BASIC MULTIVARIABLE CALCULUS

CSC311 Fall 2020
(Notes by Murat A. Erdogdu)

University of Toronto

1. Basic multivariable calculus. For a given function $f : \mathbb{R}^d \to \mathbb{R}$, we denote its partial derivative with respect to its i-th coordinate as $\partial f(x)/\partial x_i \in \mathbb{R}$. Gradient of this function is simply a vector with i-th coordinate $\partial f(x)/\partial x_i \in \mathbb{R}$. That is,

$$[\nabla f(x)]_i = \frac{\partial f(x)}{\partial x_i}. \tag{1.1}$$

The gradient of a function points in the direction of greatest increase, and its magnitude is the rate of increase in that direction. Therefore, when you are minimizing a function, it makes sense to move in the direction opposite to its gradient.

Similarly, we can define the second derivative of the function f, which is generally referred to as the Hessian of f. It is a matrix and its i, j-th entry is given by

$$[\nabla^2 f(x)]_{ij} = \frac{\partial^2 f(x)}{x_i x_j}. \tag{1.2}$$

Using the above definition, for $x, y \in \mathbb{R}^d$ and $A \in \mathbb{R}^{d \times d}$ we obtain

(a) the gradient with respect to x of $x^T y$ is y,

(b) the gradient with respect to x of $x^T x$ is $2x$,

(c) the gradient with respect to x of $x^T Ax$ is $2Ax$,

(d) the gradient with respect to x of Ax is A.

In some cases, you can see that the above gradients are transposed. This is a matter of definition. You should check the wikipedia page https://en.wikipedia.org/wiki/Matrix_calculus which contains a very detailed list of rules.

1.1. *Least squares problem.* In the least squares problem, we are given a target vector $\mathbf{t} \in \mathbb{R}^N$, a design matrix $\mathbf{X} \in \mathbb{R}^{N \times D}$. We would like to find the weights $\mathbf{w}$ that minimizes the objective function given by the least squares problem

$$\underset{\mathbf{w}}{\text{minimize}}\, \mathcal{J}(\mathbf{w}) =: \frac{1}{2}\|\mathbf{t} - \mathbf{X}\mathbf{w}\|_2^2.$$

We know that a minimum occurs at a critical at which the partial derivatives are equal to 0. i.e. $\partial \mathcal{J}(\mathbf{w})/w_j = 0$ for $j = 1, .., D$. This is equivalent to saying the gradient $\nabla \mathcal{J}(\mathbf{w}) = 0$. We can write

$$\mathcal{J}(\mathbf{w}) = \frac{1}{2}\|\mathbf{t}\|_2^2 + \frac{1}{2}\mathbf{w}^\top \mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{t}^\top \mathbf{X}\mathbf{w}.$$

Taking derivative with respect to the vector $\mathbf{w}$ and setting it equal to 0, we obtain

$$\nabla \mathcal{J}(\mathbf{w}) = \mathbf{X}^\top \mathbf{X}\mathbf{w} - \mathbf{X}^\top \mathbf{t} = 0.$$

If $\mathbf{X}^\top \mathbf{X}$ is invertible, a solution to above linear system is given by

$$\mathbf{w}^{\text{LS}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{t}.$$