Data-driven forward-inverse problems and modulational instability for Yajima-Oikawa system using deep learning with parameter regularization *

Juncai Pu[a], Yong Chen[a,b,*]

[a] *School of Mathematical Sciences, Shanghai Key Laboratory of Pure Mathematics and Mathematical Practice, East China Normal University, Shanghai, 200241, China*

[b] *College of Mathematics and Systems Science, Shandong University of Science and Technology, Qingdao, 266590, China*

Abstract

We investigate data-driven forward-inverse problems for Yajima-Oikawa system by employing two technologies which improve the performance of PINN in deep physics-informed neural network (PINN), namely neuron-wise locally adaptive activation functions and L^2 norm parameter regularization. In particular, we not only recover three different forms of vector rogue waves (RWs) in the forward problem of Yajima-Oikawa (YO) system, including bright-bright RWs, intermediate-bright RWs and dark-bright RWs, but also study the inverse problem of YO system by data-driven with noise of different intensity. Compared with PINN method using only locally adaptive activation function, the PINN method with two strategies shows amazing robustness when studying the inverse problem of YO system with noisy training data, that is, the improved PINN model proposed by us has excellent noise immunity. The asymptotic analysis of wavenumber k and the MI analysis for YO system with unknown parameters are derived systematically by applying the linearized instability analysis on plane wave.

Key words: *L^2 norm parameter regularization, improved PINN, rogue waves, Yajima-Oikawa system, modulation instability*

PACS numbers: 02.30.Ik, 05.45.Yv, 07.05.Mh.

*Corresponding authors.

E-mail addresses: ychen@sei.ecnu.edu.cn (Y. Chen)

1 Introduction

With the revolution of computer hardware equipment and software technology again and again, the increasing amount of data, model scale, accuracy, complexity and impact on the real world promote the continuous and successful application of deep learning in more and more practical problems [1, 2]. Currently, deep learning has achieved remarkable success in practical problems in various fields. In the field of object recognition, modern object recognition network can not only recognize at least 1000 different categories of objects, but also process rich high-resolution photographs without cropping photos near the objects to be recognized [3, 4]. The introduction of deep learning has a great impact on speech recognition, which makes the error rate of speech recognition drop sharply [5, 6]. Furthermore, deep networks have also had spectacular successes for pedestrian detection and image segmentation [7, 8], as well as yielded superhuman performance in traffic sign classification [9,10]. Moreover, deep learning detonated a wide range of landing applications, such as language understanding [11], medical imaging [12], face recognition [13], video surveillance [14] and mathematical physics [15].

Deep feedforward network, also often called feedforward neural network (NN) or multilayer perceptron, is the quintessential deep learning model [16]. The universal approximation theorem points out a feedforward NN with a linear output layer and at least one hidden layer with any "squashing" activation function can approximate any Borel measurable function from one finite-dimensional space to another with any desired non-zero amount of error, provided that the network is given enough hidden units [17]. With the successful use of back-propagation in training deep NNs with internal representation and the popularity of back-propagation algorithms [18], many optimization methods based on the idea of calculating gradient in back-propagation came into being, such as stochastic gradient descent [19], Adam [20] and L-BFGS [21]. Furthermore, automatic differentiation (AD), also called algorithmic differentiation or simply "autodiff", is a family of techniques similar to but more general than back-propagation for efficiently and accurately evaluating derivatives of numeric functions expressed as computer programs [22]. Recently, due to the general approximation ability of NN architecture [23] and a wide range of AD technology, after taking the NN space as a ansatz space for the solution of governing equation, a physics-informed neural network (PINN) has been successfully constructed to accurately solve both the forward problems, where the approximate solutions of governing partial differential equations are obtained, as well as the inverse problems, where parameters involved in the governing equation are discovered from the training data [15,24]. The PINN framework has been recently successfully applied to many physical problems, including discovering turbulence models from scattered/noisy measurements [25], fractional differential equations [26], high speed aerodynamic flows [27], stochastic differential equation by generative adversarial networks [28] and seeking the localized waves [29].

The type selection of hidden units is extremely significant and difficult in the design process of NN, and the activation function plays an important role during the selection of hidden units due to the derivative of the loss function depends on the optimization parameters, which depend on the derivative of the activation function [16]. The activation function of common hidden units in NNs usually acts on affine transformation, and popular activation functions include rectified linear units [30], maxout units [31], logistic sigmoid

and hyperbolic tangent function. Moreover, some unpublished activation functions perform just as well as the popular ones, such as sine and cosine functions. The choice of activation function completely depends on the problem at hand in practical application. However, these activation functions are fixed in the training process of NN, which will greatly limit the performance of NN and the convergence speed of objective function. Based on the basic framework of PINN, Jagtap et al. proposed two different adaptive activation functions, that is global adaptive activation function and locally adaptive activation function, to approximate smooth and discontinuous functions as well as solutions of linear and nonlinear partial differential equations by introducing scalable parameters into the activation function and adding a slope recovery term based on activation slope to the loss function of locally adaptive activation functions, it proved that the locally adaptive activation function further improves the performance of the NN and speeds up the training process of the NN [32, 33]. Remarkably, we have utilized the PINN with neuron-wise locally adaptive activation function to simulate abundant localized waves of the derivative Schrödinger equation, and these numerical results showcase that the PINN method is a promising and powerful method to increase the efficiency, robustness and accuracy of the NN-based approximation [34].

As is known to all, a central problem in machine learning is how to make an algorithm that will perform well not just on the training data, but also on new inputs. Many strategies, which are known collectively as regularization, are explicitly designed to reduce the test error in machine learning, possibly at the expense of increased training error. Indeed, regularization has been used for decades prior to the advent of deep learning [35]. Many regularization approaches are based on limiting the capacity of models, such as NNs, linear regression, or logistic regression, by adding a parameter norm penalty $\Omega(\boldsymbol{\theta})$ to the objective function, of which the most common and simplest is L^2 parameter norm regularization. L^2 parameter norm penalty is usually called weight decay, ridge regression [36] or Tikhonov regularization [37], which drives the weight closer to the origin by adding a regularization term to the objective function. Therefore, a natural inspiration is to further enhance the performance of the novel PINN by introducing the parameter regularization strategy into the aforementioned PINN.

Rogue wave (RW), which appears suddenly and disappear without a trace [38, 39], alternatively called wave of extremely large size, freak or giant waves, was originally coined for vividly describing the mysterious and monstrous large amplitude ocean wave [40, 41]. Due to RWs could appear in any place of the world ocean and have unpredictable characteristics [42, 43], RWs are a well-documented hazards for mariners, and these waves are responsible for loss of many ships and many human lives [44]. Recently, the study of RW has attracted extensive attention in its fundamental origin and complex dynamics [45, 46]. Hitherto, in addition to in the oceanographic background, these extreme wave events are also been observed and investigated in a wide class of spatial-temporal continuous systems including water tank [47], nonlinear optics [48], Bose-Einstein condensates [49], ultra-cold bosonic gases [50], capillary waves and surface ripples [51, 52], atmosphere [53], microwave transport [54], plasma [55,56], versatile lasers [57] and even financial systems [58]. With the rapid development of deep learning, predicting the generation and evolution of long-time RWs by observing initial boundary value data plays an important role in the these aforementioned disciplines. Recently, employing the PINN with neuron-wise locally adaptive

activation function, abundant localized waves of the derivative Schrödinger equation are successfully recovered, including soliton, rational soliton, RW, periodic wave and periodic rogue wave [59].

Due to many physical systems include interacting wave components with different modes, frequencies or polarizations, another important development is the study of coupled wave systems. Compared with scalar dynamical systems, vector systems usually allow energy transfer between their additional degrees of freedom, which potentially generates families of intricate vector RWs. Indeed, considering RW phenomena in various complex systems, such as optical fiber [60], financial systems [61] and Bose-Einstein condensates [62], it is necessary to consider multiple amplitudes rather than a single amplitude, that is, the coupled system can describe extreme waves more accurately than the scalar model. Among coupled field dynamics systems, the coupled long wave-short wave resonance equation (LSWR) [63] is a fascinating nonlinear physical system, it describes a resonant interaction between long wave in complex envelope of rapidly varying field and short wave in real low-frequency field. Once the resonance condition is satisfied, that is, the group velocity of a short wave (high-frequency wave) exactly or almost matches the phase velocity of a long wave (low-frequency wave), this coupled system can be derived from the Davey-Stewartson system [64]. In 1972, Zakharov made a theoretical investigation for LSWR for the first time when analyzing the Langmuir wave in plasma [65]. In the case of long wave unidirectional propagation, the general Zakharov system was reduced to LSWR system, which is usually called one-dimensional Yajima-Oikawa (YO) system [66]. Surprisingly, despite its simple form, this system can describe various nonlinear wave phenomena, such as capillary-gravity wave in fluid [64], optical-terahertz waves in second-order nonlinear negative refractive index medium [67] and between long and short internal waves [68], as well as between a long internal wave and a short surface wave in a two layer fluid [69]. Recently, in order to build new PINN suitable for coupled systems, we have proposed an PINN algorithm with four output functions and four nonlinear equation constraints to obtain the data-driven vector localized waves including vector solitons, breathers and RWs of Manakov system in complex space [70]. The next our goal is to establish a more powerful PINN method for such rogue waves of YO system and further reveal an interesting cross dynamics, ranging from bright rogue waves to dark counterparts.

In this paper, we will use the regularization strategy and the PINN with locally adaptive activation function to construct an improved PINN with three outputs and three physical constraints for studying the initial boundary value problem of YO system as follow

$$\begin{cases} \mathrm{i}S_t + \lambda_1 S_{xx} + SL = 0,\ x \in [X_0, X_1],\ t \in [T_0, T_1], \\ \mathrm{i}L_t - \lambda_2 (|S|^2)_x = 0,\ x \in [X_0, X_1],\ t \in [T_0, T_1], \\ S(x, T_0) = S^0(x),\ L(x, T_0) = L^0(x),\ x \in [X_0, X_1], \\ S(X_0, t) = S^{\mathrm{lb}}(t),\ S(X_1, t) = S^{\mathrm{ub}}(t),\ t \in [T_0, T_1], \\ L(X_0, t) = L^{\mathrm{lb}}(t),\ L(X_1, t) = L^{\mathrm{ub}}(t),\ t \in [T_0, T_1], \end{cases} \tag{1.1}$$

where the short wave component $S(x,t)$ stands for the complex envelope of the rapidly varying field and the long wave component $L(x,t)$ represents the real low-frequency field, with x and t the two independent evolution variables. Here, λ_1 and λ_2 are real valued

parameters, which can be known parameters or undetermined parameters. The $|S|$ represents the modulus of complex valued short wave S, which also means $|S|^2 = SS^*$ with S^* indicates the conjugate of S. For physics discussions, The first equation of YO system, namely the first formula of Eq. (1.1), is arranged in a form similar to the standard nonlinear Schrödinger equation, which clearly indicates that its nonlinearity is driven by long wave field L rather than Kerr term $|S|^2$. We note that the RWs of this system have been effectively obtained with the aid of Hirota bilinear method [71], KP hierarchy reduction method [72] and Darboux transformation [73].

This paper is organized as follows: After the introduction in section 1, section 2 gives a brief discussion of the improved PINN methodology with locally adaptive activation function and L^2 parameter norm regularization for the coupled YO systems, where we also discuss about training data, loss function, optimization methods and the operating environment. The algorithm flow schematic and algorithm steps for the YO system are also exhibited in detail. Section 3 provides the results and detailed discussions for forward problems on improved PINN approximations of data driven vector RWs in three different states. Section 4 presents experimental results with different trade-off norm penalty term coefficients in inverse problems. In Section 5, we systematically introduce the general MI analysis of YO system with unknown parameters λ_1 and λ_2. Finally, we summarize the conclusions of our work are given out in last section.

2 Methodology

From Ref. [34], one can know about the original PINN method could not accurately reconstruct some solutions with complex forms in some complicated nonlinear systems, and the PINN approach with neuron-wise locally adaptive activation function and slope recovery term can improve the convergence speed and stability of the loss function in the training process. Therefore, considering the training accuracy, performance requirements and structural complexity of multi-component coupled nonlinear systems, we further improve the deep learning algorithm by introducing parameter regularization based on the aforementioned PINN method.

2.1 NN and adaptive activation function

We establish a NN of depth D with an input layer, $D-1$ hidden-layers and an output layer, in which the dth hidden-layer contain N_d number of neurons. Each hidden-layer of the NN receives an output $\mathbf{x}^{d-1} \in \mathbb{R}^{N_{d-1}}$ from the previous layer, where an affine transformation can be written as follows form

$$\mathcal{L}_d(\mathbf{x}^{d-1}) \triangleq \mathbf{W}^d \mathbf{x}^{d-1} + \mathbf{b}^d, \tag{2.1}$$

where the network weights $\mathbf{W}^d \in \mathbb{R}^{N_d \times N_{d-1}}$ and bias term $\mathbf{b}^d \in \mathbb{R}^{N_d}$ associated with the dth layer. Specifically, in order to introduce adaptive activation function, we define such neuron-wise locally adaptive activation function as

$$\sigma\left(na_i^d\left(\mathcal{L}_d\left(\mathbf{x}^{d-1}\right)\right)_i\right),\ d = 1, 2, \cdots, D-1,\ i = 1, 2, \cdots, N_d,$$

where σ is the activation function, and $n > 1$ is a scaling factor and $\{a_i^d\}$ are additional $\sum_{d=1}^{D-1} N_d$ parameters to be optimized. Note that, there is a critical scaling factor n_c, and the optimization algorithm will become sensitive when $n \geqslant n_c$ in each problem set [33]. The neuron-wise locally activation function acts as a vector activation function in each hidden layer, and each neuron has its own slope of activation function. The improved NN with neuron-wise locally adaptive activation function can be represented as

$$\mathbf{q}(\mathbf{x};\bar{\Theta}) = \left((\mathcal{L}_D)_{i'} \circ \sigma \circ n a_i^{D-1} (\mathcal{L}_{D-1})_i \circ \cdots \circ \sigma \circ n a_i^1 (\mathcal{L}_1)_i \right) (\mathbf{x}),\ i' = 1, 2, 3, \tag{2.2}$$

where $\mathbf{x}$ and $q(\mathbf{x};\bar{\Theta})$ represent the two inputs and three outputs in the NN, respectively. The set of trainable parameters $\bar{\Theta} \in \bar{\mathcal{P}}$ consists of $\left\{\mathbf{W}^d, \mathbf{b}^d\right\}_{d=1}^{D}$ and $\left\{a_i^d\right\}_{i=1}^{D-1}, \forall i = 1, 2, \cdots, N_d$, $\bar{\mathcal{P}}$ is the parameter space.

