Math 371 - The Chain Rule - Picards theorem

Erik Kjær Pedersen

November 29, 2005

Differentiation is approximation by linear maps.
Specifically given a map

$$f : \mathbb{R} \to \mathbb{R}$$

and a fixed point $x \in \mathbb{R}$ we look for the best linear map approximating $f(x+h) - f(x)$.
Linear maps from $\mathbb{R}$ to $\mathbb{R}$ are just multiplication by some number.

So we are looking for the best number a so that

$$f(x+h) - f(x) = ah + \varepsilon(h)$$

Best means that $\varepsilon(h)$ is as small as possible.
Usually $\varepsilon(h)$ will go to 0 when h goes to 0
This is because that happens to the two other terms if f is continuous.
We would like it to be small of the second degree.
Specifically and precisely we want that

$$\frac{\varepsilon(h)}{h}$$

goes to 0 when h goes to 0.
This means that $\varepsilon(h)$ becomes very small when h is small.
When this is the case we call the constant a for $f'(x)$.

Recall the well known chain rule

Theorem

Let $g : \mathbb{R} \to \mathbb{R}$ be differentiable at x and $f : \mathbb{R} \to \mathbb{R}$ be differentiable at $g(x)$.
Then $f \circ g$ is differentable at x and the derivative is $f'(g(x))g'(x)$

Let us think of $g'(x)$ as a 1 by 1 matrix, and $f'(g(x))$ as a 1 by 1 matrix.
We see that this theorem says that when g is approximated by the linear map determined by $g'(x)$ and f is approximated by the linear map determined by $f'(g(x))$
Then the composite of f and g is approximated by the composite i. e. the product of the linear maps.
This formulation generalizes.
Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a map.
We say that f is differentiable at the point x if

$$f(x+h) - f(x) = A(h) + \varepsilon(h)$$

where $\frac{||\varepsilon(h)||}{||h||}$ goes to 0 when h goes to 0.
and A is a linear map.
We obtain the matrix for A by the following procedure:
f had m components, $f = (f_1, f_2, \ldots, f_m)$.
Each of these functions f_i, depend on n variables
$f_i = f_i(x_1, x_2, \ldots, x_n)$.
We now do the following:

- Pretend that all the variables are constant except for x_j.
- We then have f_i is a function of x_j
- We can differentiate to get the derivative of f_i with respect to x_j pretending that all the other variables are constants.
- This derivative is called a partial derivative and is written
- $$\frac{\partial f_i}{\partial x_j}$$
- These form a matrix which is the matrix for A, the derivative of f.

Theorem

Let $g : \mathbb{R}^k \to \mathbb{R}^n$ be differentiable at x with derivative B and $f : \mathbb{R}^n \to \mathbb{R}^m$ be differentiable at $g(x)$ with derivative A. Then $f \circ g$ is differentiable at x with derivative AB.

To see how this works in practice consider the following example

Example

Consider $g : \mathbb{R} \to \mathbb{R}^2$ given by $g(t) = (\cos(t), \sin(t))$ and $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x, y) = x^2 + y^2$.
The derivative of g is given by differentiating each coordinate function so it is $(-\sin(t), \cos(t))$.
The derivative of f is given by first pretending y is a constant and differentiating with respect to x giving $\frac{\partial f}{\partial x} = 2x$ and $\frac{\partial f}{\partial y} = 2y$.
The derivative of f is thus represented by the matrix $\begin{pmatrix} 2x \\ 2y \end{pmatrix}$.
The composite is $\begin{pmatrix} -\sin(t) & \cos(t) \end{pmatrix} \begin{pmatrix} 2\cos(t) \\ 2\sin(t) \end{pmatrix} = 0$.
This is of course in agreement with $f \circ g(t) = \cos^2(t) + \sin^2(t)$ which is constant 1, so the derivative should be 0.

Example

In this example we spell out what the chain rule means at the formulae level.
Suppose f depends on $x_1, x_2, \ldots x_n$ and each of the x_i's depend on t.
That means that f indirectly depends on t and the chain rule implies the formula

$$\frac{df}{dt} = \sum \frac{\partial f}{\partial x_i} \frac{d(x_i)}{dt}$$

This example is simply matrix multiplication as it occurs in the chain rule spelled out.

