
ITERATIVE TRAINING OF PHYSICS-INFORMED NEURAL NETWORKS WITH FOURIER-ENHANCED FEATURES

Yulun Wu

Division of Decision and Control Systems
Digital Futures and KTH Royal Institute of Technology
Stockholm, Sweden
yulunw@kth.se

Miguel Aguiar

Division of Decision and Control Systems
Digital Futures and KTH Royal Institute of Technology
Stockholm, Sweden
aguiar@kth.se

Karl H. Johansson

Division of Decision and Control Systems
Digital Futures and KTH Royal Institute of Technology
Stockholm, Sweden
kallej@kth.se

Matthieu Barreau

Division of Decision and Control Systems
Digital Futures and KTH Royal Institute of Technology
Stockholm, Sweden
barreau@kth.se

October 23, 2025

ABSTRACT

Spectral bias, the tendency of neural networks to learn low-frequency features first, is a well-known issue with many training algorithms for physics-informed neural networks (PINNs). To overcome this issue, we propose IFeF-PINN, an algorithm for iterative training of PINNs with Fourier-enhanced features. The key idea is to enrich the latent space using high-frequency components through Random Fourier Features. This creates a two-stage training problem: (i) estimate a basis in the feature space, and (ii) perform regression to determine the coefficients of the enhanced basis functions. For an underlying linear model, it is shown that the latter problem is convex, and we prove that the iterative training scheme converges. Furthermore, we empirically establish that Random Fourier Features enhance the expressive capacity of the network, enabling accurate approximation of high-frequency PDEs. Through extensive numerical evaluation on classical benchmark problems, the superior performance of our method over state-of-the-art algorithms is shown, and the improved approximation across the frequency domain is illustrated.

1 Introduction

Capturing high-frequency behavior is central to modeling complex phenomena such as wave propagation, turbulence, and quantum dynamics. Traditional numerical methods, including spectral approaches [Boyd, 2001], multiscale schemes [Weinan and Engquist, 2003], and oscillatory quadrature [Iserles and Nørsett, 2005], have achieved notable success but often require problem-specific adaptations or become prohibitively costly in complex or high-dimensional settings.

There is a need for new approximation strategies that capture high-frequency behavior without sacrificing stability or tractability. Deep-learning surrogates of differential equations are a promising alternative, such as Physics-Informed Neural Networks (PINNs), which offer a grid-free alternative by combining data and physical models within a neural network framework [Raissi et al., 2017]. This paradigm has shown strong performance in solving partial differential equations (PDEs) and inferring hidden dynamics, benefiting adaptability to complex geometries [Costabal et al., 2024], and high-dimensional scalability [Hu et al., 2024]. Related approaches such as Fourier Neural Operators [Li et al., 2021] and DeepONet [Lu et al., 2021] further expand its reach. Despite these advances, PINN methods remain limited by *spectral bias*—the tendency of neural networks to learn low-frequency components first—which hinders accurate recovery of oscillatory solutions [Rahaman et al., 2019, Xu et al., 2025, Lin et al., 2021, Qin et al., 2024].

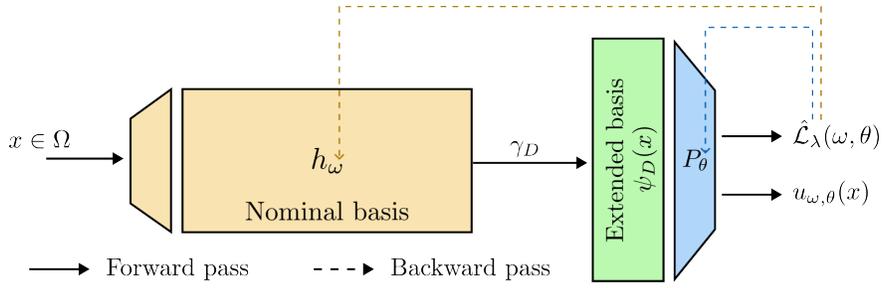


Figure 1: Architecture of IFeF-PINN. The first part (in yellow) generates the nominal basis vectors, which are then extended via γ_D generating random Fourier features ψ_D (in green), and a linear combination of the extended basis (in blue) forms the approximated solution $u_{\omega, \theta}$.

Table 1: Representative methods for approximating solutions to PDE, highlighting application domain, key idea, high-frequency handling (HF), limitations, and optimality.

Method	Domain	Key Idea	HF	Limitations / Optimality
Boyd [2001], Iserles and Nørsett [2005]	Linear	Global basis functions (Fourier, Chebyshev)	+++	Requires regular domains; global optimum
Weinan and Engquist [2003]	Multiscale	Separate scales and compute effective dynamics	++	Needs clear scale separation, problem-specific; local optimum
Raissi et al. [2017]	Generic	NN minimizing physics + data loss	-	Struggles with high-frequency components; local optimum
Li et al. [2021], Lu et al. [2021]	Operator	Learn mapping in Fourier / function space	+	Problem-specific, may require large networks; local optimum
Chai et al. [2024], Waheed [2022], Zhao et al. [2023]	Multiscale	Network architecture or training strategy	++	Problem-specific, not robust; local optimum
Lau et al. [2024], Tang et al. [2024], Song [2025]	General	Adaptive resampling	++	Computationally expensive, no convergence guarantees; local optimum
IFeF-PINN (this work)	Generic	Iterative training with extended basis via Fourier features	+++	Not adapted to resampling, high memory footprint; Global optimum (for linear PDEs)

Several strategies have been proposed to mitigate spectral bias, including weight balancing [Wang et al., 2021a, Krishnapriyan et al., 2021, Barreau and Shen, 2025], resampling [Lau et al., 2024, Tang et al., 2024, Song, 2025], and curriculum or architecture-based approaches [Sirignano and Spiliopoulos, 2018, Howard et al., 2025, Waheed, 2022, Chai et al., 2024, Mustajab et al., 2024, Eshkofti and Barreau, 2025, Wang and Lai, 2024]. Table 1 summarizes some of the most representative approaches. While effective in certain cases, these methods remain tied to single-level optimization frameworks, where feature learning and coefficient fitting are intertwined in neural networks, limiting both robustness and theoretical guarantees.

To address this gap, we take inspiration from classical numerical PDE solvers [Zienkiewicz et al., 2005], which approximate solutions using basis functions. We propose a new neural network architecture and tailored training algorithm. The key idea is to create a feed-forward neural network comprising three components, as illustrated in Figure 1. First, the hidden layers h_ω generate a nominal basis in the latent functional space. Next, this basis is extended to ψ_D with potentially higher-frequency elements to span a larger latent space. This is done through Random Fourier Features (RFF), introduced by Rahimi and Recht [2007]. Finally, the last linear layer performs regression on these extended basis vectors. The first and last blocks can be optimized separately, resulting in a two-stage iterative scheme alternating between latent basis construction and regression on output coefficients. A major feature of this framework, related to extreme learning machines [Dwivedi and Srinivasan, 2020], is that for linear differential equations, the regression stage is convex and achieves asymptotic global optimality. Unlike existing approaches, our method enriches the latent space representation, enabling systematic capture of high-frequency dynamics while leveraging the strengths of established PINN frameworks.

In this paper, we propose Iterative PINNs with Fourier-Enhanced Features (IFeF-PINN), a novel iterative two-stage training algorithm that mitigates the spectral bias of PINNs in high-frequency problems while maintaining accurate approximation on standard benchmark PDEs. Our contributions are threefold: (i) we introduce a flexible building block that augments existing PINNs architectures with improved high-frequency estimation and demonstrate its universal approximation capabilities; (ii) we propose an iterative two-stage training algorithm and prove its convergence properties; and (iii) we validate the approach through extensive simulations on benchmark problems, showing substantial improvements over existing methods.

2 Background

2.1 Physics-Informed Neural Networks

PINNs is a deep learning framework that integrates PDEs into the neural network training via the loss function, enabling data-driven learning with physical constraints [Raissi et al., 2017, Karniadakis et al., 2021].

Generally, for $n > 0$, let $\Omega \subset \mathbb{R}^n$ be a bounded domain and \mathcal{W} an appropriate Sobolev space of functions from Ω to \mathbb{R} , we consider PDEs of the form

$$\begin{aligned}\mathfrak{F}[u](x) &= f(x), & x \in \Omega, \\ \mathfrak{B}[u](s) &= g(s), & s \in \Gamma \subseteq \partial\Omega,\end{aligned}\tag{1}$$

where $u \in \mathcal{W}$ is the solution, $\mathfrak{F} : \mathcal{W} \rightarrow \mathcal{L}^2(\mathbb{R}^n, \mathbb{R})$ is the differential operator, $f \in \mathcal{L}^2(\Omega, \mathbb{R})$ is the source term, $\mathfrak{B} : \mathcal{W} \rightarrow \mathcal{Y}(\Gamma)$ is the boundary/initial operator, $g \in \mathcal{Y}(\Gamma)$ specifies the boundary/initial conditions, where $\mathcal{Y}(\Gamma)$ denotes the appropriate trace space. We assume that this problem is well-posed and therefore has a unique solution in \mathcal{W} .

The objective of PINNs is to approximate the solution u with a feedforward neural network u_ω , where ω denotes the network parameters. Yeonjong et al. [2020] and Sirignano and Spiliopoulos [2018] analyzed consistency in weak formulations under suitable assumptions, motivating the following continuum loss:

$$\mathfrak{L}_\lambda(u_\omega) = \frac{1}{|\Gamma|} \int_\Gamma \|g(s) - \mathfrak{B}[u_\omega](s)\|^2 ds + \frac{\lambda}{|\Omega|} \int_\Omega \|\mathfrak{F}[u_\omega](x)\|^2 dx,\tag{2}$$

with $\lambda > 0$ where, for A a bounded set, $|A|$ denotes its measure. However, this version is not numerically tractable and, in practice, we use the Monte Carlo approximation

$$\hat{\mathfrak{L}}_\lambda(u_\omega) = \frac{1}{N_u} \sum_{i=1}^{N_u} \|g(x_u^i) - \mathfrak{B}[u_\omega](x_u^i)\|^2 + \frac{\lambda}{N_f} \sum_{i=1}^{N_f} \|\mathfrak{F}[u_\omega](x_f^i)\|^2,\tag{3}$$

where $\{x_u^i\}_{i=1, \dots, N_u}$ and $\{x_f^i\}_{i=1, \dots, N_f}$ are uniformly sampled on Γ and Ω , respectively. Finally, the optimal parameters are found as $\omega^* = \arg \min_\omega \hat{\mathfrak{L}}_\lambda(u_\omega)$.

2.2 Random Fourier Features

In this work, we use Random Fourier Features (RFFs) introduced by Rahimi and Recht [2007] to include high-frequency terms. Grounded on Bochner’s theorem, RFF provides a way to explicitly construct a feature map that approximates a stationary kernel, enabling the scaling of kernel methods to large datasets.

RFF has been used by Tancik et al. [2020] to tackle spectral bias.

The novelty is to extend the input to the neural network using the RFF mapping

$$\gamma_D(x) = \frac{1}{\sqrt{D}} \begin{bmatrix} \cos(2\pi \mathbf{B}_D x) \\ \sin(2\pi \mathbf{B}_D x) \end{bmatrix} \in \mathbb{R}^{2D},\tag{4}$$

where the entries of the matrix $\mathbf{B}_D \in \mathbb{R}^{D \times n}$ are sampled from a given symmetric distribution. Wang et al. [2021b] adapted this method to PINNs by using u_ω from the previous section with $2D$ inputs, so that the neural network becomes $u_\omega \circ \gamma_D$. This new architecture can learn to approximate the solution from the enriched inputs. However, when the input is of low dimension, such as for PINNs, this could limit its asymptotic approximation capabilities.

3 Proposed Method

We leverage the PINNs and RFFs in a novel way. Note first that the PINN training process couples two roles within a single nonconvex objective: (i) hidden layers h_ω learn a nonlinear feature basis, and (ii) a linear regression operator

$P_\theta : h_\omega \mapsto h_\omega^\top \theta$ finds the optimal projection coefficients θ of the approximated solution onto the feature basis, thereby minimizing the loss $\hat{\mathcal{L}}_\lambda$. This coupling leads to PINN pathologies, where gradients from interior residuals can dominate and suppress boundary terms, and spectral bias drives low-frequency learning first, leaving oscillatory components underfit and slowing convergence on high-frequency modes [Wang et al., 2021b, 2022].

To overcome this coupling issue, we approximate the solution u to the PDEs in (1) as a linear combination of basis functions. We thus consider the two problems in isolation: basis generation, which we will denote as the upper-level problem, and linear regression on the basis functions, which we will call the lower-level problem.

