PROF. FRANCESCO PALMIERI
UNIV. DELLA CAMPANIA Nov. 2023

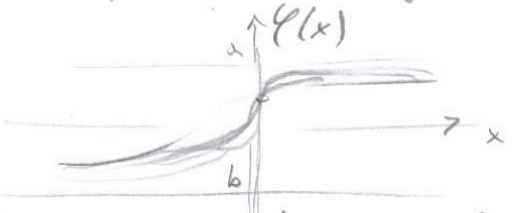
APPENDIX

ACTIVATION FUNCTIONS

In neural network architectures there are various non linear one-dimensional functions used in various ways. Their role inside the networks, that use mostly linear connections, confer to system non-linear overall characteristics that help to implement approximators to arbitrary multi-dimensional functions. Some of them are used for classifiers and some for gating operations. Their use in the architectures is discussed in the various chapters of the book. Here we limit ourselves to list many of them pointing to their gradient and main properties.

SIGMOIDAL ACTIVATION FUNCTIONS

Many activation functions used in neural network architectures, have a "sigmoidal" shape.



Their behaviour saturates for $x \to \infty$ and for $x \to -\infty$ manifesting a non decreasing behaviour $\forall x$. Their threshold is usually centered at $x=0$, but it can be shifted $\varphi(x-t)$. By controlling the scale of x, we can obtain "quasi" step functions for x large, or "quasi linear" for x small.

More specifically any sigmoidal function can be scaled as

$$f(x) = \beta\,\varphi(\alpha x) \quad \text{where usually } \beta\alpha > 0$$

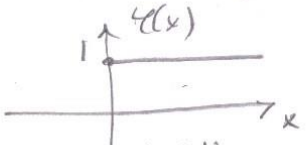
for $\alpha \to 0$ we have a linear function, i.e. $f(x) \sim \alpha\beta$ for $x \to 0$ and $f(x) \sim b\beta$ for $x < 0$ and for $\alpha \to \infty$ a linear behaviour around $x = 0$.

Peculiarities are in their derivatives and properties. That will be detailed in the following. Note that all cumulative distribution functions (cdf) have a sigmoidal behaviour. In fact some of them listed in the following are directly derived from typical random variable cdfs and pdfs.

STEP FUNCTION

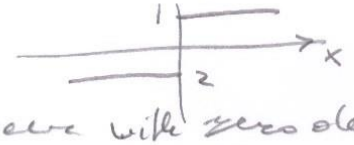
The step function is the first sigmoidal function to consider

$$\varphi(x) = u(x)$$



Also its shifted version is the SIGN FUNCTION

$$\varphi(x) = 2u(x) - 1 = \operatorname{sgn}(x)$$



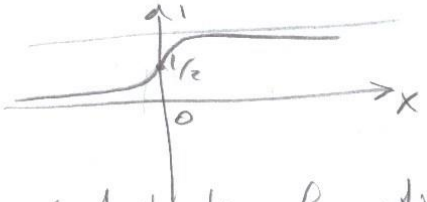
has the same behaviour with zero derivative everywhere except for $x = 0$ where they are discontinuous. In neural network architectures it is preferable to use smooth versions of these functions for more effective backpropagation of the error.

APPENDIX 1

THE LOGISTIC FUNCTION

By far the most common sigmoidal activation function is the logistic

$$\varphi(x) = \ell(x) = \frac{1}{1+e^{-x}}$$



It is a valid CDF (cumulative distribution function)
For small values of x is almost linear $\simeq x+\frac{1}{2}$
and has a number of interesting properties:

① THE DERIVATIVE:

$$\boxed{\ell'(x) = \ell(x)(1-\ell(x))}$$

Proof:

$$\ell'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \frac{e^{-x}-1+1}{(1+e^{-x})^2} = \frac{(1+e^{-x})-1}{(1+e^{-x})^2}$$

$$= \frac{1}{1+e^{-x}} - \frac{1}{(1+e^{-x})^2} = \frac{1}{1+e^{-x}}\left(1-\frac{1}{1+e^{-x}}\right) = \ell(x)(1-\ell(x))$$

