Exercise Sheet #9

Problem 1 (*Computing neural tangent kernels*)

Consider a single hidden layer network with scalar in- and output,

$$f(x;\boldsymbol{\vartheta}) = \sum_{i=1}^{N} v_i\,\sigma(w_i x), \qquad \boldsymbol{\vartheta} = (v_1,\ldots,v_N,\, w_1,\ldots,w_N),$$

where both the output weights v_i and the input weights w_i are adaptable.

(a) Compute the neural tangent kernel $\Theta(x,x') = \nabla_{\boldsymbol{\vartheta}} f(x)\cdot\nabla_{\boldsymbol{\vartheta}} f(x')$ for general N and general differentiable transfer function $\sigma(z)$.

(b) Specialize to a linear network, $\sigma(z) = z$, and $N = 1$, i.e. $f(x) = v\,w\,x$. Give the explicit tangent kernel and evaluate it for $v = w = 1$.

(c) A single training pair $(x^1, F) = (1, 2)$ is learned via gradient flow, $\dot{\boldsymbol{\vartheta}} = \eta\,(\nabla_{\boldsymbol{\vartheta}} f^1)(F - f^1)$, starting from $v(0) = w(0) = 1$. Assuming the lazy regime, in which the tangent kernel keeps its initial value, solve for $f^1(t) = f(x^1;\boldsymbol{\vartheta}(t))$ explicitly.

Problem 2 (*Attention! A problem about attention*)

In causal dot-product attention, every token i of a context of N_T tokens generates query, key and value vectors, $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i \in \mathbb{R}^D$. The output of the attention layer is

$$\mathbf{y}_i = \sum_{j\le i} \alpha_{ij}\,\mathbf{V}_j, \qquad \alpha_{ij} = \frac{1}{Z_i}\,\mathrm{e}^{-\beta e_{ij}}, \qquad Z_i = \sum_{j\le i} \mathrm{e}^{-\beta e_{ij}}, \qquad e_{ij} = \mathbf{Q}_i\cdot\mathbf{K}_j,$$

where the causality condition $j \le i$ restricts attention to the past, and β is an effective inverse temperature, cf. (10.61) in the lecture.

(a) Count the number of overlaps e_{ij} that need to be evaluated for the full context and verify the claim from the lecture notes that the computational effort of attention scales as $O(N_T^2\,D)$.

Consider now a minimal concrete example, a context of just $N_T = 2$ tokens with two-dimensional activities

$$\mathbf{x}_1 = \begin{pmatrix}1\\0\end{pmatrix}, \qquad \mathbf{x}_2 = \begin{pmatrix}1\\1\end{pmatrix},$$

for which query, key and value matrices are all given by the identity, such that $\mathbf{Q}_i = \mathbf{K}_i = \mathbf{V}_i = \mathbf{x}_i$.

(b) Compute all causally allowed overlaps e_{ij}, $j \leq i$.

(c) For $\beta = 1$, compute the attention weights α_{21} and α_{22}, together with the output $\mathbf{y}_2$ of the second token.