Linear algebra review

James Chuang

December 15, 2016

Contents

2.1 vector-vector products . . . 1
2.2 matrix-vector products . . . 2
2.3 matrix-matrix products . . . 4
3.2 the transpose . . . 5
3.3 symmetric matrices . . . 5
3.4 the trace . . . 6
3.5 norms . . . 6
3.6 linear independence and rank . . . 7
3.7 the inverse . . . 7
3.8 orthogonal matrices . . . 8
3.9 range and nullspace of a matrix . . . 8
3.10 the determinant . . . 9
3.11 quadratic forms and positive semidefinite matrices . . . 10
3.12 eigenvalues and eigenvectors . . . 11
3.13 eigenvalues and eigenvectors of symmetric matrices . . . 12
4.1 the gradient . . . 13
4.2 the Hessian . . . 14
4.3 gradients and hessians of linear and quadratic functions . . . 15
4.5 gradients of the determinant . . . 16
4.6 eigenvalues as optimization . . . 16

This is a post of my notes reviewing linear algebra, aggregating information mainly from the linear algebra review and reference section notes from the Stanford CS229 Machine Learning course and Chapter 2 of the Deep Learning Book. In many (most) places, the material is directly transcribed from these sources. Rewriting it in LaTEX just helps me to slow down and better understand the content. The notes are not comprehensive, they contain the things that weren't immediately obvious to me (though this turns out to be most of the material).

2.1 vector-vector products

inner (dot) product

Given $x, y \in \mathbb{R}^n$, $x^T y \in \mathbb{R}$ is the ***inner product***, aka ***dot product***.

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

Therefore, $x^T y = y^T x$ is always true.

outer product

Given $x \in \mathbb{R}^m, y \in \mathbb{R}^n$, $xy^T \in \mathbb{R}^{m\times n}$ is the ***outer product***.

$$xy^T \in \mathbb{R}^{m\times n} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1y_1 & x_1y_2 & \cdots & x_1y_n \\ x_2y_1 & x_2y_2 & \cdots & x_2y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_my_1 & x_my_2 & \cdots & x_my_n \end{bmatrix}$$

The CS229 notes give an example of how the outer product with a vector of ones $\mathbf{1} \in \mathbb{R}^n$ can be used to give a compact representation of a matrix $A \in \mathbb{R}^{m\times n}$ whose columns are all equal to a vector $x \in \mathbb{R}^m$:

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

The Deep Learning Book section 2.1 describes the use of an unconventional notation called *broadcasting* where the addition of a matrix and a vector to yield another matrix is allowed: $C = A + b$, where $C_{i,j} = A_{i,j} + b_j$. $(C \in \mathbb{R}^{m\times n}, A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^n)$ Explicitly writing this out:

$$\begin{aligned} C = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1} & A_{m,2} & \cdots & A_{m,n} \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} &= \begin{bmatrix} A_{1,1}+b_1 & A_{1,2}+b_2 & \cdots & A_{1,n}+b_n \\ A_{2,1}+b_1 & A_{2,2}+b_2 & \cdots & A_{2,n}+b_n \\ \vdots & \vdots & \ddots & \vdots \\ A_{m,1}+b_1 & A_{m,2}+b_2 & \cdots & A_{m,n}+b_n \end{bmatrix} \\ &= A + \begin{bmatrix} b_1 & b_2 & \cdots & b_n \\ b_1 & b_2 & \cdots & b_n \\ \vdots & \vdots & \ddots & \vdots \\ b_1 & b_2 & \cdots & b_n \end{bmatrix} \\ &= A + (b\mathbf{1}^T)^T \\ &= A + 1b^T \end{aligned}$$

So, the shorthand $C = A + b$ can be written more explicitly (but still pretty compactly) as $C = A + \mathbf{1}b^T$, where $\mathbf{1} \in \mathbb{R}^m$

