John Nachbar
Washington University
March 25, 2018

A Quick Introduction to Linear Algebra, Topology, and Multivariate Calculus[1]

1 Vector Spaces and Linear Algebra.

1.1 Overview.

Definition 1. *A set $V \subseteq \mathbb{R}^N$ is a* vector space *iff the following two properties hold.*

1. *For any $x, \hat{x} \in V$, $x + \hat{x} \in V$.*
2. *For any $x \in V$, for any $\gamma \in \mathbb{R}$, $\gamma x \in V$.*

Because γ can equal 0, we always have that the origin $(0, \ldots, 0)$ is in V. I will write the origin in $\mathbb{R}^N$ as simply "0," but to avoid confusion with the real number 0, some authors write the origin in $\mathbb{R}^N$ as, say, θ.

A vector space is a line, plane, or higher dimensional analog thereof, through the origin. Thus, for example, for $x \in \mathbb{R}$, the graph of the line $f(x) = ax$ is a vector space in $\mathbb{R}^2$.

On the other other hand, the graph of $\hat{f}(x) = ax + b$, with $b \neq 0$, is not a vector space because the graph does not go through the origin in $\mathbb{R}^2$. The graph of $\hat{f}(x)$ is instead an example of a *linear manifold*. A linear manifold is the result of taking a vector space and shifting it in a parallel fashion away from the origin.

$\mathbb{R}^N$ itself is a vector space. Thus it is also common to see a vector space $V \subseteq \mathbb{R}^N$ called a vector *sub*space of $\mathbb{R}^N$.

1.2 Spanning, Linear Independence, and Bases

In the example above, the vector space V given by the graph of $f(x) = ax$ can also be represented in the form $V = \{(x, y) \in \mathbb{R}^2 : \text{ there is a } \gamma \in \mathbb{R} \text{ such that } (x, y) = \gamma(1, a)\}$. The vector $(1, a) \in V$ is said to span V. More generally, we have the following. As a matter of notation, s^t denotes the vector $s^t = (s_1^t, \ldots, s_N^t) \in \mathbb{R}^N$.

Definition 2. *Given a vector space $V \subseteq \mathbb{R}^N$ and a set of T vectors $S = \{s^1, \ldots, s^T\}$, all in V, S* spans *V iff for any $x \in \mathbb{R}^N$ there exist $\gamma^1, \ldots, \gamma^T$, all in $\mathbb{R}$, such that*

$$x = \gamma^1 s^1 + \cdots + \gamma^T s^T.$$

In the example, note that although $\{(1,a)\}$ spans V, so does $\{(1,a),(2,2a)\}$. In particular it is always possible to take $\gamma^2 = 0$, which puts us back in the previous case. Including $(2,2a)$ in the spanning set when we already have $(1,a)$ is redundant. We are interested in spanning by sets of vectors that are minimal in the sense of not having any such redundancies. Given a set $S = \{s^1, \ldots, s^T\}$ of T vectors, there is a redundancy if it is possible to write one of the vectors, say s^1, as a linear combination of the other vectors,

$$s^1 = \gamma^2 s^2 + \ldots \gamma^T s^T.$$

This can be rewritten as

$$-s^1 + \gamma^2 s^2 + \ldots \gamma^T s^T = 0.$$

In the example above, I could write either $(1,a) = (1/2)(2,2a)$ or $(2,2a) = 2(1,a)$.

This motivates the following definition.

Definition 3. *A set $S = \{s^1, \ldots, s^T\}$ of T vectors in $\mathbb{R}^N$ is* linearly dependent *if there exists T numbers $\gamma^1, \ldots, \gamma^T$, at least one not equal to zero, such that*

$$\gamma^1 s^1 + \cdots + \gamma^T s^T = 0.$$

If S is not linearly dependent then it is linearly independent. *Equivalent, S is linearly independent iff whenever*

$$\gamma^1 s^1 + \cdots + \gamma^T s^T = 0,$$

$\gamma^1 = \cdots = \gamma^T = 0$.

In particular, if $\gamma^1 s^1 + \cdots + \gamma^T s^T = 0$ and, say, $\gamma^1 \neq 0$ then s^1 is redundant in the sense that $s^1 = -(1/\gamma^1)(\gamma^2 s^2 + \cdots + \gamma^T s^T)$.

If S contains the origin as one of its vectors then it is automatically linearly dependent. In particular, if S contains *only* the origin then it is linearly dependent. Specializing even further, in the case $N = 1$, the "matrix" [0] is linearly dependent; on the other hand, the "matrix" [1] is linearly independent.

Even if S is linearly dependent, some subset of S may be linearly independent. And when there is one linearly independent subset of S, then there is often more than one. For example, if $S = \{(1,a),(2,2a)\}$ then S is linearly dependent (take $\lambda^1 = -2$, $\lambda^2 = 1$), but both $\hat{S} = \{(1,a)\}$ or $\tilde{S} = \{(2,2a)\}$ are linearly independent.

