A chain rule:

let $f : \Omega \to R$ be a C^1 function. Here $\Omega \subset R^2$ is an open set. Suppose φ_1 and φ_2 are two real C^1 functions defined on an interval (a, b) such that for every t, the point $(\varphi_1(t), \varphi_2(t)) \in \Omega$. Then it makes sense to define the composed function.

$$F(t) = f(\varphi_1(t), \varphi_2(t)) : (a, b) \to R.$$

Thus the function is a real valued function defined on an interval. Thus given a real number t, it produces a real number $F(t)$. However, to calculate the number $F(t)$ we pass through R^2. It is natural to expect that this is again a differentiable function. What is the formula for the derivative at a point?

Theorem (Chain rule): F is C^1 and

$$F'(t) = f_1(\varphi_1(t), \varphi_2(t))\ \varphi_1'(t) + f_2(\varphi_1(t), \varphi_2(t))\ \varphi_2'(t).$$

The result can be restated in several ways. For example

$$F'(t) = \nabla f \cdot (\varphi_1', \varphi_2').$$

where ∇f is evaluated at the point $(\varphi_1(t), \varphi_2(t))$ and the derivatives φ_1' and φ_2 are evaluated at the point t.

If we denote the map $\Phi(t) : (a, b) \to R^2$ by $\Phi(t) = (\varphi_1(t), \varphi_2(t))$ and make the convention $\Phi'(t) = (\varphi_1'(t), \varphi_2'(t))$ then the formula takes the pleasing form,

$$F'(t) = \nabla f(\Phi(t)) \cdot \Phi'(t).$$

The reason it is pleasing is that it is the same formula we learnt in one dimension. There is nothing new really, is it not?

Here is another way of stating, which uses a different suggestive notation. The maps φ_1 and φ_2 are denoted by $x(t)$ and $y(t)$ respectively. This is because they denote the x and y coordinates when we proceed to compose with f. Points in R^2 are denoted as (x, y) and the function f is $f(x, y)$. With this notation,

$$\frac{dF}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}.$$

We have used ∂f to remind that f is a fucntion of more than one variable and we have differentiated w.r.t. one of those variables. On the other hand we used dF, dx, dy because these are functions of only one variable and derivative is taken w.r.t. that variable.

Proof is not difficult. But let us see why we are interested in this.

Consider a function f defined on a region $\Omega \subset R^2$ and a point $a \in \Omega$. let us fix a unit vector $u \in R^2$. We want to see if the limit

$$\lim_{t \to 0} \frac{f(a+tu) - f(a)}{t}$$

exists and equals $\nabla f \cdot u$. Remember the points $a + tu$ are on the line in the direction of u at the point a and this is the derivative at the point a in the direction u.

This is an immediate consequence of the chain rule. Take Φ to be the function $\Phi(t) = a + tu$ defined on the interval $(-1, 1)$. Do not forget to notice that a, u are vectors but t is a real number. The differentiability of the composed function at $t = 0$ is precisely the limit required above and chain rule says that it exists and equals the claimed quantity. You need to only note that $\Phi'(t) = u$.

The earlier observation combined with Cauchy-Schwarz inequality leads to an interesting interpretation of the gradient: it gives the direction in which the derivative is the largest in modulus, assuming that the gradient is non-zero vector. This is trivial because

$$|\nabla f \cdot u| \leq ||\nabla f||\ ||u||.$$

equality above holds when and only when u is a multiple of ∇f, assuming that ∇f is non-zero. If ∇f is zero vector then for any u the quantity above is zero. Since u is a unit vector, it should be the vector: normalised ∇f.

The chain rule has another interesting consequence, the mean value theorem. Suppose $\Omega \subset R^2$ open $a, b \in \Omega$ and the line joining the points a and b is contained in Ω. Then there is some point θ on this line such that

$$f(b) - f(a) = \nabla f(\theta) \cdot (b - a). \qquad (\dagger)$$

This is again easy. Consider the composition as above with

$$\phi(t) = tb + (1-t)a.$$

Then $F(1) = b$ and $F(0) = a$ and it is a continuous function on the interval $[0, 1]$ which is differentiable at every point inside the interval. Hence usual mean value theorem applies to give a number $\eta \in (0, 1)$ such that $F(1) - F(0) = F'(\eta)$. apply chain rule to see this is same as (†) with $\theta = \Phi(\eta)$.

