

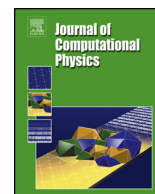


ELSEVIER

Contents lists available at ScienceDirect

## Journal of Computational Physics

www.elsevier.com/locate/jcp



# DeepParticle: Learning invariant measure by a deep neural network minimizing Wasserstein distance on data generated from an interacting particle method

Zhongjian Wang<sup>a</sup>, Jack Xin<sup>b</sup>, Zhiwen Zhang<sup>c,\*</sup>

<sup>a</sup> Department of Statistics and CCAM, The University of Chicago, Chicago, IL 60637, USA

<sup>b</sup> Department of Mathematics, University of California at Irvine, Irvine, CA 92697, USA

<sup>c</sup> Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong SAR, China

## ARTICLE INFO

## Article history:

Received 27 November 2021

Received in revised form 10 April 2022

Accepted 15 May 2022

Available online 20 May 2022

## Keywords:

Physically parameterized invariant measures

Interacting particle method

Deep learning

Wasserstein distance

Front speeds in chaotic flows

## ABSTRACT

We introduce DeepParticle, a method to learn and generate invariant measures of stochastic dynamical systems with physical parameters based on data computed from an interacting particle method (IPM). We utilize the expressiveness of deep neural networks (DNNs) to represent the transform of samples from a given input (source) distribution to an arbitrary target distribution, neither assuming distribution functions in closed form nor the sample transforms to be invertible, nor the sample state space to be finite. In the training stage, we update the weights of the network by minimizing a discrete Wasserstein distance between the input and target samples. To reduce the computational cost, we propose an iterative divide-and-conquer (a mini-batch interior point) algorithm, to find the optimal transition matrix in the Wasserstein distance. We present numerical results to demonstrate the performance of our method for accelerating the computation of invariant measures of stochastic dynamical systems arising in computing reaction-diffusion front speeds in 3D chaotic flows by using the IPM. The physical parameter is a large Péclet number reflecting the advection-dominated regime of our interest.

© 2022 Elsevier Inc. All rights reserved.

## 1. Introduction

In recent years, deep neural networks (DNNs) have achieved unprecedented levels of success in a broad range of areas such as computer vision, speech recognition, natural language processing, and health sciences, producing results comparable or superior to human experts [31,18]. The impacts have reached physical sciences where traditional first-principle based modeling and computational methodologies have been the norm. Thanks in part to the user-friendly open-source computing platforms from industry (e.g. TensorFlow and PyTorch), there have been vibrant activities in applying deep learning tools for scientific computing, such as approximating multivariate functions, solving ordinary/partial differential equations (ODEs/PDEs) and inverse problems using DNNs; see e.g. [29,52,13,14,21,45,26,2,72,50,60,73,51,25,37,36,49,66,71,1,64,39,58,70] and references therein.

There are many classical works on the approximation power of neural networks; see e.g. [12,23,15,48]. For recent works on the expressive (approximation) power of DNNs; see e.g. [10,55,69,45,39,58]. In [21], the authors showed that DNNs

\* Corresponding author.

E-mail addresses: zhongjian@statistics.uchicago.edu (Z. Wang), jxin@math.uci.edu (J. Xin), zhangzw@hku.hk (Z. Zhang).

with rectified linear unit (ReLU) activation function and enough width/depth contain the continuous piece-wise linear finite element space.

Solving ODEs or PDEs by a neural network (NN) approximation is known in the literature dating back at least to the 1990's; see e.g. [32,44,30]. The main idea in these works is to train NNs to approximate the solution by minimizing the residual of the ODEs or PDEs, along with the associated initial and boundary conditions. These early works estimate neural network solutions on a fixed mesh, however. Recently DNN methods are developed for Poisson and eigenvalue problems with a variational principle characterization (deep Ritz, [14]), for a class of high-dimensional parabolic PDEs with stochastic representations [20], and for advancing finite element methods [22,5,4]. The physics-informed neural network (PINN) method [51] and a deep Galerkin method (DGM) [60] compute PDE solutions based on their physical properties. For parametric PDEs, a deep operator network (DeepONet) learns operators accurately and efficiently from a relatively small dataset based on the universal approximation theorem of operators [40]; a Fourier neural operator method [34] directly learns the mapping from functional parametric dependence to the solutions (of a family of PDEs). Deep basis learning is studied in [43] to improve proper orthogonal decomposition for residual diffusivity in chaotic flows [42], among other works on reduced order modeling [29,62,8]. In [71,1], weak adversarial network methods are studied for weak solutions and inverse problems, see also related studies on PDE recovery from data via DNN [37,36,49,66] among others. In the context of surrogate modeling and uncertainty quantification (UQ), DNN methods include Bayesian deep convolutional encoder-decoder networks [72], deep multi-scale model learning [63], physics-constrained deep learning method [73], see also [26,55,25,68] and references therein. For DNN applications in mean field games (high dimensional optimal control problems) and various connections with numerical PDEs, see [6,7,53,3,35] and references therein. In view of the above literature, the DNN interactions with numerical PDE methods mostly occur in the Eulerian setting with PDE solutions defined in either the strong or weak (variational) sense.

Our goal here is to study *deep learning in a Lagrangian framework of multi-scale PDE problems*, coming naturally from our recent work ([41], reviewed in section 4.2 later) on a convergent interacting particle method (IPM) for computing large-scale reaction-diffusion front speeds through chaotic flows. The method is based on a genetic algorithm that evolves a large ensemble of uniformly distributed particles at the initial time to another ensemble of particles obeying an invariant measure at a large time. The front speed is readily computed from the invariant measure. Though the method is mesh-free, the computational costs remain high as advection becomes dominant at large Péclet numbers. Clearly, it is desirable to initialize particle distribution with some resemblance of the target invariant measure instead of starting from the uniform distribution. Hence learning from samples of invariant measure at a smaller Péclet number with more affordable computation to predict an invariant measure at a larger Péclet number becomes a natural problem to study.

Specifically, we shall develop a DNN  $f(\cdot; \theta)$  to map a uniform distribution  $\mu$  (source) to an invariant measure  $\nu = \nu(\kappa)$  (target), where  $\kappa$  is the reciprocal of Péclet number, and  $\theta$  consists of network weights and  $\kappa$ . The network is deep from input to output with 12 Sigmoid layers while the first three layers are coupled to a shallow companion network to account for the effects of parameter  $\kappa$  on the network weights. In addition, we include both local and nonlocal skip connections along the deep direction to assist information flow. The network is trained by minimizing the 2-Wasserstein distance (2-WD) between two measures  $\mu$  and  $\nu$  [61]. We consider a discrete version of 2-WD for finitely many samples of  $\mu$  and  $\nu$ , which involves a linear program (LP) optimizing over doubly stochastic matrices [59]. Directly solving the LP by the interior point method [65] is too costly. We devise a mini-batch interior point method by sampling smaller sub-matrices while preserving row and column sums. This turns out to be very efficient and integrated well with the stochastic gradient descent (SGD) method for the entire network training.

We shall conduct three numerical experiments to verify the performance of the proposed DeepParticle method. The first example is a synthetic data set on  $\mathbb{R}^1$ , where  $\mu$  is a uniform distribution and  $\nu$  is a normal distribution with zero mean and variance  $\sigma_1^2$ . In the second example, we compute the Kolmogorov-Petrovsky-Piskunov (KPP) front speed in a 2D steady cellular flow and learn the invariant measures corresponding to different  $\kappa$ 's. In this experiment,  $\mu$  is a uniform distribution on  $[0, 2\pi]^2$ , and  $\nu$  an empirical invariant measure obtained from IPM simulation of reaction-diffusion particles in advecting flows with Péclet number  $O(\kappa^{-1})$ . Finally, we compute the KPP front speed in a 3D time-dependent Kolmogorov flow with chaotic streamlines and study the IPM evolution of the measure in three-dimensional space. In this experiment,  $\mu$  is a uniform distribution on  $[0, 2\pi]^3$ , and  $\nu$  is an empirical invariant measure obtained from IPM simulation of reaction-diffusion particles of the KPP equation in time-dependent Kolmogorov flow. Numerical results show that the proposed DeepParticle method efficiently learns the  $\kappa$  dependent target distributions and predicts the physically meaningful sharpening effect as  $\kappa$  becomes small.

We remark that though there are other techniques for mapping distributions such as Normalizing Flows (NF) [27], Generative Adversarial Networks (GAN) [19], entropic regularization and Sinkhorn Distances [11,47], Fisher information regularization [33], our method differs either in training problem formulation or in imposing fewer constraints on the finite data samples. For example, our mapping is not required to be invertible as in NF, the training objective is not min-max as in GAN for image generation, there is no regularization effect such as blurring or noise. We believe that our method is better tailored to our problem of invariant measure learning by using a moderate amount of training data generated from the IPM. Detailed comparison will be left for future work.

The rest of the paper is organized as follows. In Section 2, we review the basic idea of DNNs and their properties, as well as Wasserstein distance. In Section 3, we introduce our DeepParticle method for learning and predicting invariant

measures based on the 2-Wasserstein distance minimization. In Section 4, we present numerical results to demonstrate the performance of our method. Finally, concluding remarks are made in Section 5.

## 2. Preliminaries

### 2.1. Artificial neural network

In this section, we introduce the definition and approximation properties of DNNs. There are two ingredients in defining a DNN. The first one is a (vector) linear function of the form  $T : \mathbb{R}^n \rightarrow \mathbb{R}^m$ , defined as  $T(x) = Ax + b$ , where  $A = (a_{ij}) \in \mathbb{R}^{m \times n}$ ,  $x \in \mathbb{R}^n$  and  $b \in \mathbb{R}^m$ . The second one is a nonlinear activation function  $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ . A frequently used activation function, known as the rectified linear unit (ReLU), is defined as  $\sigma(x) = \max(0, x)$  [31]. In the neural network literature, the sigmoid function is another frequently used activation function, which is defined as  $\sigma(x) = (1 + e^{-x})^{-1}$ . By applying the activation function in an element-wise manner, one can define (vector) activation function  $\sigma : \mathbb{R}^m \rightarrow \mathbb{R}^m$ .