2.2 YO system constraint

Especially, we consider the (1 + 1)-dimensional coupled YO system as the physical constraint of aforementioned NN to construct PINN, its specific operator representation form for YO system (1.1) is as shown below

$$\begin{aligned} &S_t + \mathcal{N}[S, L] = 0, \\ &L_t + \mathcal{N}'[S] = 0, \end{aligned} \tag{2.3}$$

where S and L are complex valued solution and real valued solution of x and t to be determined later respectively, $\mathcal{N}[\cdot,\cdot]$ and $\mathcal{N}'[\cdot]$ are nonlinear differential operators in space. Due to the complexity of the structure of the complex-valued solution $S(x,t)$ in Eq. (2.2), we decompose $S(x,t)$ into the real part $u(x,t)$ and the imaginary part $v(x,t)$ by employing real-valued functions $u(x,t)$ and $v(x,t)$, that is $S(x,t) = u(x,t) + \mathrm{i}v(x,t)$. Then substituting it into Eq. (2.3), we have

$$u_t + \mathcal{N}_u[u, v, L] = 0, \quad v_t + \mathcal{N}_v[u, v, L] = 0, \quad L_t + \mathcal{N}'_L[u, v] = 0. \tag{2.4}$$

Accordingly, the $\mathcal{N}_u$, $\mathcal{N}_v$ and $\mathcal{N}'_L$ are nonlinear differential operators in space. Then $f_u(x,t)$, $f_v(x,t)$ and $f_L(x,t)$ constitute the physics-informed parts of the NN, which can be defined as

$$f_u := u_t + \mathcal{N}_u[u, v, L], \quad f_v := v_t + \mathcal{N}_v[u, v, L], \quad f_L := L_t + \mathcal{N}'_L[u, v]. \tag{2.5}$$

2.3 Loss function and parameter regularization

We will attempt to find the optimized parameters, including the weights, biases and additional coefficients in the activation, to minimize two new loss functions $\mathscr{L}(\bar{\Theta})$ and $\widetilde{\mathscr{L}}(\bar{\Theta})$ with weights parameter regularization, which are defined as the following forms respectively

$$\begin{aligned} &\mathscr{L}(\bar{\Theta}) = Loss = Loss_S + Loss_L + Loss_{f_S} + Loss_{f_L} + Loss_a, \\ &\widetilde{\mathscr{L}}(\bar{\Theta}) = Loss_{PR} = \mathscr{L}(\bar{\Theta}) + \alpha\Omega(\bar{\Theta}), \end{aligned} \tag{2.6}$$

where $Loss_S, Loss_L, Loss_{f_S}$ and $Loss_{f_L}$ are defined as following

$$\begin{aligned} Loss_S &= \frac{1}{N_q}\left[\sum_{j=1}^{N_q}|\hat{u}(x^j,t^j)-u^j|^2+\sum_{j=1}^{N_q}|\hat{v}(x^j,t^j)-v^j|^2\right], \\ Loss_L &= \frac{1}{N_q}\sum_{j=1}^{N_q}|\hat{L}(x^j,t^j)-L^j|^2, \end{aligned} \tag{2.7}$$

and

$$\begin{aligned} Loss_{f_S} &= \frac{1}{N_f}\left[\sum_{l=1}^{N_f}|f_u(x_f^l,t_f^l)|^2+\sum_{l=1}^{N_f}|f_v(x_f^l,t_f^l)|^2\right], \\ Loss_{f_L} &= \frac{1}{N_f}\sum_{l=1}^{N_f}|f_L(x_f^l,t_f^l)|^2, \end{aligned} \tag{2.8}$$

where $\{x^j,t^j,u^j,v^j,L^j\}_{j=1}^{N_q}$ denotes the initial and boundary value inputs data on Eqs. (2.4) and (2.5). Here $\hat{u}(x^j,t^j),\hat{v}(x^j,t^j)$ and $\hat{L}(x^j,t^j)$ represent the optimal training outputs data through the NN. Furthermore, $\{x_f^l,t_f^l\}_{l=1}^{N_f}$ represent the collocation points on networks $f_u(x,t)$, $f_v(x,t)$ and $f_L(x,t)$. The last slope recovery term $Loss_a$ in the loss function (2.6) is defined as

$$Loss_a = \frac{1}{\frac{N_a}{D-1}\sum\limits_{d=1}^{D-1}\exp\left(\frac{\sum\limits_{i=1}^{N_d}a_i^d}{N_d}\right)}, \tag{2.9}$$

where $1/N_a$ is the hyperparameter for slope recovery term $Loss_a$, and we all take $N_a = 100$ for dominating the loss function and ensuring that the final loss value is not too large in this paper. Here, $Loss_a$ term forces the NN to increase the activation slope value quickly, which ensures the non-vanishing of the gradient of the loss function and improves the network's training speed. Consequently, $Loss_S$ and $Loss_L$ correspond to the loss on the initial and boundary data, the $Loss_{f_S}$ and $Loss_{f_L}$ penalizes the YO system not being satisfied on the collocation points, and the $Loss_a$ changes the topology of $Loss$ function and improves the convergence speed and network optimization ability.

Furthermore, α is a hyperparameter that weights the relative contribution of the norm penalty term Ω and loss function $\mathscr{L}(\bar{\Theta})$, and the L^2 parameter norm penalty $\Omega(\bar{\Theta})$ can be defined as following

$$\Omega(\bar{\Theta}) = \frac{1}{2}\|\mathbf{W}\|_2^2, \tag{2.10}$$

which drives the weights closer to the origin. Due to the biases typically require less data to fit accurately than the weights, we note that for NNs, we typically choose to use a parameter norm penalty Ω that penalizes only the weights of the affine transformation at each layer and leaves the biases unregularized. It is worth mentioning that the aforementioned slope recovery term $Loss_a$ can be regarded as a self-defined parameter regularization strategy for additional parameters $\{a_i^d\}$.

2.4 Optimization algorithm and improved PINN

The resulting optimization problem leads to finding the minimum value of the loss function by optimizing the parameters $\bar{\Theta}$, that is, we seek

$$\bar{\Theta}^* = \arg\min_{\bar{\Theta}\in\bar{\mathcal{P}}} \widetilde{\mathscr{L}}(\bar{\Theta}).$$

Generally, one can approximate the solutions to this minimization problem iteratively by one of the forms of gradient descent algorithm. The stochastic gradient descent (SGD) and its variants are probably the most used optimization algorithms for machine learning in general and for deep learning in particular [19]. In this work, we introduce Adam optimizer and L-BFGS optimizer to optimize the loss function. Specifically, we employ the Adam optimizer, which is a variant of the SGD algorithm, and the L-BFGS optimizer, which is a full-batch gradient descent optimization algorithm based on a quasi-Newton method to optimize the loss function [20, 21]. Moreover, in order to better measure the training error, we introduce $\mathbb{L}_2$ norm error, which is defined as follows

$$\text{Error} = \frac{\sqrt{\sum_{k=1}^{N} \left|q^{\text{exact}}(\mathbf{x}_k) - q^{\text{predict}}(\mathbf{x}_k;\bar{\Theta})\right|^2}}{\sqrt{\sum_{k=1}^{N} \left|q^{\text{exact}}(\mathbf{x}_k)\right|^2}},$$

where $q^{\text{predict}}(\mathbf{x}_k;\bar{\Theta})$ and $q^{\text{exact}}(\mathbf{x}_k)$ represent the model training prediction solution and exact analytical solution at point $\mathbf{x}_k = (x_k, t_k)$, respectively.

In order to understand the improved PINN approach more clearly, the improved PINN algorithm flow chart of the YO system is shown in following Fig.1, where one can see the NN along with the supplementary physics-informed part, and the loss function is evaluated using the contribution from the NN part as well as the residual from the governing equation given by the physics-informed part. Then, one seeks the optimal values of weights $\mathbf{W}$, biases $\mathbf{b}$ and scalable parameter a_i^d in order to minimize the loss function below certain tolerance ε until a prescribed maximum number of iterations. From Fig. 1, since the YO system contains two components $S(x,t)$ and $L(x,t)$, one can see that the "NN" part has three output functions $\{u, v, L\}$, and there are three nonlinear equation constraints in the "PDE" part, that is, in terms of the nonlinear coupled system with more components, the number of output functions and nonlinear equation constraints of the improved PINN will increase exponentially. Furthermore, in order to further understand the improved PINN, we also showcase the corresponding procedure steps of the improved PINN with adaptive activation function and L^2 norm parametric regularization in the following Tab. 1.

2.5 Training data and network environment

In supervised learning, training data is important to train the NN, which can be obtained from the exact solution (if available) or from high-resolution numerical solution using numerical methods like spectral method, finite element method, Chebfun numerical method, discontinuous Galerkin method etc, as per the problem at hand. Furthermore, training

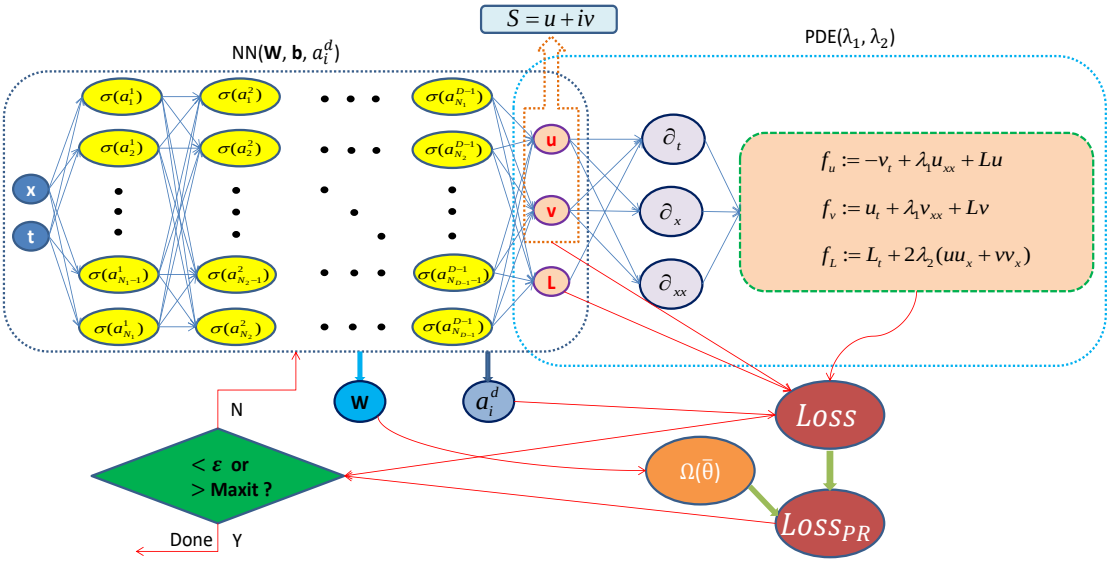


Figure 1: (Color online) Schematic of improved PINN for the YO system. The left NN is the uninformed network while the right one induced by the governing equation is the informed network. The two NNs share hyper-parameters and they both contribute to the loss function.

Table 1: Improved PINN algorithm with adaptive activation function and L^2 norm parameter regularization.

Step 1: Specification of training set in computational domain:
Training data: $\{x^j, t^j, u^j, v^j, L^j\}_{j=1}^{N_q}$, *Residual training points*: $\{x_f^l, t_f^l\}_{l=1}^{N_f}$.

Step 2: Construct neural network $\mathbf{q}(\mathbf{x}; \bar{\Theta})$ with random initialization of parameters $\bar{\Theta}$.

Step 3: Construct the residual neural network $\{f_u, f_v, f_L\}$ by substituting surrogate $\mathbf{q}(\mathbf{x}; \bar{\Theta})$ into the governing equations using automatic differentiation and other arithmetic operations.

Step 4: Specification of the loss function $\widetilde{\mathscr{L}}(\bar{\Theta})$ that includes the slope recovery term and parameter regularization term.

Step 5: Find the best parameters $\bar{\Theta}^*$ using a suitable optimization method for minimizing the loss function $\widetilde{\mathscr{L}}(\bar{\Theta})$ as

$$\bar{\Theta}^* = \arg\min_{\bar{\Theta} \in \bar{\mathcal{P}}} \mathscr{L}(\bar{\Theta}).$$

data can also be obtained from carefully performed physical experiments that may produce high- and low-fidelity data sets. Fortunately, the nonlinear integrable systems like YO system has very good integrability, and there are some effective methods to solve a variety of accurate localized wave solutions, which provide rich sample space for the extraction of training data. Here, we use the traditional finite difference scheme on the even grid to discretize these exact solutions to obtain the training data. Moreover, one can also add uniform/normal distribution random disturbance to the exact solution to select the training point, which may obtain some interesting physical phenomena and also provide a powerful numerical tool for the study of modulation instability.

In the adaptive activation function, the initialization of scalable parameters are carried out in the case of $n = 10, a_i^d = 0.1$, namely $na_i^d = 1$. In addition, we select relatively simple multi-layer perceptrons (i.e., feedforward NNs) with the Xavier initialization and the hyperbolic tangent (tanh) as activation function. All the codes in this article is based on Python 3.7 and Tensorflow 1.15, and all numerical experiments reported here are run on a DELL Precision 7920 Tower computer with 2.10 GHz 8-core Xeon Silver 4110 processor, 64 GB memory and 11 GB Nvidia GeForce GTX 1080 Ti video card.

