

# TGPT-PINN: Nonlinear model reduction with transformed GPT-PINNs

Yanlai Chen<sup>\*</sup>   Yajie Ji<sup>†</sup>   Akil Narayan<sup>‡</sup>   Zhenli Xu<sup>§</sup>

## Abstract

We introduce the *Transformed* Generative Pre-Trained Physics-Informed Neural Networks (TGPT-PINN) for accomplishing nonlinear model order reduction (MOR) of transport-dominated partial differential equations in an MOR-integrating PINNs framework. Building on the recent development of the GPT-PINN that is a network-of-networks design achieving snapshot-based model reduction, we design and test a novel paradigm for nonlinear model reduction that can effectively tackle problems with parameter-dependent discontinuities. Through incorporation of a shock-capturing loss function component as well as a parameter-dependent transform layer, the TGPT-PINN overcomes the limitations of linear model reduction in the transport-dominated regime. We demonstrate this new capability for nonlinear model reduction in the PINNs framework by several nontrivial parametric partial differential equations.

**Key words:** Nonlinear model order reduction, Physics-informed neural networks, Meta-learning, Reduced basis method, Parametric systems

## 1 Introduction

Reduced order models (ROMs) are mainstays in computational science, playing an integral role in design optimization, uncertainty quantification, and in the construction of digital twins. ROMs can accelerate the evaluation of one-shot computational models in time-dependent settings, and are also capable of substantial acceleration in multi-query modeling for, e.g., parametric problems. We investigate ROMs for partial differential equation (PDE) models in this paper. Although nonlinear model reduction has gained attention in recent years, most of the established ROM theory and algorithms focus on linear model reduction techniques that are known to be effective for diffusion-dominated PDE problems. However, such linear reduction approaches frequently fail when applied to PDEs that involve transport-dominated phenomena. This failure arises from a fundamental mathematical limitation: the slow decay of the Kolmogorov  $n$  width [27] for transport problems with discontinuities [22]. Many recent advances in ROMs have therefore sought nonlinear reduction strategies that are not bound by the limitations of linear reduction.

Recent years have witnessed a wealth of nonlinear reduction methods being developed. The first set of approaches involve transforms of the trial basis [7, 18, 21, 29–32]. While these transforms improve the approximation property of the surrogate space, they tend to rely on additional knowledge about the problem and the transforms are typically not computable in

---

<sup>\*</sup>Department of Mathematics, University of Massachusetts Dartmouth, North Dartmouth, MA 02747. Email: [yanlai.chen@umassd.edu](mailto:yanlai.chen@umassd.edu). Y. Chen is partially supported by National Science Foundation grant DMS-2208277 and by the UMass Dartmouth MUST program, N00014-22-1-2012, established by Dr. Ramprasad Balasubramanian via sponsorship from the Office of Naval Research.

<sup>†</sup>School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai 200240, China. Email: [jiyajie595@sjtu.edu.cn](mailto:jiyajie595@sjtu.edu.cn).

<sup>‡</sup>Scientific Computing and Imaging Institute and Department of Mathematics, University of Utah, Salt Lake City, UT 84112. Email: [akil@sci.utah.edu](mailto:akil@sci.utah.edu). A. Narayan is partially supported by NSF DMS-1848508 and AFOSR FA9550-23-1-0749.

<sup>§</sup>School of Mathematical Sciences, CMA-Shanghai and MOE-LSC, Shanghai Jiao Tong University, Shanghai 200240, China. Email: [xuzl@sjtu.edu.cn](mailto:xuzl@sjtu.edu.cn). Z. Xu acknowledges the support from the NSFC (grant No. 12325113) and the HPC center of Shanghai Jiao Tong University.

an automatic way. The second set adopts neural networks as a component of its algorithm [10, 12, 16, 17, 19, 23], for which Franco *et al.* [12] proposed to approximate the solution map by a neural network from parameter to the reduced representation of the solutions trained by a deep autoencoder with theoretical analysis for the choice of its ROM dimensions. While being more general, these methods tend to be purely data driven, attempting to capture low-dimensional structure in a latent space representation, with large amount of data and various degrees of success. The third is the optimal transport, specifically the Wasserstein metric, based approaches [3, 11, 15]. While still nascent and more limited in applicability, it provides a unique perspective grounded in a mathematically sound theory. Another class of approaches use adaptive techniques where linear subspaces are progressively updated, for example in time-dependent problems [25, 26, 30]. Some existing work also use a specific types of nonlinear ansatz (such as judicious combinations of polynomial-type functions and neural network functions) to inject nonlinear dynamics into a reduced model [1, 13].

This paper makes advances that sit at the intersection of physics-informed machine learning and nonlinear reduced order modeling. We introduce the Transformed Generative Pre-Trained Physics-Informed Neural Networks (TGPT-PINN), a framework that extends Physics-Informed Neural Networks (PINNs) and reduced basis methods (RBM) to the nonlinear model reduction regime while maintaining the type of network structure and the unsupervised nature of its learning. By introducing a novel transform layer in a GPT-PINN, we enable the resolution of parameter-dependent discontinuity locations. Through the use of PINNs in snapshot-based model reduction, we eliminate the need for intrusive analysis and manipulation of existing PDE solvers; in particular, the “online” phase of the multi-query contexts leverages a network of pre-trained PINN-type networks, and hence all that must be explicitly manipulated is a PDE residual term in strong form and a transform layer capturing the discontinuity. Moreover, the determination of the network parameters in the transform layer are automatically achieved via back-propagation simultaneously with weights of the pre-trained PINNs serving as the super neurons of the reduced network. Lastly, we emphasize that our proposed approach is *data-sparse*. In fact, the number of full-order queries needed of the TGPT-PINN is equal to the latent space dimension which is much smaller than the number of queries needed to train the identity map of the autoencoder [12, 19]. Because the TGPT-PINN is an instance of snapshot-based model reduction, it inherits the standard limitation of such approaches, namely that the accuracy is limited by that of the underlying snapshot solutions (here PINN solutions) when the exact solution to the PDEs is unknown. Our numerical results show that, as we improve the underlying PINN solutions, our TGPT-PINN solution becomes more accurate too. This confirms the effectiveness of the TGPT-PINN’s reduction strategy, while highlighting its dependence on accuracy of snapshots, which is a limitation similar to any other snapshot-based reduced order model.

We demonstrate the efficacy of the TGPT-PINN on several nontrivial examples. Its efficacy on explicit parametric functions is first investigated to reveal the expressive strength of the procedures for nonlinear model reduction. Next, the TGPT-PINN is applied to solutions of parametric PDEs, where explicit solutions are not available and are computed instead through a PINN-type formulation. We observe in this case as well that the TGPT-PINN is very effective at accomplishing model reduction, with very little *a priori* knowledge about a PDE, or even whether or not the solution has discontinuities.

As a novel effort in developing nonlinear reduction strategies, the main advance of our proposed method is an extension of the newly developed GPT-PINN [8] to the nonlinear reduction regime by integrating shifts and transforms into the neural network/GPT-PINN framework. This is achieved by adding a *transform layer* to the developed GPT-PINN. The benefit of this integration is two-fold. First, training inherits the greedy algorithm-based model reduction capability of the GPT-PINN automatically. Second, the addition of the transform layer allows simultaneous training of the shift-and-transform network parameters of the transform layer and the mode coefficients of the GPT-PINN layer. A particularly important practical improvement is that the involved neural network is also hyper-reduced, i.e. the architecture has orders of magnitude fewer trainable parameters than those in [10, 12, 16, 17, 19].

The newly proposed approach is tested in two steps. First, to isolate any influence due to choice of a numerical (“truth”) solver, we test the case when the parametric function set is analytically given. We show that, for the more challenging case when the Kolmogorov width decays slowly due to a moving kink or discontinuity, the TGPT-PINN is able to capture the parameter dependence exactly by one or a handful of neurons (while linear approaches, such as the Empirical Interpolation Method (EIM) [2], fail to be effective). Next, we investigate when  $u(x, \mu)$  is the numerical solution to a parametric PDE, which involves usage of a numerical solver. We show that, on examples when linear reduction does not work, the TGPT-PINN *achieves machine accuracy with one neuron* (e.g., the transport equation with a discontinuous initial condition whose speed is a parameter). On examples when linear reduction does work well, the TGPT-PINN works by employing only one (nonlinear reaction) or three (nonlinear reaction diffusion) neurons while achieving better accuracy than the GPT-PINN with 10 neurons. This significant improvement is a manifestation of the novel design of the TGPT-PINN, i.e., the addition of a trainable transform layer to a network of pre-trained networks.

The rest of this paper is organized as follows. Background materials including a motivating example, the PINN and GPT-PINN are briefly presented in Section 2. The main algorithm is given in Section 3 with numerical results on three classes of analytically-given functions and three types of equations detailed in Section 4. We draw conclusions in Section 5.

## 2 Background

This section is devoted to a brief introduction of the two ingredients leading to the TGPT-PINN, namely snapshot transformations and the GPT-PINN for neural network based model order reduction for parametric PDEs.

### 2.1 The Kolmogorov $n$ -width and barrier

Let  $U$  be a Banach space, and let  $K$  be a set (or “manifold”) of functions that we wish to approximate using elements of an  $n$ -dimensional subspace. The theoretically optimal performance of such *linear* reduced order models is quantified by the Kolmogorov  $n$ -width [27], which measures how well elements in  $K \subset U$  can be best approximated from a dimension- $n$  linear subspace  $U_n \subset U$ ,

$$d_n(K) = \inf_{\dim U_n = n} \sup_{u \in K} \inf_{\phi \in U_n} \|u - \phi\|_U.$$

The class  $K$  can be a set of functions that are encoded with a parameterization; in our setting a more salient example is to consider  $K$  as the solution manifold of a (parametric) partial differential equation,  $K = \{u_\mu \mid \mu \in \mathcal{D}\} \subset U$ , where  $\mu \mapsto u_\mu$  is the  $\mu$ -parameterized PDE solution operator, and  $\mathcal{D}$  is frequently a subset of  $\mathbb{R}^k$  for some  $k \in \mathbb{N}$ .

The  $n$ -width  $d_n(K)$  provides an accuracy lower bound for all reduced order models that rely on linear reduction, i.e., by building approximations from a(ny) dimension- $n$  subspace  $U_n$ . For example, the  $n$ -width is the best possible error when building  $U_n$  from snapshots or proper orthogonal decomposition (POD) modes. For many parametric PDEs that are elliptic or parabolic in nature,  $d_n(K)$  is known to decay exponentially in  $n$  [9]. For such cases, constructive (i.e., computationally feasible) greedy methods have been proven to generate reduced spaces that asymptotically match the Kolmogorov  $n$ -width [5, 6]. This fast decay of the  $n$ -width along with the development of effective greedy algorithms are the driving forces behind linear model reduction and its descendants, such as RBM and GPT-PINN, respectively.

However, when facing solution manifolds generated by transport-dominated problems, the Kolmogorov  $n$ -width decreases much slower. Ohlberger and Rave [22] proved that  $d_n(K) \sim n^{-1/2}$  where  $K$  is the solution manifold of a (relatively simple) transport dominated problem. This extremely slow decay for transport-dominated problems is the fundamental

reason that linear model order reduction methods are ineffective in this case, and motivates the need to develop and exercise *nonlinear* model reduction approaches.

## 2.2 A motivating example for transformed snapshots

Consider the following example from [32],

$$u(x, \mu) := \psi\left(\frac{x}{0.4 + \mu} - 1\right), \quad \psi(x) := \begin{cases} \exp\left(-\frac{1}{1-x^2}\right) & -1 \leq x < -\frac{1}{2}, \\ 0 & \text{else.} \end{cases}$$

The function  $\psi(x)$  is a discontinuous function at  $x = -1/2$ , which makes  $u(x, \mu)$  discontinuous at  $x = (\mu + 0.4)/2$ . We are therefore in a situation that the location of the discontinuity depends on the parameter. Interpolatory type techniques in the parameter  $\mu$ , including classical polynomial interpolations and the EIM which is a standard tool in model order reduction for the nonlinear and nonaffine setting, have limited effectiveness as demonstrated in Figure 1. However, if we change the interpolation ansatz

$$u(x, \mu) \approx \sum_{i=0}^n \ell_i(\mu) u(x, \mu^i) \quad \longrightarrow \quad u(x, \mu) \approx \sum_{i=0}^n \ell_i(\mu) u(\phi(\mu, \mu^i)(x), \mu^i), \quad (2.1)$$

where  $\ell_i$  are any basis (e.g., cardinal Lagrange interpolants associated to the  $\mu^i$ ), we can design  $\phi$  to eliminate the staircasing behavior. Here,  $\phi(\mu, \eta)$  represents a transform from the “source” parameter  $\eta$  to the “target” parameter  $\mu$  so that the jump locations match. For this example, the choice

$$\phi(\mu, \eta)(x) = \frac{0.4 + \eta}{0.4 + \mu} x,$$

would enable exact reconstruction with  $n = 0$ , i.e.,  $u(\phi(\mu, \eta)(x), \eta) = u(x, \mu)$ . This shows that sufficiently complicated transformation maps  $\phi$  can restore expressive power in reduced order models. However, we make a further observation that the even simpler map,

$$\phi(\mu, \eta)(x) = x + \frac{\eta - \mu}{2},$$

which is linear in  $(\eta, \mu)$ , is enough to ensure that the discontinuity locations are exactly captured, and hence even relatively simple maps  $\phi$  can provide enormous benefit. The proposed TGPT-PINN procedure attempts to identify the map  $\phi$  automatically. As a result, the TGPT-PINN approximation of this function is indistinguishable from the exact solution while the standard approaches produce noticeable staircases leading to large interpolation errors, see Figure 1.