3.1 The upper-level problem: Basis function generation

The initial step for the basis generation is to follow the classical PINN methodology and train a standard feed-forward neural network with parameters (ω, W) , denoted by

$$\tilde{u}_{\omega, W}(x) = Wh_\omega(x), \quad x \in \Omega, \quad (5)$$

to minimize $\omega, W \mapsto \hat{\mathcal{L}}_\lambda(\tilde{u}_{\omega, W})$. This is usually done using a gradient-descent numerical scheme such as ADAM [Kingma and Ba, 2014] or a more complex second-order solver such as L-BFGS [Liu and Nocedal, 1989]. Then, the neural network $h_\omega : \mathbb{R}^n \rightarrow \mathbb{R}^p$ generates a basis $h_\omega \in \mathcal{C}(\mathbb{R}, \mathbb{R}^p)$ of the latent space while W is the projection operator. This initial step acts as a warm start for the upper-level problem. Note that $\tilde{u}_{\omega, W}$ most likely contains only the low-frequency components of the original solution. Therefore, the surrogate $\tilde{u}_{\omega, W}$ might be an aliased or steady-state solution of the PDE, and the fit at the boundary points might be poor.

In our approach, the strategy is to apply an RFF mapping to the last hidden layer features h_ω . This upgrades the implicit linear kernel on h_ω to a stationary kernel, such as a radial basis function, in the adaptive feature space. Since \tilde{u} is probably a distorted version of the real solution u , the RFF extension might bring higher frequency signals that mitigate the spectral bias.

Concretely, we define $\psi_D(x) = \gamma_D(h_\omega(x)) = \frac{1}{\sqrt{D}} \begin{bmatrix} \cos(2\pi \mathbf{B}_D h_\omega(x)) \\ \sin(2\pi \mathbf{B}_D h_\omega(x)) \end{bmatrix}$ where $\mathbf{B}_D \in \mathbb{R}^{D \times p}$ is a constant matrix with entries sampled i.i.d. from $\mathcal{N}(0, \sigma^2)$.

3.2 The lower-level problem: Linear regression

The linear output layer over h_ω induces a dot-product kernel in feature space, which can limit expressivity and exacerbate spectral bias toward low frequencies. Applying RFF to h_ω equips the adaptive features with a stationary kernel without adding trainable parameters, injecting high-frequency components via random projections. Formally speaking, an approximate solution to the PDE in (1) with $\theta \in \mathbb{R}^{2D}$ becomes

$$u_{\omega, \theta}(x) = \psi_D(x)^\top \theta, \quad x \in \Omega.$$

As we show in Appendix B, assuming that the operators \mathfrak{F} and \mathfrak{B} are linear, the loss function $\hat{\mathcal{L}}_\lambda(u_{\omega, \theta})$ is quadratic in θ :

$$\mathfrak{L}_{\text{lower}}(\theta \mid \omega) := \hat{\mathcal{L}}_\lambda(u_{\omega, \theta}) = \frac{1}{2} \theta^\top Q(\omega) \theta + c(\omega)^\top \theta + b, \quad (6)$$

where Q and c collect boundary and interior residual terms.

Proposition 1. *Assume that the operators \mathfrak{F} and \mathfrak{B} are linear; $\lambda > 0$, and that the rank condition (3) from Appendix B.1 is verified. Then Q is positive definite and there is a unique solution to $\arg \min_\theta \mathfrak{L}_{\text{lower}}(\theta \mid \omega) = -Q^{-1}(\omega)c(\omega)$.*

The proof is given in Appendix B.1.

In the case of nonlinear operators \mathfrak{F} and \mathfrak{B} , the cost $\mathfrak{L}_{\text{lower}}$ is not necessarily convex in θ . Consequently, it is generally not possible to get an explicit solution of the optimal coefficients, and uniqueness is not guaranteed. In this case we use a gradient-descent update or L-BFGS to reach a local minimum.

3.3 The global bi-level problem

Combining the results from the two previous subsections, we get the following formulation that decouples basis learning (upper-level) from linear regression (lower-level):

$$\begin{aligned} \omega^*(\theta) &= \arg \min_\omega \hat{\mathcal{L}}_\lambda(u_{\omega, \theta}) := \arg \min_\omega \mathfrak{L}_{\text{upper}}(\omega \mid \theta), \\ \theta^*(\omega) &= \arg \min_\theta \hat{\mathcal{L}}_\lambda(u_{\omega, \theta}) := \arg \min_\theta \mathfrak{L}_{\text{lower}}(\theta \mid \omega). \end{aligned} \quad (7)$$

The classical bi-level optimization framework [Bard, 1991] proposes the following three-step numerical method: (i) sample w_0, θ_0 randomly; (ii) solve the upper-level problem $\omega^+ = \omega^*(\theta_0)$; (iii) solve the lower-level problem $\theta^+ = \theta^*(\omega^+)$. The final parameters (ω^+, θ^+) are the optimal solutions to the bi-level optimization.

However, this approach does not consider a warm start and is not particularly adapted to a learning problem. For better approximation capabilities, we propose an iterative scheme. We warm start using a vanilla PINN pre-training to get an initial value ω_0 for the weights of the basis generator. Then we compute $\theta_{i+1} = \theta^*(w_i)$ before performing a one-step gradient-descent on ω_i to minimize $\mathfrak{L}_{\text{upper}}(\omega_i | \theta_{i+1})$ to get ω_{i+1} . This leads to Algorithm 1. The convergence of this numerical scheme and the approximation capabilities of the new neural network architecture are studied in the next section.

Remark 1. Our approach is related to deep kernel learning, where a neural network learns a nonlinear feature transformation, and a Gaussian process is defined over the resulting feature space using a traditional kernel function. This enables learning a flexible, data-driven kernel that combines the expressiveness of deep learning with the uncertainty estimation of Gaussian processes [Wilson et al., 2016]. However, to the best of the authors’ knowledge, learning a Gaussian process with a nonlinear PDE prior is not yet possible [Jidling et al., 2017], and we propose a solution in that case.

Remark 2 (On resampling strategies). The proposed framework has broad integration potential, as it can be added to almost any existing neural architecture by changing the last layer to include the RFF extension. However, in practice, resampling strategies to approximate the integrals must be taken with care. Since the upper update is a slow process, resampling does not abruptly impact the solution, while, for the lower-level problem, its convexity implies a fast convergence. If there are not enough sampling points, we might converge to an aliased solution as a local minimum. A practical condition (given in Appendix B.1) on the minimum number of sampling points is $N_u + N_f > 2D$.

4 Theoretical Analysis

We provide convergence and approximation guarantees for the proposed bi-level training algorithm.

4.1 Convergence properties of the Bi-level algorithm

We establish convergence by showing that the optimal lower-level solution $\theta^*(\omega)$ is Lipschitz continuous with respect to the upper-level parameters ω , which ensures a well-defined Lipschitz hypergradient for gradient descent on the upper level.

Proposition 2 (Lipschitz Continuity of the Solution Map). *Let the lower-level problem be a strongly convex QP problem parameterized by ω . Assume that the mappings $\omega \mapsto Q(\omega)$ and $\omega \mapsto c(\omega)$ are locally Lipschitz continuous, and that the smallest eigenvalue of $Q(\omega)$ is uniformly bounded below by $\mu_Q > 0$ on any compact set of ω . Then, the optimal solution map $\theta^*(\omega)$ is also locally Lipschitz continuous with respect to ω . That is, for any compact set $\mathcal{W} \subset \mathbb{R}^P$, there exists a constant $K > 0$ such that for all $\omega_1, \omega_2 \in \mathcal{W}$:*

$$\|\theta^*(\omega_1) - \theta^*(\omega_2)\| \leq K \|\omega_1 - \omega_2\|. \quad (8)$$

The detailed proof is provided in Appendix C.2. This holds both in the nonlinear and linear PDE cases, though the proof is more straightforward for linear cases. Consequently, the hypergradient is L-smooth, which we use in the convergence analysis.

Theorem 1 (Convergence to a stationary point). *Assume that*

1. *The functions Q and c are continuously differentiable with respect to ω . The upper-level loss $\mathfrak{L}_{\text{upper}}$ is continuously differentiable with respect to both θ and ω .*
2. *The lower-level problem is μ -strongly convex.*
3. *The objective function $\mathfrak{L}_{\text{upper}}(\cdot | \theta)$ is bounded below and its hypergradient is L-smooth.*

together with the L-Smoothness of the hypergradient, formulated as

$$\exists L > 0, \quad \forall \omega_1, \omega_2, \quad \|\nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_1 | \theta) - \nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_2 | \theta)\| \leq L \|\omega_1 - \omega_2\|. \quad (9)$$

Algorithm 1 IFeF-PINN training

Initialize network parameter w_0, θ_0 and B
for k from 0 to N_{epoch} **do**
 Formulate extended RFF basis ψ_D
 $\psi_D = \gamma_D \circ h_{\omega}$
 $\mathfrak{L}_{\text{BLO}}(\omega; \theta) = \frac{1}{2} \theta^{\top} Q(\omega) \theta + c(\omega)^{\top} \theta + b(\omega)$.
 Lower update: $\theta_{k+1} = -Q(\omega_k)^{-1} c(\omega_k)$
 Upper update:
 $\omega_{k+1} = \omega_k - \eta \nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_k | \theta_{k+1})$
end for
return $\omega_{N_{\text{epoch}}}, \theta^*(\omega_{N_{\text{epoch}}})$

Then, the sequence of iterates $\{\omega_k\}_{k=0}^\infty$ generated by the gradient descent algorithm with a constant step size $\eta \in (0, 2/L)$ converges to a stationary point of $\mathcal{L}_{\text{upper}}(\cdot | \theta)$.

$$\lim_{k \rightarrow \infty} \|\nabla_{\omega} \mathcal{L}_{\text{upper}}(\omega_k | \theta)\| = 0. \quad (10)$$

The assumptions made are classical in learning problems and are a direct consequence of the structure of the bi-level framework. A formula for the hypergradient is derived via the Implicit Function Theorem in Appendix C.1, showing it as a composition of smooth functions. Its Lipschitz continuity is then guaranteed by the Lipschitz continuity of the solution map θ^* established in Theorem 2.

4.2 Universal approximation capabilities

To analyze the expressiveness of the RFF-augmented features, we show that the hypothesis class is not less expressive than linear readouts over the last hidden layer features. The necessary function spaces for this analysis are defined with comprehensive foundational definitions and proofs in Appendix D.

Definition 1. *The feature space \mathcal{H}_f and the composite RFF function space \mathcal{H}_{RFF} are defined as:*

$$\mathcal{H}_f := \{g \mid g(x) = h_{\omega}(x)^{\top} \theta, \theta \in \mathbb{R}^p\}, \quad \mathcal{H}_{\text{RFF}} := \{g \mid g(x) = \psi_D(x)^{\top} \theta, \theta \in \mathbb{R}^{2D}\}, \quad (11)$$

where $\psi_D = \gamma_D \circ h_{\omega}$ denotes the vector of composite RFF features defined in Equation 4.

We will show that $\overline{\mathcal{H}_{\text{RFF}}}$ strictly contains \mathcal{H}_f , and thus defines a more expressive hypothesis class. The argument constructs a bridge between the two spaces using a reproducing kernel Hilbert space.

Theorem 2. *Let f be any target function in $\mathcal{L}^2(\Omega, \mathbb{R})$. The projection error (see Definition 3 in Appendix D.1) achievable by the composite RFF Function Space \mathcal{H}_{RFF} is no greater than the projection error achieved by the original Feature Space \mathcal{H}_f when the number of RFF features D goes to infinity, that is:*

$$\inf_{g \in \mathcal{H}_{\text{RFF}, h}} \|f - g\|_{L^2} \leq \inf_{g \in \mathcal{H}_1} \|f - g\|_{L^2}. \quad (12)$$

The proof is given in Appendix D.2. This result establishes a powerful theoretical assurance that RFF embedding offers better approximation capabilities. Theorem 2 yields the universal approximation corollary presented below, the proof of which is given in Appendix D.2.1.

Corollary 1 (Universal approximation). *The projection error of the solution u to equation 1 onto \mathcal{H}_{RFF} can be made as small as desired, provided enough neurons and RFF features D .*

5 Related work

Weight-balancing strategies These methods adapt the physics weight λ in equation 3 during training. For instance, [Wang et al., 2021a] dynamically updates λ to balance the gradients of data and physics losses, while the NTK framework [Jacot et al., 2018, Krishnapriyan et al., 2021] enforces equal decay rates, theoretically recovering high-frequency solutions. Primal–dual methods [Goemans and Williamson, 1997, Barreau and Shen, 2025] instead compute λ from the PDE residual. Although simple to implement, these approaches offer weak convergence guarantees and remain tied to single-level optimization. Nonetheless, they are complementary to our framework and could be integrated as weight-balancing strategies within the upper-level problem.

Resampling strategies A second line of work reduces the gap between the true loss \mathcal{L}_{λ} and its sampled counterpart $\hat{\mathcal{L}}_{\lambda}$. Examples include NTK-informed sampling [Lau et al., 2024], adversarial sampling [Tang et al., 2024], and reinforcement learning [Song, 2025]. While effective in reducing approximation error, these methods do not explicitly target spectral bias, which is the focus of our proposed method.