3 The forward problem of the YO system

In this section, we will focus on the forward problem of the YO system, that is reveal the data-driven RWs for the YO system by means of small data set and 9 hidden layers deep improved PINN with 40 neurons per layer. Specifically, after knowing the determined unknown parameters of the YO system (1.1), which are $\lambda_1 = 0.5$ and $\lambda_2 = 1$, as well as some initial boundary value data points, we can successfully approximate the various RWs through improved PINN with parameter regularization and hyperparameter $\alpha = 0.0001$. Here the physics-informed parts Eq. (2.5) of the improved PINN for YO sysem (1.1) become the following formula

$$f_u := -v_t + 0.5u_{xx} + uL, \quad f_v := u_t + 0.5v_{xx} + vL, \quad f_L := L_t - (2uu_x + 2vv_x). \tag{3.1}$$

The exact form of these RWs for the YO system (1.1) with $\lambda_1 = 0.5$ and $\lambda_2 = 1$ have been derived by the Darboux transformation [73], the general vector form of rogue waves can be expressed as

$$\begin{aligned} S(x,t) &= a\mathrm{e}^{\mathrm{i}kx-\mathrm{i}\left(\frac{1}{2}k^2-b\right)t}\left[1-\frac{\mathrm{i}t+(\mathrm{i}x)/(2m-k)+1/\big(2(2m-k)(m-k)\big)}{(x-mt)^2+n^2t^2+1/(4n^2)}\right],\\ L(x,t) &= b+2\frac{n^2t^2-(x-mt)^2+1/(4n^2)}{[(x-mt)^2+n^2t^2+1/(4n^2)]^2}, \end{aligned} \tag{3.2}$$

and m, n, a, b and k satisfy the following relationship

$$\begin{aligned} &m=\frac{1}{6}\big[5k-\sqrt{3(k^2+\eta+\sigma/\eta)}\big],\ n=\pm\sqrt{(3m-k)(m-k)},\\ &\sigma=\frac{1}{9}k^4+6a^2k,\ \rho=\frac{1}{2}k^6-\frac{1}{54}(27a^2+5k^3)^2,\\ &\eta=\begin{cases}-\big(\rho-\sqrt{\rho^2-\sigma^3}\big)^{1/3} & k\leqslant -3k_n,\\ \big(-\rho+\sqrt{\rho^2-\sigma^3}\big)^{1/3} & -3k_n<k\leqslant \frac{3}{2}k_n,\end{cases} \end{aligned} \tag{3.3}$$

where $k_n = (2a^2)^{1/3}$, $a > 0$, $b \geqslant 0$ and $k \in \mathbb{R}$. It is clear that the short-wave RW S is characterized by the second-order polynomial of x and t, while the real long-wave RW L involves the fourth-order polynomial. One can observe that although having a form similar to that of Peregrine soliton [74], the RWs in (3.2) admit more complex dynamics than the latter.

According to the central amplitude and the relative position of two zero-amplitude points, one can divide the regime of the short-wave RWs $S(x,t)$ into three regions, that is bright RW region, intermediate RW region and dark RW region, which are corresponding to parametric conditions $k \leqslant 0$, $0 < k < (4/3)^{1/3}k_n$ (here $(4/3)^{1/3} \approx 1.1$) and $(4/3)^{1/3}k_n \leqslant k < 1.5k_n$, respectively [73].

• **Bright-Bright RWs $S_{\rm brw}$ and $L_{\rm brw1}$**

In order to obtain the bright-bright RWs $S_{\rm brw}$ and $L_{\rm brw1}$ of YO system, one can lead to $m = -\frac{1}{2}$, $n = \pm\frac{\sqrt{3}}{2}$ by taking $a = 1$, $b = 0$, $k = 0$ in Eq. (3.3) and combining the value range of k from the aforementioned results. Substituting the above parameters into the formula Eq. (3.2), the specific form of bright-bright rogue waves $S_{\rm brw}$ and $L_{\rm brw1}$ for YO system is as follows

$$\begin{aligned} S_{\rm brw}(x,t) &= \frac{-3{\rm i}t + 3{\rm i}x + 3t^2 + 3tx + 3x^2 - 2}{3t^2 + 3tx + 3x^2 + 1}, \\ L_{\rm brw1}(x,t) &= \frac{3(3t^2 - 6tx - 6x^2 + 2)}{(3t^2 + 3tx + 3x^2 + 1)^2}. \end{aligned} \tag{3.4}$$

Apparently, the complex short-wave RW $S_{\rm brw}(x,t)$ takes plane $|S_{\rm brw}| = 1$ as the background wave, while real long-wave RW $L_{\rm brw1}(x,t)$ takes plane $L_{\rm brw1} = 1$ as the background wave. Furthermore, $|S_{\rm brw}(x,t)|$ obtains the maximum amplitude at $(x,t) = (0,0)$ and the maximum amplitude is 2, but $L_{\rm brw1}(x,t)$ obtains the maximum amplitude at $(x,t) = (0,0)$ and the maximum amplitude is 6.

• **Intermediate-Bright RWs $S_{\rm irw}$ and $L_{\rm brw2}$**

Similarly, in order to obtain the intermediate-bright RWs $S_{\rm irw}$ and $L_{\rm brw2}$ of YO system, one can also take $a = 1$, $b = 0$, but parameter k have to meet the condition $0 < k < (4/3)^{1/3}k_n$, here $k_n = 2^{1/3}$. In particular, we derive the values of parameters m and n by taking $k = \frac{1}{2}2^{1/3}$, then substitute them into the Eq. (3.2), and then one can obtain the intermediate-bright RWs $S_{\rm irw}$ and $L_{\rm brw2}$ of YO system. Since the forms of expressions for m, n, $S_{\rm irw}$ and $L_{\rm brw2}$ are complex, we omit them here.

• **Dark-Bright RWs $S_{\rm drw}$ and $L_{\rm brw3}$**

In the same way, the dark-bright RWs $S_{\rm drw}$ and $L_{\rm brw3}$ of YO system are obtained by taking $a = 1$, $b = 0$ and making the parameter k satisfy the condition $(4/3)^{1/3}k_n \leqslant k < 1.5k_n$. After taking $k = \frac{6}{5}2^{1/3}$, then we obtain the dark-bright RWs $S_{\rm drw}$ and $L_{\rm brw3}$ of YO system by means of known and derived parameters. Similarly, due to the complex form, we omit the specific expression form here.

Next, we will use the above exact solution to obtain the initial boundary value condition data, so as to construct the training data set. Under different initial boundary value conditions, three different RWs dynamic structures are recovered by utilizing improved PINN with parameter regularization. In order to more intuitively exhibit the training effect of improved PINN with parameter regularization, we also compare it with the training

results from PINN without parameter regularization.

3.1 The data-driven bright-bright RWs

In this section, in order to recover the data-driven bright-bright RWs of the YO system, we will commit to introducing the initial boundary value conditions of the YO system to the 9-layer improved PINN with 40 neurons per layer. Selecting $[X_0, X_1]$ and $[T_0, T_1]$ in Eq. (1.1) as $[-5.0, 5.0]$ and $[-2.0, 2.0]$ respectively, we can write the corresponding initial value conditions as follows

$$S^0(x) = S_{\rm brw}(x, -2.0),\ L^0(x) = L_{\rm brw1}(x, -2.0),\ x \in [-5.0, 5.0], \tag{3.5}$$

and the Dirichlet boundary conditions become

$$\begin{aligned} &S^{\rm lb}(t) = S_{\rm brw}(-5.0, t),\ L^{\rm lb}(t) = L_{\rm brw1}(-5.0, t),\ t \in [-2.0, 2.0], \\ &S^{\rm ub}(t) = S_{\rm brw}(5.0, t),\ L^{\rm ub}(t) = L_{\rm brw1}(5.0, t),\ t \in [-2.0, 2.0]. \end{aligned} \tag{3.6}$$

In order to obtain the original training data set of the above initial boundary value conditions (3.5) and (3.6), we discretize the exact bright-bright RWs $S_{\rm brw}$ and $L_{\rm brw1}$ (3.4) based on the finite difference method by dividing the spatial region $[-5.0, 5.0]$ into 2000 points and the temporal region $[-2.0, 2.0]$ into 1000 points in Matlab. Furthermore, in addition to the data set composed of the aforementioned initial boundary value conditions, the residual data set is used to calculate the $\mathbb{L}_2$ norm error by comparing with the data-driven bright-bright RWs. After that, a smaller training dataset that containing initial-boundary data will be generated by randomly extracting $N_q = 1000$ from original dataset and $N_f = 20000$ collocation points which are produced by the LHS. According to 20000 Adam iterations and 50000 L-BFGS iterations, the latent bright-bright RWs $S(x, t)$ and $L(x, t)$ have been successfully learned by employing the improved PINN with parameter regularization, and the network achieved relative $\mathbb{L}_2$ error of 4.968430e-04 for the bright RW $S(x, t)$ and relative $\mathbb{L}_2$ error of 1.763312e-03 for the bright RW $L(x, t)$, and the total number of iterations is 70000.

Figs. 2 presentS the deep learning results of the data-driven bright-bright RWs based on the improved PINN with parameter regularization for the YO system with the initial boundary value problem (3.5) and (3.6). The left panels of Fig. 2 display the exact, learned and error dynamics for the bright RW $|S(x, t)|$ and bright RW $L(x, t)$, and exhibit the sectional drawings which contain the learned and explicit RWs at five different moments. From the density plots of learned dynamics and profiles which reveal amplitude and error of exact and prediction RWs in Fig. 2, we observe that the amplitude of long-wave RW is much higher than that of short-wave RW. The right panels of Fig. 2 exhibit the 3D plots of the predicted bright-bright RWs. Fig. 3 showcases curve plots of the loss function after 20000 Adam optimization iterations and 50000 L-BFGS optimization iterations in improved PINN framework. In the left panel of Fig. 3, one can see that the loss function curve $Loss_{PR}$ oscillatly descends in the process of optimizing the loss function by means of the Adam optimizer, and the gradient of the loss function descends very fast in the last about 4000 iterations. While, from the right panel of Fig. 3, the loss function curve $Loss_{PR}$ linearly descends by means of the L-BFGS optimization algorithm. Furthermore,

the loss function curves $Loss_S$ and $Loss_L$ of L-BFGS optimization are missing after a certain number of iterations, that is because the loss function value in this part is less than $1e^{-6}$, which is beyond the statistical range of "%f" in Python 3.7. During the process of optimizing the loss function $Loss_{PR}$ both in two optimization algorithms, $Loss_a$ and $\alpha\Omega(\bar{\Theta})$ descend linearly, which depend on their topological structure and mathematical form. That is, $Loss_a$ ensures the better and faster convergence of the loss function by means of PINN algorithm with neuron-wise locally adaptive activation function, and $\alpha\Omega(\bar{\Theta})$ ensures the weight decay continuously by imposing the parameter regularization term.

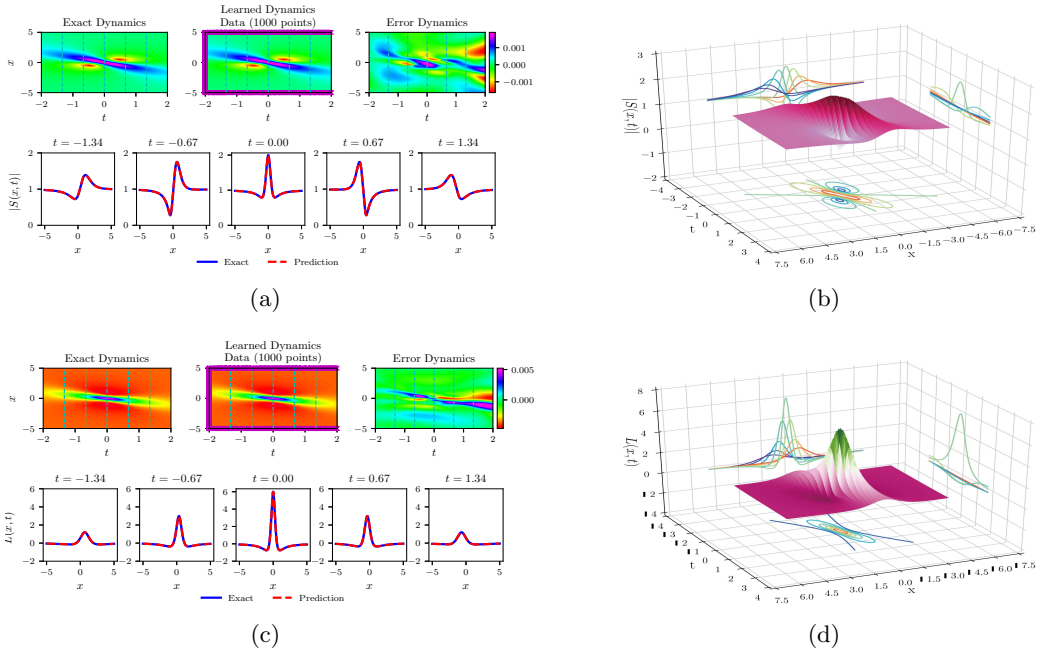
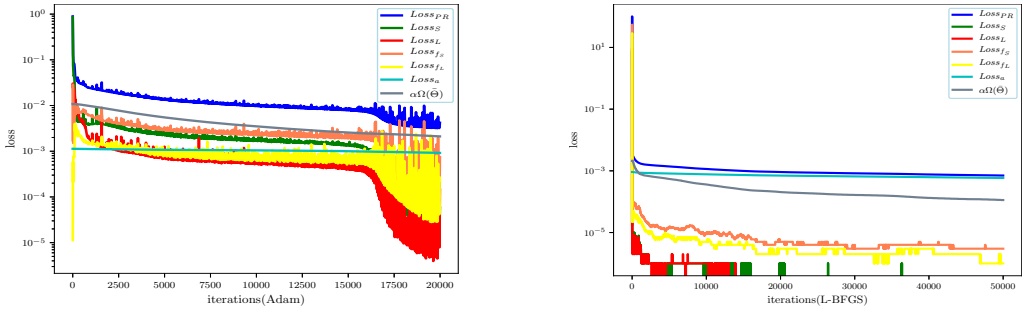


Figure 2: (Color online) The data-driven bright-bright RWs $S(x,t)$ and $L(x,t)$ resulted from the improved PINN with the randomly chosen initial and boundary points $N_q = 1000$ which have been shown by using mediumorchid " $\times$ " in learned dynamics , and $N_f = 20000$ collocation points in the corresponding spatiotemporal region: (a) and (c) The exact, learned and error dynamics density plots with five distinct tested times $t = -1.34, -0.67, 0.00, 0.67$ and 1.34 (darkturquoise dashed lines), as well as sectional drawings which contain the learned and explicit bright-bright RWs at the aforementioned five distinct times; (b) and (d) The 3D plot for the data-driven bright-bright RWs.