The chain rule also leads to Taylor expansion but since it needs higher derivatives we return to this later. Let us now prove the chain rule. Fix a t_0. Need to show

$$F'(t_0) = \nabla f(\Phi(t_0)) \cdot \Phi'(t).$$

That is,

$$\frac{f(\Phi(t)) - f(\Phi(t_0))}{t - t_0} - \nabla f(\Phi(t_0)) \cdot \Phi'(t_0) \to 0.; \quad \text{as} \quad t \to t_0.$$

Let $\epsilon > 0$ be any fixed number. We show $\delta > 0$ such that

$$|t - t_0| < \delta \Rightarrow \left| \frac{f(\Phi(t)) - f(\Phi(t_0))}{t - t_0} - \nabla f(\Phi(t_0)) \cdot \Phi'(t_0) \right| < \epsilon. \quad (\dagger)$$

We reduce it two separate simple problems. We show $\delta_1 > 0$ such that

$$|t - t_0| < \delta_1 \Rightarrow \left| \frac{f(\Phi(t)) - f(\Phi(t_0))}{t - t_0} - \nabla f(\Phi(t_0)) \cdot \left[\frac{\Phi(t) - \Phi(t_0)}{t - t_0} \right] \right| < \epsilon/2. \qquad (\spadesuit)$$

We show a $\delta_2 > 0$ such that

$$|t - t_0| < \delta_2 \Rightarrow \left| \nabla f(\Phi(t_0)) \cdot \left[\frac{\Phi(t) - \Phi(t_0)}{t - t_0} \right] - \nabla f(\Phi(t_0)) \cdot \Phi'(t_0)] \right| < \epsilon/2. \qquad (\clubsuit).$$

Then for $|t - t_0| < \min\{\delta_1, \delta_2\}$ both inequalities hold; adding them we get (†) as required.

($\clubsuit$) is simple. In fact the long expression, by Cauchy-Schwarz inequality, is at most

$$||\nabla f(\Phi(t_0))|| \quad \left\| \frac{\Phi(t) - \Phi(t_0)}{t - t_0} - \Phi'(t_0)] \right\|$$

Denote $1 + ||\nabla f(\Phi(t_0))|| = c$. By using definition of derivative (for usual one variable functions), choose $\delta_2 > 0$ so that for $|t - t_0| < \delta_2$

$$\left| \frac{\varphi_1(t) - \varphi_1(t_0)}{t - t_0} - \varphi_1'(t_0)] \right| < \frac{\epsilon}{2c}; \quad \left| \frac{\varphi_2(t) - \varphi_2(t_0)}{t - t_0} - \varphi_2'(t_0)] \right| < \frac{\epsilon}{2c}.$$

(I have used mean value theorem here while explaining in the class and you suggested definition of derivative is enough.) Of course, the choice of δ_2

implies that $(t_0 - \delta_2, t_0 + \delta_2) \subset (a, b)$. if you are not able to see, just you can as well take smaller value of δ_2 that satisfies this condition.

Then definition of norm and definition of Φ' will give you (♣).

To achieve (♠) you do exactly similar thing. Denote $\Phi(t_0) = a$. using definition of derivative (that we have learnt now for functions of two variables), first choose η so that for $||x - a|| < \eta$ we have

$$\left| \frac{f(x) - f(a) - \nabla f(a) \cdot (x - a)}{||x - a||} \right| < \epsilon/2. \qquad (*)$$

Just to remind you, here the dot in the neumerator is scalar product. By using continuity of Φ, that is, continuity of φ_1 and φ_2; get $\delta_1 > 0$ so that when $|t - t_0| < \delta_1$ then $||\Phi(t) - \Phi(a)|| < \eta$.

Let us see what happens to the right side expression of (♠) when $|t - t_0| < \delta_1$. Fix such a t. In case $\Phi(t) = \Phi(t_0)$ then that expression is zero and nothing for us to do to see the required inequality. Other wise denoting $\Phi(t) = x$, that expression equals

$$\left| \frac{f(x) - f(a) - \nabla f(a) \cdot (x - a)}{||x - a||} \right| \times \frac{||x - a||}{|t - t_0|.}$$

Here the first quantity is smaller than $\epsilon/2$ by $(*)$. What about the second term?

