Instructions. (Different from homework 0.)

• Everyone must submit an individual write-up.

• You may discuss with up to 3 other people. State their NetIDs clearly on the first page. Outside of office hours, you should not discuss with anyone but these three.

• Homework is due Tuesday, November 5, at 3:00pm; no late homework accepted.

• Please consider using the provided \LaTeX file as a template.
1. Miscellaneous short questions.

Here is some refined notation. Firstly, for the class of nodes you get via some fixed activation:

\[ \mathcal{H}_{\phi,d} := \left\{ \mathbb{R}^d \ni x \mapsto \phi(a^T x + b) \in \mathbb{R} \mid a \in \mathbb{R}^d, b \in \mathbb{R} \right\}. \]

Also this notation for uniform distance:

\[ \|f - g\|_u = \sup_{\|x\| \leq 1} |f(x) - g(x)|. \]

Lastly, a binary classification risk with loss \( \ell \):

\[ \mathcal{R}_\ell(f) := \int \ell(yf(x)) \, d\mu(x,y). \]

(a) (Justifying uniform norm: upper bound.) Suppose \( \ell \) is \( L \)-lipschitz, and \( \mu \) is a probability measure on \((x,y)\) with \( y \in \{1, -1\} \) and \( \|x\| \leq 1 \). Show

\[ \mathcal{R}_\ell(f) - \mathcal{R}_\ell(g) = \int \ell(yf(x)) \, d\mu(x,y) - \int \ell(yg(x)) \, d\mu(x,y) \leq L \|f - g\|_u. \]

(b) (Justifying uniform norm: lower bound.) Given any two continuous functions \( f \) and \( g \), construct a 1-lipschitz loss \( \ell \) and a probability measure \( \mu \) so that the previous part is tight: that is,

\[ \mathcal{R}_\ell(f) - \mathcal{R}_\ell(g) = \|f - g\|_u. \]

Remark: together, we’ve shown why we aim for uniform approximation (it implies bounds for all measures).

(c) (Justifying \( L_2(P) \).) Let \( P \) be a probability measure on \((x,y)\) with \( y \in \{\pm 1\} \), and define

\[ \|f - g\|^2_{L_2(P)} = \int (f(x) - g(x))^2 \, dP(x,y). \]

Suppose \( \ell \) is \( \beta \)-smooth (as in the first optimization lecture) and \( L \)-Lipschitz. Then

\[ \mathcal{R}_\ell(f) - \mathcal{R}_\ell(g) \leq L \|f - g\|_{L_2(P)} + \frac{\beta}{2} \|f - g\|^2_{L_2(P)}. \]

Hint: Cauchy-Schwarz works on \( \|\cdot\|_{L_2(P)} \).

Remark: this points out that \( L_2(P) \) approximation (as with Barron) is also useful.

(d) (Deep, narrow networks.) Let \( \sigma_t(z) := \max\{0, z\} \) denote the ReLU, and for convenience let \( \sigma_t \) denote the coordinate-wise version of appropriate dimension (i.e., \( \sigma_t(v) \) outputs a vector of the same dimension as \( v \), whatever it happens to be).

Suppose \( f : [0,1]^d \to \mathbb{R} \) can be written as a network with a single ReLU layer, specifically \( f(x) = A_2 \sigma_t(A_1 x + b_1) \) where \( A_1 \in \mathbb{R}^{w \times d} \) and \( A_2 \in \mathbb{R}^{1 \times w} \). Construct a network with \( w \) ReLU layers and width \( d + 3 \) which also (exactly) computes \( f \).

Remark: this reveals some convenient properties of ReLUs.

Solution.

(Your solution here.)
2. **Activation in final layer.**

Recall that the lectures on approximation of continuous functions by 2- and 3-layer networks did not include a nonlinearity on the final output. This exercise points out that we can use those as a lemma to establish that networks with final nonlinearities can also approximate continuous functions (albeit with restrictions on the range).

Throughout this exercise, suppose a function class $\mathcal{F}$ is given which fits continuous functions in our usual sense: for any continuous $g : \mathbb{R}^d \to \mathbb{R}$ and any $\tau > 0$, there exists $f \in \mathcal{F}$ with $\|f - g\|_u \leq \tau$.

The following notation will be handy. Namely, given univariate $\sigma : \mathbb{R} \to [0, 1]$, define the function class $\mathcal{F}_\sigma := \{\sigma \circ f : f \in \mathcal{F}\}$.

(a) Suppose $\sigma : \mathbb{R} \to [0, 1]$ is continuous, nondecreasing, Lipschitz, $\lim_{z \to -\infty} \sigma(z) = 0$, $\lim_{z \to +\infty} \sigma(z) = 1$, and $\sigma$ has a continuous inverse.

