

A Note on Link Functions in Generalized Linear Models and Activation Functions in Deep Learning: An Alternative Activation Function

Kazumitsu Nawata

University of Tokyo, Bunkyo-ku, Tokyo, Japan

Email: kn1016abc@gmail.com

How to cite this paper: Nawata, K. (2026)
A Note on Link Functions in Generalized
Linear Models and Activation Functions in
Deep Learning: An Alternative Activation
Function. *Open Journal of Statistics*, 16,
513-523.

<https://doi.org/10.4236/ojs.2026.165024>

Received: September 26, 2026

Accepted: October 26, 2026

Published: October 29, 2026

Copyright © 2026 by author(s) and
Scientific Research Publishing Inc.
This work is licensed under the Creative
Commons Attribution International
License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

The generalized linear model (GLM) is widely used in econometrics and statistics. In a GLM, a location parameter of the dependent variable—typically the expected value or mean—is linked to a linear combination of independent variables through a link function. The choice of link function is determined by the distribution of the error term. In contrast, activation functions are used in the convolutional layers of deep learning (DL) models. Common activation functions include the indicator (step) function, sigmoid, rectified linear unit (ReLU), and leaky ReLU. Although the universal approximation theorem shows that DL models using these activation functions can approximate a wide class of functions, major challenges remain, including overfitting, local minima, and substantial computational requirements. It is therefore essential to select an appropriate activation function. Link functions in GLMs and activation functions in DL are closely related and often coincide. In this note, we examine the similarities between these functions and propose an activation function derived from the truncated mean. We focus on the theoretical and methodological aspects rather than practical applications. The relationship between link functions and activation functions remains conjectural and warrants further research.

Keywords

Generalized Linear Model, Deep Learning, Activation Function, ReLU, Tobit Model, Truncated Mean

1. Introduction

Generalized linear models (GLMs) are widely used for data analysis in econometrics and statistics [1]-[5]. In a GLM, a location parameter (usually the expected

value or mean) of a dependent variable is linked to a linear combination of independent variables through a link function. The choice of link function is determined by either the underlying error distribution (such as logistic, Poisson, gamma, or inverse-gamma) or the observational structure (such as censoring or truncation).

In contrast, activation functions are used in deep learning (DL) models [6] [7]. Certain DL architectures typically consist of two types of layers: convolutional and pooling. Nawata [8] [9] proposed the grouping-adjusted median estimator (GAME). However, GAME focuses mainly on the pooling stage, and activation functions in the convolutional stage have not been fully examined. Common activation functions include the step (or indicator) function used in threshold models [10]-[13], the sigmoid function [14]-[17], the rectified linear unit (ReLU) [18]-[23], and the leaky ReLU [24]-[28]. Although the universal approximation theorem [29]-[32] shows that most functions can be approximated using these activation functions in DL models, several major challenges remain, including overfitting, multiple local minima, and substantial computational requirements.

These issues become particularly serious in large language models (LLMs), which may contain trillions of unknown parameters due to their numerous layers and nodes (neurons). Selecting an appropriate activation function is therefore critical. Link functions and activation functions are closely related and often coincide. In this note, we examine the similarities between these functions and propose an alternative activation function derived from the conditional expectation (truncated mean) of a truncated model. Our focus is on the theoretical and methodological aspects rather than practical applications.

2. Link and Activation Functions

In this note, we consider the expected value (hereafter, the mean) and other location parameters, such as the median. Following Nawata [33], the GLM is specified as

$$\begin{aligned} y_i^* &= x_i' \beta + u_i, \\ y_i &= g(y_i^*). \end{aligned} \quad (1)$$

y_i^* is an unobserved latent variable, and y_i is an observed dependent variable. x_i and β are vectors of independent variables (including a constant term) and unknown parameters, respectively, and u_i is an error term. $g(z)$ is a monotonically increasing function such that $g(z_1) \geq g(z_2)$ if $z_1 \geq z_2$. If it is not uniformly monotonically increasing, we consider a local interval in which $g(z)$ is monotonic. If it is monotonically decreasing, we use $-g(z)$. Let $\mathcal{L}$ denote a function that gives a suitable location parameter of y_i .