Definition 4. *S is a* basis *for V iff S is linearly independent and spans V.*

Except for the trivial case $V = \{0\}$, there will be many (uncountably infinitely many, in fact) bases. In the example above, $(1,a)$ is a basis but so is $(2,2a)$ and so is $(-1,-a)$.

One can show that if there is a basis for V with T vectors then *every* basis for V has exactly T vectors. This allows us to define the dimension of V.

Definition 5. *If V is a vector space then the dimension of V is T iff there is a basis for V with T vectors.*

In particular, the dimension of $\mathbb{R}^N$ is N, because the N unit vectors $e^1 = (1, 0, \ldots, 0)$, $e^2 = (0, 1, 0, \ldots, 0)$ and so on are a basis for $\mathbb{R}^N$. This basis is called the *standard basis*.

Finally, note that if V is the vector space and S is a basis for V then any $x \in V$ is *uniquely* represented in the form

$$x = \gamma^1 s^1 + \cdots + \gamma^T s^T.$$

To see this, suppose that we also have

$$x = \hat{\gamma}^1 s^1 + \cdots + \hat{\gamma}^T s^T.$$

Then setting these equal and rearranging,

$$0 = (\gamma^1 - \hat{\gamma}^1)s^1 + \cdots + (\gamma^T - \hat{\gamma}^T)s^T$$

If S is independent, all of the $\gamma^t - \hat{\gamma}^t$ must equal 0, which implies that for all t, $\hat{\gamma}^t = \gamma^t$.

1.3 Linear Functions and Matrices.

Definition 6. *A function $f : \mathbb{R}^N \to \mathbb{R}^M$ is* linear *iff the following hold.*

1. *For any $x, \hat{x} \in \mathbb{R}^N$,*
$$f(x + \hat{x}) = f(x) + f(\hat{x}).$$
2. *For any $x \in \mathbb{R}^N$, $\gamma \in \mathbb{R}$,*
$$f(\gamma x) = \gamma f(x).$$

When $M > 1$, linear functions are often called linear *maps*. Map and function mean the same thing here.

Setting $\gamma = 0$ implies that if f is linear then $f(0) = 0$. Thus, in the examples above, $f(x) = ax$ is linear but $\hat{f}(x) = ax + b$ is not when $b \neq 0$ ($\hat{f}$ is *affine*).

A fundamental fact is that a function f is linear iff it can be represented in matrix form: there is an $M \times N$ matrix A (note the dimensions) such that for any $x \in \mathbb{R}^N$,

$$f(x) = Ax.$$

To see that this is true, note that for any $x \in \mathbb{R}^N$, $x = x_1 e^1 + \cdots + x_N e^N$, where e^n is the unit vector with a 1 in coordinate n and 0s everywhere else. Then since f is linear

$$f(x) = x_1 f(e^1) + \cdots + x_N f(e^N).$$

Let $a^n = f(e^n)$. Let A be the $M \times N$ matrix in which column n is a^n. Then

$$f(x) = Ax.$$

In our simple example in which $f(x) = ax$, $f(1) = a$ and so $A = [a]$.

1.4 The Fundamental Spaces of a Linear Function

Given an $M \times N$ matrix A, let a^n denote column n of A. Then for the linear function $f(x) = Ax$,

$$f(x) = x_1 a^1 + \cdots + x_N a^N.$$

In words, this says that $f(x)$ is in the vector space in $\mathbb{R}^M$ that is spanned by the columns of A. This space is called the *column space*.

Similarly, one can consider the vector space spanned in $\mathbb{R}^N$ by the *rows* of A, considered as vectors (equivalently, consider the columns of A', the transpose of A).

Finally, let $K(A)$ be the set of points x such that

$$Ax = 0.$$

$K(A)$ is the *kernel* of A. (It is also called the *null space* of A.) It is easy to verify that $K(A)$ is a vector space in $\mathbb{R}^N$.

1.5 The Fundamental Dimensionality Theorem.

One can prove the following theorem.

Theorem 1 (The Dimension Theorem)**.** *Let A be an $M \times N$ matrix. The dimension of the column space of A plus the dimension of $K(A)$ equals N.*

One consequence of the Dimension Theorem is that, with some additional work, one can show that, for any given matrix, the maximum number of independent columns (i.e. the number of columns in the largest independent subset of columns) equals the maximum number of independent rows (i.e., the maximum number of independent columns of A'). This number is called the *rank* of A. In particular, this says that the column space of A and the row space of A have the same dimension. And this says that the dimension of $K(A)$ is N minus the rank of A.

The rank of A cannot exceed $\min\{M, N\}$ but could be strictly less. For example, if $N = M = 2$ and

$$A = \begin{bmatrix} 1 & 2 \\ a & 2a \end{bmatrix},$$

then the rank is 1. A matrix A has *full rank* iff its rank is $\min\{M, N\}$. Otherwise, A is *singular*. As discussed below, the linear function $f(x) = Ax$ is one-to-one, and hence invertible, iff A has full rank.

1.6 Vector Subspaces Revisited.

Consider any $M \times N$ matrix A. Suppose that $M < N$ and that A has full rank, which is M. Then by the Dimension Theorem, the vector space $K(A)$ has dimension $N - M$. In particular, if $M = 1$ (so that A is a $1 \times N$ "row matrix") then $K(A)$ has dimension $N - 1$.

