Stats 231C - Homework 3

Peter Racioppo (103953689)

May 11, 2020

Exercise 1. *(Shallow logical circuit, 5 points). Construct a two level logical circuit that computes parity for n inputs with a level of ORs followed by an AND output gate.*

An OR gate outputs 0 if and only if all its inputs are 0. Thus, an OR gate "recognizes" the unique input σ whose digits are 1 wherever the OR gate has a negation and 0 everywhere else, in which case the OR gate outputs a 0. Since the outputs of all the other OR gates are 1s, the AND gate will output a 0 if and only if one of the OR gates has recognized a particular σ with the correct parity. If we let a final output of 0 correspond to even parity, than each OR gate must recognize an even parity σ. There are $2^n/2 = 2^{n-1}$ even vectors, so we need this many OR gates. This is the reverse of the situation in a circuit in disjunctive normal form (DNF), since in the DNF case we check odd parity, while in the CNF case we check even parity.

Exercise 2. *(Saddle points, 5 points). Prove the following statement: Let $f : S \subseteq R^2 \to R$ be a piecewise constant function where the pieces are defined by lines. If the intersection x_0 of two of the lines is a saddle-type point, then f cannot be implemented by a network with a single hidden layer of linear threshold units (and a single affine output unit). Here the intersection $p \in R^2$ of two lines l_1 and l_2 is a saddle-type point of f if in any circle C centered at p, 1) in two opposite sectors S_1 and S_2 of C formed by the lines l_1 and l_2 the function f is constant with values a_1 and a_2; 2) in the other two opposite sectors $f(x) = b_1$ for all x near l_1 bordering $S_1, f(x) = b_2$ for all x near l_1 bordering S_2, and $b_1 < a_1$ and $b_2 \leq a_2$.*

We know that f must map every circle region to a unique constant. Given an input x, we can express the output of a linear threshold network as $g(x) = A^T\sigma(Wx+b)+c$, where $\sigma(s) = 0$ if $s < 0$ and $\sigma(s) = 1$ if $s \geq 0$. Since there are only two linear boundaries that define the regions in which f takes the values a_1, a_2, b_1, b_2, anymore than two linear threshold units (LTUs) will not improve the network's performance in these regions. Thus, we consider the performance of two linear threshold units in only these regions.

Suppose, without loss of generality, that the first unit returns a 0 if it is input a

point on the side of l_1 corresponding to S_1 and 1 otherwise, and that the second unit returns a 0 if it is input a point on the side of l_2 corresponding to S_1 and 1 otherwise. Thus, the outputs of the LTUs (before the affine unit) for the four circle regions are: (0,0) in S_1, where $f = a_1$, (0,1) in the region in which $f = b_2$, (1,0) in the region in which $f = b_1$, and (1,1) in S_2, where $f = a_2$. There are four possible such assignments, which we could check individually, but I think it suffices to check only one. Let A_1 be the element of A corresponding to the first LTU and A_2 the second. Then in order for g to equal f in every region, we have the following system of four equations in three unknowns:

$$A_1 \times 0 + A_2 \times 0 + c = a_1$$
$$A_1 \times 1 + A_2 \times 0 + c = b_1$$
$$A_1 \times 0 + A_2 \times 1 + c = b_2$$
$$A_1 \times 1 + A_2 \times 1 + c = a_2$$

From the first three equations, $c = a_1$, $A_1 = b_1 - a_1$, and $A_2 = b_2 - a_1$, and it follows from the last equation that $b_1 + b_2 = a_1 + a_2$. But this is a contradiction, since $b_1 < a_1$ and $b_2 \leq a_2$. Hence, g cannot represent f in all four regions.

Exercise 3. *(Compactly supported functions, 15 points). Consider functions of the form* $x \mapsto W^{(L)}\sigma(W^{(L-1)}...\sigma(W^{(2)}\sigma(W^{(1)}x + b^{(1)}) + b^{(2)})... + b^{(L-1)}) + b^{(L)}$ *represented by feedforward networks with* $d \geq 2$ *input units,* $L - 1$ *layers of ReLUs, and an affine output layer. For a given function* $f : R^d \mapsto R^{n_L}$*, let* $supp(f) = \{x \in R^d : f(x) \neq 0\}$.

• *(5 points) Show that a network with* $L = 3$ *can represent a function* f *with* $supp(f) = B_1(0) = \{x = (x_1, ..., x_d) \in R^d : \sum_{k=1}^{d} |x_k| \leq 1\}$.

