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

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

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 $w, w' \in \mathbb{R}^{md}$, where

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

Remark: You can use without proof that the ReLU function is 1-Lipschitz.
Solution: Using the triangle and the Cauchy-Schwarz inequality, we estimate

$$\begin{aligned} |F(x; w, c) - F(x; w', c)| &= \frac{1}{\sqrt{m}} \left| \sum_{i=1}^{m} c_i (\sigma(w_i^\top x) - \sigma({w_i'}^\top x)) \right| \\ &\leq \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \left| w_i^\top x - {w_i'}^\top x \right| \\ &\leq \frac{\|x\|_2}{\sqrt{m}} \sum_{i=1}^{m} \|w_i - w_i'\|_2 = \frac{\|x\|_2}{\sqrt{m}} \cdot \|w - w'\|_{1,2}. \end{aligned}$$

Further, by Cauchy-Schwarz we have

$$\|w - w'\|_{1,2} = \sum_{i=1}^{m} \|w_i - w_i'\|_2 \leq \|\mathbb{1}\|_2 \|w - w'\|_{2,2} = \sqrt{m} \cdot \|w - w'\|_{2,2},$$

where $\mathbb{1} \in \mathbb{R}^m$ denotes the all one vector.

(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).
Solution: For a symmetric Xavier initialization it holds that $F(x; w(0), c) = 0$ and hence (a) yields the claim.

(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.

Solution: Using the triangle inequality and Cauchy-Schwarz we estimate

$$|f^\star(x)| \leq \mathbb{E}_{w\sim\mathcal{N}(0,I_d)}\left[|v(w)^\top x|\mathbb{1}\{w^\top x \geq 0\}\right] \leq \mathbb{E}_{w\sim\mathcal{N}(0,I_d)}\left[\|v(w)\|_2\|x\|_2\right] = \alpha \cdot \|x\|_2.$$

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

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

Solution: We set $X_k := x^\top x' \sigma'(w^\top x)\sigma'(w^\top x')$ and want to use Hoeffding's inequality for $K^{(m)}(x,x') = \frac{1}{m}\sum X_k$. Note that $|X_k| \leq |x^\top x'| \cdot L^2$. Now Hoeffding's inequality yields the claim.

(b) Consider data points $x_1, \dots, x_n \in \mathbb{R}^d$ with $\|x_i\|_2 \leq 1$ 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

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

Solution: First, note that by the union bound and part (a) we have

$$\mathbb{P}\left(\|H - H^{(m)}\|_\infty > \delta\right) \leq \sum_{i,j=1}^n \mathbb{P}\left(|K(x_i,x_j) - K^{(m)}(x_i,x_j)| > \delta\right) \leq n^2 \exp\left(-\frac{\delta^2 m}{2L^4}\right),$$

where we also used $|x_i^\top x_j| \leq \|x_i\|_2\|x_j\|_2 \leq 1$. Further, note that we have

$$\|H - H^{(m)}\|_{2,2} \leq n\|H - H^{(m)}\|_\infty$$

and therefore $\|H - H^{(m)}\|_{2,2} > t$ implies $\|H - H^{(m)}\|_\infty > \frac{t}{n}$. Therefore, we have

$$\mathbb{P}\left(\|H - H^{(m)}\|_{2,2} > t\right) \leq \mathbb{P}\left(\|H - H^{(m)}\|_\infty > \frac{t}{n}\right) \leq n^2 \exp\left(-\frac{t^2 m}{2n^2 L^4}\right).$$

(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

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

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

Solution: First, note that

$$F_0(x; w') = F(x; w, c) + \frac{1}{\sqrt{m}} \sum_{k=1}^m c_k (w_k' - w_k)^\top x \sigma'(w_k^\top x).$$

Using the smoothness of the activation function we estimate

$$\begin{aligned}
|F(x; w', c) - F_0(x; w')| &\leq \frac{1}{\sqrt{m}} \sum_{k=1}^m \left| c_k \left(\sigma({w_k'}^\top x) - \sigma(w_k^\top x) - ({w_k'}^\top x - w_k^\top x) \sigma'(w_k^\top x) \right) \right| \\
&\leq \frac{\|c\|_\infty}{\sqrt{m}} \sum_{k=1}^m \frac{\beta \cdot |{w_k'}^\top x - w_k^\top x|}{2} \\
&\leq \frac{\beta \|c\|_\infty \|x\|_2}{2\sqrt{m}} \sum_{k=1}^m \|w_k' - w_k\|_2 \\
&\leq \frac{\beta \|c\|_\infty \|x\|_2}{2\sqrt{m}} \cdot \|w' - w\|_{1,2}.
\end{aligned}$$

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

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

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

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

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.