By MVT, of the last semester, applied to the functions φ_1 and φ_2, there are points P_1 and P_2 in the interval $(t_0 - \delta_1, t_0 + \delta_1)$ so that the second term in the above display in nothing but norm of the vector $\langle \varphi_1'(P_1), \varphi_2'(P_2) \rangle$. If $c/2$ is a bound for these derivatives of φ_1, φ_2 over this interval then this norm is smaller than c. Thus the he second term is smaller that c and hence the above expression is smaller than $c\epsilon/2$.

So it appears that a better choice of δ_1 would do, namely, choose your η so that for $||x - a|| < \eta$ we have

$$\left| \frac{f(x) - f(a) - \nabla f(a) \cdot (x - a)}{||x - a||} \right| < \epsilon/2c, \qquad (**)$$

instead of $(*)$ and then chose δ_1 for this η.

Yes, if you now go back and choose δ_1 for this η the proof seems to work perfectly. But this argument is faulty because c depended on δ_1 (see where

we got into this c) and δ_1 now depends on c. So actually nothing is achieved. A wise thinking, that involves looking ahead before you take your step, would help. Here is the precise argument to show that a δ_1 can be chosen to satisfy (♠).

First choose $\delta' > 0$ so that

$$[t_0 - \delta', t_0 + \delta'] \subset (a, b).$$

Since φ_1' and φ_2' are continuous functions, they are bounded on this interval and let $c/2$ be a bound for these functions. Choose η so that $(**)$ holds. Choose δ_1 exactly as earlier for this η. If necessary, make it smaller so that $\delta_1 < \delta'$. This choice would do to show (♠).

I have given the thought process in choosing δ_1 as part of the proof. At the end, if you are confused, ignore the thought process and the faulty argument we went through. Go to the para where we choose η earlier, replace it by the above para and then proceed with the argument. You must convince yourself that (♠) is achieved. You must also convince ourself that the proof is actually very simple and this was precisely what was done last semester too for the chin rule; a two step procedure.

Thus we have completed proof of the chain rule. You must understand what we achieved. We have a given formula to calculate the derivative the function F from R to R. Then did we not do such things already last semester? No. Here to get the value of the function you pass through R^2.

You can generalise to passing through R^n. This means, you have C^1 functions $\varphi_1, \cdots, \varphi_n$ on an interval (a, b) and you have a C^1 function $f : \Omega \to R$. Here $\Omega \subset R^n$ is an open set and it includes the vector $(\varphi_1(t), \cdots, \varphi_n(t))$ for very $t \in (a, b)$. Then it makes sense to define $F : R \to R$ by

$$F(t) = f(\varphi_1(t), \cdots, \varphi_n(t)); \quad a < t < b.$$

Then F is C^1 and

$$F'(t) = \sum f_i(\varphi_1(t), \cdots, \varphi_n(t))\varphi_i'(t) = \sum \frac{\partial f}{\partial x_i}(\varphi_1(t), \cdots, \varphi_n(t))\varphi_i'(t).$$

$$= \nabla f(\Phi(t)) \cdot \Phi'(t) = f'(\Phi(t)) \cdot \Phi'(t).$$

where

$$\Phi(t) = (\varphi_1(t), \cdots, \varphi_n(t)); \quad \Phi'(t) = (\varphi_1'(t), \cdots, \varphi_n'(t)).$$

Several new problems arise. Suppose φ_1 and φ_2 are C^1 functions defined on Ω_1 taking values in R. The only difference is that they are not defined on an interval contained in R. Then you can think of $\Phi(x) = (\varphi_1(x), \varphi_2(x))$ on as a function on Ω_1 with values in R^2, as earlier. Suppose its values fall in the open set $\Omega \subset R^2$ on which we have real valued C^1 function f. Then it makes sense to talk about the composition:

$$F(x) = f(\Phi(x)); \qquad x \in \Omega_1.$$

is this differentiable? Is

$$F'(x) = f'(\Phi(x)) \cdot \Phi'(x).$$

It should be correct, the only problem is that we do not know the meaning of $\Phi'(t)$ because Φ is a function on R^2 to R^2. We need to assign meaning to derivative of function defined on R^m and taking values in R^n.

Let us pause for a moment and see what happened so far. First we had functions from R to R and we defined derivative at a point and it is a number. Then we had function from R^n to R and we defined its derivative at a point and it is a vector. In the previous theorem we had Φ from R to R^n and we defined Φ'. This was facilitated by the fact that such a function Φ is made up of n real valued functions, namely, $\varphi_i(t)$ equals the i-th coordinate of $\Phi(t)$. It turned out that $\Phi'(t)$ is also a vector.

