

Multilayer Perceptron-based Surrogate Models for Finite Element Analysis

Lawson Oliveira Lima, Julien Rosenberger, Esteban Antier, Frédéric Magoulès
 Université Paris-Saclay, CentraleSupélec, MICS, Gif-sur-Yvette, France
 Email: frederic.magoules@hotmail.com

Abstract—Many Partial Differential Equations (PDEs) do not have analytical solution, and can only be solved by numerical methods. In this context, Physics-Informed Neural Networks (PINN) have become important in the last decades, since it uses a neural network and physical conditions to approximate any functions. This paper focuses on hypertuning of a PINN, used to solve a PDE. The behavior of the approximated solution when we change the learning rate or the activation function (sigmoid, hyperbolic tangent, GELU, ReLU and ELU) is here analyzed. A comparative study is done to determine the best characteristics in the problem, as well as to find a learning rate that allows fast and satisfactory learning. GELU and hyperbolic tangent activation functions exhibit better performance than other activation functions. A suitable choice of the learning rate results in higher accuracy and faster convergence.

Index Terms—Multilayer perceptron; Neural networks; Physics-Informed Neural Networks; Partial Differential Equations; Hypertuning;

I. INTRODUCTION

Artificial intelligence draws people attention thanks to its various spectrum of application fields such as language processing, computer vision or decision making. Lagaris described in [17] a general method to solve a PDE with a neural network instead of classical methods as Finite Element Methods (FEM). Finite Element Methods [1] [2] are powerful methods to approximate solutions of a PDE. The accuracy of these methods can be improved using various stabilization techniques like [6], [5], [4], [3]. Since FEM solution requires the solution of a large linear system, computation cost becomes prohibitive. As an alternative, using neural network [15] will help construct function from one already forced to satisfy the boundary conditions of the problem. The method is broaden in [16] where the straightforward method of solving a PDE on complex geometries without assisting the neural network to satisfy the boundary conditions is highlighted [18]. In the 1990's, Physics-Informed Neural Networks were first introduced and show great expectations in terms of efficiency and versatility [19].

The paper is organized as follows. In Section II, we present concepts in Neural Networks, Partial Differential Equations and Physics Informed Neural Networks. In Section III, we present the methodology followed to allow hypertuning of the parameters. In Section IV, we present numerical results and we conclude in Section V.

II. NEURAL NETWORKS FOR SOLVING PARTIAL DIFFERENTIAL EQUATIONS

Neural Network: We will focus on the study of Multilayer Perceptron (MLP), one of the first neural networks developed. It consists in an input layer, a vector of n components, and an output layer, a vector of p components. The input is subjected to a series of linear operations and nonlinear activations are applied through, so called hidden layers, to get the output. A MLP can be seen as a function: $\mathcal{N} : \mathbb{K}^n \rightarrow \mathbb{K}^p$. If we note g_i the (non-linear) activation function, and W_i and b_i the weight and bias matrix of the i^{th} -hidden layer. Then, for a MLP with k hidden layers we can write:

$$\forall x \in \mathbb{K}^n, \quad \mathcal{N}(x) = g_k(b_k + W_k g_{k-1}(b_{k-1} + W_{k-1} g_{k-2}(\dots (b_1 + W_1 x))))$$

Through training epochs, the MLP passes through the training dataset coupled with gradient descent so that \mathcal{N} approaches a function of interest.

Our goal is to solve a Partial Differential Equation (PDE) defined within a domain Ω :

$$(E) : \begin{cases} \mathcal{Q}(u, \nabla u, Hu, \dots)(x) = 0 & \text{inside } \Omega \\ \mathcal{R}(u, \nabla u, Hu, \dots)(x) = f(x) & \text{on } \partial\Omega \end{cases}$$

where \mathcal{Q} and \mathcal{R} are linear operators, Ω is the domain considered, $\partial\Omega$ the boundary of the domain considered, and $u : \mathbb{K}^n \rightarrow \mathbb{K}^p$ the solution to (E).

We want to replace u with a neural network \mathcal{N} and train it to get a solution of (E). However, this would mean training on both the domain and its boundary. To avoid this case, as the PDE is always defined with boundary conditions, we reformulate the problem by replacing u with a function Ψ defined as: $\Psi(x) = A(x) + F(x)\mathcal{N}(x)$ where A verifies the boundary conditions, and F is null on the boundary. As such, we unconstrained the problem, and the neural network can train inside the domain, with the boundary conditions handled by the above mentioned functions.