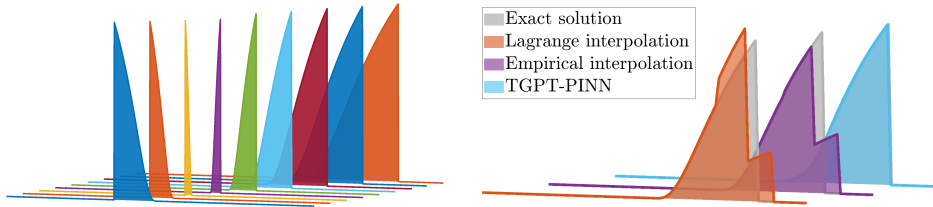


Figure 1: Shown on the left are 9 snapshots of  $u(x, \mu)$  with various  $\mu$  values. On the right are the function at  $\mu = 1.0$  and its three kinds of interpolations, the polynomial interpolation from snapshots with  $\mu = 0.5, 0.85$  and  $1.2$ , the EIM approximation, and the TGPT-PINN interpolation.

## 2.3 PINN and GPT-PINN

Physics-informed neural networks (PINNs), popularized by Raissi *et al.* [28], have emerged as an increasingly popular alternative to traditional numerical methods for PDEs in recent



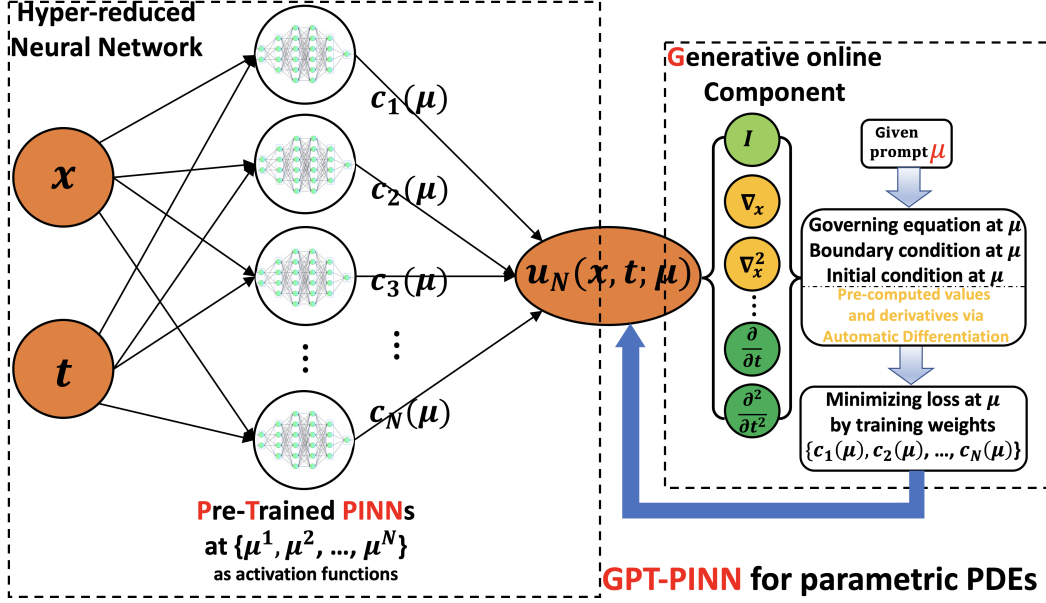


Figure 2: The GPT-PINN architecture [8]. A hyper-reduced network adaptively embeds pre-trained PINNs at the nodes of its sole hidden layer. It then allows a quick online generation of a surrogate solution at any given parameter value.

years. They adopt Deep Neural Networks (DNNs) to approximate the solutions of PDEs while incorporating a strong physics-based PDE prior encoded into the loss function to constrain the output of the DNN. In comparison to traditional numerical solvers, the advantages of PINNs include that they are able to numerically solve the PDE without discretizing the spatiotemporal domain, and that they can leverage automatic differentiation [4, 24] to algorithmically minimize the residual in the loss. However, PINNs have some weaknesses. For example, training a vanilla PINN is usually significantly slower than employing a classic numerical method to solve the corresponding PDE. To address this shortcoming, the Generative Pre-Trained PINNs (GPT-PINNs) were developed [8] as a meta-learning approach for parametric systems to reduce the architecture and corresponding number of trained parameters in a PINN, thereby reducing the computational time required to solve parametric PDEs with PINNs. The GPT-PINN requires an initial (“offline”) investment cost dedicated to learning the parametric dependence during the offline stage, which is guided by a mathematically reliable greedy algorithm. With this initial investment, The GPT-PINN is capable of providing significant computational savings in the multi-query and real-time settings thanks to the fact that their marginal cost is of orders of magnitude lower than that of an individual PINN solve [8].

The GPT-PINN architecture, depicted in Figure 2, is a network-of-networks, where activation functions in the sole hidden layer of the outer network are chosen in a customized way. The inner networks, defining the activation functions, are full pre-trained PINNs instantiated by the PDE solutions at a set of adaptively-selected parameter values  $\{\mu^1, \mu^2, \dots, \mu^n\}$  chosen by a greedy algorithm. The corresponding outer-/meta-network is hyper-reduced in comparison to the inner networks, having only one hidden layer. Moreover, this meta layer adaptively “learns” the parametric dependence of the system and can “grow” its hidden layer one super neuron/network at a time. The GPT-PINN is capable of generating approximate solutions for the parametric system across the entire parameter domain accurately and efficiently, with a cost independent of the size of the full PINN.

### 3 The TGPT-PINN algorithm

The insight provided by Section 2.2 is that, to achieve nonlinear model order reduction, one way is to compose the snapshot with a *parameter-dependent transformation*. This is the basic idea of the TGPT-PINN whose schematic design is shown in Figure 3. It seeks to approximate  $u(\mathbf{x}, t; \mu)$  as follows,

$$u(\mathbf{x}, t; \mu) \approx \sum_{i=1}^N c_i(\mu) u(T_{\mu, \mu^i}(\mathbf{x}, t); \mu^i). \quad (3.1)$$

All the quantities,  $N$ ,  $c_i(\mu)$ ,  $\mu^i$ ,  $T_{\mu, \mu^i}$  and  $u(\mathbf{x}, t; \mu^i)$ , are obtained through training in a two-step procedure. The first step is an “offline” step that trains  $\mu^i$ ,  $N$  and  $u(\mathbf{x}, t; \mu^i)$ ; this step can be expensive as we require well-resolved solutions  $u$  to PDEs and we must sweep over the parameter domain  $\mathcal{D} \subset \mathbb{R}^k$  to identify parameter values  $\mu^i$ . The second step is an “online” step that, given an arbitrary  $\mu$ , computes  $T_{\mu, \mu^i}$  and  $c_i(\mu)$  for  $i \in [N] = \{1, \dots, N\}$ . This second step is much more computationally efficient, training a neural network with only  $N(d^2 + 3d + 3)$  degrees of freedom, i.e., linear in the number of snapshots  $N$  and *substantially* fewer degrees of freedom than required for even a single snapshot  $u$ . More explicitly, the TGPT-PINN is the following two-step procedure:

$$\begin{array}{c} i = 1, \mu^1, \text{“Offline”}: \quad \text{PDE/residual} \xrightarrow{\text{Minimize } \mathcal{L}_{\text{PINN}}} u(\cdot, \cdot; \mu^i) \xrightarrow{\text{Sufficient accuracy}} N := i \\ \quad \quad \quad \uparrow \quad \quad \quad \downarrow \text{Insufficient accuracy} \\ \quad \quad \quad \mu^i \quad \quad \quad i += 1 \\ \quad \quad \quad \longleftarrow \text{Search parameter space} \end{array}$$

$$\text{“Online”}: \quad \mu \xrightarrow{\text{Minimize } \mathcal{L}_{\text{PINN}}^{\text{TGPT}}} c_i(\mu), T_{\mu, \mu^i}(\cdot, \cdot), i \in [N] \xrightarrow{(3.1)} u(\cdot, \cdot; \mu)$$

The “online” step is visualized in Figure 3. The main strength of this approach is that function composition is natural and straightforward in neural networks - simply adding a layer or block, and the network parameters defining the transformation can be trained together with the mode coefficients  $\{c_i(\mu)\}_{i=1}^N$  in the GPT-PINN block of the network. The loss functions  $\mathcal{L}_{\text{PINN}}$  and  $\mathcal{L}_{\text{PINN}}^{\text{TGPT}}$  are described in the coming sections. All minimization steps are performed using somewhat standard neural network optimization and back-propagation procedures.

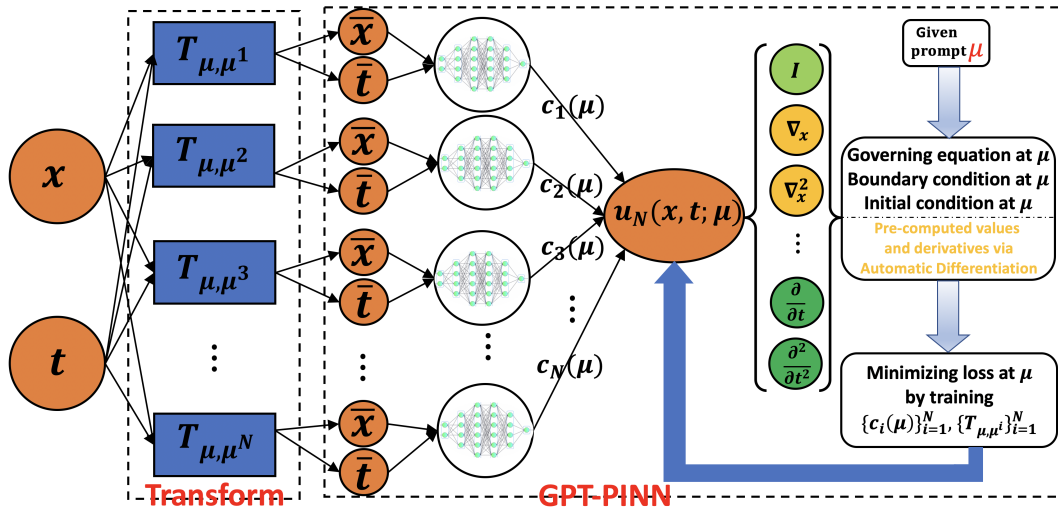


Figure 3: The TGPT-PINN design schematic. For any given parameter value  $\mu$ , a  $\mu$ -dependent loss is constructed and the coefficients  $c_j(\mu)$  and the weights and biases in  $T_{\mu, \mu^i}$  are trained.

### 3.1 PDE formulation

We define our problem to be the following time-dependent PDE on the spatial domain  $\Omega \subset \mathbb{R}^d$  with boundary  $\partial\Omega$ , and parametric domain  $\mu \in \mathcal{D}$ :

$$\begin{aligned} \frac{\partial}{\partial t} u(\mathbf{x}, t; \mu) + \mathcal{F}[u(\mathbf{x}, t; \mu)] &= 0, \quad \mathbf{x} \in \Omega, \quad t \in [0, T], \quad \mu \in \mathcal{D} \\ \mathcal{G}(u)(\mathbf{x}, t; \mu) &= 0, \quad \mathbf{x} \in \partial\Omega, \quad t \in [0, T], \quad \mu \in \mathcal{D} \\ u(\mathbf{x}, 0; \mu) &= u_0(\mathbf{x}; \mu), \quad \mathbf{x} \in \Omega, \quad \mu \in \mathcal{D}. \end{aligned} \quad (3.2)$$

Here  $\mathcal{F}$  is a differential operator and  $\mathcal{G}$  denotes a boundary operator [8]. Our goal is to compute the numerical solution to this PDE, which we shall accomplish using the TGPT-PINN methodology.

The remainder of this section is devoted to describing the particulars of the TGPT-PINN. In Section 3.2 we will introduce an approach that modifies the “offline” portion of the TGPT-PINN to enhance its ability to resolve discontinuities in the solution for transport-dominated problems.