Equipped with those definitions, we are able to define a continuous function  $F(x)$  by a composition of linear transforms and activation functions, i.e.,

$$F(x) = T^k \circ \sigma \circ T^{k-1} \circ \sigma \cdots \circ T^1 \circ \sigma \circ T^0(x), \quad (1)$$

where  $T^i(x) = A_i x + b_i$  with  $A_i$  be undetermined matrices and  $b_i$  be undetermined vectors, and  $\sigma(\cdot)$  is the element-wise defined activation function. Dimensions of  $A_i$  and  $b_i$  are chosen to make (1) meaningful. Such a DNN is called a  $(k+1)$ -layer DNN, which has  $k$  hidden layers. Denoting all the undetermined coefficients (e.g.,  $A_i$  and  $b_i$ ) in (1) as  $\theta \in \Theta$ , where  $\theta$  is a high-dimensional vector and  $\Theta$  is the space of  $\theta$ . The DNN representation of a continuous function can be viewed as

$$F = F(x; \theta). \quad (2)$$

Let  $\mathbb{F} = \{F(\cdot, \theta) | \theta \in \Theta\}$  denote the set of all expressible functions by the DNN parameterized by  $\theta \in \Theta$ . Then,  $\mathbb{F}$  provides an efficient way to represent unknown continuous functions, in contrast with a linear solution space in classical numerical methods, e.g. the space of linear nodal basis functions in the finite element methods and orthogonal polynomials in the spectral methods.

### 2.2. Wasserstein distance and optimal transportation

Wasserstein distances are metrics on probability distributions inspired by the problem of optimal mass transportation. They measure the minimal effort required to reconfigure the probability mass of one distribution in order to recover the other distribution. They are ubiquitous in mathematics, especially in fluid mechanics, PDEs, optimal transportation, and probability theory [61].

One can define the  $p$ -Wasserstein distance between probability measures  $\mu$  and  $\nu$  on a metric space  $Y$  with distance function  $dist$  by

$$W_p(\mu, \nu) := \left( \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{Y \times Y} dist(y', y)^p d\gamma(y', y) \right)^{1/p} \quad (3)$$

where  $\Gamma(\mu, \nu)$  is the set of probability measures  $\gamma$  on  $Y \times Y$  satisfying  $\gamma(A \times Y) = \mu(A)$  and  $\gamma(Y \times B) = \nu(B)$  for all Borel subsets  $A, B \subset Y$ . Elements  $\gamma \in \Gamma(\mu, \nu)$  are called couplings of the measures  $\mu$  and  $\nu$ , i.e., joint distributions on  $Y \times Y$  with marginals  $\mu$  and  $\nu$  on each axis.

In the discrete case, the definition (3) has a simple intuitive interpretation: given a  $\gamma \in \Gamma(\mu, \nu)$  and any pair of locations  $(y', y)$ , the value of  $\gamma(y', y)$  tells us what proportion of  $\mu$  mass at  $y'$  should be transferred to  $y$ , in order to reconfigure  $\mu$  into  $\nu$ . Computing the effort of moving a unit of mass from  $y'$  to  $y$  by  $dist(y', y)^p$  yields the interpretation of  $W_p(\mu, \nu)$  as the minimal effort required to reconfigure  $\mu$  mass distribution into that of  $\nu$ . A cost function  $c(y', y)$  on  $Y \times Y$  tells us how much it costs to transport one unit of mass from location  $y'$  to location  $y$ . When the cost function  $c(y', y) = dist(y', y)^p$ , the  $p$ -Wasserstein distance (3) reveals the Monge-Kantorovich optimization problem, see [61] for more exposition.

In a practical setting [47], the closed-form solution of  $\mu$  and  $\nu$  may be unknown, instead only  $N$  independent and identically distributed (i.i.d.) samples of  $\mu$  and  $\nu$  are available. We approximate the probability measures  $\mu$  and  $\nu$  by empirical distribution functions:

$$\mu = \frac{1}{N} \sum_{i=1}^N \delta_{x_i} \quad \text{and} \quad \nu = \frac{1}{N} \sum_{j=1}^N \delta_{y_j}. \quad (4)$$

Any element in  $\Gamma(\mu, \nu)$  can clearly be represented by an  $N \times N$  doubly stochastic matrix [59], denoted as transition matrix,  $\gamma = (\gamma_{ij})_{i,j}$  satisfying:

$$\gamma_{ij} \geq 0; \quad \forall j, \sum_{i=1}^N \gamma_{ij} = 1; \quad \forall i, \sum_{j=1}^N \gamma_{ij} = 1. \quad (5)$$

The empirical distribution functions allow us to approximate different measures. And the doubly stochastic matrix provides a practical tool to study the transformation of measures via minimizing the Wasserstein distance between different empirical distribution functions. Notice that when the number of particles becomes large, it is expensive to find the transition matrix  $\gamma$  to calculate discrete Wasserstein distance. We propose an efficient submatrix sampling method to overcome this difficulty in the next section.

### 3. Methodology

#### 3.1. Physical parameter dependent neural networks

Compared with general neural networks, we propose a new architecture to learn the dependency of physical parameters. Such type of network is expected to have independently batched input, and the output should also rely on some physical parameter that is shared by the batched input. Thus, in addition to concatenating the physical parameters as input, we also include some linear layers whose weights and biases are generated from a shallow network; see Fig. 1 for the layout of the proposed network used in later numerical experiments.

To be more precise, we let the network take on two kinds of input,  $X$  and  $\eta$ . Their batch sizes are denoted as  $N$  and  $n_\eta$  respectively. Then, there are  $n_\eta$  sets of  $X$ , such that  $X$  in each set shares the same physical parameter  $\eta$ . From `layer1` to `layer12` are general linear layers in which the weights and biases are randomly initialized and updated by Adams descent during training. They are 20 in width and Sigmoid function is applied as activation between adjacent layers. Arrows denote the direction in forward propagation. From `layer1_2` to `layer3_2` are linear layers with 20 in width, but every element of weight matrix and bias vector is individually generated from the `par-net` specified on the right. Input of `par-net` is the shared parameter, namely  $\eta$ . From `par-net:layer1` to `par-net:layer3` are general linear layers with 10 in width. For example, given the physical parameter  $\eta$  is  $d_\eta$  dimension, to generate a  $20 \times 20$  weight matrix, we first tile  $\eta$  to  $\text{bsz} \times 20 \times 20 \times d_\eta$  tensor  $\boldsymbol{\eta}$  (boldface for tensor). Here `bsz` denotes the shape of input batches which are independently forwarded in the network. In our numerical examples,  $\text{bsz} = n_\eta \times N$ . Then, we introduce a  $20 \times d_\eta \times 10$  tensor as weight matrix  $w$  in `par-net:layer1` and do matrix multiplication of  $\boldsymbol{\eta}$  and  $w$  on the last two dimensions while keeping dimensions in front. The third dimension in  $w$  is the width of the `par-net:layer3`. Shapes of weight tensor in `par-net:layer2` and `par-net:layer3` are  $20 \times 10 \times 10$  and  $20 \times 10 \times 1$  respectively. The weight parameters of the `par-net` are randomly initialized and updated by gradient descent during training. We also include some skip connections as in Resnet to improve the performance of deeper layers in our network. In addition to standard Resnet short cuts, we design a linear propagation path directly from the input  $x$  to the output  $f(x)$ . In numerical experiment shown later, such connection helps avoid output being over-clustered. There is no activation function between the last layer and output.

#### 3.2. DeepParticle algorithms

Given distributions  $\mu$  and  $\nu$  defined on metric spaces  $X$  and  $Y$ , we aim to construct a transport map  $f^0 : X \rightarrow Y$  such that  $f_*^0(\mu) = \nu$ , where star denotes the push forward of the map. On the other hand, given function  $f : X \rightarrow Y$ , the  $p$ -Wasserstein distance between  $f_*(\mu)$  and  $\nu$  is defined by:

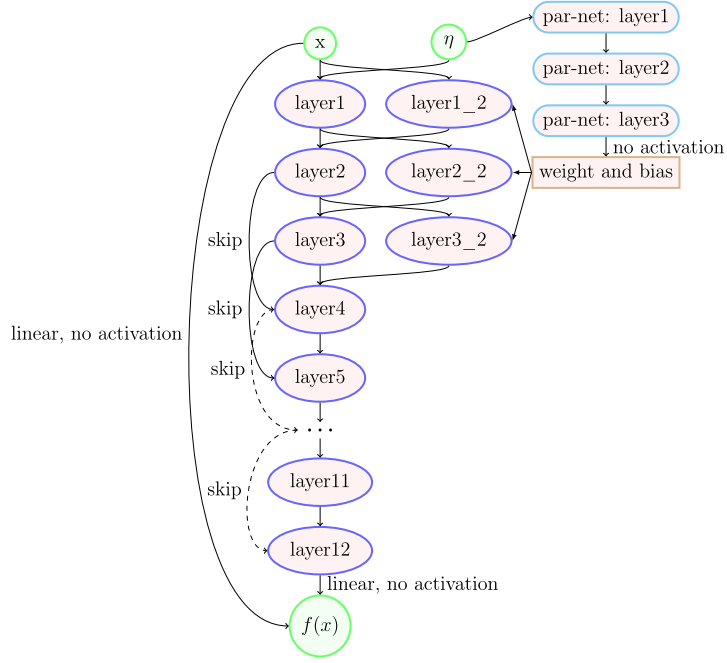
$$W_p(f_*(\mu), \nu) := \left( \inf_{\gamma \in \Gamma(f_*(\mu), \nu)} \int_{Y \times Y} \text{dist}(y', y)^p d\gamma(y', y) \right)^{1/p}, \quad (6)$$

where  $\Gamma(f_*(\mu), \nu)$  denotes the collection of all measures on  $Y \times Y$  with marginals  $f_*(\mu)$  and  $\nu$  on the first and second factors respectively and  $\text{dist}$  denotes the metric (distance) on  $Y$ . A straightforward derivation yields:

$$W_p(f_*(\mu), \nu) = \left( \inf_{\gamma \in \Gamma(\mu, \nu)} \int_{X \times Y} \text{dist}(f(x), y)^p d\gamma(x, y) \right)^{1/p}, \quad (7)$$

where  $\Gamma(\mu, \nu)$  denotes the collection of all measures on  $X \times Y$  with marginals  $\mu$  and  $\nu$  on the first and second factors respectively and still  $\text{dist}$  denotes the metric (distance) on  $Y$ .