Activation functions: Activation functions set the behaviour of a neuron at a low level in the network. Their impacts can be more or less fast to compute and drastically change the convergence speed and quality. We consider in Equations (1a)-(1e) the most rampant activation functions in the state-of-the-art articles:

$$\forall z \in \mathbb{R}, \quad \sigma(z) := \frac{1}{1 + e^{-z}} \quad (1a)$$

$$\forall z \in \mathbb{R}, \tanh(z) := \frac{e^z - e^{-z}}{e^z + e^{-z}} = \sigma(2z) - \sigma(-2z) \quad (1b)$$

$$\alpha \in [0, 1], \forall z \in \mathbb{R}, \text{ReLU}(z) := \max(\alpha z, z) \quad (1c)$$

$$\alpha \in \mathbb{R}^*, \forall z \in \mathbb{R}, \text{ELU}(z) := \begin{cases} z, & \text{if } z \geq 0 \\ \alpha(e^z - 1), & \text{if } z < 0 \end{cases} \quad (1d)$$

$$\forall z \in \mathbb{R}, \text{GELU}(z) = 0.5z \left(1 + \tanh \left[\sqrt{\frac{2}{\pi}} (z + 0.044715z^3) \right] \right) \quad (1e)$$

The first one Equation (1a) coming from an imitation of biological neurons' behaviour is the sigmoid. Inspired from the previous one, but zero-centered, is the hyperbolic tangent Equation (1b). Yet, activation functions easier and faster to compute arose. The most notorious one is the Rectified Linear Unit (ReLU) function Equation (1c) and it holds many variants. A smoother variant is the Exponential Linear Unit (ELU) function. This function is defined Equation (1d). The final variant to introduce is the Gaussian Error Linear Unit (GELU) function presented in [20]. This function tries to embed the Dropout regularizer by randomly multiplying outputs of neurons to the activation function itself. For using this function, the inputs have to be batch normalized first. The GELU function is defined Equation (1e). The different activation functions are represented in Figure 1.

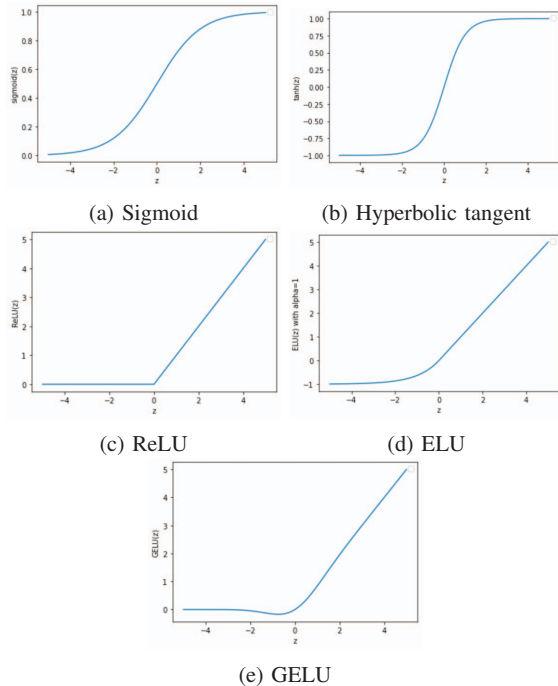


Fig. 1: Activation functions.

III. METHODOLOGY

Problem Statement: We want to solve the following PDE in a square domain $\Omega = [0, 1]^2$ using Physics-Informed Neural Networks (PINN).

$$\Delta\psi(x, y) + \psi(x, y) \cdot \frac{\partial\psi(x, y)}{\partial y} = f(x, y) \quad \text{on } \Omega = [0, 1]^2 \quad (2)$$

where $f(x, y) = \sin(\pi x)(2 - \pi^2 y^2 + 2y^3 \sin(\pi x))$ and with the boundary conditions:

$$\begin{cases} \psi(0, y) = \psi(1, y) = \psi(x, 0) = 0 \\ \frac{\partial\psi}{\partial y}(x, 1) = 2 \sin(\pi x) \end{cases}$$

