CS 540 DLT — Homework 2.

your NetID here.

Version $1 + \epsilon$.

Instructions.

- This homework is due **Wednesday, October 20, 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)**.**

- "Uniform norm" or "supremum norm" (over the cube): $\|f\|_{\mathrm{u}} := \sup_{x \in [0,1]^d} |f(x)|$. (Well, typically it's over all of $\mathbb{R}^d$, but we only need this restriction.)

  L_1 norm (with uniform measure over the cube): $\|f\|_{L_1} = \int_{[0,1]^d} |f| \, \mathrm{d}x$.

Version history.

1. Initial version.

$1 + \epsilon$. $x^{2k} \to x^{2^k}$ in 1(b).

1. Miscellaneous short questions.

The first two parts are approximation theory questions.

(a) Use Lemma 2.1 and Lemma 5.5 from the typed notes to prove the following *constructive, deep network* universal approximation theorem: for any continuous function $g : \mathbb{R}^d \to \mathbb{R}$ and any $\epsilon > 0$, there exists a ReLU network $f : \mathbb{R}^d \to \mathbb{R}$ *of depth* $\mathcal{O}(d \ln(2^d/\epsilon))$ *and unbounded width* such that $\|f - g\|_{\mathrm{u}} \leq \epsilon$.

Anti-hint. You must use Lemmas 2.1 and 5.5 to receive points; invoking a universal approximation result from the lecture notes will receive 0 points.

Remark. As mentioned in lecture, Lemma 5.5 is neat because it gives us a way to localize and thus get uniform norm control, something we couldn't do in Theorem 2.1.

(b) Prove that for any positive integer k and any real $\epsilon > 0$, there exists a ReLU network $h : \mathbb{R} \to \mathbb{R}$ of width $\mathcal{O}(1)$ and depth $\mathcal{O}(k^2 \ln(1/\epsilon))$ satisfying $\sup_{x \in [0,1]} \left| h(x) - x^{2^k} \right| \leq \epsilon$.

Hint. Use Theorem 5.2 from the typed lecture notes; if stuck, see some of the proofs in section 5.4 (although afaik nothing there is adequate to solve this directly).

In the remaining parts of the problem, let $S \subseteq \mathbb{R}^d$ be a subspace, let Π_S denote orthogonal projection onto S (thus Π_S can be written as $VV^{\intercal}$ for some orthonormal matrix $V \in \mathbb{R}^{d \times k}$, where $k \leq d$), suppose $\widehat{\mathcal{R}}$ is β-smooth and convex, and consider the *projected gradient iterates*

$$w_0 \in S, \qquad w_{i+1} := \Pi_S \left(w_i - \frac{1}{\beta} \nabla \widehat{\mathcal{R}}(w_i) \right).$$

Please use just what was proved in lecture and linear algebra to prove the following statements, rather than concepts from convexity which we did not cover.

(c) Show that we still have a smoothness inequality as in section 7.1.1 of the typed notes:

$$\widehat{\mathcal{R}}(w_{i+1}) \leq \widehat{\mathcal{R}}(w_i) - \frac{1}{2\beta} \|\Pi_S(\nabla \widehat{\mathcal{R}}(w_i))\|^2.$$

(d) Prove that Theorem 7.3 still holds: for any $z \in S$,

$$\widehat{\mathcal{R}}(w_t) - \widehat{\mathcal{R}}(z) \leq \frac{\beta}{2t} \left(\|w_0 - z\|^2 - \|w_t - z\|^2 \right).$$

Remark. These two parts hold more generally (than for subspaces), but we didn't discuss convex sets and convexity in enough detail to grant easy proofs. Thus, please just use lecture material and linear algebra.

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

2. 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 $\|h\|_{L_1} = \int_0^1 |h(x)|\,\mathrm{d}x$. The functions in this problem are strictly univariate.

(a) 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 $f : \mathbb{R} \to \mathbb{R}$ with $\leq 10L$ layers and width ≤ 10 such that for the 1-nearest-neighbor predictor given by any set of $n \leq 2^L$ points $((x_i, y_i))_{i=1}^n$, then $\|f - g_n\| \geq 1/100$.

Hint / proof sketch from lecture. We only skimmed this in lecture, so here's the idea. Use $f(x) = \Delta^{3L}(x)$ (so that we meet the width and depth requirements). This function consists of 2^{3L-1} isosceles triangles of width $1/2^{3L-1}$ and height 1. We need to lower bound the area between f and g_n. We've described the graph of f, but how does the graph of g_n look? Is there some easy way to argue that g_n can not possibly approximate f?

(b) 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 (disjoint!) interval $[a_i, a_i + \pi]$:

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

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

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

3. Automatic regularization property of GD.

In lecture, we'll use the margin maximization perspective to show that gradient descent can sometimes choose minimum norm solutions. If instead we are content with merely showing a low norm property, then in fact we've provided enough machinery in just the convex optimization lectures (section 7 of the typed notes).

