APM462H1S: Nonlinear optimization, Review of calculus and linear algebra.

This is a list, probably far from complete, of some facts about calculus and linear algebra that you should know.

1. CALCULUS

1.1. single-variable.

Let g be a real-valued function of a single variable.

(1) definition of derivative:

$$g'(x) = \lim_{h \to 0} \frac{g(x+h) - g(x)}{h} \qquad \text{(when it exists.)}$$

(2) basic facts such as the chain rule, product rule, quotient rule.

(3) derivative of polynomials, exponentials, basic trigonometric functions, inverse functions.

(4) mean value theorem: if g is C^1 then for every $x < y$ there exists some $z \in [x, y]$ such that

$$g(y) - g(x) = (y - x)g'(z).$$

(5) Fundamental theorem of calculus:

- If g is C^1 then $\int_a^b g'(x)\,dx = g(b) - g(a)$, and...
- if g is continuous and $G(x) = \int_a^x g(y)\,dy$ for all x, then $G' = g$.

(6) Among techniques for integration that you have learned, the one you see most often in higher math courses is probably integration by parts, which follows from combining the product rule and the (first part of the) Fundamental Theorem of Calculus. Thus:

$$\int_a^b g_1'(x)g_2(x)\,dx + \int_a^b g_1(x)g_2'(x)\,dx \;=\; (g_1\, g_2)|_a^b.$$

(7) First-order Taylor approximation: If g is C^1 then for any x and any h,

$$g(x+h) = g(x) + hg'(x) + \mathrm{rem}_1(x, h)$$

where, for every x, the remainder term $\mathrm{rem}_1(x, h)$ satisfies

$$\lim_{h \to 0} \frac{\mathrm{rem}_1(x, h)}{|h|} = 0$$

(8) Second-order Taylor approximation: If g is C^2 then for any x and any h,

$$g(x+h) = g(x) + hg'(x) + \frac{1}{2}h^2 g''(x) + \mathrm{rem}_2(x, h)$$

where, for every x, the second-order remainder term $\mathrm{rem}_2(x, h)$ satisfies

$$\lim_{h \to 0} \frac{\mathrm{rem}_2(x, h)}{h^2} = 0$$

(9) For every positive integer k, if g is C^k then there is also a kth order Taylor approximation, which we may not need in this course.

(10) If g is C^2 and x^* is a local minimum point of g, then

$$g'(x^*) = 0, \qquad g''(x^*) \geq 0.$$

The converse is not true: it can happen that $g'(x^*) = 0$ and $g''(x^*) \geq 0$ at a point x^* that is not a local minimum for g. However:

if $g'(x^*) = 0$ and $g''(x^*) > 0$, then x^* is a local minimum point for g.

This is a consequence of the second-order Taylor approximation, for example.

(11) Taylor approximations can be written in a concise way using the "little o" notation. In this notation, $o(h)$ denotes a quantity that is negligible, compared to h, as $h \to 0$. In other words, by definition, $o(h)$ means:

$$\text{if } (\cdots) = o(h), \quad \text{then} \quad \lim_{h\to 0} \frac{(\cdots)}{h} = 0.$$

More generally for any positive power p, the symbol $o(h^p)$ means:

$$\text{if } (\cdots) = o(h^p), \quad \text{then} \quad \lim_{h\to 0} \frac{(\cdots)}{h^p} = 0.$$

With this notation, the first- and second-order Taylor approximations can be written

$$g(x+h) = g(x) + hg'(x) + o(h)$$
$$g(x+h) = g(x) + hg'(x) + \frac{1}{2}h^2 g''(x) + o(h^2)$$

(12) In addition to the "little o" notation, there is also a "big O" notation. We will not use it in this class, probably, but please be careful, when writing little o's, to write "$o(h)$" instead of "$O(h)$".

1.2. **multi-variable.** In this discussion, I will use the notation from the textbook. Thus, the gradient is always considered to be a *row* vector (that is, an element of E_n), and points in E^n are considered to be *column* vectors.

Unlike the textbook, in these notes vectors, matrices etc are not indicated by boldface type.

