

# Why Rectified Linear Unit is Efficient in Machine Learning: One More Explanation

Barnabas Bede, Vladik Kreinovich, and Uyen Pham

**Abstract** In many applications, in particular, in econometric application, deep learning techniques are very effective. In this paper, we provide a new explanation for why rectified linear units – the main units of deep learning – are so effective. This explanation is similar to the usual explanation of why Gaussian (normal) distributions are ubiquitous – namely, it is based on an appropriate limit theorem.

## 1 Formulation of the Problem

**From traditional models to machine learning.** Traditionally, to describe econometric (and other) phenomena, researchers would:

- first come up with a generic parametric model – e.g., linear regression – and
- then find the values of these parameters for which the model provides the best fit for the data – provided, of course, that this is indeed a good fit.

In some cases, this works well, but in many other cases, no one has found a parametric model that describes the observations with a desired accuracy. To deal with such situations, it is desirable to have algorithms that do not require such a parametric model. Such algorithms are known as *machine learning*; see, e.g., [1, 2].

---

Barnabas Bede  
DigiPen Institute of Technology 9931 Willows Rd,  
Redmond, WA 98052, USA, e-mail: bbede@digipen.edu

Vladik Kreinovich  
Department of Computer Science, University of Texas at El Paso  
500 W. University, El Paso, Texas 79968, USA  
e-mail: vladik@utep.edu

Uyen Pham  
University of Economics and Law, Ho Chi Minh City, Vietnam e-mail: uyenph@uel.edu.vn

In machine learning, when we want to predict the desired future value  $y$  of some important quantity based on the known current and past values of relevant quantities  $x_1, \dots, x_n$ , we:

- first find the cases  $k = 1, \dots, K$  when we know both the values  $x_i^{(k)}$  and the value  $y^{(k)}$ , and
- then try to find the dependence  $y = f(x_1, \dots, x_n)$  for which  $y^{(k)} \approx f(x_1^{(k)}, \dots, x_n^{(k)})$  for all  $k$  – without a priori fixing the class of such dependencies.

**Neural networks and Rectified Linear Unit (ReLU).** At present, the most effective machine learning tool is a neural network – a tool that simulates how our brain processes the information. The basic unit of a neural network is a *neuron* that takes inputs  $z_1, \dots, z_m$ , and transform them into a value  $z = s(w_1 \cdot z_1 + \dots + w_m \cdot z_m - w_0)$  for some constants  $w_i$ , where  $s(t)$  is a non-linear increasing continuous function known as the *activation function*.

*Comment.* In this paper, by an increasing function, we mean a function with the property that if  $t \leq u$ , then  $s(t) \leq s(u)$ .

Outputs of neurons serve as inputs to other neurons, etc. The weights  $w_i$  are selected in such a way that for all known cases  $k = 1, \dots, K$ , the output of the neural network is close to the desired output  $y^{(k)}$ .

At first, the activation function was selected to be close to the activation function used by the biological neurons, i.e.,  $s(t) = 1/(1 + \exp(-t))$ . However, later, it turned out that in many cases, it is more effective to use a different activation function  $s(t) = \max(0, t)$ . This function is known as *Rectified Linear Unit*, or ReLU, for short.

**But why is ReLU so efficient?** There are many possible explanation of why ReLU is so efficient; see, e.g., [2, 3] and references therein, but the very fact that these new explanations appear all the time means that none of these explanations is fully convincing. It is therefore desirable to continue to come up with new explanations. This is what we do in this paper.

## 2 Our Main Idea and the Resulting Explanation

**Our main idea: looking for similar phenomena.** To come up with a desired explanation, let us recall similar situations when:

- first, some data processing technique was empirically shown to be very effective, and
- then a convincing explanation was found for this empirical success.

A natural example of this type is the ubiquity of Gaussian (normal) distributions.

**How this similar phenomenon is explained.** A usual explanation for this ubiquity comes from the Central Limit Theorem, according to which, under reasonable conditions:

- if a random variable is a sum of several small independent ones,
- then its distribution is closed to Gaussian; see, e.g., [4].

