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 $\delta \in (0,1)$ and assume that $S = (x_i)_{i=1}^n$ with $\|x_i\|_2 \le 1$ for all $i = 1, \dots, n$. Show that with probability at least $1 - \delta$ it holds that

$$\widehat{\operatorname{Rad}}_S(\mathcal{F}_{\rho,m}) \le \frac{\rho \|c\|_\infty}{\sqrt{n}} + \frac{2\rho}{\sqrt{m}} \left(\rho + \sqrt{\log\left(\frac{1}{\rho}\right)} \right).$$

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

Solution: We define the ball

$$B_{\rho,m} \coloneqq \left\{ w \in \mathbb{R}^{md} : \max_{1 \le k \le m} \|w_i - w(0)_i\|_2 \le \frac{\rho}{\sqrt{m}} \right\}.$$

By Theorem 1 of the lecture on linearization we have for any $\|x\|_2 \le 1$ and $w \in B_{\rho,m}$ that

$$\begin{aligned}
\left| F(x; w, c) - \nabla_w F(x; w(0), c)^\top (w - w(0)) \right| &\le \frac{\rho}{\sqrt{m}} (1 + \|x\|_2) \left(\rho \|x\|_2 + \sqrt{\log\left(\frac{1}{\delta}\right)} \right) \\
&\le \frac{2\rho}{\sqrt{m}} \left(\rho + \sqrt{\log\left(\frac{1}{\delta}\right)} \right)
\end{aligned}$$

with probability at least $1 - \delta$. Now, we can estimate

$$\begin{aligned}
n\widehat{\operatorname{Rad}}_S(\mathcal{F}_{\rho,m}) &= \mathbb{E}_\varepsilon \left[\sup_{w \in B_{\rho,m}} \sum_{i=1}^n \varepsilon_i F(x_i; w, c) \right] \\
&\le \frac{2\rho}{\sqrt{m}} \left(\rho + \sqrt{\log\left(\frac{1}{\delta}\right)} \right) + \mathbb{E}_\varepsilon \left[\sup_{w \in B_{\rho,m}} \sum_{i=1}^n \varepsilon_i \nabla_w F(x_i; w(0), c)^\top (w - w(0)) \right] \\
&= \frac{2\rho}{\sqrt{m}} \left(\rho + \sqrt{\log\left(\frac{1}{\delta}\right)} \right) + \mathbb{E}_\varepsilon \left[\sup_{w \in B_{\rho,m}} (w - w(0))^\top \sum_{i=1}^n \varepsilon_i \nabla_w F(x_i; w(0), c) \right].
\end{aligned}$$

Note that by Cauchy-Schwarz we can estimate the second part according to

$$
\begin{aligned}
(w-w(0))^\top \sum_{i=1}^n \varepsilon_i \nabla_w F(x_i; w(0), c) &= \frac{1}{\sqrt{m}} \sum_{k=1}^m (w_k - w(0)_k)^\top \sum_{i=1}^n \varepsilon_i x_i c \mathbb{1}\{w(0)_k^\top x_i\} \\
&\leq \frac{\rho}{m} \sum_{k=1}^m \left\| \sum_{i=1}^n \varepsilon_i x_i c_k \mathbb{1}\{w(0)_k^\top x_i\} \right\|_2 .
\end{aligned}
$$

