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 4

Due: Tue., Jan. 9th, 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.

Q1. (Uniform convergence via empirical Rademacher complexity) Let $\mathcal{H}$ be a class of functions from $\mathcal{Z}$ to $[0,1]$ and fix $\delta \in (0,1)$ and consider a probability measure P on $\mathcal{Z}$. Further, we consider a sequence $S = (z_1, \dots, z_n) \in \mathcal{Z}^n$ consisting of independent samples distributed according to P. Show that

$$\sup_{h\in\mathcal{H}} \left\{ \mathbb{E}[h(z)] - \frac{1}{n}\sum_{j=1}^{n} h(z_j) \right\} \leq 2\widehat{\mathrm{Rad}}_S(\mathcal{H}) + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}$$

with probability at least $1-\delta$.

Hint. Use the uniform convergence based on the Rademacher complexity from the lecture and bound the difference of $\widehat{\mathrm{Rad}}_S$ and Rad_n using a suitable concentration inequality.

Q2. (Rademacher calculus) Let $\mathcal{H}, \mathcal{H}_1, \mathcal{H}_2 \subseteq \{f : \mathcal{Z} \to \mathbb{R} \text{ measurable}\}$ be classes of real-valued functions on $\mathcal{Z}$ and consider $S = (z_i)_{i=1,\dots,n} \subseteq \mathcal{Z}$. Show the following properties:

(a) If $c \in \mathbb{R}$, then $\widehat{\mathrm{Rad}}_S(c\mathcal{H}) = |c| \cdot \widehat{\mathrm{Rad}}_S(\mathcal{H})$.

(b) If $\mathcal{H}_1 \subseteq \mathcal{H}_2$, then $\widehat{\mathrm{Rad}}_S(\mathcal{H}_1) \leq \widehat{\mathrm{Rad}}_S(\mathcal{H}_2)$.

(c) It holds that $\widehat{\mathrm{Rad}}_S(\mathcal{H}_1 + \mathcal{H}_2) = \widehat{\mathrm{Rad}}_S(\mathcal{H}_1) + \widehat{\mathrm{Rad}}_S(\mathcal{H}_2)$.

(d) If holds that $\widehat{\mathrm{Rad}}_S(\mathcal{H}) = \widehat{\mathrm{Rad}}_S(\mathrm{conv}(\mathcal{H}))$, where

$$\widehat{\mathrm{Rad}}_S(\mathcal{H}) = \left\{ \sum_{i=1}^{m} \lambda_i h_i : \lambda_i \geq 0, \sum_i \lambda_i = 1, h_i \in \mathcal{H}, m \in \mathbb{N} \right\}$$

denote the *convex hull* of $\mathcal{H}$.

Remark. All of the above remarks directly generalize to the Rademacher complexity Rad_n.

Q3. (Bounding the smallest eigenvalue of the NTK) Let us consider a shallow network

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

where we assume $\sigma : \mathbb{R} \to \mathbb{R}$ to be L-Lipschitz for some $L \geq 0$. We denote the linearized network by

$$F_0(x; w) \coloneqq F(x; w_0, c) + \nabla_w F(x; w_0, c)^\top (w - w_0),$$

where for symmetric initialization we have $F(x; w_0, c) = 0$. Hence, the linearized network falls under the setting of **Q1** with $\Phi(x) = \nabla_w F(x; w_0, c)$ and $\theta = (w - w_0)$. Finally, we denote the finite and infinite width NTKs by

$$K^{(m)}(x, x') := \frac{1}{m} \sum_{k=1}^{m} x^\top x' \sigma'(w_k^\top x) \sigma'(w_k^\top x') \text{ and } K^{(\infty)}(x, x') := \mathbb{E}_w \left[x^\top x' \sigma'(w^\top x) \sigma'(w^\top x') \right].$$

(a) Consider the matrix $H = \Phi(X)\Phi(X)^\top$ introduced in **Q1**. Show that $H_{ij} = K^{(m)}(x_i, x_j)$. *Remark.* This justifies the name NTK matrix used in **Q1** and we set $H^{(m)} := H$.

