§ Review of multivariable calculus ~~[illegible]~~

Main things we care about: derivatives, Taylor expansions, some single variable integrals, ... multiple integrals aren't so useful for us.

Note: the multivariable calculus you learned probably built up to Stokes theorem etc. These things are important for physics, less so for optimization. We need to be able to take derivatives of multivariable functions and reason about the behavior of these functions. You may have done this in $\mathbb{R}^2$ and $\mathbb{R}^3$, maybe saw a bit of $\mathbb{R}^n$. So let's review and build up the material you need.

First, recall the def'n of the ordinary derivative of a single-valued function of a real variable, which we may write: $f: \mathbb{R} \to \mathbb{R}$, i.e. $y = f(x)$:

$$f'(x) = \frac{df}{dx} = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

Partial derivatives are just ordinary derivatives applied to a single variable of a multiple variable function $f: \mathbb{R}^n \to \mathbb{R}$ with the remaining arguments held constant:

$$\frac{\partial f}{\partial x_i} = \lim_{h \to 0} \frac{f(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_n)}{h}.$$

Most important for us is the gradient ∇f ("∇" is pronounced "nabla" or "del" or "grad"), which is the vector in $\mathbb{R}^n$ which is the direction and rate of steepest increase in f at a point x.

How can we recover the well-known expression for ∇f:

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right) \in \mathbb{R}^n$$

from this verbal description of the gradient? Let's consider finding v that solves:

a unit vector

$$\underset{\substack{v \in \mathbb{R}^n \\ \|v\|_2 = 1}}{\text{maximize}} \ \frac{d}{dt} f(x+tv)\Big|_{t=0}.$$

That is, we restrict f to lines passing through x, take the ordinary derivative of the resulting single-variable function of t, and try to find the vector $v \in \mathbb{R}^n$ (s.t. $\|v\|_2 = 1$) which maximizes the rate of change (at x). Note:

chain rule

$$\frac{d}{dt} f(x+tv)\Big|_{t=0} = \left(\sum_{i=1}^{n} v_i \frac{\partial f}{\partial x_i}\Big|_{x+tv}\right)\Big|_{t=0}$$

$$= v_1 \frac{\partial f}{\partial x_1} + \dots + v_n \frac{\partial f}{\partial x_n}.$$

Among all $v \in \mathbb{R}^n$ s.t. $\|v\|_2 = 1$, the one which maximizes this quantity is:

$$v = \frac{\left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right)}{\sqrt{\frac{\partial f}{\partial x_1}^2 + \dots + \frac{\partial f}{\partial x_n}^2}}.$$

~~However~~ We can see that this is true by using the Cauchy-Schwarz inequality:

$$\left|\frac{d}{dt} f(x+tv)\Big|_{t=0}\right| = \left|v_1 \frac{\partial f}{\partial x_1} + \dots + v_n \frac{\partial f}{\partial x_n}\right|$$

$$\leq \|(v_1, \dots, v_n)\|_2 \cdot \left\|\left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right)\right\|.$$

We know in the C-S inequality that this inequality is obtained when $(v_1, \dots, v_n)$ is linearly dependent on $\left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right)$. Since the magnitude of $(v_1, \dots, v_n)$ is restricted to be equal to 1, the choice is clear.

Hence, v is the vector which points in the direction of steepest increase by definition. We can also now see that that direction is:

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n}\right),$$

i.e. the gradient of f, and that the slope there (or, the rate of increase) is just $\|\nabla f\|$.

The next differential operator to look at is the Jacobian:

$$Df = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n},$$

where $f : \mathbb{R}^n \to \mathbb{R}^m$ is a map from $\mathbb{R}^m$ to $\mathbb{R}^n$ — it is a vector-valued function of multiple variables. We can write f in terms of its component functions:

$$f = (f_1, \dots, f_m) \in \mathbb{R}^m.$$

Each row of Df is the transpose of the gradient of a component function. Note also that if $g : \mathbb{R}^n \to \mathbb{R}$ is a scalar-valued function, then:

$$\nabla g = \left(\frac{\partial g}{\partial x_1}, \dots, \frac{\partial g}{\partial x_n}\right) = \begin{bmatrix} \frac{\partial g}{\partial x_1} & \cdots & \frac{\partial g}{\partial x_n} \end{bmatrix}^T = Dg^T.$$