Using this in the estimate above, we obtain

$$
\begin{aligned}
n\widehat{\mathrm{Rad}}_S(\mathcal{F}_{\rho,m}) &\leq \frac{2\rho}{\sqrt{m}} \left(\rho + \sqrt{\log\left(\frac{1}{\delta}\right)}\right) + \frac{\rho}{m} \sum_{k=1}^m \mathbb{E}_\varepsilon \left[\left\| \sum_{i=1}^n \varepsilon_i x_i c_k \mathbb{1}\{w(0)_k^\top x_i\} \right\|_2 \right] \\
&\leq \frac{2\rho}{\sqrt{m}} \left(\rho + \sqrt{\log\left(\frac{1}{\delta}\right)}\right) + \frac{\rho}{m} \sum_{k=1}^m \sqrt{\mathbb{E}_\varepsilon \left[\left\| \sum_{i=1}^n \varepsilon_i x_i c_k \mathbb{1}\{w(0)_k^\top x_i\} \right\|_2^2 \right]} \\
&= \frac{2\rho}{\sqrt{m}} \left(\rho + \sqrt{\log\left(\frac{1}{\delta}\right)}\right) + \frac{\rho}{m} \sum_{k=1}^m \sqrt{\mathbb{E}_\varepsilon \left[\sum_{i=1}^n \left\| \varepsilon_i x_i c_k \mathbb{1}\{w(0)_k^\top x_i\} \right\|_2^2 \right]} \\
&\leq \frac{2\rho}{\sqrt{m}} \left(\rho + \sqrt{\log\left(\frac{1}{\delta}\right)}\right) + \frac{\rho \|c\|_\infty}{m} \sum_{k=1}^m \sqrt{\mathbb{E}_\varepsilon \left[\sum_{i=1}^n \|x_i\|_2^2 \right]} \\
&\leq \frac{2\rho}{\sqrt{m}} \left(\rho + \sqrt{\log\left(\frac{1}{\delta}\right)}\right) + \sqrt{n}\rho\|c\|_\infty .
\end{aligned}
$$

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 $P(\|\Phi(x)\|_2 \leq 1 \text{ and } |y| \leq 1) = 1$ and consider a training set $S = ((x_i, y_i))_{i=1,\dots,n} \subseteq \mathbb{X} \times \mathbb{R}$ 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 $B_R(0) = \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$ 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 $(J_t)_{t\in\mathbb{N}} \subseteq \{1, \dots, n\}$ 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 $\overline{\theta}_T := \frac{1}{T}\sum_{t=1}^T \theta_t$ it holds that

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

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

Solution: We want to apply the theorem from the lecture and use the notation there. However, we work with the iterates $\widetilde{\theta}_t$ rather than with θ_t. Note that we can express the average iterate as

$$\overline{\theta}_t = \frac{1}{T}\sum_{t=1}^{T} \Pi(\widetilde{\theta}_t) = F(\widetilde{\theta}_1, \ldots, \widetilde{\theta}_T).$$

Note that the iterates $\widetilde{\theta}_t$ satisfy the recursion

$$\widetilde{\theta}_{t+1} = \Pi_{B_R(0)}(\widetilde{\theta}_t) - \eta_t \psi(\widetilde{\theta}_t, z_{J_t}) + \xi_t,$$

where

$$\psi(\theta, z) := (\Pi_{B_R(0)}(\theta)^\top \Phi(x_{J_t}) - y_{J_t})\Phi(x_{J_t}),$$

for $z = (x, y)$. The assumption on the samplign strategy is clearly satisfied (same strategy as in the Corollary after Theorem 2). Further, we have

$$\|\psi(\theta, z)\| = |(\Pi_{B_R(0)}(\theta)^\top \Phi(x_{J_t}) - y_{J_t})| \cdot \|\Phi(x_{J_t})\| \leq |(\Pi_{B_R(0)}(\theta)^\top \Phi(x_{J_t})| + |y_{J_t})| \leq R + 1$$

for any $\theta \in \mathbb{R}^{d_f}$ and $z \in \mathbb{Z}$. Further, we check the sub-Gaussianity and estimate

$$|\ell(f_{F(\underline{\theta})}(x), y)| = \frac{1}{2}(F(\underline{\theta})^\top \Phi(x) - y)^2 \leq \frac{(R+1)^2}{2}$$

for all $\underline{\theta} = (\theta_1, \ldots, \theta_T)$. Thus, by Hoeffding's lemma, $\ell(f_{F(\underline{\theta})}(x), y)$ is sub-Gaussian with parameter $\sigma^2 = \frac{(R+1)^2}{4}$. Using Theorem 2 from the lecture, we obtain

$$\left|\mathbb{E}_{S,J,\xi}\left[\hat{\mathcal{R}}_S(f_{\overline{\theta}_T}) - \mathcal{R}(f_{\overline{\theta}_T})\right]\right| \leq \sqrt{\frac{(R+1)^4}{4n}\sum_{t=1}^{T}\frac{\eta_t^2}{\rho_t^2}}$$

