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.

Solution: By the generalization result from the lecture, it holds that

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

with probability at least $1-\delta/2$. Further, by Hoeffding's inequality we have with probability at least $1-\delta/2$ that

$$\left|\widehat{\mathrm{Rad}}_S(\mathcal{H}) - \mathrm{Rad}_n(\mathcal{H})\right| \leq \sqrt{\frac{\log(\frac{2}{\delta})}{2n}}. \tag{2}$$

Combining (1) and (2) yields the claim.

Q2. (Rademacher calculus) Let $\mathcal{H}, \mathcal{H}_1, \mathcal{H}_2 \subseteq \{f\colon \mathcal{Z} \to \mathbb{R} \text{ measurable}\}$ 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

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

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

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

Solution: The properties (a)-(c) are immediate from the definition. For (d) we refer to [2].

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 $\sigma \colon \mathbb{R} \to \mathbb{R}$ 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 **Q4** 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

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

(a) Consider the matrix $H = \Phi(X)\Phi(X)^\top$ introduced in **Q4**. Show that $H_{ij} = K^{(m)}(x_i, x_j)$.
Remark. This justifies the name NTK matrix used in **Q4** and we set $H^{(m)} \coloneqq H$.
Solution: First we compute the feature map

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

This implies

$$H_{ij} = \Phi(x_i)\Phi(x_j) = \frac{1}{m} \sum_{k=1}^{m} x_i^\top x_j \sigma'(w_k^\top x_i) \sigma'(w_k^\top x_j) = K^{(m)}(x_i, x_j).$$

(b) Assume that $H^{(\infty)} \in \mathbb{R}^{n \times n}$ defined by $H_{ij}^{(\infty)} \coloneqq K^{(\infty)}(x_i, x_j)$ with $\|x_i\|_2 \leq 1$ 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 $1 - 2\delta$ if

$$m \geq \frac{64 L^4 \log\left(\frac{n}{\delta}\right)}{\lambda_{\min}(H^{(\infty)})^2} \cdot n^2.$$

Hint. You can use **Q2** of Assignment 3. Be careful, the notation is slightly different here.
Solution: By **Q2** of Assignment 3 we have

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

Setting $t \coloneqq \frac{\lambda_{\min}(H^{(\infty)})}{4}$ and solving for m we obtain

$$m \geq \frac{32L^4 \log(\frac{n^2}{\delta})}{\lambda_{\min}(H^{(\infty)})} \cdot n^2.$$

Note now that $\delta^2 \leq \delta$ and hence

$$\log\left(\frac{n^2}{\delta}\right) \leq \log\left(\frac{n^2}{\delta^2}\right) = 2\log\left(\frac{n}{\delta}\right).$$

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 $\Phi(X)_{ij} \coloneqq \Phi(x_i)_j$ 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 $f_\theta(X) \coloneqq \Phi(X)\theta$.

(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 $\sigma(G) \setminus \{0\} = \sigma(H) \setminus \{0\}$.
Solution: Fix an eigenvalue $\lambda \in \sigma(A^\top A)$ with eigenvector $x \neq 0$. Then $0 \leq (Ax)^\top Ax = x^\top A^\top Ax = \lambda x^\top x > 0$. This yields $\lambda \geq 0$. Further, if $\lambda > 0$, for $y \coloneqq Ax$ we have $AA^\top y = AA^\top Ax = A\lambda x = \lambda y$ and hence $\lambda \in \sigma(AA^\top)$.

(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

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

Solution: Recall from **Q5** of Assignment 2 that $\nabla L(\theta) = \Phi(X)^\top \Phi(X)\theta - \Phi(X)^\top Y$. Consequently, we compute

$$f_{\theta_{t+1}}(X) = \Phi(X)\theta_{t+1} = \Phi(X)\theta_t - \eta \Phi(X) \nabla L(\theta_t) = f_{\theta_t}(X) - \eta H(f_{\theta_t}(X) - Y).$$

Substracting Y on both sides and iterating over t yields the claim.

(c) We set $f_t(X) \coloneqq f_{\theta_t}(X)$ 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.