To be more precise, as the number of small components increases, the distribution of the sum tends to Gaussian. Under other conditions, we can have other distributions in the limit, e.g., Cauchy distribution with the probability density

$$f(x) = \frac{1}{\pi \cdot \Delta} \cdot \frac{1}{1 + \frac{(x-a)^2}{\Delta^2}}.$$

How do we know which probability distributions can appear as such limits? Clearly, since we talk about probability distributions of the sums, the sum of two limit distributions is also a limit distribution. Thus, the class of all limit distributions should be closed under convolution – the operation that transforms two probability density functions of two independent random variables into a probability density function describing their sum.

In the simplest case, when we consider families of distributions

$$a^{-1} \cdot f_0(a \cdot x + b)$$

for some function  $f_0(x)$ , this condition leads to a family of so-called *infinitely divisible distributions* – that includes Gaussian and Cauchy distributions. If we allow more-parametric families, then we can get more general families, e.g., convolutions of Gaussian and Cauchy distributions.

**Towards a similar explanation for ReLU: idea.** For neural networks, we do not have random variables, we have non-linear transformations.

If one layer performs some transformation  $y = f(x)$  and then the next layer transforms  $y$  into  $z = g(y)$ , then the resulting transformation from the input  $x$  to the final value  $z$  is a composition of the two functions  $z = g(f(x))$ .

An analog of adding a very small random variable – i.e., a transformation that practically does not change anything – is a close-to-identity transformation  $f(x)$  for which  $f(x) \approx x$  for all  $x$ .

Thus, it makes sense to consider limit functions that can be obtained if we consider compositions of many close-to-identity transformation. Just like the class of limit probability density functions is closed under convolution, the class of limit transformation functions should be closed under composition:

- if two functions  $f(x)$  and  $g(x)$  belong to this class,
- then their composition  $g(f(x))$  should also belong to this class.

**Towards a similar explanation for ReLU: details.** There are many classes of functions which are closed under composition. For example, a composition of two linear

functions is always a linear function, a composition of two fractional-linear functions is always fractional-linear, etc.

Out of all possible classes of functions which are closed under composition, we want to select the simplest class, determined by the smaller number of parameters.

In principle, the smaller number of parameters is 0, when the whole class consists of a single element – or of discretely many elements. For adding random variables, we cannot have a 0-parametric class, since the only random variable  $r$  for which the sum  $r_1 + r_2$  of two independent copies of this variable is distributed exactly as  $r$  is when  $r$  is equal to 0 with probability 1 – i.e., in effect, when there is no randomness at all.

However, in our case, for compositions of functions, it is possible to have a function whose composition with itself is exactly the same function, i.e., for which  $f(f(x)) = f(x)$  for all  $x$ . For example, the function  $f(x) = \max(0, x)$  corresponding to rectified linear unit has this property.

We are interested in the simplest possible families. In our case, the simplest possible families are 0-dimensional ones. So let us describe all 0-dimensional families, i.e., all continuous functions for which  $f(f(x)) = f(x)$  for all  $x$ . This description is provided by the following result.

**Proposition.** *For any continuous increasing function  $f(x)$  from real numbers to real numbers, the following two conditions are equivalent to each other:*

- we have  $f(f(x)) = f(x)$  for all  $x$ , and
- the function  $f(x)$  is equal:
  1. either to  $f(x) = x$ ,
  2. or to  $f(x) = \max(\underline{x}, x)$  for some  $\underline{x}$ ,
  3. or to  $f(x) = \min(\bar{x}, x)$  for some  $\bar{x}$ ,
  4. or to  $f(x) = \min(\bar{x}, \max(\underline{x}, x))$  for some  $\underline{x} \leq \bar{x}$ .

