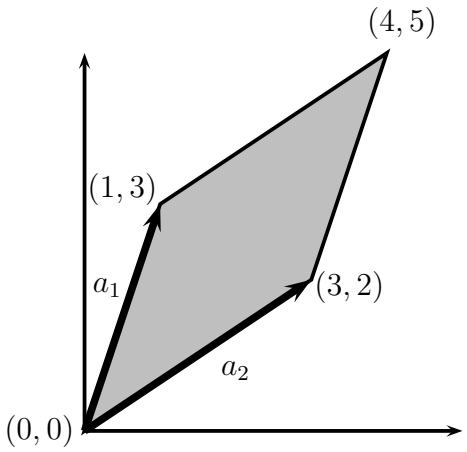


Figure 1: Illustration of the determinant for the 2×2 matrix A given in (1). Here, a_1 and a_2 are vectors corresponding to the rows of A, and the set S corresponds to the shaded region (i.e., the parallelogram). The absolute value of the determinant, $|\det A| = 7$, is the area of the parallelogram.

Algebraically, the determinant satisfies the following three properties (from which all other properties follow, including the general formula):

1. The determinant of the identity is 1, $|I| = 1$. (Geometrically, the volume of a unit hypercube is 1).

2. Given a matrix $A \in \mathbb{R}^{n \times n}$, if we multiply a single row in A by a scalar $t \in \mathbb{R}$, then the determinant of the new matrix is $t|A|$,

$$\left| \begin{bmatrix} — & t\, a_1^T & — \\ — & a_2^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} \right| = t|A|.$$

   (Geometrically, multiplying one of the sides of the set S by a factor t causes the volume to increase by a factor t.)

3. If we exchange any two rows a_i^T and a_j^T of A, then the determinant of the new matrix is $-|A|$, for example

$$\left| \begin{bmatrix} — & a_2^T & — \\ — & a_1^T & — \\ & \vdots & \\ — & a_m^T & — \end{bmatrix} \right| = -|A|.$$

In case you are wondering, it is not immediately obvious that a function satisfying the above three properties exists. In fact, though, such a function does exist, and is unique (which we will not prove here).

Several properties that follow from the three properties above include:

- For $A \in \mathbb{R}^{n\times n}$, $|A| = |A^T|$.
- For $A, B \in \mathbb{R}^{n\times n}$, $|AB| = |A||B|$.
- For $A \in \mathbb{R}^{n\times n}$, $|A| = 0$ if and only if A is singular (i.e., non-invertible). (If A is singular then it does not have full rank, and hence its columns are linearly dependent. In this case, the set S corresponds to a "flat sheet" within the n-dimensional space and hence has zero volume.)
- For $A \in \mathbb{R}^{n\times n}$ and A non-singular, $|A^{-1}| = 1/|A|$.

Before giving the general definition for the determinant, we define, for $A \in \mathbb{R}^{n\times n}$, $A_{\setminus i,\setminus j} \in \mathbb{R}^{(n-1)\times(n-1)}$ to be the *matrix* that results from deleting the ith row and jth column from A. The general (recursive) formula for the determinant is

$$\begin{aligned} |A| &= \sum_{i=1}^{n} (-1)^{i+j} a_{ij} |A_{\setminus i,\setminus j}| \quad \text{(for any } j \in 1,\ldots,n) \\ &= \sum_{j=1}^{n} (-1)^{i+j} a_{ij} |A_{\setminus i,\setminus j}| \quad \text{(for any } i \in 1,\ldots,n) \end{aligned}$$

with the initial case that $|A| = a_{11}$ for $A \in \mathbb{R}^{1\times 1}$. If we were to expand this formula completely for $A \in \mathbb{R}^{n\times n}$, there would be a total of $n!$ (n factorial) different terms. For this reason, we hardly ever explicitly write the complete equation of the determinant for matrices bigger than 3×3. However, the equations for determinants of matrices up to size 3×3 are fairly common, and it is good to know them:

$$\begin{aligned} |[a_{11}]| &= a_{11} \\ \left|\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right| &= a_{11}a_{22} - a_{12}a_{21} \\ \left|\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\right| &= \begin{array}{l} a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} \\ \quad -a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} - a_{13}a_{22}a_{31} \end{array} \end{aligned}$$

The ***classical adjoint*** (often just called the adjoint) of a matrix $A \in \mathbb{R}^{n\times n}$, is denoted $\text{adj}(A)$, and defined as

$$\text{adj}(A) \in \mathbb{R}^{n\times n}, \quad (\text{adj}(A))_{ij} = (-1)^{i+j} |A_{\setminus j,\setminus i}|$$

(note the switch in the indices $A_{\setminus j,\setminus i}$). It can be shown that for any nonsingular $A \in \mathbb{R}^{n\times n}$,

$$A^{-1} = \frac{1}{|A|}\text{adj}(A) \ .$$

While this is a nice "explicit" formula for the inverse of matrix, we should note that, numerically, there are in fact much more efficient ways of computing the inverse.

3.11 Quadratic Forms and Positive Semidefinite 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***. Written explicitly, we see that

$$x^T Ax = \sum_{i=1}^{n} x_i (Ax)_i = \sum_{i=1}^{n} x_i \left(\sum_{j=1}^{n} A_{ij} x_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j \quad .$$

Note that,

$$x^T Ax = (x^T Ax)^T = x^T A^T x = x^T \left(\frac{1}{2} A + \frac{1}{2} A^T \right) x,$$

where the first equality follows from the fact that the transpose of a scalar is equal to itself, and the second equality follows from the fact that we are averaging two quantities which are themselves equal. From this, we can conclude that only the symmetric part of A contributes to the quadratic form. For this reason, we often implicitly assume that the matrices appearing in a quadratic form are symmetric.

We give the following definitions:

- A symmetric matrix $A \in \mathbb{S}^n$ is ***positive definite*** (PD) if for all non-zero vectors $x \in \mathbb{R}^n$, $x^T Ax > 0$. This is usually denoted $A \succ 0$ (or just $A > 0$), and often times the set of all positive definite matrices is denoted $\mathbb{S}^n_{++}$.
- A symmetric matrix $A \in \mathbb{S}^n$ is ***positive semidefinite*** (PSD) if for all vectors $x^T Ax \geq 0$. This is written $A \succeq 0$ (or just $A \geq 0$), and the set of all positive semidefinite matrices is often denoted $\mathbb{S}^n_{+}$.
- Likewise, a symmetric matrix $A \in \mathbb{S}^n$ is ***negative definite*** (ND), denoted $A \prec 0$ (or just $A < 0$) if for all non-zero $x \in \mathbb{R}^n$, $x^T Ax < 0$.
- Similarly, a symmetric matrix $A \in \mathbb{S}^n$ is ***negative semidefinite*** (NSD), denoted $A \preceq 0$ (or just $A \leq 0$) if for all $x \in \mathbb{R}^n$, $x^T Ax \leq 0$.
- Finally, a symmetric matrix $A \in \mathbb{S}^n$ is ***indefinite***, if it is neither positive semidefinite nor negative semidefinite — i.e., if there exists $x_1, x_2 \in \mathbb{R}^n$ such that $x_1^T Ax_1 > 0$ and $x_2^T Ax_2 < 0$.

It should be obvious that if A is positive definite, then $-A$ is negative definite and vice versa. Likewise, if A is positive semidefinite then $-A$ is negative semidefinite and vice versa. If A is indefinite, then so is $-A$.

One important property of positive definite and negative definite matrices is that they are always full rank, and hence, invertible. To see why this is the case, suppose that some matrix $A \in \mathbb{R}^{n \times n}$ is not full rank. Then, suppose that the jth column of A is expressible as a linear combination of other $n - 1$ columns:

$$a_j = \sum_{i \neq j} x_i a_i,$$

for some $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n \in \mathbb{R}$. Setting $x_j = -1$, we have

$$Ax = \sum_{i=1}^{n} x_i a_i = 0.$$

But this implies $x^T Ax = 0$ for some non-zero vector x, so A must be neither positive definite nor negative definite. Therefore, if A is either positive definite or negative definite, it must be full rank.

Finally, there is one type of positive definite matrix that comes up frequently, and so deserves some special mention. Given any matrix $A \in \mathbb{R}^{m \times n}$ (not necessarily symmetric or even square), the matrix $G = A^T A$ (sometimes called a ***Gram matrix***) is always positive semidefinite. Further, if $m \geq n$ (and we assume for convenience that A is full rank), then $G = A^T A$ is positive definite.

3.12 Eigenvalues and Eigenvectors

Given a square matrix $A \in \mathbb{R}^{n \times n}$, we say that $\lambda \in \mathbb{C}$ is an ***eigenvalue*** of A and $x \in \mathbb{C}^n$ is the corresponding ***eigenvector***[3] if

$$Ax = \lambda x, \quad x \neq 0.$$

Intuitively, this definition means that multiplying A by the vector x results in a new vector that points in the same direction as x, but scaled by a factor λ. Also note that for any eigenvector $x \in \mathbb{C}^n$, and scalar $t \in \mathbb{C}$, $A(cx) = cAx = c\lambda x = \lambda(cx)$, so cx is also an eigenvector. For this reason when we talk about "the" eigenvector associated with λ, we usually assume that the eigenvector is normalized to have length 1 (this still creates some ambiguity, since x and $-x$ will both be eigenvectors, but we will have to live with this).

We can rewrite the equation above to state that (λ, x) is an eigenvalue-eigenvector pair of A if,

$$(\lambda I - A)x = 0, \quad x \neq 0.$$

But $(\lambda I - A)x = 0$ has a non-zero solution to x if and only if $(\lambda I - A)$ has a non-empty nullspace, which is only the case if $(\lambda I - A)$ is singular, i.e.,

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

We can now use the previous definition of the determinant to expand this expression into a (very large) polynomial in λ, where λ will have maximum degree n. We then find the n (possibly complex) roots of this polynomial to find the n eigenvalues $\lambda_1, \ldots, \lambda_n$. To find the eigenvector corresponding to the eigenvalue λ_i, we simply solve the linear equation $(\lambda_i I - A)x = 0$. It should be noted that this is not the method which is actually used

[3]Note that λ and the entries of x are actually in $\mathbb{C}$, the set of complex numbers, not just the reals; we will see shortly why this is necessary. Don't worry about this technicality for now, you can think of complex vectors in the same way as real vectors.

in practice to numerically compute the eigenvalues and eigenvectors (remember that the complete expansion of the determinant has $n!$ terms); it is rather a mathematical argument.

The following are properties of eigenvalues and eigenvectors (in all cases assume $A \in \mathbb{R}^{n \times n}$ has eigenvalues $\lambda_i, \ldots, \lambda_n$ and associated eigenvectors $x_1, \ldots x_n$):

- The trace of a A is equal to the sum of its eigenvalues,

$$\mathrm{tr}A = \sum_{i=1}^{n} \lambda_i.$$

- The determinant of A is equal to the product of its eigenvalues,

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

- The rank of A is equal to the number of non-zero eigenvalues of A.
- If A is non-singular then $1/\lambda_i$ is an eigenvalue of A^{-1} with associated eigenvector x_i, i.e., $A^{-1}x_i = (1/\lambda_i)x_i$. (To prove this, take the eigenvector equation, $Ax_i = \lambda_i x_i$ and left-multiply each side by A^{-1}.)
- The eigenvalues of a diagonal matrix $D = \mathrm{diag}(d_1, \ldots d_n)$ are just the diagonal entries $d_1, \ldots d_n$.