2.2 matrix-vector products

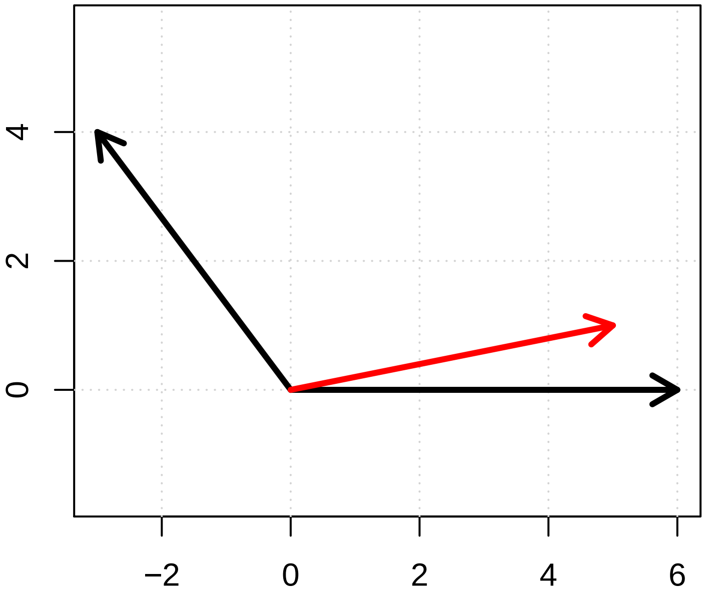
Given a matrix $A \in \mathbb{R}^{m\times n}$ and a vector $x \in \mathbb{R}^n$, their product is a vector $y = Ax \in \mathbb{R}^m$. The CS229 notes go through some different representations of the product:

representing A as rows

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

That is, the ith entry of y is the inner product of the ith row of A and x, $y_i = a_i^T x$.

Recalling that the inner product is a similarity measure, y can be interpreted as a list of how similar each row of A is to x. This is illustrated below, with the rows of the matrix $A = \begin{bmatrix} 6 & 0 \\ -3 & 4 \end{bmatrix}$ in black and the vector $x = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$ in red.



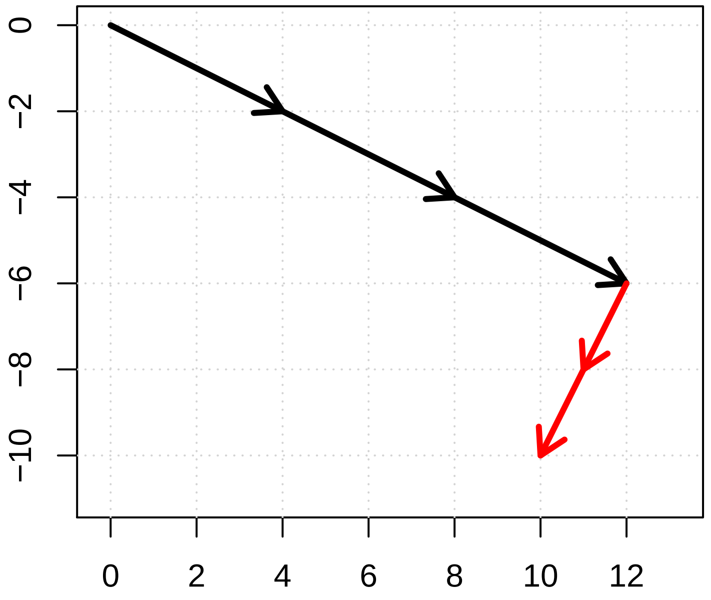
Here $y = Ax = \begin{bmatrix} 30 \\ -11 \end{bmatrix}$, reflecting the fact that $x = \begin{bmatrix} 5 \\ 1 \end{bmatrix}$ is more similar to $a_1 = \begin{bmatrix} 6 \\ 0 \end{bmatrix}$ than it is to $a_2 = \begin{bmatrix} -3 \\ 4 \end{bmatrix}$

representing A as columns

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

That is, y is a ***linear combination*** of the columns of A, where the coefficients of the linear combination are the entries of x.