### 3.2 Offline: The PINN solver

We discuss the “offline” procedure of the TGPT-PINN that computes  $u(\cdot, \cdot; \mu)$  for a given, fixed  $\mu$ . In particular, we will use  $\mu = \mu^i$  in  $u$  for the TGPT-PINN. Because  $\mu$  is fixed in this section, we frequently omit writing the  $\mu$ -dependence in what follows. The idealized loss function that we use corresponds to a standard PINN loss function,

$$\mathcal{L}(u) = \mathcal{L}_{\text{int}}(u) + \varepsilon_i \int_{\Omega} \|u(\mathbf{x}, 0) - u_0(\mathbf{x})\|_2^2 d\mathbf{x} + \varepsilon_b \int_{\partial\Omega \times [0, T]} \|\mathcal{G}(u)(\mathbf{x}, t)\|_2^2 d\mathbf{x} dt, \quad (3.3)$$

which is constructed by summing losses corresponding to the residual of the PDE, the initial values and the boundary conditions, and the domain-interior loss is given by,

$$\mathcal{L}_{\text{int}}(u) := \int_{\Omega \times (0, T)} \left\| \lambda(\mathbf{x}, t) \left( \frac{\partial}{\partial t} u(\mathbf{x}, t) + \mathcal{F}(u)(\mathbf{x}, t) \right) \right\|_2^2 d\mathbf{x} dt.$$

We choose the weighting constants as  $\varepsilon_i = 1$  and  $\varepsilon_b = 1$ . In practice, PINNs-type losses can suffer from the exploding or vanishing gradient problem when solutions are discontinuous. The factor  $\lambda$  above is introduced to ameliorate this effect, and is designed to attenuate the loss when the gradient of the solution is large. We therefore introduce the following “shock-capturing weighting” factor in the calculation of the residual of the PDE, which was originally proposed in Liu *et al.* [20],

$$\lambda(\mathbf{x}, t) = \frac{1}{\varepsilon_{\lambda} |\nabla \cdot u| + 1}.$$

We choose  $\varepsilon_{\lambda} = 0.1$ . The PINNs approach chooses a discretization for  $u$  and minimizes the loss function through optimization. Like a standard PINN, we use a fully connected neural network  $\psi_{\text{NN}}^{\mu}(\mathbf{x}, t)$  to approximate  $u$ , where the weights and biases of  $\psi_{\text{NN}}^{\mu}$  are optimized. The loss  $\mathcal{L}$  is approximated through a sampling/quadrature-type procedure. In particular, the actual discretized loss that is optimized is,

$$\begin{aligned} \mathcal{L}_{\text{PINN}}(\psi_{\text{NN}}^{\nu}) &= \frac{1}{|\mathcal{C}_o|} \sum_{(\mathbf{x}, t) \in \mathcal{C}_o} \left\| \lambda \cdot \left( \frac{\partial}{\partial t} (\psi_{\text{NN}}^{\nu})(\mathbf{x}, t) + \mathcal{F}(\psi_{\text{NN}}^{\nu})(\mathbf{x}, t) \right) \right\|_2^2 \\ &+ \varepsilon_b \cdot \frac{1}{|\mathcal{C}_{\partial}|} \sum_{(\mathbf{x}, t) \in \mathcal{C}_{\partial}} \|\mathcal{G}(\psi_{\text{NN}}^{\nu})(\mathbf{x}, t)\|_2^2 + \varepsilon_i \cdot \frac{1}{|\mathcal{C}_i|} \sum_{\mathbf{x} \in \mathcal{C}_i} \|\psi_{\text{NN}}^{\nu}(\mathbf{x}, 0) - u_0(\mathbf{x})\|_2^2. \end{aligned} \quad (3.4)$$

Above, we sample collocation/training points in certain fashion (described below) from the PDE spatiotemporal domain  $\mathcal{C}_o \subset \Omega \times (0, T)$ , spatiotemporal boundary  $\mathcal{C}_{\partial} \subset \partial\Omega \times [0, T]$ ,

and spatial interior  $\mathcal{C}_i \subset \Omega$ , and use them to form an approximation of the continuous loss (3.3). In our examples, the boundary operator  $\mathcal{G}$  is taken to be a spatially periodic one for simplicity.

The training points must be carefully chosen in the presence of discontinuities. In general it is well known that the distribution of training points can impact the network performance. In particular, dense concentration of training points near the spatiotemporal vicinity of a discontinuity can effectively enhance the network performance in approximating the PDE solutions with jumps. For the transport equation, the location of the discontinuity is often determined by the parameters. Therefore, to achieve better network performance, in practice it is necessary for the training/sampling points to be constructed in a parameter-dependent fashion.

Finally, we remark that the above construction optimizes for “meta-neurons”, i.e. the networks within the GPT-PINN. For some of our test cases, in particular those in Section 4.1, we use analytically available formulas for  $u(\cdot, \cdot; \mu^i)$ , and in those examples the entire PINN apparatus is simply replaced with this explicit function evaluation.

### 3.3 TGPT-PINN

As shown by Figure 3, the hidden layers of the TGPT-PINN include initially a transform layer and, similar to GPT-PINN [8], are followed by a meta-PINN layer with pre-trained PINNs as activation functions. The novel contribution is the transform layer that allows the TGPT-PINN to achieve nonlinear model order reduction. This transform layer allows us to further explore the potential of the full PINNs used as the activation function, thus significantly enhancing the expressivity of the (pre-trained) full PINNs, even with only one pre-trained PINN in a single neuron in the hidden layer. This boost of expressivity from the parameter-dependent transform layer allow us to construct a hyper-reduced neural network of pre-trained full networks featuring parameter-dependent discontinuous functions. In particular, we achieve sizes of the outer network that are much smaller than what the Kolmogorov width of the problem implies is optimal for linear reduction of transport-dominated problems.

#### 3.3.1 Design of the transform layer

The *source-to-target transform layer*  $T_{\mu,\eta}$  in (3.1) of the proposed TGPT-PINN is abstractly,

$$T_{\mu,\mu^i}(\mathbf{x}, t) : \Omega \times [0, T] \longrightarrow \Omega \times [0, T],$$

and in principle it should be surjective. In this paper, we set it as

$$T_{\mu,\eta}(\mathbf{x}, t) := \text{Mod}_{\Omega,T} \left( W_{\mu,\eta} \begin{pmatrix} \mathbf{x} \\ t \end{pmatrix} + b_{\mu,\eta} \right), \quad \eta = \mu^1, \dots, \mu^N. \quad (3.5)$$

Here,  $W_{\mu,\eta} \in \mathbb{R}^{(d+1) \times (d+1)}$ , and  $b_{\mu,\eta} \in \mathbb{R}^{d+1}$  making  $T_{\mu,\eta}$  a simple linear transformation that depends on  $\mu$  and  $\eta$ . In particular, the dependence on  $\mu$  is enforced through “online” training. Moreover, to make sure that the range of  $T_{\mu,\eta}$  is exactly  $\Omega \times [0, T]$ , we apply  $\text{Mod}_{\Omega,T}(\cdot)$  which is an element-wise modulo map, ensuring that each component of  $T$  outputs on the appropriate slice of  $\Omega \times [0, T]$ . Our notation of  $(\mu, \eta)$  subscripts on  $W_{\mu,\eta}$  is meant to reveal that  $W$  depends on both  $\eta$  (i.e., indexing the particular snapshot) as well as  $\mu$  (i.e., the given value of the parameter in the online phase affects the value of the weights trained through minimization).

We use the particular form of  $T$  above for all the experiments in this paper. However, the general framework of the TGPT-PINN is not limited by the relatively simple choice of  $T$  above. For example,  $T$  could itself be a deep neural network. Our simple choice is motivated by the discussion in Section 2.2, demonstrating that even simple linear-type maps can be effective for transport-based problems. In particular,  $T$  can be trained to stretch or shrink the  $(\mathbf{x}, t)$  variables, and in particular can be trained to expand spectral content in the frequency domain. For this reason, as shown by our numerical examples, the proposed architecture also works well for much more complicated problems that don’t just feature transport with constant speed.

### 3.3.2 Online phase training

The transformation (3.5) and the TGPT-PINN ansatz (3.1) mean that a given TGPT-PINN with  $n$  PINNs (pre-trained at  $\{\mu^1, \dots, \mu^n\}$ ) has the following  $n(d^2 + 3d + 3)$  network parameters to train

$$\Theta(\mu) := \{\{W_{\mu, \mu^i}\}_{i=1}^n, \{b_{\mu, \mu^i}\}_{i=1}^n, \{c_i(\mu)\}_{i=1}^n\}. \quad (3.6)$$

Similar to the GPT-PINN, this number of network parameters is *independent* of the architecture parameters used to train the individual PINNs, and depends strictly linearly on  $n$ , the number of snapshots. With the network version of the TGPT-PINN ansatz (3.1) denoted by

$$\Psi_{\text{NN}}^{\Theta(\mu)}(x, t) := \sum_{i=1}^n c_i(\mu) \Psi_{\text{NN}}^{\mu^i}(T_{\mu, \mu^i}(x, t)), \quad (3.7)$$

the loss function of the TGPT-PINN is set to be the same loss as the full PINN, consisting of three parts: the residual of the PDE and the losses corresponding to the initial value condition and the boundary condition

$$\begin{aligned} \mathcal{L}_{\text{PINN}}^{\text{TGPT}}(\Theta(\mu)) = & \frac{1}{|\mathcal{C}_o^r|} \sum_{(\mathbf{x}, t) \in \mathcal{C}_o} \left\| \lambda \cdot \left( \frac{\partial}{\partial t} (\Psi_{\text{NN}}^{\Theta(\mu)})(\mathbf{x}, t) + \mathcal{F}(\Psi_{\text{NN}}^{\Theta(\mu)})(\mathbf{x}, t) \right) \right\|_2^2 \\ & + \varepsilon_b \cdot \frac{1}{|\mathcal{C}_\partial^r|} \sum_{(\mathbf{x}, t) \in \mathcal{C}_\partial} \left\| \mathcal{G}(\Psi_{\text{NN}}^{\Theta(\mu)})(\mathbf{x}, t) \right\|_2^2 + \varepsilon_i \cdot \frac{1}{|\mathcal{C}_i^r|} \sum_{\mathbf{x} \in \mathcal{C}_i} \left\| \Psi_{\text{NN}}^{\Theta(\mu)}(\mathbf{x}, 0) - u_0(\mathbf{x}) \right\|_2^2. \end{aligned} \quad (3.8)$$

The online collocation sets  $\mathcal{C}_o^r \subset \Omega \times [0, T]$ ,  $\mathcal{C}_\partial^r \subset \partial\Omega \times [0, T]$  and  $\mathcal{C}_i^r \subset \Omega$  need not be related to their full PINN counterparts  $\mathcal{C}_o, \mathcal{C}_\partial$  and  $\mathcal{C}_i$ , but in this paper we take them to be the same sets for simplicity. The training of  $\Theta(\mu)$  is accomplished through standard automatic differentiation and back propagation. See Ref. [8] for more details on this step, such as precomputations for fast training of the reduced network, and the non-intrusiveness of the (T)GPT-PINN.

### 3.3.3 Offline training of the TGPT-PINN

---

**Algorithm 1** TGPT-PINN for parametric PDE: Offline stage

---

**Input:** A random (or given)  $\mu^1$ , training set  $\Xi_{\text{train}} \subset \mathcal{D}$ , full PINN.

---

- 1: Train a full PINN at  $\mu^1$  to obtain  $\Psi_{\text{NN}}^{\mu^1}$ . Precompute quantities necessary for  $\nabla_{\Theta(\mu)} \mathcal{L}_{\text{PINN}}^{\text{TGPT}}$  at collocation nodes  $\mathcal{C}_o^r, \mathcal{C}_\partial^r$ , and  $\mathcal{C}_i^r$ . Set  $n = 2$ .
- 2: **while** *stopping criteria not met*, **do**
- 3:   Train the  $(n-1)$ -neuron TGPT-PINN at  $\mu$  for all  $\mu \in \Xi_{\text{train}}$  and record the indicator  $\Delta_{\text{NN}}^r(\Theta(\mu))$ .
- 4:   Choose  $\mu^n = \arg \max_{\mu \in \Xi_{\text{train}}} \Delta_{\text{NN}}^r(\mu)$ .
- 5:   Train a full PINN at  $\mu^n$  to obtain  $\Psi_{\text{NN}}^{\mu^n}$ . Precompute quantities necessary for  $\nabla_{\Theta} \mathcal{L}_{\text{PINN}}^{\text{TGPT}}$  at collocation nodes  $\mathcal{C}_o^r, \mathcal{C}_\partial^r$ , and  $\mathcal{C}_i^r$ .
- 6:   Update the TGPT-PINN by adding a neuron to the hidden TGPT-PINN layer to construct the  $n$ -neuron TGPT-PINN.
- 7:   Set  $n \leftarrow n + 1$ .
- 8: **end while**

**Output:**  $N$ -neuron TGPT-PINN, with  $N$  being the terminal index.

---

With the online solver described above, the offline training amounts to the application of the greedy algorithm outlined in Algorithm 1. The meta-network adaptively “learns” the parametric dependence of the system and “grows” the TGPT-PINN hidden layer and enriches the transform layer one neuron and transformation at a time. At every step, we select the parameter value that is worst approximated by the current meta-network.

Specifically, we first randomly select, in a discretized parameter domain  $\Xi_{\text{train}} \subset \mathcal{D}$ , one parameter value  $\mu^1$  and train the associated (highly accurate) PINN  $\Psi_{\text{NN}}^{\mu^1}$ . The algorithm then decides how to “grow” its meta-network by scanning the entire discrete parameter space  $\Xi_{\text{train}}$  and, for each parameter value, training this reduced network (of 1 neuron  $\Psi_{\text{NN}}^{\mu^1}$ ). As it scans, it records an error indicator  $\Delta_{\text{NN}}^r(\Theta(\mu))$  at every location. The next parameter value  $\mu^2$  is the one generating the largest error indicator. The algorithm then proceeds by training a full PINN at  $\mu^2$  and therefore grows its hidden PINN layer into two neurons with customized (but pre-trained) activation functions  $\Psi_{\text{NN}}^{\mu^1}$  and  $\Psi_{\text{NN}}^{\mu^2}$ . This process is repeated until the stopping criteria is met which can be either that the error indicator is sufficiently small or a pre-selected size of the reduced network is met.

## 4 Numerical results

In this section, we test the proposed TGPT-PINN and report numerical results. This is done with two types of experiments. First, to separate the influence of the PINN solver which is analogous to the “truth” discretization in traditional RBM methods, we test the case when  $u(\mathbf{x}, \mu)$  is a known function with different types of regularity. Next, we treat  $u(\mathbf{x}, \mu)$  as a solution to a parameterized PDE, numerically computed using a PINN solver. The code for all these examples are published on GitHub at <https://github.com/DuktigYajie/TGPT-PINN>.

### 4.1 TGPT-PINN vs EIM as function approximators

We first test the case when

$$u(\mathbf{x}, \mu) : \Omega \times \mathcal{D} \longrightarrow \mathbb{R}$$

is analytically given, and compare the TGPT-PINN with the EIM [2] which is the standard approach in model order reduction for approximating nonlinear and nonaffine functions. Table 1 lists the results when both approaches are applied to 7 functions of different regularities.