We can write all the eigenvector equations simultaneously as

$$AX = X\Lambda$$

where the columns of $X \in \mathbb{R}^{n \times n}$ are the eigenvectors of A and Λ is a diagonal matrix whose entries are the eigenvalues of A, i.e.,

$$X \in \mathbb{R}^{n \times n} = \begin{bmatrix} | & | & & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & & | \end{bmatrix}, \quad \Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n).$$

If the eigenvectors of A are linearly independent, then the matrix X will be invertible, so $A = X\Lambda X^{-1}$. A matrix that can be written in this form is called ***diagonalizable***.

3.13 Eigenvalues and Eigenvectors of Symmetric Matrices

Two remarkable properties come about when we look at the eigenvalues and eigenvectors of a symmetric matrix $A \in \mathbb{S}^n$. First, it can be shown that all the eigenvalues of A are real. Secondly, the eigenvectors of A are orthonormal, i.e., the matrix X defined above is an orthogonal matrix (for this reason, we denote the matrix of eigenvectors as U in this case).

We can therefore represent A as $A = U\Lambda U^T$, remembering from above that the inverse of an orthogonal matrix is just its transpose.

Using this, we can show that the definiteness of a matrix depends entirely on the sign of its eigenvalues. Suppose $A \in \mathbb{S}^n = U\Lambda U^T$. Then

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

where $y = U^T x$ (and since U is full rank, any vector $y \in \mathbb{R}^n$ can be represented in this form). Because y_i^2 is always positive, the sign of this expression depends entirely on the λ_i's. If all $\lambda_i > 0$, then the matrix is positive definite; if all $\lambda_i \geq 0$, it is positive semidefinite. Likewise, if all $\lambda_i < 0$ or $\lambda_i \leq 0$, then A is negative definite or negative semidefinite respectively. Finally, if A has both positive and negative eigenvalues, it is indefinite.

An application where eigenvalues and eigenvectors come up frequently is in maximizing some function of a matrix. In particular, for a matrix $A \in \mathbb{S}^n$, consider the following maximization problem,

$$\max_{x \in \mathbb{R}^n} \; x^T Ax \quad \text{subject to } \|x\|_2^2 = 1$$

i.e., we want to find the vector (of norm 1) which maximizes the quadratic form. Assuming the eigenvalues are ordered as $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$, the optimal x for this optimization problem is x_1, the eigenvector corresponding to λ_1. In this case the maximal value of the quadratic form is λ_1. Similarly, the optimal solution to the minimization problem,

$$\min_{x \in \mathbb{R}^n} \; x^T Ax \quad \text{subject to } \|x\|_2^2 = 1$$

is x_n, the eigenvector corresponding to λ_n, and the minimal value is λ_n. This can be proved by appealing to the eigenvector-eigenvalue form of A and the properties of orthogonal matrices. However, in the next section we will see a way of showing it directly using matrix calculus.

4 Matrix Calculus

While the topics in the previous sections are typically covered in a standard course on linear algebra, one topic that does not seem to be covered very often (and which we will use extensively) is the extension of calculus to the vector setting. Despite the fact that all the actual calculus we use is relatively trivial, the notation can often make things look much more difficult than they are. In this section we present some basic definitions of matrix calculus and provide a few examples.

4.1 The Gradient

Suppose that $f : \mathbb{R}^{m \times n} \to \mathbb{R}$ is a function that takes as input a matrix A of size $m \times n$ and returns a real value. Then the ***gradient*** of f (with respect to $A \in \mathbb{R}^{m \times n}$) is the matrix of

partial derivatives, defined as:

$$\nabla_A f(A) \in \mathbb{R}^{m\times n} = \begin{bmatrix} \frac{\partial f(A)}{\partial A_{11}} & \frac{\partial f(A)}{\partial A_{12}} & \cdots & \frac{\partial f(A)}{\partial A_{1n}} \\ \frac{\partial f(A)}{\partial A_{21}} & \frac{\partial f(A)}{\partial A_{22}} & \cdots & \frac{\partial f(A)}{\partial A_{2n}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(A)}{\partial A_{m1}} & \frac{\partial f(A)}{\partial A_{m2}} & \cdots & \frac{\partial f(A)}{\partial A_{mn}} \end{bmatrix}$$

i.e., an $m \times n$ matrix with

$$(\nabla_A f(A))_{ij} = \frac{\partial f(A)}{\partial A_{ij}}.$$

Note that the size of $\nabla_A f(A)$ is always the same as the size of A. So if, in particular, A is just a vector $x \in \mathbb{R}^n$,

$$\nabla_x f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

It is very important to remember that the gradient of a function is *only* defined if the function is real-valued, that is, if it returns a scalar value. We can not, for example, take the gradient of $Ax, A \in \mathbb{R}^{n\times n}$ with respect to x, since this quantity is vector-valued.

It follows directly from the equivalent properties of partial derivatives that:

- $\nabla_x(f(x) + g(x)) = \nabla_x f(x) + \nabla_x g(x)$.
- For $t \in \mathbb{R}$, $\nabla_x(t\ f(x)) = t\nabla_x f(x)$.

In principle, gradients are a natural extension of partial derivatives to functions of multiple variables. In practice, however, working with gradients can sometimes be tricky for notational reasons. For example, suppose that $A \in \mathbb{R}^{m\times n}$ is a matrix of fixed coefficients and suppose that $b \in \mathbb{R}^m$ is a vector of fixed coefficients. Let $f : \mathbb{R}^m \to \mathbb{R}$ be the function defined by $f(z) = z^T z$, such that $\nabla_z f(z) = 2z$. But now, consider the expression,

$$\nabla f(Ax).$$

How should this expression be interpreted? There are at least two possibilities:

1. In the first interpretation, recall that $\nabla_z f(z) = 2z$. Here, we interpret $\nabla f(Ax)$ as evaluating the gradient at the point Ax, hence,

$$\nabla f(Ax) = 2(Ax) = 2Ax \in \mathbb{R}^m.$$

2. In the second interpretation, we consider the quantity $f(Ax)$ as a function of the input variables x. More formally, let $g(x) = f(Ax)$. Then in this interpretation,

$$\nabla f(Ax) = \nabla_x g(x) \in \mathbb{R}^n.$$

Here, we can see that these two interpretations are indeed different. One interpretation yields an m-dimensional vector as a result, while the other interpretation yields an n-dimensional vector as a result! How can we resolve this?

Here, the key is to make explicit the variables which we are differentiating with respect to. In the first case, we are differentiating the function f with respect to its arguments z and then substituting the argument Ax. In the second case, we are differentiating the composite function $g(x) = f(Ax)$ with respect to x directly. We denote the first case as $\nabla_z f(Ax)$ and the second case as $\nabla_x f(Ax)$.[4] Keeping the notation clear is extremely important (as you'll find out in your homework, in fact!).

4.2 The Hessian

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a function that takes a vector in $\mathbb{R}^n$ and returns a real number. Then the ***Hessian*** matrix with respect to x, written $\nabla_x^2 f(x)$ or simply as H is the $n \times n$ matrix of partial derivatives,

$$\nabla_x^2 f(x) \in \mathbb{R}^{n \times n} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}.$$

In other words, $\nabla_x^2 f(x) \in \mathbb{R}^{n \times n}$, with

$$(\nabla_x^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}.$$

Note that the Hessian is always symmetric, since

$$\frac{\partial^2 f(x)}{\partial x_i \partial x_j} = \frac{\partial^2 f(x)}{\partial x_j \partial x_i}.$$

Similar to the gradient, the Hessian is defined only when $f(x)$ is real-valued.

It is natural to think of the gradient as the analogue of the first derivative for functions of vectors, and the Hessian as the analogue of the second derivative (and the symbols we use also suggest this relation). This intuition is generally correct, but there a few caveats to keep in mind.

[4] A drawback to this notation that we will have to live with is the fact that in the first case, $\nabla_z f(Ax)$ it appears that we are differentiating with respect to a variable that does not even appear in the expression being differentiated! For this reason, the first case is often written as $\nabla f(Ax)$, and the fact that we are differentiating with respect to the arguments of f is understood. However, the second case is *always* written as $\nabla_x f(Ax)$.

First, for real-valued functions of one variable $f : \mathbb{R} \to \mathbb{R}$, it is a basic definition that the second derivative is the derivative of the first derivative, i.e.,

$$\frac{\partial^2 f(x)}{\partial x^2} = \frac{\partial}{\partial x}\frac{\partial}{\partial x} f(x).$$

However, for functions of a vector, the gradient of the function is a vector, and we cannot take the gradient of a vector — i.e.,

$$\nabla_x \nabla_x f(x) = \nabla_x \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

and this expression is not defined. Therefore, it is *not* the case that the Hessian is the gradient of the gradient. However, this is *almost* true, in the following sense: If we look at the ith entry of the gradient $(\nabla_x f(x))_i = \partial f(x)/\partial x_i$, and take the gradient with respect to x we get

$$\nabla_x \frac{\partial f(x)}{\partial x_i} = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_i \partial x_1} \\ \frac{\partial^2 f(x)}{\partial x_i \partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_i \partial x_n} \end{bmatrix}$$

which is the ith column (or row) of the Hessian. Therefore,

$$\nabla_x^2 f(x) = \begin{bmatrix} \nabla_x(\nabla_x f(x))_1 & \nabla_x(\nabla_x f(x))_2 & \cdots & \nabla_x(\nabla_x f(x))_n \end{bmatrix}.$$

If we don't mind being a little bit sloppy we can say that (essentially) $\nabla_x^2 f(x) = \nabla_x(\nabla_x f(x))^T$, so long as we understand that this really means taking the gradient of each entry of $(\nabla_x f(x))^T$, not the gradient of the whole vector.

Finally, note that while we can take the gradient with respect to a matrix $A \in \mathbb{R}^n$, for the purposes of this class we will only consider taking the Hessian with respect to a vector $x \in \mathbb{R}^n$. This is simply a matter of convenience (and the fact that none of the calculations we do require us to find the Hessian with respect to a matrix), since the Hessian with respect to a matrix would have to represent all the partial derivatives $\partial^2 f(A)/(\partial A_{ij} \partial A_{k\ell})$, and it is rather cumbersome to represent this as a matrix.

4.3 Gradients and Hessians of Quadratic and Linear Functions

Now let's try to determine the gradient and Hessian matrices for a few simple functions. It should be noted that all the gradients given here are special cases of the gradients given in the CS229 lecture notes.

For $x \in \mathbb{R}^n$, let $f(x) = b^T x$ for some known vector $b \in \mathbb{R}^n$. Then

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

so

$$\frac{\partial f(x)}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} b_i x_i = b_k.$$

From this we can easily see that $\nabla_x b^T x = b$. This should be compared to the analogous situation in single variable calculus, where $\partial/(\partial x)\ ax = a$.