Usually, we consider the mean,

$$\mathcal{L}(x_i' \beta) = E(y_i) = E_u \{g(y_i^*)\}, \quad (2)$$

where E_u denotes the expected value with respect to u_i , and $\mathcal{L}$ depends on the distribution of the error term. However, when the median is used as the loca-

tion parameter, we obtain

$$\mathcal{L}(y_i) = \text{Med}(y_i) = \text{Med}\{g(y_i^*)\} = g(x_i'\beta). \quad (3)$$

Med denotes the median with respect to the error term. The result does not depend on the distribution of the error term, provided that its median is zero. In this sense, the estimator is robust.

In the binary response model, y_i takes the value 0 or 1 and is given by $y_i = g(y_i^*) = 1(x_i'\beta + u_i > 0)$. We obtain

$$\mathcal{L}(y_i) = \begin{cases} E_u(y_i) = F(x_i'\beta) & \text{for the mean} \\ 1(x_i'\beta > 0) & \text{for the median} \end{cases} \quad (4)$$

Here, $1(A)$ is the indicator function, which takes the value 1 if A is true and 0 otherwise, and F is the distribution function of $-u_i$. Assuming a distribution symmetric about zero, we treat F as the distribution function of u_i itself. Under the standard logistic distribution,

$$F(z) = \frac{\exp(z)}{1 + \exp(z)}, \quad (5)$$

we obtain the logit model.

When the median is used, the model becomes semiparametric. No specific distribution function is required, and the resulting estimator is robust. However, Nawata [34] showed that such robust estimators cannot be $\sqrt{n}$ -consistent and are less efficient than the maximum likelihood estimator (MLE) in the binary response model.

These link functions are similar to the activation functions $h(z)$ used in DL. The step function in threshold models is equivalent to the indicator function in the median case. However, threshold models and their corresponding activation function have a major drawback: the derivative is zero except at the jump discontinuity, causing backpropagation to fail and making optimization difficult. This is consistent with the inefficiency of the median estimator in the binary response model. The sigmoid function corresponds to the logistic distribution function. Even when the sigmoid function is used as an activation function, the vanishing gradient problem—in which the derivative approaches zero—makes learning extremely slow and computationally expensive.

For censored regression or Tobit models, $g(z) = \max\{0, z\}$ and

$$y_i = g(x_i'\beta + u_i) = \max\{0, x_i'\beta + u_i\}. \quad (6)$$

The MLE, which assumes a specific distribution for the error term (usually normal or logistic), is widely used in these models. However, it is inconsistent when the distribution is misspecified. Powell [35] considers the least absolute deviations estimator (PLAD), which minimizes

$$S = \sum_{i=1}^n |y_i - \max\{0, x_i'\beta\}|. \quad (7)$$

PLAD is the GLM estimator that uses the median as the location parameter and

is consistent when the median of the error term is zero. Its drawbacks are that (i) multiple local minima make the estimator difficult to compute (Nawata [36]) and (ii) its asymptotic variance is much larger than that of the MLE, making it inefficient.

In DL, the ReLU function,

$$h(x'_i\beta) = \max\{0, x'_i\beta\}, \quad (8)$$

is widely used as an activation function, corresponding to the GLM link function when the error term has a median of zero. Its derivative satisfies $\partial h(z)/\partial z = 1$ if $z > 0$ and $\partial h(z)/\partial z = 0$ if $z < 0$; the function is not differentiable at $z = 0$. This means that (i) information from observations with $x'_i\beta < 0$ is used only through their sign, (ii) the first derivative is discontinuous and jumps at zero, and (iii) the second derivative is zero at any differentiable point. Lin *et al.* ([37], p. 2) state “ReLU-based functions that eliminate negative values appear to constrain representational capacity in ANNs”. By analogy with PLAD, problems of multiple local minima are expected. These issues make computation difficult, especially during backpropagation. Therefore, alternatives such as sigmoid linear unit (SiLU or Swish) [38], leaky ReLU, softplus [39], and Gaussian error linear units (GELUs) [40] functions are also used, as they have been reported to improve performance [38]. The leaky ReLU function is,

$$h(x'_i\beta, a) = \max\{a \cdot x'_i\beta, x'_i\beta\} \quad (9)$$

where a is a small positive constant. The derivative is 1 for $z > 0$ and a for $z < 0$. However, the modifications (including other functions previously described) are ad hoc and lack theoretical justification. Moreover, the derivative remains discontinuous at zero, and the fundamental problems of the ReLU function persist. As in the binary case, it is natural to consider the mean.