This is illustrated below, with the matrix $A = \begin{bmatrix} 4 & 1 \\ -2 & 2 \end{bmatrix}$ and $x = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$.



Here $y = Ax = \begin{bmatrix} 10 \\ -10 \end{bmatrix} = \begin{bmatrix} 4 \\ -2 \end{bmatrix}(3) + \begin{bmatrix} 1 \\ 2 \end{bmatrix}(-2)$, representing the point in $\mathbb{R}^m$ reached after taking $x_1 = 3$ "steps" of $a_1 = \begin{bmatrix} 4 \\ -2 \end{bmatrix}$ drawn as black vectors plus $x_2 = -2$ "steps" of $a_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ drawn as red vectors.

Analogous cases occur in the left multiplication of a matrix by a row vector, $y^T = x^T A$ for $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^m$, and $y \in \mathbb{R}^n$.

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

Showing that the ith entry of y^T is the inner product of x and the ith column of A.

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

So y^T is a linear combination of the rows of A, where the coefficients of the linear combination are given by the entries of x.

2.3 matrix-matrix products

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

$$C_{i,j} = \sum_{k=1}^{n} A_{i,k} B_{k,j}$$

This is ***not*** the same as a matrix containing the product of individual elements. That is the *element-wise*, or *Hadamard product*, denoted $A \odot B$.

The CS229 notes go through four different ways of viewing matrix multiplication.

First, matrix-matrix multiplication as a set of vector-vector products, where A is represented by rows and B is represented by columns. This is the way that matrix multiplication is usually taught.

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

Next, representing A by columns and B by rows:

$$C = AB = \begin{bmatrix} | & | & & | \\ a_1 & a_2 & \cdots & a_p \\ | & | & & | \end{bmatrix} \begin{bmatrix} - & b_1^T & - \\ - & b_2^T & - \\ & \vdots & \\ - & b_m^T & - \end{bmatrix} = \sum_{i=1}^{n} a_i b_i^T$$

This representation is not as intuitive. Conceptually, it calculating the matrix by summing together n matrices where each entry is the ith element of the sum in each element of C. This is in contrast to the canonical representation above, in which you go element by element in C and calculate the entire sum for each element individually.

Matrix-matrix multiplication can also be represented as a set of matrix-vector products. Representing B by columns:

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

Each column in C can then be interpreted as in section 2.2 on matrix-vector products.

Representing A by rows:

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

3.2 the transpose

Some properties:

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

This property, along with the fact that a scalar is equal to its own transpose, can be used to show that the dot product is commutative:

$$x^T y = (x^T y)^T = y^T x$$

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

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 $A = -A^T$.

For any matrix $A \in \mathbb{R}^{n \times n}$, the matrix $A + A^T$ is symmetric:

$$\begin{aligned} A + A^T &= \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n,1} & A_{n,2} & \cdots & A_{n,n} \end{bmatrix} + \begin{bmatrix} A_{1,1} & A_{2,1} & \cdots & A_{n,1} \\ A_{1,2} & A_{2,2} & \cdots & A_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{1,n} & A_{2,n} & \cdots & A_{n,n} \end{bmatrix} \\ &= \begin{bmatrix} 2A_{1,1} & A_{1,2} + A_{2,1} & \cdots & A_{1,n} + A_{n,1} \\ A_{2,1} + A_{1,2} & 2A_{2,2} & \cdots & A_{2,n} + A_{n,2} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n,1} + A_{1,n} & A_{n,2} + A_{2,n} & \cdots & 2A_{n,n} \end{bmatrix} \end{aligned}$$

, which is symmetric due to commutativity of addition.

Similarly, the matrix $A - A^T$ is anti-symmetric.

From these properties, 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:

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

Symmetric matrices have nice properties and occur often, particularly when they are generated by a function of two arguments that does not depend on the order of the arguments (e.g. a distance measure between two points). The set of all symmetric matrices of size n can be denoted as $\mathbb{S}^n$.

