# MathDNN HW 3

2018-13260 차재현

## 1 Problem 3

Let $j$ be any element in $\{1, \cdots, k\}$, and denote $\sum_j$ as $\sum_{j=1}^k$.
(a) Since $\exp(f_j) > 0$ for all $j$, we have

$$0 < \frac{\exp(f_y)}{\sum_j \exp(f_j)} < \frac{\exp(f_y)}{\exp(f_y)} = 1.$$

Therefore its $-\log$ value $l^{\mathrm{CE}}(f, y)$ is strictly larger than zero, and $0 <$ at the left side implies $l^{\mathrm{CE}}(f, y) < \infty$.
(b) Inserting $f = e_y$ in (a), we get

$$l^{\mathrm{CE}}(\lambda e_y, y) = -\log\left(\frac{\exp(\lambda y)}{(k-1)\exp(0) + \exp(\lambda y)}\right) = -\log\left(\frac{\exp(\lambda y)}{\exp(\lambda y) + (k-1)}\right)$$

Since k is fixed, as $\lambda$ goes to $\infty$, term inside the $-\log$ at the right side converges to 1. Continuity of log ($-\log$, equivalently) gives our CE loss converges to 0.

## 2 Problem 4

Choose $x$, and assume $I(x) = i$. Assume that $f_i$ is strictly larger than others. Let $\epsilon_0 = f_i(x) - \max(f_1, \cdots, \hat{f_i}, \cdots, f_k)$, where $\hat{f_i}$ means $\hat{f_i}$ is omitted. Since differentiable function is continuous, usual $\epsilon$-$\delta$ method gives $\delta$ s.t. whenever $|x - y| < \delta$, $I(x) = i$. In this range, $f(y) = f_i(y)$, so $f'(y) = f_i'(y)$.
First, assume that, for any $1 \leq i < j \leq k$ the set $\{x \in \mathbb{R} \mid f_i(x) = f_j(x)\}$ is nowhere dense. Then for almost all $x$, if $I(x) = i$ then there is a neighborhood $U$ of $x$ s.t. $I(U) = i$. Also, if we denote $V$ the union of all $U$ as $x$ varies, then $\mathbb{R} - V$ is discrete by nowhere dense property. In this case, applying first argument concludes $f$ is differentiable a.e..
If $\{x \in \mathbb{R} \mid f_i(x) = f_j(x)\} \cap U$ is dense in $U$, since differentiable function is continuous we get $U \subset \{x \in \mathbb{R} \mid f_i(x) = f_j(x)\}$. In this case, if $f \neq f_i$ nothing differs to the previous case, and if $f = f_i$ in $W \subset U$ we have $f' = f_i' = f_j'$. This

means, the property $\{x \in \mathbb{R} \mid f_i(x) = f_j(x)\} \cap U$ being dense in $U$ does not affect the differentiability of $f$ on $U$, so we can reduce the case to $\{x \in \mathbb{R} \mid f_i(x) = f_j(x)\}$ being nowhere dense. Iterating over all pair $(i, j)$, which is finite step, we return to first case.

# 3 Problem 5

(a) If $z < 0$, $\sigma(\sigma(z)) = \sigma(0) = 0$. If $z \geq 0$, $\sigma(\sigma(z)) = \sigma(z) = z$.
(b) Derivative of $\sigma'(z) = e^z/(1 + e^z)$ is $-e^{-z}/(1 + e^{-z})^2$, which is defined on $\mathbb{R}$ and

$$\lim_{z \to -\infty} \frac{-e^{-z}}{(1 + e^{-z})^2} = \lim_{z \to \infty} \frac{-e^{-z}}{(1 + e^{-z})^2} = 0.$$

Thus it is bounded. Let $M = \sup_z |-e^{-z}/(1 + e^{-z})^2|$. Then for $x < y$,

$$|\sigma'(x) - \sigma'(y)| \leq \int_x^y \left| \frac{-e^{-z}}{(1 + e^{-z})^2} \right| dz \leq M|y - x|.$$

Therefore $\sigma'$ is Lipshitz continuous.
ReLU has its derivative at $\mathbb{R} - \{0\}$, which value is 0 when $z < 0$ and 1 when $z > 0$. In this case, for any $M > 0$ if we let $x = -\frac{1}{M}, y = \frac{1}{M}$, we get

$$1 = 1 - 0 > \frac{M}{3}(y - x).$$

Letting $M \to \infty$ gives ReLU is not Lipshitz continuous.
(c) We will use the relation

$$\rho(z) = \frac{1 - e^{-2z}}{1 + e^{-2z}} = \frac{2}{1 + e^{-2z}} - 1 = 2\sigma(2z) - 1.$$