Now consider the quadratic function $f(x) = x^T A x$ for $A \in \mathbb{S}^n$. Remember that

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

To take the partial derivative, we'll consider the terms including x_k and x_k^2 factors separately:

$$\begin{aligned}
\frac{\partial f(x)}{\partial x_k} &= \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \sum_{j=1}^{n} A_{ij} x_i x_j \\
&= \frac{\partial}{\partial x_k} \left[\sum_{i \neq k} \sum_{j \neq k} A_{ij} x_i x_j + \sum_{i \neq k} A_{ik} x_i x_k + \sum_{j \neq k} A_{kj} x_k x_j + A_{kk} x_k^2 \right] \\
&= \sum_{i \neq k} A_{ik} x_i + \sum_{j \neq k} A_{kj} x_j + 2 A_{kk} x_k \\
&= \sum_{i=1}^{n} A_{ik} x_i + \sum_{j=1}^{n} A_{kj} x_j = 2 \sum_{i=1}^{n} A_{ki} x_i,
\end{aligned}$$

where the last equality follows since A is symmetric (which we can safely assume, since it is appearing in a quadratic form). Note that the kth entry of $\nabla_x f(x)$ is just the inner product of the kth row of A and x. Therefore, $\nabla_x x^T A x = 2Ax$. Again, this should remind you of the analogous fact in single-variable calculus, that $\partial/(\partial x)\ ax^2 = 2ax$.

Finally, let's look at the Hessian of the quadratic function $f(x) = x^T A x$ (it should be obvious that the Hessian of a linear function $b^T x$ is zero). In this case,

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_\ell} = \frac{\partial}{\partial x_k} \left[\frac{\partial f(x)}{\partial x_\ell} \right] = \frac{\partial}{\partial x_k} \left[2 \sum_{i=1}^{n} A_{\ell i} x_i \right] = 2 A_{\ell k} = 2 A_{k \ell}.$$

Therefore, it should be clear that $\nabla_x^2 x^T A x = 2A$, which should be entirely expected (and again analogous to the single-variable fact that $\partial^2/(\partial x^2)\ ax^2 = 2a$).

To recap,

- $\nabla_x b^T x = b$
- $\nabla_x x^T A x = 2Ax$ (if A symmetric)
- $\nabla_x^2 x^T A x = 2A$ (if A symmetric)

4.4 Least Squares

Let's apply the equations we obtained in the last section to derive the least squares equations. Suppose we are given matrices $A \in \mathbb{R}^{m\times n}$ (for simplicity we assume A is full rank) and a vector $b \in \mathbb{R}^m$ such that $b \notin \mathcal{R}(A)$. In this situation we will not be able to find a vector $x \in \mathbb{R}^n$, such that $Ax = b$, so instead we want to find a vector x such that Ax is as close as possible to b, as measured by the square of the Euclidean norm $\|Ax - b\|_2^2$.

Using the fact that $\|x\|_2^2 = x^T x$, we have

$$\begin{aligned} \|Ax - b\|_2^2 &= (Ax - b)^T (Ax - b) \\ &= x^T A^T Ax - 2b^T Ax + b^T b \end{aligned}$$

Taking the gradient with respect to x we have, and using the properties we derived in the previous section

$$\begin{aligned} \nabla_x (x^T A^T Ax - 2b^T Ax + b^T b) &= \nabla_x x^T A^T Ax - \nabla_x 2b^T Ax + \nabla_x b^T b \\ &= 2A^T Ax - 2A^T b \end{aligned}$$

Setting this last expression equal to zero and solving for x gives the normal equations

$$x = (A^T A)^{-1} A^T b$$

which is the same as what we derived in class.

4.5 Gradients of the Determinant

Now let's consider a situation where we find the gradient of a function with respect to a matrix, namely for $A \in \mathbb{R}^{n\times n}$, we want to find $\nabla_A |A|$. Recall from our discussion of determinants that

$$|A| = \sum_{i=1}^{n} (-1)^{i+j} A_{ij} |A_{\setminus i,\setminus j}| \qquad \text{(for any } j \in 1, \ldots, n)$$

so

$$\frac{\partial}{\partial A_{k\ell}} |A| = \frac{\partial}{\partial A_{k\ell}} \sum_{i=1}^{n} (-1)^{i+j} A_{ij} |A_{\setminus i,\setminus j}| = (-1)^{k+\ell} |A_{\setminus k,\setminus \ell}| = (\text{adj}(A))_{\ell k}.$$

From this it immediately follows from the properties of the adjoint that

$$\nabla_A |A| = (\text{adj}(A))^T = |A| A^{-T}.$$

Now let's consider the function $f : \mathbb{S}_{++}^n \to \mathbb{R}$, $f(A) = \log |A|$. Note that we have to restrict the domain of f to be the positive definite matrices, since this ensures that $|A| > 0$, so that the log of $|A|$ is a real number. In this case we can use the chain rule (nothing fancy, just the ordinary chain rule from single-variable calculus) to see that

$$\frac{\partial \log |A|}{\partial A_{ij}} = \frac{\partial \log |A|}{\partial |A|} \frac{\partial |A|}{\partial A_{ij}} = \frac{1}{|A|} \frac{\partial |A|}{\partial A_{ij}}.$$

From this it should be obvious that

$$\nabla_A \log |A| = \frac{1}{|A|} \nabla_A |A| = A^{-1},$$

where we can drop the transpose in the last expression because A is symmetric. Note the similarity to the single-valued case, where $\partial/(\partial x)\ \log x = 1/x$.

4.6 Eigenvalues as Optimization

Finally, we use matrix calculus to solve an optimization problem in a way that leads directly to eigenvalue/eigenvector analysis. Consider the following, equality constrained optimization problem:

$$\max_{x \in \mathbb{R}^n}\ x^T A x \quad \text{subject to } \|x\|_2^2 = 1$$

for a symmetric matrix $A \in \mathbb{S}^n$. A standard way of solving optimization problems with equality constraints is by forming the ***Lagrangian***, an objective function that includes the equality constraints.[5] The Lagrangian in this case can be given by

$$\mathcal{L}(x, \lambda) = x^T A x - \lambda x^T x$$

where λ is called the Lagrange multiplier associated with the equality constraint. It can be established that for x^* to be a optimal point to the problem, the gradient of the Lagrangian has to be zero at x^* (this is not the only condition, but it is required). That is,

$$\nabla_x \mathcal{L}(x, \lambda) = \nabla_x (x^T A x - \lambda x^T x) = 2A^T x - 2\lambda x = 0.$$

Notice that this is just the linear equation $Ax = \lambda x$. This shows that the only points which can possibly maximize (or minimize) $x^T A x$ assuming $x^T x = 1$ are the eigenvectors of A.

[5] Don't worry if you haven't seen Lagrangians before, as we will cover them in greater detail later in CS229.

Reference 2:

Introduction to Applied Linear Algebra by Stephen Boyd & Lieven Vandenberghe

https://web.stanford.edu/~boyd/vmls/

Vector spaces

a *vector space* or *linear space* (over the reals) consists of

- a set $\mathcal{V}$
- a vector sum $+ : \mathcal{V} \times \mathcal{V} \rightarrow \mathcal{V}$
- a scalar multiplication : $\mathbf{R} \times \mathcal{V} \rightarrow \mathcal{V}$
- a distinguished element $0 \in \mathcal{V}$

which satisfy a list of properties

- $x + y = y + x, \quad \forall x, y \in \mathcal{V}$ ($+$ is commutative)
- $(x + y) + z = x + (y + z), \quad \forall x, y, z \in \mathcal{V}$ ($+$ is associative)
- $0 + x = x$, $\forall x \in \mathcal{V}$ (0 is additive identity)
- $\forall x \in \mathcal{V} \;\; \exists (-x) \in \mathcal{V}$ s.t. $x + (-x) = 0$ (existence of additive inverse)
- $(\alpha\beta)x = \alpha(\beta x), \quad \forall \alpha, \beta \in \mathbf{R} \quad \forall x \in \mathcal{V}$ (scalar mult. is associative)
- $\alpha(x + y) = \alpha x + \alpha y, \quad \forall \alpha \in \mathbf{R} \;\; \forall x, y \in \mathcal{V}$ (right distributive rule)
- $(\alpha + \beta)x = \alpha x + \beta x, \quad \forall \alpha, \beta \in \mathbf{R} \quad \forall x \in \mathcal{V}$ (left distributive rule)
- $1x = x, \quad \forall x \in \mathcal{V}$

Examples

- $\mathcal{V}_1 = \mathbf{R}^n$, with standard (componentwise) vector addition and scalar multiplication
- $\mathcal{V}_2 = \{0\}$ (where $0 \in \mathbf{R}^n$)
- $\mathcal{V}_3 = \mathrm{span}(v_1, v_2, \ldots, v_k)$ where

$$\mathrm{span}(v_1, v_2, \ldots, v_k) = \{\alpha_1 v_1 + \cdots + \alpha_k v_k \mid \alpha_i \in \mathbf{R}\}$$

  and $v_1, \ldots, v_k \in \mathbf{R}^n$

Subspaces

- a *subspace* of a vector space is a *subset* of a vector space which is itself a vector space
- roughly speaking, a subspace is closed under vector addition and scalar multiplication
- examples $\mathcal{V}_1$, $\mathcal{V}_2$, $\mathcal{V}_3$ above are subspaces of $\mathbf{R}^n$

Vector spaces of functions

- $\mathcal{V}_4 = \{x : \mathbf{R}_+ \to \mathbf{R}^n \mid x \text{ is differentiable}\}$, where vector sum is sum of functions:

  $$(x + z)(t) = x(t) + z(t)$$

  and scalar multiplication is defined by

  $$(\alpha x)(t) = \alpha x(t)$$

  (a *point* in $\mathcal{V}_4$ is a *trajectory* in $\mathbf{R}^n$)

- $\mathcal{V}_5 = \{x \in \mathcal{V}_4 \mid \dot{x} = Ax\}$
  (*points* in $\mathcal{V}_5$ are *trajectories* of the linear system $\dot{x} = Ax$)

- $\mathcal{V}_5$ is a subspace of $\mathcal{V}_4$

Independent set of vectors

a set of vectors $\{v_1, v_2, \ldots, v_k\}$ is *independent* if

$$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_k v_k = 0 \Longrightarrow \alpha_1 = \alpha_2 = \cdots = 0$$

some equivalent conditions:

- coefficients of $\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_k v_k$ are uniquely determined, *i.e.*,

  $$\alpha_1 v_1 + \alpha_2 v_2 + \cdots + \alpha_k v_k = \beta_1 v_1 + \beta_2 v_2 + \cdots + \beta_k v_k$$

  implies $\alpha_1 = \beta_1, \alpha_2 = \beta_2, \ldots, \alpha_k = \beta_k$

- no vector v_i can be expressed as a linear combination of the other vectors $v_1, \ldots, v_{i-1}, v_{i+1}, \ldots, v_k$

Basis and dimension

set of vectors $\{v_1, v_2, \ldots, v_k\}$ is a *basis* for a vector space $\mathcal{V}$ if