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

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

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

Solution: Let us set $u_t := \nabla_\theta \ell(f_\theta(x_{J_t}), y_{J_t})$ then just like in the proof of the convergence results of projected SGD we use the Lyapunov function $\mathcal{L}(\theta) := \|\theta - \theta^\star\|_2^2$ for an optimizer $\theta^\star \in B_R(0)$. Note that this exists since L is a continuous function. Now we can estimate

$$\begin{aligned}
\mathcal{L}(\theta_{t+1}) &\leq \|\theta_t - \eta_t u_t + \xi_t - \theta^\star\|_2^2 \\
&= \mathcal{L}(\theta_t) - 2\eta u_t^\top(\theta_t - \theta^\star) + \eta^2\|u_t\|_2^2 - 2\eta u_t^\top \xi_t + \|\xi_t\|_2^2.
\end{aligned}$$

Note that $\|u_t\|_2 = \|(\theta_t^\top \Phi(x_{J_t}) - y_{J_t})\Phi(x_{J_t})\|_2 \leq |\theta_t^\top \Phi(x_{J_t})| + |y_{J_t}| \leq R + 1$. Taking the conditional expectation $\mathbb{E}[\cdot|\theta_t]$, using $\mathbb{E}[u_t|\theta_t] = \nabla g(\theta_t)$, and $\mathbb{E}[u_t^\top \xi_t|\theta_t] = \mathbb{E}[u_t|\theta_t]^\top \mathbb{E}[\xi_t|\theta_t] = 0$ and the convexity of L we obtain

$$\mathbb{E}[\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t)|\theta_t] \leq 2\eta(g(\theta_t) - g^\star) + \eta^2(R+1)^2 + \rho^2.$$

Taking the expectation yields

$$\mathbb{E}[\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t)] \le 2\eta\mathbb{E}[g(\theta_t) - g^\star] + \eta^2(R+1)^2 + \rho^2.$$

Summing over t, rearranging and using the convexity of g yields

$$\mathbb{E}[g(\overline{\theta}_T)] - g^\star \le \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} g(\theta_t) - g^\star\right] \le \frac{\mathcal{L}(\theta_0)}{2\eta T} + \frac{\eta(R+1)^2}{2} + \frac{\rho^2}{2\eta} \le \frac{2R^2}{\eta T} + \frac{\eta(R+1)^2}{2} + \frac{\rho^2}{2\eta}.$$

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

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

Solution: First, we note that $\hat{\mathcal{R}}(f_{\theta^\star}) = 0$ and hence $\inf_{\theta\in B_R(0)} \hat{\mathcal{R}}_S(f_\theta)$. Now, (1) is a direct consequence of the previous two questions. If $T = n^\alpha, \eta = n^\beta, \rho = n^\gamma$, then the right hand side of (1) behaves (up to constants) like n^κ for

$$\kappa \coloneqq \max\{\alpha/2 + \beta - \gamma - 1/2, -\alpha - \beta, \beta, -\beta + 2\gamma\}.$$

Note that $\beta \le \kappa$, $\alpha \ge -\kappa - \beta$ and $2\gamma \le \kappa + \beta$. This implies

$$\kappa \ge \alpha/2 + \beta - \gamma - 1/2 \ge -\frac{\kappa}{2} - \frac{\beta}{2} + \beta - \frac{\kappa}{2} - \frac{\beta}{2} - \frac{1}{2} = -\kappa - \frac{1}{2}$$

and hence $\kappa \ge -\frac{1}{4}$. Further, for $\alpha = \frac{1}{2}, \beta = -\frac{1}{4}, \gamma \in [-\frac{1}{4}, 0]$ we have $\kappa = -\frac{1}{4}$.

