Partial Derivatives

Literature:

`Calculus`, Paul Dawkins (available as PDF document)

`MATHEMATICS FOR MACHINE LEARNING` , Deisenroth et. al.

Scope

1. Review of some concepts of partial derivatives

Multivariate Functions

A multivariate function $f(x_1, x_2, \ldots, x_N)$ depends on N independent variables $[x_1, x_2, \ldots, x_N]$. To keep it simple these variables shall be real. The result y of the multivariate function can be a scalar or a vector. But here only the scalar case shall be considered. Moreover it shall be assumed that y is a real number.

$$y = f(x_1, x_2, \ldots, x_N)$$

The independent variables are summarized into a vector:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \\ \vdots \\ x_N \end{bmatrix}$$

A partial derivative is defined like this:

$$\frac{\partial}{\partial x_n} f(\mathbf{x}) = \lim_{\Delta h \to 0} \frac{f(x_1,\ \ldots,\ x_n + \Delta h,\ x_N) - f(x_1,\ \ldots,\ x_n,\ x_N)}{\Delta h}$$

A vector of all partial derivatives

$$\mathbf{g}(\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_n} f(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_N} f(\mathbf{x}) \end{bmatrix}$$

is defined as *gradient* of a the multivariate function $f(\mathbf{x})$. Here the gradient vector has been defined as a column vector. But we could have defined it as a row vector as well. It just depends on how that gradient shall be processed in subsequent steps.

Directional Derivatives

Let $\mathbf{r}$ denote a unit vector (length 1; $|\mathbf{r}| = 1$ with N components:

$$\mathbf{r} = \begin{bmatrix} r_1 \\ \vdots \\ r_n \\ \vdots \\ r_N \end{bmatrix}$$

and

$$|\mathbf{r}| = \sum_{n=1}^{N} r_n^2 = 1$$

When going from $\mathbf{x}$ to $\mathbf{x} + \mathbf{r} \cdot h$ function $f(\mathbf{x})$ changes. The amount of change $\Delta_{\mathbf{r}} f(\mathbf{x})$ is computed here:

$$\Delta_{\mathbf{r}} f(\mathbf{x}) = f(\mathbf{x} + \mathbf{r} \cdot h) - f(\mathbf{x}) = f(x_1 + r_1 \cdot h,\ \ldots,\ x_n + r_n \cdot h,\ \ldots,\ x_N + r_N \cdot h) - f(\mathbf{x}$$

Defining $\Delta x_n = r_n \cdot h$ for $1 \leq n \leq N$ and assuming *vanishingly* small value of h a reasonably good approximation of this change is:

$$\Delta_{\mathbf{r}} f(\mathbf{x}) = f(\mathbf{x} + \mathbf{r} \cdot h) - f(\mathbf{x}) \approx h \cdot \sum_{n=1}^{N} \frac{\partial}{\partial x_n} f(\mathbf{x}) \cdot r_n$$

The rate of change is obtained by dividing both sides of this equation by h:

$$\frac{\Delta_{\mathbf{r}} f(\mathbf{x})}{h} = \frac{f(\mathbf{x} + \mathbf{r} \cdot h) - f(\mathbf{x})}{h} \approx \sum_{n=1}^{N} \frac{\partial}{\partial x_n} f(\mathbf{x}) \cdot r_n$$

In the limit of $h \to 0$ the rate of change converges to the directional derivative $D_{\mathbf{r}} f(\mathbf{x})$ (in the direction of vector $\mathbf{r}$:

$$D_{\mathbf{r}} f(\mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + \mathbf{r} \cdot h) - f(\mathbf{x})}{h} = \sum_{n=1}^{N} \frac{\partial}{\partial x_n} f(\mathbf{x}) \cdot r_n$$

More commonly the directional derivative may be expressed as the dot product of the *gradient* vector and the *directional* vector:

$$D_{\mathbf{r}} f(\mathbf{x}) = \begin{bmatrix} \frac{\partial}{\partial x_1} f(\mathbf{x}) & \ldots & \frac{\partial}{\partial x_n} f(\mathbf{x}) & \ldots & \frac{\partial}{\partial x_N} f(\mathbf{x}) \end{bmatrix} \cdot \begin{bmatrix} r_1 \\ \vdots \\ r_n \\ \vdots \\ r_N \end{bmatrix}$$

Summary

1. If the directional vector $\mathbf{r}$ has the same direction as the gradient vector the directional derivative $D_{\mathbf{r}} f(\mathbf{x})$ is maximized.

2. If vector $\mathbf{r}$ is orthogonal to the gradient vector the directional derivative is 0 (no change in this direction).

3. The direction of *steepest descent* is the gradient vector with each vector component multiplied by -1.

```
In [ ]:
```