Exercises 2.5

Exercise 2.5.1

a. Show that the logistic function σ satisfies the inequality $0 < \sigma'(x) \leq \frac{1}{4}$, for all $x \in \mathbb{R}$.

b. How does the inequality changes in the case of the functions σ_c?

Exercise 2.5.2

Let $S(x)$ and $H(x)$ denote the bipolar step fucntion and the Heaviside function, respectively. Show that:

a. $S(x) = 2H(x) - 1$

b. $ReLU(x) = \frac{1}{2}x(S(x) + 1)$

Exercise 2.5.3

Show that the softplus function, $sp(x)$, satisfies the following properties:

a. $sp'(x) = \sigma(x)$, where $\sigma(x) = \frac{1}{1+e^{-x}}$

b. Show that $sp(x)$ is invertible with inverse $sp^{-1}(x) = \ln(e^x - 1)$

c. Use the softplus function to show the formula $\sigma(x) = 1 - \sigma(-x)$

Exercise 2.5.4

Show that $\tanh(x) = 2\sigma(2x) - 1$

Exercise 2.5.5

Show that the softsign function, $so(x)$, satisfies the following properties:

a. Its sctrictly increasing;

b. Its is onto $(-1, 1)$, with the inverse $so^{-1}(x) = \frac{x}{1-|x|}$, for $|x| < 1$.

c. $so(|x|)$ is subadditive, i.e., $so(|x + y|) \leq so(|x|) + so(|y|)$.

Exercise 2.5.6

Show that the softmax function is invariant with respect to the addition of constant vectors $\boldsymbol{c} = (c_1 \ldots c_n)^T$, i.e.,

$$softmax(y + \boldsymbol{c}) = softmax(y).$$

This property is used in practice by replacing $\boldsymbol{c} = -\max_i y_i$, fact that leads to a more stable numerically variant of this function.

Exercise 2.5.7

Let $\rho : \mathbb{R}^n \to \mathbb{R}^n$ defined by $\rho(y) \in \mathbb{R}^n$, with $\rho(y)_i = \dfrac{y_i^2}{\|y\|^2}$. Show that:

a. $0 \leq \rho(y)_i \leq 1$ and $\sum_i \rho(y)_i = 1$.

b. The function ρ is invariant with to multiplication by nonzero constant, i.e., $\rho(\lambda y) = \rho(y)$ for any $\lambda \in \mathbb{R}/0$. Taking $\lambda = \frac{1}{\max_i y_i}$ leads in practice to a more stable version of this function.

Exercise 2.5.8 (cosine squasher)

Show that the function $\varphi(x) = \frac{1}{2}(1 + \cos(x + \frac{3\pi}{2}))1_{[-\frac{\pi}{2},\frac{\pi}{2}]}(x) + 1_{(\frac{\pi}{2},\infty)}(x)$ is a squashing function.

Exercise 2.5.9

a. Show that any squashing function is a sigmoidal function.

b. Give an example of a sigmoidal function which is not a squashing function.

SOLUTIONS

2.5.1 (a)

Computing the derivative of σ we find: $\sigma'(x) = \frac{d}{dx}\frac{1}{1+e^{-x}} = \frac{d}{dx}\frac{e^x}{1+e^x} = \frac{e^x}{(1+e^x)^2}$. From the inequality $1 \le (1+e^x)^2$ and the non-negativeness of the exponential function follows that $0 \le \frac{e^x}{(1+e^x)^2}$.

Now lets prove that in $x = 0$ the function has a local maximum in $[-1, 1]$, this will imply $0 \le \frac{e^x}{(1+e^x)^2} \le \sigma'(0)$, $\sigma'(0) = \frac{1}{4}$. By computing the first derivative of σ' we find: $\sigma''(x) = e^x \frac{1-e^x}{(1+e^x)^3}$. The critical will be found by solving the equation $\sigma''(x) = 0$.