Conversely, any k-dimensional vector space in $\mathbb{R}^N$ can be expressed as the kernel of an $(N-k)\times N$ matrix. Constructing the matrix in general is a bit of pain, but it is trivial in the $N=2$ case. Consider once again the graph of $f(x)=ax$. This vector space is spanned by $(1,a)$. This vector space can also be viewed as the kernel of the linear map $F(x)=Ax$ where $A=\begin{bmatrix} a & -1 \end{bmatrix}$. In particular, note that $A(x,y)=0$ iff $ax-y=0$, or $y=ax$.

A related comment is that any k dimensional vector space $V\subseteq\mathbb{R}^N$ can be expressed as the graph of a linear function. Suppose that, as above, V is expressed as the kernel of an $(N-k)\times N$ matrix. Write this matrix in the form $\begin{bmatrix} A & B \end{bmatrix}$ where A is $(N-k)\times k$ and B is $(N-k)\times(N-k)$. Write x in the form $x=(p,q)$, where $p\in\mathbb{R}^k$ and $q\in\mathbb{R}^{N-k}$. Then

$$F(x)=F(p,q)=Ap+Bq.$$

If (p,q) belong to the vector space $V=K(A)$ then $Ap+Bq=0$ and hence I can write

$$q=f(p)=-B^{-1}Ap.$$

The vector space V is the graph of f. The Implicit Function Theorem is a generalization of this observation to non-linear analogs of vector spaces called differential manifolds.

1.7 Invertibility and Determinants

In general, a function $f:\mathbb{R}^N\to\mathbb{R}^N$ is invertible iff it is one-to-one. If $f(\mathbb{R}^N)$ is a proper subset of $\mathbb{R}^M$ then the domain of the inverse is $f(\mathbb{R}^N)$ rather than all of $\mathbb{R}^M$.

In the particular case of a linear function $f:\mathbb{R}^N\to\mathbb{R}^M$, $f(x)=Ax$, f is invertible iff its kernel has dimension 0 (is just the origin). To see this, note that $f(x)=f(\hat{x})$ iff $Ax=A\hat{x}$ iff $A(x-\hat{x})=0$, iff $x-\hat{x}\in K(A)$. So if $K(A)=\{0\}$ then $x=\hat{x}$. But if $K(A)$ has positive dimension, then there are $x,\hat{x}$ such that $x\neq\hat{x}$ but $f(x)=f(\hat{x})$: f is not one-to-one. Thus, if f is linear then f is one-to-one, and hence is invertible, iff $K(A)=\{0\}$.

It also follows from the Dimension Theorem that if the dimension of $K(A)$ is 0 then the dimension of the range of f is $N-0=N$. Note that this is impossible if $M<N$, since the rank can never be greater than $\min\{M,N\}$. Therefore, if $M<N$ then f *cannot* be one-to-one and hence cannot be invertible. On the other hand, if $M\geq N$ then f is one-to-one, and hence invertible, iff A has full rank, namely N.

Moreover, in the particular case where $M=N$ (A is square), this arithmetic implies that if A has full rank, and hence is invertible, then f is also *onto*: the column space of A is all of $\mathbb{R}^N$. In this case, the domain of f^{-1} is all of $\mathbb{R}^N$ rather than some proper subset.

There is a function called the determinant that is defined on square matrices (and only on square matrices). The general form is cumbersome but for 2×2 matrices

the determinant is given by

$$\mathrm{Det}\begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc.$$

The critical fact about determinants is that a square matrix A has full rank, and hence is one-to-one, and hence is invertible, iff $\mathrm{Det}A \neq 0$.

A second, less important but sometimes useful, fact is that if you multiply any column by a constant, say γ, then the entire determinant gets multiplied by γ. In particular, if you multiply *every* column by γ then the determinant gets multiplied by γ^N.

2 Topology.

2.1 The Euclidean Metric.

In order to discuss concepts like limit in $\mathbb{R}^N$, we need a way to measure the distance between points in $\mathbb{R}^N$. The standard way of doing this is the Euclidean metric and the associated Euclidean norm. The Euclidean distance between two points, say $a, b \in \mathbb{R}^N$, is defined to be

$$d(a,b) = \sqrt{\sum_n (a_n - b_n)^2} = \sqrt{(a-b)\cdot(a-b)}$$

The function $d : \mathbb{R}^2 \to \mathbb{R}$ is called the *Euclidean metric*. This is *exactly* the formula for everyday distance, subject only to a choice of units (inches, centimeters, etc). If I measure the distance from myself to the door, I am using Euclidean distance. The fact that it happens to have this precise formula is, for most applications, largely irrelevant. What does matter is that the Euclidean metric has the following properties.

1. For any $a, b \in \mathbb{R}$, $d(a,b) \geq 0$, with $d(a,b) = 0$ iff $a = b$.
2. For any $a, b \in \mathbb{R}$, $d(a,b) = d(b,a)$.
3. For any $a, b, c \in \mathbb{R}$, $d(a,b) \leq d(a,c) + d(c,b)$.

