Physics-informed deep generative models

Yibo Yang, Paris Perdikaris
Department of Mechanical Engineering and Applied Mechanics,
University of Pennsylvania,
Philadelphia, PA 19104, USA
{ybyang, pgp}@seas.upenn.edu

Abstract

We consider the application of deep generative models in propagating uncertainty through complex physical systems. Specifically, we put forth an implicit variational inference formulation that constrains the generative model output to satisfy given physical laws expressed by partial differential equations. Such physics-informed constraints provide a regularization mechanism for effectively training deep probabilistic models for modeling physical systems in which the cost of data acquisition is high and training data-sets are typically small. This provides a scalable framework for characterizing uncertainty in the outputs of physical systems due to randomness in their inputs or noise in their observations. We demonstrate the effectiveness of our approach through a canonical example in transport dynamics.

1 Introduction

Despite their immense recent success, many state-of-the art machine learning techniques (e.g., deep neural nets, convolutional networks, recurrent networks [1, 2, 3, 4], etc.) are lacking robustness and often fail to provide any guarantees of convergence or quantify the error/uncertainty associated with their predictions. Hence, the ability to quantify predictive uncertainty and learn in a sample-efficient manner is a necessity [5, 6, 7, 8], especially in data-limited domains [9]. Even less well understood is how one can constrain such algorithms to leverage domain-specific knowledge and return predictions that satisfy certain physical principles [10] (e.g., conservation of mass, momentum, etc.). In recent work, Raissi *et. al.* [11, 12, 13, 14] explored this new interface between classical scientific computing and machine learning by revisiting the idea of penalizing the loss function of deep neural networks using differential equation constraints, as first put forth by Psichogios and Ungar [15] and Lagaris *et. al.* [16]. In this work, we revisit these approaches from a probabilistic standpoint [17, 18, 19] and develop an adversarial inference framework [20, 21, 22, 23] to enable the posterior characterization [24, 25] of the uncertainty associated with the model predictions. These uncertainty estimates reflect how observation noise and/or randomness in the system's inputs and outputs are propagated through complex infinite-dimensional dynamical systems described by non-linear partial differential equations (PDEs).

2 Methods

In Raissi *et. al.* [11, 12, 13, 14], the authors have considered constructing deep neural networks that return predictions which are constrained by PDEs of the form $\boldsymbol{u}_t + \mathcal{N}_{\boldsymbol{x}}\boldsymbol{u} = 0$, where $\boldsymbol{u}(\boldsymbol{x}, t)$ is represented by a deep neural network parametrized by a set of parameters θ, i.e. $\boldsymbol{u}(\boldsymbol{x}, t) = f_\theta(\boldsymbol{x}, t)$, $\boldsymbol{x}$ is a vector of space coordinates, t is the time coordinate, and $\mathcal{N}_{\boldsymbol{x}}$ is a nonlinear differential operator. As neural networks are differentiable representations, this construction defines a so-called *physics informed neural network* that corresponds to the PDE residual, i.e. $\boldsymbol{r}_\theta(\boldsymbol{x}, t) := \frac{\partial}{\partial t}f_\theta(\boldsymbol{x}, t) + \mathcal{N}_{\boldsymbol{x}}f_\theta(\boldsymbol{x}, t)$. By construction, this network shares the same architecture and parameters

with $f_\theta(\boldsymbol{x}, t)$, but it has different activation functions corresponding to the action of the differential operator. The resulting training procedure allows us to recover the shared network parameters θ using a few scattered observations of $\boldsymbol{u}(\boldsymbol{x}, t)$, namely $\{(\boldsymbol{x}_i, t_i), \boldsymbol{u}_i\}$, $i = 1, \ldots, N_u$, along with a larger number of collocation points $\{(\boldsymbol{x}_i, t_i), \boldsymbol{r}_i = 0\}$, $i = 1, \ldots, N_r$, that aim to penalize the PDE residual at a finite set of N_r collocation nodes. This way, the resulting optimization problem can be effectively solved using standard stochastic gradient descent without necessitating any elaborate constrained optimization techniques, simply by minimizing the composite loss function

$$\mathcal{L}_{\text{PDE}}(\theta) := \frac{1}{N_u} \sum_{i=1}^{N_u} \|f_\theta(\boldsymbol{x}_i, t_i) - \boldsymbol{u}_i\|^2 + \frac{1}{N_r} \sum_{i=1}^{N_r} \|\boldsymbol{r}_\theta(\boldsymbol{x}_i, t_i) - \boldsymbol{r}_i\|^2, \tag{1}$$

where the required gradients can be readily obtained using automatic differentiation [26].