Show that for any continuous $g : \mathbb{R}^d \to (0, 1)$ and any $\epsilon > 0$, there exists $h \in \mathcal{F}_\sigma$ with

$$\|h - g\|_{L_1} = \int_{x \in [0, 1]^d} |h(x) - g(x)| \, dx \leq \epsilon.$$ 

**Hint.** Find a way to bake the inverse into the problem.

(b) (Optional; hard mode.) Suppose $\sigma : \mathbb{R} \to [0, 1]$ is continuous, nondecreasing, $\lim_{z \to -\infty} \sigma(z) = 0$, and $\lim_{z \to +\infty} \sigma(z) = 1$. (The literature refers to this as “sigmoidal”.)

Show that for any continuous $g : \mathbb{R}^d \to [0, 1]$ and any $\epsilon > 0$, there exists $h \in \mathcal{F}_\sigma$ with $\|h - g\|_u \leq \epsilon$.

**Note.** If you do this part, you must still provide a complete independent solution to the previous part. Be nice to the TA...

**Solution.**

(Your solution here.)
3. Triangle counting.

Recall the proof technique from the “benefits of depth” lectures, which allows us to say that shallow networks can not approximate deep networks in the $\| \cdot \|_{L_1}$ metric, where $\| h \|_{L_1} = \int_0^1 |h(x)| \, dx$. The functions in this problem are strictly univariate.

(a) Prove that for every $L \geq 10$, there exists a ReLU network with $10L$ layers and width $10$ which can not be approximated in $L_1$ by polynomials of degree $\leq L$.

**Hint:** how often can a polynomial of some fixed degree cross a fixed line?

(b) Let $g_n$ denote a 1-nearest-neighbor predictor over some set of points $((x_i, y_i))_{i=1}^n$, with $x_i \in \mathbb{R}$ and $y_i \in \mathbb{R}$; to predict on a new point $x$, $g_n(x) = y_i$ where $|x - x_i| = \min_j |x - x_j|$.

Prove that for every $L \geq 10$, there exists a ReLU network with $10L$ layers and width $10$ which can not be approximated in $L_1$ by $g_n$ with any $((x_i, y_i))_{i=1}^n$ when $n \leq 2^L$.

(c) In class, we mentioned that the various universal approximation results must use something like compact sets. Let’s make this rigorous.

Prove that for any $f : \mathbb{R} \to \mathbb{R}$ computed by a ReLU network of any width and depth, there exist (countably) infinitely many reals $(a_i)_{i=1}^\infty$ so that $f$ does not approximate sin along each interval $[a_i, a_i + \pi]$:

$$\inf_i \int_{a_i}^{a_i + \pi} |f(x) - \sin(x)| \, dx \geq \frac{\pi}{2}.$$

**Remark.** Note that we do not include a ReLU in the last layer, as usual. In this particular problem, if we had one, you could use negative $a_i$ for a trick solution.

**Solution.**

*(Your solution here.)*
4. **NTK with general activations.**

As in the NTK lectures, recall that the kernel corresponding to a shallow network with arbitrary activation has the form

\[ k(x, x') := x^\top x' E_w \sigma'(w^\top x) \sigma'(w^\top x'). \]

Throughout this problem, suppose \( \|x\| = 1. \)

(a) Prove \( k(x, x') = x^\top x' E_w \left[ \sigma'(w_1) \sigma'(w_1^\top x + w_2 \sqrt{1 - (x^\top x')^2}) \right]. \)

**Hint:** rotational invariance of the Gaussian! I extremely incoherently hinted the answer to this twice during lecture...  

**Technical note:** if you wish, you can assume \( \sigma \) has at most countably many points of nondifferentiability; since \( w \) has a continuous distribution, the integral may still be computed.  

**Remark:** The kernel therefore only interacts with \( x \) and \( x' \) via \( x^\top x' \), which is pretty interesting!

(b) Let points \( (x_1, \ldots, x_n) \) be given as well as labels \( (y_1, \ldots, y_n) \) with \( y_i \in \{\pm 1\} \), and suppose \( \sigma(z) = \max\{0, z\} \), the ReLU. Recall that the the NTK predictor of width \( m \) will have the form (ignoring scaling)

\[ f(x) := \sum_{j=1}^m v_j^\top x \sigma'(w_j^\top x), \]

where \( (w_1, \ldots, w_m) \) are IID Gaussian, and \( (v_1, \ldots, v_m) \) are parameters. Suppose there exists a pair \( (x_i, x_j) \) with \( y_i \neq y_j \) and the angle between \( x_i \) and \( x_j \) is at most \( \delta > 0 \). Prove that with probability at least \( 1 - m\delta^2/\pi \), it is impossible to find \( (v_1, \ldots, v_m) \) with \( \sum_i \|v_i\|^2 \leq 1/\delta \) so that \( f(x_i) = y_i \) for all \( i \).