3.2 The data-driven intermediate-bright RWs

In what follows, we will consider the initial-boundary value problem of the YO system for obtaining the data-driven intermediate-bright RWs by applying the multilayer improved PINN. Similarly, taking $[X_0, X_1]$ and $[T_0, T_1]$ in Eq. (1.1) as $[-5.0, 5.0]$ and $[-3.0, 3.0]$ respectively, we derive the corresponding initial conditions $S^0(x)$ and $L^0(x)$, and Dirichlet



(a) (b)

Figure 3: (Color online) The loss function curve figures of the bright-bright RWs $S(x,t)$ and $L(x,t)$ arising from the improved PINN with the 20000 steps Adam and 50000 steps L-BFGS optimizations: (a) The loss function curve for the 20000 Adam optimization iterations; (b) The loss function curve for the 50000 L-BFGS optimization iterations.

boundary conditions as shown in the following formulas

$$S^0(x) = S_{\rm irw}(x,-3.0),\ L^0(x) = L_{\rm brw2}(x,-3.0),\ x\in[-5.0,5.0], \tag{3.7}$$

and

$$\begin{aligned} &S^{\rm lb}(t) = S_{\rm irw}(-5.0,t),\ L^{\rm lb}(t) = L_{\rm brw2}(-5.0,t),\ t\in[-3.0,3.0],\\ &S^{\rm ub}(t) = S_{\rm irw}(5.0,t),\ L^{\rm ub}(t) = L_{\rm brw2}(5.0,t),\ t\in[-3.0,3.0]. \end{aligned} \tag{3.8}$$

By means of Matlab, we discretize the exact intermediate-bright RWs $S_{\rm irw}$ and $L_{\rm brw2}$ by utilizing the traditional finite difference scheme on even grids, and obtain the original training data which only contains initial data (3.7) and boundary data (3.8) by dividing the spatial region $[-5.0,5.0]$ into 2000 points and the temporal region $[-3.0,3.0]$ into 1000 points, the remaining data will be used to obtain training errors by comparing with predicted intermediate-bright RWs. After that, we generate a smaller training dataset containing initial-boundary data by randomly extracting $N_q = 2000$ from original training dataset and $N_f = 30000$ collocation points produced via LHS in the corresponding spatiotemporal region. Then, the intermediate-bright RWs $S(x,t)$ and $L(x,t)$ have been successfully learned by imposing a 9-hidden-layer improved PINN with 40 neurons per layer, and the related loss functions are optimized through 20000 Adam iterations and 50000 L-BFGS iterations. The relative $\mathbb{L}_2$ errors of the improved PINN model are 1.168852e-03 for $S(x,t)$ and 6.766132e-03 for $L(x,t)$, the total number of iterations is 70000.

Figs. 4 - 5 display the training results of the data-driven intermediate-bright RWs $S(x,t)$ and $L(x,t)$ based on the improved PINN related to the initial boundary value problem (3.7) and (3.8) of the YO system (1.1). The left panels of Fig. 4 depicts various dynamic density plots and sectional drawing at different moments, in which the panel (a) corresponds to short-wave intermediate RW $S(x,t)$ and panel (c) corresponds to the long-wave bright RW $L(x,t)$ for the YO system (1.1). As we can see from the right panels in Fig. 4, the 3D plots of the intermediate-bright RWs are shown in Fig. (b) and Fig. (d) respectively. Apparently, from Figs. 4, one can see that the maximum amplitudes of short

wave RW and long wave RW are lower than those of the two RWs in Figure Figs. 2. Fig. 5 showcases curve plots of the loss function after 20000 Adam optimization iterations and 50000 L-BFGS optimization iterations in improved PINN framework. Different from the curve plots of the loss function in Figs. 3, the loss function curves of Adam optimization for the intermediate-bright RWs descend steadily. However, in the process of optimizing the loss function using the L-BFGS optimizer, the loss function curve descends faster after about 20000 iterations, which is different from Figure Fig. 3.

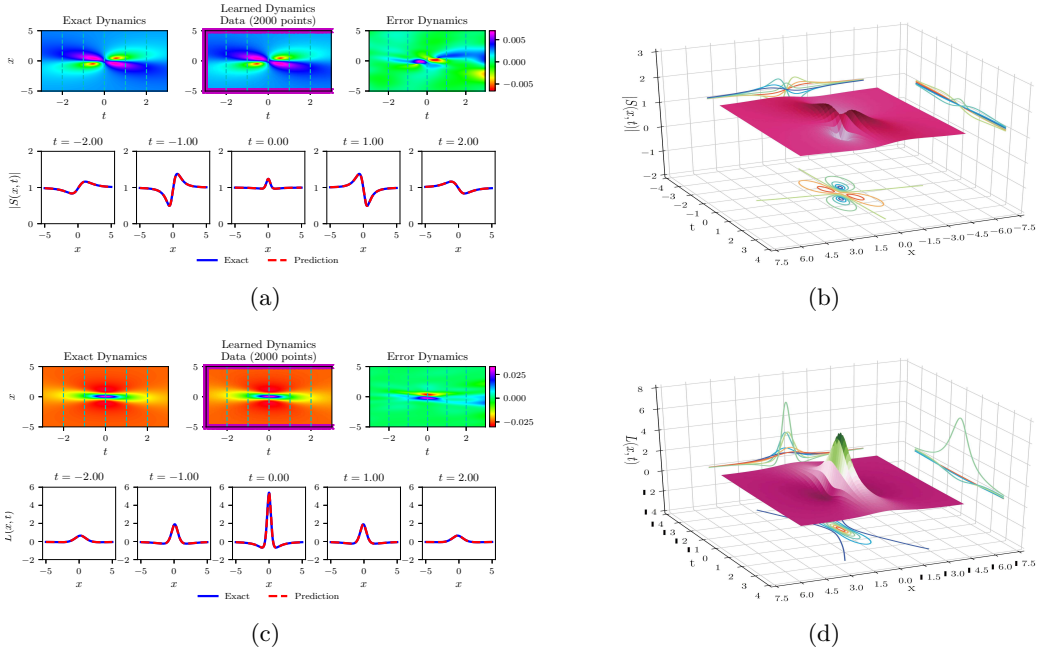
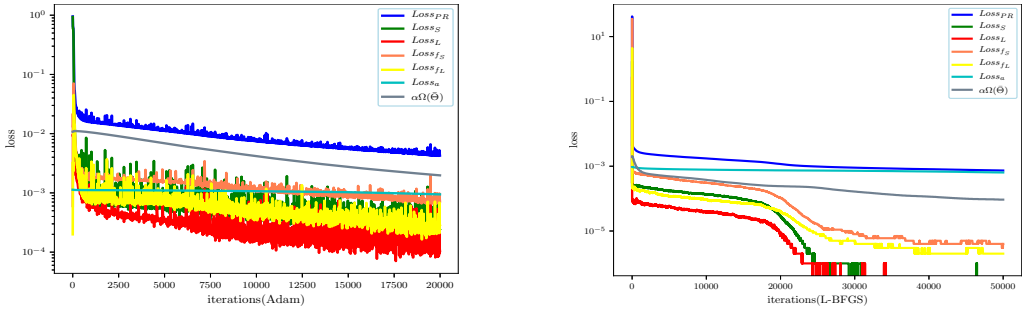


Figure 4: (Color online) The data-driven intermediate-bright RWs $S(x,t)$ and $L(x,t)$ resulted from the improved PINN with the randomly chosen initial and boundary points $N_q = 2000$ which have been shown by using mediumorchid " $\times$ " in learned dynamics , and $N_f = 30000$ collocation points in the corresponding spatiotemporal region: (a) and (c) The exact, learned and error dynamics density plots with five distinct tested times $t = -2.00, -1.00, 0.00, 1.00$ and 2.00 (darkturquoise dashed lines), as well as sectional drawings which contain the learned and explicit intermediate-bright RWs at the aforementioned five distinct times; (b) and (d) The 3D plot for the data-driven intermediate-bright RWs.

3.3 The data-driven dark-bright RWs

Similarly, considering the initial condition and Dirichlet boundary condition of the YO system to obtain the dark-bright RWs by using the 9-layer improved PINN with 40 neurons per layer, the $[X_0, X_1]$ and $[T_0, T_1]$ in Eq. (1.1) are taken as $[-5.0, 5.0]$ and $[-3.0, 3.0]$, respectively. We immediately obtain the initial value conditions

$$S^0(x) = S_{\mathrm{drw}}(x, -3.0),\ L^0(x) = L_{\mathrm{brw3}}(x, -3.0),\ x \in [-5.0, 5.0], \tag{3.9}$$



(a) (b)

Figure 5: (Color online) The loss function curve figures of the intermediate-bright RWs $S(x,t)$ and $L(x,t)$ arising from the improved PINN with the 20000 steps Adam and 50000 steps L-BFGS optimizations: (a) The loss function curve for the 20000 Adam optimization iterations; (b) The loss function curve for the 50000 L-BFGS optimization iterations.

and the Dirichlet boundary conditions

$$\begin{aligned} &S^{\rm lb}(t) = S_{\rm drw}(-5.0,t),\ L^{\rm lb}(t) = L_{\rm brw3}(-5.0,t),\ t\in[-3.0,3.0],\\ &S^{\rm ub}(t) = S_{\rm drw}(5.0,t),\ L^{\rm ub}(t) = L_{\rm brw3}(5.0,t),\ t\in[-3.0,3.0]. \end{aligned} \tag{3.10}$$

Similarly, discretizing exact dark-bright RWs $S_{\rm drw}$ and $L_{\rm brw3}$ with the aid of the traditional finite difference scheme on even grids, and we obtain the original training data which contain initial data (3.9) and boundary data (3.10) by dividing separately the spatial region $[-5.0,5.0]$ into 2000 points and the temporal region $[-3.0,3.0]$ into 1000 points. Then, one can generate a smaller training dataset that contains partial initialCboundary data by randomly extracting $N_q = 2000$ from original dataset and $N_f = 30000$ collocation points which are produced by the LHS. After that, the latent dark-bright RWs $S(x,t)$ and $L(x,t)$ have been successfully learned by tuning all learnable parameters of the improved PINN, and the network achieved relative $\mathbb{L}_2$ error of 1.964839e-03 for the dark RW $S(x,t)$ and relative $\mathbb{L}_2$ error of 1.692152e-02 for the bright RW $L(x,t)$, and the total number of iterations is 70000.

Figs. 6 - 7 provide the training results arising from the improved PINN for the data-driven dark-bright RWs $S(x,t)$ and $L(x,t)$ of the YO system with the initial boundary value problem (3.9) and (3.10). In the left panels of Fig. 6, the exact, learned and error dynamics density plots with corresponding amplitude scale size on the right side have been exhibited, it is worth mentioning that the $N_q = 2000$ training data points involved in the initial-boundary condition are marked by mediumorchid symbol " $\times$ " in the learned density plots both in (a) and (c) of Fig. 6. Meanwhile, the sectional drawings which include the learned and exact dark-bright RWs have been shown at the five distinct times pointed out in the exact, learned and error dynamics density plots by using darkturquoise dashed lines in the bottom panels of (a) and (c). The right panels of Fig. 6 display the three-dimensional plots with contour map on three planes of the predicted dark-bright RWs $S(x,t)$ and $L(x,t)$ based on the improved PINN. Fig. 7 exhibits the loss function curve figures of the dark-bright RWs $S(x,t)$ and $L(x,t)$ arising from the improved PINN with the 20000 steps Adam and 50000 steps L-BFGS optimizations on the loss function

$\widetilde{\mathscr{L}}(\bar{\Theta})$.

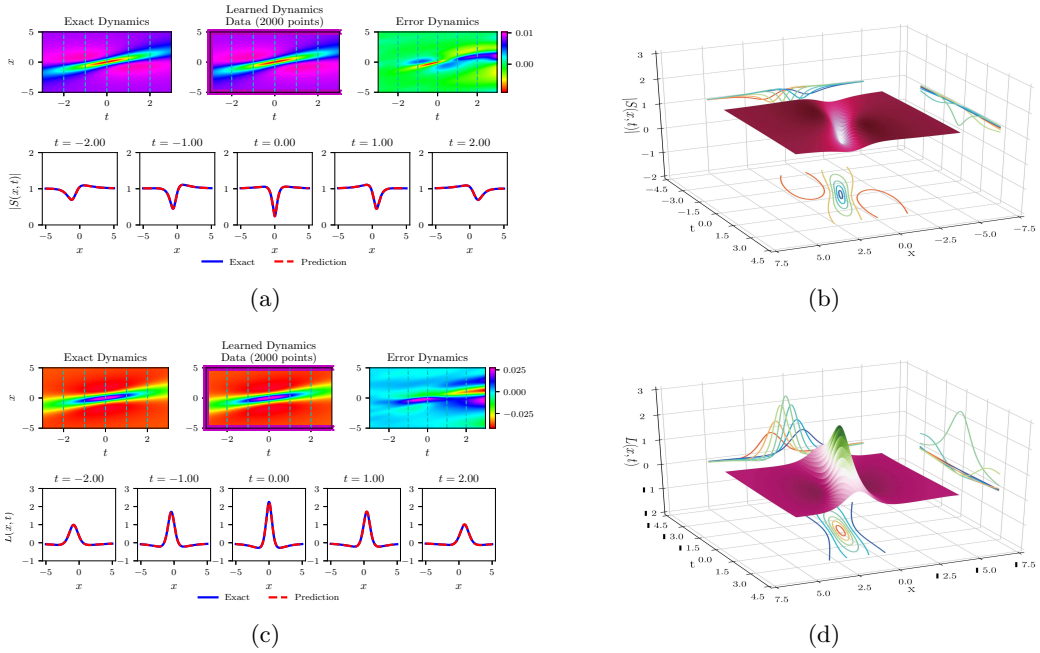
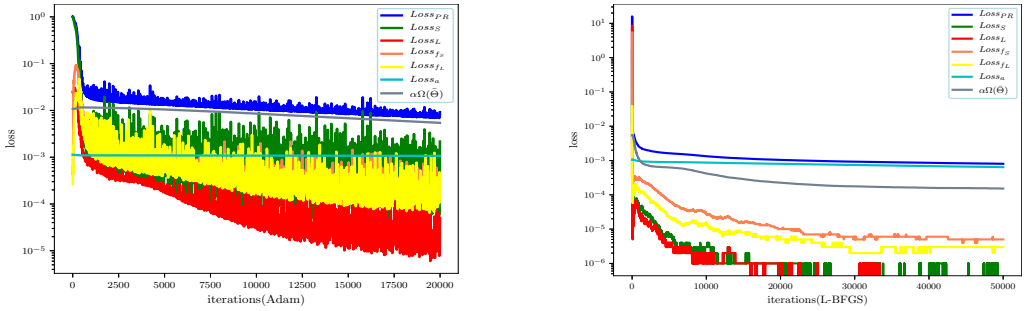


Figure 6: (Color online) The data-driven dark-bright RWs $S(x,t)$ and $L(x,t)$ resulted from the improved PINN with the randomly chosen initial and boundary points $N_q = 2000$ which have been shown by using mediumorchid " $\times$ " in learned dynamics , and $N_f = 30000$ collocation points in the corresponding spatiotemporal region: (a) and (c) The exact, learned and error dynamics density plots with five distinct tested times $t = -2.00, -1.00, 0.00, 1.00$ and 2.00 (darkturquoise dashed lines), as well as sectional drawings which contain the learned and explicit dark-bright RWs at the aforementioned five distinct times; (b) and (d) The 3D plot for the data-driven dark-bright RWs.

