

A practical PINN framework for multi-scale problems with multi-magnitude loss terms

Yong Wang^{a,b}, Yanzhong Yao^{a,*}, Jiawei Guo^b, Zhiming Gao^a

^a*Institute of Applied Physics and Computational Mathematics, Beijing 100088, China*
^b*Graduate School of China Academy of Engineering Physics, Beijing 100088, China*

Abstract

For multi-scale problems, the conventional physics-informed neural networks (PINNs) face some challenges in obtaining available predictions. In this paper, based on PINNs, we propose a practical deep learning framework for multi-scale problems by reconstructing the loss function and associating it with special neural network architectures. New PINN methods derived from the improved PINN framework differ from the conventional PINN method mainly in two aspects. First, the new methods use a novel loss function by modifying the standard loss function through a (grouping) regularization strategy. The regularization strategy implements a different power operation on each loss term so that all loss terms composing the loss function are of approximately the same order of magnitude, which makes all loss terms be optimized synchronously during the optimization process. Second, for the multi-frequency or high-frequency problems, in addition to using the modified loss function, new methods upgrade the neural network architecture from the common fully-connected neural network to special network architectures such as the Fourier feature architecture given in Ref. [1] and the integrated architecture developed by us. The combination of the above two techniques leads to a significant improvement in the computational accuracy of multi-scale problems. Several challenging numerical examples demonstrate the effectiveness of the proposed methods. The proposed methods not only significantly outperform the conventional PINN method in terms of computational efficiency and computational accuracy, but also compare favorably with the state-of-the-art methods in the recent literature. The improved PINN framework facilitates better application of PINNs to multi-scale problems. The data and code accompanying this paper are available at <https://github.com/wangyong1301108/MMPINN>.

Keywords: Deep learning, Multi-scale problems, Physics-informed neural networks, Balancing loss terms, Fourier feature architecture

*Corresponding author

Email address: yao_yanzhong@iapcm.ac.cn (Yanzhong Yao)

1. Introduction

With the rapid development of artificial intelligence technology and its increasingly widespread application in production and life, deep learning methods represented by PINN (Physics-informed neural networks) [2, 3, 4, 5] have gradually become a popular class of methods for solving partial differential equations (PDEs). Compared with traditional numerical methods, such as finite difference, finite element and finite volume methods, the PINN methods have the following obvious advantages: they are mesh-free, thus avoiding the hassles of mesh generation on complex regions and the construction of high-precision discrete schemes on meshes of poor geometric quality; they are applicable to high-dimensional problems, and the construction and implementation of the methods are dimensionality independent; they produce a prediction function over the entire computational domain, rather than a discrete solution on the meshes as in mesh-based methods. The PINN methods have received much attention for their ability to integrate physical laws and labeled data into loss functions, and have led to many notable results in computational science and engineering [6, 7, 8]. More and more researchers have realized the immense potential of deep learning algorithms.

For multi-scale problems, considering their importance in practical applications [9, 10, 11], many researchers have devoted themselves to studying the deep learning methods. In Ref. [12], based on the idea of radial scaling in the frequency domain and activation functions with compact support, Liu et al. propose Multi-scale Deep Neural Networks (MscaleDNNs) to solve elliptic PDEs with rich frequency contents. In Ref. [13], combining traditional numerical analysis and the MscaleDNNs algorithm, Li et al. construct SD²NN based on subspace decomposition to capture the smoothed and oscillatory parts of the multi-scale solution for elliptic type multi-scale PDEs. In Ref. [14], Jin et al. devise an asymptotic-preserving neural networks to solve the multi-scale uncertain linear transport equation with diffusive scaling, overcoming the computational challenges of the curse of dimensionality and the multi-scale problems. The above works have obtained distinctive research results for specific models such as elliptic equations, transport equations, etc., which provide some new ideas for solving multi-scale problems using deep learning method. However, it is still a topic for further research to construct a general deep learning framework for multi-scale problems based on the analysis of their common characteristics, so as to solve multi-scale models with high accuracy under a unified framework.

The motivation of this paper is to construct a practical deep learning framework based on PINNs for multi-scale problems. This is undoubtedly a challenging task. In Refs. [1, 15, 16], Wang et al. mention that the conventional PINN method has some weaknesses for multi-scale problems, which is proved by a large number of numerical experiments. Through investigation, we believe that the main reasons why the conventional PINN method is difficult to train multi-scale problems are as follows: For most multi-scale problems, there is usually a very large order of magnitude difference between the supervised term and the residual term of the loss function. When the loss function is optimized,

the loss term with a small order of magnitude is often not optimized, and even gets worse. Wang et al. also find this phenomenon. In Ref. [15], they point out that due to the gradient pathologies in PINNs, there will be a dominance of PDE residuals in the optimization process, which will lead to the training being heavily biased towards the residual term, and to larger errors in the fitting of the supervised term. In addition, for the multi-scale problems with multi-frequency or high-frequency features, solving them by the conventional PINN method not only encounters the mentioned problem of magnitude differences, but also struggles to learn high-frequency functions due to spectral bias [17, 18], finally resulting in inaccurate prediction results.

According to the above analysis, one approach to address the multi-scale problems is to mitigate the adverse effects caused by the large magnitude difference between the loss terms. Adjusting the weights of the loss terms is the simplest way to balance the magnitude difference, and much work has been done in this area. In Ref. [19], based on the dual-dimer method, Liu et al. present a new neural network with minimax architecture that can systematically adjust the weights of different losses. In Ref. [20], McClenney et al. give a Self-Adaptive Physics-informed Neural Networks (SA-PINN) using a soft attention mechanism, where different weights are set for each point of the residual and supervisory terms. In addition, to avoid the adverse effects caused by the magnitude difference, a possible way is the *hard constraint approach* [2, 18], which strictly enforces the initial and boundary conditions by using the neural network as part of a solution ansatz.

Another approach to improve the performance of the conventional PINN method for multi-scale problems is to replace deep neural networks with other neural network architectures. In Ref. [1], through the lens of neural tangent kernel theory, Wang et al. construct novel architectures that effectively tackle the problems with multi-scale behavior. In Ref. [15], the authors construct an improved fully-connected neural networks that could reduce the stiffness of the gradient flows, and provide more accurate results for some multi-scale problems. In Refs. [21] and [22], Jagtap et al. devise conservative PINN and extended PINN via domain decomposition, which are particularly effective for multi-scale problems. However, employing the appropriate architecture solely does not ensure accurate prediction for multi-scale problems [1].

In this paper, by constructing novel loss functions, we eliminate the adverse effects caused by the large difference in magnitude between the loss terms. Furthermore, by introducing and developing special neural network architectures, we find an effective way to address the multi-frequency problems. By combining the two approaches, we present a practical PINN framework which has the ability to well handle the multi-scale problems with multi-magnitude loss terms. The rest of this paper is organized as follows: In Section 2, the conventional PINN framework for general partial differential equations is introduced. In Section 3, we propose an improved PINN framework for multi-scale models, which is named as MMPINN. In Section 4, some numerical examples are discussed to demonstrate the performance of the new PINN framework. In the last section, we give a conclusion of our work.

2. Conventional PINN framework

The conventional PINN method presented in Ref. [3] utilizes deep neural networks to solve PDEs. The prediction function is obtained by optimizing the loss function through some optimization algorithms. The loss function consists of the supervised term and the residual term. The supervised term reflects the degree of approximation of the definite condition. The residual term, obtained by the automatic differentiation technique, reflects the degree of approximation of the governing equation. The PINN methods are applicable to a wide range of types of PDEs, including hydrodynamic equations, heat conduction equations, electromagnetism equations, and so on.

Suppose the PDE has the following general forms:

$$\begin{cases} \mathcal{P}(u) = f(x), & x \in \Omega \subset \mathbb{R}^d, \\ \mathcal{B}(u) = g(x), & x \in \partial\Omega, \end{cases} \quad (2.1)$$

where \mathcal{P} denotes the differential operator, $u = u(x)$ is the solution function on the d -dimensional domain Ω , $f(x)$ is the source term, $\mathcal{B}(u)$ represents the definite conditions, including the initial conditions (I.C.) and boundary conditions (B.C.) such as Dirichlet, Neumann, or Robin conditions. Note that, the time variable t can be considered a component of x in some cases [5].

The PINN method produces the prediction function, denoted by $u_\theta(x)$, as an approximation of the solution of Eq. (2.1), where θ is a set containing the weights and biases of the neural network, which are also the parameters we want to obtain by optimizing the loss function. For the conventional PINN method, the loss function consists of two parts as follows:

$$\mathcal{L}(\theta; \Sigma) = w_s \mathcal{L}_s(\theta; \tau_b) + w_r \mathcal{L}_r(\theta; \tau_r), \quad (2.2)$$

where w_s and w_r are the weights for the two parts of the loss function. In Eq. (2.2), the *supervised loss term* and the *residual loss term* are

$$\mathcal{L}_s(\theta; \tau_s) = \frac{1}{N_s} \sum_{i=1}^{N_s} |\mathcal{B}(u_\theta(x_i)) - g(x_i)|^2, \quad (2.3)$$

$$\mathcal{L}_r(\theta; \tau_r) = \frac{1}{N_r} \sum_{i=1}^{N_r} |\mathcal{P}(u_\theta(x_i)) - f(x_i)|^2. \quad (2.4)$$

$\Sigma = \{\tau_s, \tau_r\}$ denotes the set of training points, where $\tau_s = \{x_i|_{i=1}^{N_s}, x_i \in \partial\Omega\}$, and $\tau_r = \{x_i|_{i=1}^{N_r}, x_i \in \Omega\}$. N_s and N_r denote the number of boundary sampling points on $\partial\Omega$ and the number of inner sampling points in Ω , respectively.