Solution: We set

$$X_k := x^\top x' \mathbb{1}\{w_k^\top x \geq 0\} \mathbb{1}\{w_k^\top x' \geq 0\} + \sigma(w_k^\top x) \sigma(w_k^\top x')$$

and

$$X := x^\top x' \mathbb{1}\{w^\top x \geq 0\} \mathbb{1}\{w^\top x' \geq 0\} + \sigma(w^\top x) \sigma(w^\top x')$$

By the strong law of large numbers it holds that

$$K^{(m)}(x, x') = \frac{1}{m} \sum_{k=1}^m X_k \to \mathbb{E}_w[X] = K(x, x')$$

for $m \to \infty$ almost surely, if $\mathbb{E}_w[|X|] < +\infty$. We estimate

$$|X| \leq |x^\top x'| + |w^\top x| \cdot |w^\top x'| \leq \|x\|_2 \|x'\|_2 (1 + \|w\|_2^2).$$

Noting that $\mathbb{E}_w[\|w\|_2^2] < +\infty$ since $w \sim \mathcal{N}(0, \sigma^2 I_d)$ yields the claim.

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

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

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

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

Solution: We begin by computing

$$\begin{aligned}
F_0(x; w', c') &= \frac{1}{\sqrt{m}} \sum_{k=1}^m c_k \sigma(w_k^\top x) + c_k x^\top (w'_k - w_k) \sigma'(w_k^\top x) + (c'_k - c_k)\sigma(w_k^\top x) \\
&= \frac{1}{\sqrt{m}} \sum_{k=1}^m c_k x^\top (w'_k - w_k) \sigma'(w_k^\top x) + c'_k \sigma(w_k^\top x)
\end{aligned}$$

Now we can estimate

$$\begin{aligned}
\left| F(x; w', c') - F_0(x; w', c') \right| &= \frac{1}{\sqrt{m}} \left| \sum_{k=1}^m c'_k \sigma({w'_k}^\top x) - c_k x^\top (w'_k - w_k) \sigma'(w_k^\top x) - c'_k \sigma(w_k^\top x) \right| \\
&\leq \frac{1}{\sqrt{m}} \sum_{k=1}^m |c'_k| \cdot |\sigma({w'_k}^\top x) - \sigma(w_k^\top x)| + |c_k| \cdot |x^\top (w'_k - w_k)| \\
&\leq \frac{1}{\sqrt{m}} \sum_{k=1}^m |c'_k| \cdot |x^\top (w'_k - w_k)| + |c_k| \cdot |x^\top (w'_k - w_k)| \\
&\leq \frac{\|x\|_2}{\sqrt{m}} \sum_{k=1}^m |c'_k| \cdot \|w'_k - w_k\|_2 + |c_k| \cdot \|w'_k - w_k\|_2 \\
&\leq \frac{(\|c\|_2 + \|c'\|_2)\|w' - w\|_{2,2}}{\sqrt{m}} \cdot \|x\|_2.
\end{aligned}$$

Further, note that $\|c'\|_2 \leq \|c\|_2 + \|c - c'\|_2$.

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.