This problem has an analytical solution: $\psi_{th}(x, y) = y^2 \sin(\pi x)$. Thus, to solve it, we define our loss function as:

$$\mathcal{L} = \|\Delta\psi(x, y) + \psi(x, y) \cdot \frac{\partial\psi(x, y)}{\partial y} - f(x, y)\|_2$$

with $\psi_{ap}(x, y)$ defined as $\psi_{ap}(x, y) = A(x, y) + F(x, y)N(x, y)$ and with A and F equal to $A(x, y) = y \sin(\pi x)$ and $F(x, y) = \sin(x-1) \sin(y-1) \sin(x) \sin(y)$. A careful reader will notice the lack of knowledge about the exact solution of the PDE. Furthermore, the model is called a PINN as the loss takes into account the residual of Equation (2).

Implementation: The implementation of the multilayer perceptron is similar to [7] [8] and is done using the Python library Jax, which has high performance and low computational cost due to being built using XLA, and is thus recommended for simulations involving neural networks. Furthermore, it can be used to vectorize functions and execute them faster in Graphics Processing Unit (GPU). The steps of the model used are described in the algorithm below.

Solving PDEs using neural networks (PINN method)

Input domain and boundary-coords, tolerance, A and F

Output trained-PINN

set max-epoch, max-train-step, batch-size, validation-size

set learning rate and activation function

while loss \geq tolerance **do**

for epoch \leq max-epoch **do**

 sample-coords \leftarrow sample batch-size points + noise

 loss \leftarrow compute \mathcal{L}_2 -norm on PINN-model and the PDE

 apply gradient descent and backpropagation to reduce

 loss

end for

end while

make the validation of the model

return PINNmodel

To achieve a better performance in a shorter time, we have vectorized it to compute in parallel the neural network evaluation and make gradient descent on GPU.

Hyperparameters: The hypertuning comes when one wants to construct a model and a training that will end to a low validation error. Let's list all the hyperparameters involved in this scheme: (i) `l_units`: list of integers depicting the number of hidden layers and the number of neurons per layer of the

sequential machine learning model; (ii) `l_activations`: list depicting the activation functions of each layer of the model; (iii) `noise`: float in $[0, 1]$ conditioning the use of a Gaussian layer of mean 0 after the input layer; (iv) `stddev`: the standard deviation of the Gaussian layer when it is used; (v) `optimizer`: string depicting the optimizer used; (vi) `learning_rate`: float setting the learning rate of the previous optimizer; (vii) `epochs_max`: integer setting the number of maximum epoch loops; (viii) `batch_size`: integer setting the number of points extracted from the mesh to construct the training set at each epoch loop. Hypertuning consists of adapting the parameters of the neural network in order to obtain a model that has the lowest validation error. To do this, several models are trained and the one with the best performance is chosen

IV. NUMERICAL EXPERIMENTS

For the hyperparameters, we used `batch_size` = 50 and a multilayer perceptron with 3 completely connected layers, the first layer has 2 neurons, the second 30 and the last 1, each followed by the activation function. Thus the neural network considered has 121 trainable parameters: hidden layer = $2 * 30 + 30 = 90$ and output layer = $30 * 1 + 1 = 31$.

A. First parameter: Learning rate

Studying Equation (2) we note that each activation function has a learning rate that improve its accuracy and reduce the loss function. Thus, to get a good solution, we have to find a model that minimizes it. It's possible to see the behavior between learning rate and loss function in Figure 2 for each one of the studied functions. In order to find it, we have used grid and random search in pre-defined intervals.

Grid search: Grid search is a very common way to make hyperparameter tuning. It consists of determining a domain and then using points that are uniformly distributed. So, to use this method, we chose 5 intervals with 50 learning rates in each and then ran simulations to obtain the best loss function. The mean absolute error and respective learning rate for each activation function using grid search are presented in Table I.