② RELATIONS TO THE HYPERBOLIC TANGENT (discussed in the following)

$$\boxed{\ell(x) = \frac{1}{2}\left(1+\tanh\frac{x}{2}\right)}$$

Proof

$$\frac{1}{2}\left(1+\frac{e^{\frac{x}{2}}-e^{-\frac{x}{2}}}{e^{\frac{x}{2}}+e^{-\frac{x}{2}}}\right) = \frac{1}{2}\,\frac{e^{\frac{x}{2}}+e^{-\frac{x}{2}}+e^{\frac{x}{2}}-e^{-\frac{x}{2}}}{e^{\frac{x}{2}}+e^{-\frac{x}{2}}} = \frac{e^{\frac{x}{2}}}{e^{\frac{x}{2}}+e^{-\frac{x}{2}}} = \frac{1}{1+e^{-x}}$$

(5) THE INVERSE

Setting $p = \ell(x)$

$$\boxed{x = \ln \frac{p}{1-p}} \qquad x \in]0,1[$$

(4) Given a set of real numbers $\alpha_1 \alpha_2 \dots \alpha_N$

$$\ell(\alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_N x_N) = \frac{p_1^{\alpha_1} p_2^{\alpha_2} \dots p_N^{\alpha_N}}{p_1^{\alpha_1} p_2^{\alpha_2} \dots p_N^{\alpha_N} + (1-p_1)^{\alpha_1}(1-p_2)^{\alpha_2} \dots (1-p_N)^{\alpha_N}}$$

where $p_i = \ell(x_i)$

Proof:

If $p = \ell(x)$ $\quad e^{-x} = \frac{1-p}{p}$ and $e^{-\alpha x} = \left(\frac{1-p}{p}\right)^{\alpha}$

Therefore

$$\frac{1}{1 + e^{-\alpha_1 x_1 - \alpha_2 x_2 \dots - \alpha_N x_N}} = \frac{1}{1 + \left(\frac{1-p_1}{p_1}\right)^{\alpha_1}\left(\frac{1-p_2}{p_2}\right)^{\alpha_2} \dots \left(\frac{1-p_N}{p_N}\right)^{\alpha_N}}$$

which gives the result.

Interesting corollaries are the sum and the difference

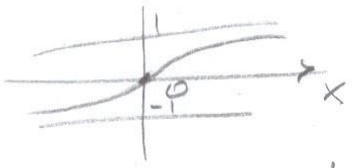
$$\ell(x_1 + x_2) = \frac{p_1 p_2}{p_1 p_2 + (1-p_1)(1-p_2)}$$

$$\ell(x_1 - x_2) = \frac{\frac{p_1}{p_2}}{\frac{p_1}{p_2} + \frac{1-p_1}{1-p_2}} = \frac{p_1(1-p_2)}{p_1(1-p_2) + p_2(1-p_1)}$$

THE HYPERBOLIC TANGENT

Strictly related to the logistic function, is the hyperbolic function.

$$\varphi(x) = \tgh x \triangleq \frac{e^x - e^{-x}}{e^x + e^{-x}}$$



Its peculiarity is in being between -1 and 1.
For small values of x is linear $\varphi(x) \simeq x$.
And for large scale $\varphi(x) \simeq \operatorname{sgn} x$
We list some of the relevant properties:

① RELATION TO THE LOGISTIC

$\tgh x$ and $\ell(x)$ are really the same function related in the argument and shifted in value

$$\tgh x = 2\,\ell(2x) - 1$$

Proof

$$2\frac{1}{1+e^{-2x}} - 1 = \frac{2 - 1 - e^{-2x}}{1+e^{-2x}} = \frac{1 - e^{-2x}}{1+e^{-2x}} = \frac{e^{-x}(e^x - e^{-x})}{e^{-x}(e^x - e^{-x})} = \tgh x$$

