02-680 Module 11

Essentials of Mathematics and Statistics

Aidan Jan

October 30, 2025

Calculus Review

Derivatives

For some function $f(x)$ where x is a scalar,

- the derivative $\frac{\mathrm{d}f}{\mathrm{d}x}$ is the change in the value of f as you increase/decrease x.
- formally, $\frac{\mathrm{d}f}{\mathrm{d}x} = \lim_{h \to 0} \frac{f(x+h)-f(x)}{h}$
- Know basic derivatives. (e.g., polynomials, trig functions, inverse trig functions, exponentials, logarithms)
- Know basic derivative rules. (e.g., chain rule, product rule, quotient rule)

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 example:

$$\begin{aligned} f'(x) &= nx^{n-1} \\ \int nx^{n-1}\mathrm{d}x &= n\int x^{n-1}\mathrm{d}x \\ &= n \cdot \frac{x^n}{n} + C \\ &= x^n + C \end{aligned}$$

Know the following:

- Rules: constant multiple rule, polynomial rule, exponent rules, log rule, sum rule.

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:

$$x^* = \arg\max_{\forall x} f(x)$$

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

Gradient Descent

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 $f(x)$.

$$\frac{\mathrm{d}f}{\mathrm{d}x} \approx \frac{f(x+\epsilon) - f(x)}{\epsilon} \rightarrow f(x+\epsilon) \sim f(x) + \epsilon f'(x) \rightarrow f(x+\epsilon) - f(x) \sim \epsilon f'(x)$$

When $f'(x) = 0$ we call this a **critical**, or **stationary**, point.

We can rewrite $\frac{\mathrm{d}f}{\mathrm{d}x}$ as $f'(x)$ for ease, so then reducing

$$\begin{aligned} \epsilon f'(x) &\approx f(x+\epsilon) - f(x) \\ f(x) + \epsilon f'(x) &\approx f(x+\epsilon) \end{aligned}$$

Notice that when $f'(x) = 0$, we can change x, but nothing happens.

Example

Suppose we want to minimize:

$$f(x) = \frac{x^2}{2}$$

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

If we change the example, so $g(x) = \frac{-x^2}{2}$, and $g'(x) = -x$, it is still the case that we want to add an ϵ opposite the sign of the slope since we awant to minimize; if $g'(x) < 0$ then $\epsilon > 0$. If $g'(x) > 0$ then $\epsilon < 0$, and $g'(x) = 0$ is the critical point.

Either way, we have a critical point at $x = 0$, but in the first example, being slightly away from zero sends us to zero, but in the second example, it sends us away. Therefore, we 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 year 1867. It is one of the most commonly used optimization procedures in machine learning.

Gradients (Multivariable Derivation)

When a function has multiple variables we cannot simply take the derivative of the whole thing, like function $f : \mathbb{R}^n \mapsto \mathbb{R}$.

Euclidean norm (instance):

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

Since x is a vector, we use gradient ∇f (sort of like $\frac{\mathrm{d}f}{\mathrm{d}x}$ when x is a vector)

$$\nabla f = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \cdots, \frac{\partial f}{\partial x_n} \right], \quad x \in \mathbb{R}^n$$

In the example of Euclidean norm,

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

Application: Least Squares Minimization

Problem: $Ax = b$ where $A \in \mathbb{R}^{m \times n}$, and $b \in \mathbb{R}^n$. But $b \notin \text{span}(\text{col}(A))$. In this condition, $\nexists x \in \mathbb{R}$ such that $Ax = b$. But we can find something "close enough".

It is helpful here to remember that for any vector:

$$x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n, \quad \|x\|_2^2 = \sum_{i=1}^{n} x_i^2 = x^T x$$

Let's minimize $\|Ax - b\|_2^2$. Applying the norm formulas,

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

Now, finding the gradient. The gradient with respect to x is

$$2A^T Ax - 2A^T b = 0 \to A^T Ax = A^T b$$

Solve for x (assuming $A^T A$ is invertible): $x = (A^T A)^{-1} A^T b$.

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 $m, n \in \mathbb{R}^{\geq 2}$, the **Jacobian** of g is:

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

For a vector map $g : \mathbb{R}^n \to \mathbb{R}^m$, the Jacobian $J_g(x)$ captures how small input changes or moves the outputs.

- **Local linearization.** Near x: $g(x + \mathrm{d}x) \approx g(x) + J_g(x)\mathrm{d}x$. So, $J_g(x)$ maps small input changes $\mathrm{d}x$ to output changes $\mathrm{d}y$. This is the multi-output generalization of the gradient.
- **Chain rule for multivariable compositions:** If $y = g(x)$ and $z = h(y)$, then $\mathrm{d}z = J_h(y) J_g(x) \mathrm{d}x$, which is the matrix form used for gradient flow through layers.

Example

Let $h : \mathbb{R}^2 \mapsto \mathbb{R}^2$

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

The Jacobian:

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