**Network training objective** Our DeepParticle algorithm does not assume the knowledge of closed form distribution of  $\mu$  and  $\nu$ , instead we have i.i.d. samples of  $\mu$  and  $\nu$  namely,  $x_i$  and  $y_j$ ,  $i, j = 1, \dots, N$ , as training data. Then a discretization of (7) is:



**Fig. 1.** Layout of the proposed deep network where  $\eta$  is physical parameter input for learning its implicit dependence in the output  $f(x)$ .

$$\hat{W}(f) := \left( \inf_{\gamma \in \Gamma^N} \sum_{i,j=1}^N \text{dist}(f(x_i), y_j)^p \gamma_{ij} \right)^{1/p}, \quad (8)$$

where  $\Gamma^N$  denotes all  $N \times N$  doubly stochastic matrices.

Let the map (represented by neural network in Fig. 1) be  $f_\theta(x; \eta)$  where  $x$  is the input,  $\eta$  is the shared physical parameter and  $\theta$  denotes all the trainable parameters in the network. Clearly  $\hat{W}(f_\theta) \geq 0$ . In case of  $X = Y = \mathbb{R}^d$  equipped with Euclidean metric, we take  $p = 2$ . The training loss function is

$$\hat{W}^2(f_\theta) := \sum_{r=1}^{n_\eta} \left( \inf_{\gamma_r \in \Gamma^N} \sum_{i,j=1}^N |f_\theta(x_{i,r}; \kappa_r) - y_{j,r}|^2 \gamma_{i,j,r} \right). \quad (9)$$

*Iterative method in finding transition matrix  $\gamma$*  To minimize the loss function (9), we update parameters  $\theta$  of  $f_\theta$  with the classical Adams stochastic gradient descent, and alternate with updating the transition matrix  $\gamma$ .

We now present a *mini-batch linear programming algorithm* to find the best  $\gamma$  for each inner sum of (9), while suppressing  $k_r$  dependence in  $f_\theta$ . Notice that the problem (9) is a linear program on the bounded convex set  $\Gamma^N$  of vector space of real  $N \times N$  matrices. By Choquet's theorem, this problem admits solutions that are extremal points of  $\Gamma^N$ . Set of all doubly stochastic matrix  $\Gamma^N$  can be referred to as Birkhoff-von Neumann theorem [54] states that such polytope is the convex hull of all permutation matrices, i.e., those matrices such that  $\gamma_{ij} = \delta_{j,\pi(i)}$  for some permutation  $\pi$  of  $\{1, \dots, N\}$ , where  $\delta_{jk}$  is the Kronecker symbol.

The algorithm is defined iteratively. In each iteration, we select columns and rows and solve a linear programming sub-problem under the constraint that maintains column-wise and row-wise sums of the corresponding sub-matrix of  $\gamma$ . To be precise, let  $\{i_k\}_{k=1}^M, \{j_l\}_{l=1}^M$  ( $M \ll N$ ) denote the index chosen from  $\{1, 2, \dots, N\}$  without replacement. The cost function of the sub-problem is

$$C(\gamma^*) := \sum_{k,l=1}^M |f_\theta(x_{i_k}) - y_{j_l}|^2 \gamma_{i_k,j_l}^* \quad (10)$$

subject to

$$\begin{cases} \sum_{k=1}^M \gamma_{i_k,j_l}^* = \sum_{k=1}^M \gamma_{i_k,j_l} & \forall l = 1, \dots, M \\ \sum_{l=1}^M \gamma_{i_k,j_l}^* = \sum_{l=1}^M \gamma_{i_k,j_l} & \forall k = 1, \dots, M \\ \gamma_{i_k,j_l}^* \geq 0 & \forall k, l = 1, \dots, M, \end{cases} \quad (11)$$

where  $\gamma_{i_k, j_l}$  are from the previous step. This way, the column and row sums of  $\gamma$  are preserved by the update. The linear programming sub-problem of much smaller size is solved by the interior point method [65].

We observe that the global minimum of  $\gamma$  in (9) is also the solution of sub-problems (10) with arbitrarily selected rows and columns, subject to the row and column partial sum values of the global minimum. The selection of rows and columns can be one's own choice. In our approach, in each step after gradient descent, apart from random sampling of rows and columns, we also perform a number of searches for rows or columns with the largest entries to speed up computation; see Algorithm 1.

---

### Algorithm 1: Random Pivot Search

---

**Result:** Given transition matrix  $\gamma$ , randomly search  $M$  rows/columns with largest elements ("pivots").  
 Randomly pick  $i_1$  as the first row;  
 find  $j_1$  such that  $\gamma_{i_1 j_1}$  is the largest among  $\{\gamma_{i_1 1}, \dots, \gamma_{i_1 N}\}$ ;  
**for**  $k \leftarrow 2$  **to**  $M$  **do**  
 | find  $i_k$  such that  $\gamma_{i_k j_{k-1}}$  is the largest among  $\{\gamma_{i' j_{k-1}} | i' \in \{1, \dots, N\} \setminus \{i_1, \dots, i_{k-1}\}\}$ ;  
 | find  $j_k$  such that  $\gamma_{i_k j_k}$  is the largest among  $\{\gamma_{i_k j'} | j' \in \{1, \dots, N\} \setminus \{j_1, \dots, j_{k-1}\}\}$ ;  
**end**

---

The cost of finding optimal  $\gamma$  increases as  $N$  increases, however, the network itself is independent of  $\gamma$ . After training, our network acts as a sampler from some target distribution  $\nu$  without assumption of closed-form distribution of  $\nu$ . At this stage, the input data is no longer limited by training data, an arbitrarily large amount of samples approximately obeying  $\nu$  can be generated through  $\mu$  (uniform distribution).

*Update of training data* Note that given any fixed set of  $\{x_i\}$  and  $\{y_j\}$  (training data), we have developed the iterative method to calculate the optimal transition matrix  $\gamma$  in (9) and update network parameter  $\theta$  at the same time. However, in more complicated cases, more than one set of data (size of which is denoted as  $\text{bsz}$ ) should be assimilated. The total number of data set is denoted  $N_{\text{dict}}$ . For the second and later set of training data, the network is supposed to establish some accuracy. So before updating the network parameter, we first utilize our iterative method to find an approximated new  $\gamma$  for the new data feed to reach a preset level measured by the *normalized Frobenius norm*  $\|\cdot\|_{F_{\text{ro}}}$  of  $\gamma$  ( $\|\gamma\|_F$  divided by the square root of its row number). The full training process is outlined in Algorithm 2.

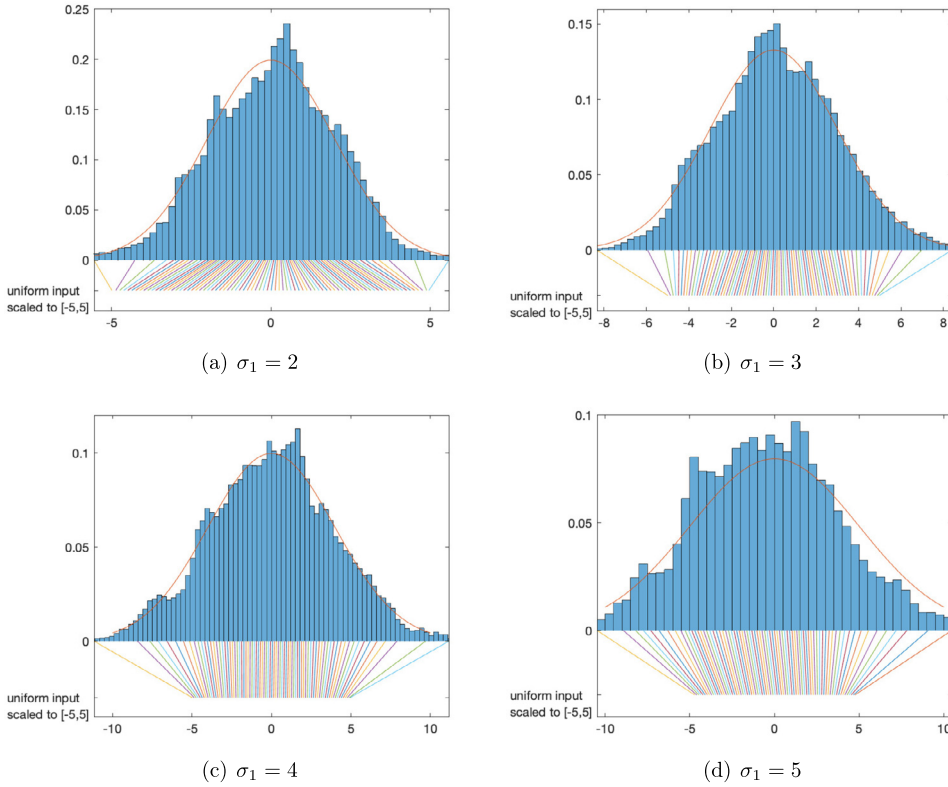
---

### Algorithm 2: DeepParticle Learning

---

Randomly initialize weight parameters  $\theta$  in network  $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^d$ ;  
**repeat**  
 | **for** physical parameter set  $r \leftarrow 0$  **to**  $n_\eta$  **do**  
 | | randomly select  $\{x_{i,r}\}, \{y_{j,r}\}, i, j = 1 : N$  from i.i.d. samples of input and target distribution with respect to physical parameter  $\eta_r$ ;  
 | |  $\gamma_{i,j,r} = 1/N$ ;  
 | **end**  
 | **if** not the first training mini-batch **then**  
 | | **for** physical parameter set  $r \leftarrow 0$  **to**  $n_\eta$  **do**  
 | | |  $P_r = \sum_{i,j=1}^N |f_\theta(x_{i,r}, \eta_r) - y_{j,r}|^2 \gamma_{i,j,r}$ ;  
 | | | **while**  $\|\gamma_r\|_{F_{\text{ro}}} < \text{tol}$  **do**  
 | | | | randomly (or by Algorithm 1) choose  $\{i_{k,r}\}_{k=1}^M, \{j_{l,r}\}_{l=1}^M$  from  $\{1, 2, \dots, N\}$  without replacement;  
 | | | | solve the linear programming sub-problem (10)-(11) to get  $\gamma_r^*$ ;  
 | | | | update  $\{\gamma_{i_{k,r} j_{l,r}}\}_{k,l=1}^M$  with  $\{\gamma_{i_{k,r} j_{l,r}}^*\}_{k,l=1}^M$ .  
 | | | **end**  
 | | **end**  
 | **end**  
 |  $P = \sum_{r=1}^{N_r} \sum_{i,j=1}^N |f_\theta(x_{i,r}, \eta_r) - y_{j,r}|^2 \gamma_{i,j,r}$ ;  
 |  $\theta \leftarrow \theta - \delta_1 \nabla_\theta P$ ,  $\delta_1$  is the learning step size;  
 | **repeat**  
 | | **for** physical parameter set  $r \leftarrow 0$  **to**  $n_\eta$  **do**  
 | | |  $P_r = \sum_{i,j=1}^N |f_\theta(x_{i,r}, \eta_r) - y_{j,r}|^2 \gamma_{i,j,r}$ ;  
 | | | randomly (or by Algorithm 1) choose  $\{i_{k,r}\}_{k=1}^M, \{j_{l,r}\}_{l=1}^M$  from  $\{1, 2, \dots, N\}$  without replacement;  
 | | | solve the linear programming sub-problem (10)-(11) to get  $\gamma_r^*$ ;  
 | | | update  $\{\gamma_{i_{k,r} j_{l,r}}\}_{k,l=1}^M$  with  $\{\gamma_{i_{k,r} j_{l,r}}^*\}_{k,l=1}^M$ .  
 | | **end**  
 | | **until** given linear programming steps,  $N_{LP}$ ;  
 | **until** given steps for each training mini-batch;  
**until** given number of training mini-batches,  $N_{\text{dict}}$ ;  
**Return**