Suppose

$$\bar{\theta} = \arg \min_{\theta} \mathcal{L}(\theta; \Sigma), \quad (2.5)$$

and then $u_{\bar{\theta}}(x)$ is the approximation of the unknown function u . $\bar{\theta}$ is obtained by using certain optimization algorithms to optimize the function $\mathcal{L}(\theta; \Sigma)$. The most commonly used optimization algorithms are Adam [23] and L-BFGS [24].

Since the heat equation is a class of equations with a very wide range of applications, the specific form of the loss function for this class of equations is given below as an example. The heat conduction problem can be formulated by the following equation:

$$\begin{cases} \frac{\partial u}{\partial t} - a^2 \Delta u = Q(x), & x \in \Omega \subset \mathbb{R}^d, t \in (0, T], \\ u(x, 0) = \phi(x), & x \in \Omega \subset \mathbb{R}^d, \\ u(x, t) = \psi(x, t), & x \in \partial\Omega, t \in (0, T]. \end{cases} \quad (2.6)$$

The loss function for Eq. (2.6) is defined as follows:

$$\mathcal{L}(\theta; \Sigma) = w_s \mathcal{L}_s(\theta; \tau_s) + w_r \mathcal{L}_r(\theta; \tau_r), \quad (2.7)$$

$$\mathcal{L}_s(\theta; \tau_s) = \frac{1}{N_0} \sum_{i=1}^{N_0} |u_\theta(x_i, 0) - \phi(x_i)|^2 + \frac{1}{N_b} \sum_{i=1}^{N_b} |u_\theta(x_i, t_i) - \psi(x_i, t_i)|^2, \quad (2.8)$$

$$\mathcal{L}_r(\theta; \tau_r) = \frac{1}{N_r} \sum_{i=1}^{N_r} \left| \frac{\partial u_\theta}{\partial t}(x_i, t_i) - a^2 \Delta u_\theta(x_i, t_i) - Q(x_i, t_i) \right|^2, \quad (2.9)$$

where N_0 is the number of sampling points for the I.C., N_b is the number of sampling points on the boundary $\partial\Omega$, and N_r is the number of residual points sampled in the computation domain Ω .

3. An improved PINN framework for multi-scale problems

In this section, we first make an analysis of the difficulties in solving the multi-scale problems by the conventional PINN method, and then discuss some strategies to overcome these difficulties. Finally, for the multi-scale problems containing multi-magnitude loss terms and multi-frequency features, we apply all these strategies to present an improved PINN framework, which we call the **MMPINN** framework.

3.1. Problems in solving multi-scale models by conventional PINN methods

Many problems in practical engineering applications have multi-scale features. For example, the radiation heat conduction problem in inertial confinement fusion is a typical multi-scale problem, where the temperature of the matter can vary from room temperature to millions or even tens of millions of degrees Celsius [25]. There are also some classical multi-scale problems in hydrodynamics, such as the fluid in a large river consisting of both low-frequency, large-amplitude waves and high-frequency, small-amplitude eddies.

For the model with multi-scale feature, the conventional PINN method has some issues in solving them. One of the most prominent issue is that, if there is a large order of magnitude difference between the loss terms of the loss function, a bad phenomenon like the following will occur in the optimization process:

If

$$\mathcal{L}_r(\theta; \tau_r) \gg \mathcal{L}_s(\theta; \tau_s),$$

and the optimization objective is

$$\mathcal{L}(\theta; \Sigma) = \mathcal{L}_s(\theta; \tau_s) + \mathcal{L}_r(\theta; \tau_r),$$

to make the value of $\mathcal{L}_r(\theta; \tau_r)$ decrease significantly, the value of $\mathcal{L}_s(\theta; \tau_s)$ will increase instead of decreasing at the beginning of the optimization process until their magnitudes are not particularly different. The risk associated with this phenomenon is that since the loss function $\mathcal{L}(\theta; \Sigma)$ is generally not convex, the optimization algorithm may fall into a local minimum, at which point the value of $\mathcal{L}_s(\theta; \tau_s)$ may still be large, indicating that the resulting network prediction function does not satisfy the initial boundary condition. We know that the functions that satisfy the equation can be a family of functions. If $u(x, t)$ satisfies the heat equation

$$u_t - u_{xx} = f(x, y),$$

then $u + ax + b$ all satisfy the equation. The initial boundary condition, also known as the definite condition, can be used to uniquely determine the function that satisfies the equation. Therefore, failure to satisfy the initial boundary condition means that the prediction obtained by the neural network is a function that deviates from the true solution.

In addition to the fact that $\mathcal{L}_r(\theta; \tau_r)$ and $\mathcal{L}_s(\theta; \tau_s)$ can have large magnitude differences, $\mathcal{L}_r(\theta; \tau_r)$ can also have large magnitude differences for different computational subdomains. In this case, a similarly bad situation can occur, where the optimization of subdomains with large magnitudes also comes at the expense of the optimization of subdomains with small magnitudes. The final prediction results are often better for subdomains with large magnitudes, while subdomains with small magnitudes have large relative errors or even distortions.

When optimizing an objective function that contains loss terms of different orders of magnitude, it is certainly possible to improve the optimization effect of the overall objective by choosing a powerful optimization algorithm or adjusting the parameters of the optimization algorithm, and further to make each loss term as small as possible, but then this is usually difficult to do. A more scientific approach is to design reasonable regularization methods to deal with each loss term, translate them to a certain range, so as to mitigate the difference in magnitude between loss terms. This enables the importance of each loss term to be balanced so that they are optimized simultaneously in the optimization process, which is the main motivation of this paper.

Another important issue of the conventional PINN method is that for the high-frequency or multi-frequency problems, the expressiveness of conventional deep neural networks is limited due to spectral bias [16]. To achieve better prediction results, it is necessary to use a neural network architecture adapted to the problem, such as the multi-scale Fourier feature network architecture in Ref. [1], the improved fully-connected neural network architecture proposed in Ref. [15] and the integrated neural network architecture developed in this paper.

3.2. A regularization strategy for multi-magnitude loss terms

The word “regularization” here means processing the values of the loss terms to approximately the same order of magnitude. In order to make each loss term of the loss function to be optimized synchronously in the optimization process, one can use the corresponding mechanism provided by the conventional PINN to reduce the difference in magnitude between different loss terms by adjusting their weights [19, 20, 26]. The weight selection is related to specific problems and has a wide range of values, so it is very difficult to give an exact selection criteria.

Consider the following formula:

$$\lim_{n \rightarrow \infty} a^{\frac{1}{n}} = 1, \quad \forall a \geq \epsilon > 0. \quad (3.1)$$

If $a = 10^9, b = 10^3$, the difference between them is 6 orders of magnitude. But $a^{\frac{1}{3}}$ has the same magnitude as b . Inspired by this, we can reduce the difference in magnitude between the loss terms by computing their $\frac{1}{n}$ power, for which we define the following loss function:

$$\tilde{\mathcal{L}}(\theta; \Sigma) = w_s \mathcal{L}_s^{\frac{1}{m}}(\theta; \tau_s) + w_r \mathcal{L}_r^{\frac{1}{n}}(\theta; \tau_r), \quad m > 0, n > 0, \quad (3.2)$$

where m and n are called regularization parameters. Clearly, if $\tilde{\mathcal{L}}(\theta; \Sigma)$ tends to 0, then the standard loss function $\mathcal{L}(\theta; \Sigma)$ in Eq. (2.2) must also tend to 0.

We call the method of obtaining the prediction function for multi-scale problems with **multi-magnitude** loss terms by optimizing Eq. (3.2) as the **MMPINN** method. The new loss function has the following characteristics:

- It is a generalization of the conventional PINN loss function, and it degenerates to the standard loss function when $m = 1, n = 1$.
- By simply adjusting m or n , one can regularize $\mathcal{L}_s(\theta; \tau_s)$ and $\mathcal{L}_r(\theta; \tau_r)$ to approximately the same order of magnitude, which allows the optimization algorithm to optimize both simultaneously rather than favoring one over the other.
- According to Eq. (3.1), the choice of m and n can be independent of the problem, and it is always possible to make $\mathcal{L}_s(\theta; \tau_s)$ and $\mathcal{L}_r(\theta; \tau_r)$ approximately the same magnitude as long as m and n are large enough.
- Depending on the actual order of magnitude of $\mathcal{L}_s(\theta; \tau_s)$ and $\mathcal{L}_r(\theta; \tau_r)$, one can suppress or accelerate the optimization of the corresponding loss terms to achieve the desired effect by adjusting m or n during the training process.

Remark 1. To reduce the order of magnitude of the loss terms, another natural idea is to use a logarithmic function for the loss terms, which is a common approach in the field of data processing. This approach can reduce the order of magnitude of the loss terms to 1 and obviates the need to tune hyper-parameters

such as m or n . However, the logarithmic approach has two problems: (1) The value range of the logarithmic function is $[-\infty, +\infty]$, which is inconsistent with the value range of the original loss function; (2) Even if we consider using the logarithmic function only for the loss terms greater than 1, because the logarithmic function excessively compresses the value of the loss function, the gradient of the corresponding loss term becomes extremely small, resulting in a case similar to the disappearance of the gradient, which causes the optimizer to obtain an invalid prediction result.

3.3. Multi-level training algorithms

Theoretically, if $\tilde{\mathcal{L}}(\theta; \Sigma) \rightarrow 0$, we can get $\mathcal{L}(\theta; \Sigma) \rightarrow 0$. Nevertheless, when the two loss functions are optimized by an optimization algorithm, the optimization process and the optimization results may be significantly different due to the different properties of the two functions.

When using Eq. (3.2) to handle loss terms of varying magnitudes, the small magnitude loss terms could be synchronously optimized. However, the large magnitude loss terms may not be fully optimized due to suppression. To address this problem, we propose the *multi-level training algorithm*, which ensures that the optimized loss function of the new algorithm exactly matches with that of the conventional PINN method. This theoretically ensures that the results of the two methods are completely equivalent.