The last property is called the Triangle Inequality, referring to the fact that the length of any one side of a triangle is always less than the sum of the lengths of the other two sides.

The *Euclidean norm* of $a \in \mathbb{R}^N$, written $\|a\|$, is defined by $\|a\| = d(a,0) = \sqrt{a \cdot a}$. Thus the norm of a is simply the distance of a to the origin. The Euclidean norm has the following properties.

1. For any $a \in \mathbb{R}^N$, $\|a\| \geq 0$, with $\|a\| = 0$ iff $a = 0$.
2. For any $a \in \mathbb{R}^N$ and any $\gamma \in \mathbb{R}$, $\|\gamma a\| = |\gamma| \|a\|$.
3. For any $a, b \in \mathbb{R}^N$, $\|a + b\| \leq \|a\| + \|b\|$.

2.2 Open Balls.

Given a point $x^* \in \mathbb{R}^N$ and a number $\varepsilon \in \mathbb{R}$, $\varepsilon > 0$, the *open ball of radius ε around* x^* is given by

$$N_\varepsilon(x^*) = \{x \in \mathbb{R}^N : d(x, x^*) < \varepsilon\}.$$

Because of the strict inequality, $N_\varepsilon(x^*)$ does not contain the sphere of radius ε that forms its boundary. Loosely, imagine a basket ball. The "open ball" is everything inside the ball, but not the rubber shell of the ball.

2.3 Sequences, Cauchy Sequences, and Completeness.

Definition 7. *Given a sequence (x_t) in $\mathbb{R}^N$, the sequence converges to x^*, written $\lim x_t = x^*$ or $x_t \to x^*$, iff for any $\varepsilon > 0$ there is a T such that for all $t > T$,*

$$x_t \in N_\varepsilon(x^*).$$

In words, (x_t) converges to x^* iff for any standard of what it means to be "close to" x^* (i.e. for any $\varepsilon > 0$), the sequence will eventually stay at least that close to x^*, forever.

Here and elsewhere, it is critical to get quantifiers correct. If I had instead written "there is an ε," I would have gotten nonsense. For example the sequence 1, -1, 1, -1, ... does not converge, but all x_t are within, say, 100, of 0. And the order of quantifiers matters. Because T comes second, the interpretation is that it is allowed to vary with ε: choose a smaller ε and you may have to choose a larger T. If you reverse the order, the convergence condition becomes hopelessly strong. For example, the sequence 1/2, 1/3, 1/4, ... converges to 0. But I can't get the above condition to hold if I have to fix T in advance, independently of ε. If I choose $T = 1000$, for example, then it will *not* be true that for all $t > T$, $x_t \in N_{1/10,000}(0)$.

Definition 8. *A sequence (x_t) in $\mathbb{R}^N$ is* Cauchy *iff for any $\varepsilon > 0$ there is a T such that for all $s, t > T$,*

$$d(x_s, x_t) < \varepsilon.$$

It is not hard to show that if (x_t) converges to x^* then (x_t) is Cauchy. Conversely, if (x_t) is Cauchy then there is an $x^* \in \mathbb{R}^N$ such that $x_t \to x^*$. This fact, that every Cauchy sequence in $\mathbb{R}^N$ converges to a point in $\mathbb{R}^N$, is called *completeness*. In contrast, the set of rational numbers, $\mathbb{Q}$, is not complete. Consider, for example, the sequence of rational numbers, 3, 3.1, 3.14, 3.141 converging to π. This sequence is Cauchy but since π is not rational, the sequence does not converge to a point in $\mathbb{Q}$.

2.4 Open Sets.

Given a point $x \in \mathbb{R}^N$ and a set $A \subseteq \mathbb{R}^N$, x is an *interior point* of A iff there is an $\varepsilon > 0$ such that $N_\varepsilon(x) \subseteq A$. Thus 1/2 is an interior point of $[0, 1)$ but 0 is not an interior point of $[0, 1)$.

Definition 9. *A set $O \subseteq \mathbb{R}^N$ is* open *iff every $x \in O$ is an interior point of O.*

Thus $[0, 1)$ is not open but $(0, 1)$ is open. A subtle issue is that $\emptyset$ is open (every point in $\emptyset$ is interior since there are *no* points in $\emptyset$). Also all of $\mathbb{R}^N$ is open.

Theorem 2. *A set $O \subseteq \mathbb{R}^N$ is open iff it is either empty or is a union of open balls.*

2.5 Closed Sets.

Definition 10. *A set $C \subseteq \mathbb{R}^N$ is* closed *iff its complement C^c is open.*

Thus $[0, 1)$ is not closed because $[0, 1)^c = (-\infty, 0) \cup [1, \infty)$ is not open (in particular, 1 is not an interior point). But $[0, 1]$ is closed because $[0, 1]^c = (-\infty, 0) \cup (1, \infty)$ is open.