Function	$\Omega$	$\mathcal{D}$	$ \Xi_{\text{train}} $	EIM		TGPT-PINN	
				#basis	$L_2$ error	#basis	$L_2$ error
$\sin(x + \mu)$	$[\pi, \pi]$	$[-5, 5]$	201	3	7.6e-15	1	1.2e-14
$\sin(\mu x)$	$[-\pi, \pi]$	$[1, 2]$	201	17	1.8e-15	1	9.9e-13
$\sin(\mu_1(x + \mu_2))$	$[-\pi, \pi]$	$[-5, 5]^2$	$20 \times 20$	21	2.0e-14	1	9.9e-11
$\max(\sin(x + \mu), 0)$	$[-\pi, \pi]$	$[-5, 5]$	201	192	4.2e-04	1	9.7e-15
				193	2.3e-15		
$ x + \mu $	$[-10, 10]$	$[-5, 5]$	201	101	8.1e-02	1	4.9e-15
				102	5.6e-14		
$\psi\left(\frac{x}{0.4+\mu} - 1\right)$	$[-1, 1]$	$[-1, 1]$	201	100	1.9e-01	10	1.5e-05
				101	6.9e-16		
$\frac{1}{\sqrt{(x-\mu_1)^2 + (y-\mu_2)^2}}$	$[-1, 1]^2$	$[-1, -0.01]^2$	$21 \times 21$	150	3.8e-15	1	9.9e-13

Table 1: Results for Sections 4.1 and 4.1.1 to 4.1.3: Comparison of the EIM and the proposed TGPT-PINN as function approximators. The parameter space  $\mathcal{D}$  is discretized to  $\Xi_{\text{train}}$  using  $|\Xi_{\text{train}}|$  equispaced points.

The first three of 7 functions in Table 1 are smooth functions. We observe that the TGPT-PINN is able to capture the parameter dependence exactly and, as a result, approximates each of these functions by one neuron to machine precision. Because the parametric



dependence for these functions is smooth, the EIM is also able to approximate the functions to machine accuracy. However, it requires many more basis functions since it is a linear reduction approach (while the TGPT-PINN is nonlinear). We discuss the remaining four functions, which have different types of regularity, in the sections below.

#### 4.1.1 Functions with moving kinks

When  $u(x, \mu) = \max(\sin(x + \mu), 0)$  or  $|x + \mu|$  (rows 4 and 5 of Table 1), we have at least one kink (point of discontinuity of the derivative) and the location of the kink depends on  $\mu$ . See Figure 4 for the convergence history for the EIM along with plots of the function snapshots for different  $\mu$  values. We observe relatively slow convergence for the EIM, reaching two digits with 99 snapshots. On the other hand, the TGPT-PINN only relies on one snapshot to achieve machine accuracy (see rows 4 and 5 of Table 1) for each of these two parametric functions.

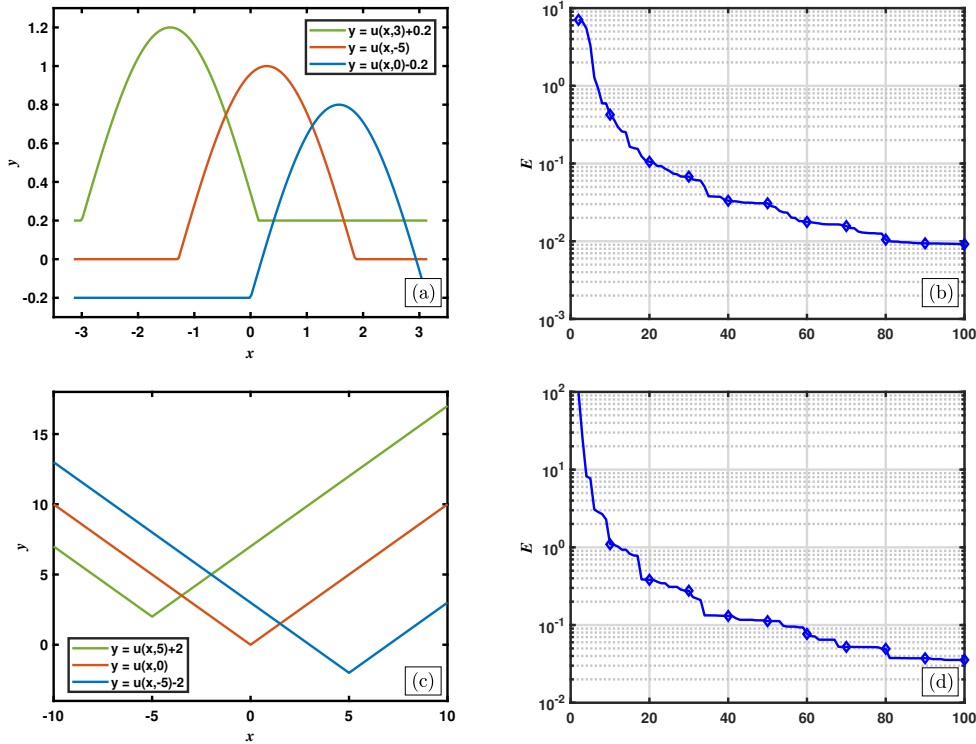


Figure 4: Results from Section 4.1.1 paired with rows 4 and 5 of Table 1: Snapshots and EIM histories of convergence for functions with moving kinks. (a, b)  $u(x, \mu) = \max(\sin(x + \mu), 0)$ ; and (c, d)  $u(x, \mu) = |x + \mu|$ .

#### 4.1.2 Functions with moving discontinuities

We next test the following example from [32] which features a moving discontinuity as explained in Section 2.2,

$$u(x, \mu) := \psi\left(\frac{x}{0.4 + \mu} - 1\right), \quad \psi(x) := \begin{cases} \exp\left(-\frac{1}{1-x^2}\right) & -1 \leq x < -\frac{1}{2}, \\ 0, & \text{else.} \end{cases}$$

The summary of results is given by row 6 of Table 1. The function  $\psi(x)$  is discontinuous at  $x = -1/2$  which means that  $u(x, \mu)$  is discontinuous at a parameter-dependent location  $x = (\mu + 0.4)/2$ . Furthermore, we note that when  $\mu > -0.4$ ,  $x_* = (\mu + 0.4)/2 > 0$  is a discontinuity point of  $u(x, \mu)$  which is clamped to zero for  $x > x_*$ . On the other hand, when

$\mu < -0.4$ , then  $x_* < 0$  is the discontinuity point of  $u(x, \mu)$ , and  $u$  is clamped to zero for  $x < x_*$ .

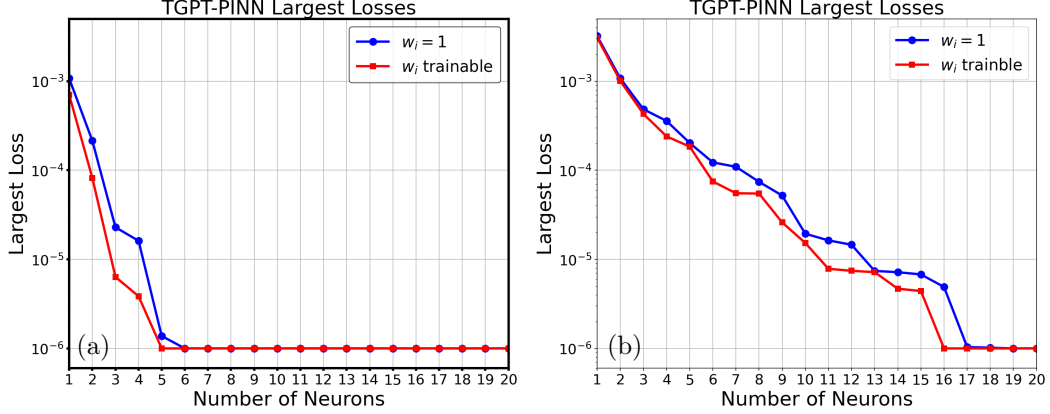


Figure 5: Results from Section 4.1.2 paired with row 6 of Table 1: TGPT-PINN histories of convergence when the number of neurons increases for functions with a moving discontinuity,  $u(x, \mu) = \psi\left(\frac{x}{0.4+\mu} - 1\right)$ . (a)  $\mu \in [0, 1]$  with 101 equispaced points; (b)  $\mu \in [-1, 1]$  with 201 equispaced points.

We observe in row 6 of Table 1 that the EIM fails to capture this family of functions even with all but one snapshots from the entire discretized domain. On the other hand, the TGPT-PINN successfully reaches 6 digits of accuracy by 10 neurons. Since this example is more difficult than the previous ones, we conduct a further example to investigate the impact of the parameter space: We compare the TGPT-PINN performance using  $(\mathcal{D}, \Xi_{\text{train}}) = ([0, 1], 101)$ , and  $(\mathcal{D}, \Xi_{\text{train}}) = ([-1, 1], 201)$ . The latter is the setup in Table 1. As we increase the number of neurons in the hidden layer, Figure 5 shows the change of the L2 norm of the error committed by the approximate TGPT-PINN solution. The exact function and its approximation with 1, 6, and 10 neurons are shown in Figure 6. The visual agreement between the exact function and its TGPT-PINN approximation is confirmed by the computed error indicating an accuracy of 5 to 6 digits.

Next, we examine the convergent network parameters for the more challenging case with  $\mu \in [-1, 1]$  to confirm that it aligns with our theoretical understanding of this discontinuous function. We examine the function approximation with 10 neurons

$$u(x; \mu) \approx \sum_{i=1}^{10} c_i(\mu) u(w_i(\mu)x + b_i(\mu), \mu^i), \quad (4.1)$$

where in order to line up the discontinuity, we know the exact shift should be  $b_i^{\text{exact}} = \frac{\mu^i + 0.4}{2} - w_i \frac{\mu + 0.4}{2}$ . We list in Table 2 the results of  $\{(\mu^i, w_i, b_i, c_i) : i = 1, \dots, 10\}$  for an unseen parameter value  $\mu = 0.595$ , i.e., this parameter was not in the training set for offline construction. The left half of the table shows the results when we fix  $w_i = 1$  and the right part corresponds to when the  $w_i$ 's are trainable parameters. We see that the network training is highly effective as all the  $b_i$ 's coincide with their exact value in both cases. It is interesting to note that, for both cases, relatively larger discrepancies  $|b^i - b_i^{\text{exact}}|$  occur when  $c_i$  is relatively small which is a testament to the robustness of the algorithm.

#### 4.1.3 2D functions close to being degenerate

We now consider the classical 2-dimensional nonlinear and nonaffine function from [14],

$$u(x, y; \mu_1, \mu_2) = \frac{1}{\sqrt{(x - \mu_1)^2 + (y - \mu_2)^2}} \text{ on } (x, y) \in [0, 1]^2,$$

parameterized by  $(\mu_1, \mu_2) \in [-1, -0.01]^2$ . Due to the setup of the physical and parameter domains, this family contains functions that are very close to being singular for certain

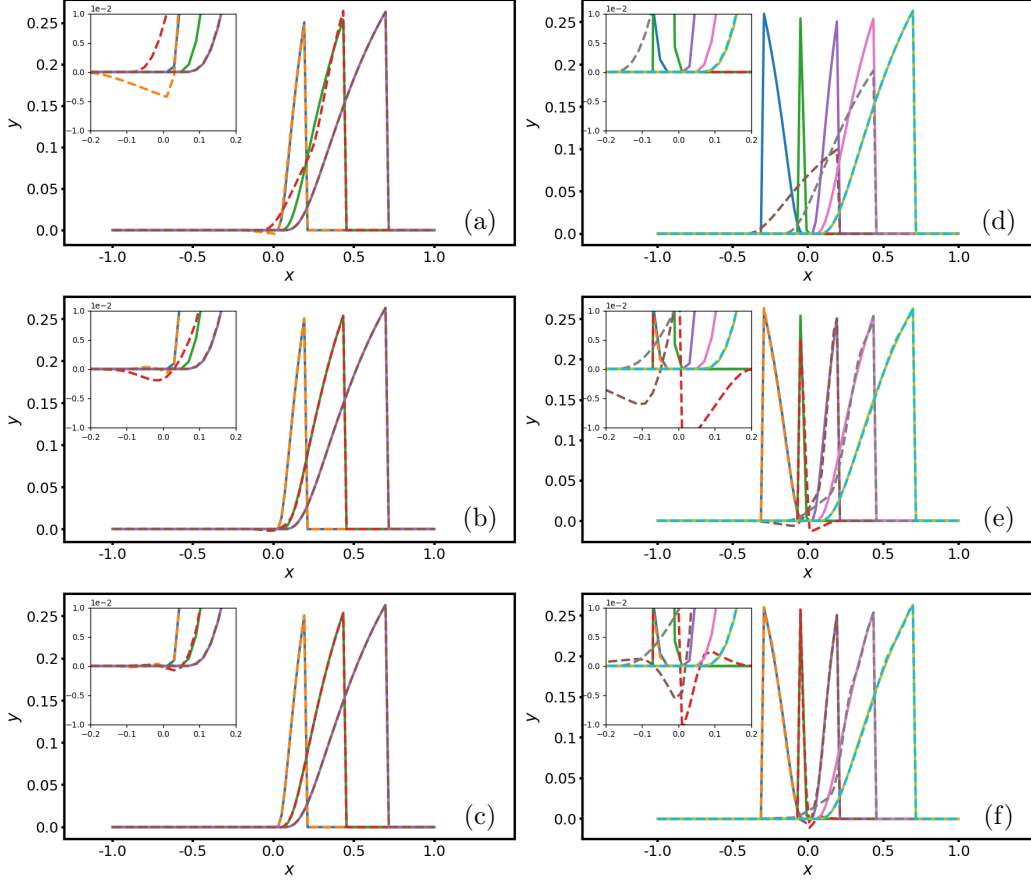


Figure 6: Results from Section 4.1.2 paired with row 6 of Table 1: Function with a moving discontinuity  $u(x, \mu) = \psi\left(\frac{x}{0.4+\mu} - 1\right)$  and its approximation with 1, 6, and 10 neurons (top to bottom). The solid lines describe the exact solutions and the dash lines are the TGPT-PINN solutions for different  $\mu$ . (a-c) training set  $\mu \in [0, 1]$  with 101 equispaced points and solutions of  $\mu = 0.005, 0.505$  and  $0.995$  being displayed; (d-f) training set  $\mu \in [-1, 1]$  with 201 equispaced points and solutions of  $\mu = -0.995, -0.505, 0.005, 0.505$  and  $0.995$  being displayed.