To clearly describe the multi-level training strategy, we assume that $\mathcal{L}_r(\theta; \tau_r)$ is a loss term with large orders of magnitude and let $m = 1, n = 3$. For this case, we can obtain different levels of prediction by the following steps:

- *Level 1*

Take the random parameters θ_0 as the initial network parameters and use the combined Adam and L-BFGS optimizers to optimize the following loss function

$$\tilde{\mathcal{L}}_1(\theta; \Sigma) = w_s \mathcal{L}_s(\theta; \tau_s) + w_r \mathcal{L}_r^{\frac{1}{3}}(\theta; \tau_r), \quad (3.3)$$

then we get θ_1 and the first level of the prediction function u_{θ_1} .

- *Level 2*

Utilize the θ_1 as the initial guess of the neural network parameter and use the L-BFGS optimizer to optimize the following loss function

$$\tilde{\mathcal{L}}_2(\theta; \theta_1, \Sigma) = w_s \mathcal{L}_s(\theta; \theta_1, \tau_s) + w_r \mathcal{L}_r^{\frac{1}{2}}(\theta; \theta_1, \tau_r), \quad (3.4)$$

then we get θ_2 and the second level of the prediction function u_{θ_2} .

- *Level 3*

Utilize θ_2 to initialize network parameter and use the L-BFGS optimizer to optimize the following loss function

$$\tilde{\mathcal{L}}_3(\theta; \theta_2, \Sigma) = w_s \mathcal{L}_s(\theta; \theta_2, \tau_s) + w_r \mathcal{L}_r(\theta; \theta_2, \tau_r), \quad (3.5)$$

then we get θ_3 and the third level of the prediction function u_{θ_3} , which is also an approximation to the solution of Eq. (2.1).

The above multi-level algorithm is only designed to theoretically achieve the same results as the conventional PINN, and practical experience shows that in most cases the results of the first level prediction are sufficient. In addition, the algorithm can also jump directly from the first level to the third level without necessarily going through the second level.

For the multi-level algorithm, MMPINN can be considered as a pre-training strategy [27] of the conventional PINN method to obtain a set of good initial guess of neural network parameters for the optimizer. In most cases, this pre-training strategy is so effective that a few iterations are required to optimize the loss function of the conventional PINN using it as the initial guess of the network parameters.

3.4. Grouping regularization approach for different subdomains

Generally, the solution of the governing equation exhibits varying features across different locations in the computational domain. Significant variations in the solution's gradient across different locations of the computational domain may lead to widely varying residual magnitudes on different subdomains. In this case, combining them together for optimization according to Eq. (2.4) necessarily results in the subdomains with smaller residual magnitude not being optimized synchronously. To solve this problem, we present the *grouping regularization approach*.

Taking the evolution equation as an example, assuming that the property of the solution of the equation has a large difference in the time direction, we divide the whole time domain $[0, T]$ into k time-subdomains, i.e,

$$[0, T] = [0, T_1] \cup (T_1, T_2) \cup \dots \cup (T_{k-1}, T_k).$$

Figure 3.1 is a schematic diagram of the division of the entire domain into several subdomains along the temporal axis.

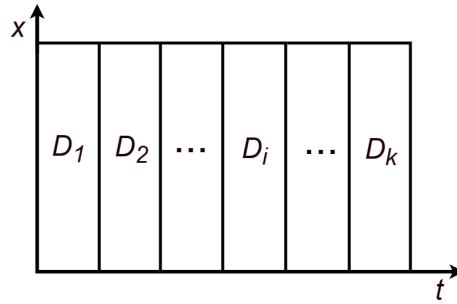


Figure 3.1: Grouping along the time direction.

Based on this grouping way, the loss function is constructed as follows:

$$\begin{cases} \mathcal{L}(\theta; \Sigma) = w_s \mathcal{L}_s(\theta; \tau_s) + \mathcal{L}_r(\theta; \tau_r), \\ \mathcal{L}_s(\theta; \tau_s) = \frac{1}{N_s} \sum_{i=1}^{N_s} |\mathcal{B}(u_\theta(x_i)) - g(x_i)|^2, \\ \mathcal{L}_r(\theta; \tau_r) = \sum_{h=1}^k \omega_{r,h} L_{r,h}(\theta; \tau_{r,h}), \\ L_{r,h}(\theta; \tau_{r,h}) = \frac{1}{N_{r,h}} \sum_{i=1}^{N_{r,h}} |\mathcal{P}(u_\theta(x_{h,i})) - f(x_{h,i})|^2, \end{cases} \quad (3.6)$$

where $x_{h,i} = (x_i, t_{h,i})$ represents a sample point in the subdomain D_h , $\tau_{r,h}$ is the set consisting of these points, and $L_{r,h}(\theta; \tau_{r,h})$ is the residual term defined on the subdomain D_h .

In the same way as in Section 3.2, we set a regularization parameter for each subdomain, e.g., assign n_h to the subdomain D_h :

$$\mathcal{L}_r(\theta; \tau_r) = \sum_{h=1}^k \omega_{r,h} [L_{r,h}(\theta; \tau_{r,h})]^{\frac{1}{n_h}}. \quad (3.7)$$

By reasonably setting the parameters n_h , we can focus more on some specific subdomain to get better prediction results when training the neural network.

Note that, the above is a grouping strategy for the whole computational domain along the temporal axis, of course it is also possible to divide the whole domain into several subdomains by spatial locations. The number of subdomains and the way they are grouped depends on the specific problem.

3.5. An integrated neural network architecture

Network architecture is important for the successful implementation of PINN methods in scientific computing [15]. As discussed in Section 3.1, the conventional deep neural networks have relatively weak expressiveness for multi-frequency problems due to spectral bias. Therefore, it is necessary to study special network architecture to handle the typical multi-scale problem with multi-frequency, especially high-frequency features.

Through in-depth investigation, we find that the Fourier feature neural networks in Ref. [1] and the improved fully-connected neural networks in Ref. [15] have their own merits in dealing with multi-scale problems. However, for multi-frequency problems with multi-scale loss terms, using either alone does not yield ideal results. To solve arbitrary types of multi-scale problems, we develop an integrated network architecture by combining these two network architectures, which takes advantage of the characteristics of both and can deal relatively well with difficult multi-scale problems.

The design idea of the integrated neural network architecture is as follows: First, we apply multiple Fourier feature embeddings to the input coordinates. Second, we construct two transform networks for each Fourier feature. Third, we pass transform networks to the fully-connected neural networks. Finally, we concatenate the outputs with a linear layer. Figure 3.2 shows the schematic of the integrated neural network architecture, which we will formulate in detail next.

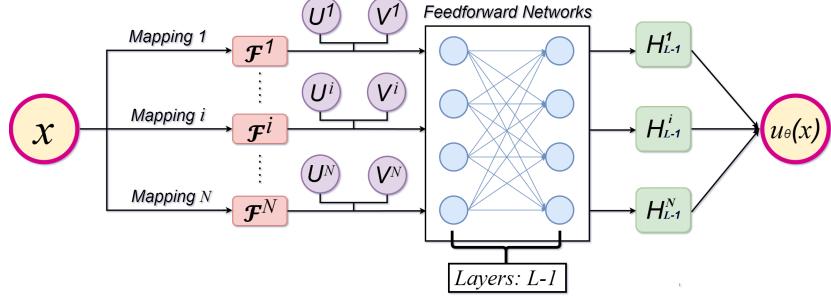


Figure 3.2: INN network architecture.

The Fourier mapping \mathcal{F} is defined as

$$\mathcal{F}^i(x) = \begin{bmatrix} \sin(B^{(i)}x) \\ \cos(B^{(i)}x) \end{bmatrix}, \quad i = 1, 2, \dots, N, \quad (3.8)$$

where N represents the number of Fourier mapping for the input coordinates x . Each entry of $B^{(i)}$ is sampled from the Gaussian distribution $\mathcal{N}(0, (\sigma_i)^2)$ and is held fixed during model training. When $\sigma = 1$, low frequencies are learned initially, which is similar to the behavior of deep neural networks. If σ is larger, high frequencies are learned first, which may result in over-fitting [1]. Therefore, selecting the appropriate σ is crucial. The effectiveness of the Fourier feature embeddings in mitigating the spectral bias is illustrated in Ref. [1]. Using this technique can improve the accuracy of solving multi-frequency problems.

The two transform networks are defined as

$$\begin{cases} U^i = \phi(W_u^i \cdot \mathcal{F}^i(x) + b_u^i), \\ V^i = \phi(W_v^i \cdot \mathcal{F}^i(x) + b_v^i), \end{cases} \quad i = 1, 2, \dots, N, \quad (3.9)$$

where ϕ denotes a nonlinear activation function and $(W_u^i, b_u^i, W_v^i, b_v^i)$ are the parameters to be optimized. The transform networks based on the Fourier mapping, formulated by Eq. (3.9), are different from those in Ref. [15]. They are the improved version by replacing the input variable x with $\mathcal{F}^i(x)$. Introducing the two networks into conventional fully-connected neural network architecture could enhance the hidden states with residual connections to alleviate the stiffness of the gradient flow [15].

The subsequent detailed forward pass is

$$\begin{cases} Z_1^i = \phi(W_1 \cdot \mathcal{F}^i(x) + b_1), \\ H_1^i = (1 - Z_1^i) \odot U^i + Z_1^i \odot V^i, \\ Z_l^i = \phi(W_l \cdot H_{l-1}^i + b_l), \\ H_l^i = (1 - Z_l^i) \odot U^i + Z_l^i \odot V^i, \end{cases} \quad i = 1, 2, \dots, N, \quad l = 2, \dots, L-1, \quad (3.10)$$

where \odot represents the point-wise multiplication and L is the number of hidden layers of the neural network. The final output is

$$u_\theta(x) = W_L \cdot [H_{L-1}^1, \dots, H_{L-1}^N] + b_L. \quad (3.11)$$

The weights W_l and the biases b_l in Eq. (3.10) and Eq. (3.11) are also the parameters of the neural network to be obtained by optimizing the loss function.