3. Modification of the ReLU Function Based on the Truncated Mean

3.1. Truncated Mean

The ReLU function is closely related to the link function in GLMs that use the median as a location parameter. However, this approach has the limitations described in the previous section. In binary models, taking the mean of a binary variable under a logistic error distribution yields the sigmoid function. We therefore consider the mean of the ReLU function with an error term, given by

$$h_i(x'_i\beta, \sigma) = E_u[\max\{0, x'_i\beta + u_i\} | y_i^* > 0], \quad (10)$$

where u_i follows the normal distribution with mean zero and variance σ^2 , $N(0, \sigma^2)$.

Considering the truncated mean when $x'_i\beta + u_i > 0$ (the truncated case in GLMs; Amemiya ([41], pp. 366-373)), we obtain

$$h_i(x'_i\beta, \sigma) = x'_i\beta + \sigma \cdot \lambda(x'_i\alpha), \quad \alpha = \beta/\sigma, \quad \text{and} \quad \lambda(z) = \frac{\phi(z)}{\Phi(z)} \quad (11)$$

where $\phi(z)$ and $\Phi(z)$ are the density and distribution functions of the standard normal distribution, respectively, and $\lambda(z)$ represents the inverse of the Mills ratio. **Figure 1** shows $h_i(z, \sigma)$ with $\sigma = 1$. It converges to z and zero as $z \rightarrow \infty$ and $z \rightarrow -\infty$, respectively. Hereafter, we refer to the function in Equation (11) as the truncated mean function (TMF). Signals are often unobservable when they are significantly smaller than the background noise, as in astronomical or medical data, making it natural to consider truncated observations. For instance, in astronomy, a zero value at an observation point can have two different meanings. One is that there are no stars or galaxies at that location. The other is that they are present, but their signals are too weak to be detected. These two situations have very different implications for analyzing the structure of space. Therefore, it is reasonable to use only observations that exceed the observational detection limit. The estimator based on the TMF, known as Heckman's two-step estimator [42] [43], is widely used in econometrics. The TMF is infinitely differentiable for any real value of z , and its first derivative is given by

$$\partial h_i / \partial z = 1 - w \cdot \lambda(w) - \lambda(w)^2, \quad w = z/\sigma. \quad (12)$$

Although multicollinearity can severely affect Heckman's two-step estimator [44]-[46], this issue does not occur when the TMF is used as an activation function.

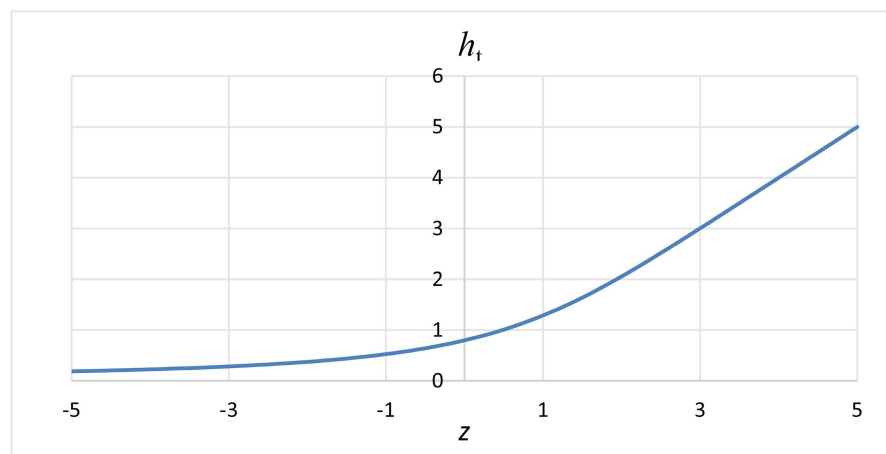


Figure 1. $h_i(z, \sigma)$ with $\sigma = 1$.

Figure 2 shows $\partial h_i / \partial z$ with $\sigma = 1$. $\partial h_i / \partial z$ converges to 1 and 0 as $z \rightarrow \infty$ and $z \rightarrow -\infty$, respectively. The first derivatives of ReLU and leaky ReLU are discontinuous at zero, and their second derivatives vanish at any differentiable point, hindering optimization. This highlights an advantage of the TMF.