In the proposed work we aim to model uncertainty in such physics-informed models by constructing conditional latent variable models of the form

$$p(\boldsymbol{u}|\boldsymbol{x}, t) = \int p(\boldsymbol{u}, \boldsymbol{z}|\boldsymbol{x}, t) d\boldsymbol{z} = \int p(\boldsymbol{u}|\boldsymbol{x}, t, \boldsymbol{z}) p(\boldsymbol{z}|\boldsymbol{x}, t) d\boldsymbol{z}, \tag{2}$$

and encourage the resulting samples to be constrained by a given physical law according to a likelihood

$$p(\boldsymbol{u}|\boldsymbol{x}, t, \boldsymbol{z}), \quad \boldsymbol{z} \sim p(\boldsymbol{z}), \quad \text{such that } \boldsymbol{u}_t + \mathcal{N}_{\boldsymbol{x}} \boldsymbol{u} = 0. \tag{3}$$

This setting encapsulates a wide range of deterministic and stochastic problems, where $\boldsymbol{u}(\boldsymbol{x}, t)$ is a potentially multi-variate field, and $\boldsymbol{z}$ is a collection of latent variables.

Following the recent findings of [27] we will train the generative model by matching the joint distribution of the generated samples $p_\theta(\boldsymbol{x}, t, \boldsymbol{u})$ with the joint distribution of the observed data $q(\boldsymbol{x}, t, \boldsymbol{u})$ by minimizing the reverse Kullback-Leibler (KL) divergence, which can be decomposed as

$$\mathbb{KL}[p_\theta(\boldsymbol{x}, t, \boldsymbol{u})\|q(\boldsymbol{x}, t, \boldsymbol{u})] = -h(p_\theta(\boldsymbol{x}, t, \boldsymbol{u})) - \mathbb{E}_{p_\theta(\boldsymbol{x}, t, \boldsymbol{u})}[\log(q(\boldsymbol{x}, t, \boldsymbol{u}))], \tag{4}$$

where $h(p_\theta(\boldsymbol{x}, t, \boldsymbol{u}))$ denotes the entropy of the generative model. This decomposition reveals the interplay between two competing mechanisms that define the model training dynamics. On one hand, minimizing the negative entropy term is encouraging the support of $p_\theta(\boldsymbol{x}, t, \boldsymbol{u})$ to spread to infinity, while the second term will penalize regions for which the support of $p_\theta(\boldsymbol{x}, t, \boldsymbol{u})$ and $q(\boldsymbol{x}, t, \boldsymbol{u})$ do not overlap. As observed in [27], this introduces a regularization mechanism for mitigating the pathology of mode collapse.

As the entropy term $h(p_\theta(\boldsymbol{x}, t, \boldsymbol{u}))$ is intractable, we can obtain a computable training objective by deriving the following lower bound (see proof in Appendix A)

$$h(p_\theta(\boldsymbol{x}, t, \boldsymbol{u})) \geq h(p(\boldsymbol{z})) + \mathbb{E}_{p_\theta(\boldsymbol{x}, t, \boldsymbol{u}, \boldsymbol{z})}[\log(q_\phi(\boldsymbol{z}|\boldsymbol{x}, t, \boldsymbol{u}))], \tag{5}$$

where the inference model $q_\phi(\boldsymbol{z}|\boldsymbol{x}, t, \boldsymbol{u})$ plays the role of a variational approximation to the true posterior over the latent variables, and appears naturally using information theoretic arguments in the derivation of the lower bound [27].