(1) definition of the gradient: if f is a real-valued function on E^n and $x \in E^n$, then ∇f, when it exists, is a (row) vector characterized by the property

$$\lim_{|v|\to 0} \frac{f(x+v) - f(x) - \nabla f(x)v}{|v|} = 0. \tag{1}$$

Here and below, we use the notation

$$|v| = (v_1^2 + \cdots + v_n^2)^{1/2}$$

for $v = (v_1, \ldots, v_n)$. In practice, the gradient of f is given by the formula

$$\nabla f(x) = [\tfrac{\partial f}{\partial x_1}(x), \ldots, \tfrac{\partial f}{\partial x_n}(x)].$$

In formula (1), since $\nabla f(x)$ is a row vector and v is a column vector,

$$\begin{aligned}\nabla f(x)v &= v_1\frac{\partial f}{\partial x_1}(x) + \ldots + v_n\frac{\partial f}{\partial x_n}(x) \\ &= \sum_{i=1}^{n} v_i\frac{\partial f}{\partial x_i}(x).\end{aligned}$$

(2) We checked in the lecture that if p is any row vector such that

$$\lim_{|v|\to 0}\frac{f(x+v) - f(x) - p\,v}{|v|} = 0.$$

then $p = \nabla f(x)$. The same fact can be expressed by saying that formula (1) uniquely determined the gradient.

(3) The Divergence Theorem and Stokes' Theorem can be understood as multivariable analogs of the Fundamental Theorem of Calculus. We may not need these theorems in this course

(4) First-order Taylor approximation: If f is a C^1 function on E^n, then for any x and any v,

$$f(x+v) = f(x) + \nabla f(x)\,v + \mathrm{rem}_1(x,v)$$

where, for every x, the remainder term $\mathrm{rem}_1(x,v)$ satisfies

$$\lim_{v\to 0}\frac{\mathrm{rem}_1(x,v)}{|v|} = 0$$

(5) Second-order Taylor approximation: if f is a C^1 function on E^n, then for any x and any v,

$$f(x+v) = f(x) + \nabla f(x)\,v + \frac{1}{2}v^T\,\nabla^2 f(x)v + \mathrm{rem}_2(x,v)$$

where, for every x, the second-order remainder term $\mathrm{rem}_2(x,v)$ satisfies

$$\lim_{v\to 0}\frac{\mathrm{rem}_2(x,v)}{|v|^2} = 0$$

In the above formula, $\nabla^2 f(x)$ is the matrix of second derinvatives, so

$$v^T\nabla^2 f(x)v = \sum_{i,j=1}^{n} v_i v_j\frac{\partial^2 f}{\partial x_i \partial x_j}(x).$$

(6) For every positive integer k, if g is C^k then there is also a kth order Taylor approximation, which we will probably not need in this course.

(7) You may not have seen this in earlier classes, but as discussed in the first lecture: if f is a C^2 function on E^n and x^* is a local minimum point of f, then

$$\nabla f(x^*) = 0, \qquad v^T\nabla^2 f(x^*)v \geq 0 \text{ for all } v \in E^n.$$

The converse is not true: it can happen that the above conditions bother hold at a point x^* that is not a local minimum. But if f is C^2 then

$$\text{if } \nabla f(x^*) = 0 \text{ and } v^T \nabla^2 f(x^*) v > 0 \text{ for all nonzero } v \in E^n, \quad \text{then } x^* \text{ is a local minimum point for } g.$$

This is a consequence of the second-order Taylor approximation, for example.

If f is only C^1, then it is still true that $\nabla f(x^*) = 0$, but we cannot say anything about second derivatives (which may not exist!)

(8) Using the "little o" notation, the first- and second-order Taylor approxmations can be written

$$f(x+v) = f(x) + \nabla f(x)\, v + o(|v|)$$
$$f(x+v) = f(x) + \nabla f(x)\, v + \frac{1}{2} v^T \nabla^2 f(x) v + o(|v|^2)$$

Similarly, the fact stated in point (2) above can be restated as:

$$\text{if } f(x+v) = f(x) + p\, v + o(|v|), \qquad \text{then } p = \nabla f(x).$$

2. some linear algebra

(1) Transpose: Recall that if A is an $n \times m$ matrix

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}$$

then the *transpose* of A is the $m \times n$ matrix, denoted A^T, defined by

$$A^T = \begin{pmatrix} a_{11} & \cdots & a_{n1} \\ \vdots & \ddots & \vdots \\ a_{1m} & \cdots & a_{nm} \end{pmatrix}.$$

In particular, the transpose of a row vector is a column vector, and vice versa.

It is clear that $(A^T)^T = A$.

Recall also that $(AB)^T = B^T A^T$ and more generally

$$(A_1 \dots A_k)^T = A_k^T \cdots A_1^T.$$

In particular, if $a, b \in E^n$ then $a^T b$ is a number (that is, a 1×1 matrix) so

$$a^T b = (a^T b)^T = b^T a.$$

(2) you should remember definitions and basic facts about *linear dependence*, *linear independence*, a *basis* for a vector space.