In Ref. [1], the authors also propose a **spatio-temporal** multi-scale Fourier mapping, which is similar to the Fourier mapping (3.8) and can also be combined with the improved fully-connected neural networks in Ref. [15] to form a new network architecture. For simplicity of description, we refer to the (spatio-temporal) Multi-scale Fourier Feature neural network architectures in Ref. [1] as MFF and the corresponding Integrated Neural Network architectures as INN in this paper.

By combining MMPINN with different network architectures, such as the commonly used deep neural networks, MFF and INN, we obtain several new PINN methods. Table 3.1 lists the names of the new methods, and their performance is examined in Section 4.

Table 3.1: MMPINN with different network architectures.

Neural Network Architecture	Method
common Deep Neural Network	MMPINN-DNN
Multi-scale Fourier Feature Network	MMPINN-MFF
Integrated Neural Network	MMPINN-INN

Numerical experiments show that MMPINN-INN obtains highly accurate computational results for difficult multi-scale problems with high-frequency features, which has obvious advantages over MMPINN-DNN and MMPINN-MFF. However, INN requires large computational resources, and improving its computational efficiency is one of our next works.

3.6. MMPINN framework

By combining all of the above strategies, we construct an improved PINN framework called the MMPINN framework. Using this framework, it is possible to design deep learning methods that obtain highly accurate predictions for a large class of multi-scale problems with multi-magnitude loss terms and multi-frequency features. Figure 3.3 is the diagram of the MMPINN framework.

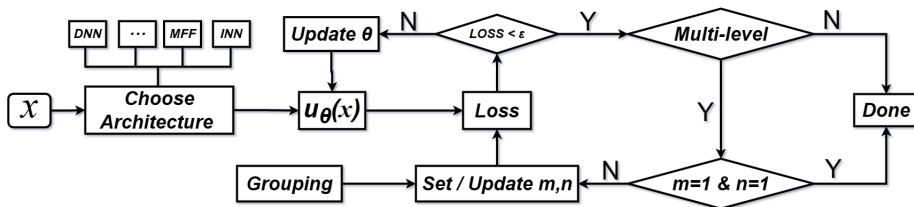


Figure 3.3: Diagram of the MMPINN framework.

4. Numerical Examples

In this section, several numerical experiments are performed to validate the performance of our method. In Section 4.1, we show the robustness of the MMPINN-DNN method by solving a large gradient heat conduction problem. The examples in Section 4.2, Section 4.3 and Section 4.4 are taken from the recent literature, and we compare our method with the state-of-the-art methods. In Section 4.5, we further investigate the performance of our method on a difficult multi-frequency model. In Section 4.6, the grouping regularization strategy is used to solve a problem with dramatic variations in different subdomains, demonstrating its effectiveness.

The L_2 relative error norm is used to measure the accuracy of the prediction, which is defined as follows:

$$\|\epsilon\|_2 = \frac{\sqrt{\sum_{i=1}^N |u_\theta(x_i) - u(x_i)|^2}}{\sqrt{\sum_{i=1}^N |u(x_i)|^2}}, \quad (4.1)$$

where $u(x_i)$ is the real solution or the reference solution, and $u_\theta(x_i)$ is the neural network prediction for a point in the test set $\{x_i |_{i=1}^N \in \Omega\}$. N is the size of test set.

We use the deep learning framework TensorFlow (version 1.13.1 or 2.11.0) to implement all experiments. The data type is `float32` and the activation function is `tanh` for all examples. We combine Adam and L-BFGS to optimize the loss function, without any special statement, first running 2000 Adam iterations and then switching to L-BFGS iterations until convergence. For our method, 2000 Adam iterations are sufficient to obtain available predictions; however, for the conventional PINN method, this setting often fails to produce available predictions, and thus it is necessary to run 10^6 Adam iterations and then switch to L-BFGS until convergence, which will be referred to as conventional PINN (10^6 Adam) in the following sections. All parameters and termination criteria of the L-BFGS optimizer are considered as suggested in Refs. [20, 24]. Before training, the parameters of the neural network are randomly initialized by using the Xavier scheme [28].

4.1. A heat conduction problem with large gradients

In practical applications such as inertial confinement fusion [11], it is often necessary to solve the heat conduction problem with a strong heat source that suddenly appears at a certain time, and this example is designed to test the effectiveness of our method for this type of problem.

Consider the following heat equation

$$\begin{cases} u_t = u_{xx} + f(x, t), & x \in (-1, 1), t \in (0, 1], \\ u(x, 0) = (1 - x^2)e^{\frac{1}{1+x}}, & x \in (-1, 1), \\ u(\pm 1, t) = 0, & t \in (0, 1]. \end{cases} \quad (4.2)$$

Its exact solution is given by

$$u(x, t) = (1 - x^2)e^{\frac{1}{(2t-1)^2+\varepsilon}}, \quad (4.3)$$

where ε is a positive constant, and $f(x, t)$ is derived from the exact solution.

For this model, if ε is large, the variation of u will be smooth over the entire computational domain; if ε is very small, the value of u will change sharply near $t = 0.5$, embodying typical multi-scale characteristics. Since the boundary value of this model is always 0 and its initial value is less than e , the value of its supervised loss term is always small, while the value of its residual term increases significantly as ε decreases. For different ε , the ratio of initial losses of the two terms from practical computation is as follows:

$$\begin{aligned} \varepsilon = 0.3, \quad \mathcal{L}_s : \mathcal{L}_r &= 1 : 10^3, \\ \varepsilon = 0.15, \quad \mathcal{L}_s : \mathcal{L}_r &= 1 : 10^6. \end{aligned}$$

As we know, the conventional PINN method can usually achieve good predictions for smooth problems, while for the problem with large difference in magnitude between the supervised term and the residual term, it is difficult to obtain valid predictions without fine-tuning hyperparameters.

We choose the MMPINN-DNN method with $m = 1, n = 3$ to solve this model for different ε , and compare its result with those of conventional PINN methods and the SA-PINN method in Ref. [20]. Table 4.1 lists the L_2 relative errors of various methods. *Using 2000 Adam iterations + L-BFGS*, the conventional PINN can give an accurate prediction for $\varepsilon = 0.3$, but can no longer give a correct prediction for $\varepsilon = 0.15$, while the MMPINN-DNN method can still give an accurate prediction. To obtain an available prediction with the conventional PINN method, we have to let it run 10^6 Adam iterations.

Table 4.1: The L_2 relative errors of different methods for solving Eq. (4.2).

ε	Method	$\ \epsilon\ _2$
0.3	conventional PINN	$3.41 \pm 1.71 \times 10^{-4}$
	MMPINN-DNN (this work)	$1.86 \pm 0.61 \times 10^{-4}$
0.15	conventional PINN	$1.07 \pm 0.11 \times 10^0$
	Hard Constraints [2] PINN	$1.10 \pm 0.12 \times 10^0$
	conventional PINN (10^6 Adam)	$5.26 \pm 3.43 \times 10^{-3}$
	SA-PINN [20]	1.76×10^{-3}
	MMPINN-DNN (this work)	$5.01 \pm 1.52 \times 10^{-4}$

Figure 4.1 shows the exact solution of Eq. (4.2) with $\varepsilon = 0.15$ and the point-wise absolute errors of the SA-PINN [20] and MMPINN-DNN methods. According to Table 4.1 and Figure 4.1, we can see that neither adjusting the weights [20] nor increasing the number of optimizations could yield better results than the MMPINN-DNN method. Figure 4.2 shows the variation curves of different loss terms during the optimization process (2000 Adam iterations + L-BFGS). The boundary condition loss and the initial condition loss of the conventional PINN method increase instead of decreasing during the optimization

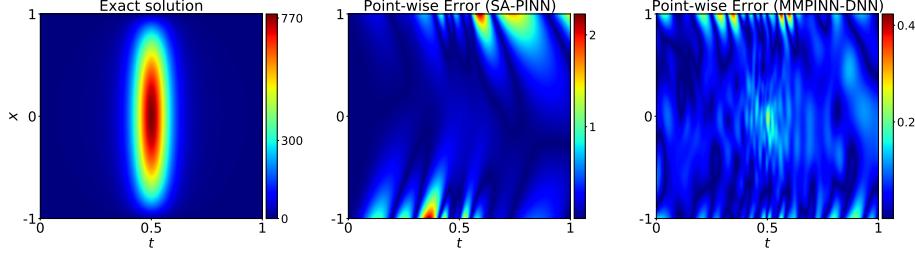


Figure 4.1: The exact solution of Eq. (4.2) with $\varepsilon = 0.15$ and the point-wise absolute errors of the SA-PINN [20] and MMPINN-DNN methods.

process, which ultimately leads to training failure; the MMPINN-DNN method can simultaneously reduce the boundary loss, initial condition loss and residual loss, which ensures accurate prediction results.

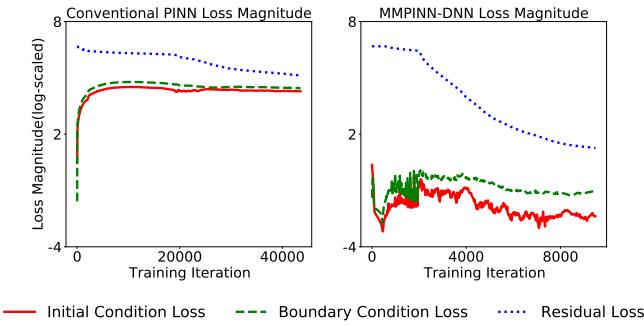


Figure 4.2: Loss curves of the conventional PINN method and the MMPINN-DNN method for solving Eq. (4.2) with $\varepsilon = 0.15$.