Curriculum learning strategies Finally, new architectures and training schedules aim to better capture high-frequency components. Attention mechanisms [Sirignano and Spiliopoulos, 2018], multi-stage networks [Howard et al., 2025, Waheed, 2022, Chai et al., 2024, Mustajab et al., 2024, Eshkofti and Barreau, 2025, Wang and Lai, 2024], or finite-basis approximation [Moseley et al., 2023] have shown improved multi-scale resolution. However, their complexity often makes training slow and delicate, and they still lack dedicated optimization algorithms.

6 Numerical Experiments

Objective. In this section we describe comprehensive experiments that establish four main advantages of IFeF-PINN. First, improved approximation over PINNs and SOTA variants on low-frequency PDEs. Second, higher accuracy on high-frequency and multi-scale linear PDEs, where standard PINNs typically show failure modes. Third, strong generalization capabilities of our framework when integrated with advanced PINN variants. Finally, a spectrum analysis experiment demonstrates that our proposed method improves the network fitting accuracy for high-frequency signals.

Experiment setup. We will use four PDEs, namely the 2D Helmholtz equation (low and high frequency), 1D convection equation (low and high frequency), 1D convection-diffusion equation, and the viscous Burgers’ equation. The baseline methods are Vanilla PINNs, NTK [Wang et al., 2022], PINNsformer [Zhao et al., 2023], and Physics-Informed Gaussians (PIG) [Kang et al., 2024], keeping their default settings for a fair comparison. For simplicity, we set $\lambda = 0.01$ for the Vanilla PINNs in Equation 3. Detailed hyperparameters for our proposed methods are in Appendix E. For low-frequency 2D Helmholtz and low-frequency 1D convection equations, we adopt the uniform sampling strategy settings of Zhao et al. [2023]. For the viscous Burgers’ equation, we follow the setup of Raissi et al. [2019]. For the high-frequency Helmholtz equation, we employ Latin hypercube sampling [McKay et al., 2000] to improve domain coverage. We evaluate two variants of our framework: IFeF (Vanilla training) and IFeF-PD (primal-dual weight-balancing proposed by Barreau and Shen [2025]). PDE definitions, datasets, and network architectures are provided in Appendix F. We measure the relative L^2 -error after convergence. Each method is run five times with independent random seeds, with the best predictions for each approach. All models are implemented in PyTorch and trained on a single NVIDIA GeForce RTX 4090 GPU. The code for all benchmarks will be released on GitHub upon paper acceptance.

6.1 Results on benchmark PDEs

We begin with three popular low-frequency benchmark PDEs: 2D Helmholtz equation, 1D convection equation, and the viscous Burgers’ equation. Figure 2 summarizes relative L^2 -errors across baseline methods; boxplots display medians and IQRs, and red diamonds denote means. Additional prediction and absolute error maps are provided in Appendix G.

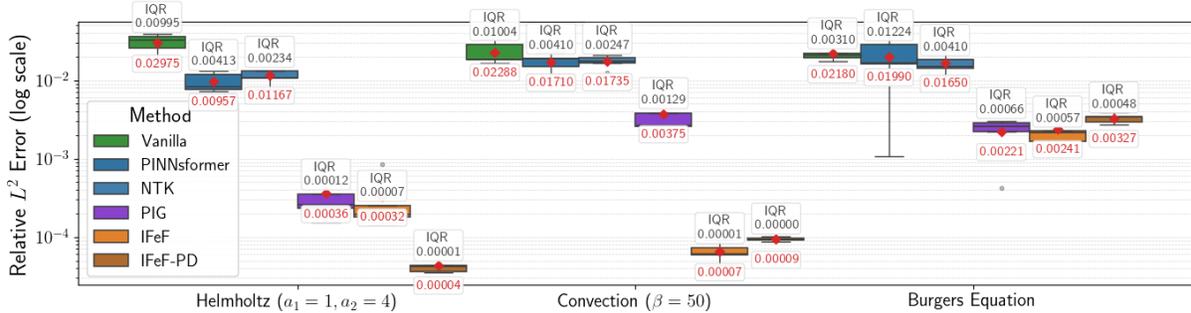


Figure 2: Boxplot of relative L^2 -errors (log10 scale) for all methods on three low-frequency benchmarks with median, inter-quartile range (IQR), and mean (red diamonds).

Across all these problems, our proposed method attains the lowest median errors with reduced variability. On Helmholtz, IFeF-PD achieves the best relative L^2 error of 3.5×10^{-5} . On convection, IFeF achieves the best error of 4.3×10^{-5} . Even in the nonlinear case of Burgers’ equation, IFeF obtains the lowest median error.

In addition, we conducted an ablation study where we discarded the RFF basis extension but performed a similar iterative two-step optimization process, obtaining similar but slightly better results than the Vanilla PINN (1.4923×10^{-2} relative L^2 -error) on the low-frequency convection problem.

Figure 3 presents the predictions for the low-frequency 2D Helmholtz case. On a logarithmic scale, the gap between IFeF-PINN and other methods is consistent with the box plot summaries. These results highlight the strong approximation capability of the proposed method, especially for linear equations, underscoring its robustness for solving diverse PDEs.

6.2 Mitigating the spectral bias

To evaluate challenging cases of spectral bias, we study the failure modes of PINNs on high-frequency and multi-scale PDEs, where vanilla PINNs typically struggle to learn rapidly oscillatory or widely separated frequency components.

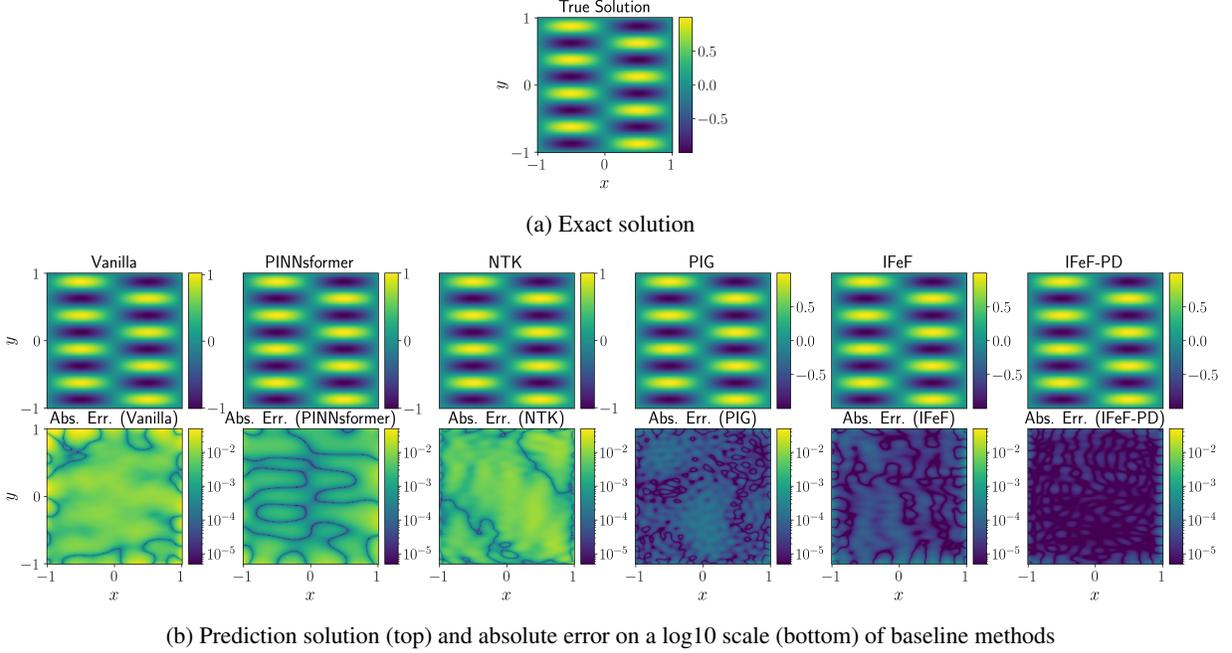


Figure 3: True solution, prediction and absolute error of baseline methods for Low-frequency Helmholtz equation

In particular, we study the high-frequency Helmholtz and convection equations, as well as a multi-scale convection-diffusion equation. Table 2 presents the mean and standard deviation of the relative L^2 -errors over baselines applied to these problems. Additional prediction and absolute error maps are provided in Appendix G.

Baseline	Helmholtz ($a_1 = a_2 = 100$)	Convection ($\beta = 200$)	Convection-Diffusion ($k_{\text{low}} = 4, k_{\text{high}} = 60$)
Vanilla	-	0.9024 (0.0239)	0.0501 (0.0030)
PINNsformer	-	1.2278 (0.2010)	0.0525 (0.0001)
NTK	-	0.8685 (0.0318)	0.0526 (0.0001)
PIG	1.6884 (0.2775)	1.0009 (0.0003)	0.0560 (0.0010)
IFeF	0.0156 (0.0055)	0.0027 (0.0010)	0.0009 (0.0003)
IFeF-PD	0.0092 (0.0031)	0.0025 (0.0005)	0.0010 (0.0002)

Table 2: Average relative L^2 -error with corresponding standard deviation for each baseline on three high-frequency PDEs. A dash '-' denotes that the baseline failed to converge.

Figure 4 depicts the high-frequency Helmholtz solutions and the corresponding log-scale absolute errors. In the considered scenarios, all baselines exhibit clear failure modes. We also conducted a similar ablation study as described in the previous section, removing the RFF basis extension, and the training did not converge for both the high-frequency Helmholtz and convection equations. In contrast, the proposed IFeF-PINN method effectively mitigates the spectral bias of neural networks. Moreover, when combined with the primal-dual method to adaptively balance the physics-based loss, our method achieves accurate solutions even under very high frequencies, which illustrates the flexibility of the proposed framework in incorporating advanced learning methods. A similar result holds for the multi-scale convection-diffusion equation in Figure 4 in Appendix G, clearly showing that only IFeF-PINN succeeds in learning both low and high frequency components of the solution. In contrast, all baselines suffer from the spectral bias failure mode, where models prioritize learning low-frequency components and tend to ignore the high-frequency components.

6.3 Spectrum Analysis

To quantitatively demonstrate our method’s ability to mitigate spectral bias, we employ the Fast Fourier transform to analyze the frequency-domain distribution of the network’s prediction. We conduct a spectrum analysis similar to Rahaman et al. [2019], designing a challenging multi-scale convection equation with an initial condition composed of a superposition of ten sinusoids of different frequencies and unit amplitude. More details of the setup are in Appendix F.2.

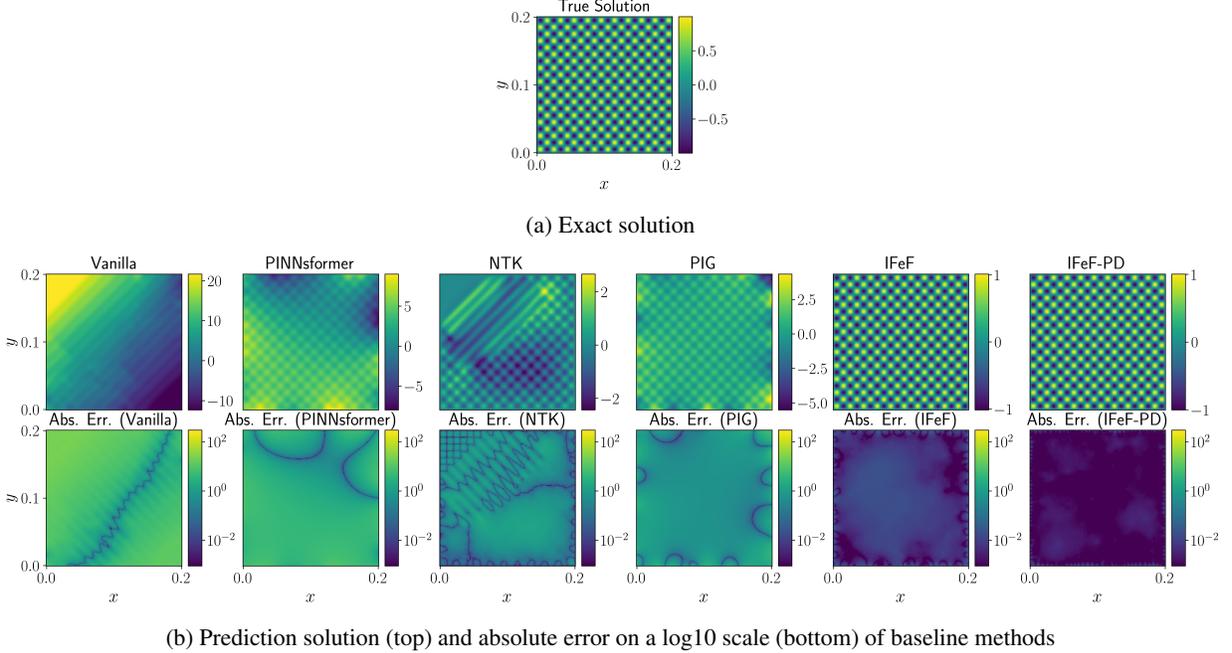


Figure 4: True solution, prediction and absolute error of baseline methods for High-frequency Helmholtz equation

During analysis, we compare the performance of Vanilla PINNs against models where the basis is extended with a varying number of random Fourier features and carry out a one-step solution of the lower-level objective in Equation 6. No additional training for the upper-level problem is performed.

