Sparsifying activation functions

Hélène Paugam-Moisy[1,2], **Sébastien Rebecchi**[1]
1 - TAO, INRIA - LRI, CNRS - Université Paris-Sud 11, F-91405 Orsay, France
2 - LIRIS, CNRS - Université Lumière Lyon 2, F-69676 Bron, France

In the infancy of backpropagation [1, 2], the shape of the (differentiable) activation function was investigated as well as other hyperparameters (the number of hidden layers, their size, the learning rate, etc.). The motive was likely that the two historical activation functions are limit cases of the family $\Phi = \{\phi_\lambda(x) = \frac{1-\exp(-\lambda x)}{1+\exp(-\lambda x)}\}_{\lambda \in \mathbb{R}^{+*}}\}$: linear function for $\lambda \to 0$ and hard threshold for $\lambda \to \infty$. Over time, people used to implement less rich models of multilayer neural networks and the choice for activation function was limited to the logistic sigmoid (in $]0, 1[$) or the hyperbolic tangent *tanh* (in $]-1, +1[$). Note that $tanh = \phi_2$ belongs to the above defined Φ family of activation functions. Such an oversimplification is not actually well-founded, as well as the reduction to 1 for the number of hidden layers. Whereas the number of hidden layers has been revisited recently with the "deep learning" current of research, the shape of activation functions is more seldom challenged, except in [3, 4].

On the other hand, learning to extract "sparse" features from data is a burning research area, in relation with compressed sensing, signal decomposition and dictionary learning. In the domain of neural networks, inspiration comes either from physics, with energy-based models [3] or free energy and linear filters [5], or from biology, with comparison to the response of a leaky integrate-and-fire (LIF) neuron model [4]. Starting from a mathematical point of view, we propose to consider a new family of activation functions, derived from Φ, which we call the *sparsifying activation functions*: $\Psi = \{\psi_{\lambda,\mu}(x) = \frac{1-\exp(-\lambda x)}{1+\exp(-\mu x)}\}_{(\lambda,\mu)\in\mathbb{R}^2} \ / \ \mu > \lambda > 0\}$.

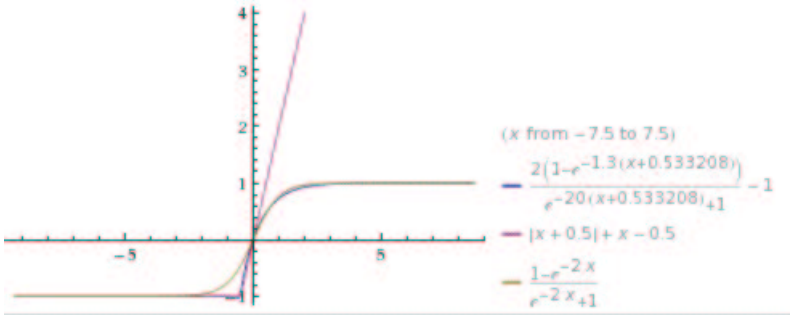

Figure 1: The sparsifier (blue) matches the rectifier (purple) for x<0 and matches the hyperbolic tangent (brown) for x>0.

Since $\mu > \lambda$, the asymptotic values of a $\psi_{\lambda,\mu}$ function are 0 in $-\infty$ and 1 in $+\infty$. However, for easier comparison with *tanh*, a $\psi_{\lambda,\mu}$ function, also called "sparsifier", as well as the "rectifier" (as defined in [4]), can be rescaled with asymptotic values -1 and $+1$. Fig. 1 shows how a $\psi_{\lambda,\mu}$ function

matches well the rectifier for $x < 0$ and *tanh* for $x > 0$. The numerator $1 - \exp^{-\lambda x}$ controls the behaviour of the positive section of the curve whereas the denominator $1 + \exp^{-\mu x}$ controls the negative section, independently. A peculiar point is that a sparsifying activation function is no longer monotonic: Its first derivative has a zero value in a negative point ($x = -0.533208$ in Fig. 1) and the function tends towards its negative asymptote by lower values. Such characteristics are not usual for artificial neuron activation functions, but henceforth the function is l_∞ nothing in the backpropagation algorithm prevents the use of a non monotonic activation function. The rectifier has no derivative in zero, which yields the authors of [4] to add several artefacts to their neural network. Moreover, the sparsifier is much more similar to the LIF neuron model response and its small negative section (around the zero of its derivative) may also find a biological justification from the shape of the well-known "mexican hat" function, observable in many biological neuron activities.

First experiments have shown several advantages of the sparsifying activation functions for backpropagation learning: similar classification rates are reached after a fewer number of epochs; the performance is more stable through time (compared to rectifying neurons); a sparsifier better prevents overfitting (compared to *tanh*); and, last but not least, the sparsifying activation functions yield learning sparse representations of the data. Finally, compared to the "sparsifying logistic" and the algorithm defined in [3], the implementation of sparsifying activation functions is much more straightforward.

References

[1] Y. LeCun. Une procédure d'apprentissage pour réseau à seuil asymétrique (a learning scheme for asymmetric threshold networks). In *Cognitiva 85*, 1985.

[2] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[3] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun. Efficient learning of sparse representations with an energy-based model. In *NIPS*, 2006.

[4] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *AISTATS*, pages 315–323, 2011.

[5] M. Ranzato, Y. Boureau, and Y. LeCun. Sparse feature learning for deep belief networks. In *NIPS*, 2007.

Topic: Learning algorithms
Preference: Oral or poster