CS 540 DLT — Homework 2.

your NetID here.

Version 1.

Instructions.

- Homework is due **Wednesday, November 16, 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.

Version history.

1. Initial version.

1. Clarke differentials.

Recall the definition of Clarke differential:

$$\partial f(w) := \operatorname{conv}\left\{\lim_i \nabla f(w_i) \ : \ w_i \to w,\ \nabla f(w_i) \text{ exists},\ \lim_i \nabla f(w_i) \text{ exists}\right\},$$

where "conv" denotes the convex hull. Additionally, given a set $U \subseteq \mathbb{R}^d$, define subgradients and supergradients *relative to* U as

$$\partial_{\mathrm{s}} f(w) := \left\{s \in \mathbb{R}^d : \forall w' \in U \centerdot f(w') \geq f(w) + \langle s, w' - w\rangle\right\},$$
$$\partial_{\mathrm{u}} f(w) := \left\{s \in \mathbb{R}^d : \forall w' \in U \centerdot f(w') \leq f(w) + \langle s, w' - w\rangle\right\}.$$

Define a function $g : \mathbb{R}^2 \to \mathbb{R}$ as

$$g(w) := |w_1| - |w_2|.$$

(a) Prove that g is locally Lipschitz.

(b) Prove that $\partial g(0) = \{w \in \mathbb{R}^2 : \|w\|_\infty \leq 1\}$.
Hint: the vector field view of this function in lecture 18 makes this much easier, and yes we gave the answer in class, but give a rigorous check here.

(c) Prove that for every $\tau > 0$, no element of $\partial g(0)$ is a subgradient or supergradient of g relative to $\{w \in \mathbb{R}^2 : \|w\|_\infty < \tau\}$.
Remark: this gives a rigorous backing to the comment in lecture that we can not use local subgradients and supergradients as the basis of a general differential.

(d) Prove that for any starting point $w \in \{(a, 0) : a \in \mathbb{R}\} \subseteq \mathbb{R}^2$, the Clarke differential inclusion has multiple solutions.
Remark: we handled $w = (0, 0)$ in class and you can just say that, don't need to reprove that case. For full points, you must explicitly provide multiple paths and show that they satisfy the differential inclusion.

(e) Prove that for any starting point $w \in \{(a, b) : a \in \mathbb{R}, b \in \mathbb{R} \setminus \{0\}\}$, the Clarke differential inclusion has a unique solution.

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

2. Eigenvalues of expected kernel.

In the strongly-convex NTK proof scheme, we needed the *kernel Gram matrix* to have nicely-behaved eigenvalues. In this problem, we will work out these eigenvalues in the infinite-width shallow network setting with differentiable activations; in homework 3, we'll convert this into a finite-width bound.

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 activation (in the NTK setting) 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.

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.)*

3. Experiments near initialization.

This problem will check some basic properties near initialization. Starter code is provided in `hw2.py`: it loads the classical *iris data*, runs (linear) logistic regression on the two chosen classes, and defines and defines a neural net class. (For further python help, see for instance my `pytorch` tutorial: `https://mjt.cs.illinois.edu/ml/pytorch_basics.pdf` , which is generated from a jupyter notebook at `https://mjt.cs.illinois.edu/ml/pytorch_basics.ipynb` .)

Note that this data is linearly separable and very small, so this problem is a toy warm-up, no more.

Coding note: some `pytorch` versions may complain about type errors and/or about the `dtype` argument not existing. You can fix both issues by using expressions of the form "`.type(X.dtype)`" or appropriate equivalents, as already exist in the code.

(a) Using the provided shallow ReLU network class, plot empirical logistic risk curves for 4096 iterations with step size 1.0 and widths $m \in \{4, 64, 256, 1024\}$. (That is, the horizontal axis is iterations, vertical axis is empirical logistic risk.)

For full points: include the plot here, and describe it qualitatively in 1-3 sentences.

Coding remark: when $m = 4$, with probability $1/8$, the output layer is all $+1$ or all -1, and the training error will stay large. If your plot happens to have such a situation (for $m = 4$ only of course), you do not need to explain it.

(b) For the same setup as the previous part (including the four choices of m), plot $\|W_t - W_0\|^{4/3}/m^{1/6}$, with t as the horizontal axis once again. (These exponents are chosen to match Lemma 4.1 in the typed notes.)

For full points: include the plot here, and describe it qualitatively in 1-3 sentences, including a discussion of whether you think it supports or negates parts of the NTK story.

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