We compute the magnitude of their discrete Fourier transform at frequencies k_i , denoted as $|\tilde{f}_{k_i}|$. Figure 5 presents the average normalized magnitudes $\frac{|\tilde{f}_{k_i}|}{A_i}$ over five independent runs. The results clearly illustrate the spectral bias of Vanilla PINNs, which struggle to accurately capture high-frequency components. In contrast, by extending the network’s basis through RFF, the network can fit high-frequency signals much more effectively, even without the subsequent bi-level training procedure of IFeF-PINN. Furthermore, we observe that increasing the number of random features enhances the network’s ability to approximate high-frequency components, confirming the effectiveness of our basis extension strategy.

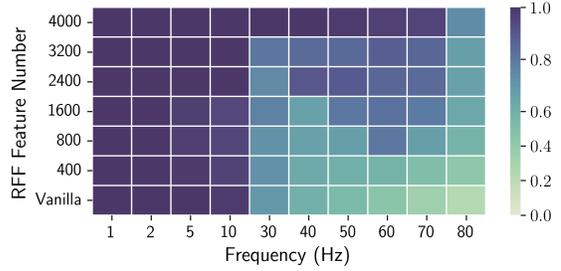


Figure 5: Prediction of the network spectrum with an increasing number of Fourier features. The x-axis represents frequency, and the colorbar shows the normalized magnitude of the predicted solution at $t = 0$. The colorbar is scaled accordingly from 0 to 1.

7 Conclusion

In this paper, we introduced IFeF-PINN, a novel iterative training of Fourier-enhanced Features PINNs. By augmenting the network with Random Fourier Features mapping as a basis extension with the bi-level problem, IFeF-PINN mitigates the spectral bias problem of standard PINNs when capturing the high-frequency and multi-scale components during training. Experimental results demonstrate that IFeF-PINN consistently outperforms advanced baselines across various scenarios, including popular low-frequency benchmarks and handling high-frequency and multi-scale PDEs. Furthermore, it has strong flexibility when integrating with different training strategies for PINNs.

Despite its strengths, IFeF-PINN faces two main challenges. First, for nonlinear PDEs, the lower-level problem becomes nonconvex, precluding a one-step solve and requiring iterative two-stage gradient descent updates that can stall in local minima. Advancing principled bi-level optimization to better reshape the nonlinear lower-level problem is a promising direction. Second, the method is sensitive to resampling because the bi-level objective depends strongly on minima found on a fixed dataset. These two non-trivial extensions are left for future work.

References

- Jonathan F Bard. Some properties of the bilevel programming problem. *Journal of optimization theory and applications*, 68(2):371–378, 1991.
- Matthieu Barreau and Haoming Shen. Accuracy and robustness of weight-balancing methods for training pinns. *arXiv preprint arXiv:2501.18582*, 2025.
- John P Boyd. *Chebyshev and Fourier spectral methods*. Courier Corporation, 2001.
- Xintao Chai, Wenjun Cao, Jianhui Li, Hang Long, and Xiaodong Sun. Overcoming the spectral bias problem of physics-informed neural networks in solving the frequency-domain acoustic wave equation. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.
- Francisco Sahli Costabal, Simone Pezzuto, and Paris Perdikaris. δ -pinns: Physics-informed neural networks on complex geometries. *Engineering Applications of Artificial Intelligence*, 127:107324, 2024.
- Vikas Dwivedi and Balaji Srinivasan. Physics informed extreme learning machine (pielm)—a rapid method for the numerical solution of partial differential equations. *Neurocomputing*, 391:96–118, 2020.
- Katayoun Eshkofti and Matthieu Barreau. Vanishing stacked-residual pinn for state reconstruction of hyperbolic systems. *IEEE Control Systems Letters*, 2025.
- Michel X Goemans and David P Williamson. The primal-dual method for approximation algorithms and its application to network design problems. *Approximation algorithms for NP-hard problems*, pages 144–191, 1997.
- Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.
- Amanda A. Howard, Sarah H. Murphy, and al. Stacked networks improve physics-informed training: Applications to neural networks and deep operator networks. *Foundations of Data Science*, 2025.
- Zheyuan Hu, Khemraj Shukla, George Em Karniadakis, and Kenji Kawaguchi. Tackling the curse of dimensionality with physics-informed neural networks. *Neural Networks*, 176:106369, 2024.
- Arieh Iserles and Syvert P Nørsett. Efficient quadrature of highly oscillatory integrals using derivatives. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 461(2057):1383–1399, 2005.
- Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.
- Carl Jidling, Niklas Wahlström, Adrian Wills, and Thomas B Schön. Linearly constrained gaussian processes. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/71ad16ad2c4d81f348082ff6c4b20768-Paper.pdf.
- Namgyu Kang, Jaemin Oh, Youngjoon Hong, and Eunbyung Park. Pig: Physics-informed gaussians as adaptive parametric mesh representations. *arXiv preprint arXiv:2412.05994*, 2024.
- G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang. Physics-informed machine learning. *Nature Reviews Physics*, 2021.
- Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Aditi Krishnapriyan, Amir Gholami, Shandian Zhe, Robert Kirby, and Michael W Mahoney. Characterizing possible failure modes in physics-informed neural networks. *Advances in neural information processing systems*, 34:26548–26560, 2021.
- Gregory Kang Ruey Lau, Apivich Hemachandra, See-Kiong Ng, and Bryan Kian Hsiang Low. PINNACLE: PINN adaptive collocation and experimental points selection. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=GzNaCp6Vcg>.
- Zongyi Li, Nikola Borislavov Kovachki, Kamyar Azizzadenesheli, Burigede liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=c8P9NQVtmm0>.
- Chensen Lin, Zhen Li, Lu Lu, Shengze Cai, Martin Maxey, and George Em Karniadakis. Operator learning for predicting multiscale bubble growth dynamics. *The Journal of Chemical Physics*, 154(10), 2021.
- Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.
- Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature machine intelligence*, 3(3):218–229, 2021.

- Michael D McKay, Richard J Beckman, and William J Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 42(1):55–61, 2000.
- Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.
- Ben Moseley, Andrew Markham, and Tarje Nissen-Meyer. Finite basis physics-informed neural networks (fbpinns): a scalable domain decomposition approach for solving differential equations. *Advances in Computational Mathematics*, 49(4):62, 2023.
- Abdul Hannan Mustajab, Hao Lyu, Zarghaam Rizvi, and Frank Wuttke. Physics-informed neural networks for high-frequency and multi-scale problems using transfer learning. *Applied Sciences*, 14(8):3204, 2024.
- Shaoxiang Qin, Fuyuan Lyu, Wenhui Peng, Dingyang Geng, Ju Wang, Naiping Gao, Xue Liu, and Liangzhu Leon Wang. Toward a better understanding of fourier neural operators: Analysis and improvement from a spectral perspective. *CoRR*, 2024.
- Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019.
- Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.
- Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations, 2017. URL <https://arxiv.org/abs/1711.10561>.
- Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.
- Claude E Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 2006.
- J. Sirignano and K. Spiliopoulos. DGM: A deep learning algorithm for solving partial differential equations. *Journal of computational physics*, 2018.
- Zhenao Song. Rl-pinns: Reinforcement learning-driven adaptive sampling for efficient training of pinns. *arXiv preprint arXiv:2504.12949*, 2025.
- Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.
- Kejun Tang, Jiayu Zhai, Xiaoliang Wan, and Chao Yang. Adversarial adaptive sampling: Unify PINN and optimal transport for the approximation of PDEs. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=7QI7tVrh2c>.
- Umair Bin Waheed. Kronecker neural networks overcome spectral bias for pinn-based wavefield computation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.
- Sifan Wang, Yujun Teng, and Paris Perdikaris. Understanding and mitigating gradient flow pathologies in physics-informed neural networks. *SIAM Journal on Scientific Computing*, 43(5):A3055–A3081, 2021a.
- Sifan Wang, Hanwen Wang, and Paris 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*, 384:113938, 2021b.
- Sifan Wang, Xinling Yu, and Paris Perdikaris. When and why pinns fail to train: A neural tangent kernel perspective. *Journal of Computational Physics*, 449:110768, 2022.
- Yongji Wang and Ching-Yao Lai. Multi-stage neural networks: Function approximator of machine precision. *Journal of Computational Physics*, 504:112865, 2024.
- E Weinan and Bjorn Engquist. The heterogenous multiscale methods. *Communications in Mathematical Sciences*, 1(1): 87–132, 2003.
- Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial intelligence and statistics*, pages 370–378. PMLR, 2016.
- Zhi-Qin John Xu, Lulu Zhang, and Wei Cai. On understanding and overcoming spectral biases of deep neural network learning methods for solving pdes. *Journal of Computational Physics*, page 113905, 2025.

Shin Yeonjong, Darbon Jérôme, and Em Karniadakis, George. On the convergence of physics-informed neural networks for linear second-order elliptic and parabolic type pdes. *Communications in Computational Physics*, 28(5):2042–2074, 2020.

Zhiyuan Zhao, Xueying Ding, and B Aditya Prakash. Pinnsformer: A transformer-based framework for physics-informed neural networks. *arXiv preprint arXiv:2307.11833*, 2023.

O.C. Zienkiewicz, R.L. Taylor, and J.Z. Zhu. *The Finite Element Method: Its Basis and Fundamentals*. Butterworth-Heinemann, 2005. ISBN 9780080472775.

A Contents

We organize this supplementary document as follows:

- Section B provides the formulation for the QP problem and proof of Proposition 1.
- Section C provides the proofs for the convergence analysis of our method.
- Section D provides the proofs for the projection error analysis of our method.
- Section E details the hyperparameter settings of our method.
- Section F describes the experimental setup for all PDEs, models, and datasets.
- Section G presents additional results considered in our numerical experiments.

B QP formulation details and proof of Proposition 1

Let $X_u = (x_u^1, \dots, x_u^{N_u}) \in \mathbb{R}^{N_u \times n}$ be boundary collocation points and $X_f = (x_f^1, \dots, x_f^{N_f}) \in \mathbb{R}^{N_f \times n}$ be interior collocation points. We define

- $B_u(\omega) \in \mathbb{R}^{N_u \times 2D}$ with rows $\mathfrak{B}[\psi \circ h_\omega](x_u^i)^\top \in \mathbb{R}^{2D}$ for $i = 1, \dots, N_u$ as boundary values;
- $G_u \in \mathbb{R}^{N_u}$ with rows $g(x_u^i)^\top \in \mathbb{R}^{2D}$ for $i = 1, \dots, N_u$ as boundary measurement;
- $R_f(\omega) \in \mathbb{R}^{N_f \times 2D}$ with rows $\mathfrak{F}[\psi \circ h_\omega](x_f^i)^\top$ for $i = 1, \dots, N_f$ as residual values;
- $F_f \in \mathbb{R}^{N_f}$ with rows $f(x_f^i)^\top$ for $i = 1, \dots, N_f$ as source terms.

Note that the linearity of the \mathfrak{B} and \mathfrak{F} operators leads to $\mathfrak{B}[u_{\omega, \theta}](X_u) = B_u \theta$ and $\mathfrak{F}[u_{\omega, \theta}](X_f) = R_f \theta$. Consequently, using the standard PIML loss with boundary and PDE residual terms, we get:

$$\hat{\mathfrak{L}}_\lambda(u_{\omega, \theta}) = \frac{1}{N_u} \|B_u(\omega)\theta - G_u\|^2 + \frac{\lambda}{N_f} \|R_f(\omega)\theta - F_f\|^2. \quad (13)$$

The loss expands to the quadratic form

$$\begin{aligned} \mathfrak{L}_{\text{lower}}(\theta | \omega) = & \frac{1}{2} \theta^\top \left(\frac{2}{N_u} B_u^\top B_u + \frac{2\lambda_{\text{LL}}}{N_f} R_f^\top R_f \right) \theta + \left(-\frac{2}{N_u} B_u^\top G_u - \frac{2\lambda_{\text{LL}}}{N_f} R_f^\top F_f \right)^\top \theta \\ & + \frac{1}{N_u} G_u^\top G_u + \frac{\lambda_{\text{LL}}}{N_f} F_f^\top F_f. \end{aligned} \quad (14)$$

Identifying

$$Q(\omega) = \frac{2}{N_u} B_u(\omega)^\top B_u(\omega) + \frac{2\lambda_{\text{LL}}}{N_f} R_f(\omega)^\top R_f(\omega), \quad c(\omega) = -\frac{2}{N_u} B_u(\omega)^\top G_u - \frac{2\lambda_{\text{LL}}}{N_f} R_f(\omega)^\top F_f,$$

leads to $\mathfrak{L}_{\text{lower}}(\theta | \omega) = \frac{1}{2} \theta^\top Q(\omega) \theta + c(\omega)^\top \theta + b$, where b is a constant term and λ_{LL} is the physics weight used in the lower-level problem.