The TMF can approximate a piecewise-linear function or the leaky ReLU function, and vice versa. **Figure 3** and **Figure 4** show the zero-adjusted TMF with

$\sigma = 0.1$ and 0.5 , respectively, together with the corresponding leaky ReLU functions. The zero-adjusted TMF equals zero at $z = 0$. The correlation coefficients between the paired functions over $(-3, 3)$ are 0.9996 and 0.9957 , respectively. Consequently,

$$h_t(z, \sigma) = z + \sigma \cdot \lambda(z) \quad \text{or} \quad (13)$$

$$h_t^*(z, \sigma) = z + \sigma \cdot \lambda(z) - \sigma \cdot \lambda(0) = z + \sigma \cdot \{\lambda(z) - 0.7979\}$$

may serve as better activation functions than the leaky ReLU functions. This fact gives rise to the hypothesis that the leaky ReLU and other alternatives might improve the performance because they are approximations of the TMF.

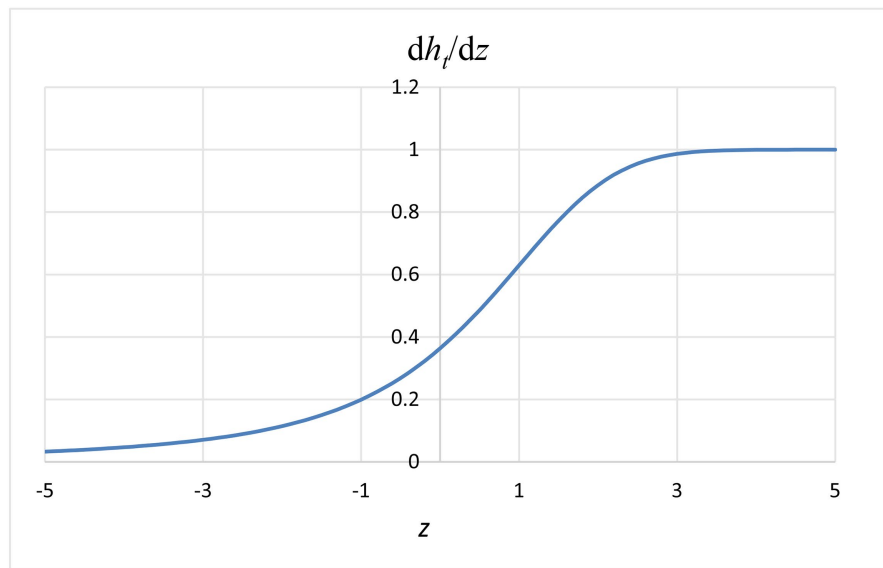


Figure 2. $\partial h_t / \partial z$ with $\sigma = 1$.

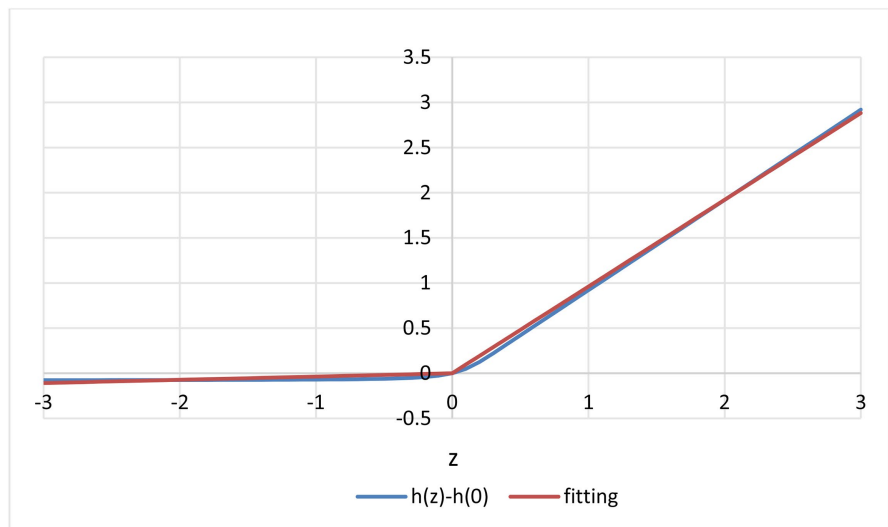


Figure 3. The function $h_t(z, \sigma) - h_t(0, \sigma)$ with $\sigma = 0.1$ and values fitted using the leaky ReLU function.