- $v_1, v_2, \ldots, v_k$ span $\mathcal{V}$, *i.e.*, $\mathcal{V} = \mathrm{span}(v_1, v_2, \ldots, v_k)$
- $\{v_1, v_2, \ldots, v_k\}$ is independent

equivalent: every $v \in \mathcal{V}$ can be uniquely expressed as

$$v = \alpha_1 v_1 + \cdots + \alpha_k v_k$$

fact: for a given vector space $\mathcal{V}$, the number of vectors in any basis is the same

number of vectors in any basis is called the *dimension* of $\mathcal{V}$, denoted $\mathbf{dim}\mathcal{V}$

(we assign $\mathbf{dim}\{0\} = 0$, and $\mathbf{dim}\mathcal{V} = \infty$ if there is no basis)

Nullspace of a matrix

the *nullspace* of $A \in \mathbf{R}^{m \times n}$ is defined as

$$\mathcal{N}(A) = \{\ x \in \mathbf{R}^n \mid Ax = 0\ \}$$

- $\mathcal{N}(A)$ is set of vectors mapped to zero by $y = Ax$
- $\mathcal{N}(A)$ is set of vectors orthogonal to all rows of A

$\mathcal{N}(A)$ gives *ambiguity* in x given $y = Ax$:

- if $y = Ax$ and $z \in \mathcal{N}(A)$, then $y = A(x + z)$
- conversely, if $y = Ax$ and $y = A\tilde{x}$, then $\tilde{x} = x + z$ for some $z \in \mathcal{N}(A)$

Zero nullspace

A is called *one-to-one* if 0 is the only element of its nullspace:
$\mathcal{N}(A) = \{0\} \iff$

- x can always be uniquely determined from $y = Ax$
  ($i.e.$, the linear transformation $y = Ax$ doesn't 'lose' information)
- mapping from x to Ax is one-to-one: different x's map to different y's
- columns of A are independent (hence, a basis for their span)
- A has a *left inverse*, $i.e.$, there is a matrix $B \in \mathbf{R}^{n \times m}$ s.t. $BA = I$
- $\det(A^T A) \neq 0$

(we'll establish these later)

Interpretations of nullspace

suppose $z \in \mathcal{N}(A)$

$y = Ax$ represents **measurement** of x

- z is undetectable from sensors — get zero sensor readings
- x and $x + z$ are indistinguishable from sensors: $Ax = A(x + z)$

$\mathcal{N}(A)$ characterizes *ambiguity* in x from measurement $y = Ax$

$y = Ax$ represents **output** resulting from input x

- z is an input with no result
- x and $x + z$ have same result

$\mathcal{N}(A)$ characterizes *freedom of input choice* for given result

Range of a matrix

the *range* of $A \in \mathbf{R}^{m \times n}$ is defined as

$$\mathcal{R}(A) = \{Ax \mid x \in \mathbf{R}^n\} \subseteq \mathbf{R}^m$$

$\mathcal{R}(A)$ can be interpreted as

- the set of vectors that can be 'hit' by linear mapping $y = Ax$
- the span of columns of A
- the set of vectors y for which $Ax = y$ has a solution

Onto matrices

A is called *onto* if $\mathcal{R}(A) = \mathbf{R}^m \iff$

- $Ax = y$ can be solved in x for any y
- columns of A span $\mathbf{R}^m$
- A has a *right inverse*, $i.e.$, there is a matrix $B \in \mathbf{R}^{n\times m}$ s.t. $AB = I$
- rows of A are independent
- $\mathcal{N}(A^T) = \{0\}$
- $\det(AA^T) \neq 0$

(some of these are not obvious; we'll establish them later)

Interpretations of range

suppose $v \in \mathcal{R}(A)$, $w \notin \mathcal{R}(A)$

$y = Ax$ represents **measurement** of x

- $y = v$ is a *possible* or *consistent* sensor signal
- $y = w$ is *impossible* or *inconsistent*; sensors have failed or model is wrong

$y = Ax$ represents **output** resulting from input x

- v is a possible result or output
- w cannot be a result or output

$\mathcal{R}(A)$ characterizes the *possible results* or *achievable outputs*

Inverse

$A \in \mathbf{R}^{n \times n}$ is *invertible* or *nonsingular* if $\det A \neq 0$

equivalent conditions:

- columns of A are a basis for $\mathbf{R}^n$
- rows of A are a basis for $\mathbf{R}^n$
- $y = Ax$ has a unique solution x for every $y \in \mathbf{R}^n$
- A has a (left and right) inverse denoted $A^{-1} \in \mathbf{R}^{n \times n}$, with $AA^{-1} = A^{-1}A = I$
- $\mathcal{N}(A) = \{0\}$
- $\mathcal{R}(A) = \mathbf{R}^n$
- $\det A^T A = \det AA^T \neq 0$

Interpretations of inverse

suppose $A \in \mathbf{R}^{n \times n}$ has inverse $B = A^{-1}$

- mapping associated with B undoes mapping associated with A (applied either before or after!)
- $x = By$ is a perfect (pre- or post-) *equalizer* for the *channel* $y = Ax$
- $x = By$ is unique solution of $Ax = y$

Dual basis interpretation

- let a_i be columns of A, and $\tilde{b}_i^T$ be rows of $B = A^{-1}$
- from $y = x_1 a_1 + \cdots + x_n a_n$ and $x_i = \tilde{b}_i^T y$, we get

$$y = \sum_{i=1}^{n} (\tilde{b}_i^T y) a_i$$

  thus, inner product with *rows of inverse matrix* gives the coefficients in the *expansion of a vector in the columns of the matrix*

- $\tilde{b}_1, \ldots, \tilde{b}_n$ and $a_1, \ldots, a_n$ are called *dual bases*

Rank of a matrix

we define the *rank* of $A \in \mathbf{R}^{m \times n}$ as

$$\mathbf{rank}(A) = \mathbf{dim}\, \mathcal{R}(A)$$

(nontrivial) **facts:**

- $\mathbf{rank}(A) = \mathbf{rank}(A^T)$
- $\mathbf{rank}(A)$ is maximum number of independent columns (or rows) of A
  hence $\mathbf{rank}(A) \leq \mathbf{min}(m, n)$
- $\mathbf{rank}(A) + \mathbf{dim}\, \mathcal{N}(A) = n$

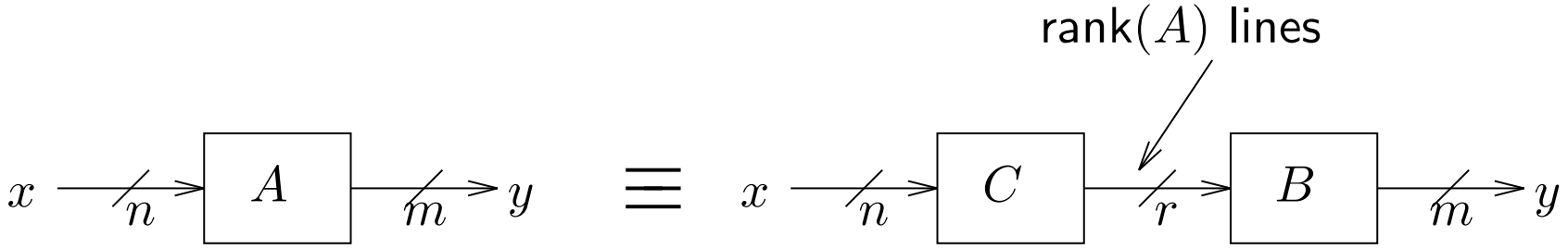
Conservation of dimension

interpretation of **rank**$(A) + $ **dim** $\mathcal{N}(A) = n$:

- **rank**(A) is dimension of set 'hit' by the mapping $y = Ax$
- **dim** $\mathcal{N}(A)$ is dimension of set of x 'crushed' to zero by $y = Ax$
- 'conservation of dimension': each dimension of input is either crushed to zero or ends up in output
- roughly speaking:
  - n is number of degrees of freedom in input x
  - **dim** $\mathcal{N}(A)$ is number of degrees of freedom lost in the mapping from x to $y = Ax$
  - **rank**(A) is number of degrees of freedom in output y

'Coding' interpretation of rank

- rank of product: $\mathbf{rank}(BC) \leq \mathbf{min}\{\mathbf{rank}(B),\ \mathbf{rank}(C)\}$
- hence if $A = BC$ with $B \in \mathbf{R}^{m\times r}$, $C \in \mathbf{R}^{r\times n}$, then $\mathbf{rank}(A) \leq r$
- conversely: if $\mathbf{rank}(A) = r$ then $A \in \mathbf{R}^{m\times n}$ can be factored as $A = BC$ with $B \in \mathbf{R}^{m\times r}$, $C \in \mathbf{R}^{r\times n}$:



- $\mathbf{rank}(A) = r$ is minimum size of vector needed to faithfully reconstruct y from x

Application: fast matrix-vector multiplication

- need to compute matrix-vector product $y = Ax$, $A \in \mathbf{R}^{m \times n}$
- A has known factorization $A = BC$, $B \in \mathbf{R}^{m \times r}$
- computing $y = Ax$ directly: mn operations
- computing $y = Ax$ as $y = B(Cx)$ (compute $z = Cx$ first, then $y = Bz$): $rn + mr = (m + n)r$ operations
- savings can be considerable if $r \ll \min\{m, n\}$

Full rank matrices

for $A \in \mathbf{R}^{m \times n}$ we always have $\mathbf{rank}(A) \leq \mathbf{min}(m, n)$

we say A is *full rank* if $\mathbf{rank}(A) = \mathbf{min}(m, n)$

- for **square** matrices, full rank means nonsingular
- for **skinny** matrices ($m \geq n$), full rank means columns are independent
- for **fat** matrices ($m \leq n$), full rank means rows are independent

Change of coordinates

'standard' basis vectors in $\mathbf{R}^n$: $(e_1, e_2, \ldots, e_n)$ where

$$e_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

(1 in ith component)

obviously we have

$$x = x_1 e_1 + x_2 e_2 + \cdots + x_n e_n$$

x_i are called the coordinates of x (in the standard basis)

if $(t_1, t_2, \ldots, t_n)$ is another basis for $\mathbf{R}^n$, we have

$$x = \tilde{x}_1 t_1 + \tilde{x}_2 t_2 + \cdots + \tilde{x}_n t_n$$

where $\tilde{x}_i$ are the coordinates of x in the basis $(t_1, t_2, \ldots, t_n)$

define $T = \begin{bmatrix} t_1 & t_2 & \cdots & t_n \end{bmatrix}$ so $x = T\tilde{x}$, hence

$$\tilde{x} = T^{-1} x$$

(T is invertible since t_i are a basis)

T^{-1} transforms (standard basis) coordinates of x into t_i-coordinates

inner product ith row of T^{-1} with x extracts t_i-coordinate of x

consider linear transformation $y = Ax$, $A \in \mathbf{R}^{n \times n}$

express y and x in terms of $t_1, t_2 \ldots, t_n$:

$$x = T\tilde{x}, \quad y = T\tilde{y}$$

so

$$\tilde{y} = (T^{-1}AT)\tilde{x}$$

- $A \longrightarrow T^{-1}AT$ is called *similarity transformation*
- similarity transformation by T expresses linear transformation $y = Ax$ in coordinates $t_1, t_2, \ldots, t_n$