---



**Fig. 2.** Generated histograms with the red lines being the ground truth pdfs. Lines under the pdfs connect inputs, scaled to  $Unif[-5, 5]$ , and their corresponding outputs of the network. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

## 4. Numerical examples

### 4.1. Mapping uniform to normal distribution

For illustration, we first apply our algorithm to learn a map from 1D uniform distribution on  $[0, 1]$  to 1D normal distribution with zero mean and various standard deviation  $\sigma_1$ . The  $\sigma_1$  refers to one dimensional  $\eta$  in the network of Fig. 1. As training data, we independently generate  $n_{\sigma_1} = 8$  sets of  $N = 1500$  uniformly distributed points, also  $N = 1500$  normally distributed sample points with zero mean and standard deviation  $\sigma_1$  equally spaced in  $[2, 3.75]$ . During training, we aim to find a  $\sigma_1$ -dependent network such that the output from the 8 input data sets along with their  $\sigma_1$  values approximate normal distributions with standard deviation  $\sigma_1$ , respectively. The layout of network is shown in Fig. 1. The total number of training steps is  $10^4$  with initial learning rate 0.02 and  $N_{dict} = 1$  data set. After each update of parameters, we solve the optimization problem of transition matrix  $\gamma$ , i.e. Eq. (10) under Eq. (11), with selection of rows and columns by Algorithm 1 for  $N_{LP} = 5$  times. After training, we generate  $N = 40000$  uniformly distributed test data points and apply the networks with  $\sigma_1 = 2, 3, 4, 5$ .

In Fig. 2, we plot the output histograms. Clearly, the empirical distributions vary with  $\sigma_1$  and fit the reference normal probability distribution functions (pdfs) in red lines. We would like to clarify that even though the dependence of  $\sigma_1$  in this example is ‘linear’, the linear propagation directly from input to output does not depend on  $\sigma_1$  and output `par-net` is no longer a linear function of  $\sigma_1$ . In addition, we plot the input-output pairs at the bottom of each empirical distribution. Each line connects the input (for better visualization we scale it to  $Unif[-5, 5]$ ) and its corresponding output from the network. In one dimensional case, it is well known that the inverse of the cumulative distribution function is the optimal transport map. It is a *monotone* function that maps uniform distribution to the quantile of the target distribution. In the plotted input-output pair, we can see there are no lines crossing. It verifies that our network finds the *monotone* transport map. Our experiments here and below are all carried out on a quad-core CPU desktop with an RTX2080 8GB GPU at UC Irvine.

### 4.2. Computing front speeds in complex fluid flows

Front propagation in complex fluid flows arises in many scientific areas such as turbulent combustion, chemical kinetics, biology, transport in porous media, and industrial deposition processes [67]. A fundamental problem is to analyze and

compute large-scale front speeds. An extensively studied model problem is the reaction-diffusion-advection (RDA) equation with Kolmogorov-Petrovsky-Piskunov (KPP) nonlinearity [28]:

$$u_t = \kappa \Delta_{\mathbf{x}} u + (\mathbf{v} \cdot \nabla_{\mathbf{x}}) u + u(1 - u), \quad t \in \mathbb{R}^+, \quad \mathbf{x} = (x_1, \dots, x_d)^T \in \mathbb{R}^d, \quad (12)$$

where  $\kappa$  is diffusion constant,  $\mathbf{v}$  is an incompressible velocity field (precise definition given later), and  $u$  is the concentration of reactant. Let us consider velocity fields  $\mathbf{v} = \mathbf{v}(t, \mathbf{x})$  to be  $T$ -periodic in space and time, which contain the celebrated Arnold-Beltrami-Childress (ABC) and Kolmogorov flows as well as their variants with chaotic streamlines [9,24]. The solutions from compact non-negative initial data spread along direction  $\mathbf{e}$  with speed [46]:

$$c^*(\mathbf{e}) = \inf_{\alpha > 0} \lambda(\alpha)/\alpha,$$

where  $\lambda(\alpha)$  is the principal eigenvalue of the parabolic operator  $\partial_t - \mathcal{A}$  with:

$$\mathcal{A}w := \kappa \Delta_{\mathbf{x}} w + (2\alpha \mathbf{e} + \mathbf{v}) \cdot \nabla_{\mathbf{x}} w + (\kappa \alpha^2 + \alpha \mathbf{v} \cdot \mathbf{e} + 1)w, \quad (13)$$

on the space domain  $\mathbb{T}^d := [0, T]^d$  (periodic boundary condition). It is known [46] that  $\lambda(\alpha)$  is convex in  $\alpha$ , and superlinear for large  $\alpha$ . The operator  $\mathcal{A}$  in (13) is a sum  $\mathcal{A} = \mathcal{L} + \mathcal{C}$ , with

$$\mathcal{L} := \kappa \Delta_{\mathbf{x}} \cdot + (2\alpha \mathbf{e} + \mathbf{v}) \cdot \nabla_{\mathbf{x}} \cdot, \quad \mathcal{C} := c(t, \mathbf{x}) \cdot = (\kappa \alpha^2 + \alpha \mathbf{v} \cdot \mathbf{e} + 1). \quad (14)$$

where  $\mathcal{L}$  is the generator of a Markov process, and  $\mathcal{C}$  acts as a potential. The Feynman-Kac (FK) formula [16] gives  $\lambda(\alpha)$  a stochastic representation:

$$\lambda = \lim_{t \rightarrow \infty} t^{-1} \ln \left( \mathbb{E} \exp \left\{ \int_0^t c(t-s, \mathbf{X}_s^{t, \mathbf{x}}) ds \right\} \right). \quad (15)$$

In Eq. (15),  $\mathbf{X}_s^{t, \mathbf{x}}$  satisfies the following stochastic differential equation

$$d\mathbf{X}_s^{t, \mathbf{x}} = \mathbf{b}(t-s, \mathbf{X}_s^{t, \mathbf{x}}) ds + \sqrt{2\kappa} d\mathbf{W}_s, \quad \mathbf{X}_0^{t, \mathbf{x}} = \mathbf{x}, \quad (16)$$

where the drift term  $\mathbf{b} = 2\alpha \mathbf{e} + \mathbf{v}$  and  $\mathbf{W}(s)$  is the standard  $d$ -dimensional Wiener process. The expectation  $\mathbb{E}(\cdot)$  in (15) is over  $\mathbf{W}(t)$ . Directly applying (15) and Monte Carlo method to compute  $\lambda(\alpha)$  is unstable as the main contribution to  $\mathbb{E}(\cdot)$  comes from sample paths that visit maximal or minimal points of the potential function  $c$ , resulting in inaccurate or even divergent results. A computationally feasible alternative is a scaled version, the FK semi-group:

$$\Phi_t^c(\nu)(\phi) := \frac{\mathbb{E}[\phi(\mathbf{X}_t^{t, \mathbf{x}}) \exp\{\int_0^t c(t-s, \mathbf{X}_s^{t, \mathbf{x}}) ds\}]}{\mathbb{E}[\exp\{\int_0^t c(t-s, \mathbf{X}_s^{t, \mathbf{x}}) ds\}]} := \frac{P_t^c(\nu)(\phi)}{P_t^c(\nu)(1)}.$$

Acting on any initial probability measure  $\nu$ ,  $\Phi_{nT}^c(\nu)$  converges weakly to an invariant measure  $\nu_c$  (i.e.  $\Phi_T^c(\nu_c) = \nu_c$ ) as  $n \rightarrow \infty$ , for any smooth test function  $\phi$ . Moreover,

$$P_t^c(\nu_c) = \exp\{\lambda t\} \nu_c \quad \text{or} \quad \lambda = t^{-1} \ln \mathbb{E}_{\nu_c}[P_t^c(\nu_c)]. \quad (17)$$

Given a time discretization step  $\Delta t$ , an interacting particle method (IPM) proceeds to discretize  $\mathbf{X}_s^{t, \mathbf{x}}$  as  $\mathbf{X}_i^{\Delta t}$ , ( $i = 1, \dots, n \times m$ , where  $m = \frac{T}{\Delta t}$ ) by Euler's method, approximates the evolution of probability measure  $\Phi_t^c(\nu)$  by a particle system with a re-sampling technique to reduce variance. Let

$$P_n^{c, \Delta t}(\nu)(\phi) := \mathbb{E} \left[ \phi(\mathbf{X}_{nm}^{\Delta t}) \exp \left\{ \Delta t \sum_{i=1}^m c((m-i)\Delta t, \mathbf{X}_{i+(n-1)m}^{\Delta t}) \right\} \right]. \quad (18)$$

Then sampled FK semi-group actions on  $\nu$ :

$$\Phi_n^{c, \Delta t}(\nu)(\phi) = \frac{P_n^{c, \Delta t}(\nu)(\phi)}{P_n^{c, \Delta t}(\nu)(1)} \longrightarrow \int_{\mathbb{T}^d} \phi d\nu_{c, \Delta t}, \quad \text{as } n \uparrow \infty, \quad \forall \text{ smooth } \phi,$$

where  $\nu_{c, \Delta t}$  is an invariant measure of the discrete map  $\Phi_1^{c, \Delta t}$ , thanks to  $\mathbf{b}$  being  $T$ -periodic in time. There exists  $q \in (0, 1)$  so that [41]:

$$\lambda_{\Delta t}^n := (nT)^{-1} \ln[P_n^{c, \Delta t}(\nu_0)(1)] \xrightarrow{n \rightarrow \infty} T^{-1} \ln[P_1^{c, \Delta t}(\nu_{c, \Delta t})(1)] = \lambda + o((\Delta t)^q). \quad (19)$$



The IPM algorithm below (Algorithm 3) approximates the evolving measure as an empirical measure of a large number of genetic particles that undergo advection-diffusion ( $\mathcal{L}$ ) and mutations ( $\mathcal{C}$ ). The mutation relies on fitness and its normalization defined as in the FK semi-group. In Algorithm 3, the evolution is phrased in  $G$  generations, each moving and mutating  $m$  times in a life span of  $T$ . The index  $n$  in (18) is changed to  $G$ .