In addition, a large number of experimental data show that the training error of improved PINN with parameter regularization ($\alpha = 0.0001$) is smaller than that of PINN without parameter regularization ($\alpha = 0$). Tab. 2 gives the training error of three different types of RWs with and without parameter regularization. From table 1, once the hyper-parameter α is small enough, one can see that the training error of the improved PINN model with parameter regularization is mostly lower than that of the PINN model without parameter regularization.

Table 2: Relative $\mathbb{L}_2$ errors of three different RW types in different PINN types

PINN Types \ **RW Types**	Bright-bright RWs	Intermediate-bright RWs	Dark-bright RWs
Hyper-parameter $\alpha = 0$	$S(x,t)$: 7.566667e-04 $L(x,t)$: 2.414187e-03	$S(x,t)$: 1.975757e-03 $L(x,t)$: 8.500360e-03	$S(x,t)$: 1.788549e-03 $L(x,t)$: 1.286998e-02
Hyper-parameter $\alpha = 0.0001$	$S(x,t)$: 4.968430e-04 $L(x,t)$: 1.763312e-03	$S(x,t)$: 1.168852e-03 $L(x,t)$: 6.766132e-03	$S(x,t)$: 1.964839e-03 $L(x,t)$: 1.692152e-02



(a) (b)

Figure 7: (Color online) The loss function curve figures of the dark-bright RWs $S(x,t)$ and $L(x,t)$ arising from the improved PINN with the 20000 steps Adam and 50000 steps L-BFGS optimizations: (a) The loss function curve for the 20000 Adam optimization iterations; (b) The loss function curve for the 50000 L-BFGS optimization iterations.

4 The inverse problem of the YO system

In this section, we focus on the inverse problem of the YO system, that is parameter discovery problem for a data-driven YO system model (1.1) by utilizing small data set. In this situation, λ_1 and λ_2 are pending parameters to be trained by means of improved PINN with parameter regularization and partial initial boundary value data points, and the physics-informed parts Eq. (2.5) of the improved PINN for YO sysem (1.1) become the following formula

$$f_u := -v_t + \lambda_1 u_{xx} + uL, \quad f_v := u_t + \lambda_1 v_{xx} + vL, \quad f_L := L_t - \lambda_2(2uu_x + 2vv_x). \quad (4.1)$$

In order to learn the parameter λ_1 and λ_2 in Eq. (1.1) with the aid of the improved PINN with neuron-wise locally adaptive activation function and parametric regularization term with different trade-off coefficients, and considering the initial conditions and Dirichlet boundary conditions of Eq. (1.1) arising from the bright-bright RWs (3.4) by using the 9-layer improved PINN with 40 neurons per layer, we set the spatial and temporal regions $(x,t) \in [-5,5]$. After that, the corresponding initial conditions can be written as belows

$$S^0(x) = S_{\rm brw}(x,-0.5),\ L^0(x) = L_{\rm brw1}(x,-0.5),\ x \in [-5.0,5.0], \quad (4.2)$$

and the Dirichlet boundary conditions

$$\begin{aligned} &S^{\rm lb}(t) = S_{\rm brw}(-5.0,t),\ L^{\rm lb}(t) = L_{\rm brw1}(-5.0,t),\ t \in [-0.5,0.5],\\ &S^{\rm ub}(t) = S_{\rm brw}(5.0,t),\ L^{\rm ub}(t) = L_{\rm brw1}(5.0,t),\ t \in [-0.5,0.5]. \end{aligned} \quad (4.3)$$

Here, employing the same data discretization method in section 3, and producing the training data which consists of initial data (4.2) and boundary data (4.3) by dividing the spatial region $[-5.0,5.0]$ into 2000 points and the temporal region $[-0.5,0.5]$ into 1000 points. We generate a smaller training dataset that containing initial-boundary data by randomly extracting $N_q = 2000$ from original dataset and $N_f = 30000$ collocation points

which are generated by the LHS method. After giving the dataset of initial and boundary points, the latent data-driven unknown parameters λ_1 and λ_2 have been successfully learned by tuning all learnable parameters of the improved PINN and utilizing 20000 Adam iterations and different number of L-BFGS iterations to regulate the loss function $\widetilde{\mathscr{L}}(\bar{\Theta})$. The unknown parameters λ_1 and λ_2 are initialized to $\lambda_1 = \lambda_2 = 0$. the relative error of unknown parameters is defined as $RE = (|\hat{\lambda}_\kappa - \lambda_\kappa|/\lambda_\kappa) \times 100\%$ $(\kappa = 1, 2)$ with the predicted value $\hat{\lambda}_\kappa$ and true value λ_κ. All noise interference in this part is added to the randomly chosen small data set, the details are as shown below

$$\begin{aligned} Data\_train =& Data\_train + noise * np.std(Data\_train) * np.random.randn \\ & (Data\_train.shape[0], Data\_train.shape[1]), \end{aligned} \tag{4.4}$$

where $Data\_train$ and $noise$ represent a small randomly chosen training data set and the noise intensity, respectively. The $np.std(\cdot)$ returns the standard deviation of an array element, and $np.random.randn(\cdot, \cdot)$ returns a set of samples with a standard normal distribution.

Next we analyze the training result of the NN from different perspectives, such as the size of hyper-parameters α, intensity of noise and the anti-interference ability. In order to more directly verify the effect of improved PINN with parameter regularization, we first showcase the training effect of PINN without parameter regularization (namely $\alpha = 0$) in Fig. 8. In the absence of a parametric regularization strategy, (a) and (b) of Fig. 8 describe the numerical variation curves of unknown parameters λ_1 and λ during iteration, one can find that λ_1 increases from 0 to more than 0.3 during the previous 20000 Adam optimizations, while λ_2 hardly increases significantly in the aforementioned iterations. Instead, λ_1 increases slowly to about 0.5 in the later L-BFGS optimization process, while λ_2 increases sharply to about 1.0 in this iterative process. The panel (c) of Fig. 8 indicates that the greater the noise intensity, the more intense the fluctuation of the loss function curve in the first 20000 Adam optimization processes, and the larger the overall value of the loss function in the later L-BFGS optimization processes. Fig. 8 (d) shows that the relative error of λ_2 is more sensitive to the change of noise intensity than that of λ_1.

Next, we impose a parametric regularization strategy to the improved PINN with a penalty coefficient $\alpha = 0.0001$, the corresponding training results are shown in Fig. 9. Due to the introduction of parameter regularization, the load of loss function will increase, thus the number of iterations will be greater than that in the PINN without parameter regularization strategy in some cases, but the maximum number of iterations is artificially set to 70000 (including 2000 Adam optimizer iterations and 50000 L-BFGS optimizer iterations). Fig. 9 (a)-(b) show the variation curves of unknown parameters λ_1 and λ_2 in the iterative process under different noise intensity conditions, in which both figures indicate that there is little difference between the final learning results of parameters λ_1 and λ_2 to be learned in the case of noise and no noise. Interestingly, panel (b) of Fig. 9 demonstrates that the value of λ_2 learned with 2% noise intensity is closer to the real value than that learned without noise case. Different from Fig. 8 (c), the Fig. 9 (c) shows that the relationship between the fluctuation degree of the loss function curve and the noise intensity is not obvious. Fig. 9 (d) exhibits that in improved PINN with parameter regularization strategy, the relative error of parameter training results to be learned is the smallest when the noise intensity is 2%, which further reveals that improved PINN with

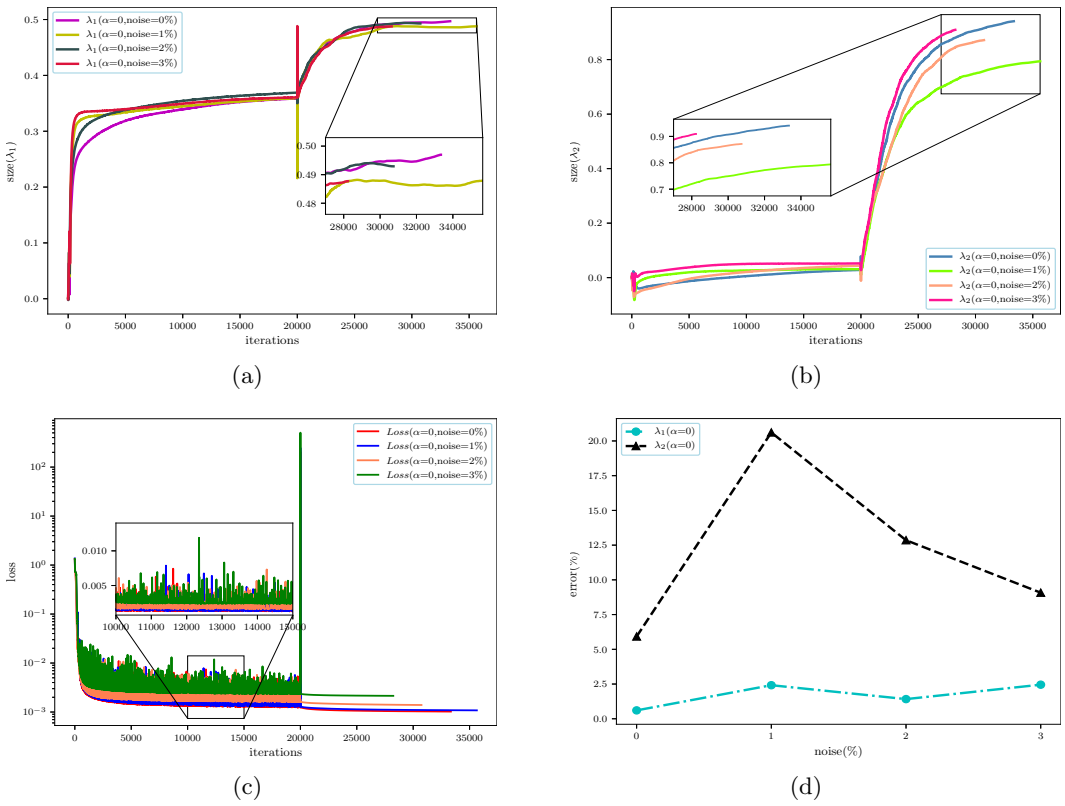


Figure 8: (Color online) The parameter discover resulted from the PINN without parameter regularization ($\alpha = 0$): (a)-(b) the variation of unknown coefficients λ_1 and λ_2 with different noise intensity; (c) the variation of loss function with different noise intensity; (d) unknown coefficient λ_1 and λ_2 error variation under different interference noise.

parameter regularization has the ability to suppress data noise interference.

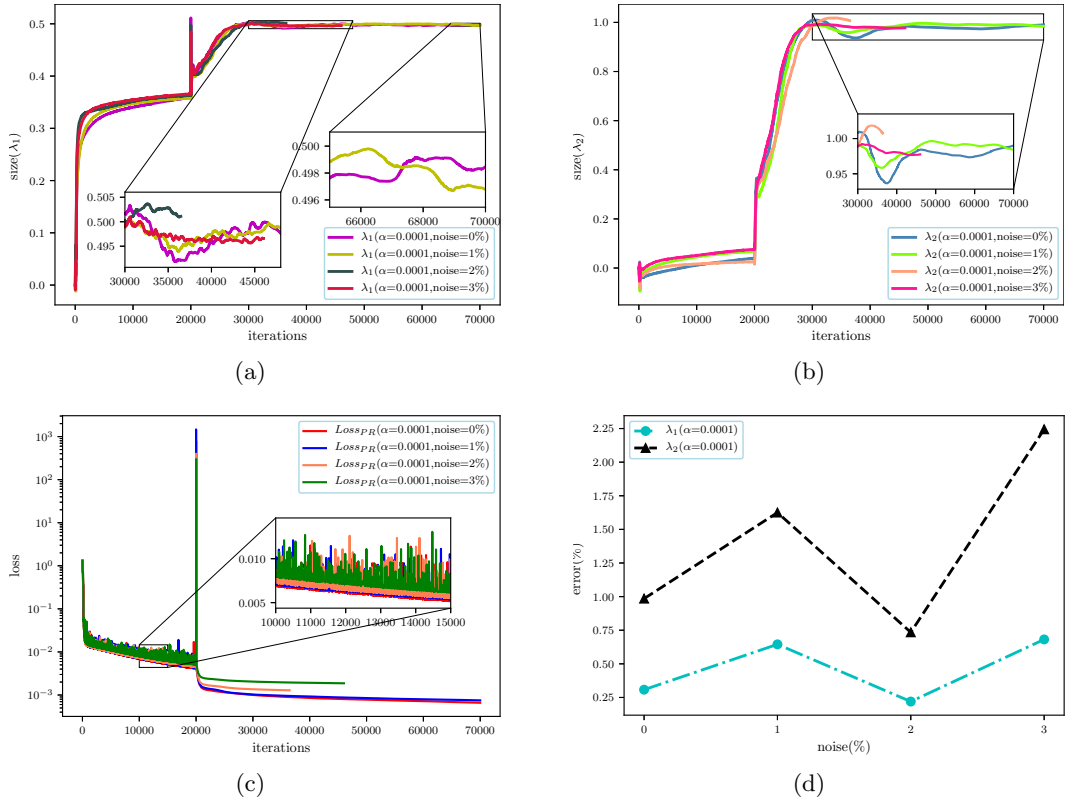


Figure 9: (Color online) The parameter discover resulted from the improved PINN with parameter regularization ($\alpha = 0.0001$): (a)-(b) the variation of unknown coefficients λ_1 and λ_2 with different noise intensity; (c) the variation of loss function with different noise intensity; (d) unknown coefficient λ_1 and λ_2 error variation under different interference noise.

Furthermore, we expand the weight coefficient of parameter regularization by one order of magnitude, namely $\alpha = 0.001$. Fig. 10 displays the dynamic behavior similar to that the case of $\alpha = 0.0001$ in Fig. 9, except that when the noise intensity is 1%, the relative training error of parameters to be learned is smaller than that in the other three cases. In order to further understand the influence of parameter regularization with larger weight ratio on improved PINN, we again expand the weight coefficient by one order of magnitude, that is, $\alpha = 0.01$, and obtain the corresponding inverse problem training results of YO system in Fig. 11. Fig. 11 (a)-(b) showcase the parameter discovery curves of λ_1 and λ_2, where panel (b) indicates that there is a large gap between the training value and the actual value of λ_2 when the noise intensities are 1% and 2%, and combined with panel (d) of Fig. 11, one can observe that the relative error of λ_2 is very large at this time. This also means that the larger the trade-off coefficient is not always better. As shown in Fig. 11, once it is expanded to a certain extent, the training effect is not ideal. Therefore, the selection of the value for the trade-off coefficient is very important for the parameter regularization strategy.

