Lecture 8

Math Foundations Team

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction

- We introduced the Taylor/MacLaurin series, partial derivatives and gradients.
- We are interested now in some aspects of Taylor's series which we have not discussed.
- Specifically, we will delve into the theory of Taylor series and derive the remainder term for a truncated Taylor series.
- We shall also develop the Taylor's series in two variables and motivate the derivation of the Hessian matrix which plays a huge role in data science, especially in neural network cost function minimization.

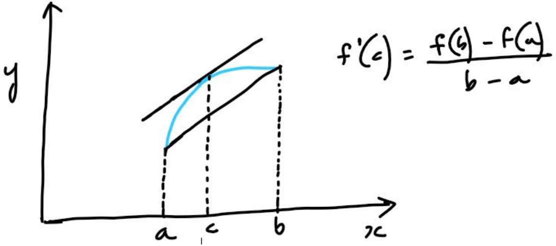
Where does the Taylor series come from?

- Our development of the Taylor series will mirror the argument given in the document "Proof of Taylor's theorem" which will also be uploaded as class material.
  Theorem: Suppose $f : (a, b) \to R$ is a function on (a, b), where a, b in R with $a < b$. Assume that f is n-times differentiable in the open interval (a, b) and $f, f', f'', \ldots f^{n-1}$ all extend continuously to the closed interval $[a, b]$, such that the extended functions are still called $f, f'' \ldots f^{n-1}$. Then there exists $c \in (a, b)$ such that

$$f(b) = \sum_{k=0}^{k=n-1} \frac{f^k(a)}{k!}(b-a)^k + \frac{f^n(c)}{n!}(b-a)^n \qquad (1)$$

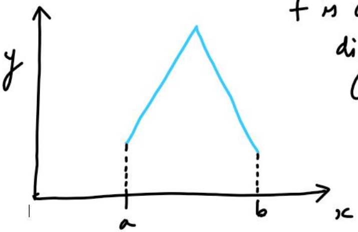
Mean-value theorem

- For $n = 1$, the statement of Taylor's theorem boils down to the mean-value theorem which is that if a function f is continuous on $[a, b]$ and differentiable on the interval (a, b), then there exists a value $c \in (a, b)$ such that $f'(c) = \frac{f(b)-f(a)}{b-a}$ as in the following figure:



Mean-value theorem

- Note that the requirement that f be a differentiable function in the mean-value theorem is needed as the theorem is not valid for functions f that are not differentiable as in the example below:



f is continuous but not differentiable over (a, b)

Mean-value theorem

- The proof of the mean-value theorem comes from Rolle's theorem whose statement follows. We shall show that the development of Taylor's series involves the repeated application of Rolle's theorem below:
  Theorem: If f is a continuous function on $[a, b]$ and differentiable on (a, b) with $f(a) = f(b) = 0$, then there exists some c in (a, b) such that $f'(c) = 0$.

Development of Taylor's series

- Let $F(x)$ be a function over the region $(a, b) \in R$ such that $F(a) = F'(a) = F^{n-1}(a) = 0$, and $F(b) = 0$. Then there exists a $c \in (a, b)$ such that $F^n(c) = 0$. Let us call this Proposition P.
- Proposition P follows from an n-fold application of Rolle's theorem as follows: since $F(a) = F(b) = 0$, an application of Rolle's theorem tells us that there is a $c_1 \in (a, b)$ such that $F'(c_1) = 0$.
- Now since $F'(a) = F'(c_1) = 0$, there exists $c_2 \in (a, c_1)$ such that $F''(c_2) = 0$.
- Continuing this argument we get $a < c_n < c_{n-1} < \ldots < c_1 < b$ such that $F^k(c_k) = 0$ for $k = 1, 2, \ldots n$.