(Euclidean) norm

for $x \in \mathbf{R}^n$ we define the (Euclidean) norm as

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{x^T x}$$

$\|x\|$ measures length of vector (from origin)

important properties:

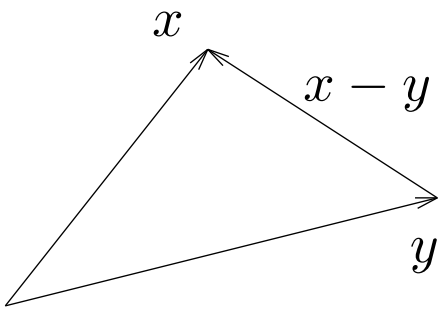
- $\|\alpha x\| = |\alpha|\|x\|$ (homogeneity)
- $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality)
- $\|x\| \geq 0$ (nonnegativity)
- $\|x\| = 0 \iff x = 0$ (definiteness)

RMS value and (Euclidean) distance

root-mean-square (RMS) value of vector $x \in \mathbf{R}^n$:

$$\mathbf{rms}(x) = \left(\frac{1}{n} \sum_{i=1}^{n} x_i^2 \right)^{1/2} = \frac{\|x\|}{\sqrt{n}}$$

norm defines distance between vectors: $\mathbf{dist}(x, y) = \|x - y\|$



Inner product

$$\langle x, y \rangle := x_1 y_1 + x_2 y_2 + \cdots + x_n y_n = x^T y$$

important properties:

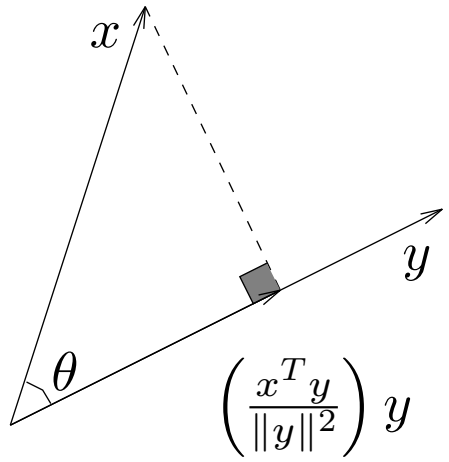
- $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$
- $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
- $\langle x, y \rangle = \langle y, x \rangle$
- $\langle x, x \rangle \geq 0$
- $\langle x, x \rangle = 0 \iff x = 0$

$f(y) = \langle x, y \rangle$ is linear function : $\mathbf{R}^n \to \mathbf{R}$, with linear map defined by row vector x^T

Cauchy-Schwartz inequality and angle between vectors

- for any $x, y \in \mathbf{R}^n$, $|x^T y| \leq \|x\|\|y\|$
- (unsigned) angle between vectors in $\mathbf{R}^n$ defined as

$$\theta = \angle(x, y) = \cos^{-1} \frac{x^T y}{\|x\|\|y\|}$$



thus $x^T y = \|x\|\|y\| \cos\theta$

special cases:

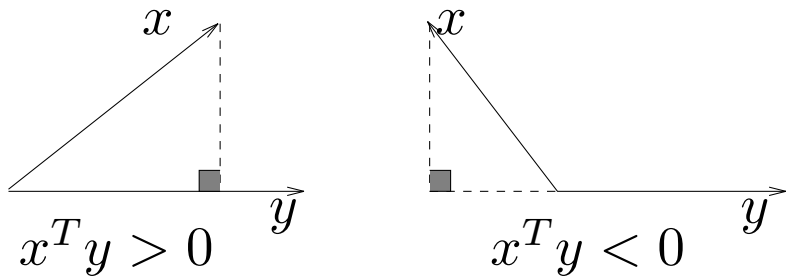
- x and y are *aligned*: $\theta = 0$; $x^T y = \|x\|\|y\|$;
  (if $x \neq 0$) $y = \alpha x$ for some $\alpha \geq 0$

- x and y are *opposed*: $\theta = \pi$; $x^T y = -\|x\|\|y\|$
  (if $x \neq 0$) $y = -\alpha x$ for some $\alpha \geq 0$

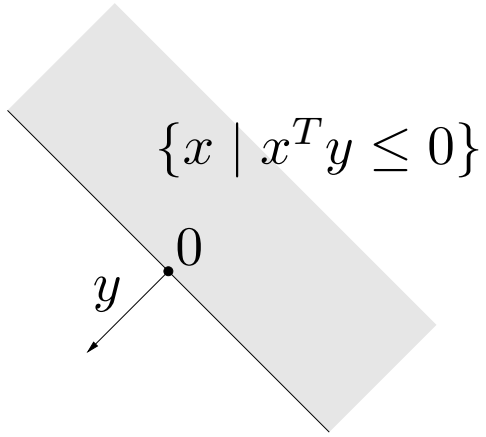
- x and y are *orthogonal*: $\theta = \pi/2$ or $-\pi/2$; $x^T y = 0$
  denoted $x \perp y$

interpretation of $x^T y > 0$ and $x^T y < 0$:

- $x^T y > 0$ means $\angle(x, y)$ is acute
- $x^T y < 0$ means $\angle(x, y)$ is obtuse



$\{x \mid x^T y \leq 0\}$ defines a *halfspace* with outward normal vector y, and boundary passing through 0



Eigenvalues of symmetric matrices

suppose $A \in \mathbf{R}^{n \times n}$ is symmetric, *i.e.*, $A = A^T$

fact: the eigenvalues of A are real

to see this, suppose $Av = \lambda v$, $v \neq 0$, $v \in \mathbf{C}^n$

then

$$\overline{v}^T A v = \overline{v}^T (Av) = \lambda \overline{v}^T v = \lambda \sum_{i=1}^{n} |v_i|^2$$

but also

$$\overline{v}^T A v = \overline{(Av)}^T v = \overline{(\lambda v)}^T v = \overline{\lambda} \sum_{i=1}^{n} |v_i|^2$$

so we have $\lambda = \overline{\lambda}$, *i.e.*, $\lambda \in \mathbf{R}$ (hence, can assume $v \in \mathbf{R}^n$)

Eigenvectors of symmetric matrices

fact: there is a set of orthonormal eigenvectors of A, *i.e.*, $q_1, \ldots, q_n$ s.t. $Aq_i = \lambda_i q_i$, $q_i^T q_j = \delta_{ij}$

in matrix form: there is an orthogonal Q s.t.

$$Q^{-1}AQ = Q^T AQ = \Lambda$$

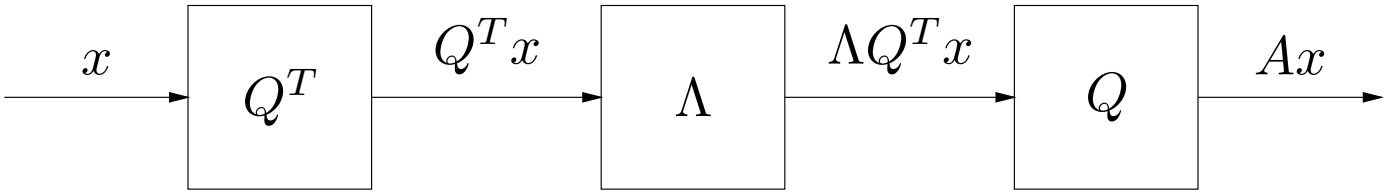
hence we can express A as

$$A = Q\Lambda Q^T = \sum_{i=1}^{n} \lambda_i q_i q_i^T$$

in particular, q_i are both left and right eigenvectors

Interpretations

$A = Q\Lambda Q^T$



linear mapping $y = Ax$ can be decomposed as

- resolve into q_i coordinates
- scale coordinates by λ_i
- reconstitute with basis q_i

or, geometrically,

- rotate by Q^T
- diagonal real scale ('dilation') by Λ
- rotate back by Q

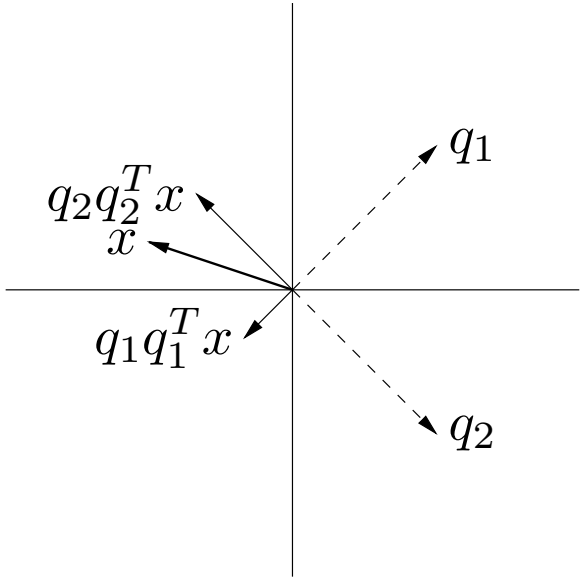
decomposition

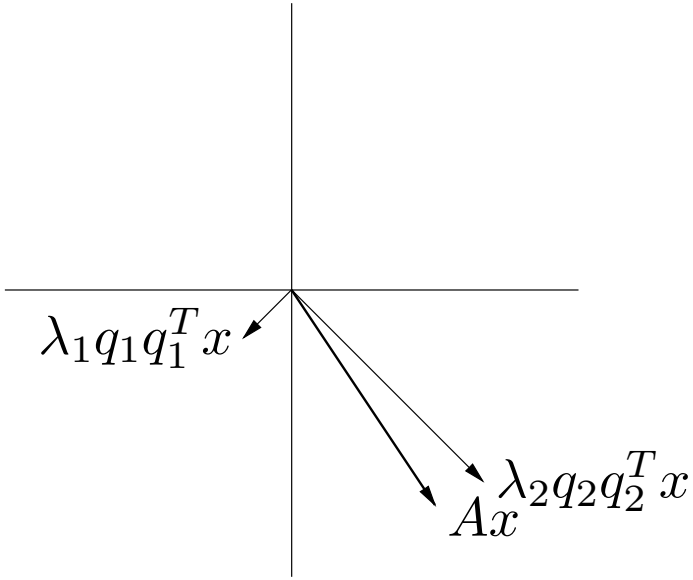
$$A = \sum_{i=1}^{n} \lambda_i q_i q_i^T$$

expresses A as linear combination of 1-dimensional projections

example:

$$
\begin{aligned}
A &= \begin{bmatrix} -1/2 & 3/2 \\ 3/2 & -1/2 \end{bmatrix} \\
&= \left(\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\right)\begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}\left(\frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\right)^T
\end{aligned}
$$





proof (case of λ_i distinct)

since λ_i distinct, can find $v_1, \ldots, v_n$, a set of linearly independent eigenvectors of A:

$$Av_i = \lambda_i v_i, \qquad \|v_i\| = 1$$

then we have

$$v_i^T(Av_j) = \lambda_j v_i^T v_j = (Av_i)^T v_j = \lambda_i v_i^T v_j$$

so $(\lambda_i - \lambda_j)v_i^T v_j = 0$

for $i \neq j$, $\lambda_i \neq \lambda_j$, hence $v_i^T v_j = 0$

- in this case we can say: eigenvectors *are* orthogonal
- in general case (λ_i not distinct) we must say: eigenvectors *can be chosen* to be orthogonal

