Foundations of Deep Learning **Fall 2024**

Exercise 7: Generalization I

Lecturer: Aurelien Lucchi

Problem 1 (Rademahcer Complexity):

Given a sample $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ drawn from some distribution $\mathcal{D}$, the empirical Rademacher complexity $\hat{\mathcal{R}}(\mathcal{F})$ of the hypothesis class $\mathcal{F}$ of binary classifier is defined as:

$$\hat{\mathcal{R}}(\mathcal{F}) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(\mathbf{x}_i)\right],$$

where $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_n)$ are independent Rademacher random variables, i.e., $\mathbb{P}(\sigma_i = +1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}$. The Rademacher complexity quantifies the richness of the function class $\mathcal{F}$ by measuring how well functions in $\mathcal{F}$ can fit random noise. It is widely used to provide generalization bounds in machine learning: with probability at least $1-\delta$ over sample draws, it holds that

$$\sup_{f\in\mathcal{F}} \left| \mathbb{E}_{x\sim\mathcal{D}}[f(x)] - \frac{1}{n}\sum_{i=1}^{n} f(x_i) \right| \leq 2\hat{\mathcal{R}}(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{n}}.$$

Now, let's consider the following problem, which involves proving some important properties of Rademacher complexity.

a) For two hypothesis classes $\mathcal{F}, \mathcal{F}'$, prove that the Rademacher complexity of their sum satisfies the following inequality:

$$\hat{\mathcal{R}}(\mathcal{F} + \mathcal{F}') \leq \hat{\mathcal{R}}(\mathcal{F}) + \hat{\mathcal{R}}(\mathcal{F}').$$

b) Consider an L-Lipschitz continuous function $\phi : \mathbb{R} \to \mathbb{R}$, i.e. for all $t, t' \in \mathbb{R}$,

$$|\phi(t) - \phi(t')| \leq L|t - t'|,$$

and define a new class $\mathcal{F}' := \{\phi \circ f : f \in \mathcal{F}\}$. Prove that the Rademacher complexity of $\mathcal{F}'$ is bounded as follows:

$$\hat{\mathcal{R}}(\mathcal{F}') \leq L\hat{\mathcal{R}}(\mathcal{F}).$$

c) Let $\mathcal{F}$ be a class of real-valued functions and let $\ell(x) = \min(1, \max(0, x))$ be the hinged loss. Show that for any $f \in \mathcal{F}$, with probability at least $1-\delta$ over the sample draw, it holds that

$$\mathbb{E}[\ell(f(x), y)] \leq \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i) + 2\hat{\mathcal{R}}(\mathcal{F}) + 3\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Problem 2 (Concentration of NTK Eigenvalues):
Consider a set of examples $\{\mathbf{x}_i\}_{i=1}^n$ and let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a data matrix containing the $\mathbf{x}_i$ as its rows. We consider a two-layer neural network of the following form,

$$f(\mathbf{W}, \mathbf{a}, \mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} a_k \sigma(\mathbf{w}_k^\top \mathbf{x}),$$

where we assume that $|a_k| \leq 1$ and the activation function σ is differentiable and $|\sigma(\cdot)| \leq B$.

We define the sampled Gram NTK matrix $\hat{\mathbf{G}} = (\hat{G}_{ij})_{i,j} \in \mathbb{R}^{n\times n}$ and the expected Gram NTK matrix $\mathbf{G} = (G_{ij})_{i,j} \in \mathbb{R}^{n\times n}$ as follows

$$\hat{G}_{ij} := \frac{1}{m} \sum_{k=1}^{m} \mathbf{x}_i^\top \mathbf{x}_j \sigma'(\mathbf{w}_k^\top \mathbf{x}_i)\sigma'(\mathbf{w}_k^\top \mathbf{x}_j), \quad G_{ij} := \mathbb{E}_{\mathbf{w}} \mathbf{x}_i^\top \mathbf{x}_j \sigma'(\mathbf{w}^\top \mathbf{x}_i)\sigma'(\mathbf{w}^\top \mathbf{x}_j).$$

We can view the matrix $\hat{\mathbf{G}}$ as an average of a set of matrices $\mathbf{H}_1, ..., \mathbf{H}_m$, i.e. $\hat{\mathbf{G}} = \frac{1}{m}\sum_{k=1}^m \mathbf{H}_k$, where

$$(\mathbf{H}_k)_{ij} := \mathbf{x}_i^\top \mathbf{x}_j \sigma'(\mathbf{w}_k^\top \mathbf{x}_i)\sigma'(\mathbf{w}_k^\top \mathbf{x}_j), \quad k = 1, ..., m.$$

