# Topic 11: Calculus Review

02-680: Essentials of Mathematics and Statistics

October 18, 2024

## 1 Derivatives

From some function $f(x)$ where $x$ is a scalar:

- the derivative $\frac{df}{dx}$ is th change in value of $f$ as you increase/decrease $x$.
- Formally $\frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$.

Lets derive one of our known rules below, for instance let $f(x) = x^n$.

$$\begin{aligned}
\frac{d(x^n)}{dx} &= \lim_{h \to 0} \frac{(x-h)^n - x^n}{h} && (1)\\
&= \lim_{h \to 0} \frac{\sum_{i=0}^{n} \binom{n}{i} x^{(n-i)} h^i - x^n}{h} && (2)\\
&= \lim_{h \to 0} \frac{\sum_{i=1}^{n} \binom{n}{i} x^{(n-i)} h^i}{h} && (3)\\
&= \lim_{h \to 0} \sum_{i=1}^{n} \binom{n}{i} x^{(n-i)} h^{i-1} && (4)\\
&= \lim_{h \to 0} \left[ \binom{n}{1} x^{(n-1)} h^{1-1} + \sum_{i=2}^{n} \binom{n}{i} x^{(n-i)} h^{i-1} \right] && (5)\\
&= n x^{(n-1)} + \lim_{h \to 0} \left[ \sum_{i=2}^{n} \binom{n}{i} x^{(n-i)} h^{i-1} \right] && (6)\\
&= n x^{(n-1)} && (7)
\end{aligned}$$

## 2 Integrals

For some function $f'(x)$ where $x$ is a scalar, the integral $\int f'(x)$ can be thought of as the inverse of differentiation

For more review, we recommend: 3blue1brown.com/topics/calculus

## 3 Application: Gradient Based Optimization

***Optimization*** is the task of either minimizing or maximizing some function. For some $f(x)$, find the $x$ that maximizes $f(x)$. Using mathematics:

$$x^* = \underset{\forall x}{\text{argmax}}\, f(x)$$

(Note that above we use maximization, but this would be equivalent to minimizing some function $g(x) = -f(x)$.)

Remember the derivative of a function is the rate of change, or slope; thus it can be used to tell us how to change $x$ in order to have a maximizing impact on $f(x)$.

$$\frac{df}{dx} \approx \frac{f(x+\varepsilon) - f(x)}{\varepsilon}$$

we can rewrite $\frac{df}{dx}$ as $f'(x)$ for ease, so then reducing

$$\varepsilon f'(x) \approx f(x+\varepsilon) - f(x)$$

$$f(x) + \varepsilon f'(x) \approx f(x+\varepsilon)$$

Notice that when $f'(x) = 0$, we can change x, but nothing happens. We call this a ***critical*** or ***stationary*** point.

Lets look at an example, lets say we want to mimize:

$$f(x) = \frac{x^2}{2} \quad \rightarrow \quad f'(x) = x.$$

If $x > 0; f'(x) > 0$ and thus to reach a critical point we need to add a negative $\varepsilon$ (since we want to minimize the function value). Alternatively, if $x < 0 : f'(x) < 0$ and thus we need to add a positive $\varepsilon$. If $x = 0, f'(x) = 0$ and thus we've found a critical point.

Looking at another example lets say we want to still find the minimum, but for

$$g(x) = \frac{-x^2}{2} \quad \rightarrow \quad g'(x) = -x.$$

Its still the case that we want to add an $\varepsilon$ opposite the sign of the slope since we want to minimize; if $g'(x) < 0$ then $\varepsilon > 0$, $g'(x) > 0$ then $\varepsilon < 0$, and $g'(x) = 0$ found a critical point.

So once again, we have a critical point at $x = 0$, but if we look at the two examples in the first if we were slightly away from 0 the derivative would have sent us *toward* 0, in the second it sends us away. So we can say the critical point of $x = 0$ is ***stable*** for $f(x)$ and ***unstable*** for $g(x)$.

What we've described here is a slight simplification of ***gradient descent*** first described by Cauchy in the 1867. One of the most commonly used optimization procedures in machine learning.

### 3.1 Example

Consider the equation

$$3x^4 - 10x^3 - 12x^2 + 18x - 3.$$

The derivative is

$$12x^3 - 30x^2 - 24x + 18 = 6(x - 3)(x + 1)(2x - 1).$$

This has 3 values that are zero: $x = 3, -1, \frac{1}{2}$.

To find out if they are local minima or maxima we can take the second derivative:

$$36x^2 - 60x - 24$$

and determine if its positive or negative at each point: $x = 3(+), -1(+), \frac{1}{2}(-)$.

Thus only $x = 3$ and 1 are candidates for minima, we can get their original values:

$$3(3)^4 - 10(3)^3 - 12(3)^2 + 18(3) - 3 = -84$$

$$3(-1)^4 - 10(-1)^3 - 12(-1)^2 + 18(-1) - 3 = -20$$

Thus the global minimum is at $x = 3$.