A common but serious mistake is to write that $[0, 1]$ is closed because it is not open (or analogously, that $(0, 1)$ is open because it is not closed). The problem with this is two fold. First, there are sets, such as $[0, 1)$ that are neither open nor closed. You cannot say that such a set is closed because it is not open or vice versa. Second, there are sets that are *both* open and closed. In particular, $\emptyset$ and $\mathbb{R}^N$ are both open and both closed. These are, however, the only subsets of $\mathbb{R}^N$ with this property.

Theorem 3. *A set $C \subseteq \mathbb{R}^N$ is* closed *iff for any sequence (x_t) in C, if there is an $x^* \in \mathbb{R}^N$ such that $x_t \to x^*$, then $x^* \in C$.*

Thus, $[0, 1)$ is not closed because the sequence $(1/2, 2/3, 3/4, \dots)$ is in $[0, 1)$ and converges to 1, but 1 is not in $[0, 1)$.

2.6 Continuity.

Informally, a function $f : \mathbb{R}^N \to \mathbb{R}^M$ is continuous iff for any $x^* \in \mathbb{R}^N$, $f(x)$ is close to $f(x^*)$ whenever x is sufficiently close to x^*. The formal definition is as follows.

Definition 11. *$f : \mathbb{R}^N \to \mathbb{R}^M$ is* continuous *iff for any $x^* \in \mathbb{R}^N$ and any $\varepsilon > 0$ there is a $\delta > 0$ such that if $x \in N_\delta(x^*)$ then $f(x) \in N_\varepsilon(f(x^*))$.*

As with the definition of convergence, it is critical to get the quantifiers correct. If I had written "there exists $\varepsilon > 0$" then the condition becomes too weak. For example, consider $f : \mathbb{R} \to \mathbb{R}$

$$f = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x \leq 0. \end{cases}$$

This function is not continuous at 0. But if I had written "there exists $\varepsilon > 0$" then f would have been been declared continuous since the condition would hold for ε equal to, say, 100.

On the other hand, if I had reversed the order of δ and ε, then I would be requiring that the same δ work for every ε no matter how small. The only function that could pass this continuity requirement is a constant function. And if I had allowed δ to depend on ε but required that the same δ work for every x^* then I would have gotten a condition called *uniform continuity*. One can show that if a function is continuous and its domain is restricted to a closed and bounded interval then it is uniformly continuous. But in general, continuity does not imply uniform continuity. For example, $f(x) = e^x$ is continuous but it is not uniformly continuous on all of $\mathbb{R}$. For any given $\varepsilon > 0$, I'll need a smaller δ the larger is x, since f is steeper the larger is x.

3 Multivariate Calculus.

3.1 Directional Derivatives, Partial Derivatives, and the Jacobian.

Given a function $f : \mathbb{R}^N \to \mathbb{R}^M$ (not assumed to be linear), let $f_m : \mathbb{R}^N \to \mathbb{R}$ be the m coordinate function: $f(x) = (f_1(x), \ldots, f_m(x))$.

Given a point $x^* \in \mathbb{R}^N$ and a vector $v \in \mathbb{R}^N$, $v \neq 0$, the *directional derivative* of f_m in the direction v is given by

$$D_v f_m(x^*) = \lim_{t \to 0} \frac{f_m(x^* + tv) - f_m(x^*)}{t},$$

provided this limit exists. (The notation $\lim_{t\to 0}$ means the limit for every sequence in $\mathbb{R}$ of non-zero elements converging to 0.) This calculation is not much different from the familiar Calc I calculation. In particular, if we define $h : \mathbb{R} \to \mathbb{R}$, $h(t) = f_m(x^* + tv)$, then

$$\frac{dh}{dt}(0) = D_v f_m(x^*).$$

In words, $D_v f_m(x^*)$ gives an infinitessimal approximation to the change in f_m as a result of changing x^* to $x^* + v$. It is natural to denote combine all M directional derivatives, for a given v, into a single matrix:

$$D_v f(x^*) = \begin{bmatrix} D_v f_1(x^*) \\ \vdots \\ D_v f_M(x^*) \end{bmatrix}.$$

If we take the direction v to be a unit vector, say $e^n = (0, \ldots, 0, 1, 0, \ldots 0)$, with the 1 in coordinate n, then we get the coordinate n *partial derivative* of f_m, written

$$D_n f_m(x^*)$$

or, in alternate notation,

$$\frac{\partial f_m}{\partial x_n}(x^*).$$

This calculation is almost exactly like the familiar Calc I calculation, with x_n the variable and all other x_k treated like constants.

It is then natural to form all the partial derivatives into an $M \times N$ matrix (note the dimensions) called the *Jacobian*:

$$Jf(x^*) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x^*) & \cdots & \frac{\partial f_1}{\partial x_N}(x^*) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_M}{\partial x_1}(x^*) & \cdots & \frac{\partial f_M}{\partial x_N}(x^*) \end{bmatrix}.$$

3.2 The Derivative.

If you work an example you will "typically" find that, provided all the required derivatives exist,

$$D_v f(x^*) = Jf(x^*)v.$$

That is, the Jacobian is a machine for computing directional derivatives in any direction for all M coordinate functions. It turns out that this equality always holds if f is differentiable at x^* in the following strong sense.

