LIMITATIONS OF PHYSICS INFORMED MACHINE LEARNING FOR NONLINEAR TWO-PHASE TRANSPORT IN POROUS MEDIA

***Olga Fuks**[*] & **Hamdi A. Tchelepi**[*]*

Department of Energy Resources Engineering, Stanford University, 367 Panama Street, Stanford, California 94305, USA

*Address all correspondence to: Hamdi A. Tchelepi or Olga Fuks, Department of Energy Resources Engineering, Stanford University, 367 Panama Street, Stanford, California 94305, USA, E-mail: tchelepi@stanford.edu; E-mail: ofuks@stanford.edu

Original Manuscript Submitted: 2/11/2020; Final Draft Received: 6/10/2020

Deep learning techniques have recently been applied to a wide range of computational physics problems. In this paper, we focus on developing a physics-based approach that enables the neural network to learn the solution of a dynamic fluid-flow problem governed by a nonlinear partial differential equation (PDE). The main idea of physics informed machine learning (PIML) approaches is to encode the underlying physical law (i.e., the PDE) into the neural network as prior information. We investigate the applicability of the PIML approach to the forward problem of immiscible two-phase fluid transport in porous media, which is governed by a nonlinear first-order hyperbolic PDE subject to initial and boundary data. We employ the PIML strategy to solve this forward problem without any additional labeled data in the interior of the domain. Particularly, we are interested in nonconvex flux functions in the PDE, where the solution involves shocks and mixed waves (shocks and rarefactions). We have found that such a PIML approach fails to provide reasonable approximations to the solution in the presence of shocks in the saturation field. We investigated several architectures and experimented with a large number of neural-network parameters, and the overall finding is that PIML strategies that employ the nonlinear hyperbolic conservation equation in the loss function are inadequate. However, we have found that employing a parabolic form of the conservation equation, whereby a small amount of diffusion is added, the neural network is consistently able to learn accurate approximation of the solutions containing shocks and mixed waves.

KEY WORDS: *two-phase transport, physics informed machine learning, partial differential equations*

1. INTRODUCTION

Machine learning (ML) techniques, specifically deep learning (LeCun et al., 2015), are at the center of attention across the computational science and engineering communities. The spectrum of deep learning architectures and techniques has already achieved notable results across applications and disciplines, including computer vision and image recognition (He et al., 2016; Karpathy et al., 2014; Krizhevsky et al., 2012), speech recognition and machine translation (Hinton et al., 2012; Sutskever et al., 2014), robotics (Lillicrap et al., 2015; Mnih et al., 2016), and

2689–3967/20/$35.00

medicine (Gulshan et al., 2016; Liu et al., 2017). There is no doubt that the range of applications will grow and the impact of ML methods will continue to spread.

Deep learning allows neural networks composed of multiple processing layers to learn representations of raw input data with multiple levels of abstraction. These networks are known to be particularly effective at supervised learning tasks, whereby the successful application of these models usually requires the availability of large amounts of labeled data. However, in many engineering applications, data acquisition is often prohibitively expensive, and the amount of labeled data is usually quite sparse. Specifically, most computational geoscience problems related to modeling subsurface flow dynamics suffer from sparse site-specific data. Consequently, in this "sparse data" regime, it is crucial to employ domain knowledge to reduce the need for labeled training data, or even aim to train ML models without any labeled data relying only on constraints (Stewart and Ermon, 2017). These constraints are used to encode the specific structure and properties of the output that are known to hold because of domain knowledge, e.g., known physics laws such as conservation of momentum, mass, and energy.

Physics informed machine learning approaches have been explored recently in a variety of computational physics problems, whereby the focus is on enabling the neural network to learn the solutions of deterministic partial differential equations (PDEs). Early works in this area date back to the 1990s (Lagaris et al., 1998; Lee and Kang, 1990; Meade Jr. and Fernandez, 1994; Psichogios and Ungar, 1992). However, in the context of modern neural network architectures, the interest in this topic has been revived (Raissi et al., 2017, 2019; Zhu et al., 2019). These so-called physics informed machine learning (PIML) approaches are designed to obtain data-driven solutions of general nonlinear PDEs, and they may be a promising alternative to traditional numerical methods for solving PDEs, such as finite-difference and finite-volume methods. The core idea of PIML is that the developed neural network encodes the underlying physical law as prior information, and then uses this information during the training process. The approach takes advantage of the neural network capability to approximate any continuous function (Cybenko, 1989; Hornik et al., 1989). Raissi et al. (2017) demonstrated the PIML capabilities for a collection of diverse problems in computational science (Burgers' equation, Navier-Stokes, etc.). They suggested that if the considered PDE is well-posed and its solution is unique, then the PIML method is capable of achieving good predictive accuracy given a sufficiently expressive neural network architecture and a sufficient number of collocation points. In the current work, we show that the neural network approach struggles and even fails for modeling the nonlinear hyperbolic PDE that governs two-phase transport in porous media. Our experience indicates that this shortcoming of PIML for hyperbolic PDEs is not related to the specific architecture, or to the choice of the hyperparameters (e.g., number of collocation points, etc.).

One important class of PDEs is that of conservation laws that describe the conservation of mass, momentum, and energy. In particular, these conservation equations describe displacement processes that are essential for modeling flow and transport in subsurface porous formations, such as water-oil or gas-oil displacements (Aziz and Settari, 1979; Orr, 2007). Numerical reservoir simulation based on solving mass conservation equations with constitutive relations for the nonlinear coefficients is used to make predictions. A major challenge in practice is that the available information/measurements (i.e., labeled data) about the specific geological formation of interest is often quite sparse. Thus, it is critical to take advantage of any prior information in order to improve the predictive reliability of the computational models. The physics of two-phase fluid transport, e.g., water-oil displacements, is described by a nonlinear hyperbolic PDE [or a system of PDEs (Orr, 2007)]. These nonlinear transport problems are known to be quite challenging for standard numerical methods (Aziz and Settari, 1979), and this is largely due

to the presence of steep saturation fronts and mixed waves (shocks and spreading waves) in the solution. Specifically, we are interested in solving Riemann problems—initial value problems, when the initial data consist of two constant states separated by a jump discontinuity at $x = 0$.

There are significant efforts aimed at figuring out the potential of machine learning in the modeling of flow processes in large-scale subsurface formations. Thus, it is extremely important to understand the limitations of PIML schemes for making computational predictions of reservoir displacement processes. Here, we investigate the application of the physics informed machine learning approach to the "pure" forward problem of nonlinear two-phase transport in porous media. We evaluate the performance of the PIML framework for this problem with different flux (fractional flow) functions. The objective is to assess how well this PIML approach performs for nonlinear flow problems with discontinuous solutions (i.e., shocks).

The paper proceeds as follows. In Section 2, we describe the two-phase transport model and the governing hyperbolic PDE that we aim to solve with a machine learning approach. In Section 3 we provide a brief overview of the physics informed machine learning framework that we use to solve the deterministic PDE. The results for the transport problem with different flux functions are presented in Section 4. Then, to understand the observed behavior of the method we provide a more detailed analysis of the trained neural networks in Section 5. Lastly, in Section 6, we summarize our findings and provide a brief discussion of the results.

