1. **(Miscellaneous short questions.)**

Provide complete proofs, but try to find short solutions.

(a) **(Justifying uniform norm: upper bound.)** Suppose ℓ is L-lipschitz, and μ is a probability measure supported on $[0,1]^d \times \{-1,+1\}$. Show

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

(b) **(Justifying uniform norm: lower bound.)** Given any two continuous functions f and g, construct an L-lipschitz loss ℓ and a probability measure μ so that the previous part is tight: that is,

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

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

(c) **(Stone-Weierstrass with cos.)** Use the Stone-Weierstrass theorem, as stated in lecture 5 (do not use another source), to prove that for every continuous function $f : \mathbb{R}^d \to \mathbb{R}$ and $\epsilon > 0$, there exists $g \in \mathrm{span}(\mathcal{H}_{\cos})$ with $\|f-g\|_u \leq \epsilon$. (**Hint:** refresh yourself on some trig identities.)

(d) **(Deep, narrow networks.)** Let $\sigma_r(z) := \max\{0,z\}$ denote the ReLU, and for convenience let $\vec{\sigma}_r$ denote the coordinate-wise version of appropriate dimension (i.e., $\vec{\sigma}_r(v)$ 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 $f(x) = A_2\vec{\sigma}_r(A_1x + b_1)$ 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.

(e) **(Uniform approximation with ReLU.)** Again define $\sigma_r(z) := \max\{0,z\}$. Construct $\phi \in \mathrm{span}(\mathcal{H}_{\sigma_r})$ which satisfies the conditions of Theorem 1.9 from Lecture 5 (and provide explicit verification of these conditions).

Remark: consequently, Theorem 1.9 may be applied, and thus shallow ReLU networks fit continuous functions.

Solution.

(Your solution here.)

1. **(Miscellaneous short questions.)**

Provide complete proofs, but try to find short solutions. Note also this refined for neural network notation specifying input dimension:

$$\mathcal{H}_{\phi,d} := \left\{\mathbb{R}^d \ni x \mapsto \phi(a^\top x - b) \in \mathbb{R} \;:\; a \in \mathbb{R}^d, b \in \mathbb{R}\right\}.$$

(a) **(Justifying uniform norm: upper bound.)** Suppose ℓ is L-lipschitz, and μ is a probability measure supported on $[0,1]^d \times \{-1,+1\}$. Show

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

(b) **(Justifying uniform norm: lower bound.)** Given any two continuous functions f and g, construct an L-lipschitz loss ℓ and a probability measure μ so that the previous part is tight: that is,

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

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

(c) **(Stone-Weierstrass with cos.)** Use the Stone-Weierstrass theorem, as stated in lecture 5 (do not use another source), to prove that for every continuous function $f : \mathbb{R}^d \to \mathbb{R}$ and $\epsilon > 0$, there exists $g \in \mathrm{span}(\mathcal{H}_{\cos,d})$ with $\|f-g\|_u \leq \epsilon$. (**Hint:** refresh yourself on some trig identities.)

(d) **(Deep, narrow networks.)** Let $\sigma_r(z) := \max\{0,z\}$ denote the ReLU, and for convenience let $\vec{\sigma}_r$ denote the coordinate-wise version of appropriate dimension (i.e., $\vec{\sigma}_r(v)$ 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 $f(x) = A_2\vec{\sigma}_r(A_1x + b_1)$ 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.

(e) **(Uniform approximation with ReLU.)** Again define $\sigma_r(z) := \max\{0,z\}$. Construct $\phi \in \mathrm{span}(\mathcal{H}_{\sigma_r,1})$ which satisfies the conditions of Theorem 1.9 from Lecture 5 (and provide explicit verification of these conditions).

Remark: consequently, Theorem 1.9 may be applied, and thus shallow ReLU networks fit continuous functions.

Solution.

(Your solution here.)

4. **(2-layer networks fit continuous functions.)**

Recall from class the definition

$$\mathcal{H}_\sigma := \left\{ x \mapsto \sigma(\langle a, x \rangle + b) : a \in \mathbb{R}^d, b \in \mathbb{R} \right\}.$$

Using Stone-Weierstrass, we proved we can approximate continuous functions with $\text{span}(\mathcal{H}_{\exp})$. It was then claimed that the rest of the proof is "essentially univariate"; this exercise completes that proof.

One more piece of notation is needed. Say that $\sigma : \mathbb{R} \to \mathbb{R}$ is *sigmoidal* when it is nondecreasing, continuous, and

$$\lim_{z \to -\infty} \sigma(z) = 0, \qquad \lim_{z \to \infty} \sigma(z) = 1.$$

(a) The first missing piece is to assert that we really are left with a univariate problem. Namely, prove the following.

Let $\sigma : \mathbb{R} \to \mathbb{R}$ and $\phi : \mathbb{R} \to \mathbb{R}$ be given. Suppose that for any interval $[r, s]$ and any $\tau > 0$, we can always find $h \in \text{span}(\mathcal{H}_\sigma)$ so that

$$\sup \left\{ |h(x) - \phi(x)| : x \in [r, s] \right\} \leq \tau.$$

(In words: we have a way to approximate ϕ along $[r, s]$ with $\text{span}(\mathcal{H}_\sigma)$.) Then for any $\epsilon > 0$ and any $g \in \mathcal{H}_\phi$ but now $g : \mathbb{R}^d \to \mathbb{R}$, we can still choose $f \in \text{span}(\mathcal{H}_\sigma)$ with $\|f - g\|_u \leq \epsilon$.