② DERIVATIVE

$$\frac{d}{dx}\tgh x = (1+\tgh x)(1-\tgh x) = 1 - \tgh^2 x = 1 - \varphi(x)^2$$

Proof

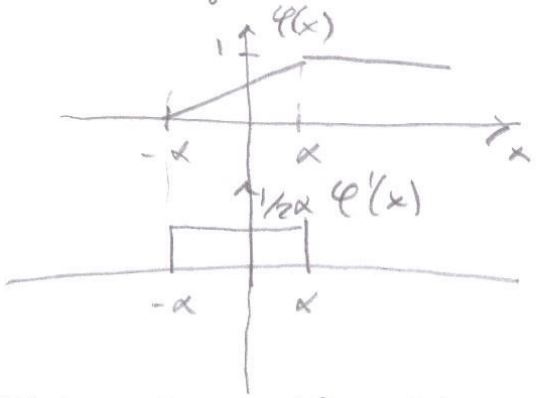
$$\frac{d}{dx}\frac{e^x - e^{-x}}{e^x + e^{-x}} = \frac{(e^x + e^{-x})(e^x + e^{-x}) - (e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2}$$

$$= \frac{(e^x + e^{-x})^2 - (e^x - e^{-x})^2}{(e^x + e^{-x})^2} = 1 - \left(\frac{e^x - e^{-x}}{e^x + e^{-x}}\right)^2 = 1 - \tgh^2 x$$

$$= (1+\tgh x)(1-\tgh x) = 1 - \tgh^2 x \;;\quad \varphi'(x) = 1 - \varphi^2(x)$$

PIECEWISE LINEAR

A sigmoidal function can be built using linear segments as



$$\varphi(x) = \begin{cases} 0 & x < -\alpha \\ \frac{x}{2\alpha} & -\alpha < x < \alpha \\ 1 & x > \alpha \end{cases}$$

$$\varphi'(x) = \begin{cases} \frac{1}{2\alpha} & -\alpha < x < \alpha \\ 0 & \text{else} \end{cases}$$

$\varphi(x)$ is clearly the cdf of a uniform random value. $\varphi(x)$ is continuous but discontinuous in its derivative

When $\alpha = 1$, it is called HARD SIGMOID and can be written as

$$\varphi(x) = \max(0, \min(1, \frac{x+1}{2}))$$

Easy to verify (and to code)

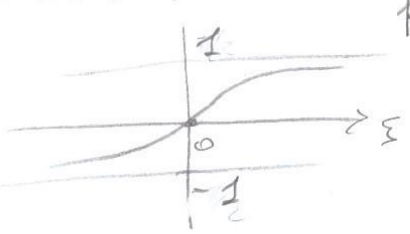
Note that the piecewise linear $\varphi(x)$ can be also written scaling the input argument

$$\varphi(x) = \max(0, \min(1, \frac{\frac{x}{\alpha}+1}{2})) .$$

TRIGONOMETRIC ARCTANGENT

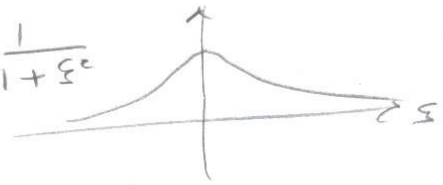
The trigonometric arctangent is naturally a sigmoidal function.

$$\varphi(x) = \frac{2}{\pi} \tan^{-1} x$$



with the well-known derivative

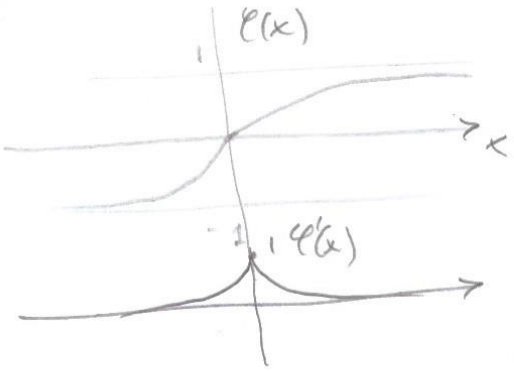
$$\varphi'(\xi) = \frac{2}{\pi} \frac{1}{1+\xi^2}$$



SOFTSIGN

The softsign is the hyperbolic tangent, but it converges to its asymptotic values polinomially rather than exponentially

$$\varphi(x) = \frac{x}{|x|+1}$$



Its derivative is

$$\varphi'(x) = \left(\frac{1}{1+|x|}\right)^2$$

Proof

$$\varphi'(x) = \frac{|x|+1-x\left(\frac{\partial |x|}{\partial x}\right)}{(|x|+1)^2} = \frac{|x|+1-x\frac{x}{|x|}}{(|x|+1)^2}$$