This construction leads to two coupled objectives that define an adversarial game for training the model parameters

$$\mathcal{L}_{\mathcal{D}}(\psi) := \mathbb{E}_{q(\boldsymbol{x}, t) p(\boldsymbol{z})}[\log \sigma(T_\psi(\boldsymbol{x}, t, f_\theta(\boldsymbol{x}, t, \boldsymbol{z})))] +$$
$$\mathbb{E}_{q(\boldsymbol{x}, t, \boldsymbol{u})}[\log(1 - \sigma(T_\psi(\boldsymbol{x}, t, \boldsymbol{u})))] \tag{6}$$
$$\mathcal{L}_{\mathcal{G}}(\theta, \phi) := \mathbb{E}_{q(\boldsymbol{x}, t) p(\boldsymbol{z})}[T_\psi(\boldsymbol{x}, t, f_\theta(\boldsymbol{x}, t, \boldsymbol{z})) + (1 - \lambda) \log(q_\phi(\boldsymbol{z}|\boldsymbol{x}, t, f_\theta(\boldsymbol{x}, t, \boldsymbol{z})))], \tag{7}$$

where $T_\psi(\boldsymbol{x}, t, \boldsymbol{u})$ is a parametrized discriminator used to approximate the KL divergence directly from samples using the density ratio trick [28], and $\sigma(x) = 1/(1 + e^{-x})$ is the logistic sigmoid function. Moreover, notice how the inference model $q_\phi(\boldsymbol{z}|\boldsymbol{x}, t, f_\theta(\boldsymbol{x}, t, \boldsymbol{z}))$ promotes cycle consistency in the latent variables, and serves as an entropic regularization term than allows us to stabilize model training and mitigate the pathology of mode collapse, as controlled by the user defined parameter λ [27]. The final training objective that encourages the generated samples to satisfy a given PDE reads as

$$\max_{\psi} \mathcal{L}_{\mathcal{D}}(\psi)$$
$$\min_{\theta, \phi} \mathcal{L}_{\mathcal{G}}(\theta, \phi) + \beta \mathcal{L}_{\text{PDE}}(\theta), \tag{8}$$

where positive values of β can be selected to place more emphasis on penalizing the PDE residual. For $\beta > 0$, the residual loss $\mathcal{L}_{\text{PDE}}(\theta)$ acts as a regularization term that encourages the generator $p_\theta(\boldsymbol{u}|\boldsymbol{x}, t, \boldsymbol{z})$ to produce samples that satisfy the underlying partial differential equation.

3 Results

Here we demonstrate the performance of the proposed methodology through the lens of a canonical problem in transport dynamics modeled by the Burgers equation with appropriate initial and boundary conditions. This equation arises in various areas of applied mathematics, including fluid mechanics, nonlinear acoustics, gas dynamics, and traffic flow [29]. In one space dimension the equation reads as

$$u_t + uu_x - \nu u_{xx} = 0, \quad x \in [-1, 1], \quad t \in [0, 1],$$
$$u(0, x) = -\sin(\pi x), \quad u(t, -1) = u(t, 1) = 0, \tag{9}$$

where $\nu = 0.01/\pi$ is a viscosity parameter, small values of which can lead to solutions developing shock formations that are notoriously hard to resolve by classical numerical methods [29].

Here, we construct a probabilistic representation for the unknown solution $u(x, t)$ using a physics-informed deep generative model $p_\theta(u|x, t, z)$, and we introduce parametric mappings corresponding to a generator $f_\theta(x, t, z)$, an encoder $q_\phi(z|x, t, u)$, and a discriminator $T_\psi(x, t, u)$ all constructed using deep feed-forward neural networks with 4 hidden layers with 50 neurons each. The activation function in all cases is chosen to be a hyperbolic tangent non-linearity. The prior over the latent variables is chosen to be a one-dimensional isotropic Gaussian distribution, i.e. $z \sim \mathcal{N}(0, 1)$. We train our probabilistic model using $30,000$ stochastic gradient Adam updates [30] using a learning rate of 10^{-4} on a data-set comprising of $N_u = 200$ input/output pairs for $u(x, t)$ – 100 points for the initial condition and 50 points for each of the domain boundaries – plus an additional $N_r = 10,000$ collocation points for enforcing the residual of the Burgers equation using the loss of equation 8 with $\lambda = 2.0$ and $\beta = 1.0$. A systematic study with respect to the model hyper-parameters is provided in Appendix B. Notice that the initial condition here is corrupted by a non-additive noise process as

$$u(x, 0) = -\sin(\pi(x + 2\delta)) + \delta, \qquad \delta = \epsilon/\exp(3|x|), \qquad \epsilon \sim \mathcal{N}(0, 0.1^2),$$

and our goal here is to propagate the effect of this uncertainty into the prediction of future system states. Notice how the noise variance is chosen to be larger around $x = 0$, therefore amplifying the effect of uncertainty on the shock formation.

