11th International Conference on Sensitivity Analysis of Model Output, April 23–25 2025, Grenoble, France

Forward Sweep Interval Sensitivity in Neural Network Functional Approximation

Dawid Ochnio
Marco de Angelis
University of Strathclyde, Glasgow, Scotland

Interval-based sensitivity is an efficient global sensitivity analysis method that is based on interval arithmetic [1]. It works by partitioning the input of interest into sub intervals, while the other inputs are intact intervals. The method is non-probabilistic and can be used to calculate the sensitivity of a function without relying on sampling, which might not capture the whole input space. For example, methods based on sampling need the definition of a probability distribution, which is often chosen arbitrarily. Interval sensitivity only need specification of the input space where the sensitivity is to be calculated [2].

Global sensitivity analysis can be an effective tool against over parametrization in neural networks. Over parametrization arises when a trained model has too many units or layers and can cause issues, including over-fitting, poor explainability, suboptimality, excessive memory usage, and more. Being able to determine sensitivities towards the output and knowing if there are parameters/units that have a negligible effect and as such can be discarded with no significant loss of performance can be consequential, leading to leaner layouts and more explainable models. Deep learning models are often deemed to be "black boxes" because of their nearly impenetrable mathematical layout. Interestingly however, neural network models always imply a clear, albeit intricate, mathematical function, whose expression is the composition of as many functions as there are layers. The model's mathematical expression can be obtained simply knowing the network's architecture and the trained weights and biases.

The network's forward sweep is a function of the network inputs t, the weights W and biases b. In this study, the inputs of the sensitivity analysis are W and b, so we can write the forward sweep as follows $y = f(t, x)$, where t is the network input and $x \in \mathrm{R}^d$ is the vector of sensitivity inputs (network parameters).

Let $f : \mathrm{R}^d \to \mathrm{R}$ be the forward sweep that is a function only of the network parameters, and let f be the its interval extension. Let $[\underline{y}, \overline{y}]_{i,n}$ be the output corresponding to the n-th subinterval when only the i-th input is partitioned and the other intervals are intact, in notation $[\underline{y}, \overline{y}]_{i,n} = \mathsf{f}([\underline{x_1}, \overline{x_1}], ..., [\underline{x_i}, \overline{x_i}]_n, ..., [\underline{x_d}, \overline{x_d}])$. The interval-based sensitivity index, for the i-th input, is

$$S_i = 1 - \frac{\sum_n^N [\underline{x_i}, \overline{x_i}]_n \cdot [\underline{y}, \overline{y}]_{i,n}}{[\underline{y}, \overline{y}] \cdot [\underline{x_i}, \overline{x_i}]}, \tag{1}$$

where N is the number of subintervals for the i-th partition, $\cdot$ is the interval multiplication used to calculate the area of the (sub) rectangles, $[\underline{x_i}, \overline{x_i}]_n$ is the n-th subinterval for the i-th partition, such that $\cup_n^N [\underline{x_i}, \overline{x_i}]_n = [\underline{x_i}, \overline{x_i}]$, and $[\underline{y}, \overline{y}] = \mathsf{f}([\underline{x}, \overline{x}])$ is the overall output range. The numerator in (1) is the sum of all subinterval areas and the denominator is the area of the box enclosing the xy graph. The sensitivity index $S_i,\ i = 1, ..., d$ ranges from 0 to 1. When $S_i = 0$, the partitioning has no effect, the numerator is equal to the denominator and so y has no functional dependence on x_i. When $S_i = 1$, the sub rectangular areas are zero and so y has full functional dependence on x_i. It is worth noticing that this sensitivity indices are immune to the curse of dimensionality because the partitioning takes place in one dimension

In the example, a neural network with two layers and five ReLU units is trained to approximate the cubic function $y = t^3 - 3t^2 + 2t + 5$. Sensitivity indices are computed for each parameter in the trained neural network, namely weights and biases $W^{(1)} \in \mathrm{R}^{1\times5}$, $b^{(1)} \in \mathrm{R}^5$, $W^{(2)} \in \mathrm{R}^{5\times1}$, $b^{(2)} \in \mathrm{R}$. The trained network settled on the following values: $W^{(1)} = ((-1, -1, 0, 1, 1))$, $b^{(1)} = (-1, 0, 0, -2, -3)$, $W^{(2)} = ((-13, -5, 0, 5, 13))$, and $b^{(2)} = 5$. The interval sensitivity is calculated by intervalizing the inputs with a radius of ±1 for each parameter and a partition of $N = 30$. The inputs are organized in the single vector $x = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)})$ of size $d = 16$.