2. TWO-PHASE TRANSPORT MODEL

We consider the standard Buckley-Leverett model with two incompressible immiscible fluids, e.g., oil and water. A nonwetting phase, e.g., oil (o), is displaced by a wetting phase, e.g., water (w), in a porous medium with permeability $k(\mathbf{x})$ and porosity $\phi(\mathbf{x})$. Gravity and capillary effects are neglected. Under these assumptions, the pressure p and fluid saturations S_α ($\alpha = o, w$) are governed by a coupled system of mass balance equations complemented by Darcy's equations for each phase. After some manipulation [see, e.g., Aziz and Settari (1979)], the system can be transformed into the incompressibility condition for the total flux, $\mathbf{u}_{tot}$:

$$\nabla \cdot \mathbf{u}_{tot} = q_t, \tag{1}$$

where q_t is a total source (sink) term, and the conservation equation for one of the phases, e.g., water:

$$\phi(\mathbf{x})\frac{\partial S_w}{\partial t} + \nabla \cdot (f_w(S_w) \cdot \mathbf{u}_{tot}) = q_w. \tag{2}$$

Here $\mathbf{u}_{tot} = \mathbf{u}_w + \mathbf{u}_o$ is the total flux and $\mathbf{u}_\alpha$ represents the Darcy's flux for a phase ($\alpha = \mathrm{o}, \mathrm{w}$); the function f_w is called the fractional flow of water or simply, flux function, and is defined as follows:

$$f_w = \frac{\lambda_w}{\lambda_w + \lambda_o}, \tag{3}$$

where $\lambda_\alpha = (k\, k_{r\alpha})/\mu_\alpha$ stands for the phase mobility, μ_α is the viscosity of the phase, $k_{r\alpha}(S_\alpha)$ is the relative phase permeability, and q_w is a source (sink) term for water. The source or sink terms represent the effect of wells. Equation (2) is supplemented with uniform initial and boundary conditions:

$$\begin{aligned} S_w(\mathbf{x}, t) &= s_{wi}, \quad \forall \mathbf{x} \quad \text{and} \quad t = 0, \\ S_w(\mathbf{x}, t) &= s_b, \quad \mathbf{x} \in \Gamma_{inj} \quad \text{and} \quad t > 0, \end{aligned} \tag{4}$$

where s_{wi} is the initial water saturation in the reservoir, and s_b is the saturation at the injection well or boundary, Γ_{inj}.

In one-dimensional space, Eq. (2) becomes

$$\phi(x)\frac{\partial S_w}{\partial t} + u_{tot}\frac{\partial f_w(S_w)}{\partial x} = 0, \tag{5}$$

and the total velocity u_{tot} is constant. After introducing the dimensionless variables $t_D = \int_0^t [(u_{tot}dt')/\phi L]$ and $x_D = x/L$, where L is the length of the one-dimensional system, we can rewrite Eq. (5) as follows:

$$\frac{\partial S_w}{\partial t_D} + \frac{\partial f_w(S_w)}{\partial x_D} = 0, \tag{6}$$

while initial and boundary conditions can be written as:

$$\begin{aligned} &S_w(x_D, 0) = s_{wi}, \quad \forall x_D \\ &S_w(x_D, t_D) = s_b, \quad x_D = 0 \quad \text{and} \quad t_D > 0. \end{aligned} \tag{7}$$

Solving this initial value problem is equivalent to solving the following nonlinear hyperbolic PDE:

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \tag{8}$$

with the piecewise constant initial condition,

$$u(t = 0, x) = u^0(x). \tag{9}$$

Here, $u(t, x)$ is the space-time dependent quantity of interest (conserved scalar) that needs to be solved for, and $f(u)$ is the flux function. The PDE (8) can be solved by the method of characteristics, and it can be shown that the characteristics are straight lines [see e.g., (Lax, 1973)]. If the initial data (9) are piecewise constant having a single discontinuity, i.e., a Riemann problem, the PDE solution is a self-similar function. The hyperbolic PDE of the general form (8) is the main subject of the current work, and in the following we solve the initial value problem, Eqs. (8) and (9), by applying the physics informed machine learning (PIML) approach.

3. PHYSICS INFORMED MACHINE LEARNING

In this section we consider the following general partial differential equation:

$$u_t + \mathcal{N}(u) = 0, \tag{10}$$

where $\mathcal{N}(\cdot)$ is a nonlinear differential operator.

Neural networks are often regarded as universal function approximators (Cybenko, 1989; Hornik et al., 1989)—which means that a feed-forward network with a single hidden layer containing a finite number of neurons can approximate any continuous function to any desired level of precision. Following the approach of Raissi et al. (2019), the solution $u(t, x)$ to the PDE is approximated by a deep neural network parameterized by a set of parameters θ. In other words,

the solution to the PDE is represented as a series of function compositions:

$$\begin{aligned} y_1(t,x) &= \sigma(W_1 X + b_1), \\ y_2(t,x) &= \sigma(W_2 y_1 + b_2), \\ &\dots \\ y_{n_l+1}(t,x) &= W_{n_l+1} y_{n_l} + b_{n_l+1}, \\ u_\theta(t,x) &= y_{n_l+1}(t,x), \end{aligned} \tag{11}$$

where the input vector X contains space and time coordinates; i.e., $X = (x,t)$, θ is the ensemble of all the model parameters:

$$\theta = \{W_1, W_2, \dots, W_{n_l+1}, b_1, b_2, \dots, b_{n_l+1}\}, \tag{12}$$

σ is an activation function (tanh in our case) and n_l is the number of hidden layers. Defining $z_i(x) = \sigma(W_i x + b_i)$ for $i = 1, ..., n_l$ and $z_i(x) = W_i x + b_i$ for $i = n_l + 1$, we can write the solution to the PDE as follows:

$$u_\theta(t,x) = z_{n_l+1}(z_{n_l}(\dots z_2(z_1(X)))). \tag{13}$$

The residual of the PDE is just the left-hand side of Eq. (10):

$$r(t,x) = u_t + \mathcal{N}(u). \tag{14}$$

When the PDE solution is approximated by a neural network $u_\theta(t,x)$, the residual of the PDE can be also represented as the neural network with the same parameters θ:

$$r_\theta(t,x) = (u_\theta)_t + \mathcal{N}(u_\theta). \tag{15}$$

This network $r_\theta(t,x)$ can be easily derived by applying automatic differentiation to the network $u_\theta(t,x)$. Then, the shared parameters θ are learned by minimizing the following loss function:

$$\begin{aligned} L(\theta) &= L_u(\theta) + L_r(\theta), \\ L_u(\theta) &= \frac{1}{N_u}\sum_{i=1}^{N_u} |u_\theta(t_u^i, x_u^i) - u_{bc}^i|^2, \\ L_r(\theta) &= \frac{1}{N_r}\sum_{i=1}^{N_r} |r_\theta(t_r^i, x_r^i)|^2, \end{aligned} \tag{16}$$

where $\{(t_u^i, x_u^i), u_{bc}^i\}_{i=1}^{N_u}$ represent the training data on initial and boundary conditions, and $\{t_r^i, x_r^i\}_{i=1}^{N_r}$ denote the collocation points for the PDE residual, $r(t,x)$, sampled randomly throughout the domain of interest. Thus, the loss function consists of two terms: one is the mean squared error coming from the initial and boundary conditions, and the other is the mean squared error from the residual evaluated at collocation points inside the physical domain.

4. NUMERICAL RESULTS

In our examples, we consider the nonlinear hyperbolic transport equation of the form

$$u_t + (f_w)_x = 0, \tag{17}$$

where $f_w = f_w(u)$ is the fractional flow function, i.e., flux function, and $x \in [0,1], t \in [0,1]$. The unknown solution u corresponds to water saturation, S_w, in Eq. (6). Different flux functions produce different types of waves in the solution. In addition, we assume the following uniform initial and boundary conditions:

$$\begin{aligned} u(x,t) &= 0, \quad \forall x \quad \text{and} \quad t = 0, \\ u(x,t) &= 1, \quad x = 0 \quad \text{and} \quad t > 0. \end{aligned} \tag{18}$$

This setting corresponds to the injection of water at one end of the oil-filled 1D reservoir, e.g., rock core, and the following parameters: $s_{wi} = 0$, $s_b = 1$. The conservation law (17) with initial and boundary conditions (18) forms a Riemann problem that has a self-similar solution, i.e., $u(x,t) = u(x/t)$.