The results of this experiment are summarized in figure 1, where we report the predicted mean solution, as well as the uncertainty associated with this prediction. We observe that the resulting generative model $p_\theta(u|x, t, z)$ can effectively capture the uncertainty in the resulting spatio-temporal solution due to the propagation of the input noise process through the complex non-linear dynamics of the Burgers equation. As expected, the uncertainty concentrates around the shock discontinuity that the solution develops around $t = 0.5$. Although we only plot the first two moments of the solution, we must emphasize that the generative model $p_\theta(u|x, t, z)$ provides a complete probabilistic characterization of its non-Gaussian statistics.
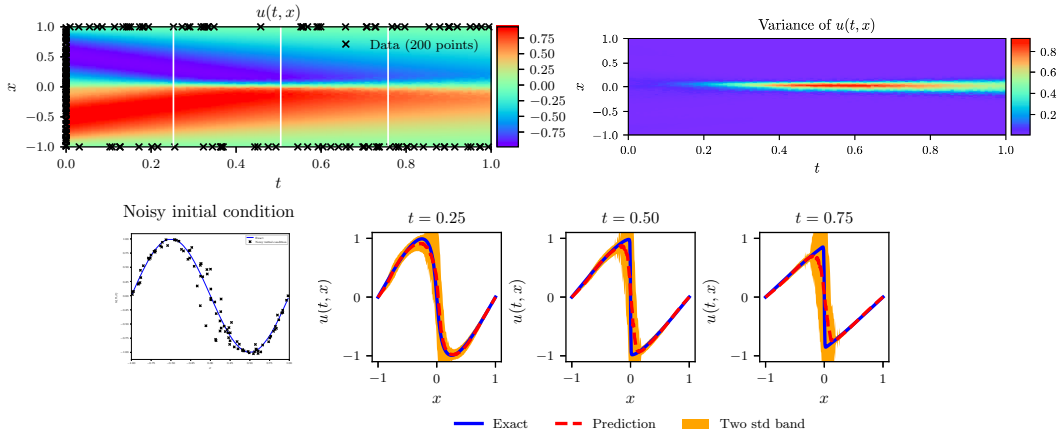

Figure 1: *Top:* Mean and variance of $p_\theta(u|x, t, z)$, along with the location of the noisy training data $\{(x_i, t_i), u_i\}_{i=1}^{N_u}$. *Bottom:* Noisy data for the initial condition, and the resulting prediction and predictive uncertainty at $t = 0.25$, $t = 0.5$ and $t = 0.75$.

References

[1] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

[2] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436.

[3] Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science, 350*(6266), 1332-1338.

[4] Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature biotechnology, 33*(8), 831.

[5] Bui, T., Hernández-Lobato, D., Hernandez-Lobato, J., Li, Y., & Turner, R. (2016, June). Deep Gaussian processes for regression using approximate expectation propagation. In *International Conference on Machine Learning* (pp. 1472-1481).

[6] Gal, Y., & Ghahramani, Z. (2016, June). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning* (pp. 1050-1059).

[7] Gal, Y. (2016). Uncertainty in deep learning. *University of Cambridge*.

[8] Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature, 521*(7553), 452.

[9] Ma, C., Li, Y., & Hernández-Lobato, J. M. (2018). Variational Implicit Processes. *arXiv preprint arXiv:1806.02390*.

[10] Welling, M., & Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pp. 681-688).

[11] Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2017). Physics Informed Deep Learning (Part I): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*.

[12] Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2017). Physics informed deep learning (Part II): data-driven discovery of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10566*.

[13] Raissi, M., & Karniadakis, G. E. (2018). Hidden physics models: Machine learning of nonlinear partial differential equations. *Journal of Computational Physics, 357*, 125-141.

[14] Raissi, M. (2018). Deep Hidden Physics Models: Deep Learning of Nonlinear Partial Differential Equations. *arXiv preprint arXiv:1801.06637*.

[15] Psichogios, D. C., & Ungar, L. H. (1992). A hybrid neural network-first principles approach to process modeling. *AIChE Journal, 38*(10), 1499-1511.

[16] Lagaris, I. E., Likas, A., & Fotiadis, D. I. (1998). Artificial neural networks for solving ordinary and partial differential equations. *IEEE transactions on neural networks, 9*(5), 987-1000.

[17] Kevin, M. (2012). Machine learning, a probabilistic perspective.

[18] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

[19] Weber, M., Welling, M., & Perona, P. (2000, June). Unsupervised learning of models for recognition. In *European conference on computer vision* (pp. 18-32). Springer, Berlin, Heidelberg.