$\mu = 0.595, w_i = 1$ fixed, final loss 1.23e-05						$\mu = 0.595, w_i$ trainable, final loss 8.12e-06					
$i$	$\mu^i$	$w_i$	$ b^i - b_i^{\text{exact}} $	$b_i$	$c_i$	$\mu^i$	$w_i$	$ b^i - b_i^{\text{exact}} $	$b_i$	$c_i$	
1	1.0	1	1.5e-08	0.20	-2.34e-02	1.0	1.14	2.98e-08	0.13	3.85e-01	
2	-0.99	1	1.8e-07	-0.80	9.20e-42	-0.99	0.47	1.79e-07	-0.53	-5.17e-42	
3	0.03	1	0.0e-00	-0.28	-1.71e-1	-0.09	0.77	5.96e-08	-0.23	1.34e-01	
4	-0.55	1	1.2e-07	-0.57	8.52e-41	-0.55	0.62	1.19e-07	-0.38	-3.45e-41	
5	-0.29	1	3.0e-08	-0.44	-2.62e-02	-0.33	0.69	0.00e-00	-0.31	7.21e-03	
6	0.44	1	1.5e-08	-0.08	8.84e-01	0.31	0.90	1.49e-08	-0.10	9.47e-01	
7	-0.47	1	1.2e-07	-0.53	-4.40e-41	-0.47	0.64	8.94e-08	-0.36	2.51e-41	
8	-0.75	1	1.2e-07	-0.67	-4.97e-41	-0.72	0.56	1.19e-07	-0.44	2.15e-41	
9	-0.17	1	3.0e-08	-0.38	-7.01e-02	-0.25	0.72	8.94e-08	-0.28	-2.79e-02	
10	-0.33	1	3.0e-08	-0.46	8.50e-03	0.08	0.83	7.45e-08	-0.17	-4.44e-01	

Table 2: Results for Section 4.1.2 paired with row 6 of Table 1: Trained TGPT-PINN network parameters at  $\mu = 0.595$  for the function with moving discontinuities.

parameter value. The EIM (see Figure 7) needs about 120 basis functions to reach an accuracy of  $10^{-12}$  while the TGPT-PINN *requires only one neuron*,  $u(x, y; (-1, -1))$ , to achieve the same accuracy for all parameters in the parameter space discretized by a  $21 \times 21$  grid. The solution, the errors, and the training history corresponding to two locations of the parameter domain committed by the TGPT-PINN surrogate are plotted in Figure 8.

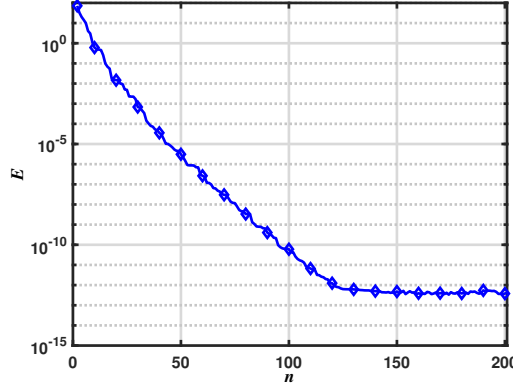


Figure 7: Results for Section 4.1.3 paired with row 7 of Table 1: EIM history of convergence for the 2D functions close to being degenerate.

## 4.2 TGPT-PINN for parametric PDEs

In this section, we test the TGPT-PINN algorithm on three pPDEs, the linear transport equation parameterized by the wave speed (first example, Section 4.2.1), the nonlinear 1D-reaction equation parameterized by the reaction coefficient (second example, Section 4.2.2), and the 1D nonlinear reaction-diffusion equation parameterized by the viscosity and reaction coefficient (third example, Section 4.2.3). The overall conclusions are as follows:

- Section 4.2.1: For wave-like problems when linear reduction does not work, the TGPT-PINN can achieve significant accuracy because transport-based parameter dependence can be efficiently approximated by the novel transform layer. For our particular example, *only one neuron is needed to achieve machine precision accuracy*.
- Sections 4.2.2 and 4.2.3: On problems when linear model reduction can work effectively, and in particular when the GPT-PINN performs well, the TGPT-PINN works using a

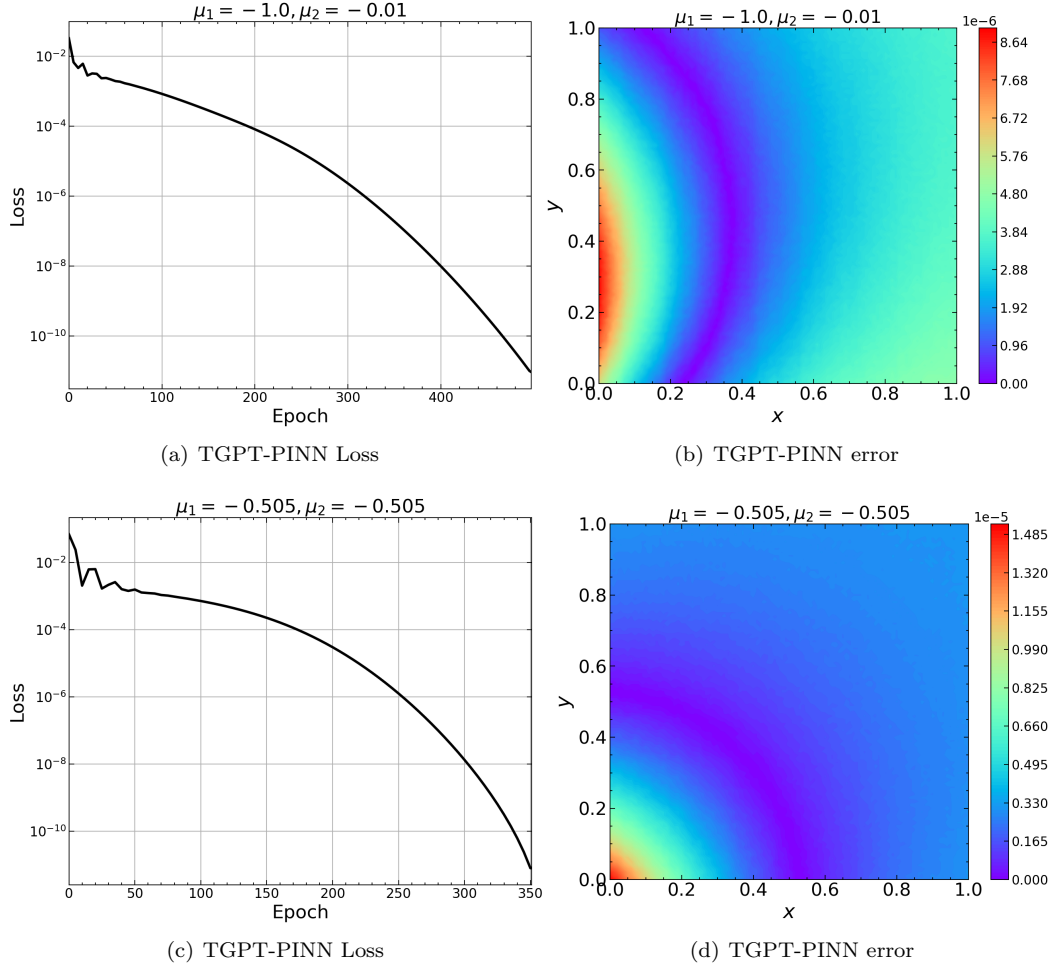


Figure 8: Results for Section 4.1.3 paired with row 7 of Table 1 for the 2D functions close to being degenerate: TGPT-PINN error (right) committed by the TGPT-PINN solution with one neuron at a corner and the center of the parameter domain. Shown on the left are the corresponding training histories.

very small number of snapshots (1 and 3, respectively), and achieves better accuracy with this small number of snapshots than the GPT-PINN with a much larger number of snapshots.

This vast improvement is a manifestation of the novel design of the TGPT-PINN: the addition of a trainable transform layer to a network of pre-trained networks.

#### 4.2.1 1D transport equation

Consider the following example [32],

$$\begin{aligned} u_t + \nu u_x &= 0, \quad \text{for } 0 < t < 2, x \in [-1, 1], \\ u(x, 0) &= g(x) := \begin{cases} 0, & -0.5 < x < 0.5, \\ 1, & \text{else.} \end{cases} \end{aligned}$$

The exact solution to the problem is given by  $u(x, t) = g(x - \nu t)$ , so that the parameter  $\nu$  determines the direction and speed of propagation for the solution of the equation.

**Sampling for the PINN loss function.** For fixed  $\nu$ , this problem is challenging to train a PINN on because of the discontinuity in  $(x, t)$ . To ameliorate this difficulty, we choose sampling points for the loss function in a standard randomized fashion over the  $(x, t)$  domain, but we additionally place extra training points near the discontinuity. For details, for the first term of the loss function (3.8) corresponding to the PDE residual in the domain interior, a  $400 \times 400$  equidistant grid is created over the domain, and 10,000 points are randomly subsampled as the training set, with a relatively concentrated distribution near the vicinity of the discontinuity bands. We choose 60% of the points by randomly subsampling within an  $(x, t)$  distance of 0.1 of the discontinuity. 20% more points are subsampled from equidistant points whose  $(x, t)$  distance from the discontinuity takes values in  $[0.1, 0.2]$ . The remaining 20% of the points are subsampled from the remaining points. A similar procedure is done for the initial condition training grid, where a total of 1000 points are randomly subsampled from  $10^4$  equispaced points in  $x$ , with a higher concentration of points subsampled close to the discontinuity.

**PINN results.** The full PINN for this transport equation is a fully connected neural network with  $[2, 20, 20, 20, 1]$  neurons across its layers. The activation function used is  $\tanh(\cdot)$ . The PINN is trained with a learning rate of 0.001, a maximum of 4e5 iterations, a tolerance of 1e-6, and the Adam optimizer. We take  $\nu = 0.0$  and train a full PINN. The reduction in loss of the network in a single random experiment is shown in Figure 9 (a-c), showing that the PINN is relatively accurate.

**TGPT-PINN results.** We conducted tests on a discretized parameter set of 41 equispaced samples over  $\mu = \nu \in [-10, 10]$ , with an accuracy tolerance of 1e-5, a learning rate of 0.05, and a maximum iteration count of  $10^5$ . We use only  $N = 1$  snapshot, manually choosing  $\mu^1 = 0$ . The variation of loss, solutions, and errors under the two extreme parameter values,  $\nu = \pm 10$  (the hardest parameter values to approximate) are shown in Figure 9 (d-i).

#### 4.2.2 1D reaction PDE

The one-dimensional reaction problem is a hyperbolic PDE that is commonly used to model chemical reactions,

$$\begin{aligned} \frac{\partial u}{\partial t} - \rho u(1 - u) &= 0, \quad (x, t) \in [0, 2\pi] \times [0, 1], \\ u(x, 0) &= h(x), \quad x \in [0, 2\pi], \\ u(0, t) &= u(2\pi, t), \quad t \geq 0, \end{aligned} \tag{4.2}$$

where  $\rho > 0$  is the reaction coefficient. We take the initial condition as,

$$h(x) = \exp\left(-\frac{(x - \pi)^2}{2(\pi/4)^2}\right). \tag{4.3}$$



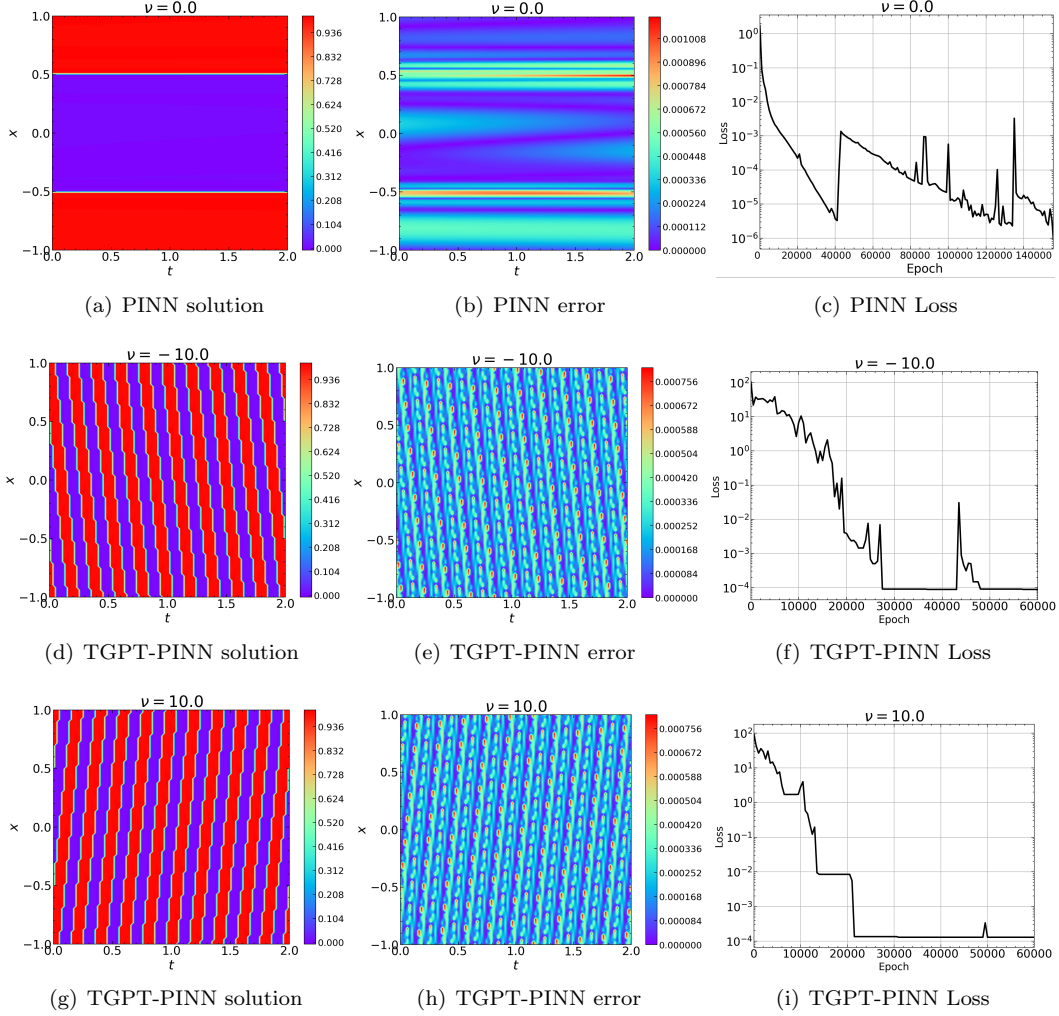


Figure 9: Results for Section 4.2.1. Top row: Full PINN for  $\nu = 0.0$ . Middle row: TGPT-PINN at  $\nu = -10.0$ . Bottom row: TGPT-PINN for  $\nu = 10.0$ . Both middle and bottom are approximated using the only PINN from the top row.