Solution: We want to apply the general convergence result from the lecture. For this we need to show that $u_t := \Phi(x_{i_t})(\theta^\top\Phi(x_{i_t}) - y_{i_t})$ is an unbiased gradient estimator and $\mathbb{E}[\|u_t\|_2^2|\mathcal{F}_t] \leq L$ for some $L > 0$. First, we note that since the index i_t is independent of $\mathcal{F}_t$ we have

$$\mathbb{E}[u_t|\mathcal{F}_t] = \mathbb{E}[u_t] = \mathbb{E}[\Phi(x_{i_t})(\theta^\top\Phi(x_{i_t}) - y_{i_t})] = \frac{1}{n}\sum_{i=1}^n \Phi(x_i)(\theta^\top\Phi(x_i) - y_i) = \nabla L(\theta).$$

Let us denote

$$B := \max\left(\{\|\Phi(x_i)\|_2 : i = 1,\dots,n\} \cup \{|y_i| : i = 1,\dots,n\}\right).$$

Using $|a-b|^2 \leq 2(a^2+b^2)$ we estimate

$$\begin{aligned}
\mathbb{E}[\|u_t\|_2^2|\mathcal{F}_t] &= \mathbb{E}_i\left[\|\Phi(x_i)(\theta^\top\Phi(x_i) - y_i)\|_2^2\right] \\
&\leq \mathbb{E}_i\left[\|\Phi(x_i)\|_2^2 \cdot |\theta^\top\Phi(x_i) - y_i|^2\right] \\
&2B^2\mathbb{E}_i\left[|\theta^\top\Phi(x_i)|^2 + |y_i|^2\right] \\
&\leq 2B^4 + 2B^2\mathbb{E}_i\left[\|\theta\|_2^2\|\Phi(x_i)|^2\right] \\
&\leq 2B^4(1+R^2) =: L^2.
\end{aligned}$$

Now the theorem on the convergence of projected SGD yields the assertion.

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.

Solution: First, note that $K = K_1 + K_2$ is symmetric. Further, K is positive semidefinite as $(K(x_i,x_j))_{1\leq i,j\leq n} = (K_1(x_i,x_j))_{1\leq i,j\leq n} + (K_2(x_i,x_j))_{1\leq i,j\leq n}$ is the sum of two positive semidefinite matrices for arbitrary $x_1,\dots,x_n \in \mathbb{X}$.

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

Solution: We endow $\mathcal{H}_{K_1} + \mathcal{H}_{K_2}$ is a with a Hilbert space structure by identifying it isometrically with

$$U := \{(g,h) \in \mathcal{H}_{K_1} \times \mathcal{H}_{K_2} : g + h = 0\}^\perp \subseteq \mathcal{H}_{K_1} \times \mathcal{H}_{K_2},$$

where $\mathcal{H}_{K_1} \times \mathcal{H}_{K_2}$ is endowed with the scalar product

$$\langle (g_1,h_1),(g_2,h_2)\rangle_{\mathcal{H}_{K_1}\times\mathcal{H}_{K_2}} := \langle g_1,g_2\rangle_{\mathcal{H}_{K_1}} + \langle h_1,h_2\rangle_{\mathcal{H}_{K_2}}.$$

Hence, for $(g_1, h_1) \in U$ and $(g_2, h_2) \in \mathcal{H}_{K_1} \times \mathcal{H}_{K_2}$ we have

$$\langle g_1 + h_1, g_2 + h_2 \rangle_{\mathcal{H}_{K_1}+\mathcal{H}_{K_2}} = \langle (g_1, h_1), (g_2, h_2) \rangle_{\mathcal{H}_{K_1}\times\mathcal{H}_{K_2}} = \langle g_1, g_2 \rangle_{\mathcal{H}_{K_1}} + \langle h_1, h_2 \rangle_{\mathcal{H}_{K_2}}. \quad (2)$$

Now, we want to show that $\mathcal{H}_{K_1} + \mathcal{H}_{K_2}$ is indeed the RKHS of K for which it suffices to show the reproducing property. For this, we first note that $K(x, \cdot) = K_1(x, 0) + K_2(x, \cdot) \in \mathcal{H}_{K_1} + \mathcal{H}_{K_2}$. Further, for a function $f \in \mathcal{H}_{K_1} + \mathcal{H}_{K_2}$ we pick $(g, h) \in U$ such that $f = g + h$. Now we can compute

$$\begin{aligned}
\langle f, K(x, \cdot) \rangle_{\mathcal{H}_{K_1}+\mathcal{H}_{K_2}} &= \langle g + h, K_1(x, \cdot) + K_2(x, \cdot) \rangle_{\mathcal{H}_{K_1}+\mathcal{H}_{K_2}} \\
&= \langle g, K_1(x, \cdot) \rangle_{\mathcal{H}_{K_1}} + \langle h, K_2(x, \cdot) \rangle_{\mathcal{H}_{K_2}} \\
&= g(x) + h(x) \\
&= f(x),
\end{aligned}$$

where we used (2) as well as the reproducing properties of K_1 and K_2.