## 4 Gradients (Multivariable Derivation)

When a function has multiple variables we cannot simply take the derivative of the whole thing. For instance, lets say we have a function $f : \mathbb{R}^n \mapsto \mathbb{R}$, as a real example think of euclidean norm:

$$e(x) = \sqrt{x_1^2 + x_2^2}.$$

Now, since $x$ is a vector, when need to take the gradient $\nabla f$ (sort of $\frac{df}{dx}$ when $x$ is a vector). We define the gradient as follows:

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & .... & \frac{\partial f}{\partial x_n} \end{bmatrix}$$

when $x \in \mathbb{R}^n$. In the example above that means

$$\nabla e = \begin{bmatrix} \frac{\partial e}{\partial x_1} & \frac{\partial e}{\partial x_1} \end{bmatrix} = \begin{bmatrix} \frac{x_1}{\sqrt{x_1^2+x_2^2}} & \frac{x_2}{\sqrt{x_1^2+x_2^2}} \end{bmatrix}$$

Note we need the chain rule then the polynomial rule for both derivatives.

They key here is to realize that the dimension of the gradient (the number of derivatives) is dependent on the dimension of the input to the function (not the output).

## 5 Application: Least Squares Minimization

Assume we have a problem which can be set up as

$$Ax = b$$

where $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^n$. But let it be the case that $b \notin span(columns(A))$, therefore we know $\nexists x \in \mathbb{R}^n$ such that $Ax = b$.

Can we find something "close enough"?

Lets say we want to minimize $\|Ax - b\|_2^2$ (we sometimes call this the squared error).

**Norm Squared**

Its helpful here to remember that for any vector

$$x \in \mathbb{R}^n : \|x\|_2^2 = \sum_{i=1}^{n} x_i^2 = x^T x.$$

The derivation is left as an exercise.

Remember from above, to find a minimum we need to find the critical points and evaluate them, so we need to find the gradient.

$$\begin{aligned} \|Ax - b\|_2^2 &= (Ax - b)^T (Ax - b) \\ &= (x^T A^T - b^T)(Ax - b) \\ &= x^T A^T Ax - 2AXb^t + b^T b \end{aligned}$$

The gradient with respect to $x$ of this is then $2A^T Ax - 2A^T b$.

For this to be 0,

$$A^T Ax = A^T b$$

solving for $x$ (assuming $A^T A$ is invertible)

$$x = (A^T A)^{-1} A^T b.$$

# 6 Jacobian (Multivariable/Multifunction Derivation)

Sometimes we also have functions that have both multiple variable input, and multiple variable output. So any function $g : \mathbb{R}^n \mapsto \mathbb{R}^m$ with $m, n \in \mathbb{R}^{\geq 2}$. We say the ***Jacobian*** of $g$ is

$$J_g(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

So for instance $h : \mathbb{R}^2 \mapsto \mathbb{R}^2$:

$$h(x) = x^T \begin{bmatrix} 3 & 1 \\ 0 & -1 \end{bmatrix} x = \begin{bmatrix} 3x_1^2 \\ x_1 x_2 - x_2^2 \end{bmatrix}$$

then the jacobian is

$$J_h(x) = \begin{bmatrix} \frac{d(3x_1^2)}{x_1} & \frac{d(3x_1^2)}{x_2} \\ \frac{d(x_1 x_2 - x_2^2)}{x_1} & \frac{d(x_1 x_2 - x_2^2)}{x_2} \end{bmatrix} = \begin{bmatrix} 6x_1 & 0 \\ x_2 & x_1 - 2x_2 \end{bmatrix}$$

# Useful References

Deisenroth, Faisal, and Ong, "Mathematics for Machine Learning". §5

Table 1: Differentiation rules

| | |
|---|---|
| constant multiple rule | $\frac{d}{dx}\left(cf\left(x\right)\right) = cf'\left(x\right)$ |
| polynomial rule | $\frac{d}{dx}\left(x^n\right) = nx^{n-1}$ |
| exponent rules | $\frac{d}{dx}\left(e^{g(x)}\right) = e^{g(x)}g'(x)$<br><br>$\frac{d}{dx}\left(c^{g(x)}\right) = c^{g(x)}g'(x)\ln c$ |
| log rule | $\frac{d}{dx}\left(\log_c g\left(x\right)\right) = \frac{g'(x)}{g(x)\ln a}$ |
| sum rule | $\frac{d}{dx}\left(f\left(x\right) + g\left(x\right)\right) = f'(x) + g'(x)$ |
| product rule | $\frac{d}{dx}\left(f\left(x\right)g\left(x\right)\right) = f'(x)g(x) + f(x)g'(x)$ |
| quotient rule | $\frac{d}{dx}\left(\frac{f(x)}{g(x)}\right) = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$ |
| chain rule | $\frac{d}{dx}\left(f\left(g\left(x\right)\right)\right) = f'\left(g\left(x\right)\right)g'(x)$ |

Table 2: Integration rules

| | |
|---|---|
| constant multiple rule | $\int cf\,(x)\,dx = c\int f(x)dx$ |
| polynomial rule | $\int x^n dx = \frac{x^{n+1}}{a+1} + c$ |
| exponent rules | $\int e^x dx = e^x$<br><br>$\int c^x dx = \frac{c^x}{\ln c}$ |
| log rule | $\int \frac{1}{x} dx = \ln\lvert x\rvert + c$ |
| sum rule | $\int (f(x) \pm g(x))\,dx = \int f(x)dx \pm \int g(x)dx$ |