[20] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems* (pp. 2672-2680).

[21] Louizos, C., Ullrich, K., & Welling, M. (2017). Bayesian compression for deep learning. In *Advances in Neural Information Processing Systems* (pp. 3288-3298).

[22] Saatci, Y., & Wilson, A. G. (2017). Bayesian gan. In *Advances in neural information processing systems* (pp. 3622-3631).

[23] Wilson, A. G., Hu, Z., Salakhutdinov, R. R., & Xing, E. P. (2016). Stochastic variational deep kernel learning. In *Advances in Neural Information Processing Systems* (pp. 2586-2594).

[24] Ghahramani, Z. (2001). An introduction to hidden Markov models and Bayesian networks. *International journal of pattern recognition and artificial intelligence, 15*(01), 9-42.

[25] Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., & Murphy, K. (2017). An information-theoretic analysis of deep latent-variable models. *arXiv preprint arXiv:1711.00464*.

[26] Baydin, A. G., Pearlmutter, B. A., Radul, A. A., & Siskind, J. M. (2017). Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research, 18*(153), 1-153.

[27] Li, C., Li, J., Wang, G., & Carin, L. (2018). Learning to Sample with Adversarially Learned Likelihood-Ratio.

[28] Sugiyama, M., Suzuki, T., & Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.

[29] Basdevant, C., Deville, M., Haldenwang, P., Lacroix, J. M., Ouazzani, J., Peyret, R., ... & Patera, A. T. (1986). Spectral and finite difference solutions of the Burgers equation. *Computers & fluids, 14*(1), 23-41.

[30] Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[31] Cover, T. M., & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

[32] Glorot, X., & Bengio, Y. (2010, March). Understanding the difficulty of training deep feed-forward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249-256).

A   Proof of the Entropy Lower Bound

Here we follow the derivation of Li [27] to construct a computable lower bound for the entropy $h(p_\theta(\boldsymbol{x}, t, \boldsymbol{u})))$. To this end, we start by considering random variables $(\boldsymbol{x}, t, \boldsymbol{u}, \boldsymbol{z})$ under the joint distribution

$$p_\theta(\boldsymbol{x}, t, \boldsymbol{u}, \boldsymbol{z}) = p_\theta(\boldsymbol{u}, \boldsymbol{x}, t | \boldsymbol{z}) p(\boldsymbol{z}) = p_\theta(\boldsymbol{u} | \boldsymbol{x}, t, \boldsymbol{z}) p(\boldsymbol{x}, t) p(\boldsymbol{z}),$$

where $p_\theta(\boldsymbol{u} | \boldsymbol{x}, t, \boldsymbol{z}) = \delta(\boldsymbol{u} - f_\theta(\boldsymbol{x}, t, \boldsymbol{z}))$, and $\delta(\cdot)$ is the Dirac delta function. The mutual information between $(\boldsymbol{x}, t, \boldsymbol{u})$ and $\boldsymbol{z}$ satisfies the information theoretic identity

$$I(\boldsymbol{x}, t, \boldsymbol{u}; \boldsymbol{z}) = h(\boldsymbol{x}, t, \boldsymbol{u}) - h(\boldsymbol{x}, t, \boldsymbol{u} | \boldsymbol{z}) = h(\boldsymbol{z}) - h(\boldsymbol{z} | \boldsymbol{x}, t, \boldsymbol{u}),$$

where $h(\boldsymbol{x}, t, \boldsymbol{u})$, $h(\boldsymbol{z})$ are the marginal entropies and $h(\boldsymbol{x}, t, \boldsymbol{u} | \boldsymbol{z})$, $h(\boldsymbol{z} | \boldsymbol{x}, t, \boldsymbol{u})$ are the conditional entropies [31]. Since in our setup $\boldsymbol{x}$ and t are deterministic variables independent of $\boldsymbol{z}$, and samples of $p_\theta(\boldsymbol{u} | \boldsymbol{x}, t, \boldsymbol{z})$ are generated by a deterministic function $f_\theta(\boldsymbol{x}, t, \boldsymbol{z})$, it follows that $h(\boldsymbol{x}, t, \boldsymbol{u} | \boldsymbol{z}) = 0$. We therefore have

$$h(\boldsymbol{x}, t, \boldsymbol{u}) = h(\boldsymbol{z}) - h(\boldsymbol{z} | \boldsymbol{x}, t, \boldsymbol{u}), \tag{10}$$

where $h(z) := -\int \log p(\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}$ is a constant with respect to the model parameters θ.