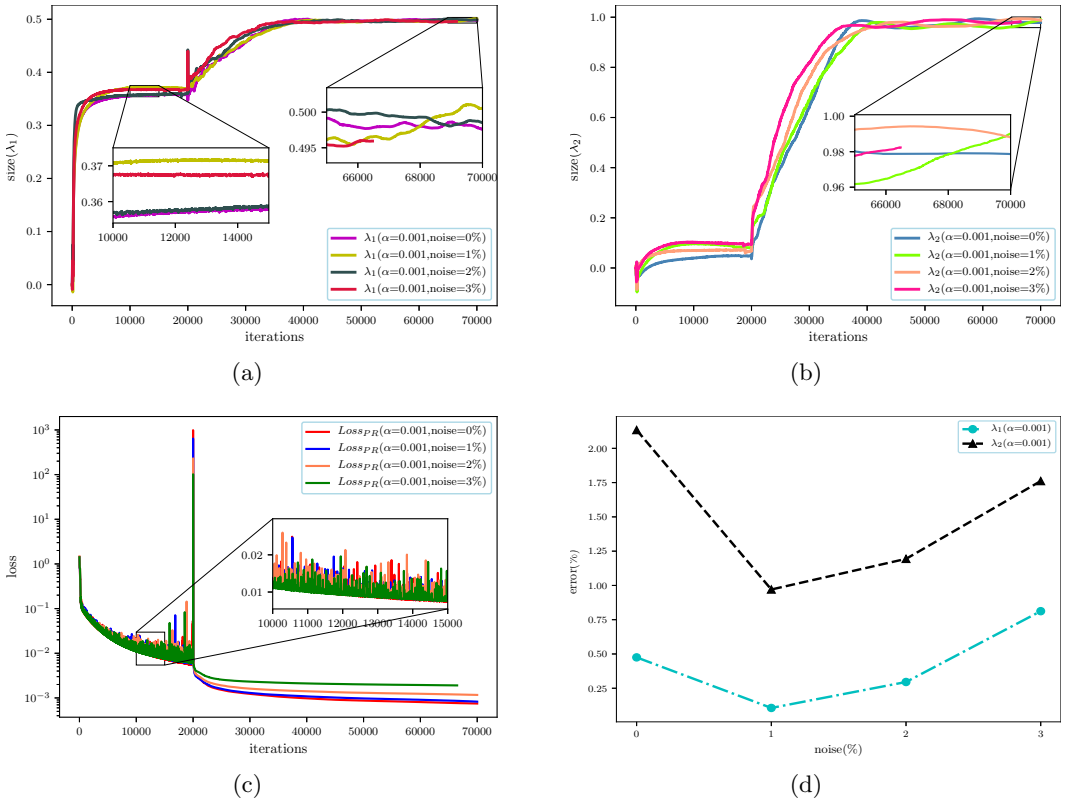


Figure 10: (Color online) The parameter discover resulted from the improved PINN with parameter regularization ($\alpha = 0.001$): (a)-(b) the variation of unknown coefficients λ_1 and λ_2 with different noise intensity; (c) the variation of loss function with different noise intensity; (d) unknown coefficient λ_1 and λ_2 error variation under different interference noise.

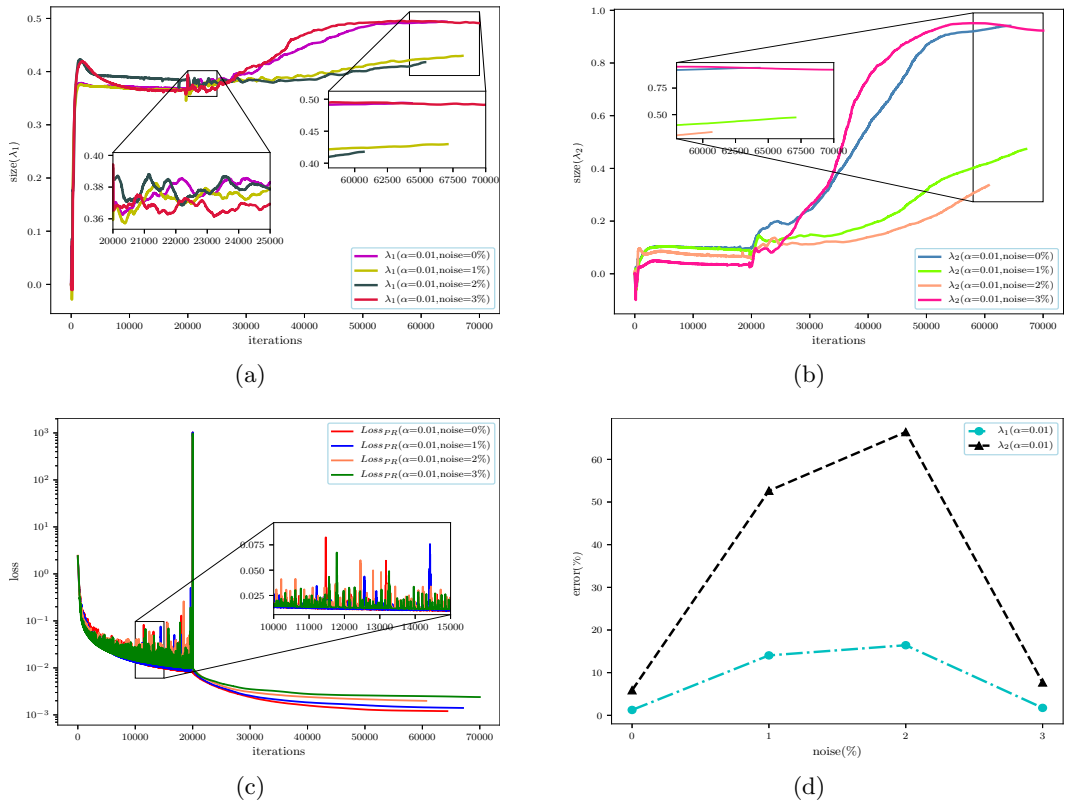


Figure 11: (Color online) The parameter discover resulted from the improved PINN with parameter regularization ($\alpha = 0.01$): (a)-(b) the variation of unknown coefficients λ_1 and λ_2 with different noise intensity; (c) the variation of loss function with different noise intensity; (d) unknown coefficient λ_1 and λ_2 error variation under different interference noise.

So to summarise, this section mainly describes the results and corresponding analysis when the parameter regularization method is not used and the parameter regularization technology with different weight ratio is used both in improved PINN model to study the inverse problem of YO system. One can also find that when using the parameter regularization strategy with appropriate weight coefficients α (generally, α should not be too large), the training effect of parameter discovery is much better than that without parameter regularization strategy, especially after adding the influence of noise with standard normal distribution. This shows that improved PINN with L^2 norm parameter regularization can not only prevent over fitting, but also have effective anti noise ability. Finally, we provide a summary of all the aforementioned training results in following Tab. 3.

Table 3: Comparison of correct YO system and identified YO system obtained by means of the PINN with different noise intensities and weight hyper-parameters α.

hyper-parameters / YO system	$\alpha = 0$	$\alpha = 0.0001$	$\alpha = 0.001$	$\alpha = 0.01$
Correct YO system	$\mathrm{i}S_t + 0.5S_{xx} + SL = 0$ $\mathrm{i}L_t - (\lvert S\rvert^2)_x = 0$ λ_1 error: 0% λ_2 error: 0%	$\mathrm{i}S_t + 0.5S_{xx} + SL = 0$ $\mathrm{i}L_t - (\lvert S\rvert^2)_x = 0$ λ_1 error: 0% λ_2 error: 0%	$\mathrm{i}S_t + 0.5S_{xx} + SL = 0$ $\mathrm{i}L_t - (\lvert S\rvert^2)_x = 0$ λ_1 error: 0% λ_2 error: 0%	$\mathrm{i}S_t + 0.5S_{xx} + SL = 0$ $\mathrm{i}L_t - (\lvert S\rvert^2)_x = 0$ λ_1 error: 0% λ_2 error: 0%
Identified YO system (clean data)	$\mathrm{i}S_t + 0.496981S_{xx} + SL = 0$ $\mathrm{i}L_t - 0.940780(\lvert S\rvert^2)_x = 0$ λ_1 error: 0.603867% λ_2 error: 5.922013%	$\mathrm{i}S_t + 0.498461S_{xx} + SL = 0$ $\mathrm{i}L_t - 0.990150(\lvert S\rvert^2)_x = 0$ λ_1 error: 0.307775% λ_2 error: 0.984997%	$\mathrm{i}S_t + 0.497624S_{xx} + SL = 0$ $\mathrm{i}L_t - 0.978666(\lvert S\rvert^2)_x = 0$ λ_1 error: 0.475115% λ_2 error: 2.133435%	$\mathrm{i}S_t + 0.493709S_{xx} + SL = 0$ $\mathrm{i}L_t - 0.941199(\lvert S\rvert^2)_x = 0$ λ_1 error: 1.258254% λ_2 error: 5.880136%
Identified YO system (1% noise)	$\mathrm{i}S_t + 0.487918S_{xx} + SL = 0$ $\mathrm{i}L_t - 0.793893(\lvert S\rvert^2)_x = 0$ λ_1 error: 2.416390% λ_2 error: 20.610661%	$\mathrm{i}S_t + 0.496775S_{xx} + SL = 0$ $\mathrm{i}L_t - 0.983750(\lvert S\rvert^2)_x = 0$ λ_1 error: 0.645000% λ_2 error: 1.625025%	$\mathrm{i}S_t + 0.500531S_{xx} + SL = 0$ $\mathrm{i}L_t - 0.990297(\lvert S\rvert^2)_x = 0$ λ_1 error: 0.106263% λ_2 error: 0.970280%	$\mathrm{i}S_t + 0.429766S_{xx} + SL = 0$ $\mathrm{i}L_t - 0.473574(\lvert S\rvert^2)_x = 0$ λ_1 error: 14.046884% λ_2 error: 52.642570%
Identified YO system (2% noise)	$\mathrm{i}S_t + 0.492921S_{xx} + SL = 0$ $\mathrm{i}L_t - 0.871463(\lvert S\rvert^2)_x = 0$ λ_1 error: 1.415741% λ_2 error: 12.853670%	$\mathrm{i}S_t + 0.501098S_{xx} + SL = 0$ $\mathrm{i}L_t - 1.007344(\lvert S\rvert^2)_x = 0$ λ_1 error: 0.219584% λ_2 error: 0.734389%	$\mathrm{i}S_t + 0.498522S_{xx} + SL = 0$ $\mathrm{i}L_t - 0.988062(\lvert S\rvert^2)_x = 0$ λ_1 error: 0.295609% λ_2 error: 1.193810%	$\mathrm{i}S_t + 0.417833S_{xx} + SL = 0$ $\mathrm{i}L_t - 0.335997(\lvert S\rvert^2)_x = 0$ λ_1 error: 16.433418% λ_2 error: 66.400314%
Identified YO system (3% noise)	$\mathrm{i}S_t + 0.487733S_{xx} + SL = 0$ $\mathrm{i}L_t - 0.909239(\lvert S\rvert^2)_x = 0$ λ_1 error: 2.453375% λ_2 error: 9.076095%	$\mathrm{i}S_t + 0.496592S_{xx} + SL = 0$ $\mathrm{i}L_t - 0.977557(\lvert S\rvert^2)_x = 0$ λ_1 error: 0.681674% λ_2 error: 2.244323%	$\mathrm{i}S_t + 0.495939S_{xx} + SL = 0$ $\mathrm{i}L_t - 0.982388(\lvert S\rvert^2)_x = 0$ λ_1 error: 0.812221% λ_2 error: 1.761216%	$\mathrm{i}S_t + 0.491294S_{xx} + SL = 0$ $\mathrm{i}L_t - 0.922920(\lvert S\rvert^2)_x = 0$ λ_1 error: 1.741284% λ_2 error: 7.707989%

5 General expression analysis of Modulational instability

Modulation instability (MI) is usually used to describe the characteristics of unstable plane waves, and conversely the one of a stable plane wave is called modulation stability (MS). The earliest MI study was the pioneering work of Benjamin and Feir in fluid dynamics in the early 1960s [75]. It is well known that the generation of rogue wave is closely related to modulation instability (MI). In optical communication systems, the interaction between dispersion and nonlinear effects will lead to MI, which is a common and very important physical phenomenon. Therefore, the research of MI is helpful to improve the performance of optical communication systems. Today, MI has played an important role in many scientific research fields, such as the fluid dynamics [76], nonlinear optics and plasma physics [77]. In fact, the MI of background wave of YO system has been studied in Ref. [78, 79]. In this part, we will systematically present the general MI analysis of YO system (1.1) with unknown parameters λ_1 and λ_2. The YO system (1.1) have following accurate plane wave solutions

$$\begin{aligned} &S = a\mathrm{e}^{\mathrm{i}(kx-\Lambda t+\delta)}, \\ &L = b, \end{aligned} \tag{5.1}$$

here a, b, k, Λ and δ are real constants. Once the spatial period of the plane wave is fixed to T_{pw}, then the wavenumber is $k = \frac{2\pi}{T_{pw}} n$ for $n \in \mathbb{Z}$. Substituting Eq. 5.1 into Eq. 1.1, k and Λ satisfy the following dispersion relation

$$\Lambda - \lambda_1 k^2 + b = 0. \tag{5.2}$$

Eq. (1.1) is linearized about the plane wave solution by substitution of

$$\begin{aligned} S &= a\mathrm{e}^{\mathrm{i}[kx-(\lambda_1 k^2 - b)t+\delta]}(1+U), \\ L &= b + V, \end{aligned} \tag{5.3}$$

where U, V are small perturbations whose nonlinear contributions are neglected. $U(x,t)$ is complex valued function, while $V(x,t)$ is real valued function. The linearized equations are

$$\begin{aligned} &\mathrm{i}U_t + \lambda_1 U_{xx} + 2\mathrm{i}\lambda_1 k U_x + V = 0, \\ &V_t - \lambda_2 a^2 (U_x + U_x^*) = 0, \end{aligned} \tag{5.4}$$

where " $*$ " denotes the complex conjugate.