Development of Taylor's series

- Thus we have $F^n(c) = 0$ for $c = c_n \in (a, b)$.
- To construct a polynomial that approximates a function f we use the ideas of the previous slide.
- Let the polynomial used to approximate the function f be of the form $P(x) = \sum_{k=0}^{k=n} a_k(x-a)^k$. We will now find the coefficients $a_0, a_1, a_2, \ldots a_n$ such that $F(x) = f(x) - P(x)$ satisfies $F(a) = F(b)$ and $F(a) = F'(a) = F^{n-1}(a) = 0$ of the previous slide. Then we have $F^n(c) = 0$ for $c \in (a, b)$.
- We can see that $F^k(a) = f^k(a) - k!a_k, k = 0, 1, 2, \ldots n-1$, and to satisfy the requirements of the function F defined on the previous slide we need to set $F(a) = F'(a) = F^{n-1}(a) = 0$, and $F(b) = 0$. This gives $a_k = \frac{f^k(a)}{k!}, k = 0, 1, 2, \ldots n-1$.

Development of Taylor's series

- We finally need to determine a_n. To do so, we note that $F(b) = f(b) - \sum_{k=0}^{k=n} a_k(b-a)^k = 0$. This becomes $F(b) = f(b) - \sum_{k=0}^{k=n-1} \frac{f^k(a)}{k!}(b-a)^k - a_n(b-a)^n = 0$ once we substitute for $a_k = \frac{f^k(a)}{k!}$, for $k = 1, 2, \ldots n-1$.
- We can then solve for a_n to get $a_n = \frac{1}{(b-a)^n}(f(b) - \sum_{k=0}^{k=n-1} \frac{f^k(a)}{k!}(b-a)^k)$.
- Our function F satisfies the requirements of proposition P, i.e $(F(a) = F(b)$ and $F'(a) = F''(a) = \ldots F^{n-1}(a) = 0)$. Therefore there exists a point c, according to proposition P such that $F^n(c) = 0$.

Development of Taylor's series

- Since $F^n(c) = f^n(c) - P^n(c)$, we have
  $F^n(c) = f^n(c) - n!a_n =$
  $f^n(c) - \frac{n!}{(b-a)^n}(f(b) - \sum_{k=0}^{k=n-1} \frac{f^k(a)}{k!}(b-a)^k) = 0,$
- Rearranging the above we have the final Taylor's series expansion with a remainder term:
  $f(b) = \sum_{k=0}^{k=n-1} \frac{f^k(a)}{k!}(b-a)^k) + \frac{f^n(c)}{n!}(b-a)^n.$
- The last term in the expression above is the remainder term which should be used when the series is truncated at a certain number of terms.

Taylor's series in two variables

- How do we develop Taylor's series in two variables?
- Let $f(x, y)$ be a function in two variables with continuous partial derivatives in an open region R containing the point $P(a, b)$ where the partial derivates f_x and f_y are both zero. Note that f_x and f_y are zero because the gradient vanishes at critical points, and (a, b) is a critical point.
- Let h and k be increments small enough to put the point $S(a + h, b + k)$ in the region R. We parameterise the line segment PS as $x = a + th, y = b + tk, t \in [0, 1]$.
- Now let $F(t) = f(a + th, b + tk)$. F is now a function of only one variable. We can compute $F'(t) = f_x \frac{dx}{dt} + f_y \frac{dy}{dt} = hf_x + kf_y$.

Taylor's series in two variables

- Now f_x and f_y are differentiable functions, F' is a differentiable function of t and we can write $F''(t) = \frac{\partial F'}{\partial x}\frac{dx}{dt} + \frac{\partial F'}{\partial y}\frac{dy}{dt}$.
- Since $x = a + th$ and $y = b + tk$, and $F' = hf_x + kf_y$, we can write
$F''(t) = \frac{\partial(hf_x+kf_y)}{\partial x}h + \frac{\partial(hf_x+kf_y)}{\partial y}k = h^2 f_{xx} + 2hkf_{xy} + k^2 f_{yy}$.
- Since F and F' are continuous on $[0,1]$ and F' is differentiable on $(0,1)$ we can apply Taylor's theorem and obtain $F(1) = F(0) + F'(0)(1-0) + \frac{1}{2}F''(c)$ for some c between 0 and 1.
- Rewriting this in terms of x, y, we have
$f(a+h, b+k) = f(a,b) + hf_x(a,b) + kf_y(a,b) + \frac{1}{2}(h^2 f_{xx} + 2hkf_{xy} + k^2 f_{yy})|_{a+ch,b+ck}$

