RWTH Aachen University **Winter 2023/24**

Instructor: Prof. Dr. Semih Çaycı
Teaching Assistant: Johannes Müller, M.Sc.

Mathematical Foundations of Deep Learning (11.80020)
Assignment 3

Due: Thursday, Dec. 7th, till 2pm as PDF via Moodle upload, TeX submission are encouraged
Each problem is worth 4 points, there are 20 points on this sheet. Submission in pairs is possible.

Throughout this assignment we consider a shallow network with NTK parametrization

$$F(x; w, c) := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} c_i \sigma(w_i^\top x) \quad \text{for } w \in \mathbb{R}^{md}, c \in \mathbb{R}^m, x \in \mathbb{R}^d.$$

Q1. (Properties of ReLU networks) Show the following statements if σ is the ReLU function and assume that $|c_i| \leq 1$ for all $i = 1, \dots, m$:

(a) For any $x \in \mathbb{R}^d$ the mapping $w \mapsto F(x; w, c)$ is $\frac{\|x\|_2}{\sqrt{m}}$-Lipschitz, i.e., it holds that

$$|F(x; w, c) - F(x; w', c)| \leq \frac{\|x\|_2}{\sqrt{m}} \cdot \|w - w'\|_{1,2} \leq \|x\|_2 \cdot \|w - w'\|_{2,2}$$

for all $w, w' \in \mathbb{R}^d$, where

$$\|w\|_{p,q} := \left(\sum_{i=1}^{m} \|w_i\|_q^p \right)^{1/p} \qquad \text{for all } w \in \mathbb{R}^{md}. \tag{1}$$

Remark: You can use without proof that the ReLU function is 1-Lipschitz.

(b) If $(w(0), c)$ are sampled from a symmetric Xavier initialization, then with probability one we have $|F(x; w, c)| \leq \|x\|_2 \cdot \|w - w(0)\|_{2,2}$ for all $x \in \mathbb{R}^d$ and $w \in \mathbb{R}^{md}$.
Hint: You can use part (a).

(c) Consider an infinitely wide neural network given by

$$f^\star(x) = \mathbb{E}_{w \sim \mathcal{N}(0, I_d)} \left[v(w)^\top x \mathbb{1}\{w^\top x \geq 0\} \right] \quad \text{for all } x \in \mathbb{R}^d$$

for a suitable transportation map $v \colon \mathbb{R}^d \to \mathbb{R}^d$ with $\alpha := \mathbb{E}_{w \sim \mathcal{N}(0, I_d)}[\|v(w)\|_2^2] < +\infty$. Show that

$$|f^\star(x)| \leq \alpha \cdot \|x\|_2 \quad \text{for all } x \in \mathbb{R}^d.$$

Remark: **Q6** shows that α is the RKHS norm of $f^\star$ in the RKHS induced by the NTK.

Q2. (NTK and linearization for smooth activation) Let $\sigma \colon \mathbb{R} \to \mathbb{R}$ be a β-smooth activation function.

(a) Assume a symmetric Xavier initialization, i.e., $w \sim \mathcal{N}(0, \sigma^2 I_d)$ and $c \sim$ Rademacher and consider the NTK

$$K(x, x') := \mathbb{E}_w \left[x^\top x' \sigma'(w^\top x) \sigma'(w^\top x') \right].$$

and the finite width NTK

$$K^{(m)}(x,x') := \frac{1}{m}\sum_{k=1}^{m} x^\top x' \sigma'(w_k^\top x)\sigma'(w_k^\top x'),$$

where $w_1,\ldots,w_k \sim \mathcal{N}(0,\sigma^2 I_d)$ are independent. Further, assume that $|\sigma'(t)| \le L$ for all $t \in \mathbb{R}$. Show that for $\delta \in (0,1)$ we have

$$\mathbb{P}\left(\left|K(x,x') - K^{(m)}(x,x')\right| > t\right) \le \exp\left(-\frac{t^2 m}{2|x^\top x'|^2 L^4}\right) \quad \text{for all } t > 0.$$