There is again some chaos, sometimes we have functions $R \mapsto R$, sometimes $R^2 \mapsto R$ and sometimes $R \mapsto R^2$. Now we have functions $R^2 \mapsto R^2$. Sometimes derivatives are numbers, sometimes they are vectors. What are they? Some order has to be brought in and a clear understanding has to be achieved.

Let us go back and see what was the purpose of derivative. We wanted to make best linear approximation of f at a point a in its domain. That is, we wanted $g(x) = L(x)+\beta$ such that $g(a) = f(a)$ and $||f(x)-g(x)||/||x-a|| \to 0$ as $x \to a$. Thus $f(x) - g(x)$ approaches zero faster than x approaches a. Of course the only non-trivial thing is $L(x)$ because $\beta = f(a) - L(a)$ since we want (or know) $\varphi(a) = f(a)$.

(**a**)

If $f : R \to R$, then the derivative at a denoted by the number c has the property that it determines g. More precisely the linear transformation from

R to R is the map $L(x) = cx$.

(**b**)

If $f : R^2 \to R$, then the derivative at $a \in R^2$ denoted by the vector $c = \nabla f(a)$ has the property that it determines g. More precisely the linear transformation from R^2 to R is the map $L(x) = c \cdot x$. Here is another way of stating the same thing. Even though everything, points in R^2 as well as derivatives are vectors, let us give them proper dress so that we can recognize them easily. Think of R^2 as space of column vectors. That is

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

Let us think of the derivative c as row vector. Then the linear transformation we are talking about is nothing but $L(x) = cx$. This makes perfect sense because c is 1×2 vector and x is 2×1 vector and so cx is 1×1 vector, or a number. Thus the derivative $\nabla f(a)$ is actually the row vector. But it is to be thought of as a linear transformation of R^2 to R.

(**c**)

If $f : R \to R^2$, then the derivative at $a \in R$ denoted by the vector c has the property that it determines g. Remember that if $f_1(t)$ is the first coordinate of $f(t)$ and $f_2(t)$ is the second coordinate of $f(t)$ then $f(t) = (f_1(t), f_2(t))$. But just now we decided to think of R^2 as column vectors. Thus

$$f(t) = \begin{pmatrix} f_1(t) \\ f_2(t) \end{pmatrix}$$

More precisely the linear transformation from R to R^2 is the map $L(x) = cx$, that is,

$$x \in R^1 \mapsto cx = \begin{pmatrix} f_1'(a) \\ f_2'(a) \end{pmatrix} x \in R^2.$$

In other words the derivative of f is the column vector

$$f'(a) = \begin{pmatrix} f_1'(a) \\ f_2'(a) \end{pmatrix}.$$

again to avoid confusion regarding column and row vectors, let us think of f as the linear transformation $L(x) = cx$. This makes sense c is 2×1 vector and x is 1×1 and cx makes sense and is 2×1 vector, in other words, an element of R^2. This is the linear transformation from R to R^2.

Before proceeding further, we should note one thing here. We have not defined derivative of $f : R \to R^2$ via best linear approximations. We defined outright $f'(a) = \langle f_1'(a), f_2'(a) \rangle$. But one can easily show that

$$\frac{||f(t) - f(a) - f'(a)(t-a)||}{|t-a|} \to 0 \quad \text{as} \quad t \to a.$$

This only depends on the fact that a sequence of vectors converges to zero iff coordinate-wise it so happens.

This brings some order into things. Derivatives are linear operators, no need to confuse whether it is row vector or column vector or a number and so on. Elements of R^n are column vectors: x and so row vectors r define linear maps on them to $R : x \mapsto rx$. again in the notation there is nothing to tell you whether a symbol is a row vector or column vector.