From $\sigma''(x) = e^x \frac{1-e^x}{(1+e^x)^3} = 0$ follows that $1 - e^x = 0$, is straigth-forward to check that the solution is $x = 0$. It rests to determine the nature of the extremizing point. To achieve this goal is necessary to calculate the second derivative of σ'.

$$\begin{aligned}\sigma'''(x) &= \frac{d}{dx}\frac{e^x - e^{2x}}{(1+e^x)^3} = \frac{(e^x - 2e^{2x})(1+e^x)^3 - 3(1+e^x)^2 e^x (e^x - e^{2x})}{(1+e^x)^6} \\ &= \frac{e^x\{1 - 4e^x + e^{2x}\}(1+e^x)^2}{(1+e^x)^6} = \frac{e^x\{1 - 4e^x + e^{2x}\}}{(1+e^x)^4}\end{aligned}$$

We clearly have $\sigma'''(0) < 0$, then $x = 0$ is a local maximum for σ', i.e $\forall x \in [-1, 1],\ \sigma'(x) \le \frac{1}{4}$. On the other hand, the function σ' decreases on the intervals $(-\infty, -1)$ and $(1, \infty)$ this implies that:

$\sup\limits_{x\in(1,\infty)} \sigma'(x) = \dfrac{e}{(1+e)^2} = \dfrac{e^{-1}}{(1+e^{-1})^2} = \sup\limits_{x\in(-\infty,-1)} \sigma'(x)$. From the fact that $\dfrac{e}{(1+e)^2} < \dfrac{1}{4}$ follows that $0 \le \sigma'(x) \le \frac{1}{4}$ is valid $\forall x \in \mathbb{R}$. $\square$

2.5.1 (b)

The inequality changes to: $0 \le \sigma_c'(x) \le \frac{c}{4},\ \forall x \in \mathbb{R}$. From the expression $\sigma_c(x) = \frac{1}{1+e^{-cx}}$, $c > 0$ one finds that $\sigma_c'(x) = \frac{d}{dx}\frac{e^{cx}}{1+e^{cx}} = c\dfrac{e^{cx}}{(1+e^{cx})^2}$. By the chain rule it can be easily verified that all the computations made for $\sigma'(x)$ in 2.5.1.a, can by applied to $\sigma_c'(x)$, having in mind the relationship $\sigma_c'(x) = c\sigma'(cx)$.

Then, one finds: $\sigma_c''(x) = c^2 e^{cx}\frac{1-e^{cx}}{(1+e^{cx})^3}$, this implies that $x = 0$ is a critical point. Using the same relationship is clear that $\sigma_c'''(x)\Big|_{x=0} = c^3 \frac{e^{cx}\{1-4e^{cx}+e^{2cx}\}}{(1+e^{cx})^4}\Big|_{x=0} < 0$. Then, $x = 0$ is a maximum.

Arguing like in 2.5.1.a, on the interval $[-1, 1]$, $\sigma_c'(0) = \frac{c}{4}$ is a local maximum. More over, the function σ_c' decreases on the intervals $(-\infty, -1)$ and $(1, \infty)$, implying:

$$\sup_{x\in(1,\infty)} \sigma_c'(x) = \frac{ce^c}{(1+e^c)^2} = \frac{ce^{-c}}{(1+e^{-c})^2} = \sup_{x\in(-\infty,-1)} \sigma_c'(x)$$

Lets now prove the inequality $\dfrac{ce^c}{(1+e^c)^2} < \dfrac{c}{4}$. We have:

$$\begin{aligned}\frac{ce^c}{(1+e^c)^2} &= \frac{c}{\frac{(1+e^c)^2}{e^{\frac{2c}{2}}}} = \frac{c}{(\frac{1+e^c}{e^{\frac{c}{2}}})^2} \\ &= \frac{c}{(e^{-\frac{c}{2}} + e^{\frac{c}{2}})^2} < \frac{c}{(1 - \frac{c}{2} + 1 + \frac{c}{2})^2} = \frac{c}{4}\end{aligned}$$

Where we have used the inequality $1 + x \le e^x,\ \forall x \in \mathbb{R}$. The later shows $\sigma_c'(0)$ is a global maximum, i.e $0 \le \sigma_c'(x) \le \frac{c}{4}$ is valid $\forall x \in \mathbb{R}$. $\square$

2.5.2 (a)

From the Heaviside function definition one has:

$$\begin{aligned} 2H(x) - 1 &= \begin{cases} 2 - 1 \text{ if } x > 0 \\ 2(0) - 1 \text{ otherwise} \end{cases} \\ &= \begin{cases} 1 \text{ if } x > 0 \\ -1 \text{ otherwise} \end{cases} = S(x). \ \square \end{aligned}$$