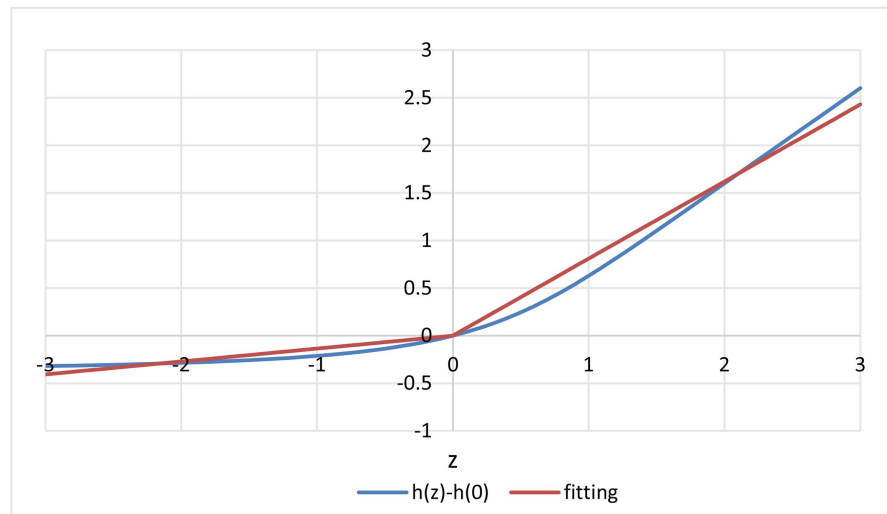


Figure 4. The function $h_i(z, \sigma) - h_i(0, \sigma)$ with $\sigma = 0.5$ and values fitted using the leaky ReLU function.

3.2. Logistic Distribution

We focus on the normal distribution because of its widespread use and well-established properties. However, the logistic distribution function can also be used. The standard logistic distribution has a standard deviation of 1.814. **Figure 5** shows the TMF for the normal distribution with parameter $\sigma = 1.814$ alongside that for the standard logistic distribution, suggesting that the two functions can be used interchangeably with similar effects.

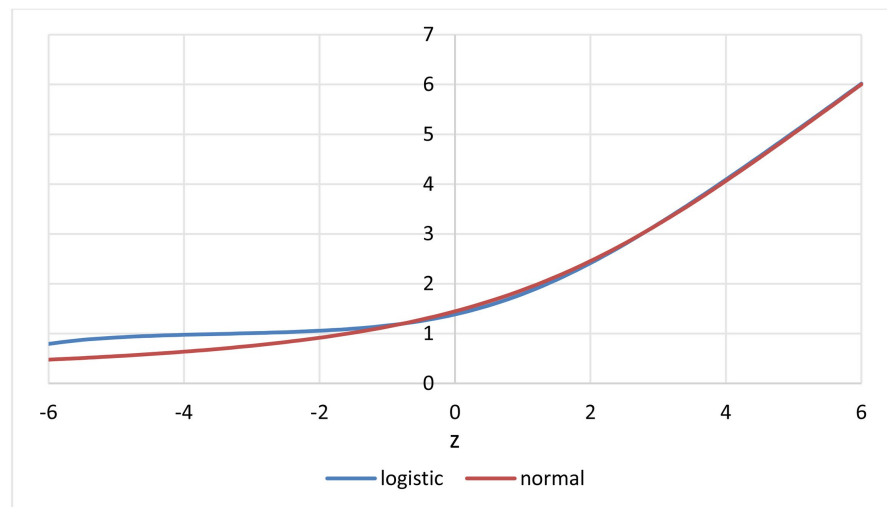


Figure 5. Truncated mean values for the logistic and normal distributions with scale parameter $\sigma = 1.814$.

3.3. Censored Mean

In the previous section, we considered the truncated mean conditional on $y_i^* > 0$. We can also consider the censored mean, or the unconditional expected

value in the censored model, given by

$$h_c(x'_i\beta, \sigma) = E_u \left[\max \{0, x'_i\beta + u_i\} \right] = \Phi(x'_i\beta/\sigma) \cdot h_t(x'_i\beta, \sigma). \quad (14)$$

This formulation is rarely used, and its properties in GLMs have not been fully investigated. **Figure 6** shows $h_t(z, \sigma)$ and $h_c(z, \sigma)$ with $\sigma = 1$. In the negative domain, the censored mean function declines rapidly, and the value of a becomes extremely close to zero when the function is approximated using a leaky ReLU function. Although this function could be a candidate activation function, we do not analyze its properties further in this note.