| MAE at epoch | Tanh | Sigmoid | GELU | ELU | ReLU |
|---------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| 10000 | $1.71 \cdot 10^{-5}$ | $3.10 \cdot 10^{-5}$ | $1.84 \cdot 10^{-4}$ | $5.38 \cdot 10^{-5}$ | 0.016 |
| 20000 | $1.43 \cdot 10^{-5}$ | $2.18 \cdot 10^{-5}$ | $9.92 \cdot 10^{-6}$ | $1.65 \cdot 10^{-4}$ | 0.016 |
| 50000 | $1.05 \cdot 10^{-5}$ | $1.08 \cdot 10^{-5}$ | $8.03 \cdot 10^{-6}$ | $1.57 \cdot 10^{-4}$ | 0.016 |
| Learning rate | $3.2040 \cdot 10^{-4}$ | $1.3673 \cdot 10^{-5}$ | $1.1837 \cdot 10^{-4}$ | $1.1837 \cdot 10^{-3}$ | $6.6939 \cdot 10^{-2}$ |

TABLE I: Grid search (grey is the best accuracy).

Random search: Random search is similar in principle to grid search, using points distributed over a domain. However, the main difference is that they are spread randomly, which allows testing parameters that would not be tested using the first method. Thus, this procedure was performed on the same intervals as previously used. Results of the mean absolute error and respective best learning rate for each activation function using random search are collected in Table II.

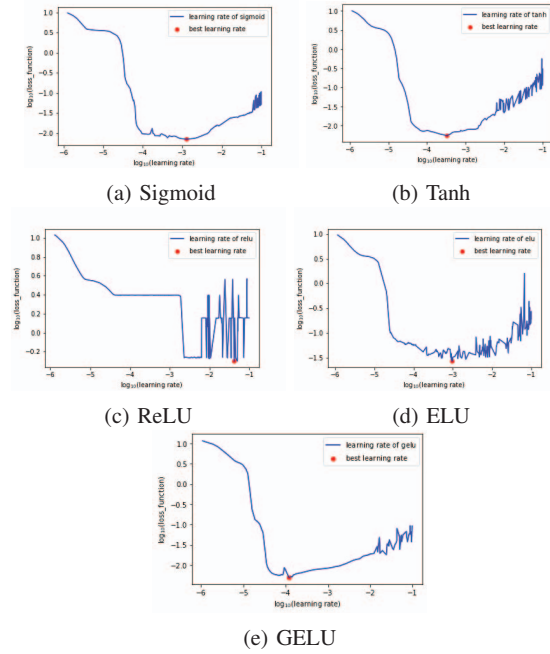


Fig. 2: Learning rates.

| MAE at epoch | Tanh | Sigmoid | GELU | ELU | ReLU |
|---------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| 10000 | $3.10 \cdot 10^{-5}$ | $6.78 \cdot 10^{-5}$ | $1.85 \cdot 10^{-4}$ | $5.48 \cdot 10^{-5}$ | 0.016 |
| 20000 | $2.18 \cdot 10^{-5}$ | $5.75 \cdot 10^{-5}$ | $1.47 \cdot 10^{-5}$ | $5.86 \cdot 10^{-5}$ | 0.016 |
| 50000 | $1.12 \cdot 10^{-5}$ | $1.08 \cdot 10^{-5}$ | $6.31 \cdot 10^{-6}$ | $1.08 \cdot 10^{-4}$ | 0.016 |
| Learning rate | $3.2915 \cdot 10^{-4}$ | $1.2626 \cdot 10^{-5}$ | $1.1891 \cdot 10^{-4}$ | $9.5295 \cdot 10^{-4}$ | $4.1886 \cdot 10^{-2}$ |

TABLE II: Random search (grey is the best accuracy).

Analysis: Mean absolute error and respective learning rate for each activation function with best learning rates are presented in Table III. We can see that random search presented the lowest errors for almost all functions, except for sigmoid. An expected result, since it does not follow a pattern to test each learning rate. Furthermore, it can be seen that the average absolute error obtained is quite small for GELU, tanh and sigmoid. We compare for the same accuracy with FEM on GPU. The linear system of equations is solved with the Alinea library [12]. The problem is reformulated in parallel with a domain decomposition method [13] and is solved on GPU [11] [9]. The local solver inside each subdomain is performed with the conjugate gradient method [10] involving an auto-tuning of the GPU memory [14]. We obtain a computational time of 25 seconds for $50 \cdot 10^3$ epochs for the MLP surrogate models and for the FEM method involving $2.5 \cdot 10^6$ nodes, more time.