In the numerical examples, we use the fully connected neural network architecture reported in Raissi et al. (2019) that consists of eight hidden layers with 20 neurons per hidden layer. The hyperbolic tangent activation function is used in all hidden layers. All weights are initialized randomly according to the Xavier initialization scheme (Glorot and Bengio, 2010). The loss function is optimized with a second-order quasi-Newton method, L-BFGS-B (Nocedal and Wright, 2006). For the training data in all examples we use $N_u = 300$ randomly distributed points on initial and boundary conditions, and $N_r = 10{,}000$ collocation points for the residual term, sampled randomly over the interior of the domain $x \in [0,1], t \in [0,1]$. Next, we consider different flux functions $f_w(u)$ in Eq. (17).

4.1 Concave Flux Function

If the relative phase permeabilities, $k_{r\alpha}(S_\alpha)$, are linear functions of saturation, and the ratio of the phase viscosities is denoted as $\mu_o/\mu_w = M$, the corresponding flux function, f_w, can be written as

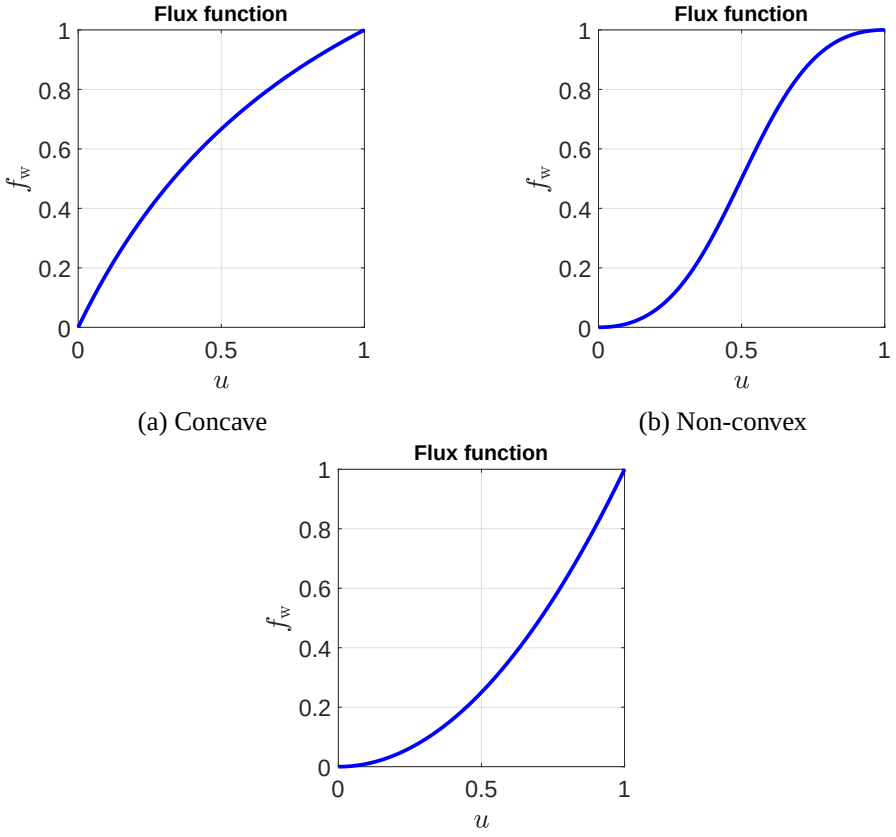
$$f_w(u) = \frac{u}{u + \dfrac{1-u}{M}}. \tag{19}$$

For $M > 1$, this flux function is concave, as shown in Fig. 1(a) for $M = 2$. The solution of the Eq. (17) for the given initial and boundary conditions (18) and the flux function (19) is a rarefaction (spreading) wave:

$$u(x,t) = \begin{cases} 0, & \dfrac{x}{t} > M \\ \dfrac{\sqrt{M\dfrac{t}{x}} - 1}{M-1}, & M \geq \dfrac{x}{t} \geq \dfrac{1}{M} \\ 1, & \dfrac{1}{M} \geq \dfrac{x}{t} \end{cases}.$$

We consider the case $M = 2$. Due to the piecewise nature of the analytical solution, there are certain locations (specifically, those along the lines $x/t = M$ and $x/t = 1/M$), where the solution is non-differentiable as derivatives of the solution are different on both sides.

However, this does not prevent the deep learning approach from learning the solution. Figure 2 presents a comparison of the exact analytical solution and the solution predicted by neural network at time instances $t = 0.25, 0.5, 0.75$. In this case, the neural network produces accurate



(a) Concave (b) Non-convex

(c) Convex

FIG. 1: Different flux functions

estimates of the PDE solution with some smoothing of the non-differentiable edges of the solution. The final loss at the end of training is $L(\theta) = 1.2 \times 10^{-3}$ and the resulting relative $\mathcal{L}^2$ norm of the prediction error of the solution (compared to the analytical solution) is 2.6×10^{-2}.

4.2 Non-Convex Flux Function

In most practical settings, the interaction between two immiscible fluids flowing through the porous medium leads to highly nonlinear relative permeabilities. A simple model that captures this characteristic is the Brooks-Corey model (Brooks and Corey, 1964), which gives the power-law relationship between the relative permeability of a fluid phase and its saturation. Specifically, we use a quadratic relationship, which leads to the following flux, i.e., fractional flow, function:

$$f_w(u) = \frac{u^2}{u^2 + \frac{(1-u)^2}{M}}, \tag{20}$$

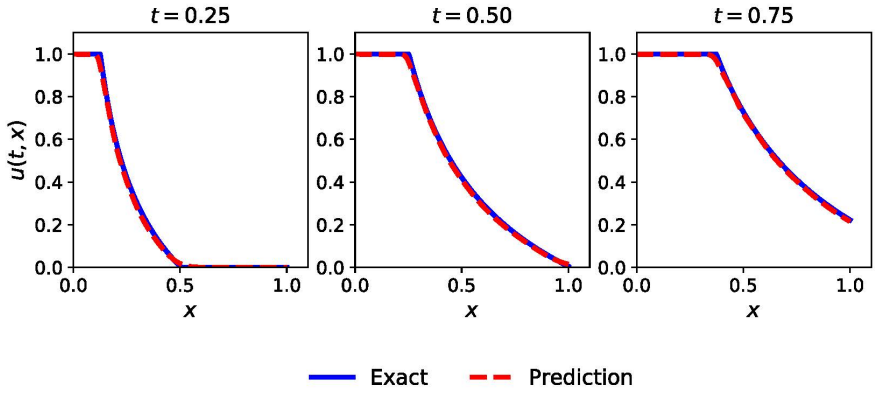


FIG. 2: Comparison of the predicted by the neural network and the exact solutions of the PDE (17) with concave flux function (19), corresponding to the three different times $t = 0.25, 0.5, 0.75$

where again M is the ratio of phase viscosities. The PDE (17) with this non-convex flux function constitutes a standard Buckley-Leverett problem in porous media flow. In our example we use $M = 1$ and the corresponding flux function is depicted in Fig. 1(b). We proceed by considering two cases with this flux function—with and without an additional diffusion term in the PDE.