Solution: Let $(v_i)_{i=1,\dots,n}$ be an orthonormal eigenbasis of H with eigenvalues $\lambda_1 \leq \dots \leq \lambda_n$. Consider the decomposition $r_0 = \sum_{i=1}^n \alpha_i v_i$. Then by the Pythagorean theorem we have

$$\begin{aligned}
\|r_t\|_2^2 &= \left\| \sum_{i=1}^n (1-\eta\lambda_i)^t \alpha_i v_i \right\|_2^2 \\
&= \sum_{i=1}^n (1-\eta\lambda_i)^t \alpha_i^2 \\
&\leq \left(1 - \frac{\lambda_{\min}(H)}{\lambda_{\max}(H)}\right)^t \sum_{i=1}^n \alpha_i^2 \\
&= \left(1 - \frac{\lambda_{\min}(H)}{\lambda_{\max}(H)}\right)^t \|r_0\|_2^2.
\end{aligned}$$

Finally, we can use $(1-s)^t \leq (e^{-s})^t = e^{-st}$ for $s \in \mathbb{R}$, $t \in \mathbb{N}$.

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

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

Further, show that there is a (not necessarily unique) minimizer $\theta^\star$ of L and that $f_{\theta^\star} = f^\star \coloneqq \Pi_F Y$, where Π_F denotes the Euclidean projection onto F.

Hint: The closed range theorem $\text{range}(A^\top) = \ker(A)^\perp$ for a matrix A might be helpful.

Solution: It is clear that $\text{range}(AA^\top) \subseteq \text{range}(A)$. Fix now $y \in \text{range}(A)$ and $x \in \ker(A)^\perp$ with $Ax = y$ (if $x \notin \ker(A)^\perp$, then deompose $x = v + w$ with $v \in \ker(A)^\perp$ and $w \in \ker(A)$; surely $Ax = Av = y$). Then by the closed range theorem we can pick z such that $A^\top z = x$. Now we have $AA^\top z = y$ and hence have shown $y \in \text{range}(AA^\top)$.

There is a unique minimizer $f^\star$ of $\|f - Y\|_2^2$ over F. Now we can take any $\theta^\star$ with $f_{\theta^\star} = f^\star$.

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

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

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

Solution: Note that the iteration from b) didn't use any assumption on the rank of H and therefore remains valid here. Hence, we can conclude just like in c) with the only difference that we expand $f_0 - f^\star$ in the eigenbasis instead of $f_0 - Y$.

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

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

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 at least $1 - \delta$.

Solution: First, recall from the lecture that for a bounded function class $\mathcal{H}$, i.e., if $\|h\|_\infty \leq R$ it holds that

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

where we consider the function class

$$\mathcal{H} = \left\{ (x, y) \mapsto \ell(\theta^\top x, y) : \|\theta|_2 \leq \rho \right\}.$$

Note that this function class is bounded as

$$|\ell(\theta^\top x, y)| = \frac{1}{2}(\theta^\top x - y)^2 \leq \frac{1}{2}\big(\|\theta\|_2 \|x\|_2 + |y|\big)^2 \leq \frac{(\rho+1)^2}{2}$$

almost surely with resepect to P and hence we can choose $R = \frac{(\rho+1)^2}{2}$. Hence, it remains to bound the Rademacher complexity of $\mathcal{H}$. First, we use the Talagrand-contraction principle, which implies

$$\mathrm{Rad}_n(\mathcal{H}) \leq L\mathrm{Rad}_n(\mathcal{F}_\rho),$$

where L denotes the Lipschitz constant of $\phi = (\hat{y} \mapsto \ell(\hat{y}, y))$, where $|\hat{y}| \leq \rho$ and $|y| \leq 1$. Note that

$$|\phi'(\hat{y})| = |\hat{y} - y| \leq |\hat{y}| + |y| \leq \rho + 1$$

and thus $L \leq \rho + 1$ and consequently

$$\mathrm{Rad}_n(\mathcal{H}) \leq (\rho + 1)\mathrm{Rad}_n(\mathcal{F}_\rho). \tag{4}$$

Hence, it remains to estimate the Rademacher complexity of the linear function class $\mathcal{F}_\rho$. We begin to compute

$$\begin{aligned} n\mathrm{Rad}_n(\mathcal{F}_\rho) &= \mathbb{E}_{\varepsilon,S}\left[\sup_{\|\theta\|_2 \leq \rho} \sum_{i=1}^{n} \varepsilon_i \theta^\top x_i \right] \\ &= \mathbb{E}_{\varepsilon,S}\left[\sup_{\|\theta\|_2 \leq \rho} \theta^\top \sum_{i=1}^{n} \varepsilon_i x_i \right] \\ &= \rho \mathbb{E}_{\varepsilon,S}\left[\left\| \sum_{i=1}^{n} \varepsilon_i x_i \right\|_2 \right]. \\ &\leq \mathbb{E}_{\varepsilon}\left[\left\| \sum_{i=1}^{n} \varepsilon_i x_i \right\|_2^2 \right]^{\frac{1}{2}}, \end{aligned} \tag{5}$$

