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 5

Due: Tue., Jan. 23th, 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. (Rademacher complexity of ReLU networks with NTK parametrization) Consider a shallow ReLU network

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

with NTK parametrization of width m and consider the restricted class

$$\mathcal{F}_{\rho,m} \coloneqq \left\{ F(\cdot; w, c) : \max_{1\le k\le m} \|w_i - w(0)_i\|_2 \le \frac{\rho}{\sqrt{m}} \right\}$$

for some $w(0) \in \mathbb{R}^{md}$ and $c \in \mathbb{R}^m$ and $\delta \in (0,1)$. Show that there is a constant $\kappa \ge 2$ with probability at least $1-\delta$ it holds that

$$\widehat{\mathrm{Rad}}_S(\mathcal{F}_{\rho,m}) \le \frac{\rho}{\sqrt{n}} + \frac{\kappa\rho}{\sqrt{m}} \left(\rho + \sqrt{\log\left(\frac{4n}{\rho}\right)} \right)$$

Hint: Theorem 1 of the lecture on linearization might be helpful.

Q2. (Generalization bound for projected SGLD) 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 $\theta \in \mathbb{R}^{d_f}$. We fix a data generating distribution P on $\mathbb{X} \times \mathbb{R}$ such that $P(\|\Phi(x)\|_2 \le 1 \text{ and } |y| \le 1) = 1$ and consider a training set $S = ((x_i, y_i))_{i=1,\ldots,n} \subseteq \mathbb{X} \times \mathbb{R}$ consisting of iid samples from P. Further, we consider the l^2 sample loss $\ell(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$ and the empirical risk

$$g(\theta) = \hat{\mathcal{R}}_S(f_\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_\theta(x_i), y_i).$$

We fix $R > 0$ and denote the Eulidean projection onto the ball $B_R(0) = \{x \in \mathbb{R}^d : \|x\|_2 \le R\}$ by $\Pi_{B_R(0)}$. We onsider the projected stochastic gradient Langevin dynamics (SGLD) given by

$$\begin{aligned} \widetilde{\theta}_{t+1} &= \theta_t - \eta_t \nabla_\theta \ell(f_{\theta_t}(x_{J_t}), y_{J_t}) + \xi_t, \\ \theta_{t+1} &= \Pi_{B_R(0)}(\widetilde{\theta}_{t+1}), \end{aligned}$$

where $(J_t)_{t\in\mathbb{N}} \subseteq \{1, \ldots, n\}$ is an iid sequence of uniformly selected indices and $(\xi_t)_{t\in\mathbb{N}}$ is a sequence independent of $(J_t)_{t\in\mathbb{N}}$ of independent Gaussian random variables with $\xi_t \sim \mathcal{N}(0; \rho_t^2 I_d)$. Show that for the average iterate $\overline{\theta}_T \coloneqq \frac{1}{T}\sum_{t=1}^{T} \theta_t$ it holds that

$$\left| \mathbb{E}_{S,J,\xi} \left[\hat{\mathcal{R}}_S(f_{\overline{\theta}_T}) - \mathcal{R}(f_{\overline{\theta}_T}) \right] \right| \le \frac{(R+1)^2}{2} \cdot \sqrt{\frac{1}{n} \sum_{t=1}^{T} \frac{\eta_t^2}{\rho_t^2}}.$$

Remark. Note that increasing the noise level ρ_t^2 improves the generalization.

Q3. (Optimization guarantee for projected SGLD) Consider the setting and projected SGLD of **Q2** and consider a constant step size η and noise variance ρ. Show that

$$\mathbb{E}\left[\hat{\mathcal{R}}_S(f_{\overline{\theta}_T}) - \inf_{\theta\in B_R(0)} \hat{\mathcal{R}}_S(f_\theta)\right] \le \frac{2R^2}{\eta T} + \frac{\eta(R+1)^2}{2} + \frac{\rho^2}{2\eta}.$$

Remark. Note that increasing the noise level ρ hurts the optimization.

Q4. (Risk bound for projected SGLD) We continue the discussion of **Q2** and **Q3** and assume realizability, i.e., assume the existence of a parameter $\theta^\star \in B_R(0)$ such that $\mathcal{R}(f_{\theta^\star}) = 0$. Show that

$$\mathbb{E}_{S,J,\xi}[\mathcal{R}(f_{\theta_T})] \le \frac{(R+1)^2}{2}\sqrt{\frac{T}{n}\cdot\frac{\eta}{\rho}} + \frac{2R^2}{\eta T} + \frac{\eta(R+1)^2}{2} + \frac{\rho^2}{2\eta}. \tag{1}$$

