2 Optimisation

2.1 Gradient of vector-valued functions

For a function J that maps a column vector $\mathbf{w} \in \mathbb{R}^n$ to $\mathbb{R}$, the gradient is defined as

$$\nabla J(\mathbf{w}) = \begin{pmatrix} \frac{\partial J(\mathbf{w})}{\partial w_1} \\ \vdots \\ \frac{\partial J(\mathbf{w})}{\partial w_n} \end{pmatrix},$$

where $\partial J(\mathbf{w})/\partial w_i$ are the partial derivatives of $J(\mathbf{w})$ with respect to the i-th element of the vector $\mathbf{w} = (w_1, \ldots, w_n)^\top$ (in the standard basis). Alternatively, it is defined to be the column vector $\nabla J(\mathbf{w})$ such that

$$J(\mathbf{w} + \epsilon \mathbf{h}) = J(\mathbf{w}) + \epsilon \left(\nabla J(\mathbf{w})\right)^\top \mathbf{h} + O(\epsilon^2) \tag{2.1}$$

for an arbitrary perturbation $\epsilon\mathbf{h}$. This phrases the derivative in terms of a first-order, or affine, approximation to the perturbed function $J(\mathbf{w} + \epsilon\mathbf{h})$. The derivative ∇J is a linear transformation that maps $\mathbf{h} \in \mathbb{R}^n$ to $\mathbb{R}$ [see Chapter 9, for a formal treatment of derivatives][1].

Use either definition to determine $\nabla J(\mathbf{w})$ for the following functions where $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times n}$ and $f : \mathbb{R} \to \mathbb{R}$ is a differentiable function.

1. $J(\mathbf{w}) = \mathbf{a}^\top \mathbf{w}$.
2. $J(\mathbf{w}) = \mathbf{w}^\top \mathbf{A} \mathbf{w}$.
3. $J(\mathbf{w}) = \mathbf{w}^\top \mathbf{w}$.
4. $J(\mathbf{w}) = ||\mathbf{w}||_2$.
5. $J(\mathbf{w}) = f(||\mathbf{w}||_2)$.

[1] Walter Rudin. Principles of Mathematical Analysis. McGraw Hill, 3rd edition edition, 1976.

2.2 Newton's method

Assume that in the neighbourhood of $\mathbf{w}_0$, a function $J(\mathbf{w})$ can be described by the quadratic approximation

$$f(\mathbf{w}) = c + \mathbf{g}^\top(\mathbf{w} - \mathbf{w}_0) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_0)^\top \mathbf{H}(\mathbf{w} - \mathbf{w}_0),$$

where $c = J(\mathbf{w}_0)$, $\mathbf{g}$ is the gradient of J with respect to $\mathbf{w}$, and $\mathbf{H}$ a symmetric positive definite matrix (e.g. the Hessian matrix for $J(\mathbf{w})$ at $\mathbf{w}_0$ if positive definite).

1. Use Task 2.1 to determine $\nabla f(\mathbf{w})$.
2. A necessary condition for $\mathbf{w}$ being optimal (leading either to a maximum, minimum or a saddle point) is $\nabla f(\mathbf{w}) = 0$. Determine $\mathbf{w}^*$ such that $\nabla f(\mathbf{w})\big|_{\mathbf{w}=\mathbf{w}^*} = 0$. Provide arguments why $\mathbf{w}^*$ is a minimiser of $f(\mathbf{w})$.
3. In terms of Newton's method to minimise $J(\mathbf{w})$, what do $\mathbf{w}_0$ and $\mathbf{w}^*$ stand for?

Newton's method is defined by:

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

which (potentially) converges to a solution of the equation $f(x) = 0$.

4. For the function $f(x) = \frac{x^4}{4} - \frac{x^3}{3} - x$ compute the first 4 iterations using Newton's method starting from $x_0 = 1$.

Gradient descent is defined by:

$$x_{n+1} = x_n - \alpha f'(x_n)$$

where α is the learning rate.

5. For the function $f(x) = \frac{x^4}{4} - \frac{x^3}{3} - x$ compute the first 4 iterations using Gradient descent starting from $x_0 = 1$. You can use $\alpha = 0, 1$.
6. what is the difference between Newton's method and gradient descent? Which one is typically preferred in machine learning and why?