Taking clue from the above, suppose we have a map $f : R^m \to R^n$ and $a \in R^m$. The best linear approximation of f at a determines the derivative. We say that a linear transformation $L(x) : R^m \to R^n$ is derivative of f at a if the map $\varphi(x) = L(x) + f(a) - L(a)$ has the property

$$\frac{||f(x) - \varphi(x)||}{||x-a||} \to 0 \quad \text{as} \quad x \to a.$$

In other words

$$\frac{||f(x) - f(a) - L(x-a)||}{||x-a||} \to 0 \quad \text{as } ||x-a|| \to 0.$$

Since linear transformations from R^m to R^n are given by $n \times m$ matrices A via $L(x) = Ax$ we can reformulate the idea of derivative as follows. A $n \times m$ matrix A is derivative at the point a if

$$\frac{||f(x) - f(a) - A(x-a)||}{||x-a||} \to 0 \quad \text{as } ||x-a|| \to 0.$$

Of course there is another way to define the derivative taking a clue from the adhoc definition we employed earlier. remember if $f : R \to R^2$ is a C^1 map we defined $f'(a) = (f_1'(a), f_2'(a))$. If

$$f = \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix}$$

then we can put

$$f'(a) = \begin{pmatrix} \nabla f_1(a) \\ \nabla f_2(a) \\ \vdots \\ \nabla f_n(a) \end{pmatrix}$$

Note that it has n rows each an m-vector. Thus it is $n \times m$ matrix.

We shall, as earlier, restrict our attention to only C^1 functions and proceed. But first we need to familiarize ourselves with maps from R^m to R^n. So you can forget about derivatives for a while.

functions from R^m to R^n:

Elements of R^n are column vectors v. But it takes space to write column vectors. One way is to think of your symbols v as row vectors and elements of R^n as their transpose: v^t. Again it is a burden in reading and so we shall not do this either. We just do as we have been doing all along. Do not scratch unless it itches. When this distinction is specifically needed, then we shall use it. Otherwise we enjoy being careless, but remember elements of R^n are column vectors.

Let $\Omega \subset R^m$. Suppose we have n real valued functions on Ω. Then we can cook up a R^n valued function on Ω as

$$f(x) = (f_1(x), f_2(x), \cdots, f_n(x)) \quad x \in \Omega.$$

Conversely, every function on Ω taking values in R^n is obtained like this. More precisely, suppose $f : \Omega \to R^n$ is given to us, then there are n uniquely determined real functions $f_1, f_2, \cdots, f_n$ on Ω such that

$$f(x) = (f_1(x), f_2(x), \cdots, f_n(x)).$$

This is obvious, $f_i(x)$ must be the i-th coordinate of the point $f(x)$ and this choice does it.

Let us say that f is continuous if each f_i is so. remember we defined continuity for real valued functions. Of course, the natural way to define is to say that $f(x)$ should be close to $f(a)$ if x is close to a. Yes, both ideas are same mathematically.

Theorem: let $f = (f_1, f_2, \cdots, f_n) : \Omega \to R^n$. Let $a \in \Omega$. Then the following are equivalent.

(i) Each of the n real valued functions $f_1, f_2, \cdots, f_n$ are continuous on Ω.
(ii) Given $\epsilon > 0$, there is a $\delta > 0$ such that

$$x \in \Omega, ||x - a|| < \delta \Rightarrow ||f(x) - f(a)|| < \epsilon.$$

(iii) If $\{x^i : i \geq 1\}$ is a sequence of points in Ω and $x^i \to a$ then $f(x^i) \to f(a)$.

We have proved it in class. Try to do so without writing.

The ideas you learnt last semester are powerful to give you theorems about continuous functions on R^m also. Here is one such. A subset $S \subset R^m$ is bounded if there is a number c such that $||x|| \leq c$ for all $x \in S$.

Theorem: Let $S \subset R^m$ be a closed bounded set and f be a continuous function $f : S \to R$. Then f is bounded, that is, there is a number M such that $|f(x)| \leq M$ for all $x \in S$. This is also same as saying that the range of the function f is a bounded subset of R.

There is no new idea. Let us execute it for $S \subset R^2$. Since the set S is bounded get a square $R_0 = [a, b] \times [a, b]$ which includes S. You only need to note that if $||x|| \leq c$ then $x \in [-c, c] \times [-c, c]$. We prove the theorem by contradiction.

suppose that the function is not bounded. Divide the square into four parts by cutting each side at the mid point. Then f must be unbounded on part of S contained in one of these smaller squares. Take one such square, $R_1 = [a_1, b_1] \times [c_1, d_1]$. Just be careful, do not be under the impression that this square is like $[c, d] \times [c, d]$ just because the earlier square was $[a, b] \times [a, b]$. That was in your hands, this is not. Do the same to R_1 and get R_2 and so on.

