Fourteenth International Conference
Financial and Actuarial Mathematics – FAM-2021

# ST activation function in dynamic systems

Pavel Stoynov,
Technical University, Sofia.
pstoynov@tu-sofia.bg

**Abstract.**

In the neural networks, the case of a single neuron arises from regression models. Logistic sigmoid is the function needed to interpret the output of a network as a probability when the hidden-units activation is governed by a member of the exponential family.

In this article a new type of non-linear activation functions is proposed based on Switch-Time (ST) distribution.

**Key words:** switch-time distribution, activation function, neural networks

## Introduction

Artificial Neural Networks (ANN) are very fast developing branch of Artificial Intelligence (AI) research. Researchers study different kind of neural networks and apply them to different areas of applied science.

In neural networks, as well as in non-linear regression (which may be considered as a very simple case of neural network), different kinds of non-linear activation functions are used.

In this article a new type of non-linear activation function is proposed based on Switch-Time (ST) distribution.

## Logistic activation function

The activation function is the function which is applied to the output of a neuron to obtain some specific, generally non-linear type of transformation of the input signals. We denote it by $g(u)$ and the neuron output is $y = g(u) = g(wx)$ where the vector $x$ is the neuron input and $w$ is the vector of weights. The function $g(u)$ should be monotonic with $u$.

Let us consider a two-class problem in which the class-conditional densities are given by Gaussian distribution with equal covariance matrices $\Sigma = \Sigma_1 = \Sigma_2$. So, for the conditional probability $p(x \mid C_k)$ for $x$ given membership of class $C_k$, we have

$$p(x \mid C_k) = \frac{1}{(2\pi)^{\frac{d}{2}} \, |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)}, k = 1,2,$$

where $C_k$, $k = 1,2,$ are the two classes into which we should classify the observations, $d$ is the dimension of the distribution and $\mu_k$, $k = 1,2,$ is the vector of the mean of the Gaussian distribution, corresponding to the class $C_k$.

Using the Bayes' theorem, the posterior probability of membership of class $C_1$ is given by Bishop (1994).

$$P(C_1 \mid x) = \frac{p(x \mid C_1)P(C_1)}{p(x \mid C_1)P(C_1) + p(x \mid C_2)P(C_2)} = \frac{1}{1+e^{-u}} = g(u)$$

where

$$u = \ln \frac{p(x \mid C_1)P(C_1)}{p(x \mid C_2)P(C_2)}.$$

So, we have that posterior probability is exactly the logistic sigmoid function

$$g(u) = \frac{1}{1+e^{-u}}.$$

Thus, the use of logistic sigmoid function provides interpretation of the output as posterior probabilities and we obtain not only a classification decision but also a probabilistic interpretation.

## Switch-Time distributions

Switch-Time (ST) distributions are presented in details by Stoynov (2019). Here we just remember some basic facts about these distributions.

The random variable $\xi$ with ST distribution has probability density function

$$f_\xi(x) = \begin{cases} C(n,\beta)e^{-\beta x}(1+x)^n, \; x \geq 0 \\ 0, x < 0, \end{cases}$$

where $C(n,\beta)$ are normalizing coefficients for which

$$C(n,\beta) = \frac{1}{I(n,\beta)}$$

and

$$I(n,\beta) = \frac{1}{\Gamma(1)} \int_0^\infty e^{-\beta t}(1+t)^n \, dt = \int_0^\infty e^{-\beta t}(1+t)^n \, dt.$$

In the special case when $n = 0,\quad n = 1$ and $n = 2$ we have $C(0,\beta) = \beta,\quad C(1,\beta) = \frac{\beta^2}{\beta+1}$ and $C(2,\beta) = \frac{\beta^3}{\beta^2 + 2\beta + 2}.$

For $I(n,\beta)$ we have

$$I(0,\beta) = \frac{1}{\beta},$$

$$I(n,\beta) = \frac{1}{\beta} + \frac{n}{\beta} I(n-1,\beta), n = 1,2,\ldots$$

The proof can be found in Stoynov (2019).

Below we can see some well known special cases of the distribution

1. Exponential distribution: $ST(0,\beta) \equiv Exp(\beta).$
2. Lindley distribution (Lindley): $ST(1,\beta) \equiv L(\beta).$ The random variable $\xi$ with Lindley distribution has probability density

$$f_\xi(x) = \begin{cases} \dfrac{\beta^2}{\beta+1} e^{-\beta x}(1+x), \; x \geq 0 \\ 0, x < 0. \end{cases}$$

3. The distribution $ST(2,\beta).$ The random variable $\xi$ with distribution $ST(2,\beta)$ has probability density

$$_\xi(x) = \begin{cases} \dfrac{\beta^3}{\beta^2 + 2\beta + 2} e^{-\beta x}(1+x), \; x \geq 0 \\ 0, x < 0. \end{cases}$$