B. Second parameter: Activation function

With the best learning rate for each activation function, we report in Figure 3, the comparison in order to find the best solution to the problem. The activation functions ELU and ReLU did not obtain satisfactory results since they reached a local minimum that did not allow for good accuracy. On the other hand, GELU has the best performance and the hyperbolic tangent has similar results. Finally, we see that sig-

| MAE at epoch | Tanh | Sigmoid | GELU | ELU | ReLU |
|---------------|------------------------|------------------------|------------------------|------------------------|------------------------|
| 10000 | $1.71 \cdot 10^{-5}$ | $3.10 \cdot 10^{-5}$ | $1.85 \cdot 10^{-4}$ | $5.48 \cdot 10^{-5}$ | 0.016 |
| 20000 | $1.43 \cdot 10^{-5}$ | $2.18 \cdot 10^{-5}$ | $1.47 \cdot 10^{-5}$ | $5.86 \cdot 10^{-5}$ | 0.016 |
| 50000 | $1.12 \cdot 10^{-5}$ | $1.05 \cdot 10^{-5}$ | $6.31 \cdot 10^{-6}$ | $1.08 \cdot 10^{-4}$ | 0.016 |
| Learning rate | $3.2040 \cdot 10^{-4}$ | $1.3673 \cdot 10^{-5}$ | $1.1891 \cdot 10^{-4}$ | $9.5295 \cdot 10^{-4}$ | $4.1886 \cdot 10^{-2}$ |

TABLE III: Best learning rates (grey is the best accuracy).

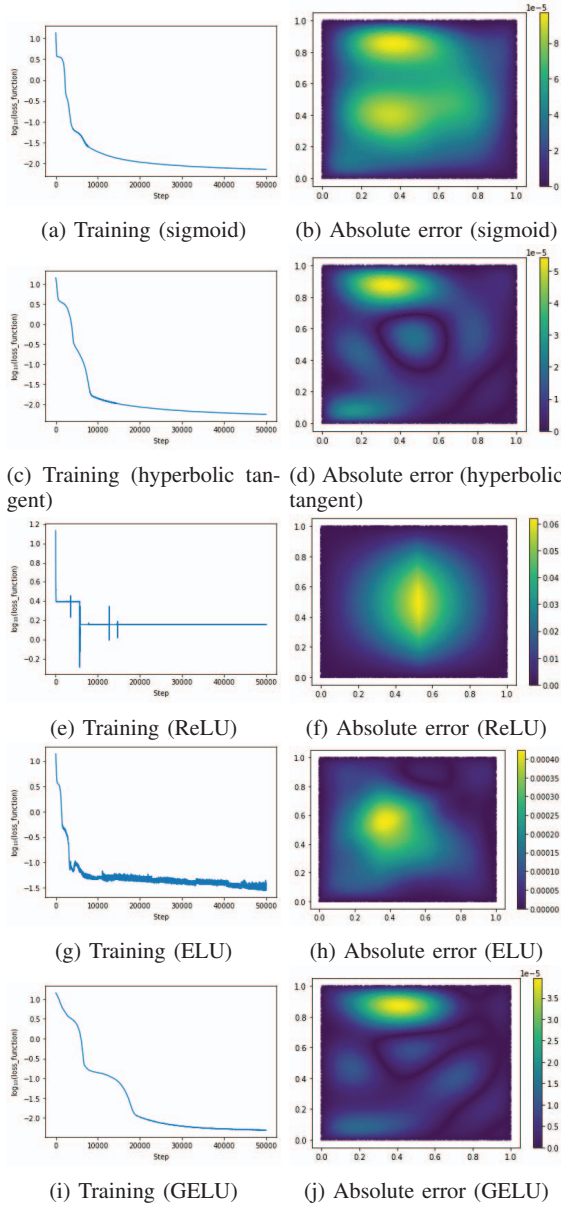


Fig. 3: Training and absolute error.

moid has performance as good as GELU and tanh. Additional simulations with varying hyperparameters did not change the performance of one over the other.

V. CONCLUSIONS

It was observed that GELU and hyperbolic tangent has lower absolute error for a smaller number of epochs, a very

favorable characteristic, since it makes the neural network easier to train. Also, sigmoid performed well, but ELU and ReLU did not obtain satisfactory results. Moreover, with the optimization of the learning rate in the hyperbolic tangent and GELU a good mean absolute error was obtained for 10 000 and 20 000 epochs, which enable us to reduce the number of epochs and the training time.