4.2.1 Without Diffusion Term

In this case, the residual term (17), representing the hyperbolic PDE, is used directly in the loss function. The analytical solution to this problem contains a shock and a rarefaction wave and is constructed as follows:

$$u(x,t) = \begin{cases} 0, & \frac{x}{t} > f'_w(u^*) \\ u\left(\frac{x}{t}\right), & f'_w(u^*) \geq \frac{x}{t} \geq f'_w(u=1) \\ 1, & f'_w(u=1) \geq \frac{x}{t} \end{cases}, \tag{21}$$

where u^* denotes the shock location, which is defined by the Rankine-Hugoniot condition $f'_w(u^*) = [f_w(u^*) - f_w(u)_{|u=0}]/(u^* - u_{|u=0})$, and $u(x/t)$ is defined for $x/t \leq f'_w(u^*)$ as $u(x/t) = (f'_w)^{-1}(x/t)$. Due to the self-similarity, the analytical solution (21) has just one governing parameter—the similarity variable x/t.

Figure 3 shows that the neural network fails in this case to provide an accurate approximation of the underlying analytical solution (21). In fact, the neural network completely misses the correct location of the saturation front, which leads to high values of the loss [at the end of training it is $L(\theta) = 0.036$] and large prediction errors. In our numerical experiments, we observed that changing the neural network architecture and/or increasing the number of collocation points had little impact on the results (details of these studies are provided in Appendix A). Thus, we think this phenomenon is not related to the choice of the network architecture or its hyperparameters.

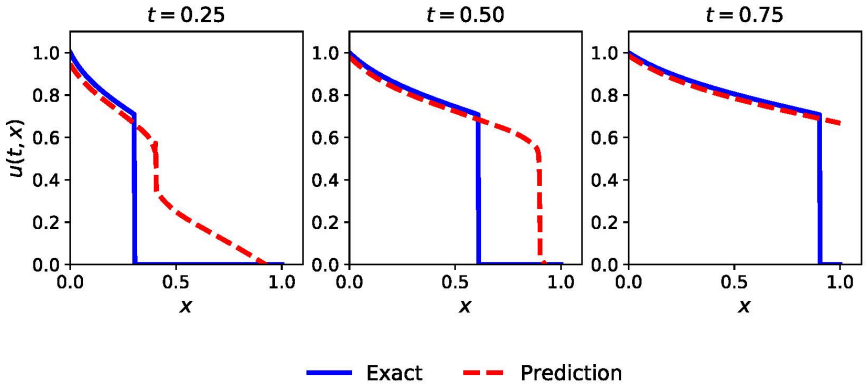


FIG. 3: Comparison of the predicted by the neural network and the exact solutions of the PDE (17) with non-convex flux function (20), corresponding to the three different times $t = 0.25, 0.5, 0.75$

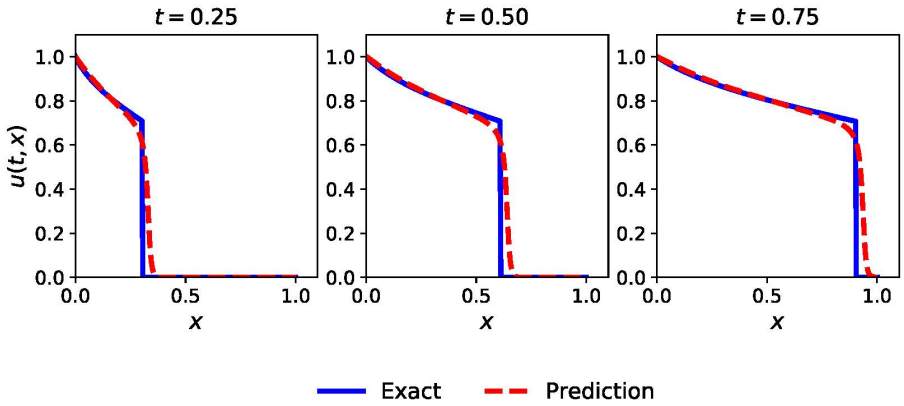
4.2.2 With Diffusion Term

The vanishing viscosity method for solving the initial value problems for hyperbolic PDEs (Crandall and Lions, 1983; Lax, 2006) is based on the fact that solutions of the inviscid equations, e.g., Eq. (17), including solutions with shocks, are the limits of the solutions of the viscous equations as the coefficient of viscosity tends to zero. Motivated by this approach, we add a second-order term, i.e., a diffusion term, to the right-hand side of Eq. (17) and consider the following equation:

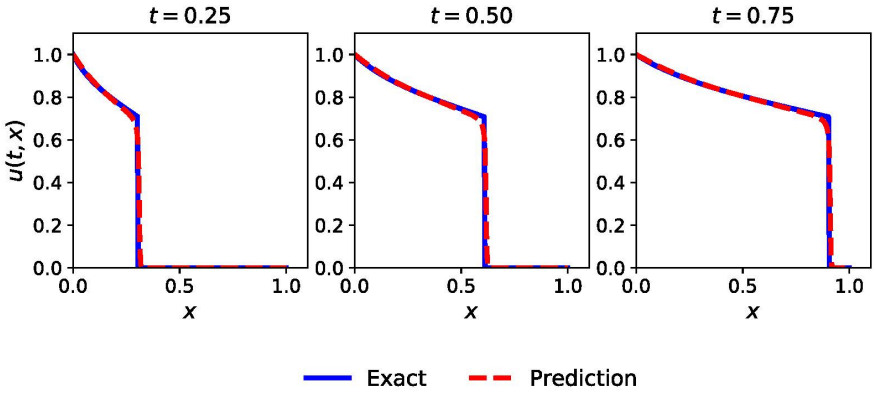
$$u_t + f'_w(u)u_x = \epsilon u_{xx}, \tag{22}$$

where $\epsilon > 0$ is a scalar diffusion coefficient that represents the inverse of the Péclet number, Pe—the ratio of a characteristic time for dispersion to a characteristic time for convection. When ϵ is small, i.e., the Péclet number is large, the effects of diffusion are negligible and convection dominates. Letting $\epsilon \to 0$ in Eq. (22) defines a vanishing diffusion solution of Eq. (17), which is the one with the correct physical behavior. Also, it should be noted that Eq. (22) is now a parabolic PDE, so its solution is smooth, i.e., it does not contain shocks.

Figure 4 shows neural network solutions for two different values of diffusion coefficient ϵ: 1×10^{-2} (Pe $= 100$) and 2.5×10^{-3} (Pe $= 400$). The loss values at the end of the training are $L(\theta) = 3.2 \times 10^{-6}$ and 2.4×10^{-5}, respectively. Note that the loss function is different in these two cases as the loss depends on the PDE residual, which is a function of ϵ according to Eq. (22). From these results, we see that adding a diffusion term to the conservation equation allows the neural network to perfectly capture the location of the saturation front even for quite small ϵ. Indeed, the solution in Fig. 4(b) for $\epsilon = 2.5 \times 10^{-3}$ is almost indistinguishable from the underlying analytical PDE solution—there is just a slight smoothing of the shock. In our numerical experiments, we also observed that if we continue to decrease the value of diffusion coefficient ϵ, e.g., $\epsilon = 1 \times 10^{-3}$, then the diffusion effects become too small, and the behavior of the neural network is the same as in the hyperbolic setting (i.e., zero diffusion) described in Section 4.2.1. It should be noted that the experiments in the current section—both for PDEs with and without the diffusion term—were all performed multiple times with different random seeds and random initializations; however, the results in terms of recovering the shock were equivalent.



