CS 540 DLT — Homework 3.

your NetID here.

Version $1 + \epsilon$.

Instructions.

- This homework is due **Wednesday, November 17, 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.

$1 + \epsilon$. $\ln 1/\delta$ swapped with $\ln n/\delta$ in 3(d); coding note about type errors in 4; coding remark about narrow width in 4(a).

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 $\partial g(0) = \{w \in \mathbb{R}^2 : \|w\|_\infty \leq 1\}$.

(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 resolves a question from lecture, regarding whether Clarke differentials must always be locally subgradients or supergradients.

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

2. Norm growth with L-homogeneous networks.

Consider the setting of Theorem 10.3 from the typed lecture notes, where we showed that the smoothed margin is nondecreasing with homogeneous networks. As was the case there, define

$$m_i(w) = y_i f(x_i; w) \qquad \text{and} \qquad \mathcal{L}(w) = \sum_{i=1}^{n} \ell(m_i(w)) = \sum_{i=1}^{n} \exp(-m_i(w)),$$

and suppose $\mathcal{L}(w_0) < 1$ (simplifying notation by taking $t_0 = 0$), and that each m_i is L-homogeneous and locally Lipschitz. Assume chain rules as needed throughout this problem.

(a) Show that $\|w_t\| \to \infty$ and $\mathcal{L}(w_t) \to 0$.

(b) Suppose additionally that f is a network with L linear layers, and is 1-homogeneous with respect to any individual layer i, meaning that for any i and any $c \geq 0$,

$$f\left(x; (W_L, \ldots, cW_i, \ldots W_1)\right) = cf\left(x; (W_L, \ldots, W_i, \ldots W_1)\right).$$

Prove that $\min_i \|W_i(t)\| \to \infty$.

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

3. Smooth projected NTK analysis.

This problem will consider shallow ReLU networks with training of only the input layer weights. In particular, the setting will match Lemma 4.1 in the typed notes, where given $a \in \{\pm 1\}^m$ chosen uniformly at random, and $W_0 \in \mathbb{R}^{m\times d}$ chosen with iid standard Gaussian coordinates, the prediction mapping f and its linearization f_0 are defined as

$$f(x;W) := \frac{1}{\sqrt{m}} \sum_j a_j \sigma(w_j^\intercal x),$$

$$f_0(x;W) = f(x;W_0) + \left\langle \nabla_W f(x;W_0), W - W_0 \right\rangle = \frac{1}{\sqrt{m}} \sum_j a_j w_j^\intercal x \sigma(w_{0,j}^\intercal x).$$

The first two parts of the problem will not use any of this structure, except that $W_0 \in \mathbb{R}^{m\times d}$ is a matrix, and we use the Frobenius norm to measure distance.

(a) Let $\mathcal{B} = \{W \in \mathbb{R}^{m\times d} : \|W - W_0\| \le R\}$ be a ball of matrices in Frobenius norm, and let $\Pi_B(V) := \arg\min_{W\in\mathcal{B}} \|V - W\|^2$ denote orthogonal projection onto $\mathcal{B}$. Prove that

$$\Pi_B V = \begin{cases} V & V \in \mathcal{B}, \\ W_0 + \frac{R(V-W_0)}{\|V-W_0\|} & V \notin \mathcal{B}. \end{cases}$$

(b) Prove that $\|\Pi_B V - \Pi_B W\| \le \|V - W\|$, meaning Π_B is nonexpansive.

Remark. This property is true for arbitary convex compact sets (and even more generally; we worked with something similar for subspaces in the last homework), but a direct geometric proof is possible here. Indeed, note that it suffices to consider the plane containing V, W, and W_0.

For the rest of the problem, let $\ell(z) := \ln(1 + \exp(-z)$ denote the logistic loss, and define the empirical logistic risk as

$$\widehat{\mathcal{R}}(W) := \frac{1}{n} \sum_{i=1}^n \ell(y_i f(x_i;W)),$$

where $y_i \in \{\pm 1\}$ and $\|x_i\| \le 1$. Consider the *projected* gradient descent iteration

$$W_{i+1} := \Pi_B \left(W_i - \nabla \widehat{\mathcal{R}}(W_i) \right).$$

(c) Show that $\|\nabla \widehat{\mathcal{R}}(W_i)\|^2 \le \widehat{\mathcal{R}}(W_i)$.

Hint. If you are stuck, see section 9.4 in the typed notes.

(d) Let $Z \in \mathbb{R}^{m\times d}$ be given, set $R := \max\{1, \|Z - W_0\|\}$ to be the radius of $\mathcal{B}$, let $\delta \in (0, 1/e)$ be given, and suppose $m \ge 32^6 R^8 \ln(n/\delta)^{3/2}$. Show that with probability at least $1 - \delta$,

$$\frac{1}{t} \sum_{i<t} \widehat{\mathcal{R}}(W_i) \le 4\widehat{\mathcal{R}}(Z) + \frac{2R^2}{t}.$$

Remark. The old Lemma 4.1 doesn't seem to be strong enough to prove this; however, Lemma 4.1 has been strengthened with a second part in the typed notes.

Remark. It's possible to use implicit bias (as in problem 3 of homework 2) to handle gradient descent without projection, but it makes the proof longer.

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

4. Experiments near initialization.

This problem will check some basic properties near initialization. Starter code is provided in `hw3.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.)*