CS 540 DLT — Homework 1.

your NetID here.

Version 2.

Instructions.

- Homework is due **Wednesday, September 22, at 11:59pm**; no late homework accepted.
- You must work individually for this homework.
- Excluding office hours, and high-level discussions on discord, 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.
- For any problem asking you to construct something, for full credit you must always formally prove your construction works.
- General course and homework policies are on the course webpage.

Notation (just for this homework, plus some reminders from lecture)**.**

- Function space norms:

$$\|f\|_{\mathrm{u}} := \sup_{x\in[0,1]^d} |f(x)|, \qquad \|f\|_{L_1} = \int_{[0,1]^d} |f(x)|\,\mathrm{d}x.$$

- Activations: $\sigma : \mathbb{R} \to \mathbb{R}$ will denote general activations, and $\sigma_{\mathrm{r}}(z) := \max\{0, z\}$ denotes the ReLU. Sometimes we'll apply the activations coordinate-wise to a vector v, and write $\vec{\sigma}(v)$ if there is some ambiguity, but simply $\sigma(v)$ otherwise.
- Function classes: bounded and unbounded (but finite!) width networks with a single hidden layer, namely

$$\mathcal{F}_{\sigma,d,m} := \left\{ x \mapsto a^{\intercal}\sigma(Wx + b) : a \in \mathbb{R}^m, W \in \mathbb{R}^{m\times d}, b \in \mathbb{R}^m \right\},$$
$$\mathcal{F}_{\sigma,d} := \bigcup_{m\geq 0} \mathcal{F}_{\sigma,d,m}.$$

Version history.

1. Initial version.

$1+\epsilon$. Strict inequality in 1(b).

$1+2\epsilon$. Wording tweak in main notation.

2. Density in 1(b) can have discontinuities; inputs in 1(d) are distinct; target width in 2(a) relaxed; coordinate notation in 3(a) clarified to $w^{\intercal}\boldsymbol{e}_1$ and $w^{\intercal}\boldsymbol{e}_2$.

1. Miscellaneous short questions.

Recall our definition of population risk for binary classification problems: given a univariate loss $\ell : \mathbb{R} \to \mathbb{R}$,

$$\mathcal{R}_\ell(f) := \int \ell(yf(x)) \, \mathrm{d}\mu(x, y) = \mathbb{E}\ell(Yf(X)).$$

(a) **(Strength of uniform norm.)** Suppose ℓ is ρ-lipschitz, and μ is a probability measure on (x, y) with $y \in \{-1, +1\}$ and $x \in [0, 1]^d$. Show

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

Remark: there exists a choice of μ, ℓ, and f and g so that this is tight.

(b) **(Weakness of L_1 norm.)** Let $B > 0$ be given, and let $\ell(z) := \ln(1 + \exp(-z))$ be the logistic loss. Construct a probability distribution over pairs (x, y) with $x \in [0, 1]^d$ and $y \in \{\pm 1\}$, where the marginal density on x is continuous, and define a pair of functions $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^d \to \mathbb{R}$ so that

$$\mathcal{R}_\ell(f) - \mathcal{R}_\ell(g) > B\|f - g\|_{L_1}.$$

Hint: "continuous" here means "continuous with respect to the Lebesgue measure"; e.g., you can make the density piecewise-constant over a partition of $[0, 1]^d$.

(c) **(Deep, narrow networks.)** Suppose $f : [0, 1]^d \to \mathbb{R}$ can be written as a network with a single ReLU layer of width m, specifically $f(x) = A_2 \vec{\sigma}_{\mathrm{r}}(A_1 x + b_1)$ where $A_1 \in \mathbb{R}^{m \times d}$ and $A_2 \in \mathbb{R}^{1 \times m}$. Construct a network with m ReLU layers and width $d + 3$ which also (exactly) computes f.

Remark: this reveals some convenient properties of ReLUs.

(d) **(Easiness of fitting finite point sets.)** Let labeled data $((x_i, y_i))_{i=1}^n$ with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ be given with all x_i distinct. Show that it is possible to choose $w \in \mathbb{R}^d$ and scalars $((a_j, b_j))_{j=1}^n$ so that

$$f(x_i) = y_i \ \forall i, \qquad \text{where } f(x) := \sum_{j=1}^n a_j \mathbb{1}[w^\intercal x + b_j \geq 0].$$

Hint: The difficulty is in showing we can use a single w across nodes; after that, you can apply Proposition 2.1 from the lecture notes (why?).

Remark: the construction may look wild, even if there exists some highly smooth function interpolating the points. Overall, there are many reasons why we try to fit functions everywhere and not just over a small finite set.

Solution. *(If using this template, please write your solution here.)*

2. Univariate approximation with shallow ReLU networks.

In lecture, we approximated differentiable or Lipschitz functions with threshold networks; here, we'll approximate twice-differentiable functions with ReLU networks.

Throughout, suppose $g : \mathbb{R} \to \mathbb{R}$ is an arbitrary twice-differentiable function with $g(0) = g'(0) = 0$.