The goal of this exercise is to bound the deviation between $\hat{\mathbf{G}}$ and $\mathbf{G}$. To do so, we will use a concentration bound that is based from what we have seen in the main lecture. We will use the Frobenius inner product notation $\langle \mathbf{A}, \mathbf{B}\rangle_F = \operatorname{tr}\left(\mathbf{A}^\top \mathbf{B}\right)$ and note that $|\langle \mathbf{A}, \mathbf{B}\rangle| \leq \|\mathbf{A}\|_F \|\mathbf{B}\|_F$.

a) Prove that $\|\mathbf{H}_k\|_F \leq B^2 \|\mathbf{X}\|_F^2$.

b) Define $\mathcal{F} := \{\mathbf{U} \to \langle \mathbf{U}, \mathbf{V}\rangle_F : \|\mathbf{V}\|_F \leq 1\}$, $\mathcal{H} := (\mathbf{H}_1, \dots \mathbf{H}_m)$. The Rademacher complexity of $\mathcal{F}_{|\mathcal{H}}$ is defined as

$$\mathcal{R}(\mathcal{F}_{|\mathcal{H}}) = \frac{1}{m} \mathbb{E}_\epsilon \sup_{\mathbf{V}} \sum_{i=1}^{m} \epsilon_i \langle \mathbf{H}_i, \mathbf{V}\rangle_F.$$

Prove that $\mathcal{R}(\mathcal{F}_{|\mathcal{H}}) \leq \frac{1}{\sqrt{m}} B^2 \|\mathbf{X}\|_F^2$.

c) Recall the concentration bound based on Rademacher complexities:

Theorem 1. *Let $\mathcal{F} : \mathbb{R}^{n\times n} \to \mathbb{R}$ be a function class such that $\forall f \in \mathcal{F}, f(\mathbf{H}) \in [a, b]$ a.s.. Then with probability at least $1 - \delta$ over the draw of $\mathbf{H}_1, ..., \mathbf{H}_m$,*

$$\sup_{f\in\mathcal{F}} f(\mathbf{H}) - \frac{1}{m}\sum_{k=1}^{m} f(\mathbf{H}_k) \leq 2\mathcal{R}(\mathcal{F}_{|\mathcal{H}}) + 3(b-a)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Prove that with probability at least $1 - \delta$ over the draw of $(\mathbf{w}_1, \dots \mathbf{w}_m)$ for $\|\mathbf{u}\|_2 \leq 1$, it holds that:

$$|\mathbf{u}^\top \mathbf{G}\mathbf{u} - \mathbf{u}^\top \hat{\mathbf{G}}\mathbf{u}| \leq \frac{2B^2\|\mathbf{X}\|_F^2}{\sqrt{m}} + 6B^2\|\mathbf{X}\|_F^2 \sqrt{\frac{\log(2/\delta)}{2m}}. \tag{1}$$

d) Denote $\star := \frac{2B^2\|\mathbf{X}\|_F^2}{\sqrt{m}} + 6B^2\|\mathbf{X}\|_F^2\sqrt{\frac{\log(2/\delta)}{2m}}$ to be the bound in Eq. (1). Prove that with probability at least $1 - \delta$,

$$\lambda_{\min}(\hat{\mathbf{G}}) \geq \lambda_{\min}(\mathbf{G}) - \star \text{ and } \lambda_{\max}(\hat{\mathbf{G}}) \leq \lambda_{\max}(\mathbf{G}) + \star.$$

Hint: Weyl's inequality bounds the eigenvalues of the sum of Hermitian matrices.

Theorem 2 (Weyl's inequality)**.** *Let $\mathbf{A}, \mathbf{B}$ be two Hermitian $n \times n$ matrices. Denote by $\lambda_1(\mathbf{A}) \geq \cdots \geq \lambda_n(\mathbf{A})$ the sorted eigenvalues of $\mathbf{A}$. Then*

$$\lambda_1(\mathbf{A} + \mathbf{B}) \leq \lambda_1(\mathbf{A}) + \lambda_1(\mathbf{B}) \tag{2}$$

and

$$\lambda_n(\mathbf{A} + \mathbf{B}) \geq \lambda_n(\mathbf{A}) + \lambda_n(\mathbf{B}) \tag{3}$$