(3) you should remember basic facts about determinants, eigenvalues, and eigenvectors:

(a) how to compute the determinant of a matrix.

(b) a square matrix is said to be *invertible* if it has an inverse (obviously), and *singular* if not.
A matrix is A is singular if and only if $\det A = 0$, and this happens if and only if there is a nonzero vector v solving the equation $Av = 0$.

(c) a number λ and vector v are called an *eigenvalue* and *eigenvector* if v is nonzero and

$$Av = \lambda v.$$

(d) It follows that

$$\lambda \text{ is an eigenvalue} \quad \text{if and only if} \quad \det(A - \lambda I) = 0,$$

since an eigenvector is a nonzero solution of the equation $(A-\lambda I)v = 0$.

(e) Thus, the eigenvalues of a $n \times n$ matrix A are exactly the roots of the polynomial $\det(A - \lambda I)$, which is a nth order polynomial in the variable λ.

(4) A matrix S is *symmetric* if $S = S^T$.
Symmetric matrices are important for us because if a function f is C^2, then the matrix $\nabla^2 f(x)$ of second derivatives is symmetric at every x.
Important fact: if S is symmetric, then all the eigenvalues are real, and there exist eigenvectors $v_1, \ldots, v_n$ which form an orthonormal basis for E^n. This means that for every $i, j \in \{1, \ldots, n\}$,

$$v_i^T v_j = \delta_{ij} := \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j. \end{cases}$$

(This implies that the vectors $v_1, \ldots, v_n$ are linearly independent, and hence a basis, since any set of n linearly independent vectors in E^n forms a basis.)

(5) A matrix Q is *orthogonal* if $Q^T = Q^{-1}$.
To check that a matrix Q is orthogonal, it suffices to check that either $QQ^T = I$ or that $Q^TQ = I$; either one of these identities implies the other (since in general, if $AB = I$ then $BA = I$ and $B = A^{-1}$.)
It follows that a matrix is orthogonal if and only if either
- its columns form an orthonormal basis for E^n, or
- its rows form an orthonormal basis for E_n.

and that if one of these conditions holds, then the other does too.
Indeed, if Q is any matrix and $c_1, \ldots, c_n$ denote the columns of Q, then Q^TQ is the matrix whose (i, j) entry equals $q_i^T q_j$. So

$$Q^TQ = I \qquad \text{if and only if} \qquad c_i^T c_j = \delta_{ij} \text{ for all } i, j.$$

(6) in particular, suppose that S is a symmetric $n \times n$ matrix, with eigenvalues $\lambda_1, \ldots, \lambda_n$ and an orthonormal set of eigenvectors $v_1, \ldots, v_n$
If we define Q to be the matrix with columns $v_1, \ldots, v_n$ (in that order) and D to be the diagonal matrix with entries $\lambda_1, \ldots, \lambda_n$ (in that order) then the set of n equations

$$Sv_i = \lambda_i v_i \quad \text{for} \quad i = 1, \ldots, n$$

can be combined into a single matrix equation

$$SQ = QD.$$

(You can check this by writing it out in more detail.) Since $Q^T = Q^{-1}$, this is equivalent to the equation $D = Q^T SQ$, and also to the equation $S = QDQ^T$.

In particular, *every symmetric matrix S can be written in the form*

$$S = Q^T DQ, \text{ where } Q \text{ is orthogonal and } D \text{ is diagonal.}$$

(7) A symmetric matric S is *positive semidefinite* (sometimes also called *nonnegative definite*) if and only if

$$v^T Sv \geq 0 \text{ for all vectors } v \in E^n$$

and *positive definite* if

$$v^T Sv > 0 \text{ for all } \textit{nonzero} \text{ vectors } v \in E^n$$

By writing S in the form $S = Q^T DQ$, one can check that s symmetric matrix S is

$$\text{positive semidefinite} \iff \text{all its eigenvalues are nonnegative, and}$$
$$\text{positive definite} \iff \text{all its eigenvalues are positive.}$$

3. some proofs (optional!)

We recall (omitting some details) why Taylor approximations are valid, starting with functions of a single variable.

These proofs will not be needed in this class, so you should read this only if you are interested.

First-order Taylor approximation, single variable It follows from the fundamental theorem of calculus that

$$g(x+h) - g(x) = \int_x^{x+h} g'(y)\ dy.$$

Also, by definition, $\text{rem}_1(x,h) = g(x+h) - g(x) - hg'(x)$. It follows that

$$\begin{aligned} \text{rem}_1(x,h) &= \int_x^{x+h} g'(y)\,dy - hg'(x) \\ &= \int_x^{x+h} (g'(y) - g'(x))\ dy \end{aligned} \tag{2}$$

since

$$hg'(x) = \int_x^{x+h} g'(x)\,dy$$

(because $g'(x)$ is a constant with respect to the variable of integration y.)