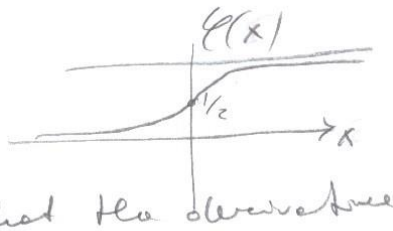
since $\frac{\partial |x|}{\partial x} = \frac{x}{|x|}$

and $\frac{x^2}{|x|} = \frac{|x|^2}{|x|} = |x|$

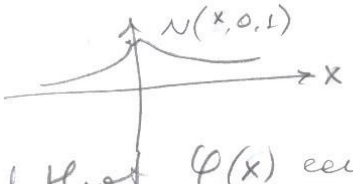
$$= \frac{\cancel{|x|}+1-\cancel{|x|}}{(1+|x|)^2} .$$

GAUSSIAN CDF

$$\varphi(x) \triangleq \int_{-\infty}^{x} N(\xi;0,1)\, d\xi$$



The advantage here is that the derivative is the gaussian function



$$\varphi'(x) = N(x;0,1)$$

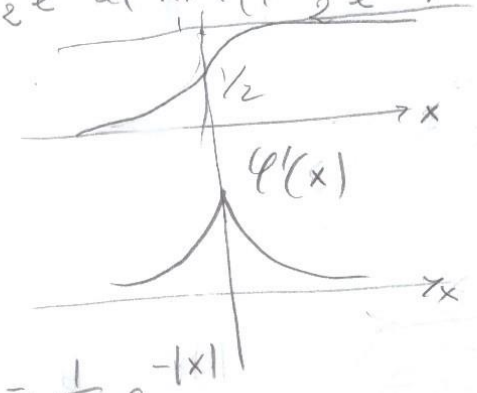
The disadvantage is in the fact that $\varphi(x)$ cannot be computed in closed-form, but needs a numerical algorithm. Most computer languages have $\varphi(x)$ available among their function library. Perhaps in the form of an error function

$$\text{erf}(\xi) = 2\varphi(\sqrt{2}\xi) - 1,$$

$$\varphi(x) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$$ (See any book on random variables)

LAPLACIAN CDF

$$\varphi(x) = \frac{1}{2}e^{x}u(-x) + \left(1 - \frac{1}{2}e^{-x}\right)u(x)$$



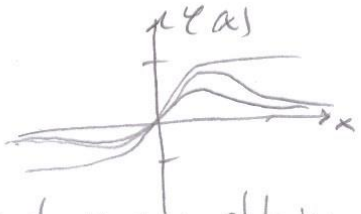
$$\varphi'(x) = \frac{1}{2}e^{-|x|}$$

This sigmoidal function has an exponential growth in the transition region.

SOBOLEVA MODIFIED HYPERBOLIC TANGENT (SMHT)

A generalization of the hyperbolic tangent is the Soboleva modified function

$$\varphi(x) = \frac{e^{ax} - e^{-bx}}{e^{cx} + e^{-dx}}$$



By controlling the values of a, b, c, d we can obtain various behaviours that go from the tgh for

$$a = b = c = d = 1$$

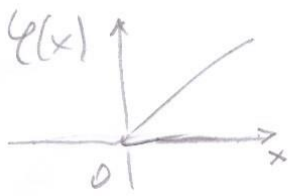
to other non-saturated behaviours.
We leave the many details here for brevity, suggesting to the interested reader to look at the specific literature.

NON SIGMOIDAL ACTIVATIONS

Since the original simplest neural network, the perceptron, has at the output a sigmoidal function, sigmoids have been used extensively in many architectures. However sigmoidal functions may saturate, and during learning, may cause large plateaux in the cost functions.
Saturation in the first layers of a neural network may limit the expressiveness of some architectures as the higher layers may have to work on saturated values.
The choice of the most appropriate activation function is application-specific and we report here some other non-sigmoidal choices that have become popular in the latest architectures.