To investigate the effect of the regularization parameters on the computation results, we test the cases of $n = 1, \dots, 8$. Meanwhile, to avoid randomness, each set of tests is performed five times independently. The test results are shown in Figure 4.3. We see that the parameter n cannot be chosen too large; large n will excessively inhibit the residual term to be reasonably optimized, thus failing to obtain valid predictions satisfying the equation.

We further test the performance of the MMPINN-DNN method in solving Eq. (4.2) with smaller ε , and the results of 5 independently repeated experiments are shown in Table 4.2. In the cases where the multi-scale property is more significant, the MMPINN-DNN method can still give accurate predictions, but the conventional PINN methods cannot give usable predictions. Figure 4.4 shows the result of the MMPINN-DNN method for the case $\varepsilon = 0.11$, where the ratio of the supervised loss term to the residual loss term differs by 8 orders of magnitude, i.e. $\mathcal{L}_s : \mathcal{L}_r = 1 : 10^8$. For this model, the MMPINN-DNN method still gives a very accurate prediction. In the case where the maximum value of u is close to 9000, the maximum prediction error is less than 0.7.

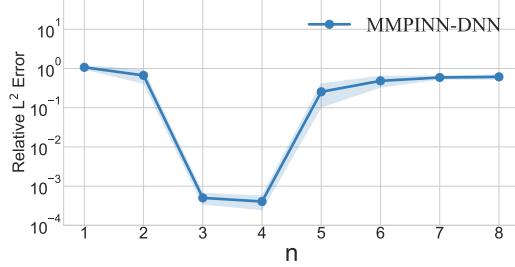


Figure 4.3: The relative L_2 error using the MMPINN-DNN method with $m = 1$ and $n = 1, \dots, 8$ for solving Eq. (4.2) with $\varepsilon = 0.15$. The line and the shaded region represent the mean and the standard deviation of 5 independent experiments, respectively.

Table 4.2: L_2 errors of the MMPINN-DNN method for Eq. (4.2) with different ε .

ε	$\ \epsilon\ _2$
0.14	$1.43 \pm 0.39 \times 10^{-4}$
0.13	$3.03 \pm 0.38 \times 10^{-5}$
0.12	$2.34 \pm 0.54 \times 10^{-5}$
0.11	$2.06 \pm 1.28 \times 10^{-5}$

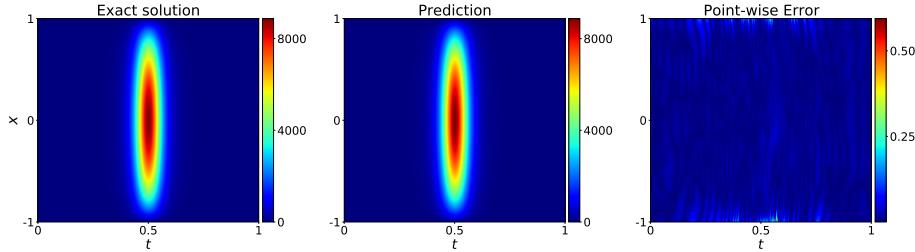


Figure 4.4: The exact solution of Eq. (4.2) ($\varepsilon = 0.11$) and the prediction and the point-wise absolute error of the MMPINN-DNN method.

Table 4.3: Hyperparameters used for the different ε in Sec. 4.1.

ε	DNN Layers	DNN Neurons	N_0	N_b	N_r
0.3	4	20	700	1400	3000
0.15	4	50	1200	2400	10000
0.14	4	100	1600	3200	20000
0.13	4	320	2000	4000	60000
0.12	4	400	3000	6000	90000
0.11	4	400	4000	8000	100000

As the complexity of the problem increases, we need to increase the number of training points and the size of the neural network. Table 4.3 shows the different hyperparameters, where N_0 , N_b and N_r are the number of initial, boundary and residual points, respectively.

4.2. A typical multi-scale problem

The example in this subsection is taken from Ref. [1]. This is a classical and pedagogical model. Although this model is very simple in form, its solution exhibits low frequency in the macro-scale and high frequency in the micro-scale, much like many practical scenarios [1].

Consider the 1D Poisson equation as follows:

$$\begin{cases} \Delta u(x) = f(x), & x \in (0, 1), \\ u(0) = u(1) = 0, \end{cases} \quad (4.4)$$

where

$$f(x) = -16\pi^2 \sin(4\pi x) - 2250\pi^2 \sin(150\pi x). \quad (4.5)$$

The exact solution for this model is

$$u(x) = \sin(4\pi x) + 0.1 \sin(150\pi x). \quad (4.6)$$

Since this is a frequency-dependent multi-scale problem, we use the MPPINN-MFF method for comparison with the method in [1]. According to the ratio of the loss terms $\mathcal{L}_s : \mathcal{L}_r = 10^1 : 10^8$, the regularization parameters are set with $m = 3$ and $n = 3$. The architecture uses 3 hidden layers with 100 neurons per hidden layer. The Fourier feature mappings are initialized with $\sigma_1 = 1$ and $\sigma_2 = 25$, respectively. Other hyperparameters are chosen as in Ref. [1], including the exponential decay of the learning rate, the choice of the optimizer, etc., except for the batch sizes ($N_b = N_r = 512$).

Table 4.4: The L_2 relative errors of different methods for solving Eq. (4.4).

Method	$\ \epsilon\ _2$
conventional PINN	$3.24 \pm 2.13 \times 10^0$
conventional PINN (10^6 Adam)	$4.25 \pm 1.92 \times 10^0$
MFF [1]	$1.41 \pm 1.43 \times 10^{-2}$
Hard Constraints [2]+MFF	$5.35 \pm 2.62 \times 10^{-3}$
MMPINN-MFF (this work)	$6.94 \pm 1.37 \times 10^{-4}$

Table 4.4 lists the results of several PINN methods. From this table, we conclude that the deep neural networks cannot adequately learn the multi-frequency function even after 10^6 iterations [1]; Moreover, using the MFF alone does not guarantee strong robustness, its relative L_2 errors ranging from 0.41% to 4.24% in 5 independently repeated experiments. Our method simultaneously captures the low-frequency components and the high-frequency oscillations of the solution more accurately than the MFF method, as shown in Figure 4.5.

Note that the exact solution for Eq. (4.4) in Ref. [1] is given by

$$u(x) = \sin(2\pi x) + 0.1 \sin(50\pi x). \quad (4.7)$$

Compared to Eq. (4.6), the multi-scale feature of Eq. (4.7) is relatively small, and the the ratio of the initial loss terms is $\mathcal{L}_s : \mathcal{L}_r = 10^1 : 10^6$. For this model, MFF is available and its relative L_2 error is 1.36×10^{-3} as reported in Ref. [1], while the error of our method is $6.43 \pm 2.52 \times 10^{-4}$ under the same parameters.

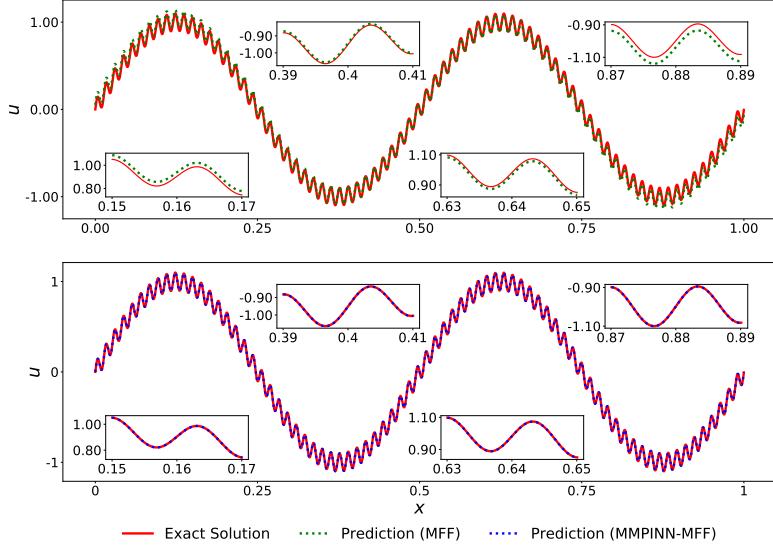


Figure 4.5: The exact solution and predictions for Eq. (4.4). The top figure is the result of the MFF method and the bottom figure is the result of the MMPINN-MFF method.

4.3. Helmholtz equation with high-frequency

The Helmholtz equation is an elliptic type partial differential equation describing electromagnetic waves. It has been used in many applications [29]. In Refs. [1, 15, 20], the authors all take this equation as an example to test the performance of their approaches. The model they consider is the following:

$$\begin{cases} \Delta u(x, y) + k^2 u(x, y) = q(x, y), & (x, y) \in \Omega = (-1, 1) \times (-1, 1), \\ u(x, y) = h(x, y), & (x, y) \in \partial\Omega. \end{cases} \quad (4.8)$$

The exact solution is assumed to be

$$u(x, y) = \sin(a_1 \pi x) \sin(a_2 \pi y), \quad (4.9)$$

$q(x, y)$ and $h(x, y)$ are derived from it.

Making accurate predictions for this mode using the conventional PINN method is challenging [15]. Due to a large initial difference in magnitude order between the residual term and the supervised term, the minimization of the residual term dominates the total training process.

Taking $k = 1$, $a_1 = 1$ and $a_2 = 8$, we use this model to test our methods, including MMPINN-MFF and MMPINN-INN. The network architecture contains 4 hidden layers, each layer with 50 neurons. Two separate Fourier feature mappings are initialized with $\sigma_1 = 1$ and $\sigma_2 = 10$, corresponding to x and y , respectively. The numbers of training points are set with $N_0 = 1200$, $N_b = 1200$ and $N_r = 10000$. The regularization parameters are set with $m = 3$ and $n = 3$, which are determined by the ratio ($\mathcal{L}_s : \mathcal{L}_r = 10^{-1} : 10^5$).