Now consider a general variational distribution $q_\phi(\boldsymbol{z} | \boldsymbol{x}, t, \boldsymbol{u})$ parametrized by a set of parameters ϕ. Then,

$$
\begin{aligned}
h(\boldsymbol{z} | \boldsymbol{x}, t, \boldsymbol{u}) &= -\mathbb{E}_{p_\theta(\boldsymbol{x}, t, \boldsymbol{u}, \boldsymbol{z})}[\log(p_\theta(\boldsymbol{z} | \boldsymbol{x}, t, \boldsymbol{u}))] \\
&= -\mathbb{E}_{p_\theta(\boldsymbol{x}, t, \boldsymbol{u}, \boldsymbol{z})}[\log(q_\phi(\boldsymbol{z} | \boldsymbol{x}, t, \boldsymbol{u}))] \\
&\quad - \mathbb{E}_{p_\theta(\boldsymbol{x}, t, \boldsymbol{u})}[\mathbb{KL}[p_\theta(\boldsymbol{z} | \boldsymbol{x}, t, \boldsymbol{u}) || q_\phi(\boldsymbol{z} | \boldsymbol{x}, t, \boldsymbol{u})]] \\
&\leq -\mathbb{E}_{p_\theta(\boldsymbol{x}, t, \boldsymbol{u}, \boldsymbol{z})}[\log(q_\phi(\boldsymbol{z} | \boldsymbol{x}, t, \boldsymbol{u}))].
\end{aligned}
\tag{11}
$$

Viewing $\boldsymbol{z}$ as a set of latent variables, then $q_\phi(\boldsymbol{z} | \boldsymbol{x}, t, \boldsymbol{u})$ is a variational approximation to the true intractable posterior over the latent variables $p_\theta(\boldsymbol{z} | \boldsymbol{x}, t, \boldsymbol{u})$. Therefore, if $q_\phi(\boldsymbol{z} | \boldsymbol{x}, t, \boldsymbol{u})$ is introduced as an auxiliary encoder associated with the generative model $p_\theta(\boldsymbol{x}, t, \boldsymbol{u})$, for which $\boldsymbol{u} = f_\theta(\boldsymbol{x}, t, \boldsymbol{z})$ and $\boldsymbol{z} \sim p(\boldsymbol{z})$, then we can use equations 10 and 11 to bound the entropy term in equation 4 as

$$h(p_\theta(\boldsymbol{x}, t, \boldsymbol{u})) \geq h(p(\boldsymbol{z})) + \mathbb{E}_{p_\theta(\boldsymbol{x}, t, \boldsymbol{u}, \boldsymbol{z})}[\log(q_\phi(\boldsymbol{z} | \boldsymbol{x}, t, \boldsymbol{u}))] \tag{12}$$

B   Systematic Studies

Here we provide results on a series of comprehensive systematic studies that aim to quantify the sensitivity of the resulting predictions on: (i) the neural network initialization, (ii) the total number of training and collocation points, (iii) the neural network architecture, and (iv) the adversarial training procedure. In all cases we have used the non-linear Burgers equation as a prototype problem.

B.1   Sensitivity with respect to the neural network initialization

In order to quantify the sensitivity of the proposed methods with respect to the initialization of the neural networks, we have considered a noise-free data set comprising of $N_u = 150$ and $N_r = 10000$ training and collocation points, respectively, and fixed the architecture for generator neural networks to include 4 hidden layers with 50 neurons each and discriminator neural networks to include 3 hidden layers with 50 neurons each , and a hyperbolic tangent activation function. Then we have trained an ensemble of 15 cases all starting from a normal Xavier initialization [32] for all network weights (with a randomized seed), and a zero initialization for all bias parameters. In table 1 we report the relative error between the predicted mean solution and the known exact solution for this problem for all 15 randomized trials using at set of 25600 randomly selected test points. Evidently, our results are robust with respect to the the neural network initialization as in all cases the stochastic gradient descent training procedure converged roughly to the same solution. We can summarize this result by reporting the mean and the standard deviation of the relative $\mathcal{L}_2$ error as

$$\hat{\mathcal{L}}_2 \in [\mu_L - \sigma_L, \mu_L + \sigma_L] = [4.7 \times 10^{-2} - 1.3 \times 10^{-2}, 4.7 \times 10^{-2} + 1.3 \times 10^{-2}].$$

Relative $\mathcal{L}_2$ error				
4.1e-02	7.9e-02	4.4e-02	4.0e-02	3.8e-02
3.2e-02	5.7e-02	4.7e-02	6.5e-02	4.0e-02
3.5e-02	3.5e-02	6.4e-02	4.0e-02	4.9e-02

Table 1: Relative $\mathcal{L}_2$ error with different seed of initialization in non-noise case.