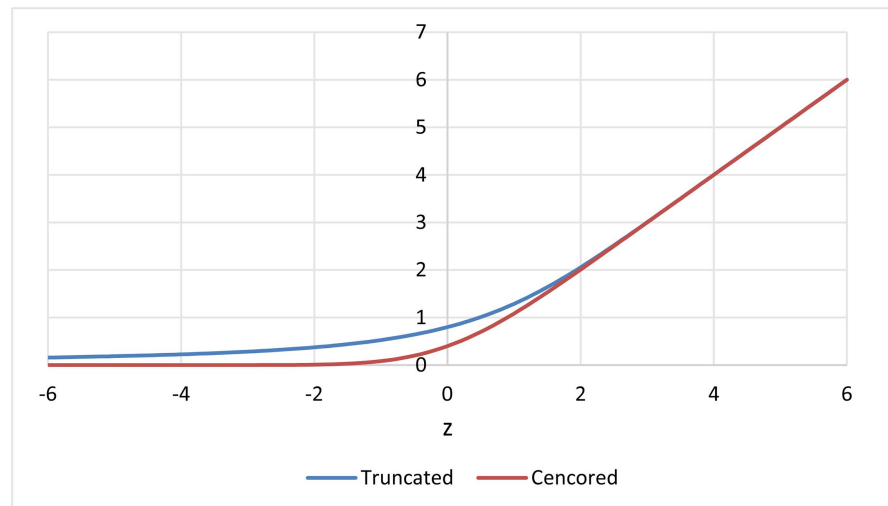


Figure 6. Truncated and censored mean values.

4. Conclusions

Link functions in GLMs and activation functions in the convolutional layers of DL share important similarities. In this note, we demonstrated that GLM link functions correspond to the activation functions used in binary models. The step function in threshold models is equivalent to the indicator function in the median case. The sigmoid function represents the mean of the indicator function under the assumption that the error term in the GLM follows a logistic distribution.

When the ReLU function is used, the model corresponds to the median regression model in GLMs. Its disadvantages include the fact that observations in the negative region have no effect and its first derivative takes only the values zero or one, complicating optimization. Powell [35] proposed the least absolute deviations estimator, but estimation remains difficult because of differentiability issues and multiple local minima. Although the leaky ReLU function is also used, it lacks theoretical justification, and computational challenges persist.

To address these issues, we propose using the TMF as an alternative activation function, derived by applying a truncated normal model. The TMF is widely used in GLMs, such as in Heckman's two-step estimator [42] [43]. It is infinitely differentiable, and its second derivative is non-zero at any point, which may facilitate optimization. It is also well approximated by the leaky ReLU function.

The relationship between link functions and activation functions remains an open question. Evaluating the performance of the TMF in DL models is both necessary and compelling, representing an important direction for future research that will be highly valuable for advancing DL methodology. In this note, we focus on the theoretical and methodological aspects, leaving practical applications for future work.

Acknowledgements

The author would like to thank an anonymous referee for his/her helpful comments and suggestions.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this note.

References

- [1] McCullagh, P. and Nelder, J.A. (1989) Generalized Linear Models. Second Edition, Chapman & Hall.
- [2] Olsson, U. (2002) Generalized Linear Models, an Applied Approach. Studentlitteratur.
- [3] Islam, M.A. and Biswas, S.C. (2025) Generalized Linear Models and Extensions. Springer Nature.
- [4] Wood, S.N. (2025) Generalized Additive Models. *Annual Review of Statistics and Its Application*, **12**, 497-526. <https://doi.org/10.1146/annurev-statistics-112723-034249>
- [5] Correia, S., Guimarães, P. and Zylkin, T. (2026) Verifying the Existence of Maximum Likelihood Estimates for Generalized Linear Models. *Econometric Reviews*, **45**, 1270-1305. <https://doi.org/10.1080/07474938.2026.2673989>
- [6] Dubey, S.R., Singh, S.K. and Chaudhuri, B.B. (2022) Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark. *Neurocomputing*, **503**, 92-108. <https://doi.org/10.1016/j.neucom.2022.06.111>
- [7] Bishop, C.M. and Bishop, H. (2024) Deep Learning: Foundation and Concepts. Springer.
- [8] Nawata, K. (1990) Robust Estimation Based on Grouped-Adjusted Data in Censored Regression Models. *Journal of Econometrics*, **43**, 337-362. [https://doi.org/10.1016/0304-4076\(90\)90124-c](https://doi.org/10.1016/0304-4076(90)90124-c)
- [9] Nawata, K. (2026) Estimating the Rectified Linear Unit Activation Function in Deep Learning Using the Grouping-Adjusted Median Estimator: An Evaluation of Max Pooling. *Open Journal of Statistics*, **16**, 1-15. <https://doi.org/10.4236/ojs.2026.161001>
- [10] Behboodi, A., Rauhut, H. and Schnoor, E. (2020) Generalization Bounds for Deep Thresholding Networks. <https://arxiv.org/abs/2010.15658v1>
- [11] Imaizumi, M. and Schmidt-Hieber, J. (2023) On Generalization Bounds for Deep Networks Based on Loss Surface Implicit Regularization. *IEEE Transactions on Information Theory*, **69**, 1203-1223. <https://doi.org/10.1109/tit.2022.3215088>
- [12] Ergen, T., Gulluk, H.I., Lacotte, J. and Pilanci, M. (2023) Globally Optimal Training of Neural Networks with Threshold Activation Functions. <https://arxiv.org/abs/2303.03382>
- [13] Dhayalkar, S.R. (2025) Neural Networks as Universal Finite-State Machines: A Con-