3.4 the trace

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

$$\text{tr}A = \sum_{i=1}^{n} A_{i,i}$$

For A, B, C such that ABC is square, tr$ABC =$ tr$BCA =$ trCAB, and so on for the product of more matrices. This holds even if the resulting products have different dimensions.

3.5 norms

A ***norm*** of a vector $||x||$ is an informal measure of the length or size of a vector. The L^p (also written as ℓ_p) norm is parameterized by p and defined as:

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

The L^2 norm, aka the Euclidean norm, is commonly used and represents the Euclidean distance from the origin to the point identified by x:

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

Note that the squared L^2 norm $||x||_2^2 = x^T x$. The squared L^2 norm is more convenient to work with mathematically and computationally than the L^2 norm itself.

The L^2 norm increases slowly near the origin. When it is important to discriminate between elements that are exactly zero and elements that are small but non-zero, the L^1 norm is often used because it increases at the same rate in all locations:

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

Another norm is the L^∞ norm, aka the max norm. This represents the absolute value of the element with the largest magnitude in the vector:

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

Norms can also be defined for matrices. In machine learning, the Frobenius norm is often used:

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

3.6 linear independence and rank

A set of vectors $\{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^m$ is ***linearly independent*** if no vector can be represented as a linear combination of the remaining vectors. 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 the vectors $x_1, \ldots, x_n$ are linearly dependent; otherwise, the vectors are linearly independent.

The ***column rank*** of a matrix $A \in \mathbb{R}^{m \times n}$ is the size of the largest subset of columns of A that constitute a linearly independent set. This often refers simply to the number of linearly independent columns of A. Similarly, the ***row rank*** is the largest number of rows of A that constitute a linearly independent set. It turns out that for any matrix $A \in \mathbb{R}^{m \times n}$, the column rank of A is equal to the row rank of A, and are referred to as the ***rank*** of A, rank(A).

Some properties of the rank:

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

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

Not all matrices have inverses. For example, non-square matrices by definition do not have inverses. A square matrix A is ***invertible*** or ***non-singular*** if A^{-1} exists and ***non-invertible*** or ***singular*** otherwise.

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

Some properties of the inverse for non-singular matrices $A, B \in \mathbb{R}^{n \times n}$:

- $(A^{-1})^{-1} = A$

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

When A^{-1} exists, there are several algorithms for finding its closed form. However, because computers can only represent it with limited precision, A^{-1} should not actually be used in practice for most software.

3.8 orthogonal matrices

Two vectors $x, y \in \mathbb{R}^n$ are ***orthogonal*** if $x^T y = 0$ (thinking geometrically, remember that the dot product $= \|x\|\|y\|\cos\theta$, where θ is the angle between the two vectors). 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*** if all its columns are ***orthonormal*** each other, i.e. the columns are orthogonal to each other and normalized. Using the definitions of orthogonality and normality,

$$\begin{aligned}
U^T U &= \begin{bmatrix} - & u_1^T & - \\ - & u_2^T & - \\ & \vdots & \\ - & u_n^T & - \end{bmatrix} \begin{bmatrix} | & | & & | \\ u_1 & u_2 & \cdots & u_n \\ | & | & & | \end{bmatrix} \\
&= \begin{bmatrix} u_1^T u_1 & u_1^T u_2 & \cdots & u_1^T u_n \\ u_2^T u_1 & u_2^T u_2 & \cdots & u_2^T u_n \\ \vdots & \vdots & \ddots & \vdots \\ u_n^T b_1 & u_n^T u_2 & \cdots & u_m^T u_n \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \\
&= I
\end{aligned}$$

Similarly, $I = UU^T$, so

$$U^T U = I = UU^T$$

That is, the inverse of an orthogonal matrix is its transpose. Orthogonal matrices are thus useful, since computing the transpose of a matrix is much cheaper than computing its inverse. Note that the columns of an orthogonal matrix must be orthogonal **and** normalized. There is no special term for a matrix whose columns are orthogonal but not normal.