B.2 Sensitivity with respect to the total number of training and collocation points

In this study our goal is to quantify the sensitivity of our predictions with respect to the total number of training and collocation points N_u and N_r, respectively. As before, we have considered noise-free data sets, and fixed the architecture for generator neural networks to include 4 hidden layers with 50 neurons each and discriminator neural networks to include 3 hidden layers with 50 neurons each, a hyperbolic tangent activation function, and a normal Xavier initialization [32] for all network weights and zero initialization for all network biases. The results of this study are summarized in table 2, indicating that as the number of collocation points are increased, a more accurate prediction is obtained. This observation is in agreement with the original results of Raissi *et. al.* [11] for deterministic physics-informed neural networks, indicating the role of the PDE residual loss as an effective regularization mechanism for training deep generative models in small data regimes.

N_u \ N_r	10	100	250	500	1000	5000	10000
60	9.3e-01	5.6e-01	4.8e-01	5.0e-02	1.9e-01	5.0e-02	5.1e-02
90	5.8e-01	5.3e-01	3.5e-01	1.5e-01	4.9e-02	1.0e-01	5.8e-02
150	6.7e-01	1.4e-01	3.0e-01	3.6e-02	4.9e-02	1.2e-01	4.7e-02

Table 2: Relative $\mathcal{L}_2$ prediction error for different number of training and collocation points N_u and N_r, respectively.

B.3 Sensitivity with respect to the neural network architecture

In this study we aim to quantify the sensitivity of our predictions with respect to the architecture of the neural networks that parametrize the generator, the discriminator, and the encoder. Here we have fixed the number of noise-free training data to $N_u = 150$ and $N_r = 10000$, and we kept the number of layers for discriminator to always be one less than the number of layers for generator (e.g., if the number of layers for generator is two then the number of layers for discriminator is one, etc.). In all cases, we have used a hyperbolic tangent non-linearity and a normal Xavier initialization [32]. In table 3 we report the relative $\mathcal{L}_2$ prediction error for different feed-forward architectures for the generator, discriminator, and encoder (i.e., different number of layers and number of nodes in each layer). The general trend suggests that as the neural network capacity is increased we obtain more accurate predictions, indicating that our physics-informed constraint on the PDE residual can effectively regularize the training process and safe-guard against over-fitting. We note number of neurons in each layer as N_n and number of layers for generator (encoder) as N_g.

N_g \ N_n	20	50	100
2	4.2e-01	3.8e-01	5.7e-01
3	6.5e-02	3.5e-02	2.1e-02
4	9.3e-02	4.7e-02	5.4e-02

Table 3: Relative $\mathcal{L}_2$ prediction error for different feed-forward architectures for the generator, encoder, and the discriminator. The total number of layers of the latter was always chosen to be one less than the number of layers for generator.

B.4 Sensitivity with respect to the adversarial training procedure

Finally, we test the sensitivity with respect to the adversarial training process. To this end, we have fixed the number of noise-free training data to $N_u = 150$ and $N_r = 10000$, and the neural network

architecture to be the same as B.2, and we vary the total number of training steps for the generator K_g and the discriminator K_d within each stochastic gradient descent iteration. The results of this study are presented in table 4 where we report the relative $\mathcal{L}_2$ prediction error. These results reveal the high sensitivity of the training dynamics on the interplay between the generator and discriminator networks, and pinpoint on the well known peculiarity of adversarial inference procedures which require a careful tuning of K_g and K_d for achieving stable performance in practice.

K_d / K_g	1	2	5
1	3.5e-01	5.0e-01	1.5e+00
2	4.3e-02	3.2e-01	5.4e-01
5	4.7e-02	2.3e-01	7.0e-01

Table 4: Relative $\mathcal{L}_2$ error with different number of training for generator and discriminator in each epoch.