B.1 Analysis of the Rank Condition and Regularization of Proposition 1

The positive semi-definiteness of $Q(\omega)$ follows from the factorization $Q(\omega) = M(\omega)^\top M(\omega) \in \mathbb{R}^{2D \times 2D}$, with the stacked design matrices $M(\omega)$ as follows:

$$M(\omega) = \begin{pmatrix} \sqrt{\frac{2}{N_u}} B_u(\omega) \\ \sqrt{\frac{2\lambda}{N_f}} R_f(\omega) \end{pmatrix},$$

where $\lambda > 0$ is always ensured. Then we have $\text{rank}(Q(\omega)) = \text{rank}(M(\omega))$. Hence, $Q(\omega)$ is strictly positive definite if and only if $M(\omega)$ has full column rank, i.e.,

$$\text{rank}(M(\omega)) = 2D. \quad (15)$$

The loss $\mathfrak{L}_{\text{lower}}(\theta | \omega)$ becomes then strongly convex, and its minimization has a unique solution $\theta^* = -Q^{-1}c$.

The rank condition in equation 15 imposes a fundamental constraint: for the stacked design matrix $M \in \mathbb{R}^{(N_u+N_f) \times 2D}$ to have full column rank $2D$, it is necessary that the number of rows is at least as large as the number of columns. This leads to the critical requirement on the number of sampling points: $N_u + N_f \geq 2D$.

When this condition is violated (i.e., $N_u + N_f < 2D$), the system becomes underdetermined. This directly causes the matrix $Q(\omega)$ to be rank-deficient. Consequently, the lower-level problem becomes ill-posed, lacking a unique solution as Q^{-1} does not exist. This might lead to an aliased solution (especially true when we add a Tikhonov regularization).

A more subtle cause of rank deficiency occurs even when $N_u + N_f \geq 2D$. This happens if the collocation points provide redundant information, failing to create $2D$ linearly independent constraints. Such a situation can arise from geometrically poor sampling (e.g., points lying on nodal lines of the basis functions) or from inherent redundancies in the randomly generated RFF basis itself (e.g., two different random vectors being nearly parallel). In these scenarios, although the design matrix M has enough rows, its columns remain linearly dependent, leading to a singular or, more commonly, a numerically ill-conditioned matrix Q . However, if D is large, a solution might be to just discard these redundancies while still keeping the same feature space \mathcal{H}_{RFF} (see equation equation 11).

To resolve this, we employ Tikhonov regularization, modifying the matrix to $Q_{\text{reg}}(\omega) = Q(\omega) + \gamma I$, where $\gamma > 0$ is a small regularization parameter. This ensures Q_{reg} is always positive definite and invertible, since for any non-zero θ , the quadratic form $\theta^\top Q_{\text{reg}} \theta = \theta^\top Q \theta + \gamma \|\theta\|^2$ is strictly positive. The lower-level problem thus regains a unique, stable solution

$$\theta^*(\omega) = -(Q(\omega) + \gamma I)^{-1} c(\omega). \quad (16)$$

This analysis reveals a fundamental trade-off: while increasing D enhances representational power, it demands proportionally more sampling points to maintain a well-posed system. When the sampling budget is limited, Tikhonov regularization provides a principled remedy to ensure algorithmic stability, at the cost of introducing a slight bias to the solution.

C Proofs of Convergence Analysis

C.1 Hypergradient Derivation via Implicit Function Theorem (IFT)

The hypergradient $\nabla_\omega \mathcal{L}_{\text{upper}}(\omega)$ is computed using the chain rule, where the Implicit Function Theorem (IFT) provides the Jacobian of the lower-level solution map $\theta^*(\omega)$.

The lower-level optimality condition is $F(\theta^*, \omega) := Q(\omega)\theta^* - c(\omega) = 0$. Taking the total derivative with respect to ω yields

$$\frac{\partial F}{\partial \theta^\top} \frac{\partial \theta^*}{\partial \omega^\top} + \frac{\partial F}{\partial \omega^\top} = 0.$$

Solving for the Jacobian $\frac{\partial \theta^*}{\partial \omega^\top}$ gives

$$\frac{\partial \theta^*}{\partial \omega^\top} = - \left(\frac{\partial F}{\partial \theta^\top} \right)^{-1} \frac{\partial F}{\partial \omega^\top} = -Q(\omega)^{-1} \left(\frac{\partial(Q(\omega)\theta^*)}{\partial \omega^\top} - \frac{\partial c(\omega)}{\partial \omega^\top} \right).$$

The full hypergradient is then obtained by the chain rule

$$\nabla_\omega \mathcal{L}_{\text{upper}}(\omega) = \frac{\partial \mathcal{L}_{\text{upper}}}{\partial \omega} + \left(\frac{\partial \theta^*}{\partial \omega^\top} \right)^\top \frac{\partial \mathcal{L}_{\text{upper}}}{\partial \theta}. \quad (17)$$

Substituting the expression for the Jacobian, we get

$$\nabla_\omega \mathcal{L}_{\text{upper}}(\omega) = \frac{\partial \mathcal{L}_{\text{upper}}}{\partial \omega} - \left(\frac{\partial(Q(\omega)\theta^*)}{\partial \omega^\top} - \frac{\partial c(\omega)}{\partial \omega^\top} \right)^\top Q(\omega)^{-1} \frac{\partial \mathcal{L}_{\text{upper}}}{\partial \theta}.$$

This provides a computable formula for the gradient used in the upper-level optimization.

C.2 Proof of Proposition 2

We first state the key properties required for our convergence analysis.

Assumption 1 (Smoothness and Boundedness Properties). The bi-level optimization problem satisfies the following regularity conditions:

1. The functions $Q(\omega)$ and $c(\omega)$ are continuously differentiable with respect to ω . The upper-level loss $\mathfrak{L}_{\text{upper}}(\theta, \omega)$ is continuously differentiable with respect to both θ and ω .
2. The lower-level problem is μ -strongly convex, i.e., $Q(\omega) \succeq \mu I$ for some constant $\mu > 0$.
3. The objective function $\mathfrak{L}_{\text{upper}}(\omega)$ is bounded below by a scalar $\mathfrak{L}_{\text{inf}}$.

Assumption 2 (L-Smoothness of the Hypergradient). The upper-level objective function $\mathfrak{L}_{\text{upper}}(\omega)$ is L -smooth, a standard assumption in gradient-based optimization analysis. This means its gradient, the hypergradient $\nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega)$, is Lipschitz continuous with constant $L > 0$:

$$\|\nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_1) - \nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_2)\| \leq L \|\omega_1 - \omega_2\|, \quad \forall \omega_1, \omega_2 \in \mathbb{R}^P. \quad (18)$$

These assumptions trivially hold when neural network activation functions are Lipschitz continuous and the loss function is smooth, which is satisfied by our choice of tanh activations with MSE losses.

Proof. Since $\mathfrak{L}_{\text{lower}}$ is strongly convex and differentiable with respect to θ , the unique optimal solution $\theta^*(\omega)$ is found by setting the gradient to zero:

$$\nabla_{\theta} \mathfrak{L}_{\text{lower}}(\theta^*(\omega)) = Q(\omega)\theta^*(\omega) - c(\omega) = 0. \quad (19)$$

This gives the closed-form solution showed in Proposition 1:

$$\theta^*(\omega) = Q(\omega)^{-1}c(\omega). \quad (20)$$

Then consider two parameter vectors ω_1, ω_2 from a compact set \mathcal{W} . We want to bound the norm of the difference $\|\theta^*(\omega_1) - \theta^*(\omega_2)\|$. This follows the structure from your provided image:

$$\begin{aligned} \|\theta^*(\omega_1) - \theta^*(\omega_2)\| &= \|Q(\omega_1)^{-1}c(\omega_1) - Q(\omega_2)^{-1}c(\omega_2)\| \\ &= \|Q(\omega_1)^{-1}c(\omega_1) - Q(\omega_1)^{-1}c(\omega_2) + Q(\omega_1)^{-1}c(\omega_2) - Q(\omega_2)^{-1}c(\omega_2)\| \\ &\leq \|Q(\omega_1)^{-1}(c(\omega_1) - c(\omega_2))\| + \|(Q(\omega_1)^{-1} - Q(\omega_2)^{-1})c(\omega_2)\| \\ &\leq \|Q(\omega_1)^{-1}\| \cdot \|c(\omega_1) - c(\omega_2)\| + \|Q(\omega_1)^{-1} - Q(\omega_2)^{-1}\| \cdot \|c(\omega_2)\|. \end{aligned} \quad (21)$$

We use the matrix identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ to bound the second term:

$$\begin{aligned} \|Q(\omega_1)^{-1} - Q(\omega_2)^{-1}\| &= \|Q(\omega_1)^{-1}(Q(\omega_2) - Q(\omega_1))Q(\omega_2)^{-1}\| \\ &\leq \|Q(\omega_1)^{-1}\| \cdot \|Q(\omega_2) - Q(\omega_1)\| \cdot \|Q(\omega_2)^{-1}\|. \end{aligned} \quad (22)$$

Substituting Equation 22 back into Equation 21:

$$\begin{aligned} \|\theta^*(\omega_1) - \theta^*(\omega_2)\| &\leq \|Q(\omega_1)^{-1}\| \cdot \|c(\omega_1) - c(\omega_2)\| \\ &\quad + \|Q(\omega_1)^{-1}\| \cdot \|Q(\omega_2) - Q(\omega_1)\| \cdot \|Q(\omega_2)^{-1}\| \cdot \|c(\omega_2)\|. \end{aligned} \quad (23)$$

By Assumption 1, on the compact set \mathcal{W} , there exist constants $L_Q, L_c > 0$ such that $\|Q(\omega_1) - Q(\omega_2)\| \leq L_Q \|\omega_1 - \omega_2\|$ and $\|c(\omega_1) - c(\omega_2)\| \leq L_c \|\omega_1 - \omega_2\|$. Furthermore, due to strong convexity, there is a $\mu_Q > 0$ such that $\|Q(\omega)^{-1}\| \leq 1/\mu_Q$ for all $\omega \in \mathcal{W}$. Finally, since $c(\omega)$ is continuous on a compact set, its norm is bounded by a constant $C_{\text{max}} = \sup_{\omega \in \mathcal{W}} \|c(\omega)\|$.

Substituting these bounds into Equation 23:

$$\begin{aligned} \|\theta^*(\omega_1) - \theta^*(\omega_2)\| &\leq \frac{1}{\mu_Q} (L_c \|\omega_1 - \omega_2\|) + \left(\frac{1}{\mu_Q} \cdot L_Q \|\omega_1 - \omega_2\| \cdot \frac{1}{\mu_Q} \right) C_{\text{max}} \\ &= \left(\frac{L_c}{\mu_Q} + \frac{L_Q C_{\text{max}}}{\mu_Q^2} \right) \|\omega_1 - \omega_2\|. \end{aligned}$$

Defining the constant $K = \frac{L_c}{\mu_Q} + \frac{L_Q C_{\text{max}}}{\mu_Q^2}$ completes the proof. \square

C.3 Proof of Theorem 1

The proof relies on the following standard lemma for L -smooth functions.

Lemma 1 (Sufficient Decrease). *If $\mathfrak{L}_{\text{upper}}(\omega)$ is L -smooth with constant L and the step size $\eta \in (0, 2/L)$, the gradient descent update rule ensures a sufficient decrease in the objective function:*

$$\mathfrak{L}_{\text{upper}}(\omega_{k+1}) \leq \mathfrak{L}_{\text{upper}}(\omega_k) - \eta \left(1 - \frac{L\eta}{2}\right) \|\nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_k)\|^2.$$

Proof. From the L -smoothness property (descent lemma) under Assumption 2, we have:

$$\mathfrak{L}_{\text{upper}}(\omega_{k+1}) \leq \mathfrak{L}_{\text{upper}}(\omega_k) + \nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_k)^{\top} (\omega_{k+1} - \omega_k) + \frac{L}{2} \|\omega_{k+1} - \omega_k\|^2$$

Substituting the gradient descent update $\omega_{k+1} - \omega_k = -\eta \nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_k)$:

$$\begin{aligned} \mathfrak{L}_{\text{upper}}(\omega_{k+1}) &\leq \mathfrak{L}_{\text{upper}}(\omega_k) - \eta \|\nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_k)\|^2 + \frac{L\eta^2}{2} \|\nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_k)\|^2 \\ &= \mathfrak{L}_{\text{upper}}(\omega_k) - \eta \left(1 - \frac{L\eta}{2}\right) \|\nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_k)\|^2. \end{aligned}$$

□

Proof of Theorem 1. Let $\delta = \eta(1 - L\eta/2)$. Since $\eta \in (0, 2/L)$, we have $\delta > 0$. Rearranging the inequality from Lemma 1 gives:

$$\delta \|\nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_k)\|^2 \leq \mathfrak{L}_{\text{upper}}(\omega_k) - \mathfrak{L}_{\text{upper}}(\omega_{k+1}).$$