Denote $x = y_0$, and $y_i^\rho = \rho(A_i y_{i-1} + b_i), y_i^\sigma = \sigma(C_i y_i + d_i)$ for $i \geq 1$. Initially, set $C_1 = 2A_1$ and $d_1 = 2b_1$. From the above relation, we get $y_1^\rho = 2y_1^\sigma - \mathbf{1}_1$, where $\mathbf{1}_i = (1 \cdots 1)^\top \in \mathbb{R}_i^n$.
Next, for all $1 < i < L$, let $C_i = 4A_i$ and $d_i = 2b_i - 2A_i \mathbf{1}_{i-1}$. Then, inductively we get

$$\begin{aligned} y_i^\rho = \rho(A_i y_{i-1} + b_i) &= 2\sigma(2A_i y_{i-1}^\rho + 2b_i) - \mathbf{1}_i \\ &= 2\sigma(4A_i y_{i-1}^\sigma + 2b_i - 2A_i \mathbf{1}_{i-1}) - \mathbf{1}_i \end{aligned}$$

$$= 2\sigma(C_i y_{i-1}^{\sigma} + d_i) - \mathbf{1}_i$$

$$= 2y_i^{\sigma} - \mathbf{1}_i.$$

In fact, this is valid for $i = 2$ by the relation $y_1^{\rho} = 2y_1^{\sigma} - \mathbf{1}_1$, and the expansion implies if it holds for $i = k - 1$, so does for $i = k$ unless $i$ exceeds $L - 1$.
Finally, let $C_L = 2A_L, d_L = b_L - A_L \mathbf{1}_{L-1}$. Then

$$y_L^{\rho} = A_L y_{L-1}^{\rho} + b_L = A_L(2y_{L-1}^{\sigma} - \mathbf{1}_{L-1}) + b_L = C_L y_{L-1}^{\sigma} + d_L = y_L^{\sigma}.$$

The last step must appears (i.e., it does not conflict to first step) because $L > 1$.

## 4 Problem 6

First of all, since every $a_i, b_i, u_i$ appears in the left entry of $l$, target function is differentiable w.r.t all of these.
When $j$-th output is dead, $\sigma(a_j X_i + b_j) = 0$. Then our target function can be rewritten as

$$\frac{1}{N}\sum_i l(f_\theta(X_i), Y_i) = \frac{1}{N}\sum_i l\left(\sum_{k\neq j} u_k \sigma(a_k X_i + b_k), Y_i\right) + \frac{1}{N}\sum_i l(u_j \sigma(a_j X_i + b_j), Y_i) \tag{1}$$

$$= \frac{1}{N}\sum_i l\left(\sum_{k\neq j} u_k \sigma(a_k X_i + b_k), Y_i\right), \tag{2}$$

where $\sum_i$ indicates summation over $i = 1$ to $i = N$. In this case, (1) is constant on $a_j, b_j, u_j$, so differential w.r.t these becomes zero. Therefore $a_j, b_j, u_j$ are unchanged after gradient step, and then $\sigma(a_j X_i + b_j) = 0$ again holds. Inductively applying this argument over iteration step, we conclude $j$-th ReLU output remains dead.

## 5 Problem 7

When we use leaky ReLU in (1), the term $l(u_j\sigma(a_j X_i + b_j), Y_i)$ is no longer trivial. We get

$$L = \frac{1}{N}\sum_i l(u_j \sigma(a_j X_i + b_j), Y_i) = \frac{1}{N}\sum_i l(\alpha u_j(a_j X_i + b_j), Y_i),$$

and differentiating w.r.t each $a_j, b_j, u_j$ gives

$$\frac{\partial L}{\partial a_j} = \frac{1}{N}\sum_i l'(\alpha u_j(a_j X_i + b_j), Y_i)\alpha u_j X_i,$$

$$\frac{\partial L}{\partial b_j} = \frac{1}{N}\sum_i l'(\alpha u_j(a_j X_i + b_j), Y_i)\alpha u_j,$$

$$\frac{\partial L}{\partial u_j} = \frac{1}{N}\sum_i l'(\alpha u_j(a_j X_i + b_j), Y_i)\alpha a_j X_i.$$

Now, if $l'(\alpha u_j(a_j X_i + b_j), Y_i) = 0$, gradient of target function is zero, which is absurd. By the same reason, $u_j \neq 0$.

Thus $L$ is not a constant in a neighborhood of $(a_j, b_j, u_j) \in \mathbb{R}^3$, not all of these three derivatives are zero. Therefore gradient is no longer exactly vanishes.