2.5.2 (b)

We know $ReLU(x) := \max(0, x)$. Consider the identities $\max(0, x) = \frac{1}{2}\{x + |x|\}$,

$$|x| = \begin{cases} 1x \text{ if } x > 0 \\ -1x \text{ otherwise} \end{cases} = xS(x).$$ Substituting the last identity into the first one yields:

$ReLU(x) = \frac{1}{2}(x + xS(x)) = \frac{1}{2}x(1 + S(x))$. $\square$

2.5.3 (a)

The identity immediately follows from the application of the chain rule to the function $\ln(1 + e^x)$. In fact, we have: $sp'(x) = \frac{d}{dx}\ln(1 + e^x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}} = \sigma(x)$. $\square$

2.5.3 (b)

The function e^x is well known to never be zero, then $1 + e^x > 0, \forall x \in \mathbb{R}$. This implies $sp'(x) \neq 0, \forall x \in \mathbb{R}$. Then by the inverse function theorem the function is invertible. We can now compute its inverse, which is given by:

$$sp(x) = y = \ln(e^x + 1) \implies e^y = e^x + 1 \implies x = sp^{-1}(y) = \ln(e^y - 1). \ \square$$

2.5.3 (c)

Let $F(x) := x + sp(-x) - sp(x)$. It happens that derivative of F is 0, then by the chain rule, the linearity of the derivative operator and the relationship proved in 2.5.3.a yield:

$\frac{d}{dx}\left[x + sp(-x)\right] = 1 - \sigma(-x) = \frac{d}{dx}sp(x) = \sigma(x)$. Lets now prove the claim aforementioned to complete the proof. We have:

$$\begin{aligned} \frac{d}{dx}F(x) &= \frac{d}{dx}\left[x + sp(-x) - sp(x)\right] = \frac{d}{dx}\left[x + \ln(\frac{e^{-x}+1}{e^x+1})\right] \\ &= 1 + \frac{e^x+1}{e^{-x}+1}\frac{d}{dx}\left[\frac{e^{-x}+1}{e^x+1}\right] = 1 + \frac{e^x+1}{e^{-x}+1}\frac{-e^{-x}(e^x+1) - e^x(1+e^{-x})}{(e^x+1)^2} \\ &= 1 + \frac{e^x+1}{e^{-x}+1}\frac{-e^{-x}(e^x+1) - (1+e^x)}{(e^x+1)^2} = 1 + \frac{e^x+1}{e^{-x}+1}\frac{-(e^x+1)(1+e^x)}{(e^x+1)^2} = 1 - 1 = 0. \ \square \end{aligned}$$

2.5.4 (a)

From the tanh definition we have:

$$\begin{aligned}\tanh(x) := \frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{e^x}{e^x + e^{-x}} - \frac{e^{-x}}{e^x + e^{-x}} = \frac{e^{2x}}{e^{2x}+1} - \frac{e^{-x}}{e^x + e^{-x}} \\ = \sigma(2x) - \frac{e^{-x}}{e^{-x}(1+e^{2x})} = \sigma(2x) - \frac{1}{1+e^{2x}} \\ = \sigma(2x) - \frac{1+e^{2x}-e^{2x}}{1+e^{2x}} = \sigma(2x) - \{1 - \frac{e^{2x}}{1+e^{2x}}\} \\ = \sigma(2x) - \{1-\sigma(2x)\} = 2\sigma(2x) - 1. \ \square\end{aligned}$$

2.5.5 (a)

<u>CASE $x > 0$</u>:
Taking the derivative of $so(x)$ we find: $\frac{d}{dx}so(x) = \frac{d}{dx}\frac{x}{1+x} = \frac{1(1+x)-x(1)}{(1+x)^2} = \frac{1}{(1+x)^2} > 0$. This implies $so(x)$ is strictly increasing on the interval $(0,\infty)$.

<u>CASE $x < 0$</u>:
Taking the derivative of $so(x)$: $\frac{d}{dx}so(x) = \frac{d}{dx}\frac{x}{1-x} = \frac{1(1-x)+x(1)}{(1-x)^2} = \frac{1}{(1-x)^2} > 0$. Therefore, the function $so(x)$ is strictly increasing on the interval $(-\infty, 0)$.