We now sum this inequality from $k = 0$ to T to form a telescoping series:

$$\begin{aligned} \sum_{k=0}^T \delta \|\nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_k)\|^2 &\leq \sum_{k=0}^T (\mathfrak{L}_{\text{upper}}(\omega_k) - \mathfrak{L}_{\text{upper}}(\omega_{k+1})) \\ &= (\mathfrak{L}_{\text{upper}}(\omega_0) - \mathfrak{L}_{\text{upper}}(\omega_1)) + (\mathfrak{L}_{\text{upper}}(\omega_1) - \mathfrak{L}_{\text{upper}}(\omega_2)) + \dots \\ &\quad + (\mathfrak{L}_{\text{upper}}(\omega_T) - \mathfrak{L}_{\text{upper}}(\omega_{T+1})) \\ &= \mathfrak{L}_{\text{upper}}(\omega_0) - \mathfrak{L}_{\text{upper}}(\omega_{T+1}). \end{aligned}$$

The objective function is bounded below by $\mathfrak{L}_{\text{inf}} \geq 0$. Therefore, $\mathfrak{L}_{\text{upper}}(\omega_{T+1}) \geq \mathfrak{L}_{\text{inf}}$. This gives us:

$$\sum_{k=0}^T \delta \|\nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_k)\|^2 \leq \mathfrak{L}_{\text{upper}}(\omega_0) - \mathfrak{L}_{\text{inf}}.$$

As $T \rightarrow \infty$, the right-hand side is a finite constant. This implies that the infinite series of squared gradient norms is bounded:

$$\sum_{k=0}^{\infty} \|\nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_k)\|^2 \leq \frac{\mathfrak{L}_{\text{upper}}(\omega_0) - \mathfrak{L}_{\text{inf}}}{\delta} < \infty.$$

For an infinite series of non-negative terms to converge to a finite value, the terms themselves must converge to zero. Therefore, we must have:

$$\lim_{k \rightarrow \infty} \|\nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_k)\|^2 = 0,$$

which implies that $\lim_{k \rightarrow \infty} \|\nabla_{\omega} \mathfrak{L}_{\text{upper}}(\omega_k)\| = 0$. This completes the proof that the algorithm converges to a stationary point. □

D Proofs for Projection Error Analysis

This appendix provides the foundational definitions and detailed proofs for the projection error analysis presented in Section 4.2.

D.1 Foundational Concepts

Definition 2 (Universal Kernel). A continuous kernel k defined on a compact metric space (\mathcal{X}, d) is called a **universal kernel** if the Reproducing Kernel Hilbert Space (RKHS) \mathcal{H}_k induced by k is dense in the space of continuous functions $C(\mathcal{X})$ with respect to the uniform norm $\|\cdot\|_\infty$.

Mathematically, this means that for any function $g \in C(\mathcal{X})$ and any $\varepsilon > 0$, there exists a function $f \in \mathcal{H}_k$ such that:

$$\sup_{x \in \mathcal{X}} |f(x) - g(x)| < \varepsilon.$$

An equivalent way to state this is that the closure of \mathcal{H}_k under the uniform norm is $C(\mathcal{X})$:

$$\overline{\mathcal{H}_k} = C(\mathcal{X}).$$

Definition 3. Let $f \in \mathcal{L}^2(\Omega, \mathbb{R})$ be a target function. The **projection error** of f onto an Hilbert space $\mathcal{H} \subseteq \mathcal{L}^2(\Omega, \mathbb{R})$ is defined as

$$\text{Err}(f, \mathcal{H}) := \inf_{g \in \mathcal{H}} \|f - g\|. \quad (24)$$

If this infimum is attained by some $g^* \in \mathcal{H}$, then g^* is the projection of f onto \mathcal{H} .

Theorem 3 (Composition of Universal Kernels). If $k(z, z')$ is a universal kernel on a space \mathcal{Z} , and the mapping $h : \mathcal{X} \rightarrow \mathcal{Z}$ is continuous and sufficiently expressive (e.g., injective), then the composite kernel $k_h(x, x') := k(h(x), h(x'))$ is universal on \mathcal{X} .

Remark 3. The universality of the composite kernel relies on the composition theorem from Micchelli et al. [2006].

Theorem 4 (RFF Approximation). Let $k : \mathbb{R}^m \times \mathbb{R}^m \rightarrow \mathbb{R}$ be a continuous, translation-invariant, positive definite kernel function, i.e.,

$$k(x, x') = k(x - x') = \int_{\mathbb{R}^d} e^{iw^T(x-x')} d\mu(w),$$

where μ is a probability measure with compact support. Define the Random Fourier Feature (RFF) map $\phi_{\text{RFF}} : \mathcal{X} \rightarrow \mathbb{R}^{2D}$ as

$$\phi_{\text{RFF}}(x) := \sqrt{\frac{1}{D}} \begin{bmatrix} \cos(w_1^\top x + b_1) \\ \cos(w_D^\top x + b_D) \\ \vdots \\ \sin(w_1^\top x + b_1) \\ \sin(w_D^\top x + b_D) \end{bmatrix}, \quad (25)$$

where $w_i \stackrel{\text{i.i.d.}}{\sim} \mu$ and $b_i \stackrel{\text{i.i.d.}}{\sim} \text{Uniform}[0, 2\pi]$, with $\{w_i\}$ independent of $\{b_i\}$.

Let \mathcal{H}_k be the RKHS corresponding to k . For any $f \in \mathcal{H}_k$ and any probability distribution ρ :

$$\lim_{D \rightarrow \infty} \inf_{\theta \in \mathbb{R}^D} \|f - \phi_{\text{RFF}}^T \theta\|_{L^2(\rho)} = 0,$$

i.e., the function space spanned by RFF features is dense in \mathcal{H}_k .

Remark 4. This result is a direct consequence of the uniform convergence of the RFF kernel approximation to the true kernel Rahimi and Recht [2007].

Theorem 5 (RFF Approximation for Composite Kernels). Let k be a continuous, translation-invariant, positive definite kernel function on \mathbb{R}^m and $h : \mathbb{R}^d \rightarrow \mathbb{R}^m$ be a continuous mapping. Let \mathcal{H}_{k_h} be the RKHS of the composite kernel $k_h(x, x') = k(h(x), h(x'))$. The function space spanned by the composite RFF features, \mathcal{H}_{RFF} defined in Equation 11, is dense in \mathcal{H}_{k_h} with respect to the $L^2(\rho)$ norm when $D \rightarrow \infty$.

$$\lim_{D \rightarrow \infty} \inf_{\theta \in \mathbb{R}^D} \|f(x) - \psi(x)^\top \theta\|_{L^2(\rho)} = 0$$

Proof. The proof connects the approximation properties in the base space \mathcal{H}_k to the composite space \mathcal{H}_{k_h} through the mapping h . The composite RKHS \mathcal{H}_{k_h} consists of functions formed by composing elements from the base RKHS \mathcal{H}_k with the mapping h , that is, $\mathcal{H}_{k_h} = \{g(h(\cdot)) \mid g \in \mathcal{H}_k\}$. Thus, for any function $f \in \mathcal{H}_{k_h}$, there exists a corresponding function $g \in \mathcal{H}_k$ such that $f(x) = g(h(x))$ for all $x \in \mathbb{R}^d$.

To proceed, define a standard Random Fourier Feature (RFF) map for the base kernel $k(z, z')$ on \mathbb{R}^m :

$$\psi_D(z) := \sqrt{\frac{1}{D}} \begin{bmatrix} \cos(w_1^T z) \\ \vdots \\ \cos(w_D^T z) \\ \sin(w_1^T z) \\ \vdots \\ \sin(w_D^T z) \end{bmatrix},$$

where $w_i \stackrel{\text{i.i.d.}}{\sim} \mu$ and $b_i \stackrel{\text{i.i.d.}}{\sim} \text{Uniform}[0, 2\pi]$, with $\{w_i\}$ independent of $\{b_i\}$. The classic RFF approximation (Theorem 4) guarantees that the linear span of these features is dense in \mathcal{H}_k with respect to the L^2 norm under suitable measures. Let ρ_h be the pushforward probability measure of ρ under the map h . Then, for the function $g \in \mathcal{H}_k$,

$$\lim_{D \rightarrow \infty} \inf_{\theta \in \mathbb{R}^{2D}} \|g - \psi_D(x)^T \theta\|_{L^2(\rho_h)} = 0.$$

The goal is to show that $f(x)$ can be approximated by a function of the form $\psi(x)^T \theta$. We analyze the squared $L^2(\rho)$ norm of the error:

$$\|f(x) - \psi(x)^T \theta\|_{L^2(\rho)}^2 = \int_{\mathbb{R}^d} |f(x) - \psi(x)^T \theta|^2 d\rho(x).$$

Substituting $f(x) = g(h(x))$ and $\psi(x) = \psi_D(h(x))$ yields

$$\int_{\mathbb{R}^d} |g(h(x)) - \psi_D(h(x))^T \theta|^2 d\rho(x).$$

By the change of variables (or pushforward measure property), this integral equals

$$\int_{\mathbb{R}^m} |g(z) - \psi_D(z)^T \theta|^2 d\rho_h(z) = \|g - \psi_D^T \theta\|_{L^2(\rho_h)}^2.$$

From the density in $L^2(\rho_h)$, this error can be made arbitrarily small as $D \rightarrow \infty$ by choosing appropriate θ . Therefore, for any $f \in \mathcal{H}_{k_h}$,

$$\lim_{D \rightarrow \infty} \inf_{\theta \in \mathbb{R}^{2D}} \|f - \psi_D^T \theta\|_{L^2(\rho)} = 0,$$

completing the proof. \square

Lemma 2 (Expressive Power of the RFF-Enhanced Space). *When the number of the RFF features $D \rightarrow \infty$, the Feature Space \mathcal{H}_f is a subset of the closure of the Composite RFF Function Space \mathcal{H}_{RFF} with respect to the $\mathcal{L}^2(\Omega, \mathbb{R})$ norm.*

Proof. Let g_1 be an arbitrary function in \mathcal{H}_f . Since g_1 is a linear combination of the continuous functions in $h(x)$, g_1 is a continuous function on a compact set \mathcal{X} , i.e., $g_1 \in C(\mathcal{X})$. By Theorem 3, the composite kernel k_h is universal. Thus, its RKHS \mathcal{H}_{k_h} is dense in $C(\mathcal{X})$ under the uniform norm. This means that for any $\epsilon > 0$, there exists a function $f_{k_h} \in \mathcal{H}_{k_h}$ such that:

$$\|g_1 - f_{k_h}\|_{\infty} = \sup_{x \in \mathcal{X}} |g_1(x) - f_{k_h}(x)| < \frac{\epsilon}{2}.$$

For any probability measure ρ , the $L^2(\rho)$ norm is bounded by the L_{∞} norm, that is:

$$\begin{aligned} \|g_1 - f_{k_h}\|_{L^2(\rho)}^2 &= \int_{\mathcal{X}} |g_1(x) - f_{k_h}(x)|^2 d\rho(x) \\ &\leq \int_{\mathcal{X}} \left(\sup_{z \in \mathcal{X}} |g_1(z) - f_{k_h}(z)| \right)^2 d\rho(x) \\ &= \|g_1 - f_{k_h}\|_{\infty}^2 \int_{\mathcal{X}} d\rho(x) = \|g_1 - f_{k_h}\|_{\infty}^2. \end{aligned}$$

Thus we have $\|g_1 - f_{k_h}\|_{L^2(\rho)} \leq \|g_1 - f_{k_h}\|_{\infty} < \frac{\epsilon}{2}$.

From Theorem 5, the composite RFF space \mathcal{H}_{RFF} is dense in the RKHS \mathcal{H}_{k_h} under the $L^2(\rho)$ norm. Therefore, for our function f_{k_h} from the previous step, there exists a function $f_{\text{RFF}} \in \mathcal{H}_{\text{RFF}}$ such that:

$$\|f_{k_h} - f_{\text{RFF}}\|_{L^2(\rho)} < \frac{\epsilon}{2}.$$

Combining the results using the triangle inequality for the L^2 norm:

$$\|g_1 - f_{\text{RFF}}\|_{L^2(\rho)} \leq \|g_1 - f_{k_h}\|_{L^2(\rho)} + \|f_{k_h} - f_{\text{RFF}}\|_{L^2(\rho)} < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Since for any $g_1 \in \mathcal{H}_f$ and any $\epsilon > 0$, we have found an element $f_{\text{RFF}} \in \mathcal{H}_{\text{RFF}}$ that is ϵ -close in the L^2 norm, we have proven that $\mathcal{H}_f \subseteq \overline{\mathcal{H}_{\text{RFF}}}$. \square

D.2 Proof of Theorem 2 (Projection Error Comparison)

We have shown in Lemma 2 that the premise $\mathcal{H}_f \subseteq \overline{\mathcal{H}_{\text{RFF}}}$ holds. The proof now proceeds as follows.

For any function $g \in L^2$, define its L^2 approximation error with respect to f as $E(g) := \|f - g\|_{L^2}$. The function $E : L^2 \rightarrow \mathbb{R}$ is continuous, which follows from the reverse triangle inequality, establishing that the norm is a continuous function.