Then by constructing a complete basis for the solutions of the linearized equations (5.4), the stability of the plane wave solution (5.1) is determined directly. Due to V is real, it is convenient to construct the small-amplitude Fourier modes in the following formula [78], that is

$$\begin{aligned} U &= f_+ \mathrm{e}^{\mathrm{i}\sqrt{\mu}(x-\Omega t)} + f_-^* \mathrm{e}^{-\mathrm{i}\sqrt{\mu}(x-\Omega^* t)}, \\ V &= g\mathrm{e}^{\mathrm{i}\sqrt{\mu}(x-\Omega t)} + g^* \mathrm{e}^{-\mathrm{i}\sqrt{\mu}(x-\Omega^* t)}, \end{aligned} \tag{5.5}$$

where $\sqrt{\mu} = \frac{2\pi}{T_{pw}} m$ and $m \in \mathbb{Z}$. Substituting Fourier modes (5.5) into linearized equations (5.4), one can obtain following linear system

$$\begin{pmatrix} -\lambda_1\mu - 2\lambda_1\sqrt{\mu}k + \Omega\sqrt{\mu} & 0 & 1 \\ 0 & -\lambda_1\mu + 2\lambda_1\sqrt{\mu}k - \Omega\sqrt{\mu} & 1 \\ -\mathrm{i}a^2\lambda_2\sqrt{\mu} & -\mathrm{i}a^2\lambda_2\sqrt{\mu} & -\mathrm{i}\Omega\sqrt{\mu} \end{pmatrix} \begin{pmatrix} f_+ \\ f_- \\ g \end{pmatrix} = 0. \tag{5.6}$$

If $\mu \neq 0$, the linearized dispersion relation can be presented by setting the determinant of the above matrix to zero, then

$$\Omega\big[(\Omega - 2\lambda_1 k)^2 - \lambda_1^2 \mu\big] - 2a^2\lambda_1\lambda_2 = 0. \tag{5.7}$$

In general, Eq. (5.7) has three distinct Ω roots, corresponding to three linearly independent complex-valued vectors (f_+, f_-, g) satisfying matrix system (5.6), we have

$$\begin{pmatrix} f_+ \\ f_- \\ g \end{pmatrix} = \begin{pmatrix} \mu\Omega(\Omega + \lambda_1\sqrt{\mu} - 2\lambda_1 k) + a^2\sqrt{\mu}\lambda_2) \\ -\lambda_2\sqrt{\mu}a^2 \\ -\lambda_2\mu a^2(\Omega + \lambda_1\sqrt{\mu} - 2\lambda_1 k) \end{pmatrix}. \tag{5.8}$$

The collection of all such Fourier modes composes a wavebasis for the linearized problem, there is a three-complex-dimensional subspace of eigenmodes (5.5) associated with

each wavenumber $\sqrt{\mu} \neq 0$ by means of the complex-valued vectors (f_+, f_-, g) and the linearized dispersion relation (5.7) for Ω. In order to complete the wavebasis of the linearized problem, the three-real-dimensional subspace associated with $\sqrt{\mu} = 0$ must also be constructed, i.e.,

$$\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathrm{i} \\ 0 \end{pmatrix}, \begin{pmatrix} \mathrm{i}t \\ 1 \end{pmatrix}. \tag{5.9}$$

Unstable Fourier modes occur if and only if Ω is nonreal. On account of the coefficients of the linearized dispersion relation are real, nonreal Ω occur in complex conjugate pairs, with at most one complex conjugate pair possible for any given value μ. Hence, each unstable wavenumber pair $\pm\sqrt{\mu}$ has a complex one-dimensional unstable manifold.

We know that the MI arises from a nonreal value of Ω and can be defined via the growth rate $G = \sqrt{\mu}\mathrm{Im}(\Omega) > 0$. Thus, to initiate it, the discriminant of Eq. (5.7),

$$\Delta = 16\,{\lambda_1}^4 k^4 \mu + 16\,{\lambda_1}^2 k^3 a^2 \lambda_2 - 8\,{\lambda_1}^4 k^2 \mu^2 - 36\,{\lambda_1}^2 \mu\, k\, a^2 \lambda_2 + {\lambda_1}^4 \mu^3 - 27\, a^4 {\lambda_2}^2. \tag{5.10}$$

Thus, for given λ_1, λ_2 and a, the marginal stability curves occur at $\Delta = 0$, the MI region demand $\Delta < 0$, and the MS region require $\Delta > 0$, correspondingly.

After setting fixed λ_1 and λ_2, an asymptotic analysis of the roots μ as $k \to \pm\infty$ can be obtained, as shown bellow:

- $\lambda_1 = 0.5$ and $\lambda_2 = 1$

The marginal stability curves occur when the discriminant (5.10) satisfy

$$\Delta = -27\, a^4 + 4\, a^2 \omega^3 - 9\, a^2 \mu\, \omega + \omega^4 \mu - \frac{1}{2}\,\omega^2 \mu^2 + \frac{1}{16}\,\mu^3 = 0. \tag{5.11}$$

Obviously, μ has three roots, thus an asymptotic analysis of the roots as $k \to \pm\infty$ shows that the positive roots μ are given by

$$\begin{aligned} &\mu \sim 4k^2 \pm 8\sqrt{2} a \sqrt{k}, \qquad && k \to \infty, \\ &\mu \sim \frac{-a^2}{k}\left(1 + \frac{a^2}{k^3}\right), \qquad && k \to -\infty. \end{aligned} \tag{5.12}$$

In other words, independently of the value of a^2, Eq. (5.11) has two positive roots and one negative root as $k \to \infty$ and one positive root and two nonreal roots as $k \to -\infty$. Also $\mu = 0$ occurs at precisely one value of $k = k_c > 0$, where $k_c^3 = \frac{27}{4}a^2$ (if $a = 1$, thus $k_c = \frac{3}{2}2^{1/3}$, $k_n = (2a^2)^{1/3}$ and $k_c/k_n = \frac{3}{2}$). At $(k, \mu) = (0, 0)$, the linearized dispersion relation (5.7) has one real root $(a^{2/3})$ and two nonreal roots $(-1/2\, a^{2/3} + 1/2\,\mathrm{i}\sqrt{3}a^{2/3}, -1/2\, a^{2/3} - 1/2\,\mathrm{i}\sqrt{3}a^{2/3})$ for Ω, that is, there is a one-dimensional unstable manifold for each $\sqrt{\mu}$ sufficiently close to zero. In general, the distinguishing features of the MI in the $(k, \sqrt{\mu})$ plane (a deformation of the plane (k, μ)) can be summarized as follows:

1. There is an interval $k \in (-\infty, k_c)$ for wavenumbers modulationally unstable to long wavelength perturbations, and this interval always contains the spatially independent plane waves $k = 0$.
2. The MI band becomes very narrow as $k \to -\infty$ according to the scaling $\sqrt{\mu} \sim \frac{a}{\sqrt{-k}}$.

3. The narrow bands of unstable intermediate wavelengths asymptotically approach the line $\sqrt{\mu} = 2k$ as $k \to \infty$.
4. There is an interval of wavenumbers $k \in (k_c, \infty)$ that are unstable to intermediate wavelength perturbations but stable to long wavelength perturbations.

- $\lambda_1 = -1$ and $\lambda_2 = -4$

The marginal stability curves occur when the discriminant (5.10) satisfy

$$\Delta = -64\,a^2\omega^3 + 16\,\omega^4\mu - 432\,a^4 + 144\,a^2\mu\,\omega - 8\,\omega^2\mu^2 + \mu^3 = 0. \tag{5.13}$$

Obviously, μ has three roots, thus an asymptotic analysis of the roots as $k \to \pm\infty$ shows that the positive roots μ are given by

$$\begin{aligned} &\mu \sim \frac{a^2}{k}\left(4 - \frac{a^2}{k^3}\right), \qquad k \to \infty, \\ &\mu \sim 4k^2 \pm 8\sqrt{2}a\sqrt{-k}, \ k \to -\infty. \end{aligned} \tag{5.14}$$

In other words, independently of the value of a^2, Eq. (5.13) has one positive root and two nonreal roots as $k \to \infty$ and two positive roots and one negative root as $k \to -\infty$. Also $\mu = 0$ occurs at precisely one value of $k = k_c < 0$, where $k_c^3 = -\frac{27}{4}a^2$ (if $a = 1$, thus $k_c = -\frac{3}{2}2^{1/3}$). At $(k, \mu) = (0, 0)$, the linearized dispersion relation (5.7) has one real root and two nonreal roots (the form of the three roots is complex and omitted here) for Ω, that is, there is a one-dimensional unstable manifold for each $\sqrt{\mu}$ sufficiently close to zero. In general, the distinguishing features of the MI in the $(k, \sqrt{\mu})$ plane (a deformation of the plane (k, μ)) can be summarized as follows:
1. There is an interval $k \in (k_c, \infty)$ for wavenumbers modulationally unstable to long wavelength perturbations, and this interval always contains the spatially independent plane waves $k = 0$.
2. The MI band becomes very narrow as $k \to \infty$ according to the scaling $\sqrt{\mu} \sim \frac{2a}{\sqrt{k}}$.
3. The narrow bands of unstable intermediate wavelengths asymptotically approach the line $\sqrt{\mu} = -2k$ as $k \to -\infty$.
4. There is an interval of wavenumbers $k \in (-\infty, k_c)$ that are unstable to intermediate wavelength perturbations but stable to long wavelength perturbations.

Next, in order to more intuitively understand the distribution of MI and MS regions, we exhibit the linear stability diagram and growth rate density plot of two specific examples for plane-wave solution (5.3) with $a = 1$ by taking two sets of specific parameters λ_1 and λ_2 in Fig.12. Specifically, after taking $\lambda_1 = 0.5$ and $\lambda_2 = 1$, one can obtain the corresponding plot in Fig. 12 (a), the region of MI is surrounded by two marginal stability curves (black dash-dotted curves) and the line $\mu^{1/2} = 0$, the density plot of the growth rate G in the whole $(k/k_n, \mu^{1/2})$ plane (for comparing with the results of Ref. [79], we make a scale transformation on the plane (k, μ)) can be calculated and shown within the MI region with the $\Delta < 0$. In the section on the forward problem of YO system, there different RWs are showcased in different k intervals. Therefore, we also display the areas where these three different forms of RWs exist, which are separated by pink dotted lines in Fig. 12 (a). It is worth mentioning that these results shown by the aforementioned special case is consistent with the research results of Chen et al [79].

Similarly, if taking $\lambda_1 = -1$ and $\lambda_2 = -4$, we obtain the corresponding region diagram of MI, MS and three different forms of RWs in the whole (k, μ) plane, as well as the density

plot of growth rate G are exhibit in Fig. 12 (b). According to the marginal stability curve on the left when $\Delta = 0$, one can easily calculate $k = -\frac{3}{2}2^{1/3}$ at the intersection of the aforementioned curve and the horizontal ordinate. From the MI region in Fig. 12 (b), one can obtain that RWs exist in the region of $k > -\frac{3}{2}2^{1/3}$, this result is consistent with the conclusion obtained by Chen et al. via the analysis of RWs with the aid of KP reduction method and related characteristic points [72]. In order to reveal the existence regions for three different forms of RWs in Fig. 12 (b), we showcase the existence ranges for three kinds of RWs in the MI region of Fig. 12 (b) by means of purple dotted line and the relevant conclusions of Ref. [72], in which bright-bright RWs: $-\frac{3}{2}2^{1/3} < k \leqslant -\frac{2}{3}3^{2/3}$, intermediate-bright RWs: $-\frac{2}{3}3^{2/3} < k < 0$, dark-bright RWs: $k \geqslant 0$.

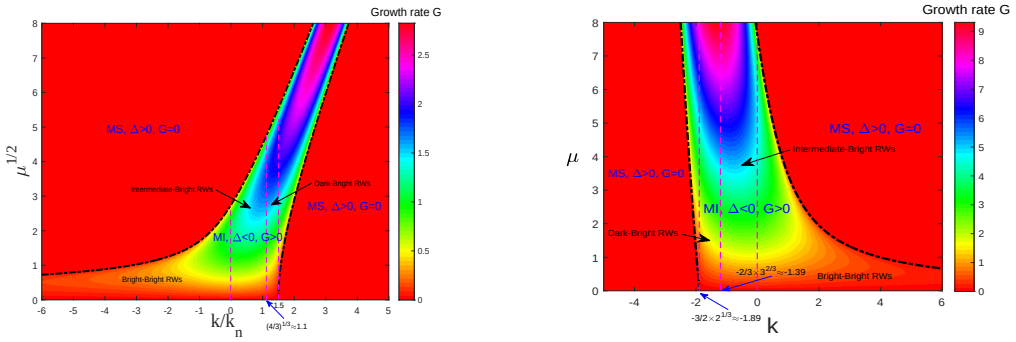


Figure 12: (Color online) MI, MS and RW existence region diagram for the plane-wave solutions with $a = 1$, in which the black dash-dotted curves stand for the marginal stability defined by $\Delta = 0$, and the contour plot shows the growth rate G of the MI: (a) $\lambda_1 = 0.5$ and $\lambda_2 = 1$; (b) $\lambda_1 = -1$ and $\lambda_2 = -4$.

6 Conclusions and discussions

In this paper, the data-driven forward-inverse problems and MI analysis of YO system are showcased by means of deep learning method based on the improved PINN with parameter regularization strategy, as well as linearized instability analysis technique on the plane waves. From many of our previous work [34,59,70], we realize that by introducing scalable hyper-parameters into the activation function, the improved PINN employed to simulate localized waves of nonlinear integrable systems not only improves the convergence of the network, but also obtains better accuracy and network performance. Therefore, for the data-driven forward problems of YO system, we find that improved PINN can well reveal three different forms of RWs, including bright RW, intermediate RW and dark RW from the perspective of short wave. Although the relative $\mathbb{L}_2$ norm errors of RWs of YO system simulated arising from improved PINN model are smaller after introducing L^2 norm parameter regularization technique into PINN, the effect is not so obvious, as shown in Tab. 2. However, in the further study of the inverse problem of YO system, we find that the data-driven unknown parameters trained by improved PINN with L^2 norm parameter regularization are more accurate than those trained by means of the PINN

without parameter regularization. Compared with the general PINN model, the improved PINN with parameter regularization shows excellent noise immunity in the inverse problem of YO system in section 4. Furthermore, Asymptotic analysis of wavenumber k and the MI analysis of YO system (1.1) with unknown parameters λ_1 and λ_2 are derived systematically by applying the linearized instability analysis on plane wave, and the density plots of MI, MS and RW regions are vividly displayed in Fig. 12. Moreover, due to the exponential growth characteristics of MI, it will still develop into a large value after a certain distance and eventually interfere with the spatiotemporal localized RWs.