Taylor's series in two variables

- Since $f_x(a, b) = f_y(a, b) = 0$, the expression for $f(a + h, b + k)$ simplifies to $f(a + h, b + k) - f(a, b) = \frac{1}{2}(h^2 f_{xx} + 2hkf_{xy} + k^2 f_{yy})|_{a+ch,b+ck}$
- The presence of an extremum at $f(a, b)$ is dependent on the sign of $f(a + h, b + k) - f(a, b)$ for arbitrary h and k.
- This is the same as the sign of $Q(c) = (h^2 f_{xx} + 2hkf_{xy} + k^2 f_{yy})|_{a+ch,b+ck}$.
- We shall now study the sign of $Q(c)$.

Hessian matrix

- If $Q(0) \neq 0$, the sign of $Q(c)$ for small c will be the same as the sign of $Q(0)$ for sufficiently small values of h and k.
- We can predict the sign of $Q(0) = (h^2 f_{xx}(a, b) + 2hkf_{xy}(a, b) + k^2 f_{yy}(a, b))$ from the signs of f_{xx} and $f_{xx}f_{yy} - f_{xy}^2$ evaluated at (a, b).
- Multiply both sides of the equation for $Q(0)$ by f_{xx}, and rearrange the right hand side to get $f_{xx}Q(0) = (hf_{xx} + kf_{xy})^2 + (f_{xx}f_{yy} - f_{xy}^2)k^2$.
- What can we now conclude about the nature of the neighbourhood of the function at (a, b)?

Hessian matrix

- If $f_{xx} < 0$ and $f_{xx}f_{yy} - f_{xy}^2 > 0$ at (a, b) then $Q(0) < 0$ for all sufficiently small non-zero values of h and k, then f has a local maximum value at (a, b).
- If $f_{xx} > 0$ and $f_{xx}f_{yy} - f_{xy}^2 > 0$ at (a, b) then $Q(0) > 0$ for all sufficiently small non-zero values of h and k, then f has a local minimum value at (a, b).
- If $f_{xx}f_{yy} - f_{xy}^2 < 0$ at (a, b) there are combinations of small values for h and k for which $Q(0) > 0$ and other combinations of h and k for which $Q(0) < 0$. This means that f has a saddle point at (a, b).
- If $f_{xx}f_{yy} - f_{xy}^2 = 0$ at (a, b) another test is needed.

Hessian matrix

- The considerations of the previous slide show that determining whether there is a minimum or maximum at the point (a, b) boils down to looking at the following matrix and asking if it is positive-definite or not. This is the Hessian matrix.

$$\begin{bmatrix} f_{xx} & f_{xy} \\ f_{xy} & f_{yy} \end{bmatrix} \tag{2}$$

Hessian matrix

- A necessary and sufficient criterion of positive-definiteness for a Hermitian matrix (such as the Hessian matrix) is Sylvester's criterion - the determinant of every upper left $m \times m$ submatrix should be positive which means that $f_{xx} > 0$ and $f_{xx} f_{yy} - f_{xy}^2 > 0$ at (a, b). This condition corresponds to bullet point 2 on Slide 15.
- Note that the Hessian matrix is symmetric, so that in case of a local minimum it is a symmetric, positive-definite matrix which we know from the linear algebra part of this course is one that has positive eigenvalues.

Hessian Matrix

- For a local maximum we need to have a negative-definite matrix which means that $f_{xx} < 0$ and $f_{xx}f_{yy} - f_{xy}^2 > 0$ at (a, b). In this case, the determinant of every upper left $m \times m$ submatrix is negative if m is odd, and positive if m is even. The eigenvalues of the Hessian matrix are all negative in this case.
- The Hessian is the collection of all second-order partial derivatives. If f(x, y) is a twice (continuously) differentiable function, then $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}$ i.e., the order of differentiation does not matter, and the corresponding Hessian matrix is symmetric. The Hessian is denoted as $\nabla^2_{x,y} f(x, y)$