(c) Show that

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

Solution: It suffices to show that

$$\|f\|_{\mathcal{H}_{K_1}+\mathcal{H}_{K_2}} = \inf \left\{ \sqrt{\|g\|_{K_1}^2 + \|h\|_{K_2}^2} : g + h = f \right\}.$$

Fix $f \in \mathcal{H}_{K_1} + \mathcal{H}_{K_2}$ and pick $(g, h) \in U$ such that $g + h = f$. Then by the Pythagorean theorem for any $(g', h') \in \mathcal{H}_{K_1} \times \mathcal{H}_{K_2}$ with $g' + h' = f$ it holds that

$$\begin{aligned}
\|(g', h')\|_{\mathcal{H}_{K_1}\times\mathcal{H}_{K_2}}^2 &= \|(g, h)\|_{\mathcal{H}_{K_1}\times\mathcal{H}_{K_2}}^2 + \|(g' - g, h' - h)\|_{\mathcal{H}_{K_1}\times\mathcal{H}_{K_2}}^2 \\
&\geq \|(g, h)\|_{\mathcal{H}_{K_1}\times\mathcal{H}_{K_2}}^2 = \|g\|_{\mathcal{H}_{K_1}}^2.
\end{aligned}$$

Note that $\|(g', h')\|_{\mathcal{H}_{K_1}\times\mathcal{H}_{K_2}}^2 = \|g'\|_{\mathcal{H}_{K_1}}^2 + \|h'\|_{\mathcal{H}_{K_2}}^2$. Overall, this shows

$$\|f\|_{\mathcal{H}_K}^2 = \|f\|_{\mathcal{H}_{K_1}+\mathcal{H}_{K_2}}^2 = \inf \left\{ \|g\|_{K_1}^2 + \|h\|_{K_2}^2 : g + h = f \right\}.$$

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**.

Solution: The symmetry is immediate. For $x_1,\dots,x_n$ we define $A \in \mathbb{R}^{d_f\times n}$ via $A_{ij} := \mathbb{E}_\theta[\phi(x_j;\theta)_i]$. Then $(K(x_i,x_j))_{1\le i,j\le n} = A^T A$ which surely is positive semidefinite. This shows that K is indeed a Mercer kernel.

It is clear that

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

with the inner product

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

for $f(x) = \mathbb{E}_{\theta\sim\mu}\left[u(\theta)^\top\phi(x;\theta)\right]$ and $g(x) = \mathbb{E}_{\theta\sim\mu}\left[v(\theta)^\top\phi(x;\theta)\right]$ for $u, v \in \{\phi(x;\cdot) : x \in \mathbb{X}\}^\perp$ is a Hilbert space. Hence, it suffices to check the reproducing property. First, we note that $K(x,\cdot) \in \mathcal{H}$ as

$$K(x,x') = \mathbb{E}_\theta\left[\phi(x;\theta)^\top\phi(x';\theta)\right]$$

and $\phi(x;\cdot) \in L^2(\mu;\mathbb{R}^{d_f})$. Further, we can check the reproducing property

$$\langle f, K(x,\cdot)\rangle_{\mathcal{H}} = \mathbb{E}_\theta\left[u(\theta)^\top\phi(x;\theta)\right] = f(x).$$

Finally, note that $\mathbb{E}_{\theta\sim\mu}\left[u(\theta)^\top\phi(x;\theta)\right] = \mathbb{E}_{\theta\sim\mu}\left[v(\theta)^\top\phi(x;\theta)\right]$ for all $x \in \mathbb{X}$ if and only if $u - v \in \{\phi(x;\cdot) : x \in \mathbb{X}\}^\perp$ and hence

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

by the Pythagorean theorem.