Quadratic forms

a function $f : \mathbf{R}^n \to \mathbf{R}$ of the form

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

is called a *quadratic form*

in a quadratic form we may as well assume $A = A^T$ since

$$x^T A x = x^T ((A + A^T)/2) x$$

($(A + A^T)/2$ is called the *symmetric part* of A)

uniqueness: if $x^T A x = x^T B x$ for all $x \in \mathbf{R}^n$ and $A = A^T$, $B = B^T$, then $A = B$

Examples

- $\|Bx\|^2 = x^T B^T B x$
- $\sum_{i=1}^{n-1} (x_{i+1} - x_i)^2$
- $\|Fx\|^2 - \|Gx\|^2$

sets defined by quadratic forms:

- $\{\ x \mid f(x) = a\ \}$ is called a *quadratic surface*
- $\{\ x \mid f(x) \leq a\ \}$ is called a *quadratic region*

Inequalities for quadratic forms

suppose $A = A^T$, $A = Q\Lambda Q^T$ with eigenvalues sorted so $\lambda_1 \geq \cdots \geq \lambda_n$

$$\begin{aligned} x^T A x &= x^T Q \Lambda Q^T x \\ &= (Q^T x)^T \Lambda (Q^T x) \\ &= \sum_{i=1}^{n} \lambda_i (q_i^T x)^2 \\ &\leq \lambda_1 \sum_{i=1}^{n} (q_i^T x)^2 \\ &= \lambda_1 \|x\|^2 \end{aligned}$$

i.e., we have $x^T A x \leq \lambda_1 x^T x$

similar argument shows $x^T A x \geq \lambda_n \|x\|^2$, so we have

$$\lambda_n x^T x \leq x^T A x \leq \lambda_1 x^T x$$

sometimes λ_1 is called $\lambda_{\max}$, λ_n is called $\lambda_{\min}$

note also that

$$q_1^T A q_1 = \lambda_1 \|q_1\|^2, \qquad q_n^T A q_n = \lambda_n \|q_n\|^2,$$

so the inequalities are tight

Positive semidefinite and positive definite matrices

suppose $A = A^T \in \mathbf{R}^{n \times n}$

we say A is *positive semidefinite* if $x^T A x \geq 0$ for all x

- denoted $A \geq 0$ (and sometimes $A \succeq 0$)
- $A \geq 0$ if and only if $\lambda_{\min}(A) \geq 0$, *i.e.*, all eigenvalues are nonnegative
- **not** the same as $A_{ij} \geq 0$ for all i, j

we say A is *positive definite* if $x^T A x > 0$ for all $x \neq 0$

- denoted $A > 0$
- $A > 0$ if and only if $\lambda_{\min}(A) > 0$, *i.e.*, all eigenvalues are positive

Matrix inequalities

- we say A is *negative semidefinite* if $-A \geq 0$
- we say A is *negative definite* if $-A > 0$
- otherwise, we say A is *indefinite*

matrix inequality: if $B = B^T \in \mathbf{R}^n$ we say $A \geq B$ if $A - B \geq 0$, $A < B$ if $B - A > 0$, etc.

for example:

- $A \geq 0$ means A is positive semidefinite
- $A > B$ means $x^T A x > x^T B x$ for all $x \neq 0$

many properties that you'd guess hold actually do, $e.g.$,

- if $A \geq B$ and $C \geq D$, then $A + C \geq B + D$
- if $B \leq 0$ then $A + B \leq A$
- if $A \geq 0$ and $\alpha \geq 0$, then $\alpha A \geq 0$
- $A^2 \geq 0$
- if $A > 0$, then $A^{-1} > 0$

matrix inequality is only a *partial order*: we can have

$$A \not\geq B, \qquad B \not\geq A$$

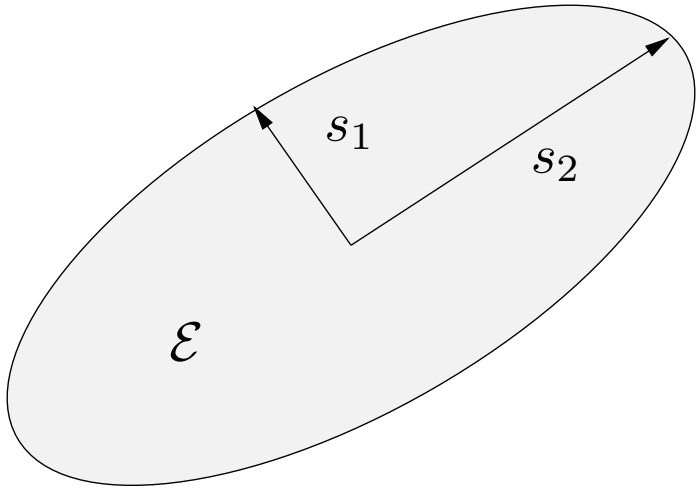
(such matrices are called *incomparable*)

Ellipsoids

if $A = A^T > 0$, the set

$$\mathcal{E} = \{\, x \mid x^T A x \leq 1 \,\}$$

is an *ellipsoid* in $\mathbf{R}^n$, centered at 0



semi-axes are given by $s_i = \lambda_i^{-1/2} q_i$, *i.e.*:

- eigenvectors determine directions of semiaxes
- eigenvalues determine lengths of semiaxes

note:

- in direction q_1, $x^T A x$ is *large*, hence ellipsoid is *thin* in direction q_1
- in direction q_n, $x^T A x$ is *small*, hence ellipsoid is *fat* in direction q_n
- $\sqrt{\lambda_{\max}/\lambda_{\min}}$ gives maximum *eccentricity*

if $\tilde{\mathcal{E}} = \{\, x \mid x^T B x \leq 1 \,\}$, where $B > 0$, then $\mathcal{E} \subseteq \tilde{\mathcal{E}} \iff A \geq B$

Gain of a matrix in a direction

suppose $A \in \mathbf{R}^{m \times n}$ (not necessarily square or symmetric)

for $x \in \mathbf{R}^n$, $\|Ax\|/\|x\|$ gives the *amplification factor* or *gain* of A in the direction x

obviously, gain varies with direction of input x

questions:

- what is maximum gain of A
  (and corresponding maximum gain direction)?
- what is minimum gain of A
  (and corresponding minimum gain direction)?
- how does gain of A vary with direction?

Matrix norm

the maximum gain

$$\max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

is called the *matrix norm* or *spectral norm* of A and is denoted $\|A\|$

$$\max_{x \neq 0} \frac{\|Ax\|^2}{\|x\|^2} = \max_{x \neq 0} \frac{x^T A^T A x}{\|x\|^2} = \lambda_{\max}(A^T A)$$

so we have $\|A\| = \sqrt{\lambda_{\max}(A^T A)}$

similarly the minimum gain is given by

$$\min_{x \neq 0} \|Ax\|/\|x\| = \sqrt{\lambda_{\min}(A^T A)}$$

note that

- $A^TA \in \mathbf{R}^{n\times n}$ is symmetric and $A^TA \geq 0$ so $\lambda_{\min},\ \lambda_{\max} \geq 0$
- 'max gain' input direction is $x = q_1$, eigenvector of A^TA associated with $\lambda_{\max}$
- 'min gain' input direction is $x = q_n$, eigenvector of A^TA associated with $\lambda_{\min}$

example: $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}$

$$A^TA = \begin{bmatrix} 35 & 44 \\ 44 & 56 \end{bmatrix}$$

$$= \begin{bmatrix} 0.620 & 0.785 \\ 0.785 & -0.620 \end{bmatrix} \begin{bmatrix} 90.7 & 0 \\ 0 & 0.265 \end{bmatrix} \begin{bmatrix} 0.620 & 0.785 \\ 0.785 & -0.620 \end{bmatrix}^T$$

then $\|A\| = \sqrt{\lambda_{\max}(A^TA)} = 9.53$:

$$\left\| \begin{bmatrix} 0.620 \\ 0.785 \end{bmatrix} \right\| = 1, \qquad \left\| A \begin{bmatrix} 0.620 \\ 0.785 \end{bmatrix} \right\| = \left\| \begin{bmatrix} 2.18 \\ 4.99 \\ 7.78 \end{bmatrix} \right\| = 9.53$$

min gain is $\sqrt{\lambda_{\min}(A^TA)} = 0.514$:

$$\left\| \begin{bmatrix} 0.785 \\ -0.620 \end{bmatrix} \right\| = 1, \qquad \left\| A \begin{bmatrix} 0.785 \\ -0.620 \end{bmatrix} \right\| = \left\| \begin{bmatrix} 0.46 \\ 0.14 \\ -0.18 \end{bmatrix} \right\| = 0.514$$

for all $x \neq 0$, we have

$$0.514 \leq \frac{\|Ax\|}{\|x\|} \leq 9.53$$

Properties of matrix norm

- consistent with vector norm: matrix norm of $a \in \mathbf{R}^{n\times 1}$ is $\sqrt{\lambda_{\max}(a^T a)} = \sqrt{a^T a}$
- for any x, $\|Ax\| \leq \|A\|\|x\|$
- scaling: $\|aA\| = |a|\|A\|$
- triangle inequality: $\|A + B\| \leq \|A\| + \|B\|$
- definiteness: $\|A\| = 0 \quad \Leftrightarrow \quad A = 0$
- norm of product: $\|AB\| \leq \|A\|\|B\|$

Singular value decomposition

more complete picture of gain properties of A given by *singular value decomposition* (SVD) of A:

$$A = U\Sigma V^T$$

where

- $A \in \mathbf{R}^{m\times n}$, $\mathbf{Rank}(A) = r$
- $U \in \mathbf{R}^{m\times r}$, $U^TU = I$
- $V \in \mathbf{R}^{n\times r}$, $V^TV = I$
- $\Sigma = \mathbf{diag}(\sigma_1, \ldots, \sigma_r)$, where $\sigma_1 \geq \cdots \geq \sigma_r > 0$

with $U = [u_1 \cdots u_r]$, $V = [v_1 \cdots v_r]$,

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

- σ_i are the (nonzero) *singular values* of A
- v_i are the *right* or *input singular vectors* of A
- u_i are the *left* or *output singular vectors* of A

$$A^T A = (U\Sigma V^T)^T (U\Sigma V^T) = V\Sigma^2 V^T$$

hence:

- v_i are eigenvectors of $A^T A$ (corresponding to nonzero eigenvalues)
- $\sigma_i = \sqrt{\lambda_i(A^T A)}$ (and $\lambda_i(A^T A) = 0$ for $i > r$)
- $\|A\| = \sigma_1$

similarly,

$$AA^T = (U\Sigma V^T)(U\Sigma V^T)^T = U\Sigma^2 U^T$$

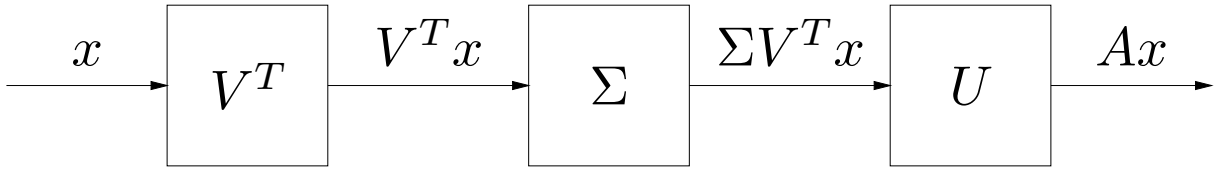
hence:

- u_i are eigenvectors of AA^T (corresponding to nonzero eigenvalues)
- $\sigma_i = \sqrt{\lambda_i(AA^T)}$ (and $\lambda_i(AA^T) = 0$ for $i > r$)

- $u_1, \ldots u_r$ are orthonormal basis for $\mathrm{range}(A)$
- $v_1, \ldots v_r$ are orthonormal basis for $\mathcal{N}(A)^\perp$

Interpretations

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



linear mapping $y = Ax$ can be decomposed as

- compute coefficients of x along input directions $v_1, \ldots, v_r$
- scale coefficients by σ_i
- reconstitute along output directions $u_1, \ldots, u_r$

difference with eigenvalue decomposition for symmetric A: input and output directions are *different*

- v_1 is most sensitive (highest gain) input direction
- u_1 is highest gain output direction
- $Av_1 = \sigma_1 u_1$

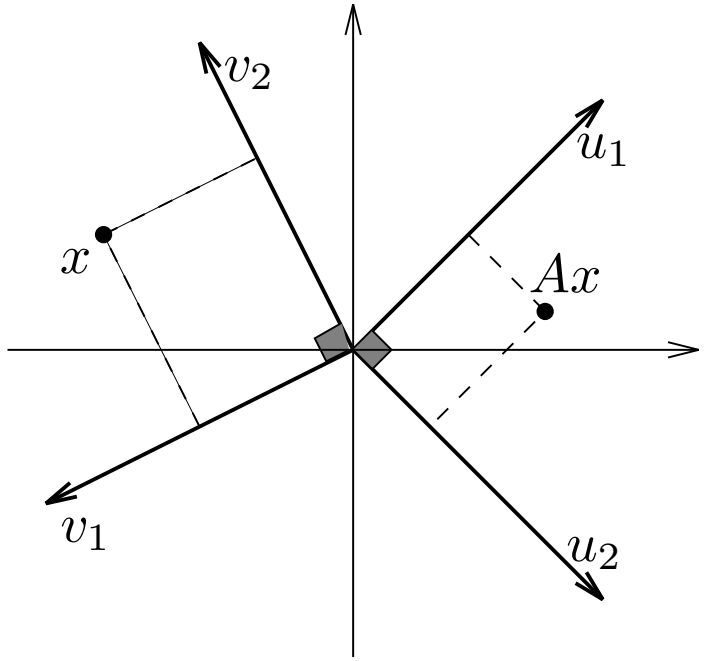
SVD gives clearer picture of gain as function of input/output directions

example: consider $A \in \mathbf{R}^{4\times 4}$ with $\Sigma = \mathbf{diag}(10,\ 7,\ 0.1,\ 0.05)$

- input components along directions v_1 and v_2 are amplified (by about 10) and come out mostly along plane spanned by u_1, u_2
- input components along directions v_3 and v_4 are attenuated (by about 10)
- $\|Ax\|/\|x\|$ can range between 10 and 0.05
- A is nonsingular
- for some applications you might say A is *effectively* rank 2

example: $A \in \mathbf{R}^{2\times 2}$, with $\sigma_1 = 1$, $\sigma_2 = 0.5$

- resolve x along $v_1,\ v_2$: $v_1^T x = 0.5$, $v_2^T x = 0.6$, *i.e.*, $x = 0.5v_1 + 0.6v_2$
- now form $Ax = (v_1^T x)\sigma_1 u_1 + (v_2^T x)\sigma_2 u_2 = (0.5)(1)u_1 + (0.6)(0.5)u_2$



General pseudo-inverse

if $A \neq 0$ has SVD $A = U\Sigma V^T$,

$$A^\dagger = V\Sigma^{-1}U^T$$

is the *pseudo-inverse* or *Moore-Penrose inverse* of A

if A is skinny and full rank,

$$A^\dagger = (A^TA)^{-1}A^T$$

gives the least-squares approximate solution $x_{\rm ls} = A^\dagger y$

if A is fat and full rank,

$$A^\dagger = A^T(AA^T)^{-1}$$

gives the least-norm solution $x_{\rm ln} = A^\dagger y$

in general case:

$$X_{\rm ls} = \{\ z \mid \|Az - y\| = \min_w\ \|Aw - y\|\ \}$$

is set of least-squares approximate solutions

$x_{\rm pinv} = A^\dagger y \in X_{\rm ls}$ has minimum norm on $X_{\rm ls}$, *i.e.*, $x_{\rm pinv}$ is the minimum-norm, least-squares approximate solution

Pseudo-inverse via regularization

for $\mu > 0$, let x_μ be (unique) minimizer of

$$\|Ax - y\|^2 + \mu\|x\|^2$$

i.e.,

$$x_\mu = \left(A^T A + \mu I\right)^{-1} A^T y$$

here, $A^T A + \mu I > 0$ and so is invertible

then we have $\lim_{\mu \to 0} x_\mu = A^\dagger y$

in fact, we have $\lim_{\mu \to 0} \left(A^T A + \mu I\right)^{-1} A^T = A^\dagger$

(check this!)

Full SVD

SVD of $A \in \mathbf{R}^{m\times n}$ with $\mathbf{Rank}(A) = r$:

$$A = U_1\Sigma_1 V_1^T = \begin{bmatrix} u_1 & \cdots & u_r \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_r^T \end{bmatrix}$$

- find $U_2 \in \mathbf{R}^{m\times(m-r)}$, $V_2 \in \mathbf{R}^{n\times(n-r)}$ s.t. $U = [U_1\ U_2] \in \mathbf{R}^{m\times m}$ and $V = [V_1\ V_2] \in \mathbf{R}^{n\times n}$ are orthogonal

- add zero rows/cols to Σ_1 to form $\Sigma \in \mathbf{R}^{m\times n}$:

$$\Sigma = \left[\begin{array}{c|c} \Sigma_1 & 0_{r\times(n-r)} \\ \hline 0_{(m-r)\times r} & 0_{(m-r)\times(n-r)} \end{array}\right]$$

then we have

$$A = U_1 \Sigma_1 V_1^T = \left[\begin{array}{c|c} U_1 & U_2 \end{array}\right] \left[\begin{array}{c|c} \Sigma_1 & 0_{r \times (n-r)} \\ \hline 0_{(m-r) \times r} & 0_{(m-r) \times (n-r)} \end{array}\right] \left[\begin{array}{c} V_1^T \\ \hline V_2^T \end{array}\right]$$

i.e.:

$$A = U \Sigma V^T$$

called *full SVD* of A

(SVD with positive singular values only called *compact SVD*)

Image of unit ball under linear transformation

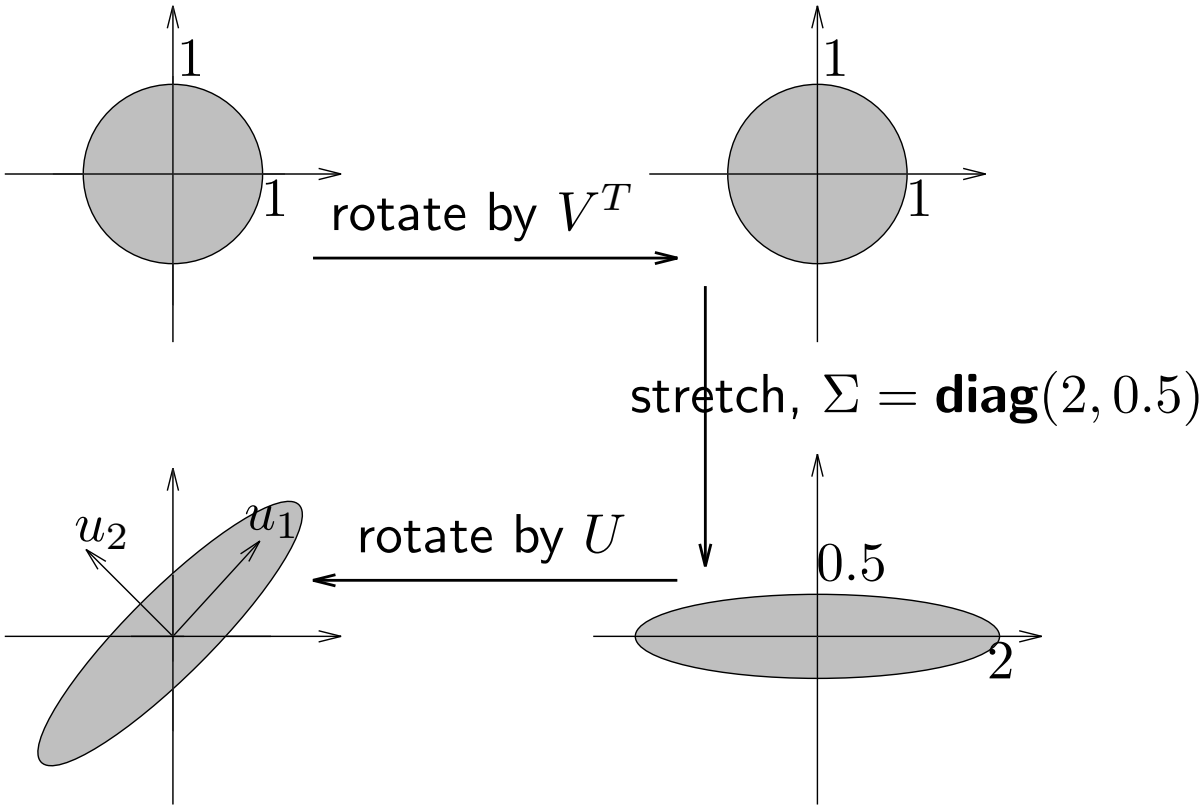
full SVD:

$$A = U\Sigma V^T$$

gives intepretation of $y = Ax$:

- rotate (by V^T)
- stretch along axes by σ_i ($\sigma_i = 0$ for $i > r$)
- zero-pad (if $m > n$) or truncate (if $m < n$) to get m-vector
- rotate (by U)

Image of unit ball under A



$\{Ax \mid \|x\| \leq 1\}$ is *ellipsoid* with principal axes $\sigma_i u_i$.