*Comment.*

- The first function does not change anything at all, so no transformation is performed.
- The second function is equal to  $\max(0, x - \underline{x}) + \underline{x}$  and can, therefore, be easily implemented by using a single ReLU unit plus appropriate linear transformations:
  - the transformation  $x \mapsto x - \underline{x}$  that precedes ReLU, and
  - the transformation  $y \mapsto y + \underline{x}$  that follows ReLU.
- The third function is equal to  $\bar{x} - \max(0, \bar{x} - x)$  and can, thus, also be implemented by using a single ReLU unit plus appropriate linear transformation.
- The fourth function is a composition of the second and the third ones and this, can be represented by two consequent ReLU units.

In all these cases, we have, modulo linear transformations, a ReLU unit. Thus, this result explains that the ReLU transformation is indeed, under appropriate conditions, a limit of compositions – i.e., naturally appears if we apply a large number of transformations one after another.

This limit result explains why ReLU units are so effective – they correspond to real-life situations where each signal transformation is performed continuously, step by step and is, therefore, in effect, a composition of many close-to-identity transformations corresponding to changes occurring during small time intervals.

**Proof.**

1°. It is easy to check that all four functions described in the formulation of the Proposition are continuous, increasing, and satisfy the condition that  $f(f(x)) = f(x)$  for all  $x$ .

So, to complete the proof, it is sufficient to prove that every continuous function  $f(x)$  that satisfies this condition has one of these four forms.

2°. Let us show that for all real numbers  $v$  from the range  $f(\mathbb{R}) \stackrel{\text{def}}{=} \{f(x) : x \in \mathbb{R}\}$ , we have  $f(v) = v$ .

Indeed, if the value  $v$  belongs to the range, this means that  $v = f(x)$  for some  $x$ . For this  $x$ , the condition  $f(f(x)) = f(x)$  means exactly  $f(v) = v$ .

3°. Since the function  $f(x)$  is continuous, its range is a connected set of real numbers, i.e., a finite or infinite interval.

This interval can be bounded from below and/or bounded from above. Let us consider all possible cases.

3.1°. Let us first consider the case when the range is neither bounded from below nor bounded from above.

In this case, the range coincides with the set of all real numbers. Thus, due to Part 2 of this proof, we have  $f(v) = v$  for all real numbers  $v$ .

3.2°. Let us now consider the case when the range is bounded from below but not from above.

Let  $\underline{x}$  denote the greatest lower bound of this range. Then, the range is equal to either  $(\underline{x}, \infty)$  or to  $[\underline{x}, \infty)$ .

For all  $v$  from this interval, we have  $f(v) = v$ . In particular, for every positive integer  $n$ , we have  $f(\underline{x} + 1/n) = \underline{x} + 1/n$ . Since the function  $f(x)$  is continuous, in the limit when  $n \rightarrow \infty$ , we get  $f(\underline{x}) = \underline{x}$ . Thus, the value  $\underline{x}$  also belongs to the range. Thus, the range has the form  $[\underline{x}, \infty)$ .

What will be the value  $f(x)$  for  $x \leq \underline{x}$ ? This value must belong to the range, i.e., it must be greater than or equal to  $\underline{x}$ :  $f(x) \geq \underline{x}$ . On the other hand, since the function  $f(x)$  is increasing, we must have  $f(x) \leq f(\underline{x}) = \underline{x}$ . Thus, for all such  $x$ , we must have  $f(x) = \underline{x}$ . So:

- for  $x \leq \underline{x}$ , we have  $f(x) = \underline{x}$ , and
- for  $x \geq \underline{x}$ , we have  $f(x) = x$ .

So, indeed, for all  $x$ , we have  $f(x) = \max(\underline{x}, x)$ .

3.3°. Let us consider the case when the range is bounded from above but not from below.

Let  $\bar{x}$  denote the least upper bound of this range. Then, the range is equal to either  $(-\infty, \bar{x})$  or to  $(-\infty, \bar{x}]$ .

For all  $v$  from this interval, we have  $f(v) = v$ . In particular, for every positive integer  $n$ , we have  $f(\bar{x} - 1/n) = \bar{x} - 1/n$ . Since the function  $f(x)$  is continuous, in the limit when  $n \rightarrow \infty$ , we get  $f(\bar{x}) = \bar{x}$ . Thus, the value  $\bar{x}$  also belong to the range. Thus, the range has the form  $(-\infty, \bar{x}]$ .