(b) Assume that $H^{(\infty)} \in \mathbb{R}^{n \times n}$ defined by $H_{ij}^{(\infty)} := K^{(\infty)}(x_i, x_j)$ has full rank or equivalently $\lambda_{\min}(H^{(\infty)}) > 0$ and fix $\delta \in (0, 1)$. Show that

$$\|H^{(m)} - H^{(\infty)}\|_{2,2} = \|H^{(m)} - H^{(\infty)}\|_F \leq \frac{\lambda_{\min}(H^{(\infty)})}{4}$$

and hence $\lambda_{\min}(H^{(m)}) \geq \frac{3\lambda_{\min}(H^{(\infty)})}{4} > 0$ with probability at least $1 - \delta$ if

$$m \geq \frac{64L^2 \log\left(\frac{n}{\delta}\right)}{\lambda_{\min}(H^{(\infty)})^2} \cdot n^2.$$

Hint. You can use **Q2** of Assignment 3. Be careful, the notation is slightly different here.

Remark. In combination with **Q4** this shows that the linearized network with (up to log factors) *quadratic overparametrization* $m = O(\frac{n^2 \log n}{\lambda_{\min}})$ converges linearly. By showing that the optimization of the linearized and the original network stay close (on a scale of $O(\frac{1}{\sqrt{m}})$) one can generalize the linear convergence result to the full model, see [1].

Q4. (Linear convergence of GD for a linear model) Consider a linear model, i.e., $f_\theta(x) = \theta^\top \Phi(x)$ for a fixed feature function $\Phi \colon \mathbb{X} \to \mathbb{R}^{d_f}$, where $\theta \in \mathbb{R}^{d_f}$. 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 $\Phi(X)_{ij} := \Phi(x_i)_j$ and $Y_i = y_i$. We consider the Gramian matrix $G = \Phi(X)^\top \Phi(X)$ as well as the NTK matrix $H = \Phi(X)\Phi(X)^\top$, i.e., $H_{ij} = \Phi(x_i)^\top \Phi(x_j)$, and set $f_\theta(X) := \Phi(X)\theta$.

(a) Let us denote the spectrum, i.e., the set of eigenvalues of a matrix A by $\sigma(A)$. Show that $\sigma(G), \sigma(H) \subseteq \mathbb{R}_{\geq 0}$ and $\sigma(G) \setminus \{0\} = \sigma(H) \setminus \{0\}$.

(b) We consider gradient descent $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$ with step size $\eta > 0$ and denote the residuum by $r_t = f_{\theta_t}(X) - Y$. Show that

$$r_t = (I - \eta H)^t r_0 \quad \text{for all } t \geq 0.$$

(c) We set $f_t(X) := f_{\theta_t}(X)$ and assume that $\operatorname{rank}(H) = n$ or equivalently $\lambda_{\min}(H) > 0$. Show that with step size $\eta = 1/\lambda_{\max}(H)$ it holds that

$$\|f_t - Y\|_2 \leq \left(1 - \frac{\lambda_{\min}(H)}{\lambda_{\max}(H)}\right)^t \|f_0 - Y\|_2 \leq e^{-\frac{\lambda_{\min}(H)}{\lambda_{\max}(H)} \cdot t} \|f_0 - Y\|_2.$$

Hint: Expand r_0 in a suitable eigenbasis.

(d) **Bonus (1 point):** Consider the function space $F := \{f_\theta(X) : \theta \in \mathbb{R}^{d_f}\} \subseteq \mathbb{R}^n$. Show that

$$F = \operatorname{range}(H) = \{Hy : y \in \mathbb{R}^n\}.$$

Further, show that there is a (not necessarily unique) minimizer $\theta^\star$ of L and that $f_{\theta^\star} = f^\star := \Pi_F Y$, where Π_F denotes the Euclidean projection onto F.
Hint: The closed range theorem $\operatorname{range}(A^\top) = \ker(A)^\perp$ for a matrix A might be helpful.

(e) **Bonus (1 point):** Without assuming $\operatorname{rank}(H) = n$, show that

$$\|f_t - f^\star\|_2 \leq \left(1 - \frac{\lambda_{\min}}{\lambda_{\max}(H)}\right)^t \|f_0 - f^\star\|_2,$$

where $\lambda_{\min} := \min(\sigma(H) \setminus \{0\})$ denotes the largest non zero eigenvalue of H.

Remark: In Assignment 2 you showed linear convergence under the assumption that G was full rank, which requires $p \leq n$ which we call the problem *underparametrized.* Here, we show linear convergence with essentially the same rate if the NTK matrix H is full rank, which requires $p \geq n$, i.e., *overparametrization.* Note that we study the functions f_t rather than the parameters θ_t, where the optimization dynamics are described by the NTK.