(b) Consider data points $x_1,\ldots,x_n \in \mathbb{R}^d$ with $\|x_i\|_2 \le 1$ and consider the NTK matrices $H, H^{(m)} \in \mathbb{R}^{n\times n}$ given by $H_{ij} := K(x_i,x_j)$ and $H_{ij}^{(m)} := K^{(m)}(x_i,x_j)$. Show that

$$\mathbb{P}\left(\|H - H^{(m)}\|_{2,2} > t\right) \le n^2 \exp\left(-\frac{t^2 m}{2n^2 L^4}\right) \quad \text{for all } t > 0.$$

(c) Let us fix $w \in \mathbb{R}^{md}$ and $c \in \mathbb{R}^m$ and consider the linearized network

$$F_0(x;w') := F(x;w,c) + \nabla_w F(x;w,c)^\top (w' - w).$$

Show that for all $w' \in \mathbb{R}^{md}, x \in \mathbb{R}^d$ we have

$$|F(x;w',c) - F_0(x;w')| \le \frac{\beta\|c\|_\infty \|x\|_2}{2\sqrt{m}} \cdot \|w' - w\|_{1,2}$$

where $\|\cdot\|_{2,2}$ is defined in (1).

Q3. (NTK linearization when training all weights) Let σ be the ReLU.

(a) Assume that $w \sim \mathcal{N}(0,\sigma^2)$ and $c \sim$ Rademacher and consider the NTK

$$K(x,x') := \mathbb{E}_w\left[x^\top x' \mathbb{1}\{w^\top x \ge 0\}\mathbb{1}\{w^\top x' \ge 0\}\right] + \mathbb{E}_w\left[\sigma(w^\top x)\sigma(w^\top x')\right]$$

when training all weights. Further, consider the finite width NTK

$$K^{(m)}(x,x') := \frac{1}{m}\sum_{k=1}^{m} x^\top x' \mathbb{1}\{w_k^\top x \ge 0\}\mathbb{1}\{w_k^\top x' \ge 0\} + \frac{1}{m}\sum_{k=1}^{m}\sigma(w_k^\top x)\sigma(w_k^\top x'),$$

where $w_1,\ldots,w_k \sim \mathcal{N}(0,\sigma^2)$ are independently sampled. Show that for any $x, x' \in \mathbb{R}^d$ it holds that $K^{(m)}(x,x') \to K(x,x')$ for $m \to \infty$ almost surely.

(b) Fix $(w,c) \in \mathbb{R}^{md} \times \mathbb{R}^m$ and consider the linearized neural network

$$F_0(x;w',c') := F(x;w,c) + \nabla_w F(x;w,c)^\top (w' - w) + \nabla_c F(x;w,c)^\top (c' - c).$$

Show that for any $w' \in \mathbb{R}^{md}, c' \in \mathbb{R}^m, x \in \mathbb{R}^d$ it holds that

$$\left|F(x;w',c') - F_0(x;w',c')\right| \le \frac{(2\|c\|_2 + \|c - c'\|_2)\|w' - w\|_{2,2}}{\sqrt{m}} \cdot \|x\|_2.$$

Q4. (Convergence of SGD for underparametrized linear l^2-regression) Consider a linear model, i.e., $f_\theta(x) = \theta^\top \Phi(x)$ for a fixed feature function $\Phi\colon \mathbb{X} \to \mathbb{R}^p$, where $\theta \in \mathbb{R}^p$. Further, we consider the l^2 sample loss $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$, which leads to the empirical risk

$$L(\theta) = \hat{\mathcal{R}}_S(f_\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(\theta^\top \Phi(x_i) - y_i\right)^2 = \frac{1}{2n} \|\Phi(X)\theta - Y\|_2^2,$$

where $\Phi(X)_{ij} \coloneqq \Phi(x_i)_j$ and $Y_i = y_i$ is convex and consider the Gramian $G = \Phi(X)^\top \Phi(X)$. We fix some $R > 0$ and consider the projected stochastic gradient descent update

$$\begin{aligned} \widetilde{\theta}_{t+1} &= \theta_t - \eta \Phi(x_{i_t})(\theta^\top \Phi(x_{i_t}) - y_{i_t}), \\ \theta_{t+1} &= \Pi_{B_2(0,R)} \widetilde{\theta}_{t+1} \end{aligned}$$

where $i_t \sim \mathcal{U}(\{1, \dots, n\})$ be indices that are drawn independently and uniformly over $\{1, \dots, n\}$. Show that choosing $\eta = \frac{1}{L\sqrt{T}}$ we have that

$$\mathbb{E} L\left(\frac{1}{T} \sum_{t=0}^{T-1} \theta_t\right) - \min_{\theta \in B_2(0,R)} L(\theta) \leq \frac{2RL}{\sqrt{T}},$$

for a suitable constant $L \geq 0$ that bounds the noise level of the gradient estimates and might depend on the training data as well as on R.