THE RECTIFYING NON LINEARITY (ReLU - Rectifying Logic Unit)

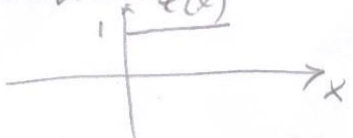
By far this is the most common choice



$$\varphi(x) = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases}$$

$$= \max(0, x)$$

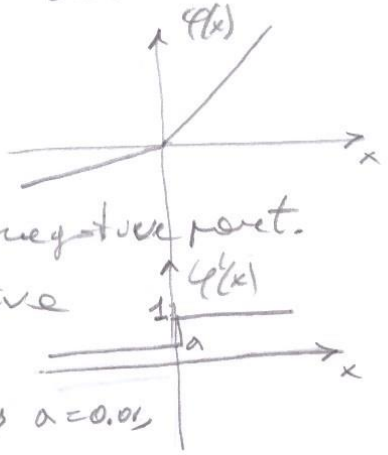
It is piecewise linear and its derivative is the step function $\varphi'(x)$



This function essentially blocks all the negative values and lets the positive ones to be copied at the output.

PARAMETRIC RECTIFIED LINEAR UNIT (PReLU)

$$\varphi(x) = \begin{cases} ax & x \le 0 \\ x & x > 0 \end{cases} \qquad 0 < a < 1$$



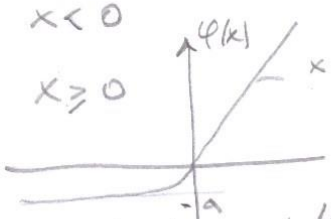
this function is a slight modification of the ReLU allowing a "leakage" in the negative part.

the derivative is a positive stepwise constant function.

When a is small, typically $a = 0.01$, this is called LEAKY RECTIFIED LINEAR UNIT (Leaky ReLU)

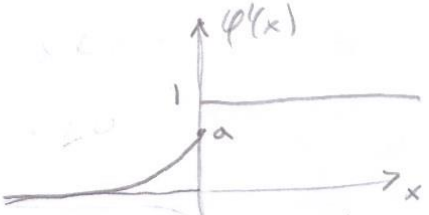
EXPONENTIAL LINEAR UNIT (ELU)

$$\varphi(x) = \begin{cases} a(e^x - 1) & x < 0 \\ x & x \ge 0 \end{cases} \qquad a > 0$$



this unit is a continuous modification of the ReLU in the negative part where it tends exponentially to $-a$. The derivative is

$$\varphi'(x) = \begin{cases} a e^x = \varphi(x) + a & x < 0 \\ 1 & x \ge 0 \end{cases}$$



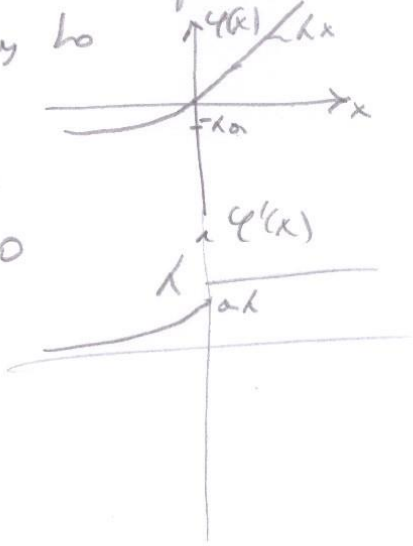
there is a discontinuity in the derivative for $x = 0$ unless $a = 1$.

this function can be scaled to get the SCALED EXPONENTIAL LINEAR UNIT (SELU)

$$\varphi(x) = \begin{cases} \lambda a(e^x - 1) & x < 0 \\ \lambda x & x \ge 0 \end{cases}$$

to control the slope of the linear part.
the derivative is similarly to above

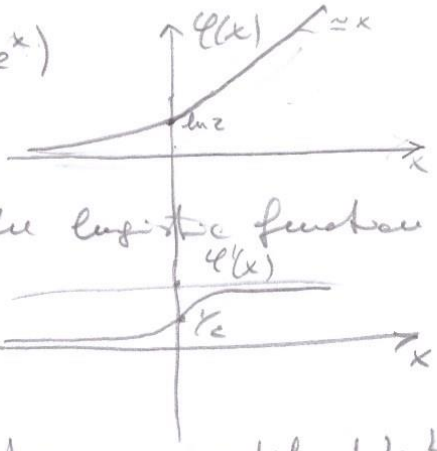
$$\varphi'(x) = \begin{cases} \lambda a e^x = \varphi(x) + \lambda a \\ \alpha \qquad x \geq 0 \end{cases}$$