Table 4.5 lists the results of different methods. Compared to the approaches in the recent literature, our approaches have better performance. The MMPINN-INN method gives the best prediction among them, and its absolute error is shown in Figure 4.6. Figure 4.7 shows the comparison of the exact solution and the prediction using the MMPINN-INN method at different positions. The MMPINN-INN method captures the high frequency components of the solution very well.

Table 4.5: The L_2 relative errors of different methods for solving Eq. (4.8).

Method	$\ \epsilon\ _2$
LRA with IFNN [15]	$2.13 \pm 0.68 \times 10^{-2}$
SA-PINN [20]	$1.27 \pm 1.11 \times 10^{-2}$
Hard Constraints [2] PINN	$7.03 \pm 6.27 \times 10^{-3}$
MFF [1]	$3.46 \pm 0.51 \times 10^{-3}$
MMPINN-MFF (this work)	$6.56 \pm 4.17 \times 10^{-4}$
MMPINN-INN (this work)	$2.13 \pm 0.43 \times 10^{-4}$

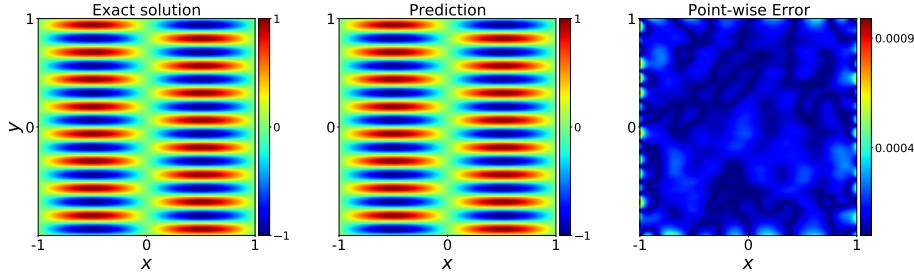


Figure 4.6: The exact solution of Eq. (4.8) and the prediction and the point-wise absolute error of the MMPINN-INN method.

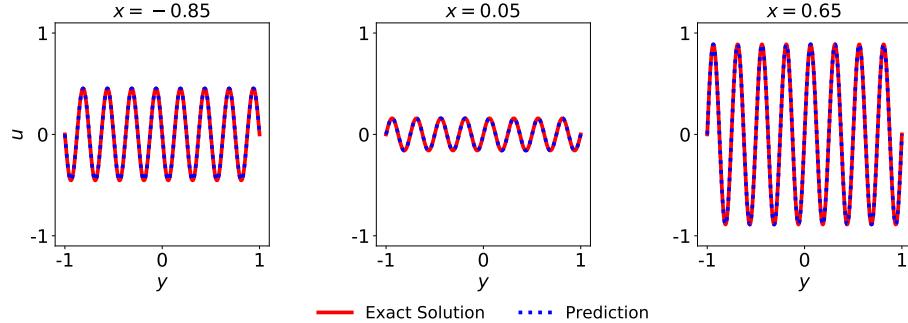


Figure 4.7: Comparison of the exact solution and the prediction using MMPINN-INN.

4.4. Klein-Gordon equation

The Klein-Gordon equation is a relativistic counterpart of the Schrödinger equation, and has received considerable attention in its numerical and analytical aspects [30]. The example in this subsection is taken from Ref. [15].

Consider the 1D Klein-Gordon equation as follows:

$$\begin{cases} u_{tt} + \alpha u_{xx} + \beta u + \gamma u^k = f(x, t), & (x, t) \in \Omega \times (0, T), \\ u(x, 0) = x, & x \in \Omega, \\ u_t(x, 0) = 0, & x \in \Omega, \\ u(x, t) = h(x, t), & (x, t) \in \partial\Omega \times [0, T]. \end{cases} \quad (4.10)$$

The model parameters are set with $\Omega = [0, 1]$, $T = 1$, $\alpha = -1$, $\beta = 0$, $\gamma = 1$, $k = 3$. We use the fabricated solution

$$u(x, t) = x \cos(a\pi t) + (xt)^3 \quad (4.11)$$

to assess the accuracy of the PINN methods. $f(x, t)$ and $h(x, t)$ are derived from Eq. (4.11).

When the parameter a in Eq. (4.11) is large, the conventional PINN method fails to produce an accurate prediction for this model. Wang et al. attribute the failure to the imbalanced gradient pathology. They separately propose a Learning Rate Annealing(LRA) algorithm and an Improved Fully-connected Neural Network (IFNN), and combine them to improve the prediction accuracy [15].

With $a = 5$ and $a = 10$, we test the performance of different PINN methods. The network architecture we use in this example contains 5 hidden layers, each layer with 150 neurons. Two separate Fourier feature mappings are initialized with $\sigma_1 = 5$ and $\sigma_2 = 1$, corresponding to t and x , respectively. The regularization parameters are set with $m = 3$ and $n = 3$. The batch sizes are set to $N_b = N_r = 256$. In order to make a fair comparison with the methods proposed in Ref. [15], the other hyperparameters are set according to the specifications of the article [15]. The training of the neural networks uses the Adam optimizer to perform 40,000 iterations. The multi-level training strategy presented in section 3.3 is not implemented in this example.

Table 4.6: The L_2 relative errors of different methods for solving Eq. (4.10).

Method	$\ \epsilon\ _2$	
	$a = 5$	$a = 10$
Conventional PINN	$2.81 \pm 0.76 \times 10^{-2}$	$2.56 \pm 0.57 \times 10^{-1}$
LRA with IFNN [15]	$1.42 \pm 0.61 \times 10^{-3}$	$9.55 \pm 3.73 \times 10^{-3}$
MMPINN-MFF (this work)	$8.80 \pm 2.03 \times 10^{-4}$	$2.17 \pm 0.57 \times 10^{-3}$
MMPINN-INN (this work)	$2.47 \pm 0.50 \times 10^{-4}$	$6.55 \pm 1.18 \times 10^{-4}$

Table 4.6 shows that as the frequency increases, the accuracy of the method in Ref. [15] obviously decreases, while the MMPINN-INN method remains very good. Figure 4.8 provides a detailed comparison of the prediction accuracy, which also proves that the MMPINN-INN method is superior. The integrated neural network(INN) architecture plays a crucial role in improving the accuracy.

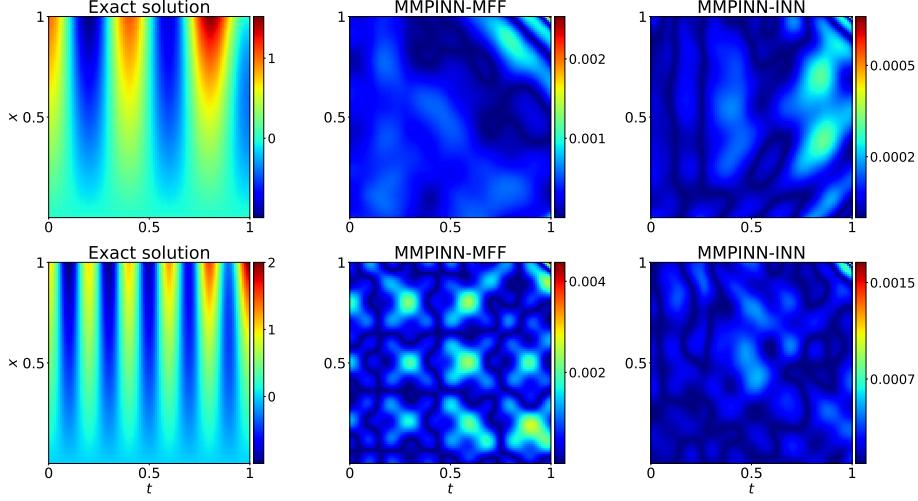


Figure 4.8: The exact solution of Eq. (4.10) and the point-wise absolute errors of the MMPINN-MFF and MMPINN-INN methods (Line 1: $a = 5$, Line 2: $a = 10$).

4.5. A multi-frequency model

Both traditional numerical methods and deep learning methods are challenged by multi-frequency problems where the frequency varies significantly with position x or time t . This example is designed to test the ability of our approaches to handle models with varying frequencies. Consider a heat equation as follows:

$$\begin{cases} u_t = u_{xx} + f(x, t), & x \in (-1, 1), t \in (0, 1], \\ u(x, 0) = 0, & x \in (-1, 1), \\ u(\pm 1, t) = \sin(2\pi t), & t \in (0, 1]. \end{cases} \quad (4.12)$$

The exact solution for this model is given by

$$u(x, t) = \sin \frac{20\pi t}{1 + 9x^2}, \quad (4.13)$$

$f(x, t)$ is derived from the exact solution.

Figure 4.9 shows the shape of the exact solution. The solution has a different frequency at different x . In addition, this solution has a high frequency feature at the position $x = 0$, which results in a large difference in order of magnitude between the supervised term and the residual term. These two factors lead to the failure of training this model with the conventional PINN method. To obtain a valid prediction for this model, we use three improved PINN methods to train it. A separate Fourier feature mapping is applied and initialized with $\sigma = 10$. The network architecture we use contains 4 hidden layers, each layer with 300 neurons. The numbers of training points are set with $N_0 = 1200$, $N_b = 1200$ and $N_r = 120000$. The regularization parameters are set with $m = 3$, $n = 3$.