**Theorem 1.** The probability density function $f_{\xi}(x)$ for $x \geq 0$ of the random variable $\xi \in ST(n, \beta)$ can be presented as a mixture of Erlang distributions:

$$f_{\xi}(x) = \sum_{k=1}^{n+1} P(D^n = k) f_{G^k}(x),$$

where $D^n$ are random variables with parameter $n$, taking values $k = 1, \ldots, (n+1)$ with probabilities $P(D^n = k) = \frac{C(n,\beta)n!}{\beta^k (n-k+1)!}$, $k = 1, \ldots, (n+1)$, and $G^k$ is a random variable with parameter $k$, having distribution $G^k \in \Gamma\left(k, \frac{1}{\beta}\right) \equiv Erlang\left(k, \frac{1}{\beta}\right)$.

**Theorem 2.** Let $\xi \in ST(2, \beta)$. Then for the moment generating function $M_{\xi}(s) = Ee^{s\xi}$ of $\xi$ we have

$$M_{\xi}(s) = \frac{C(\beta)}{C(\beta - s)},$$

where $C(x) = \frac{x^3}{x^2 + 2x + 2}$ and $\beta > s$.

**Theorem 3.** The characteristic function of the random variable $\xi \in ST(n, \beta)$ is $\psi_{\xi}(t) = C(n, \beta) \sum_{k=0}^{n} \binom{n}{k} \frac{k!}{(\beta - it)^{k+1}}$.

The proofs of Theorem 1, Theorem 2 and Theorem 3 can be seen in Stoynov (2019).

## Probabilistic interpretation of Switch-Time activation function

Here we present the probabilistic interpretation of one specific type activation function which uses Switch-Time distribution.

Let us again consider a two-class problem in which the class-conditional densities are given by Gaussian distribution with equal covariance matrices $\Sigma = \Sigma_1 = \Sigma_2$. So, for the conditional probability $p(x \mid C_k)$ for $x$ given membership of class $C_k$, we have

$$p(x \mid C_k) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)},$$

where $C_k$, $k = 1,2,$ are the two classes into which we should classify the observations, $d$ is the dimension of the distribution and $\mu_k$, $k = 1,2,$ is the vector of the mean of the Gaussian distribution, corresponding to the class $C_k$.

Using the Bayes' theorem, the posterior probability of membership of class $C_1$ can be presented as

$$P(C_1 \mid x) = \frac{p(x \mid C_1)P(C_1)}{p(x \mid C_1)P(C_1) + p(x \mid C_2)P(C_2)} =$$

$$= \frac{1}{1 + C(n,\beta)e^{-\beta u}(1+u)^n} = g(u; n, \beta)$$

where

$$C(n,\beta)e^{-\beta u}(1+u)^n = \frac{p(x \mid C_2)P(C_2)}{p(x \mid C_1)P(C_1)}.$$

So, we have that posterior probability is exactly the Switch-Time sigmoid function which is a generalization of the logistic sigmoid function because if we take the parameters $n=0$ and $\beta=1$, we obtain exactly the logistic sigmoid function.

Thus, the use of Switch-Time sigmoid function also provides interpretation of the output as posterior probabilities.

**ST activation function**

Here we present different types of activation functions based on ST-distribution.
Using ST-function as activation function was for first time proposed in Stoynov (2021).
Here we present different kinds of ST-functions as activation functions.
ST activation function with threshold zero which uses one-side ST distribution is

$$N_{kt} = \begin{cases} \int_0^{n_{kt}} C(n,b) e^{-bt}(1+t)^n\, dt,\ n_{kt} \geq 0, \\ 0,\ n_{kt} < 0. \end{cases}$$

This function depends on two parameters - $n$ and $b$. Standard ST-activation functions can be obtained by different choice of these two parameters.

With $n=0$ and $b=1$ we obtain standard exponential activation function with threshold zero:

$$N_{kt} = \begin{cases} \int_0^{n_{kt}} e^{-t} dt = 1 - e^{-n_{kt}},\ n_{kt} \geq 0, \\ 0,\ n_{kt} < 0. \end{cases}$$

The graphic of this function is presented on Figure 1.

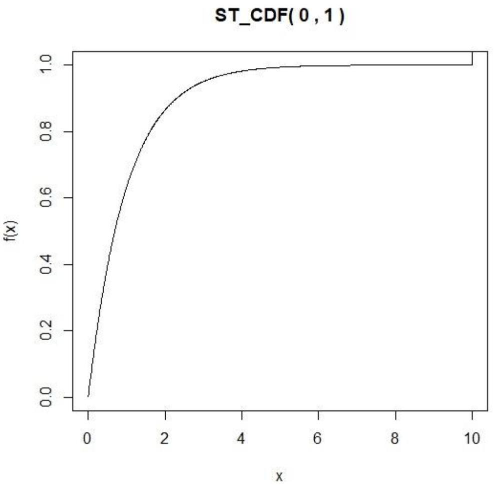


Fig1. Standard exponential activation function with zero threshold and parameters $n=0$ and $b=1$.