(a) Suppose $|g''| \leq \beta$ over $[0, 1]$, and let $\epsilon > 0$ be given. Prove that there exists a ReLU network $f(x) := \sum_{i=1}^{m} a_i \sigma_{\mathrm{r}}(x - b_i)$ with

$$m \leq \left\lceil \frac{\beta}{\epsilon} \right\rceil \qquad \text{and} \qquad \|f - g\|_u \leq \epsilon.$$

Remark: you may assume g'' is continuous, though it is not necessary.
Remark: can you get away with smaller widths $\lceil \sqrt{\beta/\epsilon} \rceil$ or even $\lceil \beta/\sqrt{\epsilon} \rceil$?

(b) Using the previous part, show that for any $\epsilon > 0$ and for any $h : \mathbb{R}^d \to \mathbb{R}$ of the form $h(x) := \sum_{j=1}^{m} a_j \exp(w_j^{\mathsf{T}} x + b_j)$, there exists $f(x) := \sum_{i=1}^{N} r_i \sigma_{\mathrm{r}}(v_i^{\mathsf{T}} x + s_i)$ with $\|f - h\|_{\mathrm{u}} \leq \epsilon$.
Remark: combining this with Lemma 2.3 from lecture implies shallow ReLU networks are universal approximators.

(c) Prove the infinite-width exact representation $g(x) = \int \sigma_{\mathrm{r}}(x - b) g''(b) \, \mathrm{d}b$.

(d) Using the previous part and the sampling tools in section 3.3 of the lecture notes, prove that for any $\epsilon > 0$, there exists a ReLU network $f(x) := \sum_{i=1}^{m} a_i \sigma_{\mathrm{r}}(x - b_i)$ with

$$m \leq \left\lceil \frac{1}{3\epsilon} \left[\int_0^1 |g''(x)| \, \mathrm{d}x \right]^2 \right\rceil \qquad \text{and} \qquad \int_0^1 \big(f(x) - g(x)\big)^2 \, \mathrm{d}x \leq \epsilon.$$

Solution. *(If using this template, please write your solution here.)*

3. NTK with general activations.

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

$$k(x, x') := x^\intercal x' \mathbb{E}_w \sigma'(w^\intercal x)\sigma'(w^\intercal x'),$$

where $w \in \mathbb{R}^d$ is a standard Gaussian random vector, thus $\mathbb{E}w = 0$ and $\mathbb{E}ww^\intercal = I$.

Throughout this problem, suppose $\|x\| = 1$ (this includes $\|x'\| = 1$ in part (a)).

(a) Prove $k(x, x') = x^\intercal x' \mathbb{E}_w \left[\sigma'(w^\intercal \boldsymbol{e}_1)\sigma'\left(w^\intercal \boldsymbol{e}_1 x^\intercal x' + w^\intercal \boldsymbol{e}_2 \sqrt{1 - (x^\intercal x')^2}\right)\right]$, where $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are standard basis vectors.

Hint: rotational invariance of the Gaussian!

Technical note: if you wish, you can assume σ 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 $x^\intercal x'$, 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)

$$f(x) := \sum_{j=1}^{m} v_j^\intercal x \sigma'(w_j^\intercal x),$$

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 $1 - m\delta/\pi$, 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. *(If using this template, please write your solution here.)*

4. 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. Moreover, as this problem develops an alternative to the Stone-Weierstrass approach, do not apply Stone-Weierstrass within your proof!

Recall the notation $\mathcal{F}_{\sigma,d}$ from section 2.2 of the lecture notes. Here are some useful analysis facts for this question:

- 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 $f \in \mathcal{F}_{\sigma,1}$.

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 $f \in \mathcal{F}_{\sigma,1}$ 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 $g \in \mathcal{F}_{\sigma,1}$ so that $\|h - g\|_{\mathrm{u}} \leq \epsilon$.
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 $g \in \mathcal{F}_{\sigma,1}$, 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 $g \in \mathcal{F}_{\sigma,1}$ with $\|g - h_{n,w,b}\|_{\mathrm{u}} \leq \epsilon$.
Hint: combine the previous part with an induction on n and some careful reasoning about approximations.

(c) **(Monomials.)** Prove that for any positive integer n and real $\epsilon > 0$, there exists $g \in \mathcal{F}_{\sigma,1}$ so that $\|g - p_n\|_{\mathrm{u}} \leq \epsilon$ where $p_n(x) = x^n$.
Hint: use the previous part, and double check the conditions on σ.

(d) **(Universal approximation.)** Show that for any $\epsilon > 0$ and any $g : \mathbb{R}^d \to \mathbb{R}$ of the form $g(x) := \sum_{j=1}^m a_j \exp(w_j^\intercal x + b_j)$, there exists $f \in \mathcal{F}_{\sigma,d}$ with $\|f - g\|_{\mathrm{u}} \leq \epsilon$.
Hint: use Taylor's theorem to approximate exp with polynomials and invoke the previous part, and otherwise proceed similarly to question 2(b).
Remark: from here we can again apply Lemma 2.3 from the lecture notes, and obtain an incomparable universal approximation theorem as compared with Theorem 2.1 in the lecture notes, since we have required σ to be C^∞ on the one hand, but do not need the well-behaved limits required there. Can we relax the C^∞ condition in this proof?

Solution. *(If using this template, please write your solution here.)*

5. 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. *(If using this template, please write your solution here.)*