(b) Let sigmoidal $\sigma : \mathbb{R} \to \mathbb{R}$, target error $\tau > 0$, interval $[r, s]$, and a function $\psi : \mathbb{R} \to \mathbb{R}$ be given with ψ Lipschitz continuous along $[r, s]$. Show that there exists $h \in \text{span}(\mathcal{H}_\sigma)$ satisfying

$$\sup \left\{ |h(x) - \psi(x)| : x \in [r, s] \right\} \leq \tau.$$

Hints. (a) Note that for large M, $\sigma(Mx) \approx \mathbb{1}[x \geq 0]$; (b) consider drawing a picture for the simpler case of nonincreasing ψ, with special attention to the meaning of Lipschitz continuity.

(c) Prove that exp and cos are Lipschitz continuous along any bounded interval. (Yup, that's really all you need to do for this part.)

Though you don't need to write anything about it here, I urge you to verify that the preceding steps can be combined with the material in lecture to complete the proof.

Solution.

(Your solution here.)

4. **(2-layer networks fit continuous functions.)**

Recall the refined definition

$$\mathcal{H}_{\sigma,d} := \left\{ x \mapsto \sigma(\langle a, x \rangle + b) : a \in \mathbb{R}^d, b \in \mathbb{R} \right\}.$$

Using Stone-Weierstrass, we proved we can approximate continuous functions with $\text{span}(\mathcal{H}_{\exp,d})$. It was then claimed that the rest of the proof is "essentially univariate"; this exercise completes that proof.

One more piece of notation is needed. Say that $\sigma : \mathbb{R} \to \mathbb{R}$ is *sigmoidal* when it is nondecreasing, continuous, and

$$\lim_{z \to -\infty} \sigma(z) = 0, \qquad \lim_{z \to \infty} \sigma(z) = 1.$$

(a) The first missing piece is to assert that we really are left with a univariate problem. Namely, prove the following.

Let $\sigma : \mathbb{R} \to \mathbb{R}$ and $\phi : \mathbb{R} \to \mathbb{R}$ be given. Suppose that for any interval $[r, s]$ and any $\tau > 0$, we can always find $h \in \text{span}(\mathcal{H}_{\sigma,1})$ so that

$$\sup \left\{ |h(x) - \phi(x)| : x \in [r, s] \right\} \leq \tau.$$

(In words: we have a way to approximate ϕ along $[r, s]$ with $\text{span}(\mathcal{H}_{\sigma,1})$.) Then for any $\epsilon > 0$ and any $g \in \mathcal{H}_{\phi,d}$ but now $g : \mathbb{R}^d \to \mathbb{R}$, we can still choose $f \in \text{span}(\mathcal{H}_{\sigma,d})$ with $\|f - g\|_u \leq \epsilon$.

(b) Let sigmoidal $\sigma : \mathbb{R} \to \mathbb{R}$, target error $\tau > 0$, interval $[r, s]$, and a function $\psi : \mathbb{R} \to \mathbb{R}$ be given with ψ Lipschitz continuous along $[r, s]$. Show that there exists $h \in \text{span}(\mathcal{H}_{\sigma,1})$ satisfying

$$\sup \left\{ |h(x) - \psi(x)| : x \in [r, s] \right\} \leq \tau.$$

Hints. (a) Note that for large M, $\sigma(Mx) \approx \mathbb{1}[x \geq 0]$; (b) consider drawing a picture for the simpler case of nonincreasing ψ, with special attention to the meaning of Lipschitz continuity.

(c) Prove that exp and cos are Lipschitz continuous along any bounded interval. (Yup, that's really all you need to do for this part.)

Though you don't need to write anything about it here, I urge you to verify that the preceding steps can be combined with the material in lecture to complete the proof.

Solution.

(Your solution here.)

6. **(Monomials and uniform approximation via derivatives.)**

This problem will provide an approach to uniform approximation that avoids 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 σ, define $\mathcal{G}_\sigma := \text{span}(\mathcal{H}_\sigma)$. 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^∞ means all derivatives exist (and are continuous). If σ is C^∞, then so is every $f \in \mathcal{G}_\sigma$.

Throughout this problem, suppose σ is C^∞ and $\sigma^{(n)} \neq 0$, meaning the $n^{(\text{th})}$ 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 $x \mapsto \partial/\partial r f(rx)|_{r=w}$). Prove that there exists $g \in \mathcal{G}_\sigma$ so that $\|h - g\|_u \leq \epsilon$.
Hint. Consider the definition of $\partial/\partial r f(rx)|_{r=w}$ 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 $w \in \mathbb{R}$ and positive integer n, define

$$h_{n,w}(x) := x^n \sigma^{(n)}(wx) = \partial^n/\partial r^n \sigma(rx)|_{r=w}.$$

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

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 $x \mapsto \sigma(\langle w, x \rangle)$, and take different partial derivatives to get various monomials).

Solution.

(Your solution here.)

6. **(Monomials and uniform approximation via derivatives.)**

This problem will provide an approach to uniform approximation that avoids 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 σ, 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^∞ means all derivatives exist (and are continuous). If σ is C^∞, then so is every $f \in \mathcal{G}_\sigma$.

Throughout this problem, suppose σ is C^∞ and $\sigma^{(n)} \neq 0$, meaning the $n^{(\text{th})}$ 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 $x \mapsto \partial/\partial r f(rx)|_{r=w}$). Prove that there exists $g \in \mathcal{G}_\sigma$ so that $\|h - g\|_u \leq \epsilon$.
Hint. Consider the definition of $\partial/\partial r f(rx)|_{r=w}$ 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

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

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

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.)