Thus to verify that $\lim_{h\to 0} \frac{1}{h}\text{rem}_1(x,h) = 0$, it suffices to check that

$$\lim_{h\to 0} \frac{1}{h} \int_x^{x+h} (g'(y) - g'(x))\ dy \ = \ 0,$$

and this can be done using the continuity of g'. To do this in detail, one would need to recall the definition of continuity and of limit in terms of δ and ϵ.

The first-order Taylor approximation can be derived more easily as a direct consequence of the definition of the derivative, but the longer derivation that we have given here, with the explicit formula for the remainder term $\mathrm{rem}_1(x,h)$, is needed below.

Second-order Taylor approximation, single variable Our discussion of the first-order approximation shows that

$$g(x+h) = g(x) + hg'(x) + \int_x^{x+h} (g'(y) - g'(x))\, dy. \tag{3}$$

Since g is C^2, clearly g' is C^1, so the formula remains true if we replace g by g' everywhere. If we further replace $x+h$ by y and rewrite things a little, we get

$$g'(y) - g'(x) = (y-x)g''(x) + \int_x^{y} (g''(z) - g''(x))\, dz. \tag{4}$$

Now substitute (4) in the integral on the right-hand side of (3) to get

$$g(x+h) = g(x) + hg'(x) + \int_x^{x+h} \left((y-x)g''(x) + \int_x^{y} (g''(z) - g''(x))\, dz \right) dy.$$

We simplify to find that

$$g(x+h) = g(x) + hg'(x) + \frac{1}{2}h^2 g''(x) + \mathrm{rem}_2(x,h), \tag{5}$$

where

$$\mathrm{rem}_2(x,h) = \int_x^{x+h} \int_x^{y} (g''(z) - g''(x))\, dz\, dy \tag{6}$$

Then one can check, using the continuity of g'', that $\lim_{h\to 0} \frac{1}{h^2}\mathrm{rem}_2(x,h) = 0$. To do this in detail, one would need to recall the definition of continuity and of limit in terms of δ and ϵ.

Taylor approximations, several variables

We will discuss only the case of a second-order approximation. First-order Taylor approximations are similar but easier.

Suppose that f is a C^2 function on E^n. Fix a point $x \in E^n$. Given any nonzero vector $v \in E_n$, we can write

$$h := |v|, \qquad d := \frac{v}{h} = \frac{v}{|v|}.$$

Thus d is a unit vector (that is, it satisfies $|d| = 1$) pointing in the same direction as v, and h is the length of v.

Now let us define

$$g(t) := f(x + td).$$

Then $f(x+v) = f(x+hd) = g(h)$.

We know from (5), (6) that

$$g(h) - \left[g(0) + hg'(0) + \frac{1}{2}h^2 g''(0) \right] = \mathrm{rem}_2(0,h) = \int_0^h \int_0^s (g''(t) - g''(0))\, dt\, ds \tag{7}$$

Also, by the chain rule (for functions of several variables)

$$hg'(0) = h \sum_{i=1}^n \frac{\partial f}{\partial x_i}(x) d_i = \sum_{i=1}^n \frac{\partial f}{\partial x_i}(x) v_i = \nabla f(x) v,$$

and similarly one can check that

$$g''(t) = \sum_{i,j=1}^{n} \frac{\partial^2}{\partial x_i \partial x_j} f(x + t\frac{v}{|v|}) v_i v_j \;=\; v^T \nabla^2 f(x + t\frac{v}{|v|}) v$$

for every t Thus rewriting (7) in terms of f, we find that

$$f(x+v) = f(x) + \nabla f(x) v + v^T \nabla^2 f(x)\, v + \mathrm{rem}_2(x, v)$$

where

$$\mathrm{rem}_2(x, v) = \int_0^{|v|} \int_0^s \sum_{i,j=1}^{n} \left[\frac{\partial^2 f}{\partial x_i \partial x_j}(x + t\frac{v}{|v|}) - \frac{\partial^2 f}{\partial x_i \partial x_j}(x) \right] v_i v_j \; dt \; ds.$$

Then as before, one can check using the continuity of the second partial derivatives (a consequence of the fact that f is C^2) that

$$\lim_{|v| \to 0} \frac{1}{|v|^2} \mathrm{rem}_2(x, v) = 0.$$