Thus you get a sequence of squares R_n such that length of each side of R_n is half length of previous square side. By cantor intersection theorem, we get a point (a_1, a_2) common to all these squares. Indeed all the sides $[a_n, b_n]$ have a point x in common and all sides $[c_n, d_n]$ have a point y in common. Since each square R_n contains points of S, we see that (x, y) is a limit point of S and must be in S because S is closed.

But then continuity of f tells that there is a $\delta > 0$ such that $x \in S, ||x - a|| < \delta$ implies $||f(x) - f(a)|| < 1$, thus $||f(x)|| \leq ||f(a)) + 1$. In particular, f is bounded by this number at all points of S in this disc. Now one of your R_n must be contained in this disc, because their lengths are

converging to zero. This contradicts the fact that f is not bounded on the part of S contained in R_n.

This completes the proof.

Let us now consider $S \subset R^m$ and $f : S \to R^n$. Say that f is bounded if there is number M such that $||f(x)|| \leq M$ for all $x \in S$. If $F = (f_1, f_2, \cdots, f_n)$ then f is bounded iff each f_n is so. This is because, if f is bounded by M then each f_i is also bounded by M and if each f_i is bounded by M then f is bounded by $M\sqrt{n}$.

Say that $f : R^m \to R^n$ is a C^1 function if each f_i is so, Note that f_i is C^1 map means that each of its m partial derivatives are continuous functions.

let $\Omega \subset R^m$ be an open set and $F : \Omega \to R^n$ be C^1 map. Let $a \in \Omega$ and let A be the $n \times m$ matrix whose i-th row consists of $\nabla f_i(a)$. In other words, (i, j)-th entry of A is $D_j f_i(a)$; the partial derivative of the i-th function f_i w.r.t. the j-th coordinate.

Theorem:

$$\frac{||f(x) - f(a) - A(x-a)||}{||x-a||} \to 0 \quad \text{as} \quad x \to a. \qquad (\diamondsuit)$$

In other words, A confirms to our intuition as a suitable candidate for being the derivative of f at A. Indeed we *define* A to be the derivative of f at a. Of course, what we mean is that the linear map $x \mapsto Ax$ of R^m to R^n is the derivative at the point a. Of course, you can also think of the matrix A itself as the derivative. Then you need to keep track of the confusion that some times derivatives are numbers, sometimes they are vectors and yet other times they are matrices.

proof of the above theorem is simple, no work is needed. Note that the quantity in $(\diamondsuit)$ is simply norm of a vector. Carefully decipher the notation, to see that the i-th entry of the above vector is nothing but

$$\frac{f_i(x) - f_i(a) - \nabla f_i(a) \cdot (x-a)}{||x-a||}$$

and by definition of ∇f_i this quantity does converge to zero as $||x-a|| \to 0$. Hence $(\diamondsuit)$ is verified.

Can there be another matrix A_1 satisfying $(\diamondsuit)$? If so, fix $r > 0$ so that $B(a,r) \subset \Omega$. Then for $h \in B(0,r)$

$$\frac{||Ah - A_1 h||}{||h||} \leq \frac{||f(a+h) - f(a) - Ah||}{||h||} + \frac{||f(a+h) - f(a) - A_1 h||}{||h||}$$

$$\longrightarrow 0.$$

If you take any non-zero vector $v \in R^m$ then for all large integers i, we see $v/i \in B(0,r)$ so that

$$\frac{||A(v/i) - A_1(v/i)||}{||v/i||} \to 0, \quad \text{as} \quad i \to \infty.$$

In other words $Av - A_1 v = 0$. This is true for every vector $v \in R^m$ and hence the two matrices/linear transformations are same.

Thus our definition of derivative is a good definition. Of, course you can say that an f is differentiable if $(\diamondsuit)$ holds for some linear transformation A; without assuming that f is C^1 to begin with. Then also you can show that such a transformation is unique, by the same argument s above. However, the derivative may exist but the function f may not be C^1. We have already seen such examples earlier. It is to avoid some pathologies we started restricting to C^1 maps.

We denote the derivative of f at a by $Df(a)$ or $(D_j f_i(a))_{ij}$ or the more compact notation $f'(a)$.

examples:

1. Fix a vector $u \in R^n$ and consider the map $f(x) = u$ for all $x \in R^m$. This is a C^1 map and $f'(a) \equiv 0$. That is, it is the zero linear transformation or it is the matrix with all entries zero.