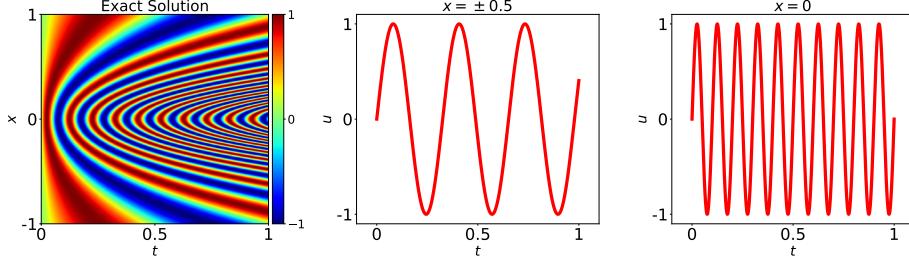


Figure 4.9: The exact solution of Eq. (4.12).

Table 4.7: The L_2 relative errors of different methods for solving Eq. (4.12).

Method	$\ \epsilon\ _2$
MFF [1]	$1.03 \pm 0.05 \times 10^{-2}$
Hard Constraints [2] + MFF	$3.13 \pm 0.77 \times 10^{-3}$
MMPINN-MFF (this work)	$7.95 \pm 1.41 \times 10^{-4}$
MMPINN-INN (this work)	$5.60 \pm 1.79 \times 10^{-4}$

Table 4.7 and Figure 4.10 show the results of the different methods. As can be seen from Figure 4.10, the MFF method predicts the boundary conditions inaccurately, with large errors concentrated near the boundary, due to the lack of measures to reduce the difference in magnitude between the loss terms. Our methods are much more accurate than the MFF method, further demonstrating the perfection and necessity of combining MMPINN with INN.

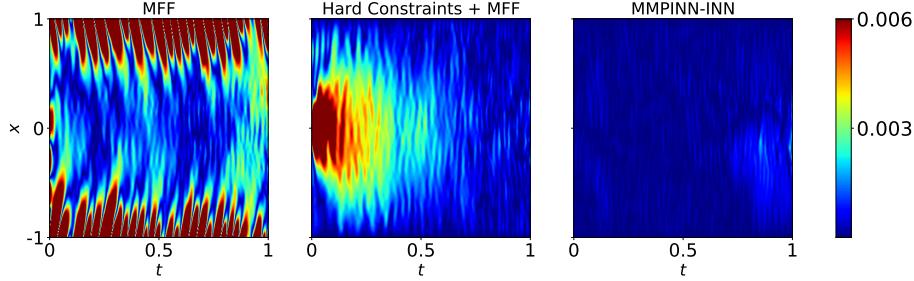


Figure 4.10: The point-wise absolute errors of the MFF, Hard Constraints + MFF and MMPINN-INN methods for solving Eq. (4.12).

4.6. A problem with dramatic variations in different subdomains

This example is designed to illustrate the importance of grouping regularization for the model whose residuals have large differences in different subdomains. Consider a Possion problem as follows:

$$\begin{cases} \Delta u(x, y) = f(x, y), & (x, y) \in \Omega = (0, 1) \times (0, 1), \\ \mathcal{B}(x, y) = g(x, y), & (x, y) \in \partial\Omega. \end{cases} \quad (4.14)$$

Suppose the exact solution is given by

$$u(x, y) = 1 + (1000 + y^2)e^{-\frac{(x-\frac{1}{2})^2}{2h^2}}, \quad (4.15)$$

where $h = 0.02$. $f(x, y)$ and $g(x, y)$ are derived from the exact solution.

As shown in Figure 4.11, the solution can be seen as a normal distribution function along the x -axis whose value reaches a maximum close to 1000 at $x=0.5$ and then decreases rapidly to a value close to 1 in a narrow interval. Compared to other examples, this model has more obvious differences in different subdomains. Therefore, to obtain an accurate prediction for this model, we have to use the grouping regularization strategy discussed in Section 3.4. Based on the characteristics of the solution, we divide the whole domain Ω into three subdomains, as shown in Figure 4.12. We refer to the MMPINN-DNN method using the grouping regularization strategy as *MMPINN-DNN-GRS* in this benchmark.

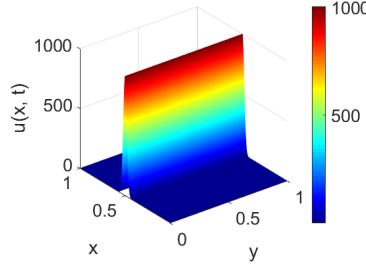


Figure 4.11: The exact solution of Eq. (4.14).

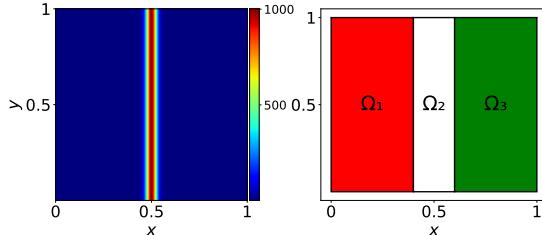


Figure 4.12: Grouping regularization mode.

The network architecture we use in this example contains 4 hidden layers, each layer with 50 neurons. The numbers of training points are set with $N_b = 4800$ and $N_r = 30000$. Note that for MMPINN-DNN-GRS, we set 10000 points in each subdomain. The regularization parameters are set with $m = 1$ and $n = 3$ for the MMPINN-DNN method and $m = 1$ and $n = (2, 4, 2)$ for the MMPINN-DNN-GRS method, which are determined by the ratios of the initial loss terms $\mathcal{L}_s : \mathcal{L}_r = 10^4 : 10^{11}$ and $\mathcal{L}_{\Omega_1} : \mathcal{L}_{\Omega_2} : \mathcal{L}_{\Omega_3} = 10^2 : 10^{11} : 10^2$.

Table 4.8 lists the results of the different methods. We see that it is difficult to obtain satisfactory predictions using conventional PINN methods. There are two factors that cause the PINN methods to perform poorly. The first factor is the large difference between the values of the supervised term and the residual term, which differ by about seven orders of magnitude. The second factor is that the distribution of the residual term is extremely uneven, with the first and third parts differing by about nine orders of magnitude from the second part.

Table 4.8: The L_2 relative errors of different methods for solving Eq. (4.14).

Method	$\ \epsilon\ _2$
conventional PINN	$1.12 \pm 0.10 \times 10^0$
conventional PINN (10^6 Adam)	$4.51 \pm 2.81 \times 10^{-2}$
MMPINN-DNN (this work)	$4.14 \pm 3.29 \times 10^{-3}$
MMPINN-DNN-GRS (this work)	$7.53 \pm 1.64 \times 10^{-4}$

The differences between the supervised term and the residual term could be balanced by MMPINN-DNN and Table 4.8 also shows that MMPINN-DNN can give more accurate prediction than PINNs. However, because MMPINN-DNN does not address the issue of large differences between subdomains, so it cannot give satisfactory predictions for the subdomains with small residual terms, leading to large relative errors for $x = 0.2$ in Ω_1 and $x = 0.8$ in Ω_3 , as shown in Figure 4.13. The MMPINN-DNN-GRS method gives an accurate prediction for the entire domain, demonstrating the importance of the grouping regularization strategy.

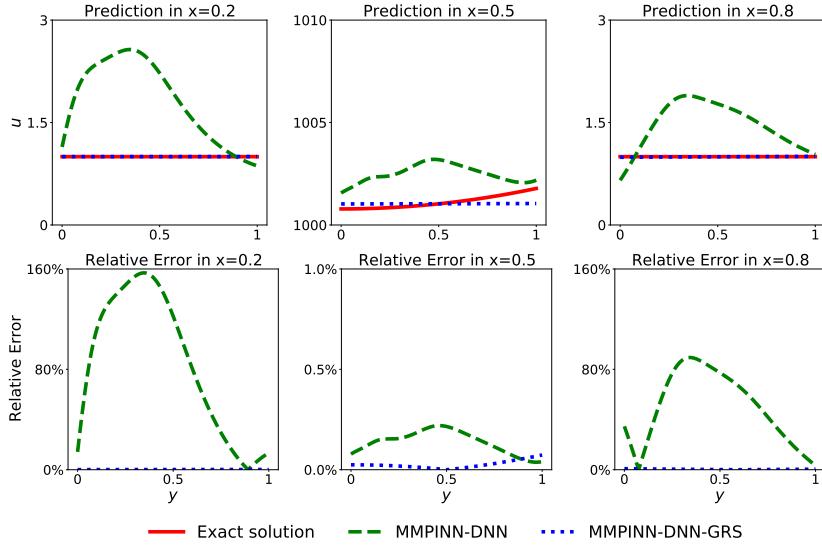


Figure 4.13: Comparison of the MMPINN-DNN and MMPINN-DNN-GRS methods.

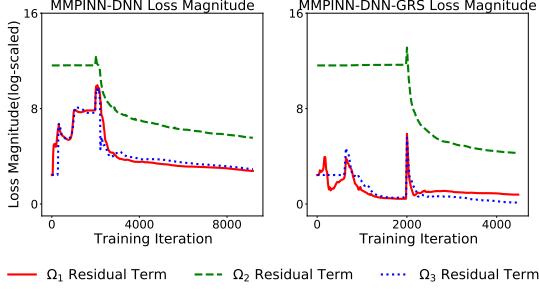


Figure 4.14: Loss variation curves of MMPINN-DNN and MMPINN-DNN-GRS.

Figure 4.14 shows that MMPINN-DNN-GRS further balances the difference of the residual terms between subdomains, which can cause the losses of all subdomains to decrease synchronously, resulting in more accurate predictions over the entire domain. In addition, the computational efficiency of MMPINN-DNN-GRS is higher than that of MMPINN-DNN. MMPINN-DNN-GRS requires about 4000 iterations to converge, while MMPINN-DNN requires nearly 10000 iterations. Note that the jump at 2000 iterations is due to the optimizer switching from Adam to L-BFGS.