SOFTPLUS

Another smooth version of the ReLU is the Softplus activation

$$\varphi(x) = \ln(1+e^x)$$



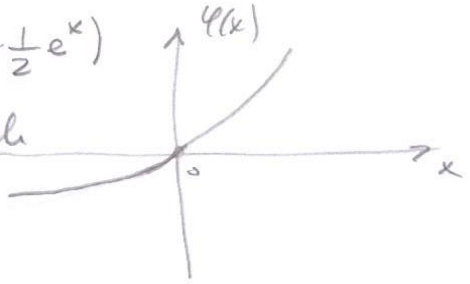
Note that for large x the function is $\simeq x$.

Its derivative is the logistic function

$$\varphi'(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

Sometimes this function is modified to be the SHIFTED SOFTPLUS

$$\varphi(x) = \ln\left(\tfrac{1}{2}+\tfrac{1}{2}e^x\right)$$

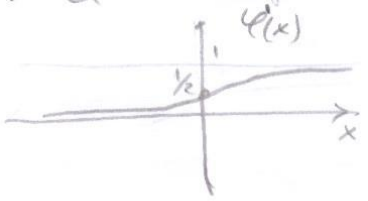


It is made to go through $x=0$

Obviously

$$\varphi(x) = \ln\tfrac{1}{2}(1+e^x) = \ln(1+e^x) - \ln 2$$

and the derivative is still the logistic function.



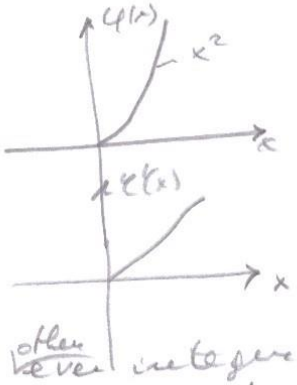
SQUARED RELU

AF14

Where the growth in the positive part needs to be longer than linear, we could use a longer order power, for example 2.

$$\varphi(x) = \begin{cases} x^2 & x \geq 0 \\ 0 & \text{else} \end{cases}$$

$$\varphi'(x) = \begin{cases} 2x & x > 0 \\ 0 & \text{else} \end{cases}$$



Clearly the exponent can be any other even integer. Some authors have found benefits in transformer architectures using squared Relu.

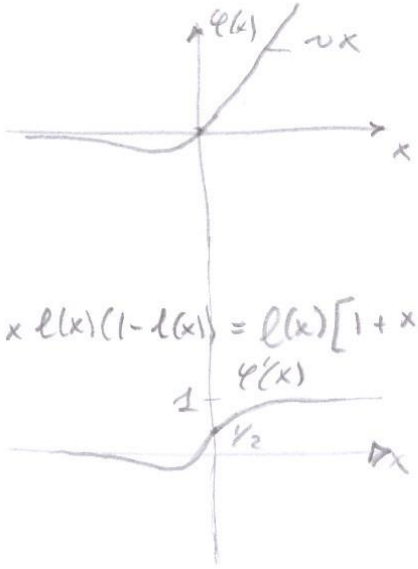
SIGMOID LINEAR UNIT (SiLU) or (SWISH-1)

Any sigmoidal function can be easily modified to give a linear behaviour for large values of x and tend to zero with negative values of x.

$$\varphi(x) = \frac{x}{1+e^{-x}}$$

Since
$\varphi(x) = x\,\ell(x)$
↑ logistic

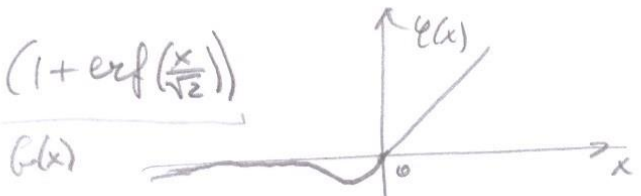
$$\varphi'(x) = \ell(x) + x\,\ell'(x) = \ell(x) + x\,\ell(x)(1-\ell(x)) = \ell(x)\left[1 + x(1-\ell(x))\right]$$