**Solution.**

*(Your solution here.*)
5. Converting activations.

Consider the two activations

\[
\sigma_1(z) := \max\{0, z\} \quad \text{ReLU},
\]
\[
\sigma_2(z) := \frac{1}{1 + \exp(-z)} \quad \text{sigmoid}.
\]

(a) Construct \( f_1 \in \text{span}(H_{\sigma_1,1}) \) of width \( m_1 \) so that \( \|f_1 - \sigma_2\|_u \leq \frac{1}{100} \). Provide an exact upper bound on \( m_1 \) and rigorously verify your construction.

(b) Construct \( f_2 \in \text{span}(H_{\sigma_2,1}) \) of width \( m_2 \) so that \( \|f_2 - \sigma_1\|_u \leq \frac{1}{100} \). Provide an exact upper bound on \( m_2 \) and rigorously verify your construction.

(c) Suppose

\[
g_1(x) = \sum_{j=1}^{m} c_j \sigma_1(a_j^T x) \in \text{span}(H_{\sigma_1,d}).
\]

Construct \( g_2 \in \text{span}(H_{\sigma_2,d}) \) with

\[
\|g_1 - g_2\|_u \leq \frac{\sum_j |c_j| \|a_j\|}{100}.
\]

Solution.

(Your solution here.)
6. **Monomials and uniform approximation via derivatives.**

This problem gives the basics of an approach that says: if your activation is *not* a polynomial, then you can approximate anything. The idea is that non-polynomial activations can approximate polynomials of arbitrary degree.

This problem thus constitutes an alternative to invoking Stone-Weierstrass; **do not** use Stone-Weierstrass or Weierstrass or anything similar in any step of the proof!

The problem will consider only the univariate case, but essentially the same proof works in the multivariate case (as discussed at the end).

For convenience, for any activation \( \sigma \), define \( \mathcal{G}_\sigma := \text{span}(\mathcal{H}_{\sigma,1}) \). Here are some useful analysis facts for this problem:

- Continuous functions are uniformly continuous and bounded (moreover attaining their suprema/infima) on compact sets.
- To say a function \( f \) is \( C^\infty \) means all derivatives exist (and are continuous). If \( \sigma \) is \( C^\infty \), then so is every \( f \in \mathcal{G}_\sigma \).

Throughout this problem, suppose \( \sigma \) is \( C^\infty \) and \( \sigma^{(n)} \neq 0 \), meaning the \( n \)th derivative is not identically the zero function for every nonnegative integer \( n \).

(a) **(Closed under a single derivative.)** Let \( f \in \mathcal{G}_\sigma \) and any \( w \in \mathbb{R} \) and any \( \epsilon > 0 \) be given, and define \( h(x) := xf'(wx) \) (the mapping \( x \mapsto \frac{\partial}{\partial r} f(rx)|_{r=w} \)). Prove that there exists \( g \in \mathcal{G}_\sigma \) so that \( \|h - g\|_u \leq \epsilon \).

**Hint.** Consider the definition of \( \frac{\partial}{\partial r} f(rx)|_{r=w} \) in terms of limits, and see how it interacts with an exact (integral remainder) Taylor expansion. Via the analysis facts above, you can conveniently bound the remainder term. Use this to construct an appropriate \( g \in \mathcal{G}_\sigma \), and prove that it works.

(b) **(Closed under derivatives.)** For every real \( w, b \in \mathbb{R} \) and positive integer \( n \), define

\[
h_{n,w,b}(x) := x^n \sigma^{(n)}(wx - b) = \frac{\partial^n}{\partial r^n} \sigma(rx - b)|_{r=w}.
\]

Show that for any \((w, b, \epsilon, n)\), there exists \( g \in \mathcal{G}_\sigma \) with \( \|g - h_{n,w,b}\|_u \leq \epsilon \).

**Hint.** Combine the previous part with an induction on \( n \) and some careful reasoning about approximations. Be wary of circularity...

(c) **(Monomials.)** Prove that for any positive integer \( n \) and real \( \epsilon > 0 \), there exists \( g \in \mathcal{G}_\sigma \) so that \( \|g - p_n\|_u \leq \epsilon \) where \( p_n(x) = x^n \).

**Hint.** Use the previous part, and double check the conditions on \( \sigma \)...

Now that we have monomials, we can use the Weierstrass Theorem (which has a simple constructive proof). Also, the proof above goes through no problem in the multivariate case (now use \( x \mapsto \sigma(\langle w, x \rangle - b) \), and take different partial derivatives to get various monomials).

**Solution.**

(Your solution here.)
7. **Why?**

You receive full credit for this question so long as you write at least one sentence for each answer. Please be honest and feel free to be critical.

(a) Why are you taking this class?
(b) What is something the instructor can improve?

**Solution.**

*(Your solution here.)*