Tutorial 01: Mathematical Background

1 Linear Algebra

1.1 Multiple Transpositions (2P)

$$((\mathbf{A}^{\mathrm{T}})^{\mathrm{T}})_{ij} = (\mathbf{A}^{\mathrm{T}})_{ji} = \mathbf{A}_{ij} \Rightarrow (\mathbf{A}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{A}.$$

Transposing a matrix corresponds to a reflection with respect to the diagonal. The second reflection returns the matrix to its original form.

1.2 Transposing a Matrix Product 2 (3P)

$$((\mathbf{AB})^{\mathrm{T}})_{ij} = (\mathbf{AB})_{ji} = \sum_{l}^{n} \mathbf{A}_{jl}\mathbf{B}_{li} = \sum_{l}^{n} \mathbf{B}_{li}\mathbf{A}_{jl} = \sum_{l}^{n} (\mathbf{B}^{\mathrm{T}})_{il}(\mathbf{A}^{\mathrm{T}})_{lj} = (\mathbf{A}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}})_{ij}.$$

$$(\mathbf{AB})^{\mathrm{T}} \in \mathbb{R}^{p \times m}.$$

1.3 Brackets in Matrix Multiplications (4P)

A simple implementation of the sum of the matrix multiplication (5) in the handout takes $2n-1$ steps for each element ij: n multiplications and $n-1$ additions. This sum needs to be calculated for every of the mp elements in the resulting matrix. Therefore the total number of steps $mp(2n-1)$ which lies in $\mathcal{O}(mnp)$.
There are two ways to calculate the result: $(\mathbf{AB})\mathbf{C}$ and $\mathbf{A}(\mathbf{BC})$. In general consider $\mathbf{C} \in \mathbb{R}^{p \times q}$, $\mathbf{A}$ and $\mathbf{B}$ as before. For a concrete example: $m = 16$, $n = 2$, $p = 4$, $q = 8$
$(\mathbf{AB})\mathbf{C}$:

$$\#_{\text{steps}} = n(2p-1)q + m(2n-1)q \stackrel{\text{insert}}{=} 1088$$

$\mathbf{A}(\mathbf{BC})$:

$$\#_{\text{steps}} = m(2n-1)p + m(2p-1)q \stackrel{\text{insert}}{=} 496$$

$\mathbf{A}(\mathbf{BC})$ is more than twice as fast.

2 Differential Calculus

2.1 Quotient Rule (3P)

$$f(x) = g(x)h(x)^{-1}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}f(x) \overset{\text{product}}{=} h(x)^{-1}\frac{\mathrm{d}}{\mathrm{d}x}g(x)+g(x)\frac{\mathrm{d}}{\mathrm{d}x}(h(x)^{-1}) \overset{\text{chain}}{=} h(x)^{-1}\frac{\mathrm{d}}{\mathrm{d}x}g(x)+g(x)(-1)h(x)^{-2}\frac{\mathrm{d}}{\mathrm{d}x}h(x) =$$

$$= \frac{\frac{\mathrm{d}}{\mathrm{d}x}g(x)}{h(x)} - \frac{g(x)\frac{\mathrm{d}}{\mathrm{d}x}h(x)}{h(x)^2} \quad \rightarrow \quad \frac{\mathrm{d}}{\mathrm{d}x}f(x) = \frac{h(x)\frac{\mathrm{d}}{\mathrm{d}x}g(x) - g(x)\frac{\mathrm{d}}{\mathrm{d}x}h(x)}{h(x)^2}$$

2.2 Derivative of the Sigmoid Function (4P)

$$\sigma(x) = (1 + e^{-x})^{-1}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}\sigma(x) = -(1+e^{-x})^{-2}(-e^{-x}) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{1+e^{-x}-1}{(1+e^{-x})^2} =$$
$$= \frac{1}{1+e^{-x}} - \frac{1}{(1+e^{-x})^2} = \sigma(x) - \sigma(x)^2 = \sigma(x)(1-\sigma(x))$$

2.3 Applying Gradients (4P)

Gradient of f:

$$\nabla_{\mathbf{x}}f(\mathbf{x}) = \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}x_1}(x_1+x_2) \\ \frac{\mathrm{d}}{\mathrm{d}x_2}(x_1+x_2) \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Normalised gradient (we leave out the dependency of $\mathbf{x}$ for a shorter notation):

$$\frac{\nabla_{\mathbf{x}}f}{\|\nabla_{\mathbf{x}}f\|} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Gradient of magnitude ϵ

$$\boldsymbol{\epsilon} = \frac{\epsilon}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \epsilon \ll 1$$

Change of f at $\mathbf{a}$ when changing $\mathbf{a}$ by gradient of magnitude ϵ.

$$\Delta = f(\mathbf{a}) - f(\mathbf{a} - \boldsymbol{\epsilon}) = (a_1 + a_2) - (a_1 + \frac{\epsilon}{\sqrt{2}} + a_2 + \frac{\epsilon}{\sqrt{2}}) = \epsilon\sqrt{2}$$

In contrast: Let $\boldsymbol{\epsilon}'$ point in directory of x_1 (with magnitude ϵ):

$$\boldsymbol{\epsilon}' = \epsilon\begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Accordingly, we calculate the change of f when going in x_1-direction:

$$\Delta' = f(\mathbf{a}) - f(\mathbf{a} - \boldsymbol{\epsilon}') = (a_1 + a_2) - (a_1 - \epsilon + a_2) = \epsilon$$

For the x_2 we obtain the same result as for the x_1 direction.
Therefore, if we take a small step in direction of the gradient at a given point, the increase of f is higher compared to if we take an equally big step in direction of x_1 or x_2. This is an expected result, as the gradient is supposed to point into the direction of the steepest incline.