CS 540 DLT — Homework 1.

your NetID here.

Version $1 + 3\epsilon$.

Instructions.

- Homework is due **Wednesday, September 28, 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. For convenience, given a univariate activation $\sigma : \mathbb{R} \to \mathbb{R}$, define 2-layer biased networks of arbitrary width as

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

Additionally, let $\sigma_\mathrm{r}(z) := \max\{0, z\}$ denote the ReLU.

Version history.

1. Initial version.

$1 + \epsilon$. (1a.) removed ℓ subscript. (1d.) "biased" explicitly stated. (4a.) $\mathcal{F}_{d,1}$ should have been $\mathcal{F}_{\sigma,1}$. (4b.) $\mathrm{d}r \to \mathrm{d}w$.

$1 + 2\epsilon$. Due date pushed back one week; (3a.) nullified.

$1 + 3\epsilon$. (2c.) Clarified/strengthened continuity simplification.

1. Miscellaneous short questions.

(a) **(Strength of uniform norm.)** Let $\mathcal{C}$ denote continuous functions over $\mathbb{R}^d$, fix a continuous activation $\sigma : \mathbb{R}^d \to \mathbb{R}$, and suppose $\mathcal{F}_{\sigma,d}$ is a *universal approximator*, meaning $\sup_{g\in\mathcal{C}} \inf_{f\in\mathcal{F}_{\sigma,d}} \sup_{x\in[0,1]^d} |f(x)-g(x)| = 0$. Show that for any loss ℓ which is ρ-Lipschitz in its first argument and any probability distribution μ over $[0,1]^d \times \mathbb{R}$, the future error $\mathcal{R}$ defined as

$$\mathcal{R}(f) := \mathbb{E}_{(x,y)\sim\mu}\ell(f(x), y)$$

satisfies $\inf_{f\in\mathcal{F}} \mathcal{R}(f) = \inf_{g\in\mathcal{C}} \mathcal{R}(g)$.

Remark: this claim follows lecture, specifically the discussion of the varying approximation goals and how they imply each other; this one shows that universal approximation implies the weaker "future risk" approximation.

(b) **(Weakness of L_1 norm.)** Suppose $\mathcal{R}$ and $\mathcal{F}_{\sigma,d}$ and $\mathcal{G}$ as in the preceding problem (with ReLU $\sigma = \sigma_{\mathrm{r}}$ for concreteness), and the logistic loss $\ell(\hat{y}, y) = \ln(1 + \exp(-\hat{y}y))$. Given any $\epsilon > 0$, construct a discrete probability distribution over $[0,1]^d \times \{\pm 1\}$ (inputs and labels) and two functions $f \in \mathcal{F}_{\sigma_{\mathrm{r}},d}$ and $g \in \mathcal{C}$ so that

$$\mathcal{R}(f) \geq \frac{1}{\epsilon} + \mathcal{R}(g) \qquad \text{and} \qquad \int_{[0,1]^d} |f(x) - g(x)|\, \mathrm{d}x \leq \epsilon.$$

Remark: continuing the remark from the preceding problem, this problem establishes that it would not be enough to approximate continuous functions in L_1, we need something stronger.

(c) **(Compactness.)** Show that $\inf_{f\in\mathcal{F}_{\sigma_{\mathrm{r}},1}} \sup_{x\in\mathbb{R}} |f(x) - \sin(x)| \geq 1$.

Remark: this shows compactness is necessary.

(d) **(Deep, narrow networks.)** Suppose $f : [0,1]^d \to \mathbb{R}$ can be written as a 2-layer biased ReLU network of width m, meaning $f(x) = a^{\mathsf{T}}\sigma_{\mathrm{r}}(Vx + b) \in \mathcal{F}_{\sigma_{\mathrm{r}},d,m}$. Construct a biased network with $m + 1$ ReLU layers and width $d + 3$ which also (exactly) computes f.

Remark: this reveals some convenient properties of ReLUs.

Remark: if f itself was constructed to approximate some continuous function, together we conclude that deep, narrow networks are also universal approximators.

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

2. Constructive data-adaptive univariate approximation.

This question considers functions $f : [0, 1] \to \mathbb{R}$; that is, over the unit interval. Define a *bounded variation* norm $\|f\|_{\text{BV}}$ in the following two steps.

- If f is monotone (nondecreasing or nonincreasing), define $\|f\|_{\text{BV}} := |f(0) - f(1)|$.
- Otherwise, given any f, define a family of decompositions into monotone functions as

$$\mathcal{S}_f := \left\{(g, h) \ : \ f = g + h \text{ where } g \text{ and } h \text{ are monotone}\right\},$$

  and finally a norm $\|f\|_{\text{BV}} = \inf\{\|g\|_{\text{BV}} + \|h\|_{\text{BV}} : (g, h) \in \mathcal{S}_f\}$, with the convention $\|f\|_{\text{BV}} = \infty$ when $\mathcal{S}_f = \emptyset$.