The equation has a simple analytical solution:

$$u(x, t) = \frac{h(x) \exp(\rho t)}{h(x) \exp(\rho t) + 1 - h(x)}. \quad (4.4)$$

**PINN results:** The PINN we employ is a fully connected neural network with layer widths  $[2, 20, 20, 20, 1]$  with  $\lambda \equiv 1$  in  $\mathcal{L}_{\text{int}}(u)$  of the PINN loss function (3.3). The training samples are randomly selected, but we use an optimized activation function introduced in [33]:

$$\text{WaveAct}(x) = w_1 * \sin(x) + w_2 * \cos(x), \quad (4.5)$$

where  $w_1$  and  $w_2$  are trainable hyperparameters. Numerical experiments demonstrate that such an activation function is more suitable for solving reaction equations compared to traditional activation functions. Figure 10 presents the accuracy and loss reduction of the full PINN for parameters  $\rho = 1$  and 10. This problem is difficult to numerically solve for large values of the parameter  $\rho$ . One can observe this difficulty in Figure 10 where  $\rho = 10$  requires substantially more epochs to successfully train; this is due to the more complex nature of the phase transition layer for larger  $\rho$ .

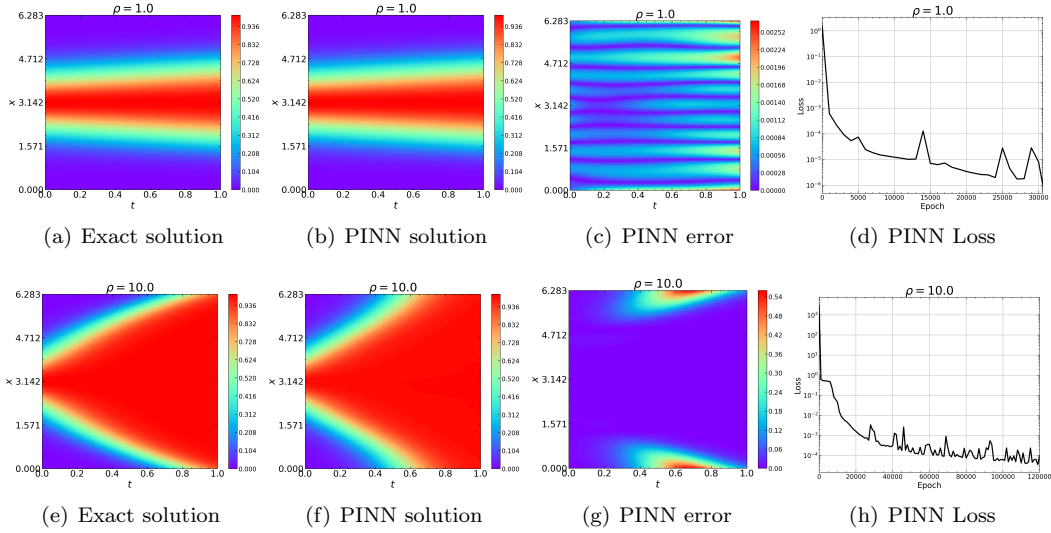


Figure 10: Results for Section 4.2.2: PINN snapshots for  $\rho = 1.0$  (top) and 10.0 (bottom).

**GPT/TGPT-PINN results:** We solve the problem using both the GPT-PINN and TGPT-PINN for  $\rho \in [1, 10]$ , and compare the histories of convergence for the loss as the hyper-reduced neural networks gradually grow. The results of the loss and parameter selection are shown in Figure 11. From the analytical solution of the problem, it can be observed that all solutions can be obtained through  $\rho$ -dependent  $(x, t)$  alignment and shift operations from one reference solution. Figure 11 demonstrates this in practice: the TGPT-PINN with just one neuron can capture the parametric dependence and achieve better approximation than the (linear) GPT-PINN with 10 neurons. This is true even though we have manually selected the single snapshot for the TGPT-PINN as the “easiest” (i.e., smallest) value of  $\rho$  in the parameter domain. In Figure 12 and Figure 13 we show a comparison of solutions obtained by the GPT-PINN and TGPT-PINN for the relatively large values of  $\rho = 3.25$  and  $\rho = 9.85$ . We observe that although the TGPT-PINN solution with 1 neuron is indeed more accurate than the GPT-PINN solution with 10 neurons, they are both significantly large in parts of the space-time domain for the more challenging case of  $\rho = 9.85$  (around 40%, see Figure 13 (c) and (g)). However, we remark that this large error is purely inherited from the underlying PINN solution. Indeed, in Figure 14, we replace the full PINN solution by a network of the same structure but trained via supervised learning from the exact solution (4.4). The resulting TGPT-PINN error decreased from 40% to below 2%. This confirms the effectiveness of the TGPT-PINN’s reduction strategy, while highlighting its dependence on accurate snapshots, similar to any other snapshot-based reduced order model.

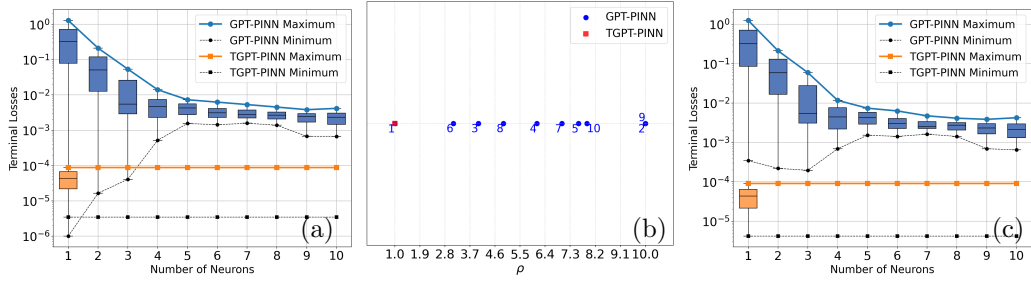


Figure 11: Results for Section 4.2.2: Comparison of the GPT-PINN and TGPT-PINN for the reaction equation in terms of histories of convergence (a, c) and parameter selection (b).

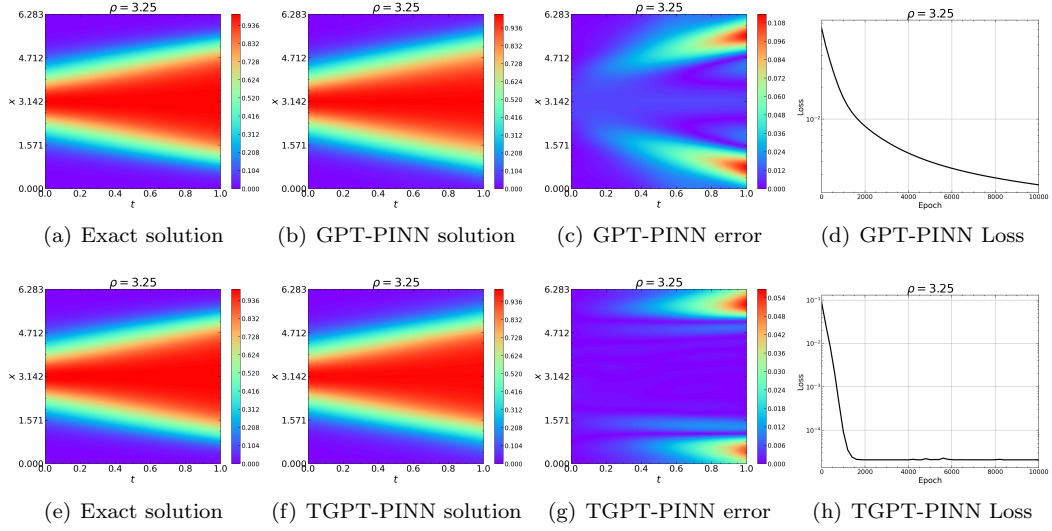


Figure 12: Results for Section 4.2.2: GPT-PINN (top, with 10 neurons) and TGPT-PINN (bottom, with 1 neuron) results for  $\rho = 3.25$ .

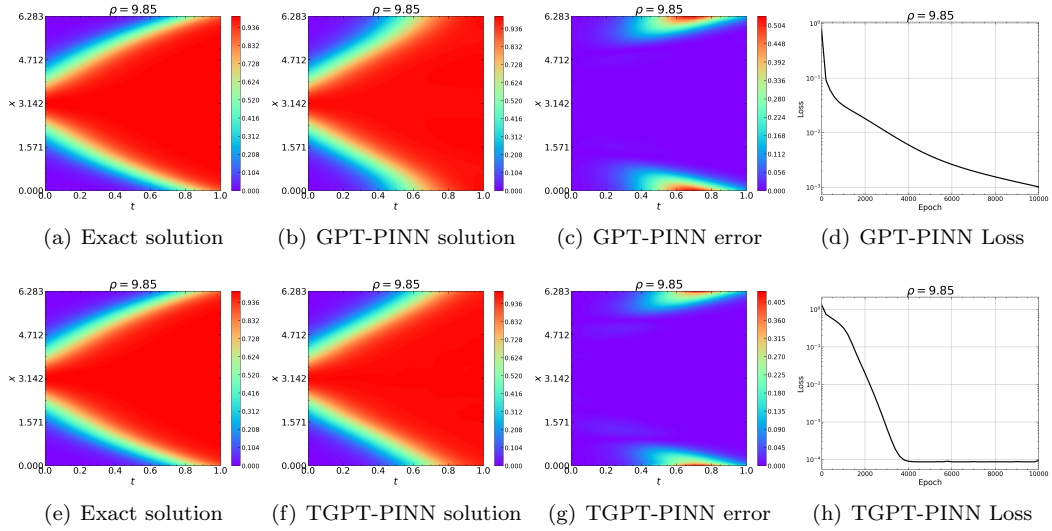


Figure 13: Results for Section 4.2.2: GPT-PINN (top, with 10 neurons) and TGPT-PINN (bottom, with 1 neuron) results for  $\rho = 9.85$ .

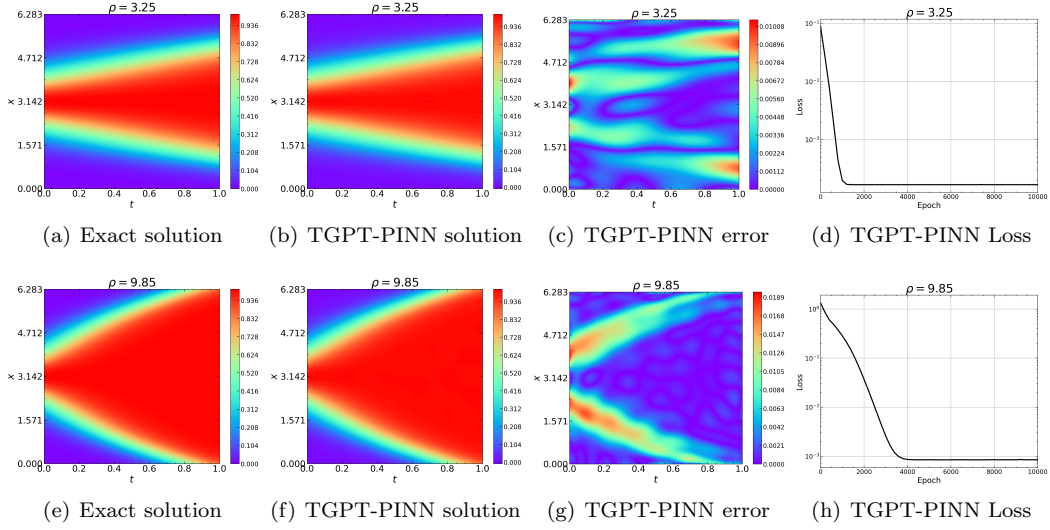


Figure 14: TGPT-PINN results for Section 4.2.2 with a more accurate underlying PINN:  $\rho = 3.25$  (top) and  $\rho = 9.85$  (bottom). The error is much smaller than those in Figures 12 and 13.