Definition 12. $f : \mathbb{R}^N \to \mathbb{R}^M$ *is* differentiable *at* $x^* \in \mathbb{R}^N$ *iff there is an* $M \times N$ *matrix* A *such that for* $w \in \mathbb{R}^N$,

$$\lim_{w \to 0} \frac{\|f(x^* + w) - f(x^*) - Ax^*\|}{\|w\|} = 0.$$

The matrix A *is the* derivative *of* f *at* x^*, *also written* $Df(x^*)$. f *is* differentiable *iff it is differentiable at every* x^*.

In practice, no one actually uses this definition to compute $Df(x^*)$. This is because *if* $Df(x^*)$ exists then $Jf(x^*)$ exists and $Df(x^*) = Jf(x^*)$. To see this, I first note that one can easily show, although I will not do so explicitly, that f is differentiable at x^* iff each coordinate function f_m is differentiable at x^*, in which case,

$$Df(x^*) = \begin{bmatrix} Df_1(x^*) \\ \vdots \\ Df_M(x^*) \end{bmatrix},$$

where each $Df_m(x^*)$ is a $1 \times N$ matrix. From the definition of differentiability, taking $w = tv$, one can then show that for each m and any $v \neq 0$,

$$D_v f_m(x^*) = Df_m(x^*)v,$$

which implies

$$D_v f(x^*) = Df(x^*)v.$$

Taking v equal to the unit vectors then implies that for each n and each m, $D_n f_m(x^*) = Df_m(x^*)e^n$, which implies that, indeed

$$Df(x^*) = Jf(x^*).$$

Again, this assumes $Df(x^*)$ exists. The next example illustrates that $Jf(x^*)$ can exist even if $Df(x^*)$ does not.

Example 1. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x) = \begin{cases} \frac{x_1^3}{x_1^2+x_2^2} & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

One can show that $Jf(0,0) = \begin{bmatrix} 1 & 0 \end{bmatrix}$. Let $v = (1,1)$. Then $Jf(0,0)v = 1$. On the other hand, one can compute that $D_v f(0,0) = 1/2$. So $D_v f(0,0) \neq Jf(0,0)v$. Since $D_v f(0,0)$ equals $Jf(0,0)$ if $Df(0,0)$ exists, the lack of equality implies that $Df(0,0)$ does not exist. □

The following theorem gives a sufficient condition for existence of $Df(x^*)$ that is often met, and is easily checked.

Theorem 4. *If all partial derivatives of $f : \mathbb{R}^N \to \mathbb{R}^M$ are continuous, then f is continuously differentiable.*

I won't formalize what "continuously differentiable" means for a multivariate function. In the above example, the partial derivatives are *not* continuous at 0.

3.3 The Chain Rule.

One can prove the following.

Theorem 5 (The Chain Rule)**.** *Let $g : \mathbb{R}^N \to \mathbb{R}^M$, $f : \mathbb{R}^M \to \mathbb{R}^L$, $x^* \in \mathbb{R}^N$, $y^* = g(x^*) \in \mathbb{R}^M$. Define $h = g \circ g$, $h(x) = f(g(x))$. If f and g are differentiable then h is differentiable and*

$$Dh(x^*) = Df(y^*)Dg(x^*).$$

Note that in the Chain Rule, the matrices conform: Dh is $L \times N$, Dg is $L \times M$, and Dg is $M \times N$.

3.4 Real-valued Functions.

If $M = 1$ then $Df(x^*)$ is a $1 \times M$ "row" matrix. The transpose of $Df(x^*)$ is called the *gradient*, written

$$\nabla f(x^*) = [Df(x^*)]' = \begin{bmatrix} \frac{\partial f}{\partial x_1}(x^*) \\ \vdots \\ \frac{\partial f}{\partial x_N}(x^*) \end{bmatrix}.$$

$\nabla f(x^*)$ can be interpreted as a point in $\mathbb{R}^N$. Therefore ∇f can be interpreted as a function from $\mathbb{R}^N$ to $\mathbb{R}^N$. If this function is differentiable, then its derivative is called the *Hessian* of f and is written,

$$D^2 f(x^*) = D\nabla f(x^*) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 x_1}(x^*) & \cdots & \frac{\partial^2 f}{\partial^2 x_1 x_N}(x^*) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial^2 x_N x_1}(x^*) & \cdots & \frac{\partial^2 f}{\partial x_N x_N}(x^*) \end{bmatrix}.$$

Theorem 6. *If all second-order partial derivatives of $f : \mathbb{R}^N \to \mathbb{R}^N$ exist and are continuous then the Hessian is symmetric.*

3.5 The Inverse Function Theorem

Recall that if $f : \mathbb{R}^N \to \mathbb{R}^N$ is linear then a necessary and sufficient condition for f to be invertible is that it have full rank, namely N. The Inverse Function Theorem says that something similar is true for *non-linear* (but continuously differentiable) functions.

