Instructions.

- Homework is due **Friday, December 18, at 11:59pm**; no late homework accepted.
- You must work individually for this homework.
- Excluding office hours, you may discuss with at most three other people; please state their NetIDs clearly on the first page of your submission.
- Homework must be typed, and submitted via gradescope. Please consider using the provided \LaTeX file as a template.
- Each part of each problem is worth 3 points.
- General course and homework policies are on the course webpage.

Version history.

1. Initial version.

1 + \epsilon. Forgot $x_i$ in p2.
1. Concentration of NTK eigenvalues.

In hw2, we established that the infinite-width Gram matrix from the NTK lectures has positive minimum eigenvalue. In this problem, we will prove that the finite width Gram matrix has similar (positive) eigenvalues. This settles the remaining pieces necessary to make the NTK gradient flow proof concrete.

Throughout this problem let examples $(x_i)_{i=1}^n$ be given and fixed, and collect all examples as rows of a matrix $X \in \mathbb{R}^{n \times d}$. Suppose for simplicity the activation $\sigma$ is differentiable, and $|\sigma'| \leq B$.

Define the sampled Gram matrix $\hat{G} \in \mathbb{R}^{n \times n}$ and expected gram matrix $G \in \mathbb{R}^{n \times n}$ via

$$\hat{G}_{ij} := \frac{1}{m} \sum_{k=1}^{m} x_i^T x_j \sigma'(w_i x_i) \sigma'(w_k x_j), \quad G_{ij} := \mathbb{E}_{w} x_i^T x_j \sigma'(w^T x_i) \sigma'(w^T x_j).$$

In this problem, the random draw is over the weights $(w_j)_{j=1}^m$, not over the examples. Consequently, it is useful to define another family of random matrices: $(H_k)_{k=1}^m$, with

$$(H_k)_{ij} := x_i^T x_j \sigma'(w_i^T x_i) \sigma'(w_k^T x_j), \quad \text{whereby } \hat{G} := \frac{1}{m} \sum_{k=1}^{m} H_k.$$  

The approach of this problem is to apply Rademacher complexity. Let $W = (w_1, \ldots, w_m)$ denote the full random draw, analogous to the data in our usual applications of Rademacher complexity.

We will use matrix inner products:

$$\langle H_k, V \rangle = \text{tr} \left( H_k^T V \right), \quad \text{and} \quad |\langle H_k, V \rangle| \leq \|H_k\|_F \cdot \|V\|_F.$$

(a) Prove $\|H_k\|_F \leq B^2 \|X\|_F^2$.

(b) Define

$$\mathcal{F} := \{U \mapsto \langle U, V \rangle : \|V\|_F \leq 1 \}, \quad \mathcal{H} := (H_1, \ldots, H_m).$$

The relevant (unnormalized) Rademacher complexity for us is

$$\text{URad}(\mathcal{F} | \mathcal{H}) = \text{URad} \left( \{ (\langle H_1, V \rangle, \ldots, \langle H_m, V \rangle) : \|V\|_F \leq 1 \} \right).$$

Prove $\text{URad}(\mathcal{F} | \mathcal{H}) \leq B^2 \|X\|_F^2 \sqrt{m}$.

**Hint.** Revisit how we handled the vector case in lecture (Theorem 21.3 in the typed lecture notes).

(c) Prove $\{ v v^T : \|v\|_2 \leq 1 \} \subseteq \{ V : \|V\|_F \leq 1 \}$.

(d) Prove that with probability at least $1 - \delta$ over the draw of $(w_1, \ldots, w_m)$, simultaneously for every $\|v\|_2 \leq 1$,

$$|v^T G v - v^T \hat{G} v| \leq \frac{2B^2 \|X\|_F^2}{\sqrt{m}} + 6B^2 \|X\|_F^2 \sqrt{\frac{\ln(4/\delta)}{2m}} =: \star.$$  

**Hint.** Use our big Rademacher generalization theorem together with the preceding parts (Theorem 21.1 in the typed lecture notes).

(e) Prove that, with probability at least $1 - \delta$, simultaneously

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

where $\star$ is the right hand side of the bound in the previous part.

**Remark.** The point is that we can make $\star$ as small as we want just by increasing $m$.

Solution.

Consider the setting of Theorem 17.3 from the typed lecture notes, where we showed that the smoothed margin is nondecreasing with homogeneous networks. As was the case there, define

$$m_i(w) = y_i f(x_i; w)$$

and

$$L(w) = \sum_{i=1}^{n} \ell(m_i(w)) = \sum_{i=1}^{n} \exp(-m_i(w)),$$

and suppose $L(w_0) < 1$.

(a) Show that $\|w_t\| \to \infty$ and $L(w_t) \to 0$.

(b) Suppose additionally that $f$ is a network with $L$ linear layers, and is 1-homogeneous with respect to any individual layer $i$, meaning that for any $i$ and any $c \geq 0$,

$$f(x; (W_L, \ldots, cW_i, \ldots W_1)) = cf(x; (W_L, \ldots, W_i, \ldots W_1)).$$

Prove that $\min_i \|W_i(t)\| \to \infty$.

**Hint.** Look over the theorems in chapter 16 of the typed notes.

**Solution.**
3. A posteriori bounds.

In lecture, we mentioned that sometimes we work with a big function class $F$ for which $\text{URad}(F|S) = \infty$ (e.g., all functions computed by some fixed deep network architecture as we vary the weights), which at first seems to cause problems with our uniform deviation approach. We mentioned that a workaround is to replace $F$ with some $F'$ which represents the tightest class of functions output by an algorithm. (Our running example has been gradient descent, which prefers low norm solutions in certain settings.)

In this question, we’ll work through a simple way to make this rigorous. Let’s carve up $F$ into a sequence of sets
\[ F_1 \subseteq F_2 \subseteq \cdots, \quad \text{where } F = \bigcup_{i \geq 1} F_i. \]

Moreover, suppose that for each $F_i$, we have a Rademacher-style uniform deviation bound: there exist constants $a_i$ and $b_i$ so that, with probability at least $1 - \delta$, every $f \in F_i$ satisfies
\[ \mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + \sqrt{\frac{a_i}{n} + b_i \sqrt{\ln(1/\delta)}}. \]

Prove the following claim: with probability at least $1 - \delta$, for every $f \in F$,
\[ \mathcal{R}(f) \leq \hat{\mathcal{R}}(f) + \inf_{i \geq 1} \left( \sqrt{\frac{a_i}{n} + b_i \sqrt{\frac{2 \ln(i+1)}{n}}} + b_i \sqrt{\frac{\ln(1/\delta)}}{n}} \right). \]

**Hint.** Union bound.

**Remark.** This bound says that we aren’t much worse off than if we had magically known the tightest class in advance ("a priori").

**Solution.**
4. **Are we still friends?**

You will receive full credit for this problem even if you leave it blank.

(a) What are some things you liked about this class (as in, things I should try to preserve next year).
(b) What are some things you did not like (and which you’d like me to change for next year).

**Remark.** My current main plan for next year, in addition to expanding and refining the notes, is to vastly reduce the number of “approximation” lectures.

**Solution.**