SVD in estimation/inversion

suppose $y = Ax + v$, where

- $y \in \mathbf{R}^m$ is measurement
- $x \in \mathbf{R}^n$ is vector to be estimated
- v is a measurement noise or error

'norm-bound' model of noise: we assume $\|v\| \leq \alpha$ but otherwise know nothing about v (α gives max norm of noise)

- consider estimator $\hat{x} = By$, with $BA = I$ ($i.e.$, unbiased)
- estimation or inversion error is $\tilde{x} = \hat{x} - x = Bv$
- set of possible estimation errors is ellipsoid

$$\tilde{x} \in \mathcal{E}_{\rm unc} = \{\ Bv \mid \|v\| \leq \alpha\ \}$$

- $x = \hat{x} - \tilde{x} \in \hat{x} - \mathcal{E}_{\rm unc} = \hat{x} + \mathcal{E}_{\rm unc}$, $i.e.$:
  true x lies in *uncertainty ellipsoid* $\mathcal{E}_{\rm unc}$, centered at estimate $\hat{x}$
- 'good' estimator has 'small' $\mathcal{E}_{\rm unc}$ (with $BA = I$, of course)

semiaxes of $\mathcal{E}_{\rm unc}$ are $\alpha\sigma_i u_i$ (singular values & vectors of B)

e.g., maximum norm of error is $\alpha\|B\|$, *i.e.*, $\|\hat{x} - x\| \leq \alpha\|B\|$

optimality of least-squares: suppose $BA = I$ is any estimator, and $B_{\rm ls} = A^\dagger$ is the least-squares estimator

then:

- $B_{\rm ls}B_{\rm ls}^T \leq BB^T$
- $\mathcal{E}_{\rm ls} \subseteq \mathcal{E}$
- in particular $\|B_{\rm ls}\| \leq \|B\|$

i.e., the least-squares estimator gives the *smallest* uncertainty ellipsoid

Sensitivity of linear equations to data error

consider $y = Ax$, $A \in \mathbf{R}^{n \times n}$ invertible; of course $x = A^{-1}y$

suppose we have an error or noise in y, *i.e.*, y becomes $y + \delta y$

then x becomes $x + \delta x$ with $\delta x = A^{-1}\delta y$

hence we have $\|\delta x\| = \|A^{-1}\delta y\| \leq \|A^{-1}\|\|\delta y\|$

if $\|A^{-1}\|$ is large,

- small errors in y can lead to large errors in x
- can't solve for x given y (with small errors)
- hence, A can be considered singular in practice

a more refined analysis uses *relative* instead of *absolute* errors in x and y

since $y = Ax$, we also have $\|y\| \leq \|A\|\|x\|$, hence

$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta y\|}{\|y\|}$$

$$\kappa(A) = \|A\|\|A^{-1}\| = \sigma_{\max}(A)/\sigma_{\min}(A)$$

is called the *condition number* of A

we have:

relative error in solution $x \leq$ condition number $\cdot$ relative error in data y

or, in terms of # bits of guaranteed accuracy:

$$\#\text{ bits accuracy in solution} \approx \#\text{ bits accuracy in data} - \log_2 \kappa$$

we say

- A is well conditioned if κ is small
- A is poorly conditioned if κ is large

(definition of 'small' and 'large' depend on application)

same analysis holds for least-squares approximate solutions with A nonsquare, $\kappa = \sigma_{\max}(A)/\sigma_{\min}(A)$

Low rank approximations

suppose $A \in \mathbf{R}^{m\times n}$, $\mathbf{Rank}(A) = r$, with SVD $A = U\Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T$

we seek matrix $\hat{A}$, $\mathbf{Rank}(\hat{A}) \leq p < r$, s.t. $\hat{A} \approx A$ in the sense that $\|A - \hat{A}\|$ is minimized

solution: optimal rank p approximator is

$$\hat{A} = \sum_{i=1}^{p} \sigma_i u_i v_i^T$$

- hence $\|A - \hat{A}\| = \left\|\sum_{i=p+1}^{r} \sigma_i u_i v_i^T\right\| = \sigma_{p+1}$
- interpretation: SVD dyads $u_i v_i^T$ are ranked in order of 'importance'; take p to get rank p approximant

proof: suppose $\mathbf{Rank}(B) \leq p$

then $\dim \mathcal{N}(B) \geq n - p$

also, $\dim \operatorname{span}\{v_1, \ldots, v_{p+1}\} = p + 1$

hence, the two subspaces intersect, *i.e.*, there is a unit vector $z \in \mathbf{R}^n$ s.t.

$$Bz = 0, \qquad z \in \operatorname{span}\{v_1, \ldots, v_{p+1}\}$$

$$(A - B)z = Az = \sum_{i=1}^{p+1} \sigma_i u_i v_i^T z$$

$$\|(A - B)z\|^2 = \sum_{i=1}^{p+1} \sigma_i^2 (v_i^T z)^2 \geq \sigma_{p+1}^2 \|z\|^2$$

hence $\|A - B\| \geq \sigma_{p+1} = \|A - \hat{A}\|$

Distance to singularity

another interpretation of σ_i:

$$\sigma_i = \min\{\ \|A - B\| \mid \mathbf{Rank}(B) \leq i - 1\ \}$$

$i.e.$, the distance (measured by matrix norm) to the nearest rank $i - 1$ matrix

for example, if $A \in \mathbf{R}^{n \times n}$, $\sigma_n = \sigma_{\min}$ is distance to nearest singular matrix

hence, small $\sigma_{\min}$ means A is near to a singular matrix

application: model simplification

suppose $y = Ax + v$, where

- $A \in \mathbf{R}^{100\times 30}$ has SVs

$$10,\ 7,\ 2,\ 0.5,\ 0.01,\ \ \ldots, 0.0001$$

- $\|x\|$ is on the order of 1
- unknown error or noise v has norm on the order of 0.1

then the terms $\sigma_i u_i v_i^T x$, for $i = 5, \ldots, 30$, are substantially smaller than the noise term v

simplified model:

$$y = \sum_{i=1}^{4} \sigma_i u_i v_i^T x + v$$

EE263 Prof. S. Boyd

Crimes Against Matrices

In this note we list some matrix crimes that we have, sadly, witnessed too often. Be very careful to avoid committing any of these crimes; in EE263 we have a *zero-tolerance* policy for *crimes against matrices*, at least on things you hand in to us. (What you do with matrices in your spare time, or on scratch paper, is of course your own business. But we recommend you avoid these crimes at all times, in order to not build bad habits.)

Check your work — don't become just another sad statistic!

Syntax crimes

In a syntax crime, the perpetrator attempts to combine matrices (or other mathematical objects) in ways that violate basic syntax rules. These are serious crimes of negligence, since it is so easy to check your work for potential violations. We list some typical examples below.

- Adding, subtracting, or equating, matrices (or vectors) of different dimensions.

  Example: writing $A + B$, when $A \in \mathbf{R}^{2\times 3}$ and $B \in \mathbf{R}^{3\times 3}$.

- Violating the rules of constructing block matrices (*e.g.*, the submatrices in any row of a block matrix must have the same number of rows).

  Example: forming the block matrix $[A\ B]$, when $A \in \mathbf{R}^{2\times 3}$ and $B \in \mathbf{R}^{3\times 3}$.

- Multiplying matrices with incompatible dimensions (*i.e.*, forming AB, when the number of columns of A does not equal the number of rows of B).

  Example: forming $A^T B$, when $A \in \mathbf{R}^{2\times 3}$ and $B \in \mathbf{R}^{3\times 3}$.

- Taking the inverse, determinant, trace, or powers of a nonsquare matrix.

  Example: forming A^{-1}, when $A \in \mathbf{R}^{2\times 3}$.

Semantic crimes

In a semantic crime, the perpetrator forms an expression or makes an assertion that does not break any syntax rules, but is wrong because of the meaning. These crimes are a bit harder to detect than syntax crimes, so you need to be more vigilant to avoid committing them.

- Taking the inverse of a square, but singular matrix. (Taking the inverse of a nonsquare matrix is a syntax crime—see above.)

  Example: forming $(ww^T)^{-1}$, where $w \in \mathbf{R}^2$.

  Note: writing $(ww^T)^{-1} = (w^T)^{-1}w^{-1}$, when $w \in \mathbf{R}^2$, involves both a syntax and semantic crime.

- Referring to a left inverse of a strictly fat matrix or a right inverse of a strictly skinny matrix.

  Example: writing $QQ^T = I$, when $Q \in \mathbf{R}^{5\times 3}$.

- Cancelling matrices on the left or right in inappropriate circumstances, *e.g.*, concluding that $B = C$ from $AB = AC$, when A is not known to be one-to-one (*i.e.*, have independent columns).

  Example: concluding $x = y$ from $a^Tx = a^Ty$, when $a,\ x,\ y \in \mathbf{R}^4$.

- *Dimension crimes.* Alleging that a set of m vectors in $\mathbf{R}^n$ is independent, when $m > n$. Alleging that a set of m vectors in $\mathbf{R}^n$ span $\mathbf{R}^n$, when $m < n$.

Miscellaneous crimes

Some crimes are hard to classify, or involve both syntax and semantic elements. Incorrect use of a matrix identity often falls in this class.

- Using $(AB)^T = A^TB^T$ (instead of the correct formula $(AB)^T = B^TA^T$).

  Note: this also violates syntax rules, if A^TB^T is not a valid product.

- Using $(AB)^{-1} = A^{-1}B^{-1}$ (instead of the correct formula $(AB)^{-1} = B^{-1}A^{-1}$).

  Note: $(AB)^{-1} = A^{-1}B^{-1}$ violates syntax rules, if A or B is not square; it violates semantic rules if A or B is not invertible.

- Using $(A+B)^2 = A^2 + 2AB + B^2$. This (false) identity relies on the very useful, but unfortunately false, identity $AB = BA$.

An example

Let's consider the expression $(A^TB)^{-1}$, where $A \in \mathbf{R}^{m\times n}$ and $B \in \mathbf{R}^{k\times p}$. Here's how you might check for various crimes you might commit in forming this expression.

- We multiply A^T, which is $n \times m$, and B, which is $k \times p$, so we better have $m = k$ to avoid a syntax violation.

  Note: if A is a scalar, then $A^T B$ might be a strange thing to write, but can be argued to not violate syntax, even though $m \neq k$. In a similar way, when B is scalar, you can write $A^T B$, and argue that syntax is not violated.

- The product $A^T B$ is $n \times p$, so we better have $n = p$ in order to (attempt to) invert it. At this point, we know that the dimensions of A and B must be the same (ignoring the case where one or the other is interpreted as a scalar).

- If $A^T B$ is a strictly skinny–strictly fat product (*i.e.*, A and B are strictly fat), then $A^T B$ cannot possibly be invertible, so we have a semantic violation. To avoid this, we must have A and B square or skinny, *i.e.*, $m \geq n$.

Summary: to write $(A^T B)^{-1}$ (assuming neither A nor B is interpreted as a scalar), A and B must have the same dimensions, and be skinny or square.

Of course, even if A and B have the same dimensions, and are skinny or square, the matrix $A^T B$ can be singular, in which case $(A^T B)^{-1}$ is meaningless. The point of our analysis above is that if A and B don't have the same dimension, or if A and B are strictly fat, then $(A^T B)^{-1}$ is *guaranteed* to be meaningless, no matter what values A and B might have in your application or argument.