(a) $\epsilon = 1 \times 10^{-2}$



(b) $\epsilon = 2.5 \times 10^{-3}$

FIG. 4: Predictions of the neural network for the PDE (22) for different values of diffusion coefficient ϵ. Exact solution corresponds to the PDE (17) without diffusion term.

Then, we conducted similar experiments for other values of phase viscosity ratio M, such as $M = 0.5, 5, 10$, that are also common in the subsurface transport domain. Under these settings the solutions differ in size and speed of the shock, i.e., for larger M the shock size decreases but its speed increases. However, the solution structure stays exactly the same—the solution still consists of a shock followed by a rarefaction wave. We considered also two cases for each value of the phase viscosity ratio M—with and without the diffusion term. The results of these tests and conclusions were the same as for $M = 1$ described above; thus we conclude that the observed behavior of the PIML approach is not sensitive to the value of parameter M.

It is worth mentioning that the obtained results are consistent with the previously reported results of the PIML approach in Raissi et al. (2017). The authors of Raissi et al. (2017) studied Burgers' equation with the diffusion term (so the shock was smoothed) and the diffusion coefficient (ϵ in our notation) was equal to $\epsilon = 0.01/\pi \approx 3.2 \times 10^{-3}$. However, if one applies the PIML approach for the same settings of Burgers' equation as in Raissi et al. (2017) but decreases

the diffusion coefficient to 0.5×10^{-3} or less (or sets it to zero altogether), then the network fails in a similar way as was described in Section 4.2.1.

4.3 Convex Flux Function

Now, we move to the convex flux function, shown in Fig. 1(c), which is simply a quadratic function $f_w(u) = u^2$. The solution is a self-sharpening wave, propagating as shock with a unit speed.

The prediction of the neural network for $t = 0.5$ in the case of hyperbolic PDE (17) is shown in the left plot of Fig. 5. As in the case of the non-convex flux function, the PIML approach fails for this problem. And similar to the non-convex flux case, adding a small diffusion term, e.g., with $\epsilon = 2.5 \times 10^{-3}$, to the PDE allows the neural network to reconstruct the solution and determine the location of the (smoothed) shock correctly (Fig. 5, on the right).

5. ANALYSIS

It is quite surprising that the neural network with several thousands of parameters is not able to yield a reasonable approximation to the analytical solution of the 1D hyperbolic PDE (17) with a non-convex flux function (20)—the solution that can be represented using a relatively simple piecewise continuous function of one parameter (21). This is surprising, especially because according to the universal approximation theorem (Cybenko, 1989) there should exist a network that can provide a close approximation of the continuous solution of (22) for any arbitrarily small ϵ (because the solution is smooth in this case); however, this is not what is observed in practice. Thus, this leads us to the conclusion that the problem is not with the solution itself, but rather with *how* we attempt to find this solution, i.e., with the optimization process, or the loss function.

For the examples described above, we provide the analysis of the obtained neural networks. Our aim here is to get a better understanding of the observed behavior of the neural network approach—why it can find a solution to the problem with the additional diffusion term, i.e., the parabolic form of the PDE, but fails to do so in the case of the underlying hyperbolic PDE, i.e., when its solution contains a discontinuity. Is this due to some fundamental reasons that prevent the neural network from finding a reasonable approximate solution (non-uniqueness of

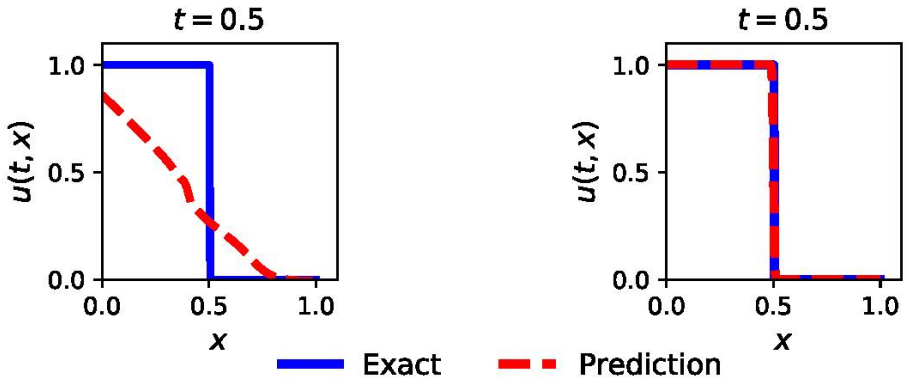


FIG. 5: Predictions of the neural network at $t = 0.5$ for the case of convex flux function: on the left the prediction for the PDE without diffusion term, on the right – with added diffusion term, as in Eq. (22), and diffusion coefficient $\epsilon = 2.5 \times 10^{-3}$. Exact solutions in both cases are shown for the PDE (17) without diffusion term.

the solution of the weak form), or is it because the employed optimization algorithm just cannot reach the solution? The latter can be due to the complicated nature of the non-convex landscape of the loss function, or other inherent limitations of the optimization algorithm.

First, we investigate the training process and study the behavior of the loss and its gradients with respect to the network parameters. Then, through 2D visualizations of the loss surface, we study how the diffusion term affects the loss landscape and the convexity of the loss near the final optimization point, i.e., optimized set of network parameters.

5.1 Training Process

Figure 6 shows the evolution of the loss function during the training process for models with different amounts of diffusion, i.e., different values of the diffusion coefficient ϵ before the second-order term in Eq. (22). The x-axis in the figure denotes the steps of the L-BFGS-B optimization method. Note that the loss function being minimized is different for each model, as part of the loss, corresponding to the residual term, is directly proportional to ϵ. In Fig. 6 we observe a clear trend. For larger values of ϵ the convergence rate of the optimization improves significantly, i.e., the loss is minimized in far fewer steps. On the other hand, for smaller values (i.e., $\epsilon = 0$ or 1×10^{-3}) the corresponding loss curve flattens out quite early during the training, and the optimization method fails to minimize the loss (the final loss is only of order 10^{-2}).

The training of the neural network can also be studied by observing the gradient of the loss with respect to the different parameters of the network, i.e., weights and biases of different layers. Figure 7 shows the $\mathcal{L}^2$ norm of the loss gradient with respect to the weights in the first layer versus the number of optimization steps (some curves were smoothed for better visualization). The curves for the models that achieve good approximation accuracy of the solution, i.e., the models with $\epsilon = 5 \times 10^{-2}$, 5×10^{-3} and 2.5×10^{-3}, show a steady decrease in the norm

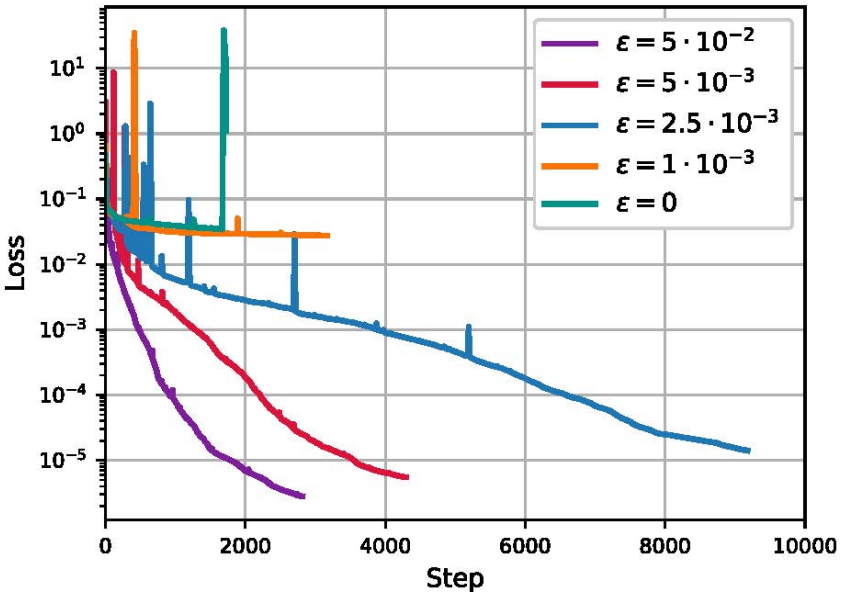


FIG. 6: The loss function during training for models with different amount of added diffusion according to Eq. (22). The x-axis denotes the steps of the L-BFGS-B optimization method.