GAUSSIAN ERROR LINEAR UNIT (GELU) AF15

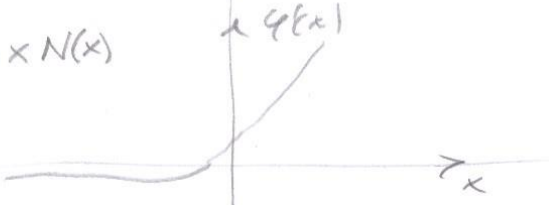
Also the gaussian cdf sigmoid can be modified to be

$$\varphi(x) = x\,\frac{1}{2}\left(1+\operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right)$$



Also the derivative can be computed using the results on the gaussian cdf sigmoid

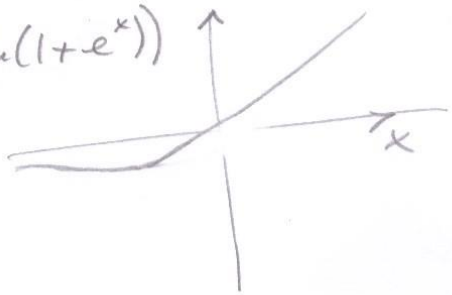
$$\varphi'(x) = x\,G(x) = G(x) + x\,N(x)$$



clearly one has to have available the functions $G(x)$ and $N(x)$.

LOG-SOFTPLUS ERROR activation FUNCTION (SERF)

$$f(x) = x \operatorname{erf}\left(\ln\left(1+e^{x}\right)\right)$$

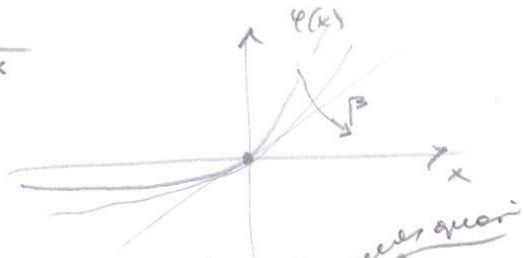


SWISH

AFIT

The swish function is a generalization of the Sigmoid linear Unit. (SiLU) because it has a parameter to control its smoothness.

$$\varphi(x) = \frac{x}{1+e^{-\beta x}}$$



For large values of x the function becomes quasi linear, but β can control the overall behaviour. For large β the function is almost linear.

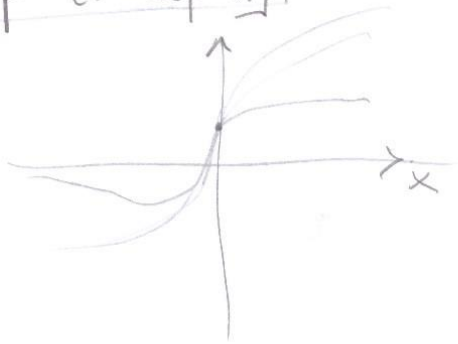
The parameter β is often learned allowing the unit to adjust itself to a linear/nonlinear behavior. Note that changing β is not the same as scaling x because the coefficient is only at the denominator.

Since $\frac{1}{1+e^{-\beta x}} = \ell(\beta x)$ and $\ell'(\beta x) = \beta\, \ell(\beta x)(1-\ell(\beta x))$

$$\varphi(x) = x\,\ell(\beta x)$$

$$\varphi'(x) = \ell(\beta x) + x\,\ell'(\beta x) = \ell(\beta x) + x\,\beta\,\ell(\beta x)(1-\ell(\beta x))$$

$$\boxed{\varphi'(x) = \ell(\beta x)\left[1+\beta x(1-\ell(\beta x))\right]}$$



More activation functions have been proposed for various architectures. Various authors have reported advantages of using a function rather than another. The results are unfortunately very specific to the application and to the algorithms utilized. Most of the time the best choice is determined at the end of a trial-and-error phase.

There are other parametric activation functions proposed that learn their parameters. They are based on superpositions of basis functions or properly-chosen parametrizations. In these cases the idea is that the most appropriate shape of non linearity has to emerge from training.