Let g_1 be an arbitrary element in \mathcal{H}_f . Since $\mathcal{H}_f \subseteq \overline{\mathcal{H}_{\text{RFF}}}$, by the definition of closure, there exists a sequence of functions $\{g_2^{(n)}\}_{n=1}^{\infty}$ in \mathcal{H}_{RFF} such that $g_2^{(n)} \rightarrow g_1$ in the L^2 sense, that is,

$$\lim_{n \rightarrow \infty} \|g_2^{(n)} - g_1\|_{L^2} = 0.$$

Because the error function $E(\cdot)$ is continuous, we can interchange the function with the limit:

$$\lim_{n \rightarrow \infty} E(g_2^{(n)}) = E\left(\lim_{n \rightarrow \infty} g_2^{(n)}\right) = E(g_1).$$

For each n , $g_2^{(n)}$ is an element of \mathcal{H}_{RFF} , so its error $E(g_2^{(n)})$ must be greater than or equal to the infimum of errors over \mathcal{H}_{RFF} :

$$\inf_{g \in \mathcal{H}_{\text{RFF}}} E(g) \leq E(g_2^{(n)}).$$

This inequality holds for all n . Taking the limit as $n \rightarrow \infty$ on both sides yields

$$\inf_{g \in \mathcal{H}_{\text{RFF}}} E(g) \leq \lim_{n \rightarrow \infty} E(g_2^{(n)}).$$

Substituting the continuity result gives

$$\inf_{g \in \mathcal{H}_{\text{RFF}}} E(g) \leq E(g_1).$$

The inequality holds for any arbitrary $g_1 \in \mathcal{H}_f$. This implies that $\inf_{g \in \mathcal{H}_{\text{RFF}}} E(g)$ is a lower bound for the set of values $\{E(g) \mid g \in \mathcal{H}_f\}$. By the definition of an infimum (greatest lower bound), this value must be less than or equal to the infimum of the set:

$$\inf_{g \in \mathcal{H}_{\text{RFF}}} E(g) \leq \inf_{g \in \mathcal{H}_f} E(g).$$

D.2.1 Proof of the universal approximation corollary

The following inequality holds:

$$\|u - u_{\omega, \theta}\|^2 \leq \|u - \tilde{u}_{\omega, W}\|^2 + \|\tilde{u}_{\omega, W} - u_{\omega, \theta}\|^2.$$

The universal approximation theorem by Hornik [1991] guarantees that for any $\epsilon > 0$, there exists p and ω_*, W_* such that $\|u - \tilde{u}_{\omega_*, W}\|^2 \leq \epsilon/2$. Theorem 2 ensures that there exists D and θ_* such that $\|\tilde{u}_{\omega_*, W} - u_{\omega_*, \theta_*}\|^2 \leq \epsilon/2$. Consequently, the norm between u and u_{ω_*, θ_*} is smaller than ϵ and that concludes the proof.

E IFeF-PINN Hyperparameters

This section details the hyperparameters used by the proposed IFeF-PINN method in each experiment (Table 3). Here, D denotes the number of Fourier-enhanced features; σ is the standard deviation of the sampled frequencies in random Fourier features (RFF); γ is the regularization parameter defined in Eq. equation 16; and Pre-training indicates a warm-up stage where a vanilla PINN is trained for several thousand epochs to provide a good initialization for basis selection in IFeF-PINN.

As discussed in Wang et al. [2021b], the selection of σ should align with the target function's frequency content. However, we fix $\sigma = 1$ across all cases in our experiments and obtain accurate approximations. In future work, a more detailed analysis of σ will be presented.

Problem	Pre-training	D	σ	λ_{LL}	γ
2D Helmholtz ($a_1 = 1, a_2 = 4$)	No	800	1	1e-2	1e-7
2D Helmholtz ($a_1 = a_2 = 100$)	No	2400	1	1e-7	1e-4
1D Convection ($\beta = 50$)	Yes	800	1	1e-2	1e-7
1D Convection ($\beta = 200$)	Yes	1600	1	1e-2	1e-4/1e-7
Viscous Burgers ($\nu = \frac{0.01}{\pi}$)	Yes	800	1	1e-1	0
Convection-Diffusion ($k_{low} = 4, k_{high} = 60$)	Yes	800	1	1e-2	1e-7

Table 3: Hyperparameters setting for IFe F-PINN under each experiment

F Experiment Setup

F.1 PDEs Setup

In this section, we provide detailed PDE settings used as our benchmarks.

2D Helmholtz Equation. The Helmholtz equation is an elliptic PDE that commonly arises in the study of wave propagation, acoustics, and electromagnetic fields. We consider the 2D Helmholtz equation as follows:

$$\begin{aligned} \nabla^2 u + u &= f, & (x, y) \in \Omega, \\ u(x, y) &= 0, & (x, y) \in \partial\Omega, \end{aligned} \quad (26)$$

corresponding to a source term

$$f(x, y) = -\pi^2 (a_1^2 \sin(a_1 \pi x) \sin(a_2 \pi y) - a_2^2 \sin(a_1 \pi x) \sin(a_2 \pi y)) + \sin(a_1 \pi x) \sin(a_2 \pi y).$$

The parameters a_1 and a_2 define the frequency of the analytic solution. We will investigate the following different frequency cases:

- low-frequency: $a_1 = 1, a_2 = 4, \Omega = [-1, 1] \times [-1, 1]$.
- high-frequency: $a_1 = 100, a_2 = 100, \Omega = [0, 0.2] \times [0, 0.2]$.

The high-frequency case uses a reduced domain to maintain computational tractability. By the Nyquist-Shannon sampling criterion Shannon [2006], resolving such high-frequency oscillations over a larger domain would require prohibitively dense collocation. Despite the smaller domain, the configuration covers 10×10 wavelengths, capturing the extreme oscillatory behavior.

1D Convection Equation. The Convection equation is a hyperbolic PDE that describes the movement of a substance through fluids. We consider the periodic boundary conditions system as follows:

$$\begin{aligned} \frac{\partial u}{\partial t} + \beta \frac{\partial u}{\partial x} &= 0, & (t, x) \in [0, 1] \times [0, 2\pi], \\ u(x, 0) &= \sin x, \\ u(0, t) &= u(2\pi, t). \end{aligned} \quad (27)$$

The closed-form solution is $u(x, t) = \sin(x - \beta t)$. We consider a low-frequency case $\beta = 50$ and a high-frequency case $\beta = 200$ on the same domain.

1D Convection-Diffusion Equation. The Convection–Diffusion equation is a parabolic PDE that models the combined effects of transport by fluid motion and spreading due to diffusion. We consider the multi-scale system with periodic boundary conditions as follows:

$$\begin{aligned} \frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} &= d \frac{\partial^2 u}{\partial x^2}, & (t, x) \in [0, 1] \times [0, 1], \\ u(t, 0) &= u(t, 1), \\ \frac{\partial u}{\partial x}(t, 0) &= \frac{\partial u}{\partial x}(t, 1), \\ u(0, x) &= A_{low} \sin(k_{low} x) + A_{high} \sin(k_{high} x). \end{aligned} \quad (28)$$

The analytic multi-scale solution is

$$u(t, x) = A_{low} e^{-dk_{low}^2 t} \sin(k_{low}(x - ct)) + A_{high} e^{-dk_{high}^2 t} \sin(k_{high}(x - ct)).$$

To set a multi-scale problem consists of both low- and high-frequency components, the parameters are chosen as follows:

$$c = 1, d = 0.00005, A_{\text{low}} = 1, A_{\text{high}} = 0.1, k_{\text{low}} = 4\pi, k_{\text{high}} = 60\pi.$$

Viscous Burgers' Equation. The Viscous Burgers' equation is a nonlinear parabolic PDE that models fluid motion by combining convection and diffusion effects. We consider the nonlinear system as follows:

$$\begin{aligned} \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} &= \nu \frac{\partial^2 u}{\partial x^2}, \quad (t, x) \in [0, 1] \times [-1, 1], \\ u(0, x) &= -\sin(\pi x), \\ u(t, -1) &= u(t, 1) = 0, \end{aligned} \tag{29}$$

where $\nu = \frac{0.01}{\pi}$.

F.2 Spectrum Analysis Setup

Given frequencies $\kappa = \{f_i\}_{i=1}^{i=10} = \{1, 2, 5, 10, 30, 40, 50, 60, 70, 80\}$, where all amplitudes are chosen as $A_i = 1$, we consider the Convection equation in 27 with $\beta = 1$ and initial condition as follows:

$$u(x, 0) = \sum_{i=1}^{10} A_i \sin(2\pi f_i x).$$

The corresponding analytic solution is then given by

$$u(x, t) = \sum_{i=1}^{10} A_i \sin(2\pi f_i (x - t)).$$

The problem domain is defined as $(t, x) \in [0, 1] \times [0, 1]$. Our objective is to evaluate and compare the ability of Vanilla PINNs and Fourier-enhanced Features to capture all frequency components at $t = 0$.

For the neural network architecture, we employ an 8-layer fully connected network with tanh activation functions and 64 neurons per layer. The training data consist of 201 uniformly sampled points along the two spatial boundaries ($x = 0$ and $x = 1$) and at the final time ($t = 1$). In addition, 201×201 collocation points are uniformly sampled within the interior domain to enforce the physics constraints.

To further design this experiment as an ablation study and demonstrate that the incorporation of Fourier-enhanced Features for basis extension is a necessary component of our proposed IFeF-PINN, we discard the iterative training procedure and retain only the basis extension step. Specifically, we first train the Vanilla PINN for 40,000 epochs using the Adam optimizer with a learning rate of 10^{-3} . We then extend the basis with varying numbers of Fourier-enhanced Features, $D_j \in \{400, 800, 1600, 2400, 3200, 4000\}$, and then solve the lower-level problem defined in Equation 6. Moreover, to ensure a more rigorous analysis, we impose the relation $B_{D_i} \subset B_{D_j}$ whenever $D_j > D_i$, so that the RFF mapping matrices are nested.

F.3 Model Setup

This section details the model setup for all baselines. Unless otherwise specified, we use a multi-layer perceptron (MLP) whose depth and width are determined by the experimental setting. For the 2D Helmholtz case ($a_1 = 1, a_2 = 4$), we follow the network structure in Barreau and Shen [2025]; for the viscous Burgers' equation, we follow Raissi et al. [2019]. For PINNsformer [Zhao et al., 2023] and PIG [Kang et al., 2024], since they both have special network architectures, we adopt the original architecture. All experiments use the tanh activation function and the Adam optimizer with a learning rate of 10^{-3} for the network parameters. For IFeF-PD, we adopt the Primal-Dual weight balancing strategy proposed in Barreau and Shen [2025], and optimize the dynamic physics weight with the same setting for Adam at a learning rate of 10^{-4} . The network architectures used in each experiment are summarized in Table 4.

F.4 Dataset setup

In this section, we detail the dataset setup for each equation and experiment. For the 1D Convection equation, 2D Helmholtz equation (low-frequency), and Convection-Diffusion equation, we follow the setting and strategy of Zhao et al. [2023]. For the Viscous Burgers' Equation, we follow Raissi et al. [2019]. The detailed settings are summarized in Table 5.

Problem	Hidden layers	Hidden width
2D Helmholtz ($a_1 = 1, a_2 = 4$)	3	[50,50,20]
2D Helmholtz ($a_1 = a_2 = 100$)	6	64
1D Convection ($\beta = 50$)	6	64
1D Convection ($\beta = 200$)	6	64
Viscous Burgers ($\nu = \frac{0.01}{\pi}$)	8	20
Convection-Diffusion ($k_{\text{low}} = 4, k_{\text{high}} = 60$)	6	64

Table 4: Network architecture for all problems.

Problem	Sampling	Boundary points	Physics points
2D Helmholtz ($a_1 = 1, a_2 = 4$)	Uniform	1000	71×71
2D Helmholtz ($a_1 = a_2 = 100$)	LHS	3000	23000
1D Convection ($\beta = 50$)	Uniform	404^\dagger 204^\ddagger	$(101 \times 101)^\dagger$ $(51 \times 51)^\ddagger$
1D Convection ($\beta = 200$)	Uniform	404	101×101
Viscous Burgers ($\nu = \frac{0.01}{\pi}$)	LHS	100	10000
Convection-Diffusion ($k_{\text{low}} = 4, k_{\text{high}} = 60$)	Uniform	404	101×101

Table 5: Dataset settings for each PDE problem.

Notes: † Vanilla/NTK/PIG; ‡ PINNsformer/IFeF.

G Experimental Results

In this section, we present the true solutions, model predictions, and absolute error maps for all baselines considered in our numerical experiments. Results for the viscous Burgers’ equation, the low- and high-frequency convection equations, and the multi-scale convection-diffusion equation are shown in separate figures. For clarity, each figure contains three panels: (i) the true solution, (ii) the model prediction, and (iii) the absolute error on a log10 scale.