The introduction of L^2 norm parameter regularization strategy in deep learning is a very mature and commonly used means. Naturally, we successfully applied this technology to improved PINN algorithm and found that it has a positive effect in the positive problem of YO system, and achieves an amazing effect in the parameter discovery process of inverse problem in this paper. In addition to L^2 norm parameter regularization, there are other parameter regularization strategies, such as L^1 norm parameter regularization. In comparison to L^2 regularization, L^1 regularization results in a solution that is more sparse, here sparsity in this context refers to the fact that some parameters have an optimal value of zero. The L^2 regularization does not cause the parameters to become sparse, while L^1 regularization may do so for large enough α. However, when we introduce L^1 norm parameter regularization strategy into improved PINN model, we find that the simulation effect is not ideal, so how to better combine L^1 parameter regularization technology with improved PINN to deal with the various data-driven problems of nonlinear integrable systems needs further research. Of course, there are other parameter regularization methods, but how to properly introduce these regularization methods into deep learning methods to solve the problem at hand is an eternal topic.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors gratefully acknowledge the support of the National Natural Science Foundation of China (No. 12175069) and Science and Technology Commission of Shanghai Municipality (No. 21JC1402500 and No. 18dz2271000).

References

[1] G.E. Hinton, S. Osindero, Y. Teh, A fast learning algorithm for deep belief nets, Neural Computation, 18 (2006) 1527-1554.

[2] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436-444.

[3] A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, In Advances in Neural Information Processing Systems 25 (NIPS2012) (2012).

[4] X.Y. Cao, J. Yao, Z.B. Xu, D.Y. Meng, Hyperspectral Image Classification With Convolutional Neural Network and Active Learning, IEEE Trans. Geosci. Remote Sens. 58 (2020) 4604-4616.

[5] A.B. Nassif, I. Shahin, I. Attili, M. Azzeh, K. Shaalan, Speech recognition using deep neural networks: a systematic review, IEEE Access, 7 (2019) 19143-19165.

[6] G.E. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, B. Kingsbury, Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process. Mag. 29(6) (2012) 82-97.

[7] G.F. Li, Y.F. Yang, X.D. Qu, Deep learning approaches on pedestrian detection in hazy weather, IEEE Trans. Ind. Electron. 67 (2020) 8889-8899.

[8] N.Y. Zeng, H. Li, Z.D. Wang, W.B. Liu, S.M. Liu, F.E. Alsaadi, X.H. Liu, Deep-reinforcement-learning-based images segmentation for quantitative analysis of gold immunochromatographic strip, Neurocomputing 425 (2021) 173-180.

[9] D. Ciresan, U. Meier, J. Masci, J. Schmidhuber, Multi-column deep neural network for traffic sign classification. Neural Networks 32 (2012) 333-338.

[10] J.M. Zhang, W. Wang, C.Q. Lu, J. Wang, A.K. Sangaiah, Lightweight deep network for traffic sign classification, Ann. Telecommun. 75 (2020) 369-379.

[11] R. Collobert, J. Weston, G. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, J. Mach. Learn. Res. 12 (2011) 2493-2537.

[12] B. Alipanahi, A. Delong, M.T. Weirauch, B.J. Frey, Predicting the sequence specificities of DMA- and RNA-binding proteins by deep learning, Nat. Biotechnol. 33 (2015) 831-838.

[13] X.D. Sun, P.C. Wu, S.C.H. Hoi, Face detection using deep learning: an improved faster RCNN approach, Neurocomputing 299 (2018) 42-50.

[14] Z.F. Shao, L.G. Wang, Z.Y. Wang, W. Du, W.J. Wu, Saliency-aware convolution neural network for ship detection in surveillance video, IEEE Trans. Circuits Syst. Video Technol. 30 (2020) 781-794.

[15] M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, J. Comput. Phys. 378 (2019) 686-707.

[16] S. Haykin, Neural Networks and Learning Machines, Prentice-Hall, New York, 2008.

[17] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Networks 2 (1989) 359-366.

[18] D.E. Rumelhart, G.E. Hinton, G.E. Williams, Learning representations by back-propagating errors. Nature, 323 (1986) 533-536.

[19] S. Ruder, An overview of gradient descent optimization algorithms, arXiv:1609 .04747v2, 2017.

[20] D.P. Kingma, J.L. Ba, Adam: a method for stochastic optimization, arXiv: 1412.6980, 2014.

[21] D.C. Liu, J. Nocedal, On the limited memory BFGS method for large scale optimization, Math. Program. 45 (1989) 503-528.

[22] A.G. Baydin, B.A. Pearlmutter, A.A. Radul, J.M. Siskind, Automatic differentiation in machine learning: a survey, Journal of Machine Learning Research 18 (2018) 1-43.

[23] I.E. Lagaris, A. Likas, D.I. Fotiadis, Artificialneural network for solving ordinary and partial differential equations, IEEE Trans. Neural Netw. 9(5) (1998) 987-1000.

[24] M. Raissi, G.E. Karniadakis, Hidden physics models: machine learning of nonlinear partial differential equations, J. Comput. Phys. 357 (2018) 125-141.

[25] M. Raissi, H. Babaee, P. Givi, Deep learning of turbulent scalar mixing, Phys. Rev. Fluids 4(12) (2019) 124501.

[26] G. Pang, L. Lu, G.E. Karniadakis, FPINNs: Fractional physics-informed neural networks, SIAM J. Sci. Comput. 41(4) (2019) A2603-A2626.

[27] Z. Mao, A.D. Jagtap, G.E. Karniadakis, Comput. Methods Appl. Mech. Engrg. 360 (2020) 112789.

[28] L. Yang, D. Zhang, G.E. Karniadakis, Physics-informed generative adversarial networks for stochastic differential equations, SIAM J. Sci. Comput. 42(1) (2020) A292-A317.

[29] J.C. Pu, J. Li, Y. Chen, Soliton, breather and rogue wave solutions for solving the nonlinear Schrödinger equation using a deep learning method with physical constraints, Chin. Phys. B 30 (2021) 060202.

[30] Nair, V. and Hinton, G. (2010). Rectified linear units improve restricted Boltzmann machines. In Proceedings of the 27^{th} International Conference on Machine Learning, Haifa, Israel, 2010. (In ICML'2010.)

[31] I.J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, Y. Bengio, Maxout networks, Proceedings of the 30 th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. (ICML'13 , pages 1319-1327 (2013)).

[32] A.D. Jagtap, K. Kawaguchi, G.E. Karniadakis, Adaptive activation functions accelerate convergence in deep and physics-informed neural networks, J. Comput. Phys. 404 (2020) 109136.

[33] A.D. Jagtap, K. Kawaguchi, G.E. Karniadakis, Locally adaptive activation functions with slope recovery for deep and physics-informed neural networks, Proc. R. Soc. A 476 (2020) 20200334.

[34] J.C. Pu, J. Li, Y. Chen, Solving localized wave solutions of the derivative nonlinear Schrödinger equation using an improved PINN method, Nonlinear Dyn. 105 (2021) 1723-1739.

[35] T. Poggio, F. Girosi, Regularization algorithms for learning that are equivalent to multilayer networks, Science, 247(4945) (1990) 978-982.

[36] R. Zhang, X.L. Li, T. Xu, Y. Zhao, Data clustering via uncorrelated ridge regression, IEEE Trans. Neural Netw. Learn. Syst. 32(1) (2021) 450-456.

[37] N.K. Chada, A.M. Stuart, X.T. Tong, Tikhonov regularization within ensemble kalman inversion, SIAM J. Numer. Anal. 58(2) (2020) 1263-1294.

[38] C. Kharif, E. Pelinovsky, Physical mechanisms of the rogue wave phenomenon, Eur. J. of Mech. B/Fluids 22 (2003) 603-634.

[39] E. Pelinovsky, C. Kharif, Extreme Ocean Waves, Springer, Berlin, 2008.

[40] L. Draper, Freak ocean waves, Weather 21 (1966) 2-4.

[41] D.H. Peregrine, Interaction of water waves and currents, Adv. Appl. Mech. 16 (1976) 9-117.

[42] M.L. Grundlingh, Evidence of surface wave enhancement in the southwest Indian ocean from satellite altimetry, J. Geophys. Res. 99 (1994) 7917-7927.

[43] N. Mori, P.C. Liu, T. Yasuda, Analysis of freak wave measurements in the sea of Japan, Ocean Eng. 29 (2002) 41399-41414.

[44] R. Smith, Giant waves, Fluid Mech. 77 (1976) 417-431.

[45] M. Onorato, S. Residori, U. Bortolozzo, A. Montina, F.T. Arecchi, Rogue waves and their generating mechanisms in different physical contexts, Phys. Rep. 528 (2013) 47-89.

[46] J.M. Dudley, G. Genty, A. Mussot, A. Chabchoub, F. Dias, Rogue waves and analogies in optics and oceanography, Nat. Rev. Phys. 11 (2019) 675-689.

[47] A. Chabchoub, N.P. Hoffmann, N. Akhmediev, Rogue wave observation in a water wave tank, Phys. Rev. Lett. 106 (2011) 204502.

[48] D.R. Solli, C. Ropers, P. Koonath, B. Jalali, Optical rogue waves, Nature 450 (2007) 1054-1057.

[49] Y.V. Bludov, V.V. Konotop, N. Akhmediev, Matter rogue waves, Phys. Rev. A 80 (2009) 033610.

[50] E.G. Charalampidis, J. Cuevas-Maraver, D.J. Frantzeskakis, P.G. Kevrekidis, Rogue waves in ultracold Bosonic seas, Rom. Rep. Phys. 70 (2018) 504.

[51] M. Shats, H. Punzmann, H. Xia, Capillary rogue waves, Phys. Rev. Lett. 104 (2010) 104503.

[52] H. Xia, T. Maimbourg, H. Punzmann, M. Shats, Oscillon dynamics and rogue wave generation in faraday surface ripples, Phys. Rev. Lett. 109 (2012) 114502.

[53] L. Stenflo, P.K. Shukla, Nonlinear acoustic-gravity waves, J. Plasma Phys. 75 (2009) 841-847.

[54] R. Höhmann, U. Kuhl, H.J. Stöckmann, L. Kaplan, E.J. Heller, Freak waves in the linear regime: a microwave study, Phys. Rev. Lett. 104 (2010) 093901.

[55] W.M. Moslem, P.K. Shukla, B. Eliasson, Surface plasma rogue waves, Europhys. Lett. 96 (2011) 25002.

[56] H. Bailung, S.K. Sharma, Y. Nakamura, Observation of Peregrine Solitons in a Multicomponent Plasma with Negative Ions, Phys. Rev. Lett. 107 (2011) 255005.

[57] C. Lecaplain, Ph. Grelu, J.M. Soto-Crespo, N. Akhmediev, Dissipative rogue waves generated by chaotic pulse bunching in a mode-locked laser, Phys. Rev. Lett. 108 (2012) 233901.

[58] Z.Y. Yan, Vector financial rogue waves, Phys. Lett. A 375 (2011) 4274-4279.

[59] J.C. Pu, W.Q. Peng, Y. Chen, The data-driven localized wave solutions of the derivative nonlinear Schrödinger equation by using improved PINN approach, Wave Motion 107 (2021) 102823.

[60] D.J. Kaup, B.A. Malomed, R.S. Tasgal, Internal dynamics of a vector soliton in a nonlinear optical fiber, Phys. Rev. E 48 (1993) 3049-3053.

[61] V.G. Ivancevic, Adaptive-wave alternative for the black-scholes option pricing model, Cogn. Comput. 2 (2010) 17-30.

[62] Y.V. Bludov, V.V. Konotop, N. Akhmediev, Vector rogue waves in binary mixtures of Bose-Einstein condensates, Eur. Phys. J. Special Topics 185 (2010) 169-180.

[63] D.J. Benney, A general theory for interactions between short and long waves, Stud. Appl. Math. 56 (1977) 81-94.

[64] V.D. Djordjevic, L.G. Redekopp, On the two-dimensional packets of capillary-gravitywaves, J. Fluid Mech. 79 (1977) 703-714.

[65] V.E. Zakharov, Collapse of Langmuir waves, Sov. Phys. JETP 35 (1972) 908-914.

[66] N. Yajima, M. Oikawa, Formation and interaction of Sonic-Langmuir solitons: inverse scattering method, Prog. Theor. Phys. 56 (1976) 1719-1739.

[67] A. Chowdhury, J.A. Tataronis, Long wave-short wave resonance in nonlinear negative refractive index media, Phys. Rev. Lett. 100 (2008) 153905.

[68] R.H.J. Grimshaw, The modulation of an internal gravity-wave packet, and the resonance with the mean motion, Stud. Appl. Math. 56 (1977) 241-266.

[69] M. Funakoshi, M. Oikawa, The resonant interaction between a long internal gravity wave and a surface gravity wave packet, J. Phys. Soc. Jpn. 52 (1983) 1982-1995.

[70] J.C. Pu, Y. Chen, The data-driven vector localized waves of Manakov system using improved PINN approach, arXiv: 2109.09266, 2021.

[71] K.W. Chow, H.N. Chan, D.J. Kedziora, R.H.J. Grimshaw, Rogue wave modes for the long wave-short wave resonance model, J. Phys. Soc. Jpn. 82 (2013) 074001.

[72] J.C. Chen, Y. Chen, B.F. Feng, K.I. Maruno, Y. Ohta, General high-order rogue waves of the (1+1)-dimensional Yajima-Oikawa System, J. Phys. Soc. Jpn. 87 (2018) 094007.

[73] S.H. Chen, Darboux transformation and dark rogue wave states arising from two-wave resonance interaction, Phys. Lett. A 378 (2014) 1095-1098.

[74] D.H. Peregrine, Water waves, nonlinear Schrödinger equations and their solutions, J. Austral. Math. Soc. Ser. B Appl. Math. 25 (1983) 16-43.

[75] T.B. Benjamin, J.E. Feir, The disintegration of wave trains on deep water part 1, Theory J. Fluid Mech. 27 (1967) 417-430.

[76] G.B. Witham, Non-linear dispersive waves, Proc. R Soc. Lond. A 283 (1965) 238-261.

[77] T. Taniuti, H, Washimi, Self-trapping and instability of hydromagnetic waves along the magnetic field in a cold plasma, Phys. Rev. Lett. 21 (1968) 209.

[78] O.C. Wright, III, Homoclinic connections of unstable plane waves of the long-waveCshort-wave equations, Stud. Appl. Math. 117 (2006) 71-93.

[79] S.H. Chen, P. Grelu, J.M. Soto-Crespo, Dark- and bright-rogue-wave solutions for media with long-wave-short-wave resonance, Phys. Rev. E 89 (2014) 011201(R).