The double-sided ST-distribution has a density function

$$s_\xi(x) = \begin{cases} \frac{1}{2} C(n,b) e^{-bx}(1+x)^n,\ x \geq 0, \\ \frac{1}{2} C(n,b) e^{bx}(1-x)^n, x < 0. \end{cases}$$

This distribution permits to define activation functions without thresholds with general formula

$$N_{kt} = \int_{-\infty}^{n_{kt}} s(x)dx = \begin{cases} \dfrac{1}{2} + \dfrac{1}{2}\displaystyle\int_{0}^{n_{kt}} C(n,b)e^{-bx}(1+x)^n dx, \; n_{kt} \geq 0, \\ \dfrac{1}{2}\displaystyle\int_{-\infty}^{n_{kt}} C(n,b)e^{bx}(1-x)^n dx, \; n_{kt} < 0. \end{cases}.$$

The function depends on two parameters – $n$ and $b$. Standard ST-activation functions without thresholds are obtained by selecting different values for these parameters.

The graphics of the density function with $n = 0$ and $b = 1$ is presented in Figure 2.

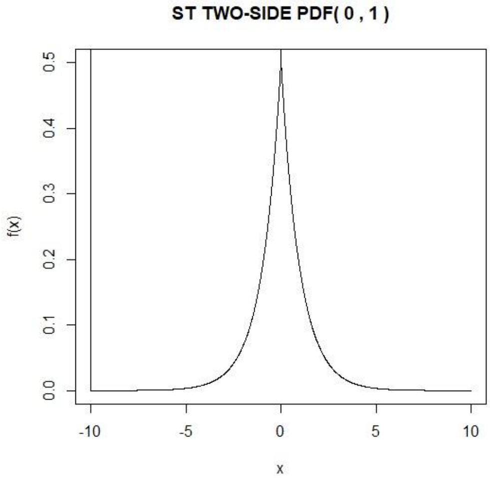


Fig 2. Two-sided ST-density function parameters $n = 0$ and $b = 1$.

With $n = 0$ and $b = 1$ we obtain standard two-sided exponential activation function, called also Laplace activation function:

$$N_{kt} = \begin{cases} \dfrac{1}{2} + \dfrac{1}{2}\displaystyle\int_{0}^{n_{kt}} e^{-x} dx = 1 - \dfrac{1}{2} e^{-n_{kt}}, \; n_{kt} \geq 0, \\ \dfrac{1}{2}\displaystyle\int_{-\infty}^{n_{kt}} e^{x} dx = \dfrac{1}{2} e^{n_{kt}}, \; n_{kt} < 0. \end{cases}$$

The graphics of the cumulative distribution function of two-sided ST distribution with parameters $n = 0$ and $b = 1$ is presented in Figure 3.

## Conclusion

ST distributions proposed for activation function are alternative of the currently used functions and distributions. It may be used in two forms - with and without threshold – based on one-sided and two-sided ST distributions. Also, different values of the parameter $n$ may be chosen with values 0, 1, and 2 giving three standard types ST activation functions for every of the two forms.

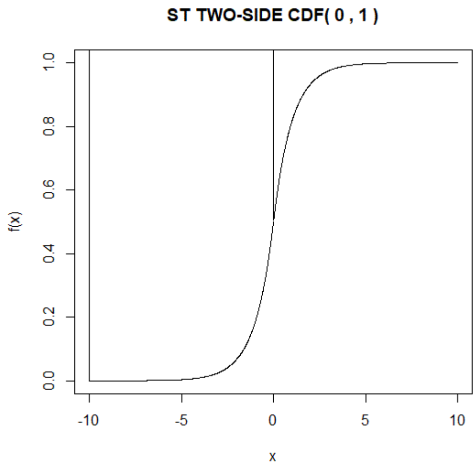


Fig 3. Two-sided ST cumulative distribution function parameters $n = 0$ and $b = 1$.

## References

Bishop, P. (1994) Neural Networks for Pattern Recognition.1994. Oxford University Press. ISBN 0 19 853864.

Stoynov, P. (2019) Switch-time distributions and processes and their applications in finance and insurance. Pliska Studia Mathematica 30, (2019), pp. 201-214.

Stoynov, P. (2021) Applications of ST Distributions to Neural Networks and Regression Models. Proceedings of Eighth International Conference on New Trends in the Applications of Differential Equations in Science (NTADES’21), 6-9 September, St. Constantine and Helena, Bulgaria. (accepted for publishing in AIP Conference Proceedings)