where we used Jensen's inequality. Further, using the independence of the Rademacher variables ε_i we obtain

$$\mathbb{E}_\varepsilon \left[\left\| \sum_{i=1}^n \varepsilon_i x_i \right\|_2^2 \right] = \sum_{i,j=1}^n x_i^\top x_j \mathbb{E}[\varepsilon_i \varepsilon_j] = \sum_{i=1}^n \|x_i\|_2 \leq n.$$

Combining this with (5) we obtain

$$\mathrm{Rad}_n(\mathcal{F}_\rho) \leq \frac{\rho}{\sqrt{n}}$$

and thus $\mathrm{Rad}_n(\mathcal{H}) \leq \frac{\rho(\rho+1)}{\sqrt{n}}$. Putting everything together, the bound (3) takes the form

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

Finally, we can estimate

$$\frac{(\rho+1)^2}{2} = \frac{\rho^2 + 2\rho + 1}{2} \leq \frac{\rho^2 + \frac{1}{2}(2^2 + \rho^2) + 1}{2} \leq \frac{2\rho^2 + 4}{2} \leq 2(\rho^2 + 1).$$

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

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

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

Solution: See for example [2].

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

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

for some $\rho > 0$. Show that

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

Solution: Using the reproducing property of K, we estimate

$$\begin{aligned}
\mathrm{Rad}_n(\mathcal{H}_\rho) &= \mathbb{E}_{S,\epsilon} \left[\sup_{h \in \mathcal{H}_\rho} \frac{1}{n} \sum_{i=1}^n \epsilon_i f(z_i) \right] \\
&= \mathbb{E}_{S,\epsilon} \left[\sup_{h \in \mathcal{H}_\rho} \left\langle \frac{1}{n} f, \sum_{i=1}^n \epsilon_i K(z_i, \cdot) \right\rangle_{\mathcal{H}} \right] \\
&\leq \mathbb{E}_{S,\epsilon} \left[\sup_{h \in \mathcal{H}_\rho} \frac{\|f\|_{\mathcal{H}}}{n} \cdot \left\| \sum_{i=1}^n \epsilon_i K(z_i, \cdot) \right\|_{\mathcal{H}} \right] \\
&= \frac{\rho}{n} \cdot \mathbb{E}_{S,\epsilon} \left[\left\| \sum_{i=1}^n \epsilon_i K(z_i, \cdot) \right\|_{\mathcal{H}} \right],
\end{aligned} \tag{6}$$

where we used Cauchy-Schwarz. By Jensen's inequality and $\langle K(z_i,\cdot), K(z_j,\cdot)\rangle_{\mathcal{H}} = K(z_i,z_j)$, we estimate further

$$\begin{aligned}
\mathbb{E}_{S,\epsilon}\left[\left\|\sum_{i=1}^n K(z_i,\cdot)\right\|_{\mathcal{H}}\right] &\leq \sqrt{\mathbb{E}_{S,\epsilon}\left[\left\|\sum_{i=1}^n K(z_i,\cdot)\right\|_{\mathcal{H}}^2\right]} \\
&= \sqrt{\mathbb{E}_{S,\epsilon}\left[\sum_{i,j=1}^n \epsilon_i\epsilon_j K(z_i,z_j)\right]} \\
&= \sqrt{\mathbb{E}_{S}\left[\sum_{i,j=1}^n \mathbb{E}_\epsilon[\epsilon_i\epsilon_j] K(z_i,z_j)\right]} \\
&= \sqrt{\mathbb{E}_{S}\left[\sum_{i=1}^n K(z_i,z_i)\right]} \\
&= \sqrt{n}\cdot\sqrt{\mathbb{E}\left[K(z,z)\right]},
\end{aligned} \tag{7}$$

where we also used the Rademacher property $\mathbb{E}_\epsilon[\epsilon_i\epsilon_j] = \delta_{ij}$. Note that (6) and (7) together yield the result.