Further, show that if $T = n^\alpha, \eta = n^\beta, \rho = n^\gamma$ the right hand side of (1) is lower bounded (up to positive constants) by $n^{-\frac{1}{4}}$. Finally, show that for a specific choice of α, β and γ we have

$$\mathbb{E}_{S,J,\xi}[\mathcal{R}(f_{\theta_T})] \le O(n^{-\frac{1}{4}}).$$

Q5. (Fast rates via Tikhonov regularization) Assume an L-Lipschitz-continuous convex sample loss ℓ and linear prediction functions with $\mathcal{F} = \{f_\theta(x) = \theta^\top\phi(x), \theta \in \mathbb{R}^d\}$, where $\|\phi(x)\|_2 \le R$. Let $\hat{\theta}_\lambda \in \mathbb{R}^d$ be the minimizer of the regularized empirical risk

$$\hat{\mathcal{R}}_S(f_\theta) + \frac{\lambda}{2}\cdot\|\theta\|_2^2.$$

Show that

$$\mathbb{E}\left[\mathcal{R}(f_{\hat{\theta}_\lambda})\right] \le \inf_{\theta\in\mathbb{R}^d}\left\{\mathcal{R}(f_\theta) + \frac{\lambda}{2}\|\theta\|_2^2\right\} + \frac{32L^2R^2}{\lambda n}. \tag{2}$$

For this, you can proceed in the following steps, where $\mathcal{R}_\lambda(f_\theta) \coloneqq \mathcal{R}(f_\theta) + \frac{\lambda}{2}\|\theta\|_2^2$ denotes the regularized risk with optimal value $\mathcal{R}_\lambda^\star$ attained at $\theta_\lambda^\star$:

(a) For $\varepsilon > 0$, show that

$$C_\varepsilon \coloneqq \left\{\theta \in \mathbb{R}^d : \mathcal{R}_\lambda(\theta) - \mathcal{R}_\lambda^\star \le \varepsilon\right\} \subseteq B_r(\theta_\lambda^\star)$$

for $r = \sqrt{\frac{2\varepsilon}{\lambda}}$. Further, show that

$$\mathbb{P}(\mathcal{R}_\lambda(f_{\hat{\theta}_\lambda}) - \mathcal{R}_\lambda^\star > \varepsilon) \le \mathbb{P}\left(\sup_{\theta\in B_r(\theta_\lambda^\star)}\left\{\mathcal{R}_\lambda(f_\theta) - \mathcal{R}_\lambda^\star - (\hat{\mathcal{R}}_\lambda(f_\theta) - \hat{\mathcal{R}}_\lambda(f_{\theta_\lambda^\star})\right\} \ge \varepsilon\right).$$

(b) Show that

$$\mathbb{E}\left[\sup_{\theta\in B_r(\theta_\lambda^\star)}\left\{\mathcal{R}_\lambda(f_\theta) - \mathcal{R}_\lambda^\star - \left(\hat{\mathcal{R}}_\lambda(f_\theta) - \hat{\mathcal{R}}_\lambda(f_{\theta_\lambda^\star})\right)\right\}\right] \le 2LR\sqrt{\frac{2\varepsilon}{n\lambda}}.$$

Remark: You can use (without proof) the generalization bound in expectation

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{n}\sum_{i=1}^n h(z_i) - \mathbb{E}[h(z)]\right)\right] \le 2\,\mathrm{Rad}_n(\mathcal{H}).$$

Then show that for a linear model with bounded parameters $\mathcal{F}_\rho = \{f_\theta : \|\theta\|_2 \le \rho\}$ and bounded features $\|\Phi\|_2 \le R$ it holds that $\mathrm{Rad}_n(\mathcal{F}_\rho) \le \frac{R\rho}{\sqrt{n}}$.

(c) Use McDiamird's inequality to show

$$\mathbb{P}(\mathcal{R}_\lambda(f_{\hat{\theta}_\lambda}) - \mathcal{R}_\lambda^\star > \varepsilon) \leq e^{-t^2} \quad \text{for } t > 0$$

if $\varepsilon \geq 8\frac{L^2R^2}{\lambda n}(2+t^2)$. Use this to conclude the proof.

Remark: You can use a one-sided McDiamird inequality without proof.

Remark: Compare the $O(\frac{1}{n})$ guarantee to the $O(\frac{1}{\sqrt{n}})$ bound for a constrained linear model given in **Q5** of Assignment 4. However, note that we make a regularization error.

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

Q6. (Bonus: Covering number of Lipschitz functions) Consider the set of pinned Lipschitz functions

$$\mathcal{F} = \Big\{f\colon [a,b] \to \mathbb{R} : f(a) = 0, |f(t) - f(s)| \leq L|t-s| \text{ for all } t,s \in [a,b]\Big\}$$

for some $L > 0$ and some $a < b$, where $a, b \in \mathbb{R}$. We consider the uniform norm

$$\|f - g\|_\infty := \sup_{t \in [a,b]} |f(t) - g(t)|$$

and the covering number $\mathcal{N}(\mathcal{F}, \varepsilon, \|\cdot\|_\infty)$. Show that

$$\log_2 \mathcal{N}(\mathcal{F}, \varepsilon, \|\cdot\|_\infty) = \left\lceil \frac{(b-a)L}{\varepsilon} \right\rceil,$$

where $\lceil x \rceil$ denotes the smallest integer not smaller than x.

Hint. Consider piecewise linear functions on a fixed grid with slope $\pm L$ in every linear region.

Bonus (2 points): Give (essentially matching) upper and lower bounds on the covering number of

$$\mathcal{F}_R = \Big\{f\colon [a,b] \to \mathbb{R} : \|f\|_\infty \leq R, |f(t) - f(s)| \leq L|t-s| \text{ for all } t,s \in [a,b]\Big\}?$$