Theorem 7 (Inverse Function Theorem)**.** *If $f : \mathbb{R}^N \to \mathbb{R}^N$ is continuously differentiable and if $Df(x^*)$ has full rank then f has a continuously differentiable local inverse. Explicitly, there is an open ball U around x^*, an open ball V around $y^* = f(x^*)$, and a continuously differentiable function $f^{-1} : V \to U$ such that the following hold.*

1. *$f(U) \subseteq V$.*
2. *For any $x \in U$, $f^{-1}(f(x)) = x$.*
3. *For any $y \in V$, $f(f^{-1}(y)) = y$.*

To take a trivial example, consider $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$. Then there is no inverse at $x = 0$ and indeed $Df(0) = 0$, which does not have full rank (recall that a matrix containing the origin is linearly dependent and hence has rank 0). But at $x = 1$, $Df(1) = 2$, which does have full rank, and indeed f is invertible near $x = 1$: $f^{-1}(y) = \sqrt{y}$. As this example illustrates, one difficult with dealing with non-linear functions is that we have to be content with only local invertibility.

Since

$$f^{-1}(f(x)) = x,$$

for any $x \in U$, the Chain Rule implies that, setting $y = f(x)$,

$$Df^{-1}(y)Df(x) = I,$$

where I is the $N \times N$ identity matrix. Hence

$$Df^{-1}(y) = [Df(x)]^{-1}.$$

In words, the derivative of the inverse is the inverse of the derivative.

If $Df(x^*)$ is *not* of full rank then f is *singular* at x^*. It is possible for f to be invertible at x^* even if f is singular at x^*. The function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = x^3$ provides an example. In particular, this f is singular at $x^* = 0$ but it is invertible; the inverse is $f^{-1}(y) = (y)^{1/3}$. But note that the derivative of f^{-1} is not defined at $y = 0 = f(0)$. This is a general fact: if a function is singular at a point x^*, its inverse *may* still exist, but even if the inverse exists, it will not be differentiable at $f(x^*)$.

3.6 The Implicit Function Theorem

Recall the discussion in Section 1.6 (I will use somewhat different notation here; sorry for that). Let $F : \mathbb{R}^{N+M} \to \mathbb{R}^M$ be linear. Write $x \in \mathbb{R}^{N+M}$ in the form $x = (p, q)$, where $p \in \mathbb{R}^N$ and $q \in \mathbb{R}^M$. Since F is linear, there is an $M \times N$ matrix A and an $M \times M$ matrix B such that

$$F(x) = F(p, q) = Ap + Bq.$$

If B, which is square, has full rank, then there is a linear function $f : \mathbb{R}^N \to \mathbb{R}^M$ given by

$$f(p) = -B^{-1}Ap$$

with the property that for all $p \in \mathbb{R}^N$,

$$F(p, f(p)) = Ap + B(-B^{-1}Ap) = 0.$$

Put differently, f expresses the kernel of the matrix $\begin{bmatrix} A & B \end{bmatrix}$ as the graph of a linear function of p. f is "implicitly" defined by the expression $F(p, f(p)) = 0$.

The Implicit Function Theorem Generalizes this observation to non-linear functions. In the theorem statement, $D_M F(x^*)$ refers to the $M \times M$ submatrix of $DF(x^*)$ formed by the last M columns of $DF(x^*)$.

Theorem 8 (Implicit Function Theorem)**.** *Let $F : \mathbb{R}^{N+M} \to \mathbb{R}^M$ be continuously differentiable and let $x^* = (p^*, q^*)$ be such that $F(x^*) = 0$. If $D_M F(x^*)$ has full rank then there is an open ball U around p^* and a continuously differentiable function $f : U \to \mathbb{R}^M$ such that following hold.*

1. *$f(p^*) = q^*$.*
2. *For all $p \in U$,*
$$F(p, f(p)) = 0.$$

In words, the Implicit Function Theorem says that, subject to some technical conditions, the "kernel" of $F : \mathbb{R}^{N+M} \to \mathbb{R}^M$ (for non-linear F, the "kernel" is called the *zero set*) can be locally described as the graph of a continuously differentiable function f defined on an open ball in $\mathbb{R}^N$. Any such graph is called a *differentiable*

manifold. It is a non-linear analog of a vector space. Thus the Implicit Function Theorem says that, subject to some technical conditions, the zero set of a function is, locally, a differentiable manifold.

Since

$$F(p, f(p)) = 0,$$

for any $p \in U$, the Chain Rule implies that, for $x = (p, f(p))$ and writing $D_N f(x)$ for the $M \times N$ submatrix formed by the first N columns of $DF(x)$,

$$D_N F(x) + D_M F(x) Df(p) = 0,$$

or

$$Df(p) = -[D_M F(x)]^{-1} D_N F(x).$$

Note that this is equivalent to the expression we found in the linear case. If you can remember the linear case, then you can remember the Implicit Function Theorem.