VI. ACKNOWLEDGEMENTS

The authors would like to thank M. Besbes, A. Paun, and G. Ruault for the useful discussions.

REFERENCES

- [1] P.G. Ciarlet. The Finite Element Method for Elliptic Problems. SIAM. 2002.
- [2] T. Hughes. The Finite Element Method: Linear Static and Dynamic Finite Element Analysis. Dover Publications Inc. 2003.
- [3] J.E. Dolbow and L.P. Franca. Residual-free bubbles for embedded Dirichlet problems. Computer Methods in Applied Mechanics and Engineering. 197(45-48):3751-3759, 2008.
- [4] F. Magoulès and H. Zhang. Three-dimensional dispersion analysis and stabilised finite element methods for acoustics. Computer Methods in Applied Mechanics and Engineering, 335:563-583, 2018.
- [5] I. Harari and F. Magoulès. Numerical investigations of stabilised finite element computations for acoustics. Wave Motion, 39(4):339-349, 2004.
- [6] A. Russo. Streamline-upwind Petrov/Galerkin method (SUPG) vs residual-free bubbles (RFB). Computer Methods in Applied Mechanics and Engineering. 195(13-16):1608-1620, 2006.
- [7] F. Magoulès, H.-X. Zhao, and D. Elizondo. Development of an RDP neural network for building energy consumption fault detection diagnosis. Energy and Buildings. 62:133-138, 2013.
- [8] H.-X. Zhao and F. Magoulès. A review on the prediction of building energy consumption. Renewable and Sustainable Energy Reviews, 16(6):3586-3592, 2012.
- [9] A.-K. Cheik Ahamed and F. Magoulès. Efficient implementation of Jacobi iterative method for large sparse linear systems on graphic processing units. The Journal of Supercomputing, 73(8):3411-3432, 2017.
- [10] A.-K. Cheik Ahamed and F. Magoulès. Conjugate gradient method with graphics processing unit acceleration: CUDA vs OpenCL. Advances in Engineering Software, 111:32-42, 2017.
- [11] F. Magoulès, A.-K. Cheik Ahamed, and A. Suzuki. Green computing on graphics processing units. Concurrency and Computation: Practice and Experience, 28(16):4305-4325, 2016.
- [12] F. Magoulès and A.-K. Cheik Ahamed. Alinea: An advanced linear algebra library for massively parallel computations on graphics processing units. International Journal of High Performance Computing Applications, 29(3):284-310, 2015.
- [13] F. Magoulès, A.-K. Cheik Ahamed, and R. Putanowicz. Optimized Schwarz method without overlap for the gravitational potential equation on cluster of graphics processing unit. International Journal of Computer Mathematics, 93(6):955-980, 2016.
- [14] F. Magoulès, A.-K. Cheik Ahamed, and R. Putanowicz. Auto-tuned Krylov methods on cluster of graphics processing unit. International Journal of Computer Mathematics, 92(6):1222-1250, 2015.
- [15] J. Blechschmidt and O.G. Ernst. Three Ways to Solve Partial Differential Equations with Neural Networks - A Review. GAMM-Mitteilungen, 44:2 (e202100006), 2021.
- [16] J. Berg and K. Nystrom. A unified deep artificial neural network approach to partial differentiate equations in complex geometries. Neurocomputing, 317 (28-41), 2019.
- [17] I.E. Lagaris, A. Likas, and D.I. Fotiadis. Artificial Neural Networks for Solving Ordinary and Partial Differential Equations. IEEE Transactions on Neural Networks, volume 9, number 5 (987-1000), 1997.
- [18] K. Rudd and S. Ferrari. A constrained integration (CINT) approach to solving partial differential equations using artificial neural networks. Neurocomputing, 155 (277-285), 2015.
- [19] M. Raissi, P. Perdikaris, and G. Em Karniadakis. Physics Informed Deep Learning (Part I): Data-driven Solutions of Nonlinear Partial Differential Equations. Journal of Computational Physics, 378 (686-707), 2019.
- [20] D. Hendrycks, K. Gimpel. Gaussian Error Linear Units (GELUs). arXiv e-prints, arXiv:1606.08415, 2018.