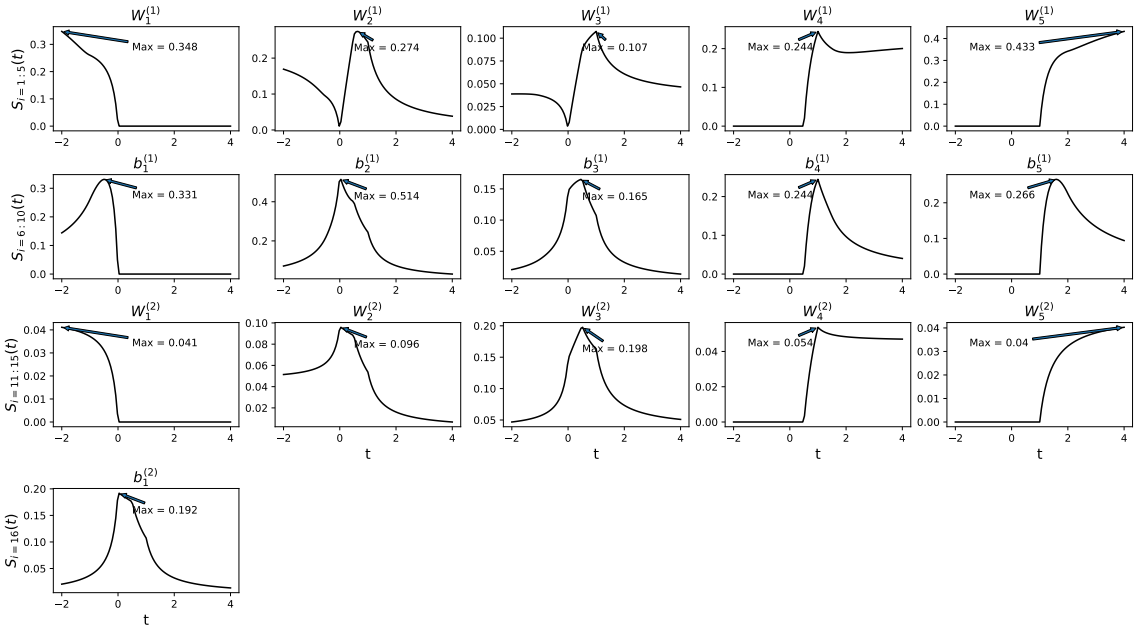

Figure 1: Sensitivity indices across network's input t for $x = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)})$.

The neural network is trained to approximate the above cubic function, whose graph changes sign at $t = 1$. All sensitivity indices reflect this displaying peaked values around it, as shown in Figure 1. We notice that the weights describing the negative values of the cubic function, namely $W_1^{(1)}$, $b_1^{(1)}$, $W_1^{(2)}$ display high sensitivities for the negative values and zero sensitivities for the positive values. Similarly, the weights describing the positive values $W_4^{(1)}$, $b_4^{(1)}$, $W_4^{(2)}$, $W_5^{(1)}$, $b_5^{(1)}$, $W_5^{(2)}$ show the same pattern for positive and negative values respectively. It is worth noticing that these network parameters display zero sensitivities for the region of the space that they do not describe, as expected. Other parameters have more complex dependencies on the output. This study has also shown that for this particular example, the network is not over parametrized so, none of the inputs can be ignored without affecting the network's accuracy in approximating the cubic function. The sensitivities can also be used to see what units are active in the regions of interest, providing a diagnostic tool to reason about the parameters role in the overall architecture.

References

[1] "IEEE Standard for Interval Arithmetic," in *IEEE Std 1788-2015*, pp.1-97, 30 June 2015

[2] Miralles-Dolz E., Gray A., de Angelis M., Patelli E. "Interval-Based Global Sensitivity Analysis for Epistemic Uncertainty", in: *Proceedings of the 32nd European Safety and Reliability Conference (ESREL 2022).* Research Publishing, IRL, pp. 2545-2552, 2022.

[Dawid Ochnio; University of Strathclyde, dawid.ochnio.2020@uni.strath.ac.uk]
[`dawid.ochnio.2020@uni.strath.ac.uk` –]