Suppose $\widehat{\mathcal{R}}$ is β-smooth and convex, let w_0 be arbitrary, and let $(w_t)_{t\geq 0}$ denote the sequence given by gradient descent with step size $1/\beta$. Define two related sequences:

$$u_t := \arg\min \left\{ \|z - w_0\| : z \in \mathbb{R}^p, \widehat{\mathcal{R}}(z) \leq \widehat{\mathcal{R}}(w_t) \right\}, \qquad t \geq 0,$$
$$v_t := \arg\min \left\{ \widehat{\mathcal{R}}(z) : z \in \mathbb{R}^p, \|z - w_0\| \leq \|w_t - w_0\| \right\}, \qquad t \geq 0.$$

At an intuitive level, these sequences seem "explicitly regularized" when compared with the gradient descent path, and one might expect them to have much smaller norms.

Prove that in fact the gradient descent path has nearly the same regularization level automatically: show that for all $t \geq 1$,

$$\|w_t - w_0\| \leq 2 \min \left\{ \|u_t - w_0\|, \|v_t - w_0\| \right\}.$$

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

4. Eigenvalues of expected kernel.

In the smooth shallow NTK proof in lecture, we will need the *infinite-width Gram matrix*, there written as $J_0 J_0^\intercal$ to have nicely-behaved eigenvalues. In this problem, we will work out these eigenvalues in the infinite-width, shallow smooth network setting.

Throughout this problem let examples $(x_i)_{i=1}^n$ be given and fixed with $\|x_i\| \le 1$, and collect all examples as rows of a matrix $X \in \mathbb{R}^{n\times d}$. Given a differentiable activation function σ, the shallow *gram matrix* $G \in \mathbb{R}^{n\times n}$ corresponding to this kernel is

$$G_{ij} := \mathbb{E}_w x_i^\intercal x_j \sigma'(w^\intercal x_i)\sigma'(w^\intercal x_j),$$

where w is a standard Gaussian random vector.

We will not prove meaningful bounds, we will merely show that G has full rank, though the bounds on the eigenvalues which can be extracted from this proof are sufficient for the setup in lecture (after adjusting for finite width, which we'll do in homework 3 or 4).

First let's establish some basic sanity checks.

(a) Suppose a *linear network*, meaning $\sigma(z) = z$. Prove that G being full rank implies $d \ge n$.

(b) Suppose there exists a pair $x_i = x_j$ with $i \ne j$; prove in the general case (σ possibly nonlinear) that G does not have full rank.

To handle the activations in the nonlinear case, we will need the *(normalized) Probabilist's Hermite polynomials* $(p_k)_{k=0}^\infty$. These satisfy many magical properties, but the ones we will need are as follows.

- p_k is a polynomial of degree k.
- If w is a standard Gaussian random vector, then $\mathbb{E}p_k(w^\intercal x_i)p_l(w^\intercal x_j) = (x_i^\intercal x_j)^k \mathbb{1}[k = l]$; this equality goes a little beyond the usual claim that Hermite polynomials are *orthonormal* with respect to an inner product defined by Gaussian integration.
- If h is a function with $\mathbb{E}|h(g)| < \infty$ where g is a standard univariate Gaussian random variable, then there exist *Hermite coefficients* $(c_k)_{k=0}^\infty$ with $c_k = \mathbb{E}h(g)p_k(g)$ such that $h(x) = \sum_{k=0}^\infty c_k p_k(x)$.

For the remaining parts of the problem, fix an activation σ (potentially nonlinear) and let $(c_k)_{k\ge 0}$ denote the Hermite coefficients of σ'.

(c) Prove that $G_{ij} = \sum_{k\ge 0} c_k^2 (x_i^\intercal x_j)^{k+1}$.

(d) Prove that if $G_{jj} > \sum_{i\ne j} |G_{ij}|$ for each j, then G is positive definite (and thus has full rank).
Hint. Since G is real and symmetric, it has real eigenvalues. Take any pair (λ, v) with $\lambda v = Gv$, and choose $j = \arg\max_j |v_j|$. After some algebra and the provided condition, it follows that $\lambda > 0$.

(e) Suppose $\|x_i\| = 1$, and $x_i \ne \pm x_j$ whenever $i \ne j$, and that σ' has infinitely many nonzero Hermite coefficients (meaning $\sup\{k : c_k \ne 0\} = \infty$).
Prove that G is positive definite (and thus has full rank).
Hint. You may use the *Schur product theorem* without proof: if $A, B \succeq 0$, then $\det(A \circ B) \ge \det(A)\cdot\det(B)$, where $A \circ B$ denote element-wise product.

Remark 1. One of our "universal approximation" themes was that we are fine so long as our activation is not a polynomial. In this setting, if the activation is not a polynomial, it will have an infinite Hermite expansion. Consequently, this result holds for the sigmoid, and also the ReLU (after being careful about the nondifferentiability).

Remark 2. If we try to use this proof technique to give a lower bound on the eigenvalues, it will be pretty bad, since standard activations all have fast decay of Hermite coefficients.

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