This problem will use $\|\cdot\|_{\text{BV}}$ to give data-adaptive univariate approximation bounds. For comparison, define the (tightest) Lipschitz constant $\|f\|_{\text{LIP}}$ as

$$\sup_{x \neq y \in [0,1]} \frac{|f(x) - f(y)|}{|x - y|}.$$

If f is differentiable, then $\|f\|_{\text{LIP}} = \sup\{|f'(x)| : x \in (0, 1)\}$.

(a) Suppose f is continuously differentiable. Show that $\|f\|_{\text{BV}} \le \|f\|_{\text{LIP}}$.
Note: it's still true without differentiability, but more painful.

(b) Show that for any $\epsilon > 0$, there exists f so that $\|f\|_{\text{LIP}} \ge 1/\epsilon$ but $\|f\|_{\text{BV}} \le \epsilon$.

(c) Show that for any $g : [0, 1] \to \mathbb{R}$ and any $\epsilon > 0$, there exists a 2-layer threshold network (meaning activation $\sigma(r) = \mathbb{1}[z \ge 0]$) with at most $4\lceil \|g\|_{\text{BV}}/\epsilon \rceil$ nodes such that $|f(x) - g(x)| \le \epsilon$ for all $x \in [0, 1]$.
Simplification: feel free to assume without proof that g is continuous, $\|g\|_{\text{BV}} < \infty$, and $\mathcal{S}_g$ can be restricted to continuous pairs (without changing $\|g\|_{\text{BV}}$).

(d) Show that for any continuously differentiable $g : [0, 1] \to \mathbb{R}$ with $g(0) = 0$ and any $\epsilon > 0$, there exists a 2-layer ReLU network with at most $4\lceil \|g'\|_{\text{BV}}/\epsilon \rceil$ nodes such that $|f(x) - g(x)| \le \epsilon$ for all $x \in [0, 1]$.
Simplification: again, when considering monotone pairs, feel free to assume continuity.
Remark: we can use this to approximate $r \mapsto \exp(r)$ with ReLUs, and plug this into the proof scheme from the lecture notes to show $\mathcal{F}_{\sigma_{\text{r}},d}$ is a universal approximator.

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) ***(The answer to this part of this problem is in the notes now, feel free to skip.)*** 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 shows that we can perform univariate approximation with any activation $\sigma : \mathbb{R} \to \mathbb{R}$ which is not a polynomial, under an additional technical condition (that it is C^∞, which will be explained shortly). This can be plugged into the proof in the typed lecture notes to imply that any such $\mathcal{F}_{\sigma,d}$ is a universal approximator.

Note: you may not use the Stone-Weierstrass Theorem when solving this problem.

In more detail, σ being C^∞ means that it (and every element of $\mathcal{F}_{\sigma,1}$) have continuous derivatives of all order, and moreover they are uniformly bounded over compact sets (that is, the n^{th} derivative $\sigma^{(n)}$ satisfies $\sup_{|x|\le r} \sigma^{(n)}(x) < \infty$ for all $r < \infty$. This fact will be used a few times in the proofs.

Additionally, as a consequence of σ not being a polynomial, then for every n, there exists x so that $\sigma^{(n)}(x) \neq 0$.

Lastly, for convenience, define *uniform norm* $\|f - g\|_{\mathrm{u}} := \sup_{x\in[0,1]} |f(x) - g(x)|$.

(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 $h(x) := xf'(wx) = ({}^{\mathrm{d}}\!/\!{}_{\mathrm{d}w})f(wx)$. Prove that there exists $g \in \mathcal{F}_{\sigma,1}$ so that $\|h - g\|_{\mathrm{u}} \le \epsilon$.

Hint: consider writing h as a limit (via the definition of derivative); we are trying to show that the limit point is well-approximated in $\mathcal{F}_{\sigma,1}$. Writing out that limit, is there a natural candidate with whichn to approximate h? To control the error of this approximation, it may be helpful to use the above properties, and a second-order Taylor expansion with an exact remainder.

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

$$h_{n,r,b}(x) := x^n \sigma^{(n)}(wx + b) = \frac{\mathrm{d}^n}{\mathrm{d}w^n}\sigma(wx + b).$$

Show that for any (w, b, ϵ, n), there exists $g \in \mathcal{F}_{\sigma,1}$ with $\|g - h_{n,w,b}\|_{\mathrm{u}} \le \epsilon$.

Hint: set up an induction over n; a key to the proof is choosing a (powerful) inductive hypothesis, be sure to state yours clearly. The inductive step can be done similarly to the previous part, albeit more elaborately.

(c) **(Monomials.)** Prove that for any positive integer n and real $\epsilon > 0$, there exists $g \in \mathcal{F}_{\sigma,1}$ so that $\|g - p_n\|_{\mathrm{u}} \le \epsilon$ where $p_n(x) = x^n$.

Hint: this one should be short, you can directly use the previous part.

(d) **(Universal approximation.)** Show that for any $r > 0$ and any $\epsilon > 0$, there exists $f \in \mathcal{F}_{\sigma,1}$ with $\sup_{|x|\le r} |f(x) - \exp(x)| \le \epsilon$.

Remark: This can now be plugged into the universal approximation proof from the lecture notes, and finishes the proof except when σ is not C^∞. There is a trick to drop C^∞ which I'll mention in some update to the notes.

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.)*