---

**Algorithm 3:** Genetic Interacting Particle Method
 

---

Initialize first generation of  $N_0$  particles  $\xi_1^0 = (\xi_1^{0,1}, \dots, \xi_1^{0,N_0})$ , uniformly distributed over  $\mathbb{T}^d$  ( $d \geq 2$ ). Let  $g$  be the generation number in approximating  $\nu_{c, \Delta t}$ . Each generation moves and replicates  $m$ -times, with a life span  $T$  (time period), time step  $\Delta t = T/m$ .

```

while  $g = 1 : G - 1$  do
  while  $i = 0 : m - 1$  do
     $\xi_g^i \leftarrow$  one-step-advection-diffusion update on  $\xi_g^i$ . Define fitness  $F := \exp\{c(T - i\Delta t, \xi_g^i) \Delta t\}$ .
     $E_{g,i} := \frac{1}{\Delta t} \ln$  (mean population fitness).
    Normalize fitness to weight.  $\mathbf{p} := F/SUM(F)$ .
     $\xi_g^{i+1} \leftarrow$  resample  $\xi_g^i$  via multinomial distribution with weight  $\mathbf{p}$ .
  end
   $\xi_{g+1}^0 \leftarrow \xi_g^m$ ,  $E_g \leftarrow \text{mean}(E_{g,i})$  over  $i$ .
end
Output: approximate  $\lambda_{\Delta t} \leftarrow \text{mean}(E_g)$ , and  $\xi_G^0$ .

```

---

Much progress has been made in the finite element computation [57,56,74] of the KPP principal eigenvalue problem (13) particularly in steady 3D flows  $\mathbf{v} = \mathbf{v}(\mathbf{x})$ . However, when  $\kappa$  is small and the spatial dimension is 3, adaptive FEM can be extremely expensive. For 2D time periodic cellular flows, adaptive basis deep learning is found to improve the accuracy of reduced order modeling [42,43]. Extension of deep basis learning to 3D in the Eulerian setting has not been attempted partly due to the costs of generating a sufficient amount of training data.

An advantage of the IPM is that given the same particle number, the computational cost of generating training data scales linearly with the dimension of spatial variables. As a comparison, in the Eulerian framework, one needs to solve the principal eigenvalue of a parabolic operator  $\partial_t - \mathcal{A}$  by discretizing the eigenvalue problem on mesh grids. The number of mesh grids depends exponentially on the spatial dimension, which becomes expensive for 3D problems. Though adaptive mesh techniques can alleviate the computational burden, their design and implementation are challenging for time dependent 3D problems when large gradient regions are dynamically changing and repeated mesh adaptations must be performed. In contrast, the IPM is spatially mesh-free and self-adaptive. The computational bottleneck remains in IPM at small  $\kappa$  when we need a large number of particles running for many generations to approach the invariant measure. To address this issue, we apply our DeepParticle method to generate particle samples by initiating at a *learned distribution* resembling the invariant measure and accelerating Algorithm 3 (i.e. reduce  $G$  to reach convergence).

**2D steady cellular flow** We first consider a 2D cellular flow  $\mathbf{v} = (-\sin x_1 \cos x_2, \cos x_1 \sin x_2)$ . In this case, we apply the physical parameter dependent network described in Section 3.1 to learn the invariant measure corresponding to different  $\kappa$  values. To generate training data, we first use the IPM, Algorithm 3 with  $\kappa_i = 2^{-2-0.25*(i-1)}$ ,  $i = 1 : 8 (= n_\eta)$ , and  $N_0 = 40000$  particle evolution for  $G = 2048$ ,  $\Delta t = 2^{-8}$ ,  $T = 1$ , to get samples of invariant measure at different  $\kappa$  values. The value of  $G$  is chosen so that the direct IPM simulation of principal eigenvalue converges, see blue line of Fig. 4(d). From each set of sample points with different  $\kappa$ 's, we randomly pick  $N = 2000$  sample points without replacement, denoted as  $\mathcal{Y}_1, \dots, \mathcal{Y}_8$ . Under such setting, we then seek neural networks  $f_\theta$  such that given  $\kappa_i$  and a set  $\mathcal{X}_i$  of i.i.d. uniformly distributed points on  $[0, 2\pi]^2$ , the network output  $f_\theta(\mathcal{X}_i; \kappa_i)$  is distributed near  $\mathcal{Y}_i$  in  $W^2$  distance. The set  $\{\mathcal{Y}_1, \dots, \mathcal{Y}_8\}$  is called one mini-batch of training data and in total we have  $N_{dict} = 5$  mini-batches through 50000 steps of network training. As stated in Algorithm 2, after each mini-batch of network training (10000 steps of gradient descent), we re-optimize  $\gamma$  until its normalized Frobenius norm is greater than  $tol = 0.7$ . Note that only a quarter of the 40,000 particles in the IPM simulations have been used in training.

During training, we apply Adams gradient descent with initial learning rate 0.002 and a weight decay with hyperparameter 0.005 for trainable weights in the network. In each step, we select  $M = 25$  columns and rows by Algorithm 1 in solving the sub-optimization problem (10) and repeat the linear programming for  $N_{LP} = 10$  times.

In Fig. 3, we present the performance of our algorithm in sampling distribution in the original training data. Each time, the training algorithm only has access to at most  $N \times N_{dict} = 10000$  samples of each target distribution with  $\kappa_i = 2^{-2-0.25*(i-1)}$ ,  $i = 1, \dots, 8 = n_\eta$ . After training, we generate  $N_0 = 40000$  uniformly distributed points as input of the network. We compare the histogram of network output and that of the  $N_0 = 40000$  sample points obeying the invariant measure from the IPM Algorithm 3. We see that our trained parameter-dependent network indeed reproduces the sharpening effect on the invariant measure when  $\kappa$  becomes small.

In Fig. 4, we present the performance of algorithm in predicting the invariant measure of IPM with  $\kappa = 2^{-4}$ . First we compare the invariant measure generated from direct simulation with IPM (Fig. 4(a)) and from the trained network (Fig. 4(b)). In Fig. 4(c), we show the  $W_2$  distance between the generated distribution and the target distribution of the

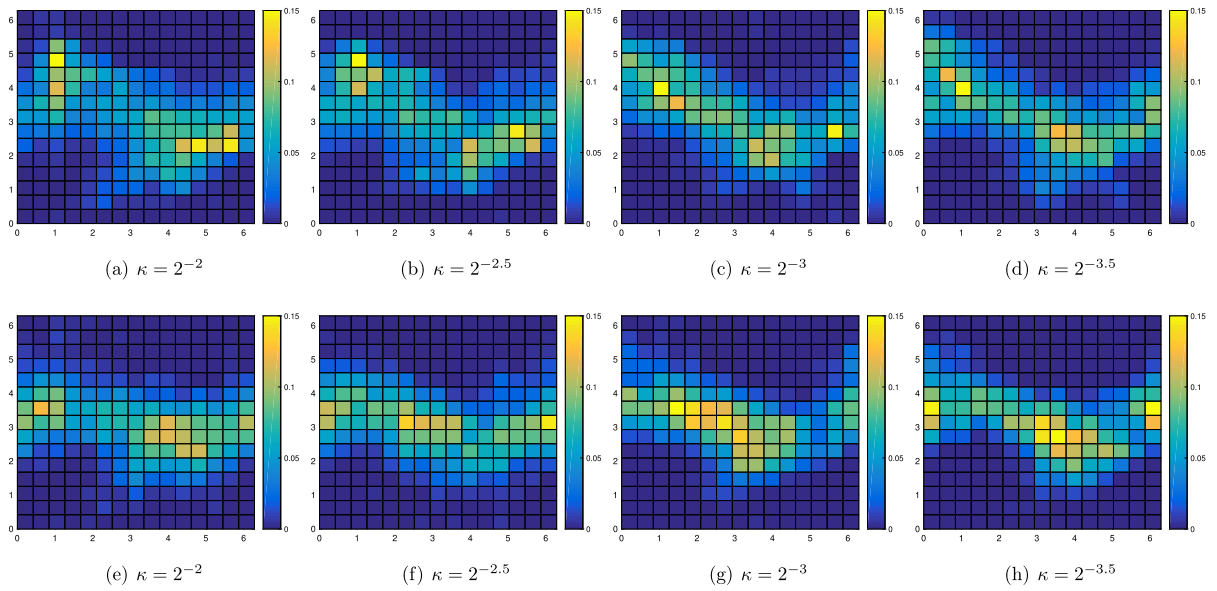
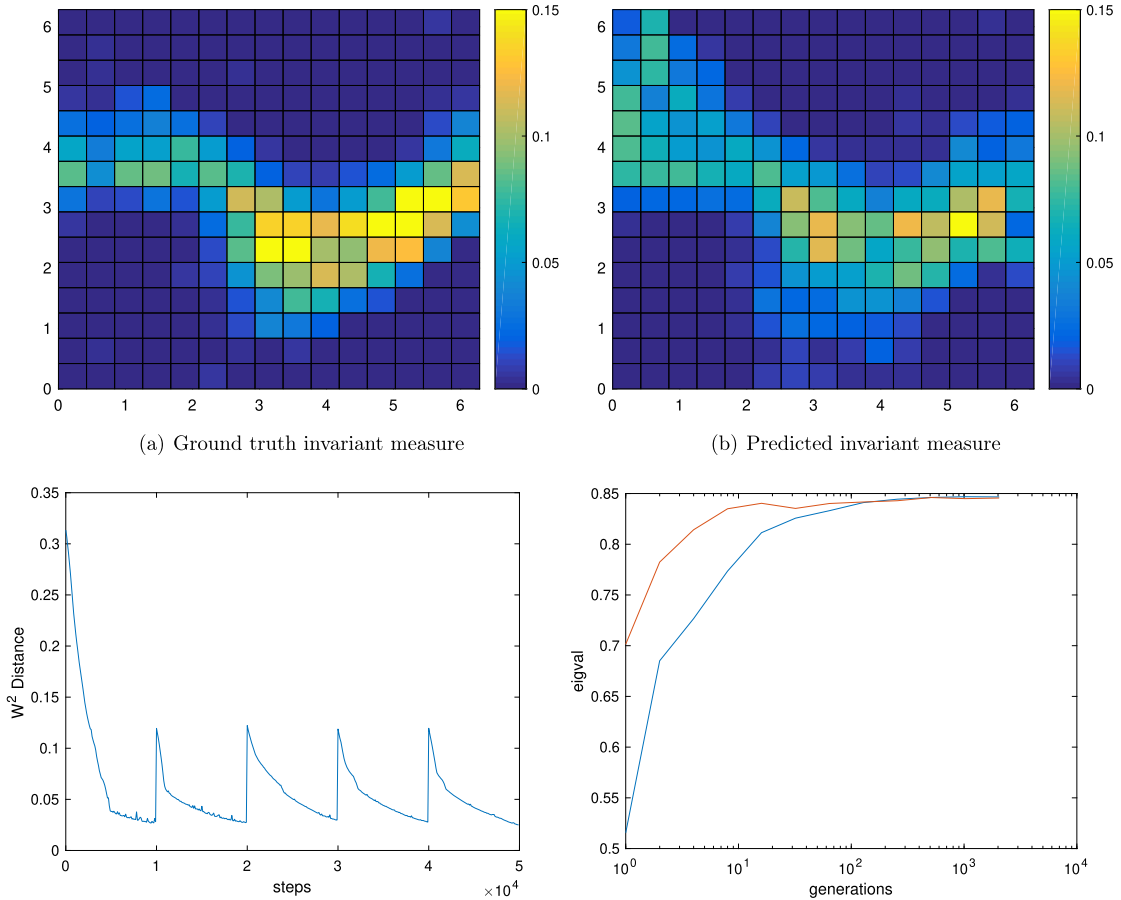