That is, the gradient of a scaler-field is the transpose of its Jacobian. Note carefully that we use two notations for vectors in $\mathbb{R}^n$:

$$\underbrace{(x_1, \dots, x_n)}_{\text{"tuple notation"}} = \underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}}_{\text{"column-vector notation"}} \in \mathbb{R}^n.$$

It is important to differentiate between row and column vectors, since they correspond to different geometric objects. A column vector, by our identification with $\mathbb{R}^n$ above, is the usual geometric vector — it is an "arrow" in $\mathbb{R}^n$, with a direction and magnitude we can write the polar representation of a vector as:

$$x = \underbrace{\|x\|}_{\text{magnitude}} \cdot \underbrace{\frac{x}{\|x\|}}_{\text{direction}} \in \mathbb{R}^n.$$

On the other hand, a row vector is better thought of as a linear map, i.e. as a function. E.g., for $a \in \mathbb{R}^{1 \times n}$, we have the mapping:

$$x \overset{a\cdot}{\longmapsto} ax = \sum_{i=1}^{n} a_i x_i.$$

Clearly, there is a close relationship between the two, since we can get one from the other by taking a transpose. In fact, they are dual to each other. To understand the relationship, consider the level sets of the function $x \mapsto ax$ and how they relate to the vector $a^T \in \mathbb{R}^{n \times 1}$.

One omission so far is the transpose of Df, in the case that $f : \mathbb{R}^n \to \mathbb{R}^m$, and $m > 1$. In this case, to define;

$$\nabla f = Df^T = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} = [\nabla f_1 \cdots \nabla f_m] \in \mathbb{R}^{n \times m},$$

we need to regard the matrix ∇f as a vector and suitably interpret it as the ~~~~ vector indicating the direction ~~~~ of steepest increase of the vector field f. Since this is more abstract and not important for the class, we won't go into this.

~~The next quantity to consider~~

The next quantity to consider is the Hessian of a scalar-valued function $f : \mathbb{R}^n \to \mathbb{R}$, defined as:

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \in \mathbb{R}^{n \times n},$$

that is, the matrix of second partials. Note that if each mixed partial commutes:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}, \quad i \neq j,$$

then the Hessian is symmetric. The geometry of the Hessian is extremely important in the theory of optimization, The gradient, Hessian, and Jacobian allow us to write down Taylor expansions of single and multiple-valued functions of multiple variables. First, let $f : \mathbb{R}^n \to \mathbb{R}$, let $x \in \mathbb{R}^n$, and let $h \in \mathbb{R}^n$. Then, the TE of f about x in the variable h is:

$$f(x+h) = f(x) + \nabla f(x)^T h + \frac{1}{2} h^T \nabla^2 f(x) h + O(\|h\|_2^3).$$

Note that the notation $O(\cdots)$ is defined as:

$$f(x) = O(g(x)) \ \left[\text{as } x \to x_0\right] \text{ if } \lim_{x \to x_0} \frac{f(x)}{|g(x)|} < \infty.$$

This is an example of Landau big-O notation. In numerical analysis, we will nearly always use $O(\cdots)$ to describe errors or Taylor expansion remainders, so the tacit assumption is that $x_0 = 0$. Hence;

$O(\|h\|_2^3) \iff$ a stand-in for a function which has magnitude dominated by $C \cdot \|h\|_2^3$ for some $C \geq 0$ as $\|h\|_2 \to 0$.

Along the same lines, we define the TE of a vector-valued function $f : \mathbb{R}^n \to \mathbb{R}^m$ by:

$$f(x+h) = f(x) + Df(x)h + O(\|h\|_2^2).$$

Note that these are vector-valued quantities — i.e., $Df(x)h$ is a matrix-vector product, and $O(\|h\|_2^2)$ is to be interpreted as standing in for a vector-valued quantity whose magnitude (i.e. $\|\cdot\|_2$) is dominated by a constant $C > 0$ times $\|h\|_2^2$ as $\|h\|_2 \to 0$.

Throughout the class, we will fill in this picture of multivariable calculus, adding new tools as we need them.