Example

In applied sciences it is common not to name the functions, by just saying that $y_1, y_1, \ldots, y_k$ depend on $x_1, x_2, \ldots x_n$, and $z_1, z_2, \ldots, z_n$ depend on $y_1, y_1, \ldots, y_k$
Therefore they indirectly depend on $x_1, x_2, \ldots x_n$.
In this notation the formula for the chain rule becomes

$$\frac{\partial z_i}{\partial x_j} = \sum_k \frac{\partial z_i}{\partial y_k} \frac{\partial y_k}{\partial x_j}$$

This phenomenon of summing over an index that appears down and up is so common that Einstein invented the so called "sum convention"
The Einstein sum convention says that you always sum over an index that appears both high and low so you do not have to write the sum symbol.

We will now consider Picards theorem

Picards theorem was originally a purely theoretical statement.

Consider the differential equation

$$y' = f(x, y), \qquad y(x_0) = y_0$$

Does it have a solution, and if yes, how many.

In many cases we can guess solutions, but it is of course important to know whether there are some we have overlooked.

Picard proved two theorems, an existence theorem and a uniqueness theorem

The existence theorem says that if $f(x, y)$ is continuous in some rectangle around (x_0, y_0), then there is a solution in some interval around x_0.

The uniquesness theorem says that if $f(x, y)$ and $\partial f \partial y$ are continuous in some rectangle around (x_0, y_0), then any two solutions will be the same in some interval around x_0.

We shall not be concerned so much about these theorem here. Instead we will discuss Picards method for finding solutions. This is because it is constructive, and can be used to program computers to find solutions.
Rewrite the equation

$$y' = f(x, y), \qquad y(x_0) = y_0$$

as

$$y = y_0 + \int_{x_0}^{x} f(t, y(t))dt$$

Notice we have to change to a different variable t in the integration because the upper limit is x, so we can not use x as integration variable.
This does not help much since y still appears on both sides of the equation.

The idea is know to start out with a reasonable guess for y on the right side and hope the answer on the left side is a little better. So we replace y in the integral by the constant y_0 and define as a first step

$$y_1(x) = y_0 + \int_{x_0}^{x} f(t, y_0)dt$$

The hope is now that $y_1(x)$ is a little closer to the real y, and that putting it in on the right hand side gets you even closer

$$y_2(x) = y_0 + \int_{x_0}^{x} f(t, y_1(t))dt$$

Notice that t now appears in both variables in the function f. It is only in the special first step that we only have dependence on t in the first variable.

Iteratively we now define

$$y_k(x) = y_0 + \int_{x_0}^{x} f(t, y_{k-1}(t))dt$$

It turns out that under the conditions set out by Picard the sequence of functions $\{y_k\}$ converge to a function that solves the differentiable equation
Furthermore it turns out that this description of the function can prove uniqueness with the extra requirement on $\partial f/\partial y$.
But the significance to us nowadays is, that it provides us with a solution on an explicit form.

Example

Consider the differentiable equation

$$y' = x^2 - y \qquad y(0) = 1$$

So in this case we have

$$f(x, y) = x^2 - y$$

The first step in finding a solution by Picards method is to calculate

- y_1
- $= y_0 + \int_{x_0}^{x} f(t, y_0)dt$
- $= 1 + \int_0^x (t^2 - 1)dt$
- $= 1 + \frac{1}{3}x^3 - x = 1 - x + \frac{1}{3}x^3$

The second step is to use y_1 instead of $y(0)$ to get

- y_2
- $= y_0 + \int_{x_0}^{x} f(t, y_1(t))dt$
- $= 1 + \int_0^x (t^2 - 1 + t - \frac{1}{3}t^3)dt$
- $= 1 + \frac{1}{3}x^3 - x + \frac{1}{2}x^2 - \frac{1}{12}x^4$
- $= 1 - x + \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{12}x^4$

I hope
Computers get better at this than humans very quickly.