Remark: Note that since we are optimizing a quadratic function over a bounded domain, the objective is β-smooth and hence choosing $\eta = \beta^{-1}$ would yield a $O(\frac{1}{T})$ convergence rate.

Q5. (Sum of kernels) Consider two Mercer kernels K_1 and K_2 and let $K = K_1 + K_2$.

(a) Show that K is a Mercer kernel.

(b) Show that $\mathcal{H}_K = \mathcal{H}_{K_1} + \mathcal{H}_{K_2} \coloneqq \{f + g : f \in \mathcal{H}_{K_1}, g \in \mathcal{H}_{K_2}\}$, where $\mathcal{H}_K, \mathcal{H}_{K_1}$ and $\mathcal{H}_{K_2}$ denotes the RKHS of K, K_1 and K_2, respectively.

(c) Show that

$$\|f\|_{\mathcal{H}_K} = \inf\left\{\sqrt{\|g\|_{K_1}^2 + \|h\|_{K_2}^2} : g + h = f\right\} \quad \text{for all } f \in \mathcal{H}_K.$$

Remark: In particular, this shows that the RKHS of the NTK of training both w and c is the sum of the RKHS of the NTKs when only training w or c, see also **Q3**.

Note: The following are bonus problems worth 4 points per problem.

Q6. (Bonus problem: Random feature RKHS) Consider an arbitrary set $\mathbb{X}$ a parameter set Θ, a probability measure μ on Θ as well as a feature map $\phi\colon \mathbb{X} \times \Theta \to \mathbb{R}^{d_f}$ such that

$$\mathbb{E}_{\theta \sim \mu}\left[\|\phi(x;\theta)\|_2^2\right] < +\infty \quad \text{for every } x \in \mathbb{X}.$$

We call

$$K(x, x') \coloneqq \mathbb{E}_{\theta \sim \mu}\left[\phi(x;\theta)^\top \phi(x';\theta)\right] \quad \text{for } x, x' \in \mathbb{X}'$$

the *random feature kernel* induced by ϕ. Show that K is a Mercer kernel, i.e., symmetric and positive semi-definite. Further, show that the RKHS of K is given by

$$\mathcal{H}_K = \left\{f(x) = \mathbb{E}_{\theta \sim \mu}\left[u(\theta)^\top \phi(x;\theta)\right] : u \in L^2(\mu; \mathbb{R}^{d_f})\right\}$$

and show that the inner product is given by

$$\langle f, g\rangle_{\mathcal{H}_K} = (u, v)_{L^2(\mu;\mathbb{R}^{d_f})} = \mathbb{E}_{\theta\sim\mu}\left[u(\theta)^\top v(\theta)\right]$$

if $f(x) = \mathbb{E}_{\theta\sim\mu}\left[u(\theta)^\top \phi(x;\theta)\right]$ and $g(x) = \mathbb{E}_{\theta\sim\mu}\left[v(\theta)^\top \phi(x;\theta)\right]$ for $u, v \in \{\phi(x;\cdot) : x \in \mathbb{X}\}^\perp$. Consequently, it holds that

$$\|f\|_{\mathcal{H}_K} = \inf\left\{\|u\|_{L^2(\mu;\mathbb{R}^{d_f})} : f(x) = \mathbb{E}_{\theta\sim\mu}\left[u(\theta)^\top \phi(x;\theta)\right]\right\}.$$

Remark: Note that the NTK is by definition a random feature RKHS, where the features are given by $\phi(x; w) = \nabla_w \sigma(w^\top x) = x\sigma'(w^\top x) \in \mathbb{R}^d$ when training w or

$$\phi(x; w, c) = \begin{pmatrix}\nabla_w c\sigma(w^\top x)\\ \nabla_c c\sigma(w^\top x)\end{pmatrix} = \begin{pmatrix}x^\top c\sigma'(w^\top x)\\ \sigma(w^\top x)\end{pmatrix} \mathbb{R}^{d+1}$$

when training both w and c, respectively. See also **Q1**.