Q5. (A generalization bound for constrained linear regression) Consider the constrained linear model

$$\mathcal{F}_\rho := \left\{x \mapsto \theta^\top x : \|\theta\|_2 \leq \rho\right\}$$

for some $\rho > 0$ and consider a training set $S = \{(x_i, y_i) : i = 1, \dots, n\}$ that we assume to consist of iid samples from some data distribution P on $\mathbb{R}^d \times \mathbb{R}$ and we assume that $\|x\|_2 \leq 1$ and $|y| \leq 1$ almost surely with respect to P. Further, we consider the l^2-sample loss $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$ and denote the empirical and population risk by $\hat{\mathcal{R}}_S$ and $\mathcal{R}$, respectively. Show that

$$\sup_{f \in \mathcal{F}_\rho} \mathcal{R}(f) - \hat{\mathcal{R}}_S(f) \leq \frac{2\rho(1+\rho)}{\sqrt{n}} + 2(\rho^2 + 1) \cdot \sqrt{\frac{\log(\frac{2}{\delta})}{2n}}$$

with probability $1 - \delta$.

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

Q6. (Massart's finite class lemma) Consider $\mathcal{H} \subseteq \{f \colon \mathcal{Z} \to \mathbb{R} \text{ measurable}\}$ and a training set $S = (z_i)_{i=1,\dots,n} \subseteq \mathcal{Z}$ that is iid with respect to some probability measure P on $\mathcal{Z}$. Show that

$$\widehat{\operatorname{Rad}}_S(\mathcal{H}) \leq \frac{R}{n} \cdot \sqrt{2\log|\mathcal{H}|} \quad \text{and} \quad \operatorname{Rad}_n(\mathcal{H}) \leq \frac{R}{n} \cdot \sqrt{2\log|\mathcal{H}|},$$

where $|\mathcal{H}|$ denotes the cardinality of $\mathcal{H}$ and where $R := \max\left\{\|h\|_\infty : h \in \mathcal{H}\right\}$.

Hint. You can use the maximal inequality from **Q2** of the Assignment 1.

Q7. (Bounding the Rademacher complexity by the RKHS norm) Let $\mathcal{Z}$ be an arbitrary set and let $K \colon \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ be a Mercer kernel with corresponding RKHS $\mathcal{H}$ and consider

$$\mathcal{H}_\rho := \{h \in \mathcal{H} : \|h\|_{\mathcal{H}} \leq \rho\}$$

for some $\rho > 0$. Show that

$$\operatorname{Rad}_n(\mathcal{H}_\rho) \leq \frac{\rho\sqrt{\mathbb{E}[K(z,z)]}}{\sqrt{n}}.$$

Q8. (Sublinear convergence under a generalized PL condition) Consider a function $g\colon \mathbb{R}^d \to \mathbb{R}$ that is bounded from below, i.e., $g^\star \coloneqq \inf_{\theta\in\mathbb{R}^d} g(\theta) > -\infty$ and satisfies the following p-PL inequality

$$\|\nabla g(\theta)\|_2^p \geq 2\mu(g(\theta) - g^\star) \quad \text{for all } \theta \in \mathbb{R}^d, \tag{1}$$

where $\mu > 0$ and $p \in [1, 2)$. Assume that g is β-smooth and consider the gradient descent iterates

$$\theta_{k+1} \coloneqq \theta_k - \frac{1}{\beta}\nabla g(\theta_k)$$

with step size $\frac{1}{\beta}$. Show that for any $\varepsilon > 0$ it holds that

$$g(\theta_k) - g^\star \leq \max\left\{\varepsilon, \left(1 - \frac{(2\mu)^{\frac{2}{p}}}{2\beta} \cdot \varepsilon^{\frac{2}{p}-1}\right)^k (g(\theta_0) - g^\star)\right\} \quad \text{for all } k \in \mathbb{N}.$$

Use this to show gradient descent achieves $g(\theta_k) - g^\star \leq \varepsilon$ if

$$k \geq c \cdot \frac{\log(\varepsilon^{-1})}{\varepsilon^{\frac{2}{p}-1}}$$

for $\varepsilon \to 0$ for a suitable constant $c > 0$.

Additional reflection question (2 points). What happens if $p > 2$?

References

[1] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.