2. Fix one $n \times n$ matrix A and consider $f(x) = Ax + u$. Then for every a, we have $f'(a) = A$.

3. $f(x,y) = x + y$ from R^2 to R. Then $f'(a) = (1,1)$.

   More generally, if we take the map from R^n to R given by $f(x) = \sum x_i$, then it is C^1 and $f'(a)$ is the vector with all entries equal to one.

   Or if you take the map R^5 to R^2 given by

   $$f(x_1, x_2, \cdots, x_5) = (x_1 + x_2 + x_3.x_4 + x_5).$$

then it is C^1 and

$$f'(a) = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}.$$

Or define f from R^{2n} to R^n by $f(x,y) = x + y$. Note that here we are denoting point of R^{2n} by $(x_1, \cdots, x_n, y_1, \cdots, y_n)$. then $f'(a,b) = (\mathbf{I}, \mathbf{I})$where I is the $n \times n$ identity matrix. thus the matrix $(\mathbf{I}, \mathbf{I})$ has n rows and $2n$ columns.

4. Define from R^2 to R; $f(x,y) = x.y$. Then

$$f'(a,b) = (b,a).$$

For example $f'(3,4)$ is the linear transformation given by $(4,3)$; in other words $L(x,y) = 4x + 3y$.

Consider $f(x,y) = \sum_1^n x_i y_i$ from R^{2n} to R^n. Again see how we denoted points of R^{2n}, not as $(x_i : 1 \leq i \leq 2n)$ but as $(x_1, \cdots, x_n, y_1, \cdots, y_n)$. Then

$$f'(a_1, \cdots, a_n, b_1, \cdots, b_n) = (b_1, \cdots, b_n, a_1, \cdots, a_n).$$

5. Let us consider a symmetric 2×2 matrix A and consider $f(x) = x^t A x$ a function from R^2 to R. Here x^t is the transpose of the column vector x. In other words with usual notation of (x,y) for points of R^2

$$f(x,y) = \alpha x^2 + 2\beta xy + \gamma y^2 \qquad A = \begin{pmatrix} \alpha & \beta \\ \beta & \gamma \end{pmatrix}$$

Then $f'(a,b) = (2\alpha a + 2\beta b, 2\beta a + 2\gamma b)$. Reverting back to the notation $x = (x_1, x_2)$ we see $f'(a) = 2Aa$. This is pleasing, just like derivative of x^2 at a being $2a$.

You can take a symmetric $n \times n$ matrix A and define $f(x) = x^t A x$ for $x \in R^n$. Then the same argument as above shows you

$$f'(x) = 2Ax, \quad x \in R^n.$$

6. You can think of more complicated functions. For example you can consider 2×2 matrix as a point in R^4.

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} = (a_{11}, a_{12}, a_{21}, a_{22}) \in R^4.$$

You can think of matrix multiplication as a map from R^8 to R^4

$$f(a_{11}, a_{12}, a_{21}, a_{22}, b_{11}, b_{12}, b_{21}, b_{22}) = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}$$

$$= (a_{11}b_{11} + a_{12}b_{21}, a_{11}b_{12} + a_{12}b_{22}, - - -, - - -).$$

Those are not difficult to handle, but we need to wait till we gain experience and and not get confused easily.

Now let us return to the main problem with which we started, namely, the chain rule.

Let $\Omega \subset R^m$ and $\Omega_1 \subset R^n$ be open sets.

$$\Phi : \Omega \to R^n; \quad f : \Omega_1 \to R^k$$

be C^1 maps. Assume that $\Phi(x) \in \Omega_1$ for all $x \in \Omega$. Define the composition

$$F(x) = f(\Phi(x)), \qquad x \in \Omega.$$

Theorem: F is a C^1 map from $\Omega \to R^k$. Further

$$F'(a) = f'(\Phi(a))\Phi'(a).$$

Note that f' is $k \times n$ matrix and Φ' is $n \times m$ matrix and so the product is $k \times m$ matrix and defines linear transformation from R^m to R^k.

proof is exactly like the earlier situation.