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

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

If $\{x_1, \ldots, x_n\}$ is a set of n linearly independent vectors, where each $x_i \in \mathbb{R}^n$, then 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 $\{x_1, \ldots, x_n\}$ (assuming $x_i \in \mathbb{R}^m$) is the vector $v \in$ span$(\{x_1, \ldots, x_n\})$ such that v is as close as possible to y, as measured by the Euclidean norm $\|v - y\|_2$. The projection is denoted Proj$(y; \{x_1, \ldots, x_n\})$ and is defined

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

The ***range*** (aka the columnspace) of a matrix $A \in \mathbb{R}^{m \times n}$, denoted $\mathcal{R}(A)$, is the span of the columns of A. That is,

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

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

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

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.,

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

3.10 the determinant

The ***determinant*** of a square matrix $A \in \mathbb{R}^{n \times n}$ is a function det : $\mathbb{R}^{n \times n} \to \mathbb{R}$ denoted $|A|$ or det$A$. The CS229 notes begin with a geometric interpretation of the determinant reproduced here.

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 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 is a measure of the "volume" of the set S. It can also be thought of as a measure of how much multiplication by the matrix expands or contracts space. If the determinant is 0, then space is contracted completely along at least one dimension, causing it to lose all of its volume. If the determinant is 1, then the transformation is volume-preserving.

The notes then go through a graphical example of a 2×2 matrix, in which the set S has the shape of a *parallelogram*. In three dimensions, S corresponds to a *parallelepiped*, and in n dimensions, S corresponds to an n-dimensional *parallelotope*.

Algebraically, the determinant satisfies the following three properties:

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 $t|A|$,

$$\left| \begin{bmatrix} - & ta_1^T & - \\ - & a_2^T & - \\ & \vdots & \\ - & a_n^T & - \end{bmatrix} \right| = t|A|$$

(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

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

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 $A \in \mathbb{R}^{n\times n}$, $|A| = 0$ iff 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, $|A^{-1}| = 1/|A|$.

In order to define the determinant, it is useful to define, for $A \in \mathbb{R}^{n\times n}$, the matrix $A_{\setminus i,\setminus j} \in \mathbb{R}^{(n-1)\times(n-1)}$ to be the matrix that results from deleting the ith row and the jth column from A. The determinant of this matrix is called the (i,j)***minor*** of A, sometimes denoted M_{ij} (i.e. $|A_{\setminus i,\setminus j}| = M_{ij}$). The general (recursive) formula for the determinant is

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

with the initial case that $|A| = A_{11}$ for $A \in \mathbb{R}^{1\times 1}$. The formula expanded completely for $A \in \mathbb{R}^{n\times n}$ would have a total of $n!$ different terms, so it's usually never written out explicitly.

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

$$\begin{aligned} \text{adj}(A) \in \mathbb{R}^{n\times n}, (\text{adj}(A))_{ij} &= (-1)^{i+j} |A_{\setminus j,\setminus i}| \\ &= (-1)^{i+j} M_{ji} \end{aligned}$$

Note the switch in the indices M_{ji}. This is because the classical adjoint is the transpose of the cofactor matrix $C \in \mathbb{R}^{n\times n}, C_{ij} = (-1)^{i+j} M_{ij}$.

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 matrix inverse, in practice there are more efficient ways of computing the inverse numerically.

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 $x^T A x$ is called a ***quadratic form***. Written explicitly, we see that

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

Note that,

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

where the first equality follows from the fact that the transpose of a scalar is equal to itself, and the third equality follows from the fact that $A = A^T$, so 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 (see section 3.3 on symmetric matrices). For this reason, we often implicitly assume that the matrices appearing in a quadratic form are symmetric.

We (the CS229 notes, that is) 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 non-zero $x \in \mathbb{R}^n, x^T Ax \leq 0$.
- Finally, a symmetric matrix $A \in \mathbb{S}^n$ is ***indefinite*** if it 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 $x_2^T Ax_2 < 0$.