Much as with the Inverse Function Theorem, the qualification that f is defined only locally can be essential. A standard example is the circle: $F : \mathbb{R}^2 \to \mathbb{R}$, $F(p, q) = p^2 + q^2 - 1$. The zero set of F is the unit circle. At $(p^*, q^*) = (0, 1)$, for example, the derivative with respect to q is 2, so the full rank condition holds, and indeed we can write,

$$q = f(p) = \sqrt{1 - p^2}.$$

On the other hand, the full rank condition fails at the point (1,0): it is not possible to represent the unit circle near (1,0) as the graph of a function of p.

It *is*, however, possible to represent the circle near (1,0) as the graph of a function of q:

$$p = f(q) = \sqrt{1 - q^2}.$$

This brings up a larger point. The stated version of the Implicit Function Theorem is unnecessarily limited. A more general version says that as long as $DF(x^*)$ has full rank, namely M, then it is possible to represent the zero set of the function, locally, as the graph of function that gives M of the variables as a function of the other N variables.

4 Optimization

4.1 Unconstrained Optimization.

Let $f : \mathbb{R}^N \to \mathbb{R}$ be twice continuously differentiable (so that the Hessian exists and is symmetric). If x^* solves

$$\max_{x \in \mathbb{R}^N} f(x)$$

then, in particular, it must be that x^* is optimal in each coordinate. That is, for each unit vector e^n,

$$f(x^* + te^n) \geq f(x^*)$$

for any $t \in \mathbb{R}$. If you manipulate this a bit you will see that this implies that for each n,

$$\frac{\partial f}{\partial x_n}(x^*) = 0.$$

This implies that $Df(x^*) = [\ 0 \cdots 0\ ]$ and hence that for any direction $v \neq 0$, $D_v f(x^*) = Df(x^*)v = 0$. Informally, at the top of the hill, the hill is flat.

An x^* such that $Df(x^*) = [\ 0 \cdots 0\ ]$, or, equivalently, such that $\nabla f(x^*) = 0$, is called a *critical point*. In summary, a necessary condition for x^* to be a solution for an unconstrained maximization problem is for x^* to be a critical point. This is often called the *first order condition*.

The requirement that x^* be critical is not sufficient. For $f(x) = x^2$, $x^* = 0$ is critical but x^* is a minimum, not a maximum. For $f(x) = x^3$, x^* is critical but it is an inflection point: neither a minimum nor a maximum.

A sufficient condition for a critical point x^* to be a maximum is for the function to be *concave*. If $N = 1$, one can establish that f is strictly concave if (but not only if) $D^2 f(x) < 0$ for all x. If $N > 1$, one can similarly establish that f is strictly concave if the second derivative is negative in every direction v. Formally, given $x \in \mathbb{R}^N$ and given $v \neq 0$, define $h : \mathbb{R} \to \mathbb{R}$ by $h(t) = f(x + tv)$. Then the condition is that $h''(0) < 0$. One can compute via the Chain Rule that,

$$h''(0) = v' D^2 f(x) v,$$

where $D^2 f(x)$ is the Hessian of f at x. Therefore, the condition is that f is strictly concave if for every $v \neq 0$,

$$v' D^2 f(x) v < 0.$$

A matrix that has this property is called *negative definite*. Thus a sufficient condition for f to be strictly concave is for $Df^2(x)$ to be negative definite for all x. This is called the *second order condition*.

If instead we are working with a minimization problem then it is again necessary that a solution be a critical point. A sufficient condition for a critical point to be a solution is that f be *convex*. A sufficient condition for f to be strictly convex is for $D^2 f(x)$ to be *positive definite* for all x: for all $x \in \mathbb{R}^N$ and all $v \neq 0$,

$$v' D^2 f(x) v > 0.$$

4.2 Constrained Maximization.

Suppose that instead there are constraints on what x can be. One gets a problem in the form

$$\begin{array}{cc} \max_{x \in \mathbb{R}^N} & f(x) \\ \text{s.t.} & g_1(x) = 0 \\ & \vdots \\ & g_K(x) = 0. \end{array}$$

To take an economics example, $g_1(x) = 0$ might be the budget constant that $p \cdot x = m$, where $p \in \mathbb{R}^N$ is the vector of prices and $m \in \mathbb{R}$ is "money" (wealth). This constraint can be rewritten in the above form: $p \cdot x - m = 0$. (In economics, constraints are typically inequality constraints. Thus, for example, the budget constraint is really $p \cdot x \leq m$: you *can* spend less than wealth, but it will not be optimal to do so. There is an easy extension of this machinery to handle inequality constraints.)

Subject to a technical condition that I won't go into, one can show that a necessary condition for x^* to be a solution to this problem is that there are numbers $\lambda_1, \ldots, \lambda_K$, called *Lagrange multipliers*, such that

$$\nabla f(x^*) = \lambda_1 \nabla g_1(x^*) + \cdots + \lambda_K \nabla g_K(x^*).$$

The intuition is that the reason why $\nabla f(x^*) \neq 0$ at a solution is that one or more constraints are in the way. The $\nabla g_k(x^*)$ express the local behavior of the constraints. So you should expect to have the first order condition combine $\nabla f(x^*)$ and the $\nabla g_k(x^*)$. I can give good intuition for why the condition takes this particular form, but it would take me too far afield.