- structive Deterministic Finite Automaton Theory. <https://arxiv.org/abs/2505.11694>
- [14] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) Learning Representations by Back-Propagating Errors. *Nature*, **323**, 533-536. <https://doi.org/10.1038/323533a0>
 - [15] Roodschild, M., Gotay Sardiñas, J. and Will, A. (2020) A New Approach for the Vanishing Gradient Problem on Sigmoid Activation. *Progress in Artificial Intelligence*, **9**, 351-360. <https://doi.org/10.1007/s13748-020-00218-y>
 - [16] Mulindwa, D.B. and Du, S. (2023) An N-Sigmoid Activation Function to Improve the Squeeze-And-Excitation for 2D and 3D Deep Networks. *Electronics*, **12**, Article No. 911. <https://doi.org/10.3390/electronics12040911>
 - [17] Huang, H., Feng, S., Li, J., Guan, T. and Zhu, H. (2025) Adaptive Belief Rule Base Modeling of Complex Industrial Systems Based on Sigmoid Functions. *Entropy*, **27**, Article No. 1157. <https://doi.org/10.3390/e27111157>
 - [18] Bai, Y. (2022) ReLU-Function and Derived Function Review. *SHS Web of Conferences*, **144**, Article No. 02006. <https://doi.org/10.1051/shsconf/202214402006>
 - [19] Nworu, C.C., Ekpenyong, E.J., Chisimkwuo, J., Onyeukwu, C.N., Okwara, G. and Agwu, O.J. (2022) The Effects of Modified ReLU Activation Functions in Image Classification. *Biomedical Engineering and Medical Devices*, **7**, Article ID: 1000237.
 - [20] Sooksatra, K., Hamerly, G. and Rivas, P. (2024) Is ReLU Adversarially Robust? <https://arxiv.org/abs/2405.03777>
 - [21] Zhang, S., Lu, J. and Zhao, H. (2024) Deep Network Approximation: Beyond ReLU to Diverse. *Activation Functions Journal of Machine Learning Research*, **25**, 1-39.
 - [22] Horuz, C.C., Kasenbacher, G., Higuchi, S., Kairat, S., Stoltz, J., Pesl, M., et al. (2025) The Resurrection of the ReLU. <https://arxiv.org/abs/2505.22074>
 - [23] Agarap, A.F. (2026) Deep Learning Using Rectified Linear Units (ReLU). <https://arxiv.org/abs/1803.08375>
 - [24] Maniatopoulos, A. and Mitianoudis, N. (2021) Learnable Leaky ReLU (LeLeLU): An Alternative Accuracy-Optimized Activation Function. *Information*, **12**, Article No. 513. <https://doi.org/10.3390/info12120513>
 - [25] Mastromichalakis, S. (2021) ALReLU: A Different Approach on Leaky ReLU Activation Function to Improve Neural Networks Performance. <https://arxiv.org/abs/2012.07564>
 - [26] Frei, S., Vardi, G., Bartlett, P.L., Srebro, N. and Hu, W. (2022) Implicit Bias in Leaky ReLU Networks Trained on High-Dimensional Data. <https://arxiv.org/abs/2210.07082>
 - [27] Mujhid, A., Surono, S., Irsalinda, N. and Thobirin, A. (2022) Comparison and Combination of Leaky ReLU and ReLU Activation Function and Three Optimizers on Deep CNN for COVID-19 Detection. In: Tallón-Ballesteros, A.J., Ed., *Fuzzy Systems and Data Mining VIII*, IOS Press, 50-57. <https://doi.org/10.3233/faia220369>
 - [28] Neufeld, A. and Nguyen, T.A. (2026) Multilevel Picard Approximations and Deep Neural Networks with ReLU, Leaky ReLU, and Softplus Activation Overcome the Curse of Dimensionality When Approximating Semilinear Parabolic Partial Differential Equations in L_p -Sense. *Journal of Computational and Applied Mathematics*, **487**, Article ID: 117736. <https://doi.org/10.1016/j.cam.2026.117736>
 - [29] Tikk, D., Kóczy, L.T. and Gedeon, T.D. (2003) A Survey on Universal Approximation and Its Limits in Soft Computing Techniques. *International Journal of Approximate Reasoning*, **33**, 185-202. [https://doi.org/10.1016/s0888-613x\(03\)00021-5](https://doi.org/10.1016/s0888-613x(03)00021-5)
 - [30] Kidger, P. and Lyons, T. (2020) Universal Approximation with Deep Narrow Net-