We know that f is such that $f(x) \neq 0 \iff \sum_{k=1}^{d} |x_k| \leq 1$. Let us construct a network in which every $x_k \in x$, $k \in [1, d]$ is routed to two ReLUs, which compute $\sigma(W_1 x_k + b_1)$ and $\sigma(W_2 x_k + b_2)$, respectively. Let $W_1 = I$ and $b_1 = 0$ and let $W_2 = -I$ and $b_2 = 0$. Then the output of the first ReLU will be x_k if $x_k \geq 0$ and 0 if $x_k < 0$. The second ReLU will return 0 if $x_k \geq 0$ and $-x_k$ if $x_k < 0$. The sum of the two units is $\text{sign}(x_k) \times x_k + 0 = |x_k|$. In total, there are thus $2d$ ReLUs on the first layer. On the second layer, we need an additional ReLU to sum all of the $2d$ outputs from the first layer and determine whether the sum is less than or equal to 1. Let W and b be the weights of this ReLU, and let $W = -I$ and $b = 1$. Then this ReLU computes $\sigma(-\sum_{k=1}^{d} |x_k| + 1)$, which returns $-\sum_{k=1}^{d} |x_k| + 1$ if $\sum_{k=1}^{d} |x_k| \leq 1$ and 0 otherwise. Our affine unit in the third layer maps this constant input to a constant output.

• *(10 points) Show that if* f *is a function represented by a network with* $L = 2$ *and* $supp(f)$ *is compact, then* $f \equiv 0$.

If $L = 2$, the network has only a single ReLU layer. Let us represent the network's output as $g(x) = a^T \sigma(w^T x + b) + c$. If $g(x)$ is to represent $f(x)$, we

must have that $S := \{x|a^T\sigma(w^Tx + b) + c \neq 0\}$ is compact. Now, $g(x) = a^T\sigma(w^Tx+b)+c = a^T(w^Tx+b)+c$ if $w^Tx+b \geq 0$ and $g(x) = c$ if $w^Tx+b < 0$. We can rewrite $a^T(w^Tx + b) + c$ as $(wa)^Tx + (a^Tb + c) = w_2^Tx + b_2$.

Now suppose that $c \neq 0$. Then, $\{x|w^Tx + b < 0\} \in \text{supp}(f)$ and $\{x|w^Tx + b \geq 0, a_2^Tw + b_2 \neq 0\} \in \text{supp}(f)$, but $\{x|w^Tx + b \geq 0, a_2^Tw + b_2 = 0\} \notin \text{supp}(f)$. Thus, $\text{supp}(f)$ is the entire input space except for a line, so $\text{supp}(f)$ is not closed, and therefore not compact. If $c = 0$, $\{x|w^Tx + b < 0\} \notin \text{supp}(f)$ and $\{x|w^Tx + b \geq 0, a_2^Tw + b_2 \neq 0\} \in \text{supp}(f)$, while $\{x|w^Tx + b \geq 0, a_2^Tw + b_2 = 0\} = \{x|w^Tx + b \geq 0, a^T(w^Tx + b) = 0\} \notin \text{supp}(f)$. This last condition implies that either $a = 0$ or $\{x|w^Tx + b = 0\} \notin \text{supp}(f)$. If $\{x|w^Tx + b = 0\} \notin \text{supp}(f)$, then $\text{supp}(f)$ is again not closed, and therefore not compact. Thus, $\text{supp}(f)$ is compact if and only if $c = 0$ and $a = 0$, which implies that $g(x) = 0$ for all x. Since, by assumption, the network represents f, we have that $f \equiv 0$.

• *(voluntary) For every d large enough, show that there is an input data distribution p on R^d and a function f that can be represented by a network with $L = 3$ and width polynomial in d, for which, if g is any function expressed by a network with $L = 2$ and width at most $c_1 exp(c_2 d)$, then $E_p[(f(x)g(x))^2] \geq c_3$.*

Exercise 4. *(Linear pieces and approximation, 5 points). Prove the following statement: Let $d, L, N \in N$, and $f \in C^2([0,1]^d)$, where f is not an affine linear function. Consider an activation function σ that is piecewise linear with p pieces. If g is a function represented by a neural network with L - 1 layers of ReLUs of width at most N and an affine output layer, then*

$$\|f - g\|_\infty \geq c(pN)^{-2(L-1)}.$$

Hint: Bound the number of linear regions of any function represented by the network and then bound the error from a piecewise linear to the target function. You can use the following proposition.

Proposition. *Let $f \in C^2([a,b]), a < b < \infty$ not affine. Then there exists a constant $c = c(f) > 0$ so that, for every $p \in N$, $\|f - g\|_\infty \geq cp^{-2}$ for all g which are piecewise linear with at most p pieces.*

Each unit computes p pieces and each unit connects to at most N units in the next layer. Thus, there are at most $(pN)^{L-1}$ pieces in total. (To be more explicit, the pieces are multiplicative because each linear boundary can divide a region into at most two subregions. Thus, at each layer, we multiply the number of pieces computed by all previous layers by pN, and it follows by induction that the total is $(pN)^{L-1}$.) By the proposition, $\|f - g\|_\infty \geq c((pN)^{L-1})^{-2} = c(pN)^{-2(L-1)}$.