D Sigmoid-like Nonlinear Functions

In a large class of networks, each node computes a function $f(\mathbf{w}^T\mathbf{x})$ of its inputs $\mathbf{x}$. In most cases, $f(u)$ is chosen to be a bounded nondecreasing function of u; the sigmoid and tanh functions are common choices. Back-propagation and other gradient based training methods require that f be differentiable.

Table D.1 lists some functions commonly used for node nonlinearities. In general, scaled and translated functions $g(u) = af(ku) + b$, for constants a, b, and k, yield networks with equivalent representational properties. The tanh and sigmoid functions are related by $\tanh(u) = 2\text{sigmoid}(2u) + 1$, for example. There may be practical reasons however, for choosing one form over another.

Sigmoid The sigmoid, or logistic, function

$$y(u) = \frac{1}{1 + e^{-\lambda u}} \tag{D.1}$$

is a bounded, nondecreasing function of u. It approaches 0 for $u \to -\infty$, is 1/2 at $u = 0$, and approaches 1 for $u \to \infty$. It is approximately linear for small inputs ($u \approx 0$), but saturates for large positive or negative inputs. The name derives from this "s" shape. Other monotonic functions with similar shapes are often called sigmoidal. The optional parameter λ controls the slope in the linear region; with large values the response approximates a step function. Normally $\lambda = 1$ unless otherwise specified since equivalent results can be obtained by scaling the magnitude of the weight vector.

A useful property of the usual form (D.1) is that its derivative is easily calculated given the output

$$\begin{aligned}\frac{\partial y}{\partial u} &= \frac{\lambda e^{-\lambda u}}{(1 + e^{-\lambda u})^2} \\ &= \lambda y(1 - y).\end{aligned} \tag{D.2}$$

The derivative is a bell shaped function, positive everywhere and largest at $u = 0$ where the slope is $\lambda/4$. For large positive and negative values of u, it approaches 0.

The inverse is

$$u = \frac{1}{\lambda} \ln\left(\frac{y}{1 - y}\right). \tag{D.3}$$

Tanh The tanh function is

$$y(u) = \tanh(\lambda u) \qquad = \frac{e^{\lambda u} - e^{-\lambda u}}{e^{\lambda u} + e^{-\lambda u}} \tag{D.4}$$

Table D.1
Common Node Nonlinearities

Name	Function $y(u)$	Derivative $\partial y/\partial u$
Sigmoid	 $1/(1+e^{-\lambda u})$	 $\lambda y(1-y)$
Tanh	 $\tanh(\lambda u)$	 $\lambda(1-y^2)$
Step	 $\begin{cases} 0 & u \le 0 \\ 1 & u > 0 \end{cases}$	 $\delta(u)$
Sign	 $\begin{cases} -1 & u \le 0 \\ 1 & u > 0 \end{cases}$	 $2\delta(u)$
Clipped linear	 $\begin{cases} -1 & u \le -1/\lambda \\ \lambda u & -1/\lambda < u \le 1/\lambda \\ 1 & u > 1/\lambda \end{cases}$	 $\begin{cases} 0 & u \le -1/\lambda \\ \lambda & -1/\lambda < u \le 1/\lambda \\ 0 & u > 1/\lambda \end{cases}$
Inverse Abs	 $u/(1+\lvert u\rvert)$	 $1/(1+\lvert u\rvert)^2$

is a centered version of the sigmoid. It is -1 for $u=-\infty$, 0 for $u=0$, and $+1$ for $u=+\infty$. The functions are related by $\tanh(u) = 2\text{sigmoid}(2u) + 1$. Its derivative, in terms of its output, is

$$\frac{\partial y}{\partial u} = \lambda(1-y^2). \tag{D.5}$$

At $u=0$, the slope is λ. Its inverse is

$$u = \frac{1}{2\lambda} \ln\left(\frac{1+y}{1-y}\right). \tag{D.6}$$

Step and Sign The unit step function is

$$y(u) = \begin{cases} 0 & u \leq 0 \\ 1 & u > 0. \end{cases} \tag{D.7}$$

In engineering, this is sometimes called the 'Heaviside' step function. A node implementing $f(\mathbf{w}^T\mathbf{x})$ where f is a step function is also called a *linear threshold unit* (LTU). The derivative of the step function is the Dirac delta function $\delta(u)$, which is ∞ at $u = 0$ and zero everywhere else.

The sign function is the bipolar equivalent of the step function

$$y(u) = \begin{cases} -1 & u \leq 0 \\ 1 & u > 0. \end{cases} \tag{D.8}$$

and has derivative $2\delta(u)$.

Clipped Linear The output of the clipped linear function is equal to its input for small inputs, but clips at large positive and negative values

$$y(u) = \begin{cases} -1 & u \leq -1/\lambda \\ \lambda u & -1/\lambda < u \leq 1/\lambda \\ 1 & u > 1/\lambda. \end{cases} \tag{D.9}$$

This may also be called a semilinear ramp function.

Its derivative is constant in the linear region and zero elsewhere

$$\frac{\partial y}{\partial u} = \begin{cases} 0 & u \leq -1/\lambda \\ \lambda & -1/\lambda < u \leq 1/\lambda \\ 0 & u > 1/\lambda. \end{cases} \tag{D.10}$$

Other Functions There are a number of alternative sigmoid-like functions occasionally used in special cases. One is

$$y(u) = \frac{u}{1 + |u|}. \tag{D.11}$$

In table D.1 this is called the Inverse Abs function, but it does not have a generally recognized name. The shape is similar to the tanh function, but convergence to the ± 1 asymptotes is slower. (The horizontal axis of the thumbnail figure in table D.1 spans $-10 < u < 10$.) Its derivative is

$$\frac{\partial y}{\partial u} = \frac{1}{(1 + |u|)^2}. \tag{D.12}$$

At $u = 0$, $\partial y/\partial u = 1$, but the derivative is more sharply peaked near 0 and has wider tails than the tanh function.

An advantage of this function is that it does not require transcendental functions that may be time consuming to calculate on some computers. This may be useful in digital implementations, but is not a particular advantage for analog electronic implementations because tanh functions are easily realized with differential amplifiers. The slower convergence to the asymptotes may help prevent paralysis during learning due to saturation of node nonlinearities.