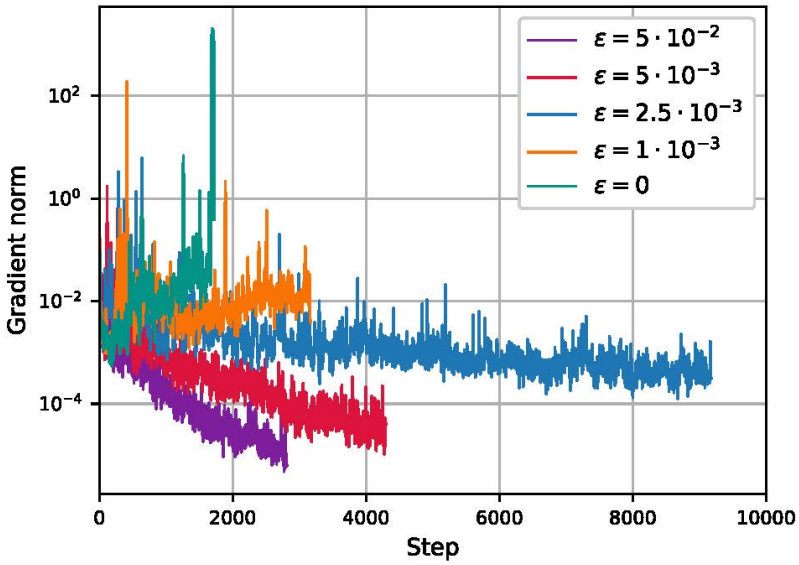


FIG. 7: The evolution of the $\mathcal{L}^2$ norm of the loss gradient with respect to the weights of the first network layer during training. The x-axis denotes the steps of the L-BFGS-B optimization method.

of the gradient during training, indicating convergence of the optimization process; on the other hand, for the models that have large prediction errors, i.e., for $\epsilon = 0$ and $\epsilon = 1 \times 10^{-3}$, the gradients do not decrease with time, and sometimes even increase, indicating failure of the optimization process. The loss gradients with respect to the parameters in other layers of the network showed similar trends. Again, from the results shown in Fig. 7 it is obvious that the magnitude of ϵ significantly affects the behavior of the loss gradients. This behavior for $\epsilon \sim 0$ may be explained with the complicated objective function landscape, so that the quasi-Newton method fails to minimize the loss. It may also be due to the poor conditioning of the Hessian of the loss, so that the desired solution lies in a very local and narrow region. Nevertheless, it is clear that the presence of the second-order term u_{xx}, i.e., presence of diffusion in the PDE, and the amount of diffusion strongly influence the training process of the physics informed network and its ability to yield accurate approximations of the solution.

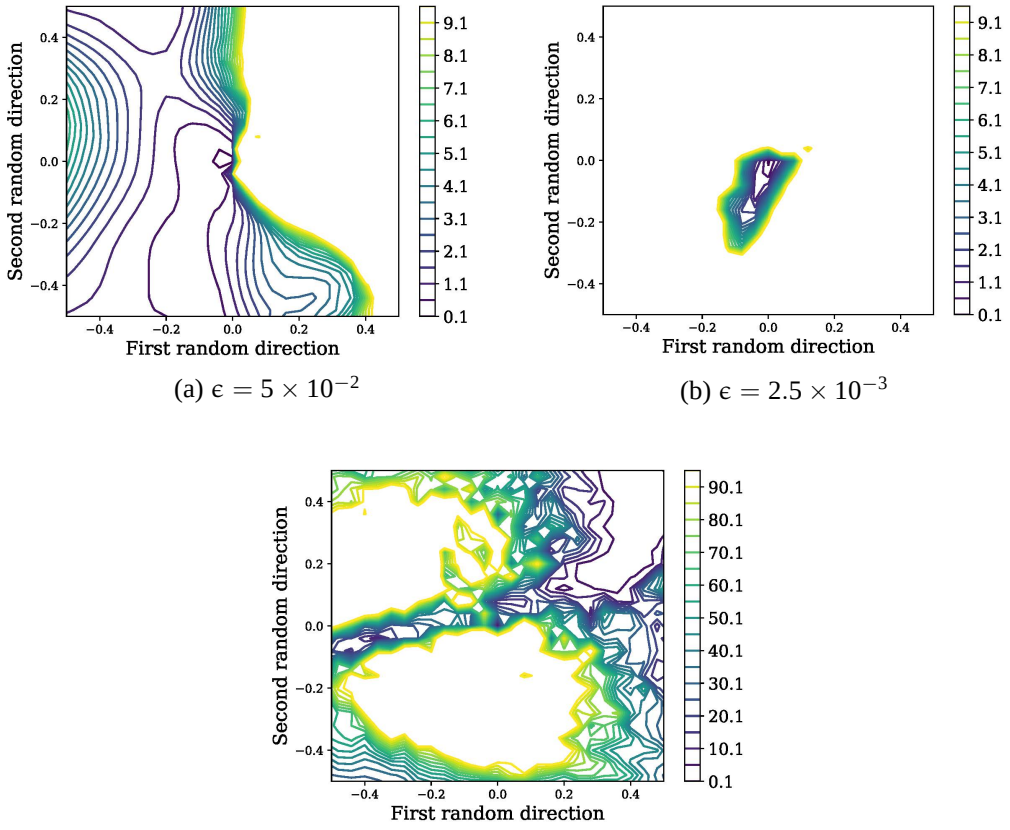
5.2 Loss Landscape

To visualize the surface of the loss, which is a function in the high-dimensional parameter space, one must restrict the space to a low-dimensional one (1D or 2D), amenable to visualization. Here, we choose to follow the approach of Li et al. (2018), whereby to get a 2D projection of the loss surface we choose a center point θ, corresponding to the final optimization point (i.e., final parameters of the model reshaped into a single vector) and two direction vectors, δ and η, of the same dimension as θ. Then, we can plot the following function:

$$f(\alpha, \beta) = L(\theta + \alpha\delta + \beta\eta), \tag{23}$$

where α and β are scalar parameters along vectors δ and η, respectively. The direction vectors are sampled randomly from Gaussian distribution—in the high-dimensional space these vectors with a high probability will be almost orthogonal to each other. Then, Li et al. (2018) suggest "filter-wise" normalizing the random directions to capture the natural distance scale of the loss surface. This step ensures that elements in random vectors, δ and η, are of the same scale as the corresponding parameters of the network, i.e., weights and biases of different network layers.