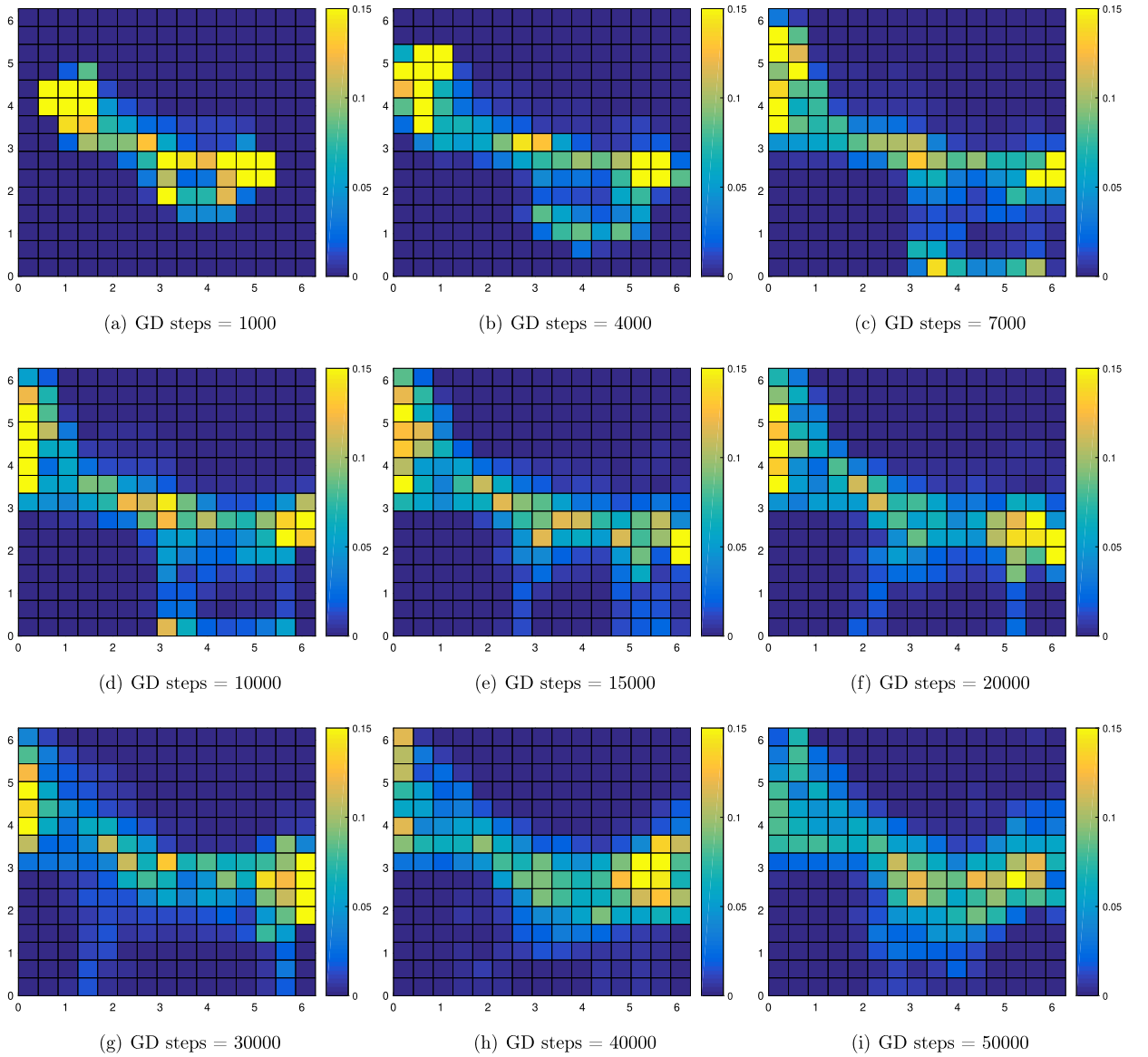
Fig. 3. Generated invariant measure (upper row) and corresponding training data (reference, lower row).



(c)  $W^2$  training error vs. gradient descent steps: spikes of height  $\approx 0.1$  occur due to transition matrix  $\gamma$  re-optimized in response to mini-batching of input training data.

(d) Acceleration: convergence to  $\lambda_{\Delta t}$  value computed by Alg.3 with warm/cold start by DeepParticle prediction (red)/uniform distribution (blue).

Fig. 4. DeepParticle prediction (top), training/acceleration (bottom) at test value  $\kappa = 2^{-4}$  in a 2D front speed computation. Loss reduction in c) shows fast (steps before  $1e4$ ) and slow phases [38].



**Fig. 5.** Measures generated at different gradient descent (GD) steps in network training for KPP front speeds in 2D steady cellular flows.

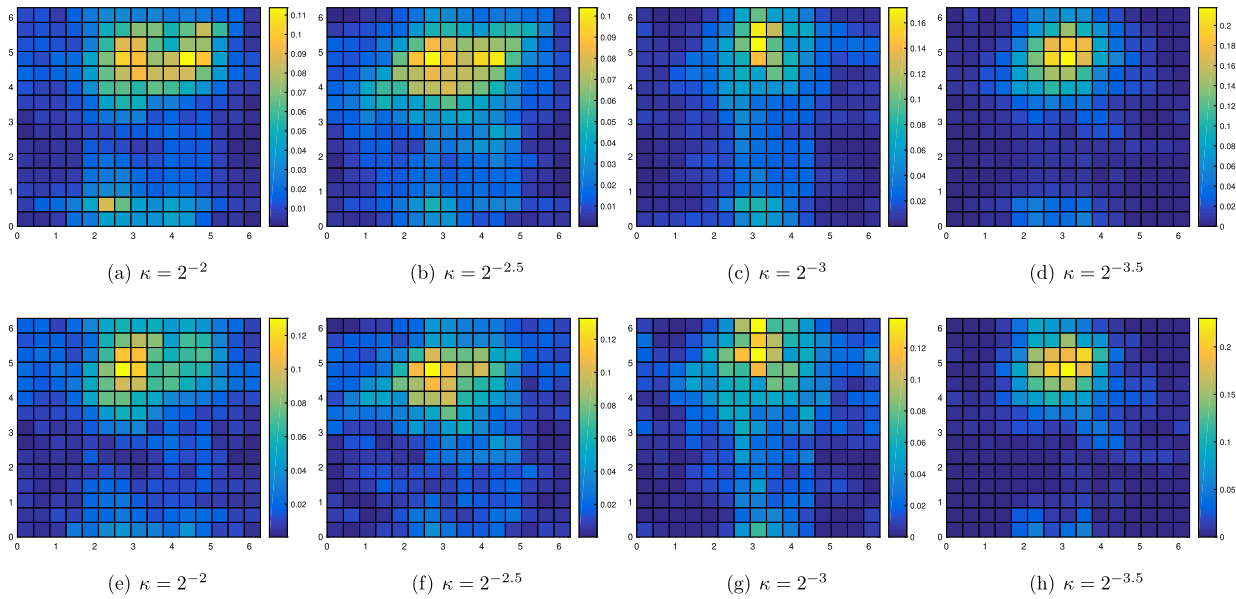
training data. The  $W_2$  distance goes up when we re-sample the training data. This is because at this moment, the transition matrix  $\gamma$  in the definition of the  $W_2$  distance is inaccurate. Note that the test  $\kappa = 2^{-4}$  value is no longer within the training range  $\kappa = 2^{-p}$ ,  $p \in [2, 3.75]$ . Our trained network predicts an invariant measure. Such prediction also serves as a ‘warm start’ for IPM and can accelerate its computation to quickly reach a more accurate invariant measure. As the invariant measure has no closed-form analytical solution, we plot in Fig. 4(d) the principal eigenvalue approximations by IPM with warm and cold starts, i.e. one generated from DeepParticle network (warm) and the i.i.d. samples from a uniform distribution (cold). The warm start by DeepParticle achieves 4x to 8x speedup.