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., $g^\star := \inf_{\theta\in\mathbb{R}^d} g(\theta) > -\infty$ and satisfies the following p-PL inequality

$$\|\nabla g(\theta)\|_2^p \geq 2\mu(g(\theta) - g^\star) \quad \text{for all } \theta \in \mathbb{R}^d, \tag{8}$$

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

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

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 $p > 2$?

Solution: By the descent lemma for β-smooth functions and the p-PL inequality, we have

$$\begin{aligned} g(\theta_{k+1}) - g^\star &\leq g(\theta_k) - g^\star - \frac{1}{2\beta}\|\nabla g(\theta_k)\|_2^2 \\ &\leq g(\theta_k) - g^\star - \frac{1}{2\beta} \cdot (2\mu(g(\theta_k) - g^\star))^{\frac{2}{p}} \\ &= \left(1 - \frac{(2\mu)^{\frac{2}{p}}}{2\beta} \cdot (g(\theta_k) - g^\star)^{\frac{2}{p}-1}\right) \cdot (g(\theta_k) - g^\star). \end{aligned}$$

We set $T_\varepsilon \coloneqq \inf\{k \in \mathbb{N} : g(\theta_k) - g^\star \leq \varepsilon\}$. For $k \geq T_\varepsilon$ we have $g(\theta_k) - g^\star \leq \varepsilon$ and for $k < T_\varepsilon$

$$g(\theta_{k+1}) - g^\star \leq \left(1 - \frac{(2\mu)^{\frac{2}{p}}}{2\beta} \cdot \varepsilon^{\frac{2}{p}-1}\right) \cdot (g(\theta_k) - g^\star) \leq \left(1 - \frac{(2\mu)^{\frac{2}{p}}}{2\beta} \cdot \varepsilon^{\frac{2}{p}-1}\right)^{k+1} (g(\theta_0) - g^\star).$$

Hence, for all $k \in \mathbb{N}$ we have

$$g(\theta_k) - g^\star \leq \max\left\{\varepsilon, \left(1 - \frac{(2\mu)^{\frac{2}{p}}}{2\beta} \cdot \varepsilon^{\frac{2}{p}-1}\right)^{k} (g(\theta_0) - g^\star)\right\}.$$

Note that this implies that $g(\theta_k) - g^\star \leq \varepsilon$ if

$$\varepsilon \geq \left(1 - \frac{(2\mu)^{\frac{2}{p}}}{2\beta} \cdot \varepsilon^{\frac{2}{p}-1}\right)^{k} (g(\theta_0) - g^\star). \tag{9}$$

For notational simplicity, set $c_1 \coloneqq \frac{(2\mu)^{\frac{2}{p}}}{2\beta}$ and $c_2 \coloneqq g(\theta_0) - g^\star$. We start solving the inequality (9) for k and take the logarithm which yields

$$\log(c_2^{-1}\varepsilon) \geq k \log(1 - c_1\varepsilon^{\frac{2}{p}-1}),$$

which is well defined if $c_1\varepsilon^{\frac{2}{p}-1} < 1$ or equivalently $\varepsilon < c_1^{-\frac{2-p}{p}}$. Note that $\log(1 - c_1\varepsilon^{\frac{2}{p}-1}) < 0$ and hence, we obtain

$$k \geq \frac{\log(c_2^{-1}\varepsilon)}{\log(1 - c_1\varepsilon^{\frac{2}{p}-1})}. \tag{10}$$

Next, we use $\log(1-t) \leq -t$ for $t \in (0,1)$ as well as $\log(c_2^{-1}\varepsilon) < 0$ for $\varepsilon < c_2^{-1}$ small, which implies that (10) is satisfied if

$$k \geq \frac{\log(c_2\varepsilon^{-1})}{c_1\varepsilon^{\frac{2}{p}-1}} = \frac{\log(\varepsilon^{-1}) + \log(c_2)}{c_1\varepsilon^{\frac{2}{p}-1}}. \tag{11}$$

Finally, for $\varepsilon < c_2^{-1}$ we have $\log(\varepsilon^{-1}) + \log(c_2) \leq 2\log(\varepsilon^{-1})$ and hence inequality (11) is satisfied if

$$k \geq \frac{2}{c_2} \cdot \frac{\log(\varepsilon^{-1})}{\varepsilon^{\frac{2}{p}-1}}. \tag{12}$$

Overall, we have shown (12) under the assumption that

$$\varepsilon < \min\left\{c_1^{-\frac{2-p}{p}}, c_2^{-1}\right\}.$$

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.

[2] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.