For visualizations we vary both scalar parameters, α and β, in the range $(-0.5, 0.5)$. Figure 8 shows the loss surface plots for different networks near their final optimization point, i.e., set of optimized parameters. This point corresponds to $(0, 0)$ in the surface plots, and the two axes represent the two random directions, respectively. The results are shown as contour plots to make it easier to see the non-convex structures of the loss landscape. The networks differ in the amount of the added diffusion, i.e., value of diffusion coefficient ϵ. For large diffusion, for example, $\epsilon = 5 \times 10^{-2}$, in Fig. 8(a), we observe quite a large convex region, whereas for a small amount of diffusion, e.g., $\epsilon = 2.5 \times 10^{-3}$, this region shrinks significantly, as shown in



(a) $\epsilon = 5 \times 10^{-2}$

(b) $\epsilon = 2.5 \times 10^{-3}$

(c) $\epsilon = 0$

FIG. 8: 2D visualizations of the loss surface near the final optimization point for neural networks trained with different values of diffusion coefficient ϵ in Eq. (22). Note the change of scale for $\epsilon = 0$.

Fig. 8(b). Note the change of scale in Fig. 8(c), which depicts the loss surface for the hyperbolic PDE, i.e., $\epsilon = 0$,—for proper visualization the scale of the loss had to be increased by 10 times compared to the cases with diffusion. No convex region is observed in this instance. Moreover, the loss landscape is not as smooth as with the diffusion present—indeed, it has a lot of chaotic features, as can be seen in Fig. 8(c). For visualization of the same loss surface on a larger slice of the parameter space, refer to Fig. 9. From these observations, we can conclude that the presence of the discontinuity, i.e., the shock, in the PDE solution strongly affects the properties of the resulting landscape of the corresponding loss function – specifically, its smoothness and convexity. It is not surprising that the optimization procedure struggles with this loss landscape and is unable to reach the proper solution, i.e., the one that gives a close continuous approximation of the discontinuous PDE solution (21). For comparison, we also show in Fig. 10 the loss surface of the network approximating a smooth PDE solution in case of concave flux function (19). The wide convex region of the loss surface is evident here.

6. DISCUSSION AND CONCLUSION

We investigated the application of a physics informed machine learning (PIML) approach to the solution of one-dimensional hyperbolic PDEs that describe the nonlinear two-phase transport in porous media. The PIML approach encodes the underlying PDE into the loss function and learns the solution to the PDE without any labeled data—only using the knowledge of the initial/boundary conditions and the PDE. Our experiments with different flux functions demonstrate that the neural network approach provides accurate estimates of the solution of the hyperbolic PDE when the solution does not contain discontinuities. However, the PIML approach fails to provide reasonable approximate solution of the PDE when shocks are present. We found that it is necessary to add a diffusion term to the underlying PDE, so that the network can recover the proper location and size of the shock, which is smoothed by diffusion. Thus, the network actually solves the parabolic form of the conservation equation, which leads to the correct solution with smoothing around the shock. It is interesting to note here the resemblance of this effect with finite-volume methods, whereby the conservative finite-volume discretization adds a

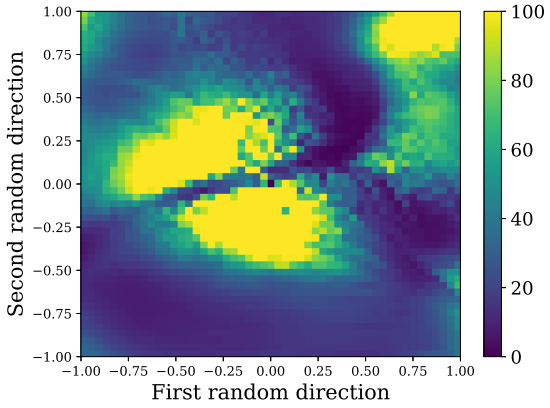


FIG. 9: 2D visualization of the loss surface of the neural network for the hyperbolic PDE (17) with non-convex flux function (20)

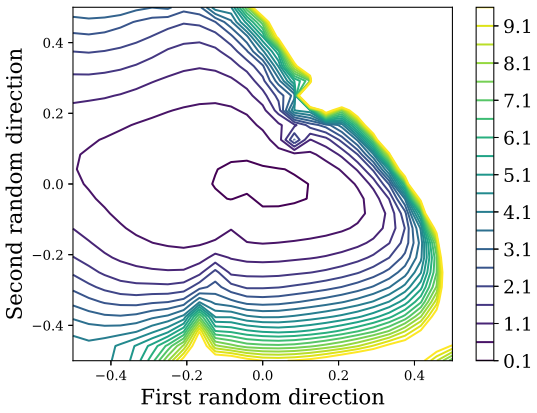


FIG. 10: 2D visualization of the loss surface of the neural network for the hyperbolic PDE (17) with concave flux function (19) (the PDE solution is smooth in this case)

numerical diffusion term, and as a result, the numerical solution corresponds to a parabolic equation with a finite amount of diffusion. This diffusion term can be controlled through refinement in space-time and by the use of higher-order discretization schemes.

Then, we analyzed the network training process for cases with and without diffusion in the PDE. Our study shows that the amount of added diffusion strongly affects the training of the network (e.g., the convergence rate, the behavior of the loss gradients). Moreover, we provided 2D visualizations of the loss landscape of the neural networks near their final optimization point, which indicate that the diffusion term in the PDE smooths the loss surface and makes it more convex, while the loss surface of the hyperbolic PDE with discontinuous solution demonstrates significant chaotic and non-convex features. However, the reasons for such behavior of the loss function are not perfectly understood yet. It would be certainly interesting to derive some analytical explanation of the observed phenomena as well. Nevertheless, through the experiments and analysis conducted in the current work we show that the physics informed machine learning framework is not suited for the hyperbolic PDEs with discontinuous solutions considered here.

ACKNOWLEDGMENTS

We thank Total for their financial support of our research on “Uncertainty Quantification.” The authors are also grateful to the Stanford University Petroleum Research Institute for Reservoir Simulation (SUPRI-B) for financial support of this work.

REFERENCES

Aziz, K. and Settari, A., *Petroleum Reservoir Simulation*, London: Elsevier/8, Applied Science Publishers, 1979.

Brooks, R. and Corey, T., Hydraulic Properties of Porous Media, *Hydrology Papers, Colorado State University*, vol. **24**, 1964.

Crandall, M.G. and Lions, P.L., Viscosity Solutions of Hamilton-Jacobi Equations, *Trans. Am. Math. Soc.*, vol. **277**, no. 1, pp. 1–42, 1983.

Cybenko, G., Approximation by Superpositions of a Sigmoidal Function, *Math. Control, Signals Syst.*, vol. **2**, no. 4, pp. 303–314, 1989.

Glorot, X. and Bengio, Y., Understanding the Difficulty of Training Deep Feedforward Neural Networks, in *Proc. of the 13th Int. Conf. on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

Gulshan, V., Peng, L., Coram, M., Stumpe, M.C., Wu, D., Narayanaswamy, A., Venugopalan, S., Widner, K., Madams, T., Cuadros, J., Kim, R., Raman, R., Nelson, P.C., Mega, J.L., and Webster, D.R., Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs, *J. Am. Med. Assoc.*, vol. **316**, no. 22, pp. 2402–2410, 2016.

He, K., Zhang, X., Ren, S., and Sun, J., Deep Residual Learning for Image Recognition, in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., and Kingsbury, B., Deep Neural Networks for Acoustic Modeling in Speech Recognition, *IEEE Signal Process. Mag.*, vol. **29**, no. 6, pp. 82–97, 2012.

Hornik, K., Stinchcombe, M., and White, H., Multilayer Feedforward Networks are Universal Approximators, *Neural Networks*, vol. **2**, no. 5, pp. 359–366, 1989.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L., Large-Scale Video Classification with Convolutional Neural Networks, *Proc. of the 2014 IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

Krizhevsky, A., Sutskever, I., and Hinton, G.E., Imagenet Classification with Deep Convolutional Neural Networks, in *Adv. Neural Inf. Process. Syst.*, vol. **25**, no. 2, pp. 1097–1105, 2012.