Fig. 5 displays histograms of  $N_0 = 40000$  samples generated from our DeepParticle algorithm at different stage of training. In Fig. 5(d) with only one set of data, we already get a rough prediction of target distribution. With more data sets added, the prediction quality improves; see Fig. 5(e) to Fig. 5(i).

**3D time-dependent Kolmogorov flow** Next, we compute KPP front speed in a 3D time-dependent Kolmogorov flow [17]:

$$\mathbf{v} = (\sin(x_3 + \sin(2\pi t)), \sin(x_1 + \sin(2\pi t)), \sin(x_2 + \sin(2\pi t))).$$

The hyper-parameter setting for network training remains the same as in the 2D case. The physical parameter-dependent network in Section 3.1 learns the invariant measure corresponding to eight different  $\kappa$  training values. Besides the in-



**Fig. 6.** Generated invariant measure (upper row) and the corresponding training data (reference, lower row), viewed as projections to 2nd and 3rd dimensions in a 3D KPP front speed computation.

put  $\mathbf{x}$  and output  $\mathbf{f}(\mathbf{x})$  are both in 3 dimensions, the network layout and training procedure are the same as in the 2D example.

Fig. 6 shows the performance of the network interpolating training data from  $\kappa = 2^{-2}$  to  $\kappa = 2^{-3.5}$ . Fig. 7 displays network prediction at  $\kappa = 2^{-4}$ . All the histograms here are 2D projections of the 3D histogram to the second and third dimensions. Results of the projection to other choices of two dimensions are similar and are not shown here for brevity.

## 5. Conclusions

We developed a DeepParticle method to generate invariant measures of stochastic dynamical (interacting particle) systems by a physically parameterized DNN that minimizes the Wasserstein distance between the source and target distributions. Our method is very general in the sense that we do not require distributions to be in closed-form and the generation map to be invertible. Thus, our method is fully data-driven and applicable in the fast computation of invariant measures of other interacting particle systems. During the training stage, we update network parameters based on a discretized Wasserstein distance defined on finite samples. We proposed an iterative divide-and-conquer algorithm that allows us to significantly reduce the computational cost in finding the optimal transition matrix in the Wasserstein distance. We carried out numerical experiments to demonstrate the performance of the proposed method. Numerical results show that our method is very efficient and helps accelerate the computation of invariant measures of interacting particle systems for KPP front speeds. In the future, we plan to apply the DeepParticle method to learn and generate invariant measures of other stochastic particle systems.

### CRedit authorship contribution statement

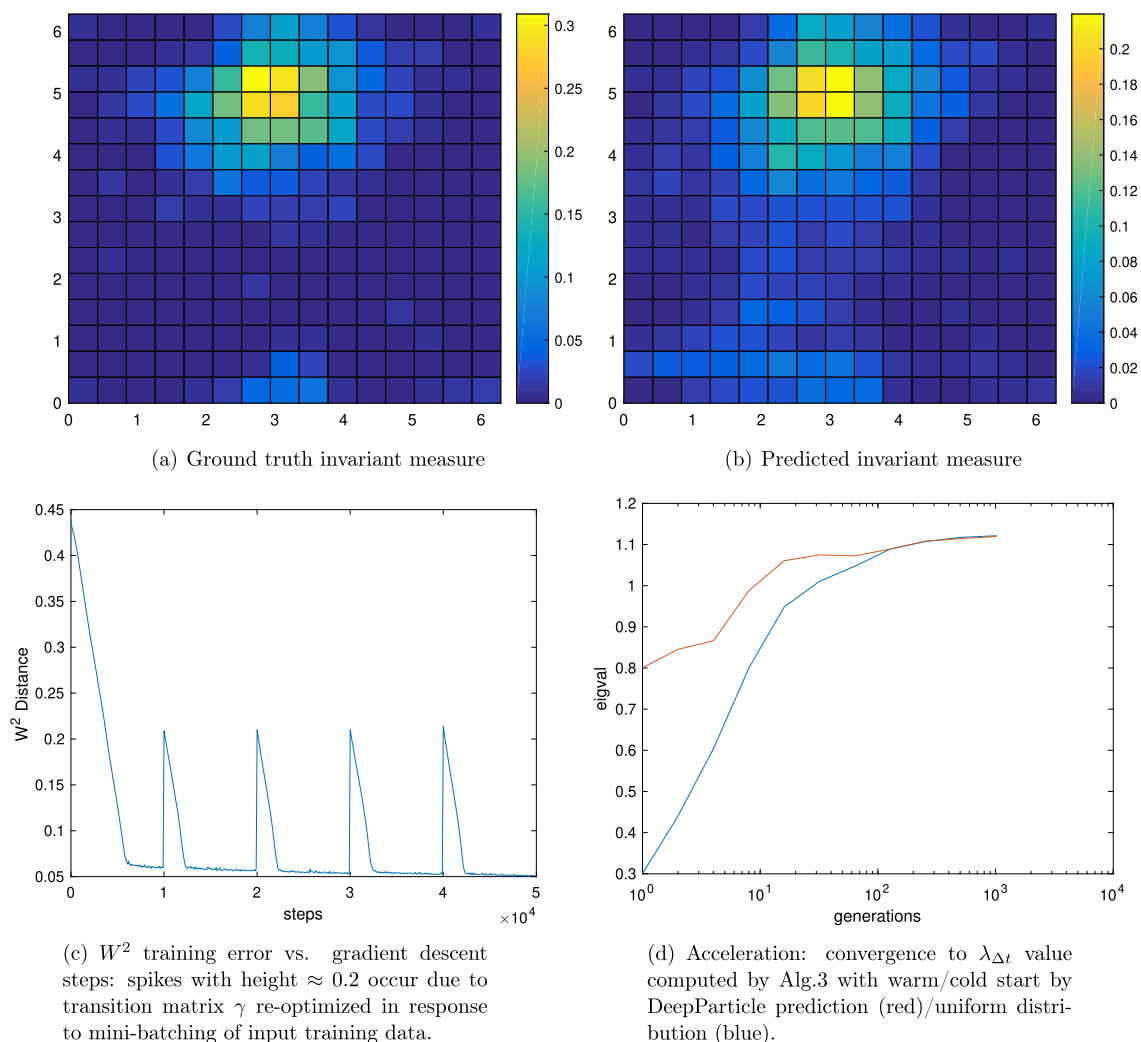
**Zhongjian Wang:** Conceptualization, Programming, Methodology, Writing – Original draft preparation. **Jack Xin:** Conceptualization, Methodology, Writing – Reviewing and Editing. **Zhiwen Zhang:** Conceptualization, Writing – Reviewing and Editing.

### 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 research of JX is partially supported by NSF grants DMS-1924548 and DMS-1952644. The research of ZZ is supported by Hong Kong RGC grant projects 17300318 and 17307921, National Natural Science Foundation of China No. 12171406,



**Fig. 7.** DeepParticle prediction (viewed in 2nd/3rd dim, top) at test value  $\kappa = 2^{-4}$  in a 3D computation. The  $W^2$  distance minimization in c) shows fast (steps before  $1e4$ ) and slow phases [38].

Seed Funding Programme for Basic Research (HKU), and Basic Research Programme (JCYJ20180307151603959) of The Science, Technology and Innovation Commission of Shenzhen Municipality.

## References

- [1] G. Bao, X. Ye, Y. Zang, H. Zhou, Numerical solution of inverse problems by weak adversarial networks, *Inverse Probl.* 36 (11) (2020) 115003.
- [2] Y. Bar-Sinai, S. Hoyer, J. Hickey, M. Brenner, Learning data-driven discretizations of PDEs, *Bull. Am. Phys. Soc.* 63 (2018).
- [3] M. Burger, L. Ruthotto, S. Osher, Connections between deep learning and partial differential equations, *Eur. J. Appl. Math.* 32 (3) (2021) 395–396.
- [4] Z. Cai, J. Chen, M. Liu, Least-squares ReLU neural network (LSNN) method for linear advection-reaction equation, *J. Comput. Phys.* (2021) 110514.
- [5] Z. Cai, J. Chen, M. Liu, X. Liu, Deep least-squares methods: an unsupervised learning-based numerical method for solving elliptic PDEs, *J. Comput. Phys.* 420 (2020) 109707.
- [6] R. Carmona, M. Lauriere, Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games: I – the ergodic case, arXiv:1907.05980, 13 July 2019.
- [7] R. Carmona, M. Lauriere, Convergence analysis of machine learning algorithms for the numerical solution of mean field control and games: II – the finite horizon case, arXiv:1908.01613, 5 August 2019.
- [8] W. Chen, Q. Wang, J. Hesthaven, C. Zhang, Physics-informed machine learning for reduced-order modeling of nonlinear problems, *J. Comput. Phys.* 446 (2021) 110666.
- [9] S. Childress, A. Gilbert, *Stretch, Twist, Fold: The Fast Dynamo*, Lecture Notes in Physics Monographs, vol. 37, Springer, 1995.
- [10] N. Cohen, O. Sharir, A. Shashua, On the expressive power of deep learning: a tensor analysis, in: *Conference on Learning Theory*, 2016, pp. 698–728.
- [11] M. Cuturi, Sinkhorn distances: lightspeed computation of optimal transport, *Adv. Neural Inf. Process. Syst.* 26 (2013) 2292–2300.
- [12] G. Cybenko, Approximation by superpositions of a sigmoidal function, *Math. Control Signals Syst.* 2 (4) (1989) 303–314.
- [13] W. E, J. Han, A. Jentzen, Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations, *Commun. Math. Stat.* 5 (4) (2017) 349–380.