Let $a \in \Omega$ be fixed. let $b = \Phi(a)$. Let $f'(b) = B$ and $\Phi'(a) = A$. Fic $\epsilon > 0$. We need to show $\delta > 0$ so that

$$||x - a|| < \delta \Rightarrow \frac{||F(x) - F(a) - BA(x - a)||}{||x - a||} < \epsilon.$$

that is

$$||x - a|| < \delta \Rightarrow \frac{||f(\Phi(x)) - f(\phi(a)) - BA(x - a)||}{||x - a||} < \epsilon.$$

We show $\delta_1 > 0$ so that

$$||x - a|| < \delta_1 \Rightarrow \frac{||f(\Phi(x)) - f(\phi(a)) - B[\Phi(x) - \Phi(a)]||}{||x - a||} < \epsilon/2.$$

We show $\delta_2 > 0$ so that

$$||x-a|| < \delta_2 \Rightarrow \frac{||B[\Phi(x)-\phi(a)] - BA(x-a)||}{||x-a||} < \epsilon/2.$$

If $||x-a|| < \delta = \min\{\delta_1.\delta - 2\}$ then both the inequalities hold and adding them gives the desired inequality.

To get δ_2:

First note that given any $k \times n$ matrix B, there is a number c so that $||Bx|| < c||x||$ for any $x \in R^n$. In fact let

$$M = \max\{|b_{i,j} : i \leq k, j \leq n\}$$

then,

$$Bx = (\sum b_{1j}x_j, \sum b_{2j}x_j, \cdots \sum b_{kj}x_j)$$

Since

$$(\sum b_{ij}x_j)^2 \leq (\sum |b_{ij}||x_j|)^2 \leq M^2k||x||^2.$$

Here we have used the following fact: if you square the sum, you get square terms and cross products, but $2\alpha\beta \leq \alpha^2 + \beta^2$. Thus cross products are again bounded by square terms. This is true for each of the k coordinates of Bx. Hence

$$||Bx||^2 \leq k^2M^2||x||^2.$$

Thus $c = kM$ would do.

Returning to our problem, fix $c > 0$ as above. Using differentiability of ϕ get $\delta_2 > 0$ so that

$$||x-a|| < \delta_2 \Rightarrow \frac{||\Phi(x) - \Phi(a) - A(x-a)||}{||x-a||} < \epsilon/(2c).$$

This will satisfy requirement of δ_2.

To get δ_1:

First fix an $r > 0$ so that the closed ball around a of radius r is contained in Ω. Since $\Phi = (\Phi_1, \Phi_2, \cdots, \Phi_n)$ is a C^1 function, we see that $\nabla\Phi_i$ is bounded on the closed ball. Now fix M so that

$$||x-a|| \leq r \Rightarrow ||\nabla\Phi_i(x)|| \leq M/n, \quad i = 1, 2, \cdots n.$$

Using differentiability of f get $\eta > 0$ so that

$$||y-b|| < \eta \Rightarrow \frac{||f(y)-f(b)-B(y-b)||}{||y-b||} < \epsilon/(2M).$$

Npw choose $\delta_1 > 0$ so that $||x-a|| < \delta_1$ implies $||\Phi(x)-b|| < \eta$. This is just by continuity of Φ. By reducing if necessary, we shall assume $\delta_1 < r$.

Let now x be such that $||x-a|| < \delta_1$. Need to show

$$\frac{||f(\Phi(x))-f(\phi(a))-B[\Phi(x)-\Phi(a)]||}{||x-a||} < \epsilon/2.$$

if $\Phi(x) = \Phi(a)$ there is nothing to be done. Otherwise, denote the point $\Phi(x)$ by y the above expression equals

$$\frac{||f(y)-f(b)-B(y-b)||}{||y-b||} \ \frac{||\Phi(x)-\Phi(a)||}{||x-a||}$$

By choice of δ_1, we conclude that $||\phi(x)-b|| < \eta$ so that choice of η tells us that the first term above is at most $\epsilon/(2M)$. The second term is norm of

$$\left(\frac{\Phi_i(x)-\Phi_i(a)}{||x-a||} : 1 \leq i \leq k\right).$$

By the mean values theorem, there are points P_i such that this vector is

$$\left(\frac{\nabla\Phi_i(P_i)\cdot(x-a)}{||x-a||} 1 \leq i \leq k\right).$$

Note that we can apply the mean slue theorem, the points x and a are all in a disc which is contained in ω and hence the lines joining are also contained in Ω. Since $\delta_1 < r$, choice of M tells us, with Cauchy-Schwarz that each entry of the vector above is at most M/n and hence its norm is at most M. so the product is at most $\epsilon/2$.

This completes the proof of chain rule.