What will be the value  $f(x)$  for  $x \geq \bar{x}$ ? This value must belong to the range, i.e., it must be smaller than or equal to  $\bar{x}$ :  $f(x) \leq \bar{x}$ . On the other hand, since the function  $f(x)$  is increasing, we must have  $f(x) \geq f(\bar{x}) = \bar{x}$ . Thus, for such  $x$ , we must have  $f(x) = \bar{x}$ . So:

- for  $x \leq \bar{x}$ , we have  $f(x) = x$ , and
- for  $x \geq \bar{x}$ , we have  $f(x) = \bar{x}$ .

So, indeed, for all  $x$ , we have  $f(x) = \min(\bar{x}, x)$ .

3.4°. Finally, let us consider the remaining case when the range is bounded both from below and from above.

Let  $\underline{x}$  denote the greatest lower bound of this range, and  $\bar{x}$  denote its least upper bound. Then, the range is equal to one of the four possible intervals:  $(\underline{x}, \bar{x})$ ,  $(\underline{x}, \bar{x}]$ ,  $[\underline{x}, \bar{x})$ , and  $[\underline{x}, \bar{x}]$ .

For all  $v$  from the corresponding interval, we have  $f(v) = v$ . In particular, for every positive integer  $n$ , we have  $f(\underline{x} + 1/n) = \underline{x} + 1/n$ . Since the function  $f(x)$  is continuous, in the limit when  $n \rightarrow \infty$ , we get  $f(\underline{x}) = \underline{x}$ . Thus, the value  $\underline{x}$  also belong to the range.

Similarly, for every positive integer  $n$ , we have  $f(\bar{x} - 1/n) = \bar{x} - 1/n$ . Since the function  $f(x)$  is continuous, in the limit when  $n \rightarrow \infty$ , we get  $f(\bar{x}) = \bar{x}$ . Thus, the value  $\bar{x}$  also belong to the range. Thus, the range has the form  $[\underline{x}, \bar{x}]$ .

What will be the value  $f(x)$  for  $x \leq \underline{x}$ ? This value must belong to the range, i.e., it must be greater than or equal to  $\underline{x}$ :  $f(x) \geq \underline{x}$ . On the other hand, since the function  $f(x)$  is increasing, we must have  $f(x) \leq f(\underline{x}) = \underline{x}$ . Thus, for such  $x$ , we must have  $f(x) = \underline{x}$ .

What will be the value  $f(x)$  for  $x \geq \bar{x}$ ? This value must belong to the range, i.e., it must be smaller than or equal to  $\bar{x}$ :  $f(x) \leq \bar{x}$ . On the other hand, since the function  $f(x)$  is increasing, we must have  $f(x) \geq f(\bar{x}) = \bar{x}$ . Thus, for such  $x$ , we must have  $f(x) = \bar{x}$ . So:

- for  $x \leq \underline{x}$ , we have  $f(x) = \underline{x}$ ,
- for  $\underline{x} \leq x \leq \bar{x}$ , we have  $f(x) = x$ , and
- for  $x \geq \bar{x}$ , we have  $f(x) = \bar{x}$ .

So, indeed, for all  $x$ , we have  $\min(\bar{x}, \max(\underline{x}, x))$ .

The proposition is proven.

## Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science), and HRD-1834620 and HRD-2034030 (CAHSI Includes), and by the AT&T Fellowship in Information Technology.

It was also supported by the program of the development of the Scientific-Educational Mathematical Center of Volga Federal District No. 075-02-2020-1478, and by a grant from the Hungarian National Research, Development and Innovation Office (NRDI).

## References

1. C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.
2. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Cambridge, Massachusetts, 2016.
3. V. Kreinovich and O. Kosheleva, “Optimization under uncertainty explains empirical success of deep learning heuristics”, In: P. Pardalos, V. Rasskazova, and M. N. Vrahatis (eds.), *Black Box Optimization, Machine Learning and No-Free Lunch Theorems*, Springer, Cham, Switzerland, 2021, pp. 195–220.
4. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.