- [14] W. E, B. Yu, The deep Ritz method: a deep learning-based numerical algorithm for solving variational problems, *Commun. Math. Stat.* 6 (1) (2018) 1–12.
- [15] S. Ellacott, Aspects of the numerical analysis of neural networks, *Acta Numer.* 3 (1994) 145–202.
- [16] M. Freidlin, *Functional Integration and Partial Differential Equations*, Princeton Univ. Press, 1985.
- [17] D. Galloway, M. Proctor, Numerical calculations of fast dynamos in smooth velocity fields with realistic diffusion, *Nature* 356 (6371) (1992) 691–693.
- [18] I. Goodfellow, Y. Bengio, A. Courville, Y. Bengio, *Deep Learning*, vol. 1, MIT Press, Cambridge, 2016.
- [19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, *Generative adversarial nets*, *Adv. Neural Inf. Process. Syst.* 27 (2014).
- [20] J. Han, A. Jentzen, W. E, Solving high-dimensional partial differential equations using deep learning, *Proc. Natl. Acad. Sci.* 115 (34) (2018) 8505–8510.
- [21] J. He, L. Li, J. Xu, C. Zheng, Relu deep neural networks and linear finite elements, *J. Comput. Math.* 38 (3) (2020) 502–527.
- [22] J. He, J. Xu, Mgnet: a unified framework of multigrid and convolutional neural network, *Sci. China Math.* 62 (7) (2019) 1331–1354.
- [23] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, *Neural Netw.* 2 (5) (1989) 359–366.
- [24] C. Kao, Y.Y. Liu, J. Xin, A semi-Lagrangian computation of front speeds of g-equation in ABC and Kolmogorov flows with estimation via ballistic orbits, *SIAM J. Multiscale Model. Simul.* 20 (1) (2022) 107–117.
- [25] S. Karumuri, R. Tripathy, I. Biliotis, J. Panchal, Simulator-free solution of high-dimensional stochastic elliptic partial differential equations using deep neural networks, *J. Comput. Phys.* 404 (2020) 109120.
- [26] Y. Khoo, J. Lu, L. Ying, Solving parametric PDE problems with artificial neural networks, in: *Special Issue 3: Connections Between Deep Learning and Partial Differential Equations*, *Eur. J. Appl. Math.* 32 (2021) 421–435.
- [27] I. Kobayzev, S. Prince, M. Brubaker, Normalizing flows: an introduction and review of current methods, *IEEE Trans. Pattern Anal. Mach. Intell.* (2020).
- [28] A. Kolmogorov, I. Petrovsky, N. Piskunov, Investigation of the equation of diffusion combined with increasing of the substance and its application to a biology problem, *Bull. Moscow State Univ. Ser. A: Math. Mech.* 1 (6) (1937) 1–25.
- [29] J.N. Kutz, Deep learning in fluid dynamics, *J. Fluid Mech.* 814 (1–4) (2017).
- [30] I. Lagaris, A. Likas, D. Fotiadis, Artificial neural networks for solving ordinary and partial differential equations, *IEEE Trans. Neural Netw.* 9 (5) (1998) 987–1000.
- [31] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (7553) (2015) 436.
- [32] H. Lee, Neural algorithm for solving differential equations, *J. Comput. Phys.* 91 (1990) 110–131.
- [33] W. Li, P. Yin, S. Osher, Computations of optimal transport distance with Fisher information regularization, *J. Sci. Comput.* 75 (3) (2018) 1581–1595.
- [34] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, A. Anandkumar, Fourier neural operator for parametric partial differential equations, *arXiv preprint, arXiv:2010.08895*, 2020.
- [35] A. Lin, S. Fung, W. Li, L. Nurbekyana, S. Osher, Alternating the population and control neural networks to solve high-dimensional stochastic mean-field games, *Proc. Natl. Acad. Sci.* 118 (31) (2021) e2024713118.
- [36] Z. Long, Y. Lu, B. Dong, PDE-Net 2.0: learning PDEs from data with a numeric-symbolic hybrid deep network, *J. Comput. Phys.* 399 (2019) 108925.
- [37] Z. Long, Y. Lu, X. Ma, B. Dong, PDE-Net: learning PDEs from data, in: *International Conference on Machine Learning*, 2018, pp. 3208–3216.
- [38] Z. Long, P. Yin, J. Xin, Global convergence and geometric characterization of slow to fast weight evolution in neural network training for classifying linearly non-separable data, *Inverse Probl. Imaging* 15 (1) (2021) 41–62.
- [39] J. Lu, Z. Shen, H. Yang, S. Zhang, Deep network approximation for smooth functions, *SIAM J. Math. Anal.* 53 (5) (2021) 5465–5506.
- [40] L. Lu, P. Jin, G. Karniadakis, Deepont: learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators, *arXiv:1910.03193*, 2019.
- [41] J. Lyu, Z. Wang, J. Xin, Z. Zhang, A convergent interacting particle method and computation of KPP front speeds in chaotic flows, *SIAM J. Numer. Anal.* (2022), <https://doi.org/10.1137/21M1410786>, in press.
- [42] J. Lyu, J. Xin, Y. Yu, Computing residual diffusivity by adaptive basis learning via spectral method, *Numer. Math., Theory Methods Appl.* 10 (2) (2017) 351–372.
- [43] J. Lyu, J. Xin, Y. Yu, Computing residual diffusivity by adaptive basis learning via super-resolution deep neural networks, in: H.A. Le Thi, et al. (Eds.), *Advanced Computational Methods for Knowledge Engineering, ICCSAMA 2019*, in: *Advances in Intelligent Systems and Computing*, vol. 1121, 2020, pp. 279–290.
- [44] A. Meade, A. Fernandez, The numerical solution of linear ordinary differential equations by feedforward neural networks, *Math. Comput. Model.* 19 (12) (1994) 1–25.
- [45] H. Montanelli, Q. Du, New error bounds for deep ReLU networks using sparse grids, *SIAM J. Math. Data Sci.* 1 (1) (2019) 78–92.
- [46] J. Nolen, M. Rudd, J. Xin, Existence of KPP fronts in spatially-temporally periodic advection and variational principle for propagation speeds, *Dyn. Partial Differ. Equ.* 2 (1) (2005) 1–24.
- [47] G. Peyré, M. Cuturi, Computational optimal transport, *Found. Trends Mach. Learn.* 11 (5–6) (2019) 355–607.
- [48] A. Pinkus, Approximation theory of the MLP model in neural networks, *Acta Numer.* 8 (1999) 143–195.
- [49] T. Qin, K. Wu, D. Xiu, Data driven governing equations approximation using deep neural networks, *J. Comput. Phys.* 395 (2019) 620–635.
- [50] M. Raissi, P. Perdikaris, G. Karniadakis, Multistep neural networks for data-driven discovery of nonlinear dynamical systems, *arXiv:1801.01236*, 2018.
- [51] M. Raissi, P. Perdikaris, G. Karniadakis, Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, *J. Comput. Phys.* 378 (2019) 686–707.
- [52] S. Rudy, J.N. Kutz, S. Brunton, Deep learning of dynamics and signal-noise decomposition with time-stepping constraints, *J. Comput. Phys.* 396 (2019) 483–506.
- [53] L. Ruthotto, S. Osher, W. Li, L. Nurbekyan, S. Fung, A machine learning framework for solving high-dimensional mean field game and mean field control problems, *Proc. Natl. Acad. Sci.* 117 (17) (2020) 9183–9193.
- [54] A. Schrijver, *Combinatorial Optimization: Polyhedra and Efficiency*, vol. 24, Springer Science & Business Media, 2003.
- [55] C. Schwab, J. Zech, Deep learning in high dimension, *Res. Rep.* 2017 (2017).
- [56] L. Shen, J. Xin, A. Zhou, Finite element computation of KPP front speeds in 3D cellular and ABC flows, *Math. Model. Nat. Phenom.* 8 (3) (2013) 182–197.
- [57] L. Shen, J. Xin, A. Zhou, Finite element computation of KPP front speeds in cellular and cat's eye flows, *J. Sci. Comput.* 55 (2) (2013) 455–470.
- [58] Z. Shen, H. Yang, S. Zhang, Deep network with approximation error being reciprocal of width to power of square root of depth, *Neural Comput.* 33 (4) (2021) 1005–1036.
- [59] R. Sinkhorn, A relationship between arbitrary positive matrices and doubly stochastic matrices, *Ann. Math. Stat.* 35 (2) (1964) 876–879.
- [60] J. Sirignano, K. Spiliopoulos, DGM: a deep learning algorithm for solving partial differential equations, *J. Comput. Phys.* 375 (2018) 1339–1364.
- [61] C. Villani, *Topics in Optimal Transportation*, vol. 58, American Math. Soc., 2021.
- [62] Q. Wang, N. Ripamonti, J. Hesthaven, Recurrent neural network closure of parametric POD-Galerkin reduced-order models based on the Mori-Zwanzig formalism, *J. Comput. Phys.* 410 (2020) 109402.
- [63] Y. Wang, S. Cheung, E. Chung, Y. Efendiev, M. Wang, Deep multiscale model learning, *J. Comput. Phys.* 406 (2020) 109071.
- [64] Z. Wang, Z. Zhang, A mesh-free method for interface problems using the deep learning approach, *J. Comput. Phys.* 400 (2020) 108963.
- [65] S. Wright, *Primal-Dual Interior-Point Methods*, SIAM, 1997.

- [66] K. Wu, D. Xiu, Data-driven deep learning of partial differential equations in modal space, *J. Comput. Phys.* (2020) 109307.
- [67] J. Xin, *An Introduction to Fronts in Random Media*, vol. 5, Springer Science & Business Media, 2009.
- [68] L. Yang, X. Meng, G.E. Karniadakis, B-PINNs: Bayesian physics-informed neural networks for forward and inverse PDE problems with noisy data, *J. Comput. Phys.* 425 (2021) 109913.
- [69] D. Yarotsky, Error bounds for approximations with deep ReLU networks, *Neural Netw.* 94 (2017) 103–114.
- [70] G. Yoo, H. Owhadi, Deep regularization and direct training of the inner layers of neural networks with kernel flows, arXiv:2002.08335, 2020.
- [71] Y. Zang, G. Bao, X. Ye, H. Zhou, Weak adversarial networks for high-dimensional partial differential equations, *J. Comput. Phys.* 411 (2020) 109409.
- [72] Y. Zhu, N. Zabaras, Bayesian deep convolutional encoder-decoder networks for surrogate modeling and uncertainty quantification, *J. Comput. Phys.* 366 (2018) 415–447.
- [73] Y. Zhu, N. Zabaras, P. Koutsourelakis, P. Perdikaris, Physics-constrained deep learning for high-dimensional surrogate modeling and uncertainty quantification without labeled data, *J. Comput. Phys.* 394 (2019) 56–81.
- [74] P. Zu, L. Chen, J. Xin, A computational study of residual KPP front speeds in time-periodic cellular flows in the small diffusion limit, *Physica D* 311 (2015) 37–44.