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

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

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

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

(b) Show that

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

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

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

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

(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.

Solution: See Proposition 4.6. in [1].

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

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

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

$$\left\lceil \frac{(b-a)L}{\varepsilon} \right\rceil - 1 \leq \log_2 \mathcal{N}(\mathcal{F}, \varepsilon, \|\cdot\|_\infty) \leq \left\lceil \frac{(b-a)L}{\varepsilon} \right\rceil,$$

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

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

Solution: Let without loss of generality $[a,b] = [0,1]$ as the general statement follows via an easy transformation. We set $n := \lceil \frac{L}{\varepsilon} \rceil$ and $h := n^{-1}$ and consider the following set of functions

$$\mathcal{G} := \Big\{ g \in \mathcal{F} : g \text{ is linear with slope } \pm L \text{ on } [x_k, x_{k+1}] \text{ for every } k = 0, \dots, n-1 \Big\},$$

where $x_k := kh$. It is clear from the definition that $|\mathcal{G}| = 2^n$. It remains to show that $\mathcal{G}$ is a minimal ε-cover of $\mathcal{F}$.

First, we show that it ε-covers $\mathcal{F}$ for which we fix a function $f \in \mathcal{F}$. We construct $g \in \mathcal{G}$ via the recursion

$$g(x_{k+1}) \in \arg\min \Big\{ |f(x_{k+1}) - y| : y \in \big\{ g(x_k) \pm hL \big\} \Big\}. \tag{3}$$

where we interpolate linearly between the points x_k. It is immediate that $g \in \mathcal{G}$. First, we show inductively, that $|f(x_k) - g(x_k)| \le hL$ for all $k = 0, \dots, n$. For $k = 0$, we have $f(0) = g(0) = 0$. Further if $|f(x_k) - g(x_k)| \le hL$, then surely

$$f(x_{k+1}) \le f(x_k) + hL \le g(x_k) + 2hL$$

and similarly $f(x_{k+1}) \ge g(x_k) - 2hL$. Hence, we have

$$f(x_{k+1}) \in [g(x_k) - 2hL, g(x_k) + 2hL]$$

which ensures

$$|f(x_{k+1}) - g(x_{k+1})| \le hL.$$

Further, for $x \in [0,1]$, we have $x \in [x_k, x_{k+1}]$. Let us assume without loss of generality that $g(x_{k+1}) = g(x_k) + hL$, then

$$f(x) \le f(x_k) + L|x - x_k| \le g(x_k) + hL + L|x - x_k| = g(x) + hL$$

and similarly

$$f(x) \ge f(x_{k+1}) - L|x - x_k| \le g(x_{k+1}) - hL - L|x - x_k| = g(x) - hL.$$

Overall, this shows that

$$\|f - g\|_\infty \le hL \le \varepsilon$$

and consequently

$$\log_2 \mathcal{N}(\mathcal{F}, \varepsilon, \|\cdot\|_\infty) \le \left\lceil \frac{L}{\varepsilon} \right\rceil.$$

For the lower bound, we consider again a class $\mathcal{H}$ of piecewise linear functions like above, however with $h = m^{-1}$ for $m := \lceil \frac{L}{\varepsilon} \rceil - 1$. Then $|\mathcal{H}| = 2^m$ and for $h_1 \ne h_2 \in \mathcal{H}$ we have

$$\|h_1 - h_2\|_\infty \ge 2hL > 2\varepsilon.$$

In particular, this shows that any ε-net of $\mathcal{F}$ has to contain at least 2^m elements.

For a generalization to arbitrary Lipschitz classes, we refer to [**?**].

References

[1] Francis Bach. Learning theory from first principles. *Draft of a book, version of Sept*, 6:2021, 2021.