Homework Week 4

Mathematics of deep learning
MASH & IASD 2025

Lecturer: Bruno Loureiro, bruno.loureiro@di.ens.fr

Instructions: This homework is **due on Monday 17/02/2025**. Please send your solutions in a PDF file named HW4_Nom_Prenom.pdf to the above address with the subject "*[MATHSDL2025] Homework 4*". Formats accepted: LaTeX or a **readable** scan of handwritten solutions.

1 Exercises

Exercise 1.

Consider a two-layer neural network with ReLU activation $\sigma(x) = x_+$:

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{\sqrt{p}}\sum_{j=1}^{p} a_j \sigma(\langle \boldsymbol{w}, \boldsymbol{x}\rangle). \tag{1}$$

Assume that the weights are initialised as $a_j^0 \sim \text{Unif}(\{-1,1\})$, $\boldsymbol{w}^0 \sim \mathcal{N}(\underline{0}, \boldsymbol{I}_d)$.

(a) Show that the NTK kernel is given by:

$$\begin{aligned} K_{\text{NTK}}(\boldsymbol{x},\boldsymbol{x}') &:= \langle \boldsymbol{x},\boldsymbol{x}'\rangle \mathbb{E}_{a,\boldsymbol{w}}\left[a^2\sigma'(\langle \boldsymbol{w},\boldsymbol{x})\sigma'(\langle \boldsymbol{w},\boldsymbol{x}\rangle)\right] \\ &= \frac{1}{2} - \frac{1}{2\pi}\arccos\left(\frac{\langle \boldsymbol{x},\boldsymbol{x}'\rangle}{||\boldsymbol{x}||_2 \cdot ||\boldsymbol{x}'||_2}\right) \end{aligned} \tag{2}$$

(b) Let $\boldsymbol{x}_i \in \mathbb{R}^d$ denote a batch of n independently sampled covariates, and assume $\boldsymbol{x}_i \in B(\underline{0}, 1)$. Using Hoeffding's inequality, show that if $p \geq \Omega(\epsilon^{-2} n^2 \log {}^n/_\delta)$, then with probability at least $1-\delta$ over the random initialisation we have:

$$||\hat{\boldsymbol{K}}_{\text{NTK}} - \boldsymbol{K}_{\text{NTK}}||_{\text{F}} \leq \epsilon \tag{3}$$

where $\hat{\boldsymbol{K}}_{\text{NTK}}, \boldsymbol{K}_{\text{NTK}} \in \mathbb{R}^{n\times n}$ with:

$$\begin{aligned} \hat{\boldsymbol{K}}_{\text{NTK},ij} &= \frac{\langle \boldsymbol{x}_i,\boldsymbol{x}_j\rangle}{p}\sum_{k=1}^{p} a_k^2 \sigma'(\langle \boldsymbol{w}_k,\boldsymbol{x}_i)\sigma'(\langle \boldsymbol{w}_k,\boldsymbol{x}_j\rangle), \\ \boldsymbol{K}_{\text{NTK},ij} &= K_{\text{NTK}}(\boldsymbol{x}_i,\boldsymbol{x}_j) \end{aligned} \tag{4}$$

(c) Conclude that for large enough width p, we have that $\lambda_{\min}(\hat{\boldsymbol{K}}_{\text{NTK}}) > 0$ with high-probability.

Exercise 2.

Let $g(f(\boldsymbol{x};\boldsymbol{\theta})) = \phi(f(\boldsymbol{x};\boldsymbol{\theta}))$ where ϕ is a twice differentiable function as $f(\boldsymbol{x};\boldsymbol{\theta})$ a two-layer neural network:

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \frac{1}{\sqrt{p}}\sum_{j=1}^{p} a_j \sigma(\langle \boldsymbol{w}_j, \boldsymbol{x}\rangle) \tag{5}$$

where σ is a twice-differentiable function.

(a) Considering a_j to be fixed, show that for any $\boldsymbol{\theta}$, the Hessian matrix $\boldsymbol{H}_g$ of g can be related to the Hessian matrix $\boldsymbol{H}(\boldsymbol{\theta})$ of f by:

$$\boldsymbol{H}_g(\boldsymbol{\theta}) = \phi'(f(\boldsymbol{x};\boldsymbol{\theta}))\boldsymbol{H}(\boldsymbol{\theta}) + \phi''(f(\boldsymbol{x};\boldsymbol{\theta}))\nabla_{\boldsymbol{w}} f(\boldsymbol{x};\boldsymbol{\theta})\nabla_{\boldsymbol{w}} f(\boldsymbol{x};\boldsymbol{\theta})^\top \tag{6}$$

(b) Under standard initialisation $a^0 \sim \mathrm{Unif}([-1,1])$, what is the scaling in p of the operator norm of each of the terms above?

(c) Conclude that $g(\boldsymbol{x};\boldsymbol{\theta})$ does not linearise as $p \to \infty$.

Exercise 3.

Generalise the argument leading to Proposition 1 to the case where the second layer weights a_j are not fixed.