### 4.2.3 Reaction-diffusion PDE

The reaction-diffusion system is where a diffusion operator is added to the reaction equation above. The system has the formulation with periodic boundary conditions as follows:

$$\begin{aligned} \frac{\partial u}{\partial t} - \nu \frac{\partial^2 u}{\partial x^2} - \rho u(1 - u) &= 0, \quad (x, t) \in [0, 2\pi] \times [0, 1], \\ u(x, 0) &= h(x), \quad x \in [0, 2\pi], \\ u(0, t) &= u(2\pi t), \quad t \geq 0, \end{aligned} \quad (4.6)$$

where  $h(x)$  is as in (4.3), and  $\nu > 0$  is the diffusion coefficient.

This equation has the following analytical solution:

$$u(x, t) = F^{-1} \left( F \left( \frac{h(x) \exp(\rho t)}{h(x) \exp(\rho t) + 1 - h(x)} \right) e^{-\nu k^2 t} \right), \quad (4.7)$$

where  $F$  and  $F^{-1}$  are the Fourier transform pair, and  $k$  denotes frequency in the Fourier domain.

**PINN results:** The PINN snapshot solver employs a fully connected neural network with dimensions  $[2, 40, 40, 40, 1]$  with  $\lambda \equiv 1$  in  $\mathcal{L}_{\text{int}}(u)$  of the PINN loss function (3.3). The training samples are randomly selected, but we again employ the optimized activation function from [33] in (4.5). In Figure 15 we show the accuracy and loss reduction of PINN for the parameter choice  $(\nu, \rho) = (1, 1)$  (top row) and  $(5, 5)$  (bottom row). Note that as the parameters  $\nu, \rho$  increase, the difficulty of solving the problem significantly increases.

**GPT/TGPT-PINN results:** We set the parameter domain to be  $(\rho, \nu) \in [1, 5] \times [1, 5]$ , and the training set to be an  $11 \times 11$  equispaced grid on this domain. Figure 16 shows the histories of convergence (left) and the locations of the two hyper-reduced networks (middle) as the number of neurons increases when they are trained offline. On the right is the result when they are tested online at 49 random locations not seen during training. It is clear that the TGPT-PINN consistently outperforms the GPT-PINN. In fact, with barely 3 neurons, it already outperforms the GPT-PINN with 10 neurons in terms of loss decay. With 2 neurons, its worst case performance is about the same as the best case for GPT-PINN. We present the solutions obtained at two unseen parameter values for both the GPT-PINN (first row) and TGPT-PINN (second row) in Figure 17 and Figure 18. Once again, we note that the large TGPT-PINN error in parts of the space-time domain (around 40%, see Figure 18 (c) and (g)) is inherited from the underlying PINN solver. In Figure 19, we replace the full

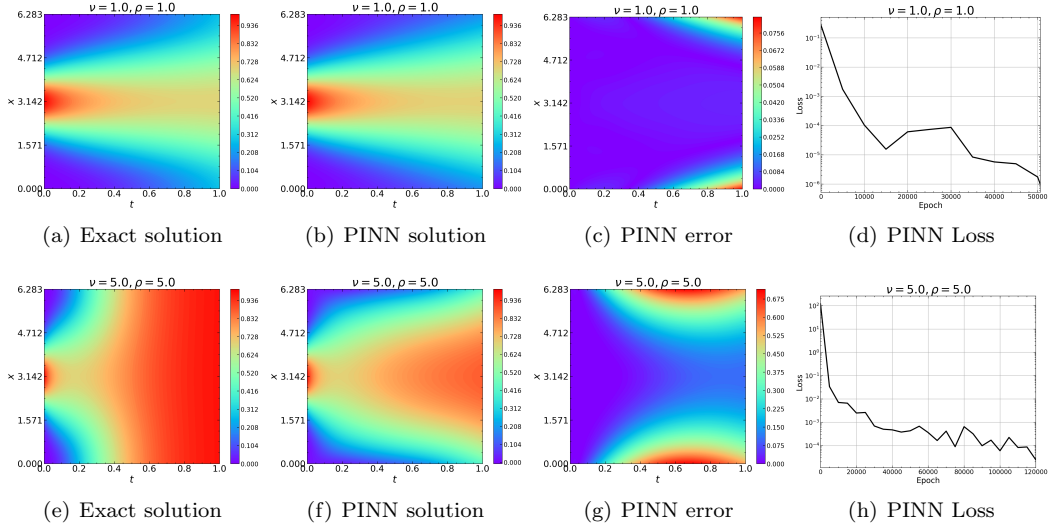


Figure 15: Results for Section 4.2.3: Full PINN for  $\nu = 1, \rho = 1$  (top) and  $\nu = 5, \rho = 5$  (bottom).

PINN solutions by a network of the same structure but trained via supervised learning from the exact solution (4.7). The resulting TGPT-PINN error decreased from 40% to about 6%, confirming the effectiveness of the TGPT-PINN.

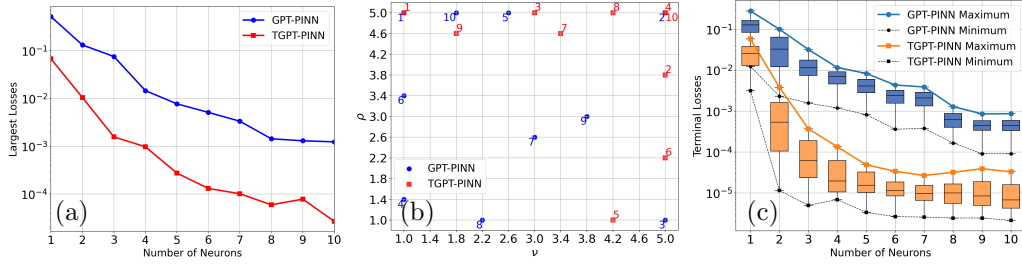


Figure 16: Results for Section 4.2.3 for (T)GPT-PINN for the reaction-diffusion equation: (a) Worst-case history of convergence during training, (b) parameter domain locations determined by the methods, and (c) box plots of the losses when tested online on unseen locations.

#### 4.2.4 2D transport equation

To show the effectiveness of the TGPT-PINN algorithm for higher dimensions, we consider the 2D counterpart of the transport equation tested in Section 4.2.1,

$$u_t + (\mu u_{x_1} + \nu u_{x_2}) = 0, \quad \text{for } 0 < t < 2, \mathbf{x} \in [-1, 1] \times [-1, 1],$$

$$u(\mathbf{x}, 0) = g(\mathbf{x}) := \begin{cases} 1, & -0.5 < x_1, x_2 < 0.5, \\ 0, & \text{else.} \end{cases}$$

Here  $(\mu, \nu) \in [-10, 10] \times [-10, 10]$  delineate the speeds of propagation along each direction. The exact solution to the problem is given by  $u(\mathbf{x}, t) = g(x_1 - \mu t, x_2 - \nu t)$ . The full PINN has architecture  $[3, 40, 40, 40, 40, 1]$  for the number of neurons in the six layers. The collocation points, the activation function, and the learning rate remain the same as 1D while the maximum number of iterations is  $6 \times 10^4$ . We take  $\mu = \nu = 0.0$  and train a full PINN as the single neuron in the GPT-PINN layer of the TGPT-PINN. The reduction in loss of the network in a single random experiment is shown in Figure 20 (a-f).

**TGPT-PINN results.** We conducted tests on a uniform  $21 \times 21$  equi-spaced points over the parameter domain  $[-10, 10] \times [-10, 10]$  with an accuracy tolerance of  $10^{-4}$ , a learning

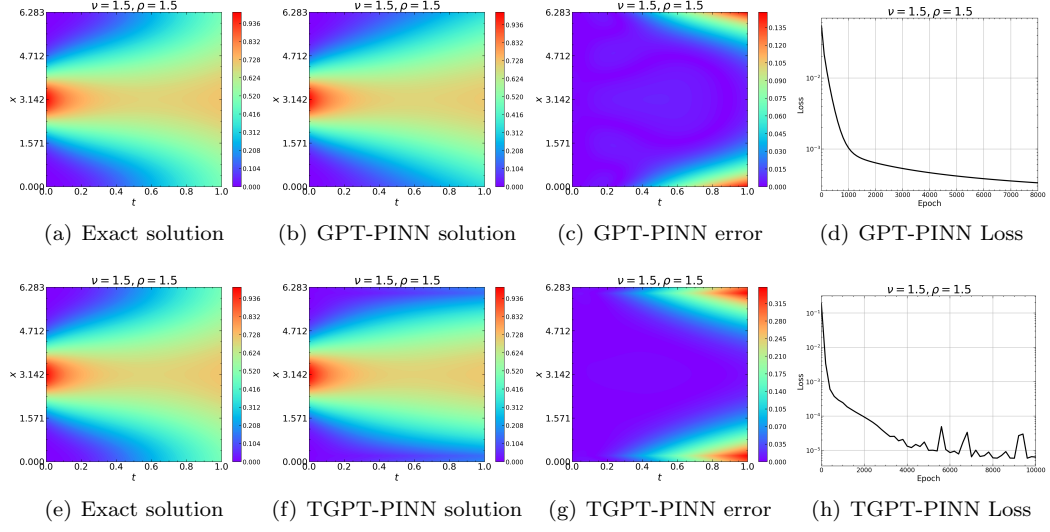


Figure 17: Results for Section 4.2.3: GPT-PINN (top, with 10 neurons) and TGPT-PINN (bottom, with 3 neurons) results for  $\rho = 1.5$  and  $\nu = 1.5$ .

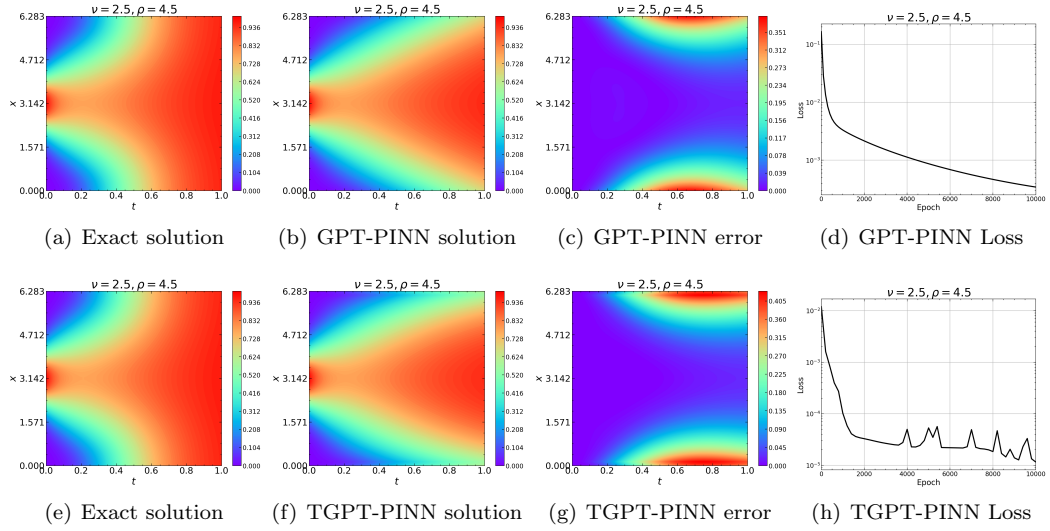


Figure 18: Results for Section 4.2.3: GPT-PINN (top, with 10 neurons) and TGPT-PINN (bottom, with 3 neurons) results for  $\rho = 2.5$  and  $\nu = 4.5$ .



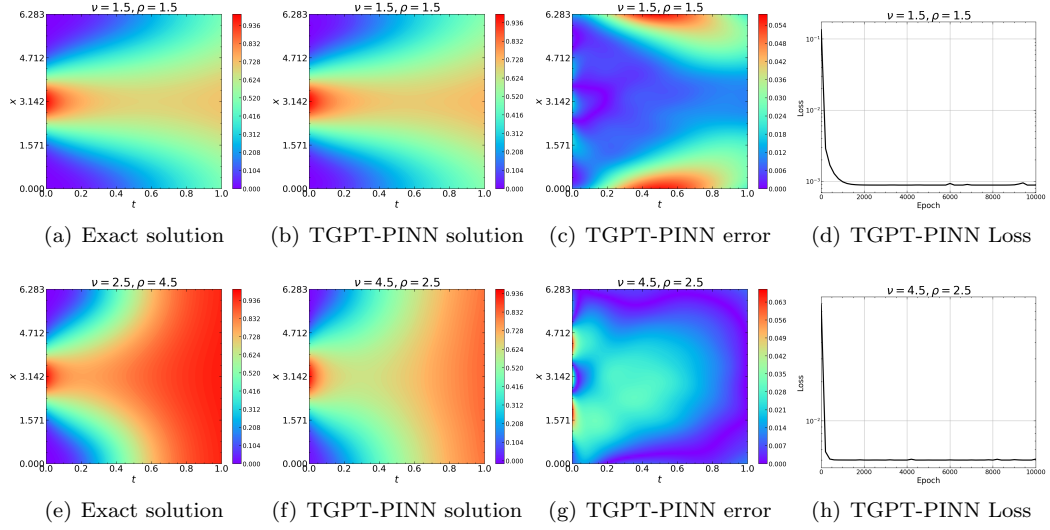


Figure 19: TGPT-PINN results for Section 4.2.3 with a more accurate underlying PINN: (a-d)  $\rho = 1.5$  and  $\nu = 1.5$ ; and (e-h)  $\rho = 2.5$  and  $\nu = 4.5$ . 3 neurons are used for both cases. The error is much smaller than those in Figures 17 and 18.

rate of 1, and a maximum iteration count of  $10^5$ . Figure 21(a-d) displays the results of the loss variation under the four extreme parameter values  $\mu = \pm 10$  and  $\nu = \pm 10$ . Figure 21(e-g) illustrates the exact solutions, the TGPT-PINN solutions, and the errors under the extreme parameter value  $(-10, 10)$ , which is one of the most challenging cases. These results clearly demonstrate that the TGPT-PINN is highly effective for this 2D (in space) case.