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 the 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

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

(If this is not immediately clear, write out the matrix-vector product explicitly, representing A as columns as in section 2.2). $Ax = 0$ implies that $x^T Ax = 0$ 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 attention. 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*** if

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

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 leaves some ambiguity, since x and $-x$ will both be eigenvectors, but we will live with this).

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

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

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

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

This determinant is a (very large) polynomial in λ, where λ will have maximum degree n. The n (possibly complex) roots of this polynomial are the eigenvalues $\lambda_1, \ldots, \lambda_2$. To find the eigenvector corresponding to the eigenvalue λ_i , we solve the linaer equation $(\lambda_i I - A)x = 0$. It should be noted that in practice this method is not used to numerically compute the eigenvalues and eigenvectors (remember that the complete expansion of the determinant has $n!$ terms).

The following are properties of eigenvalues and eigenvectors, in all cases assuming $A \in \mathbb{R}^{n\times n}$ with eigenvalues $\lambda_1, \ldots, \lambda_n$ and associated eigenvectors $x_1, \ldots, x_n$:

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

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

- 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, left-multiply each side of the eigenvector equation $Ax_i = \lambda_i x_i$ by A^{-1}).
- The eigenvalues of a diagonal matrix $D = \text{diag}(d_1, \ldots, d_n)$ 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.,

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

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

The eigenvalues and eigenvectors of a symmetric matrix $A \in \mathbb{S}^n$ have two nice properties. First, it can be shown that all of the eigenvalues of A are real. Secondly, the eigenvectors of A are orthonormal, i.e. the matrix X as 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 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 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, 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,

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

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.

4.1 the gradient

Suppose that $f : \mathbb{R}^{m \times n} \rightarrow$ 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 $x \in \mathbb{R}^n$ (note: this is a typo in the original notes) is a vector of fixed coefficients. Let $f : \mathbb{R}^m \rightarrow \mathbb{R}$ 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 $\nabla f(z) = 2z$. Here, we interpret $\nabla(Ax)$ 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, while the other interpretation yields an n-dimensional vector.

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 $\nabla_x f(Ax)$.

4.2 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. This intuition is generally correct, but there are a few caveats to keep in mind.

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.,

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

as 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

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

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) $\nabla_x^2 f(x) = \nabla_x (\nabla_x f(x))^T$, as 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.

4.3 gradients and hessians of linear and quadratic functions

Now let's determine the gradient and Hessian matrices for a few simple functions.

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 see that $\nabla_x b^T x = b$. This is analogous to the situation in single-variable calculus, where $\frac{\partial}{\partial x} ax = a$.

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

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

(see section 3.11). 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, $\nabla_x x^T Ax = 2Ax$. Again, this is analogous to single-variable calculus, where $\frac{\partial}{\partial x} ax^2 = 2ax$.

Finally, let's look at the Hessian of the quadratic function $f(x) = x^T Ax$ (the Hessian of a linear function $b^T x$ is just zero). In this case,

$$\frac{\partial^2 f(x)}{\partial x_k \partial x_l} = \frac{\partial}{\partial x_k} \left[\frac{\partial f(x)}{\partial x_l} \right] = \frac{\partial}{\partial x_k} \left[2 \sum_{i=1}^{n} A_{li} x_i \right] = 2A_{lk} = 2A_{kl}.$$

Therefore, $\nabla_x^2 x^T Ax = 2A$, which is again analogous to the single-variable case where $\frac{\partial}{\partial x^2} ax^2 = 2a$.

To recap,

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

4.5 gradients of the determinant

Let's consider a situation where we want to 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 (section 3.10) that

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

so

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

From this and the properties of the adjugate it follows 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 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 (from single-variable calculus) to see that

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

So,

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

where we can drop the transpose in the last expression because A is symmetric. Note the similarity to the single-variable case, where $\frac{\partial}{\partial x} \log x = \frac{1}{x}$.

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 constrains is by forming the ***Lagrangian***, an objective function that includes the equality constraints. The Lagrangian in this case is given by

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

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

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

Notice that this is just the linear eigenvalue 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 or A.