Lagaris, I.E., Likas, A., and Fotiadis, D.I., Artificial Neural Networks for Solving Ordinary and Partial Differential Equations, *IEEE Trans. Neural Networks*, vol. **9**, no. 5, pp. 987–1000, 1998.

Lax, P.D., *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, Philadelphia: Society for Industrial and Applied Mathematics, vol. **11**, 1973.

Lax, P.D., *Hyperbolic Partial Differential Equations*, Providence, RI: American Mathematical Soc., vol. **14**, 2006.

LeCun, Y. and Bengio, Y., Convolutional Networks for Images, Speech, and Time Series, in *The Handbook of Brain Theory and Neural Networks*, M.A. Arbib, Ed., Cambridge MA: The MIT Press, vol. **3361**, no. 10, 1995.

LeCun, Y., Bengio, Y., and Hinton, G., Deep Learning, *Nature*, vol. **521**, no. 7553, pp. 436–444, 2015.

Lee, H. and Kang, I.S., Neural Algorithm for Solving Differential Equations, *J. Comput. Phys.*, vol. **91**, no. 1, pp. 110–131, 1990.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T., Visualizing the Loss Landscape of Neural Nets, *Adv. Neural Inf. Process. Syst.*, pp. 6389–6399, 2018.

Lillicrap, T.P., Hunt, J.J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D., Continuous Control with Deep Reinforcement Learning,, 2015. arXiv: 1509.02971

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G.E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P.Q., and Corrado, G.S., Detecting Cancer Metastases on Gigapixel Pathology Images, 2017. arXiv: 1703.02442

Meade Jr., A.J. and Fernandez, A.A., The Numerical Solution of Linear Ordinary Differential Equations by Feedforward Neural Networks, *Math. Comput. Model.*, vol. **19**, no. 12, pp. 1–25, 1994.

Mnih, V., Badia, A.P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K., Asynchronous Methods for Deep Reinforcement Learning, in *Proc. of the 33rd Int. Conf. on Machine Learning*, pp. 1928–1937, 2016.

Nocedal, J. and Wright, S., *Numerical Optimization*, Berlin: Springer Science & Business Media, 2006.

Orr, F., *Theory of Gas Injection Processes*, Holte, Denmark: Tie-Line Publications, 2007.

Psichogios, D.C. and Ungar, L.H., A Hybrid Neural Network-First Principles Approach to Process Modeling, *AIChE J.*, vol. **38**, no. 10, pp. 1499–1511, 1992.

Raissi, M., Perdikaris, P., and Karniadakis, G.E., Physics Informed Deep Learning (Part I): Data-Driven Solutions of Nonlinear Partial Differential Equations, 2017. arXiv: 1711.10566

Raissi, M., Perdikaris, P., and Karniadakis, G.E., Physics-Informed Neural Networks: A Deep Learning Framework for Solving Forward and Inverse Problems Involving Nonlinear Partial Differential Equations, *J. Comput. Phys.*, vol. **378**, pp. 686–707, 2019.

Stewart, R. and Ermon, S., Label-Free Supervision of Neural Networks with Physics and Domain Knowledge, in *Proc. of the 31st AAAI Conference on Artificial Intelligence*, pp. 2576–2582, 2017.

Sutskever, I., Vinyals, O., and Le, Q.V., Sequence to Sequence Learning with Neural Networks, *Adv. Neural Inf. Process. Syst.*, pp. 3104–3112, 2014.

Zhu, Y., Zabaras, N., Koutsourelakis, P.S., and Perdikaris, P., Physics-Constrained Deep Learning for High-Dimensional Surrogate Modeling and Uncertainty Quantification without Labeled Data, *J. Comput. Phys.*, vol. **394**, pp. 56–81, 2019.

APPENDIX A. SENSITIVITY STUDY FOR THE BUCKLEY-LEVERETT PROBLEM

For the case described in Section 4.2.1, we perform a sensitivity study. Our aim here is to understand whether the result obtained in Section 4.2.1 for the Buckley-Leverett problem (nonlinear transport with a non-convex flux function) is strongly dependent on the particular choice of the network architecture and the different hyperparameters of the method, such as the number of training points in initial and boundary data N_u and the number of collocation points N_r in the interior of the domain.

First, we fix the network architecture to eight hidden layers with 20 neurons per hidden layer, and we vary the number of initial and boundary training data N_u in the range $(100, 600)$ and the number of collocation points N_r in the range (1000, 20,000). The final values of the loss function at the end of the training for these experiments are shown in Table A1. In all these cases, the network failed to yield a reasonable approximation of the shock; as the result, we observe a relatively large value of the loss function (i.e., $\sim 10^{-2}$). From Table A1, it is also clear that the network performance is not a strong function of the number of initial and boundary training data and the number of collocation points.

In the next experimental set, we kept the total number of training and collocation points fixed to $N_u = 300$ and $N_r = 10{,}000$, and varied the number of hidden layers in the range $(2, 12)$ and the number of neurons per hidden layer in the range $(10, 40)$. With these ranges, the total number of network parameters varied from 151 to over 18,000. Table A2 reports the value of the loss function at the end of the training for these different architectures. Again,

TABLE A1: Final loss at the end of training for different number of initial and boundary training data points N_u and different number of collocation points N_r. The network architecture is fixed to 8 hidden layers with 20 neurons per hidden layer

N_u \ N_r	**1000**	**5000**	**10,000**	**20,000**
100	1.6×10^{-2}	3.4×10^{-2}	3.0×10^{-2}	2.6×10^{-2}
300	2.2×10^{-2}	2.6×10^{-2}	3.4×10^{-2}	3.2×10^{-2}
600	1.3×10^{-2}	2.0×10^{-2}	3.1×10^{-2}	3.0×10^{-2}

TABLE A2: Final loss at the end of training for different number of hidden layers and different number of neurons per hidden layer. The total number of training and collocation points is fixed to $N_u = 300$ and $N_r = 10{,}000$

Layers \ Neurons	10	20	40
2	3.4×10^{-2}	3.2×10^{-2}	3.2×10^{-2}
4	1.6×10^{-2}	3.2×10^{-2}	3.1×10^{-2}
8	3.3×10^{-2}	3.4×10^{-2}	3.3×10^{-2}
12	3.5×10^{-2}	2.9×10^{-2}	1.9×10^{-2}

the observed trend is quite consistent—the final result is not weakly sensitive to the particular network architecture. Moreover, the PDE solutions $u(t, x)$ predicted by the neural networks in all these cases were quite similar to the ones reported in Section 4.2.1, where the network completely fails to approximate the shock.

In addition, we experimented with application of standard regularization of the network weights—the technique typically used in machine learning to decrease overfitting. Specifically, we added to the loss function $L(\theta)$ a regularization term of the form $l_{\text{reg}} = \beta W^T W$ (where W denotes the weights of the network) and considered a range of regularization constants $\beta = [1 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 5 \times 10^{-4}, 1 \times 10^{-3}]$. However, in these experiments we did not see any improvement in the PIML results for hyperbolic PDE.

Next, we also tested the PIML approach with different types of networks—a residual network architecture (He et al., 2016) and a convolutional neural network (CNN) (LeCun et al., 1995). For the residual network we added skip connections after each layer in the original fully connected architecture. With CNN architecture we used eight convolutional layers with 20 filters each, that perform 1D convolutions and have a kernel size of 1×1 (in this case, the number of parameters is the same as in the standard fully connected architecture reported in the paper). In these experiments we observed similar behavior—that the PIML approach fails for hyperbolic PDE but performs well for PDE with added diffusion term.