- works. *Proceedings of Machine Learning Research*, **125**, 1-22.
- [31] Augustine, M.T. (2024) A Survey on Universal Approximation Theorems. <https://arxiv.org/abs/2407.12895>
 - [32] Kratsios, A., Cheng, T.S. and Roy, D. (2025) Beyond Universal Approximation Theorems: Algorithmic Uniform Approximation by Neural Networks Trained with Noisy Data. <https://arxiv.org/abs/2509.00924>
 - [33] Nawata, K. (1997) Estimation of Generalised Regression Models by the Grouping Method. *Mathematics and Computers in Simulation*, **43**, 503-510. [https://doi.org/10.1016/s0378-4754\(97\)00038-4](https://doi.org/10.1016/s0378-4754(97)00038-4)
 - [34] Nawata, K. (1989) Semiparametric Estimation and Efficiency Bounds of Binary Choice Models When the Models Contain One Continuous Variable. *Economics Letters*, **31**, 21-26. [https://doi.org/10.1016/0165-1765\(89\)90105-5](https://doi.org/10.1016/0165-1765(89)90105-5)
 - [35] Powell, J.L. (1986) Censored Regression Quantiles. *Journal of Econometrics*, **32**, 143-155. [https://doi.org/10.1016/0304-4076\(86\)90016-3](https://doi.org/10.1016/0304-4076(86)90016-3)
 - [36] Nawata, K. (1994) Notes on Estimation of the Tobit Models by Powell's Least Absolute Deviations Estimator. *Economic Studies Quarterly*, **45**, 339-346.
 - [37] Lin, W., Piao, Z. and Fung, C.C.A. (2025) Delving into Unsupervised Hebbian Learning from Artificial Intelligence Perspectives. *Machine Learning and Knowledge Extraction*, **7**, Article No. 143. <https://doi.org/10.3390/make7040143>
 - [38] Ramachandran, P., Zoph, B. and Le, B.Q.V. (2017) Searching for Activation Functions. <https://arxiv.org/abs/1710.05941>
 - [39] Zhao, H., Liu, F., Li, L. and Luo, C. (2018) A Novel Softplus Linear Unit for Deep Convolutional Neural Networks. *Applied Intelligence*, **48**, 1707-1720. <https://doi.org/10.1007/s10489-017-1028-7>
 - [40] Hendrycks, D. and Gimpel, K. (2016) Gaussian Error Linear Units (GELUs). <https://arxiv.org/abs/1606.08415>
 - [41] Amemiya, T. (1985) *Advanced Econometrics*. Harvard University Press.
 - [42] Heckman, J.J. (1976) The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement*, **5**, 475-492.
 - [43] Heckman, J.J. (1979) Sample Selection Bias as a Specification Error. *Econometrica*, **47**, 153-161. <https://doi.org/10.2307/1912352>
 - [44] Nawata, K. (1993) A Note on the Estimation of Models with Sample-Selection Biases. *Economics Letters*, **42**, 15-24. [https://doi.org/10.1016/0165-1765\(93\)90167-b](https://doi.org/10.1016/0165-1765(93)90167-b)
 - [45] Nawata, K. (1994) Estimation of Sample Selection Bias Models by the Maximum Likelihood Estimator and Heckman's Two-Step Estimator. *Economics Letters*, **45**, 33-40. [https://doi.org/10.1016/0165-1765\(94\)90053-1](https://doi.org/10.1016/0165-1765(94)90053-1)
 - [46] Nawata, K. and Nagase, N. (1996) Estimation of Sample Selection Bias Models. *Econometric Reviews*, **15**, 387-400. <https://doi.org/10.1080/07474939608800363>