<u>CASE $x < y, x < 0, y > 0$</u>:
Let $u, u' \in (x, y)$ with $u < 0, u' > 0$. Then is clear that $u < u'$ From the condition $u < |u|$ follows that $uu' = u|u'| < |u||u'|$. Summing this inequality to the inequality $u < u'$ we get: $u + u|u'| = u(1+|u'|) < u' + |u||u'| = u'(1+|u|)$ which implies $\frac{u}{1+|u|} < \frac{u'}{1+|u'|}$, i.e so is strictly incresing on intervals of the type $(x, y), x < 0, y > 0$.

The aforementioned 3 cases imply $so(x),\ \forall x \in \mathbb{R}$ is sctrictly increasing. $\square$

2.5.5 (b)

From $x < 1 + |x|$ follows $\frac{x}{1+|x|} < 1,\ \forall x \in \mathbb{R}^{>0}$. To get the inequality $-1 < \frac{x}{1+|x|}$ apply the second inequality to u and then multiply by -1. In summary, we have that the image of $so(x)$ is the interval $(-1, 1)$

On the other hand, note that $S(x) = S(so(x))$. Now, let $u \in (-1, 1), u > 0$. Suppose there is a x such that $so(x) = u$, we have:

$$u = \frac{x}{1+|x|} \implies x = u + |x|u = u(1+x) \implies x(1-u) = u \implies x = \frac{u}{1-u} = so^{-1}(u)$$

If in the contrary, $u < 0$, suppose it exists an x such that $so(x) = u$:

$$u = \frac{x}{1+|x|} \implies x = u - xu = u(1-x) \implies x(1+u) = u \implies x = \frac{u}{1+u} = so^{-1}(u)$$

Both cases can be complactly written as: $x = so^{-1}(u) = \frac{u}{1-|u|}$. $\square$

2.5.5 (c)

From 2.5.5.a and by the triangle inequality ($|x+y| < |x| + |y|$) follows:

$$\begin{aligned}so(|x+y|) \leq so(|x|+|y|) = \frac{|x|+|y|}{1+|x|+|y|} = \frac{|x|}{1+|x|+|y|} + \frac{|y|}{1+|x|+|y|} \\ \leq \frac{|x|}{1+|x|} + \frac{|y|}{1+|y|} = so(|x|) + so(|y|). \ \square\end{aligned}$$

2.5.6 (a)

To be more consistent with notation lets write $softmax(\mathbf{c}; y) := \dfrac{(e^{c_1+y}, \dots, e^{c_j+y}, \dots, e^{c_n+y})}{\sum_{j=0}^{n} e^{c_j+y}}$ i.e $softmax$ with a scalar shift, instead of $softmax(y + \mathbf{c})$. Because otherwise, one should be precise to define objects of the type $y + \mathbf{c}$ with $y \in \mathbb{R}$, $\mathbf{c} \in \mathbb{R}^n$. That said, $softmax(\mathbf{c}; 0) = softmax(\mathbf{c})$; Continuing with the proof, from the functional form of the function $softmax(y; \mathbf{c})$ it is clear that:

$$\begin{aligned} softmax(\mathbf{c}; y) &= \frac{(e^{c_1+y}, \dots, e^{c_j+y}, \dots, e^{c_n+y})}{\sum_{j=0}^{n} e^{c_j+y}} \\ &= \frac{(e^{c_1}e^{y}, \dots, e^{c_j}e^{y}, \dots, e^{c_n}e^{y})}{e^{y}\sum_{j=0}^{n} e^{c_j}} = \frac{e^{y}(e^{c_1}, \dots, e^{c_j}, \dots, e^{c_n})}{e^{y}\sum_{j=0}^{n} e^{c_j}} \\ &= \frac{(e^{c_1}, \dots, e^{c_j}, \dots, e^{c_n})}{\sum_{j=0}^{n} e^{c_j}} = softmax(\mathbf{c}; 0). \quad \square \end{aligned}$$

2.5.7 (a)

By the definition of the L_2 norm follows the claim. Indeed $\forall\ \mathbf{y} \in \mathbb{R}^n/\{\mathbf{0}\}$:

- $0 \le y_k^2 \le \|\mathbf{y}\|^2 = \sum_{k=0}^{n} \dfrac{y_i^2}{\|y\|^2}\|\mathbf{y}\|^2 \implies 0 \le \rho(\mathbf{y})_k\|\mathbf{y}\|^2 \le \|\mathbf{y}\|^2 \implies 0 \le \rho(\mathbf{y})_k \le 1$
- $\|\mathbf{y}\|^2 = \sum_{k=0}^{n} y_k^2 = \sum_{k=0}^{n} \dfrac{y_i^2}{\|\mathbf{y}\|^2}\|\mathbf{y}\|^2 = \sum_{k=0}^{n} \rho(\mathbf{y})_k\|\mathbf{y}\|^2 \implies \sum_{k=0}^{n} \rho(\mathbf{y})_k = 1. \quad \square$

0.1 2.5.7 (b)

The claim follows easily by the properties of the norm. Let $\forall \lambda \neq 0$:

$$\rho(\boldsymbol{\lambda}\mathbf{y})_k = \frac{(\lambda y_k)^2}{\|\lambda \mathbf{y}\|^2} = \frac{(\lambda y_k)^2}{\|\lambda \mathbf{y}\|^2} = \frac{\lambda^2 (y_k^2)}{\lambda^2\|\mathbf{y}\|^2} = \rho(\mathbf{y})_k. \ \square$$

2.5.8 (a)

First, from the function definition is evident that is sigmoidal. Indeed:

- $\lim_{x\to\infty} \varphi(x) = \lim_{x\to\infty} 1_{(\frac{\pi}{2},\infty)}(x) = 1$
- $\lim_{x\to-\infty} \varphi(x) = \lim_{x\to-\infty} \dfrac{1}{2}(1 + \cos(x + \dfrac{3\pi}{2}))1_{[-\frac{\pi}{2},\frac{\pi}{2}]}(x) = 0$

Lets now prove that the function is not decreasing. On the intervals $(\frac{\pi}{2}, \infty)$ and $(-\infty, \frac{-\pi}{2})$ is evidently non decreasing. On the other hand, $\forall\ x \in [\frac{-\pi}{2}, \frac{\pi}{2}]$ we have:

$$\tfrac{d}{dx}\tfrac{1}{2}(1 + \cos(x + \tfrac{3\pi}{2})) = -\tfrac{1}{2}sin(x + \tfrac{3\pi}{2}) = -\tfrac{1}{2}(cos(x)sin(\tfrac{3\pi}{2}) + cos(\tfrac{3\pi}{2})sin(x)) = \tfrac{1}{2}cos(x).$$

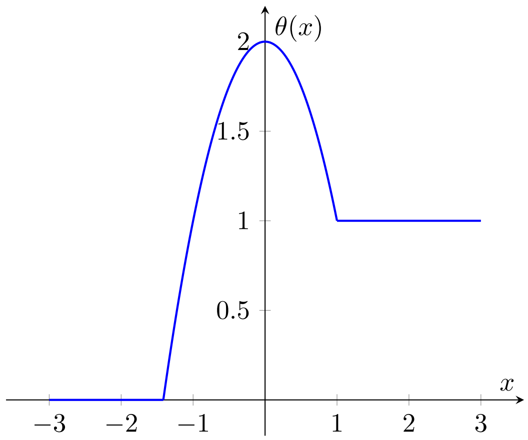
It is well known that $cos(x)$ is positive on the interval $[\frac{-\pi}{2}, \frac{\pi}{2}]$, then the function φ is non decreasing on the interval $[\frac{-\pi}{2}, \frac{\pi}{2}]$ (Because its derivative is postive on such interval). Then, $\varphi(x)$ is a squashing function.

2.5.9 (a)

The claim is obvious, it follows from the fact that by definition a squashing function is a nondecreasing sigmoidal function.

2.5.9 (b)

It is enough to give a function that is non increasing, but satisfies the sigmoidal condition. An example of such type of function is: $\theta(x) := (-x^2+2)1_{[-\sqrt{2},1]}(x) + 1_{(1,\infty)}(x)$



As one can see from the picture the function is clearly non-incresing and satisfies the sigmoidal condition, i.e by definition its limits in $+\infty$ and $-\infty$ are 1 and -1, respectively.