5. Conclusion

In this paper, we analyze and investigate the difficulties in solving multi-scale problems using the conventional PINN method. For typical multi-scale models whose solutions have large gradients or high-frequency features, the supervised and residual terms of the loss function have large differences in magnitude. In this case, a pathology of the conventional PINN method is that there is a dominance of the PDE residuals in the optimization process. We eliminate this pathology by reconstructing the loss function using a novel regularization strategy for the loss terms, and propose an improved PNINN method denoted by MMPINN, which leads to a significant improvement not only in prediction accuracy but also in convergence speed. Compared with commonly used weighting methods, MMPINN makes the loss functions of physical neural networks more diverse, which opens a new door for tackling the magnitude difference in the loss function. In addition, for the multi-scale problems with multi-frequency features, the common deep neural networks used by conventional PINNs have inherent deficiencies in dealing with this type of problem. To address this issue, we embed IFNN into MFF and develop an Integrated Neural Network(INN) architecture, which simultaneously mitigates spectral bias and gradient flow stiffness, and improves the computational accuracy of the multi-frequency problem. Furthermore, by combining the above two approaches, we present the improved PINN framework, which is also called the MMPINN framework. The new methods derived from the MMPINN framework handle the multi-scale problems very well, and their performance outperforms many similar methods.

Although we achieve some interesting results in this paper, we acknowledge that the MMPINN framework is still in its early stages. In order to advance the MMPINN framework, we focus on the following issues in our future work: How to adaptively select the regularization parameters (m, n)? How to choose an appropriate neural network architecture without prior knowledge? Since the INN architecture requires much more computational resources, how to prune the neurons while maintaining relatively high accuracy? Can the grouping strategy be automated during training?

Acknowledgements

The work is supported by the National Key R&D Program of China under Grant No.2022YFA1004500, the National Science Foundation of China under Grant No.12271055 and No.12171048. We thank the authors of the papers [1, 15] for making their codes public, which helps to compare different methods.

References

- [1] S. Wang, H. Wang, P. Perdikaris, On the eigenvector bias of fourier feature networks: From regression to solving multi-scale pdes with physics-informed neural networks, Computer Methods in Applied Mechanics and Engineering 384 (2021) 113938. doi:[10.1016/j.cma.2021.113938](https://doi.org/10.1016/j.cma.2021.113938).
- [2] I. Lagaris, A. Likas, D. Fotiadis, Artificial neural networks for solving ordinary and partial differential equations, IEEE Transactions on Neural Networks 9 (5) (1998) 987–1000. doi:[10.1109/72.712178](https://doi.org/10.1109/72.712178).
- [3] M. Raissi, P. Perdikaris, G. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, Journal of Computational Physics 378 (2019) 686–707. doi:[10.1016/j.jcp.2018.10.045](https://doi.org/10.1016/j.jcp.2018.10.045).
- [4] G. Karniadakis, Y. Kevrekidis, L. Lu, et al., Physics-informed machine learning, Nature Reviews Physics 3 (2021) 1–19. doi:[10.1038/s42254-021-00314-5](https://doi.org/10.1038/s42254-021-00314-5).
- [5] L. Lu, X. Meng, Z. Mao, et al., DeepXDE: A deep learning library for solving differential equations, SIAM Review 63 (1) (2021) 208–228. doi:[10.1137/19M1274067](https://doi.org/10.1137/19M1274067).
- [6] M. Raissi, A. Yazdani, G. E. Karniadakis, Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations, Science 367 (6481) (2020) 1026–1030. doi:[10.1126/science.aaw4741](https://doi.org/10.1126/science.aaw4741).
- [7] L. Sun, H. Gao, S. Pan, et al., Surrogate modeling for fluid flows based on physics-constrained deep learning without simulation data, Computer Methods in Applied Mechanics and Engineering 361 (2020) 112732. doi:[10.1016/j.cma.2019.112732](https://doi.org/10.1016/j.cma.2019.112732).

- [8] X. Zhang, Y. Zhu, J. Wang, et al., Gw-pinn: A deep learning algorithm for solving groundwater flow equations, *Advances in Water Resources* 165 (2022) 104243. doi:[10.1016/j.advwatres.2022.104243](https://doi.org/10.1016/j.advwatres.2022.104243).
- [9] S. Y. Kang, S. H. Kim, J. Park, S. P. Jang, Thermal performance of a thin flat vapor chamber with a multiscale wick fabricated by sac305/sn58bi micro-sized particles, *International Journal of Heat and Mass Transfer* 212 (2023) 124305. doi:[10.1016/j.ijheatmasstransfer.2023.124305](https://doi.org/10.1016/j.ijheatmasstransfer.2023.124305).
- [10] H. Wang, H. Shin, A multiscale model to predict fatigue crack growth behavior of carbon nanofiber/epoxy nanocomposites, *International Journal of Fatigue* 168 (2023) 107467. doi:[10.1016/j.ijfatigue.2022.107467](https://doi.org/10.1016/j.ijfatigue.2022.107467).
- [11] L. Ke, Dream fusion in octahedral spherical hohlraum, *Matter and Radiation at Extremes* 7 (2022) 055701. doi:[10.1063/5.0103362](https://doi.org/10.1063/5.0103362).
- [12] Z. Liu, W. Cai, Z. J. Xu, Multi-scale deep neural network (mscalednn) for solving poisson-boltzmann equation in complex domains, *Communications in Computational Physics* 28 (5) (2020) 1970–2001. doi:[10.4208/cicp.OA-2020-0179](https://doi.org/10.4208/cicp.OA-2020-0179).
- [13] X. Li, Z. J. Xu, L. Zhang, Subspace decomposition based dnn algorithm for elliptic type multi-scale pdes, *Journal of Computational Physics* 488 (2023) 112242. doi:[10.1016/j.jcp.2023.112242](https://doi.org/10.1016/j.jcp.2023.112242).
- [14] S. Jin, Z. Ma, K. Wu, Asymptotic-preserving neural networks for multiscale time-dependent linear transport equations, *Journal of Scientific Computing* 94 (3) (2023). doi:[10.1007/s10915-023-02100-0](https://doi.org/10.1007/s10915-023-02100-0).
- [15] S. Wang, Y. Teng, P. Perdikaris, Understanding and mitigating gradient flow pathologies in physics-informed neural networks, *SIAM Journal on Scientific Computing* 43 (5) (2021) A3055–A3081. doi:[10.1137/20M1318043](https://doi.org/10.1137/20M1318043).
- [16] S. Wang, X. Yu, P. Perdikaris, When and why pinns fail to train: A neural tangent kernel perspective, *Journal of Computational Physics* 449 (2022) 110768. doi:[10.1016/j.jcp.2021.110768](https://doi.org/10.1016/j.jcp.2021.110768).
- [17] N. Rahaman, A. Baratin, D. Arpit, et al., On the spectral bias of neural networks, in: K. Chaudhuri, R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning*, *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 5301–5310.
- [18] B. Moseley, A. Markham, T. Nissen-Meyer, Finite basis physics-informed neural networks (fbpinns): a scalable domain decomposition approach for solving differential equations, *Advances in Computational Mathematics* 49 (62) (2023). doi:[10.1007/s10444-023-10065-9](https://doi.org/10.1007/s10444-023-10065-9).
- [19] D. Liu, Y. Wang, A dual-dimer method for training physics-constrained neural networks with minimax architecture, *Neural Networks* 136 (2021) 112–125. doi:[10.1016/j.neunet.2020.12.028](https://doi.org/10.1016/j.neunet.2020.12.028).

- [20] L. D. McClenney, U. M. Braga-Neto, Self-adaptive physics-informed neural networks, *Journal of Computational Physics* 474 (2023) 111722. doi:10.1016/j.jcp.2022.111722.
- [21] A. D. Jagtap, E. Kharazmi, G. E. Karniadakis, Conservative physics-informed neural networks on discrete domains for conservation laws: Applications to forward and inverse problems, *Computer Methods in Applied Mechanics and Engineering* 365 (2020) 113028. doi:10.1016/j.cma.2020.113028.
- [22] A. D. Jagtap, G. E. Karniadakis, Extended physics-informed neural networks (xpinns): A generalized space-time domain decomposition based deep learning framework for nonlinear partial differential equations, *Communications in Computational Physics* 28 (5) (2020) 2002–2041. doi:10.4208/cicp.OA-2020-0164.
- [23] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization (2017). arXiv:1412.6980.
- [24] R. H. Byrd, P. Lu, J. Nocedal, C. Zhu, A limited memory algorithm for bound constrained optimization, *SIAM Journal on Scientific Computing* 16 (5) (1995) 1190–1208. doi:10.1137/0916069.
- [25] Y. Chen, Z. Li, H. Cao, et al., Determination of laser entrance hole size for ignition-scale octahedral spherical hohlraums, *Matter and Radiation at Extremes* 7 (6), 065901 (10 2022). doi:10.1063/5.0102447.
- [26] G. Zhang, H. Yang, F. Zhu, et al., Dasa-pinns: Differentiable adversarial self-adaptive pointwise weighting scheme for physics-informed neural networks, *SSRN Electronic Journal* (2023).
- [27] J. Guo, Y. Yao, H. Wang, et al., Pre-training strategy for solving evolution equations based on physics-informed neural networks, *Journal of Computational Physics* 489 (2023) 112258. doi:10.1016/j.jcp.2023.112258.
- [28] C. Xu, B. T. Cao, Y. Yuan, G. Meschke, Transfer learning based physics-informed neural networks for solving inverse problems in tunneling (2022). arXiv:arXiv:2205.07731.
- [29] Z. Xiang, W. Peng, X. Liu, W. Yao, Self-adaptive loss balanced physics-informed neural networks, *Neurocomputing* 496 (2022) 11–34. doi:10.1016/j.neucom.2022.05.015.
- [30] C. Liu, A. Iserles, X. Wu, Symmetric and arbitrarily high-order birkhoff–hermite time integrators and their long-time behaviour for solving nonlinear klein–gordon equations, *Journal of Computational Physics* 356 (2018) 1–30. doi:10.1016/j.jcp.2017.10.057.