## 5 Conclusion

We have introduced and investigated TGPT-PINN, a physics-informed nonlinear model reduction framework. By combining the practical efficacy of PINNs-based PDE solutions with model reduction using the GPT-PINN template and introducing new discontinuity-approximating strategies and nonlinear transform layers, the TGPT-PINN can overcome the limitations of linear model reduction in the transport-dominated regime and is effective on a wide range of practical PDE problems.

## References

- [1] J. Barnett, C. Farhat, and Y. Maday. Neural-network-augmented projection-based model order reduction for mitigating the Kolmogorov barrier to reducibility. *Journal of Computational Physics*, 492:112420, 2023.
- [2] M. Barrault, N. C. Nguyen, Y. Maday, and A. T. Patera. An “empirical interpolation” method: Application to efficient reduced-basis discretization of partial differential equations. *C. R. Acad. Sci. Paris, Série I*, 339:667–672, 2004.
- [3] B. Battisti, T. Blickhan, G. Enchery, V. Ehrlacher, D. Lombardi, and O. Mula. Wasserstein model reduction approach for parametrized flow problems in porous media. *ESAIM: Proceedings and Surveys*, 73:28–47, 2023.
- [4] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind. Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 18:1–43, 2018.

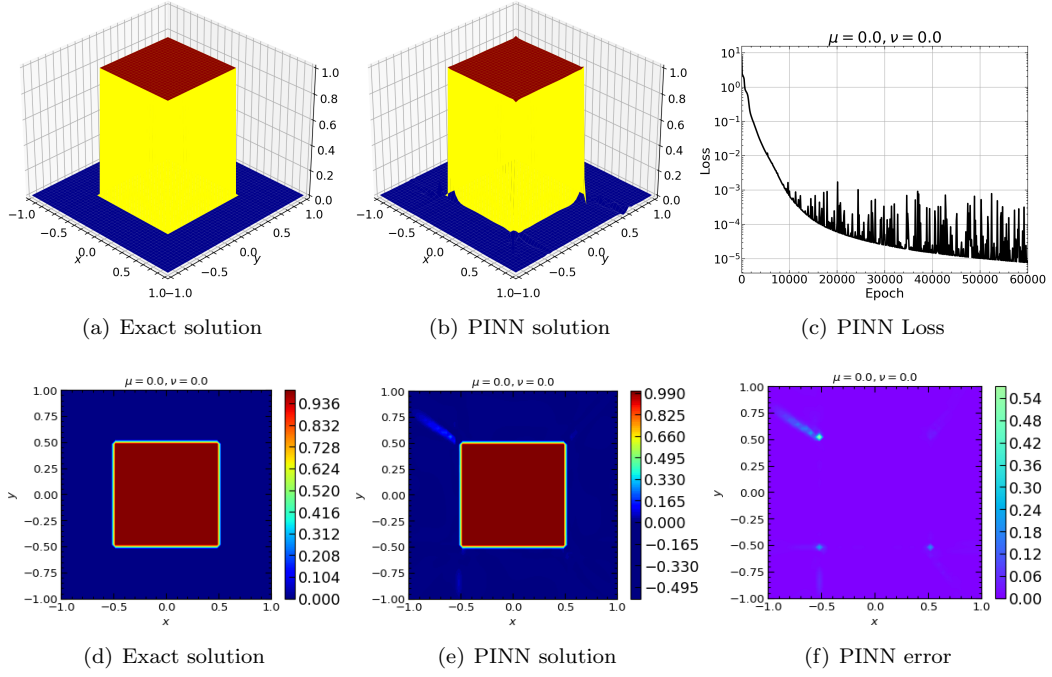


Figure 20: PINN results for Section 4.2.4.

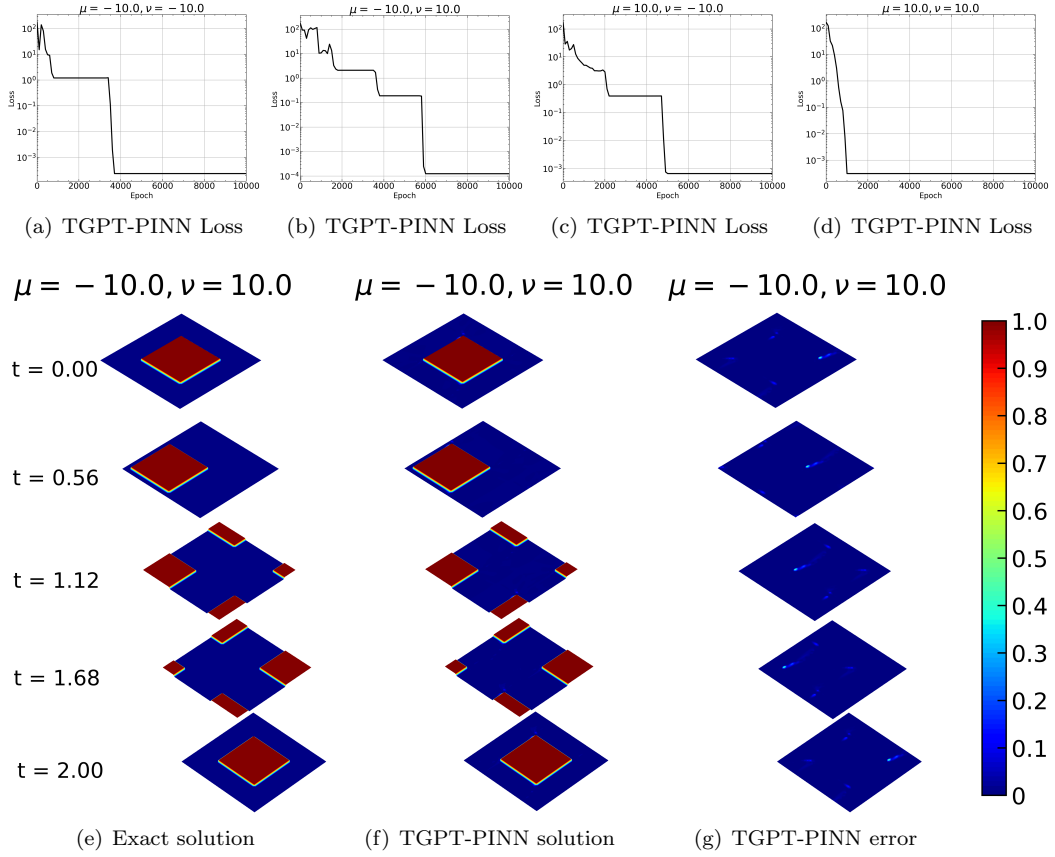
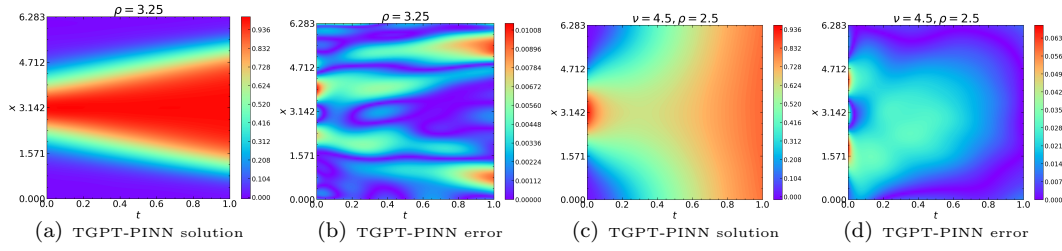


Figure 21: TGPT-PINN results for Section 4.2.4. The history of convergence for the loss at four extremal parameter values (top) and the snapshots along time for the exact and approximate solutions together with the error (bottom).



- [5] P. Binev, A. Cohen, W. Dahmen, R. Devore, G. Petrova, and P. Wojtaszczyk. Convergence rates for greedy algorithms in reduced basis methods. *SIAM Journal on Mathematical Analysis*, 43(3):1457–1472, 2011.
- [6] A. Buffa, Y. Maday, A. T. Patera, C. Prud’homme, and G. Turinici. A priori convergence of the greedy algorithm for the parametrized reduced basis method. *ESAIM: Mathematical Modelling and Numerical Analysis*, 46(3):595–603, 2012.
- [7] N. Cagniard, Y. Maday, and B. Stamm. Model order reduction for problems with large convection effects. *Contributions to Partial Differential Equations and Applications*, pages 131–150, 2019.
- [8] Y. Chen and S. Koohy. GPT-PINN: Generative pre-trained physics-informed neural networks toward non-intrusive meta-learning of parametric PDEs. *Finite Elements in Analysis and Design*, 228:104047, 2024.
- [9] A. Cohen and R. DeVore. Approximation of high-dimensional parametric PDEs. *Acta Numerica*, 24:1–159, 2015.
- [10] W. Dahmen. Compositional sparsity, approximation classes, and parametric transport equations. *arXiv:2207.06128*, 2022.
- [11] V. Ehrlacher, D. Lombardi, O. Mula, and F.-X. Vialard. Nonlinear model reduction on metric spaces. Application to one-dimensional conservative PDEs in Wasserstein spaces. *ESAIM: Mathematical Modelling and Numerical Analysis*, 54(6):2159–2197, 2020.
- [12] N. Franco, A. Manzoni, and P. Zunino. A deep learning approach to reduced order modelling of parameter dependent partial differential equations. *Mathematics of Computation*, 92(340):483–524, 2023.
- [13] R. Geelen, S. Wright, and K. Willcox. Operator inference for non-intrusive model reduction with quadratic manifolds. *Computer Methods in Applied Mechanics and Engineering*, 403:115717, 2023.
- [14] M. A. Grepl, Y. Maday, N. C. Nguyen, and A. T. Patera. Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations. *Mathematical Modelling and Numerical Analysis*, 41(3):575–605, 2007.
- [15] A. Iollo and D. Lombardi. Advection modes by optimal mass transfer. *Physical Review E*, 89(2):022923, 2014.
- [16] Y. Kim, Y. Choi, D. Widemann, and T. Zohdi. Efficient nonlinear manifold reduced order model. *arXiv:2011.07727*, 2020.
- [17] Y. Kim, Y. Choi, D. Widemann, and T. Zohdi. A fast and accurate physics-informed neural network reduced order model with shallow masked autoencoder. *Journal of Computational Physics*, 451:110841, 2022.

- [18] P. Krah, M. Sroka, and J. Reiss. Model order reduction of combustion processes with complex front dynamics. In Numerical Mathematics and Advanced Applications ENUMATH 2019: European Conference, Egmond aan Zee, The Netherlands, September 30–October 4, pages 803–811. Springer, 2020.
- [19] K. Lee and K. T. Carlberg. Model reduction of dynamical systems on nonlinear manifolds using deep convolutional autoencoders. Journal of Computational Physics, 404:108973, 2020.
- [20] L. Liu, S. Liu, H. Xie, F. Xiong, T. Yu, M. Xiao, L. Liu, and H. Yong. Discontinuity computing using physics-informed neural networks. Journal of Scientific Computing, 98(1):22, 2024.
- [21] N. J. Nair and M. Balajewicz. Transported snapshot model order reduction approach for parametric, steady-state fluid flows containing parameter-dependent shocks. International Journal for Numerical Methods in Engineering, 117(12):1234–1262, 2019.
- [22] M. Oehlberger and S. Rave. Reduced basis methods: Success, limitations and future challenges. arXiv:1511.02021, 2015.
- [23] D. Papapicco, N. Demo, M. Girfoglio, G. Stabile, and G. Rozza. The neural network shifted-proper orthogonal decomposition: A machine learning approach for non-linear reduction of hyperbolic equations. Computer Methods in Applied Mechanics and Engineering, 392:114687, 2022.
- [24] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. In 31st Conference on Neural Information Processing Systems, 2017.
- [25] B. Peherstorfer. Model reduction for transport-dominated problems via online adaptive bases and adaptive sampling. SIAM Journal on Scientific Computing, 42(5):A2803–A2836, 2020.
- [26] B. Peherstorfer and K. Willcox. Online adaptive model reduction for nonlinear systems via low-rank updates. SIAM Journal on Scientific Computing, 37(4):A2123–A2150, 2015.
- [27] A. Pinkus. N-Widths in Approximation Theory, volume 7. Springer Science & Business Media, 2012.
- [28] M. Raissi, P. Perdikaris, and G. 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.
- [29] J. Reiss, P. Schulze, J. Sesterhenn, and V. Mehrmann. The shifted proper orthogonal decomposition: A mode decomposition for multiple transport phenomena. SIAM Journal on Scientific Computing, 40(3):A1322–A1344, 2018.
- [30] D. Rim, B. Peherstorfer, and K. T. Mandli. Manifold approximations via transported subspaces: Model reduction for transport-dominated problems. SIAM Journal on Scientific Computing, 45(1):A170–A199, 2023.
- [31] T. Taddei. A registration method for model order reduction: Data compression and geometry reduction. SIAM Journal on Scientific Computing, 42(2):A997–A1027, 2020.
- [32] G. Welper. Interpolation of functions with parameter dependent jumps by transformed snapshots. SIAM Journal on Scientific Computing, 39(4):A1225–A1250, 2017.
- [33] L. Z. Zhao, X. Ding, and B. A. Prakash. PINNsFormer: A transformer-based framework for physics-informed neural networks. arXiv:2307.11833, 2023.