The case presented in Figure 9 is particularly interesting. We observe that the analytical solution is the sum of the two frequencies (20 and 60). If we zoom in on the error plot, it is possible to see that for all other methods than IFeF, the high-frequency component is not caught. Despite visually similar plots, the error can be quite large. This phenomenon does not appear with IFeF-PINN.

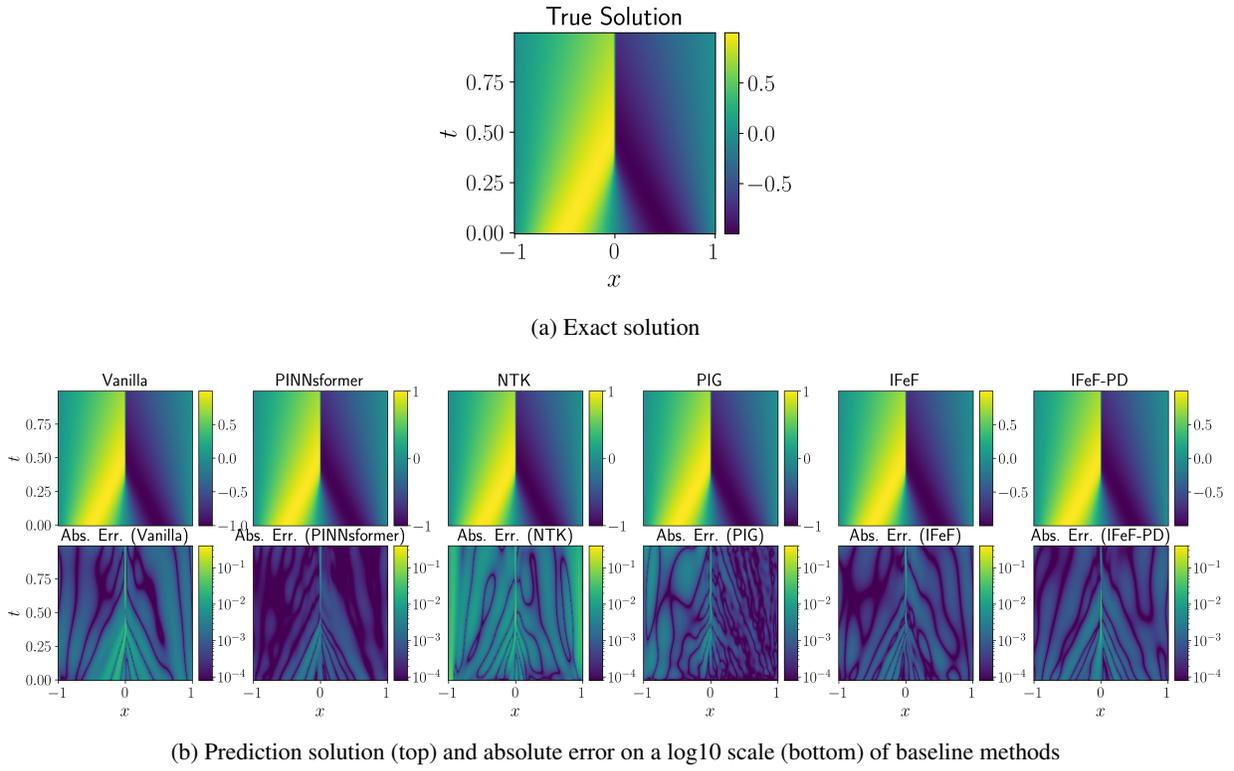


Figure 6: True solution, prediction and absolute error of baseline methods for viscous Burgers' equation

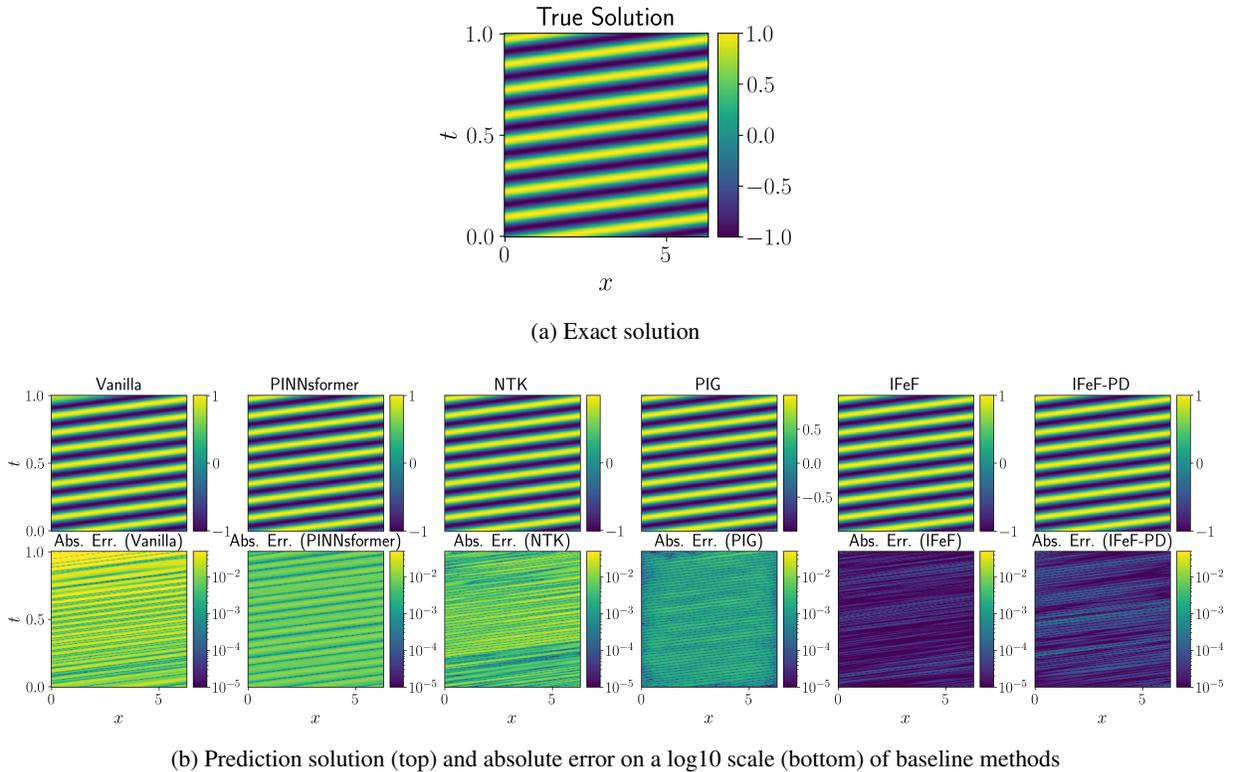


Figure 7: True solution, prediction, and absolute error of baseline methods for low-frequency convection equation

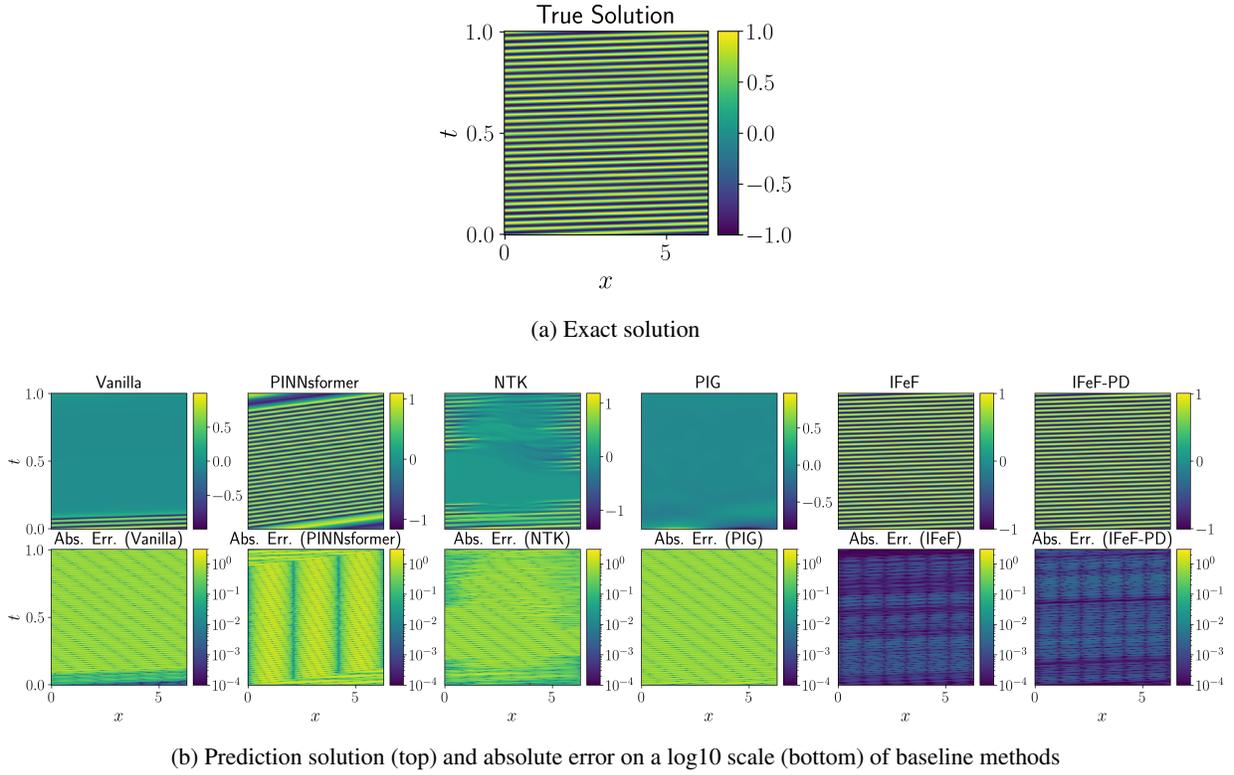


Figure 8: True solution, prediction, and absolute error of baseline methods for high-frequency convection equation

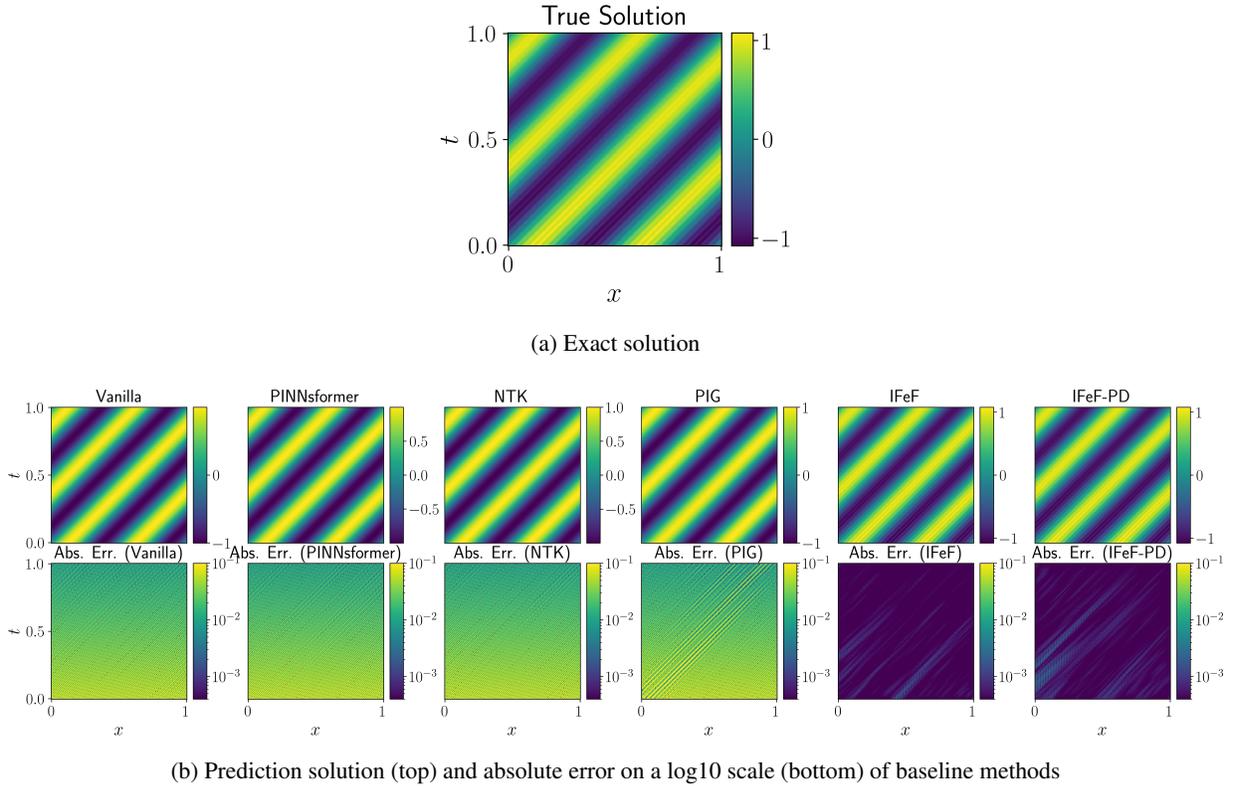


Figure 9: True solution, prediction, and absolute error of baseline methods for multi-scale convection-diffusion equation