1.4 Learning-enriched Differential Equation Models in Optimal Control and Inverse Problems

Guozhi Dong, Michael Hintermüller, and Kostas Papafitsoros

Differential equations, ordinary (ODEs) or partial (PDEs), that is, equations that involve functions and their (partial) derivatives, have been essential tools in many fields of science. They describe the dynamics of the physical world phenomena allowing scientists to get useful insights and make predictions. Very often in applications, it is desirable to focus on specific constituents of the differential equations, for instance, some *parameters* or otherwise called *controls*, that significantly affect how the solution will look like. This need leads to the widely applicable field of *optimal control of differential equations*, where one is seeking for suitable values of these parameters that result in the solution being close to some desirable *state*, e.g., a specific temperature distribution in a room or a specific configuration of a fluid flow. In certain medical applications, these parameters can be some tissue-specific biophysical variables, whose precise value can tell clinicians more about the nature of the tissue, e.g., tumor vs. healthy tissue. Applied mathematicians play a vital role in developing techniques that facilitate these diagnoses. A first key step is to identify as precisely as possible the differential equations related to a given imaging technique whose solutions depend on these parameters. By obtaining measured data that are related to these solutions and correspond to a specific small tissue area, one is able to make a link to some specific biophysical parameter values, and achieve a precise classification of that tissue area. This workflow is done, for instance, in *quantitative magnetic resonance imaging* (MRI); see Figure 1, [4], and the corresponding Scientific Highlights article of the Annual Research Report 2019 of the Weierstrass Institute.

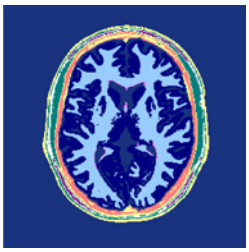


Fig. 1: *Quantitative imaging techniques aim to measure precise values of biophysical quantities of fixed units associated to different types of tissues (here color-coded)*

However, very often a differential equation is merely a simplification of a far more complex ground-truth physical process. This physical process can be unknown or too complicated to be precisely modeled. Nevertheless, experimental data can provide some glimpse into the true process itself. One can achieve that, for instance, by considering a family of the differential equations parameters, the *input data*, and experimentally measuring the response of the system, the *output data*, that corresponds to each one of these parameters. It is then desirable to have a tool — a *learned map* — that generalizes this input-output relation to input data that have not been used in this experiment and eventually approximates the physical process. With regard to the optimal control problems, this learned map will substitute the *control-to-state* map. It turns out that this versatile learning from input-output data can be realized via *artificial neural networks* (ANNs). The use of ANNs and the general field of *deep learning* — one of the cores of *artificial intelligence* (AI) — is nowadays ubiquitous. Their remarkable versatility and good approximation properties that are the result of being trained in a set of data, generalizing well in "unseen" data, have made them a powerful tool essentially in any area that involves some type of data classification and interpolation.

There has, therefore, been a need for the introduction, analysis, and application of versatile data-driven frameworks for learning totally or partially unknown physical models via ANNs. This was recently realized in [1] within the EF3 project "Direct reconstruction of biophysical parameters using dictionary learning and robust regularization" in the frame of MATH+, the Berlin Mathematics Re-

search Center, which is a cross-institutional and interdisciplinary Cluster of Excellence. ANNs were employed either to learn some unknown nonlinearity in physical models or to represent the complex parameter-to-solution maps of differential equations and subsequently be embedded into optimal control problems. For example in the case of qMRI, the dynamics that map certain tissue-dependent biophysical parameters to the acquired signal, are learned and incorporated into the reconstruction process, yielding more accurate values for the tissue parameter maps with the obvious benefits for clinicians and patients.

As it is a case for any new mathematical framework, it needs to be shown to be mathematically sound and viable. Important questions arise, such as in what degree the approximation quality of a given ANN (stemming from the quantity and quality of the available data) affects the final solution of the optimal control problem. It is also vital to design and develop robust numerical methods for the solution of these learning-informed differential equations and their corresponding optimal control problems. The recent work [1] also addressed these challenges and showed the versatility of the framework in key applications, such as qMRI and the modeling of phase transitions in fluids.

Deep learning and artificial neural networks in brief

Mathematically, a neural network is a function $\mathcal{N} : \mathbb{R}^r \to \mathbb{R}^s$, with a feed-forward architecture, in the sense that the input is successively propagated into L *layers*. Every layer consists of a series of *neurons* that perform weighted averages of their inputs that have been fed from the neurons of the previous layer. An activation function σ then decides if the output of each neuron will be passed to the neurons of the next layer, by assigning relatively large values to it. For a more precise example, a standard feed-forward ANN with one hidden layer has the following form:

$$\mathcal{N}(x) = W_0\sigma(W_1x + b_1) + b_0, \quad x \in \mathbb{R}^r, \tag{1}$$

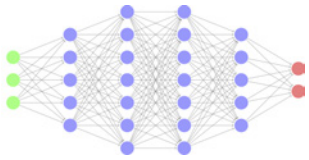


Fig. 2: *Visual example of a 4-hidden layer network*

where $W_1 \in \mathbb{R}^{l\times r}$, $W_0 \in \mathbb{R}^{s\times l}$ are weight matrices and $b_1 \in \mathbb{R}^l$, $b_0 \in \mathbb{R}^s$ are bias vectors. In that case, we say that the hidden layer has l neurons. A visual example of a 4-hidden layer network is shown in Figure 2, with the neurons in hidden layers depicted as blue nodes. The activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is some nonlinear map that acts component-wise on a vector in $\mathbb{R}^l$. Popular activation functions are Sigmoid-type functions, e.g., $\sigma(z) = \arctan(z)$, and the rectified linear unit (ReLU), $\sigma(z) = \max(0, z)$,

Given some data pairs $\{(x_i, f_i) \in \mathbb{R}^r \times \mathbb{R}^s,\ i = 1, \dots, N\}$, one of the main tasks of deep learning is to identify suitable choices for weight matrices and bias vectors, collectively denoted by θ, such that the corresponding neural network $\mathcal{N}_\theta$ satisfies $\mathcal{N}_\theta(x_i) \simeq f_i$, $i = 1, \dots, N$. In other words, the target is for $\mathcal{N}_\theta$ to *learn* a map that corresponds to the input-output data pairs. This learning is typically achieved via the so-called *supervised learning approach*, essentially by solving a minimization problem (*network training*) with respect to θ.

The success of deep learning in several applications is mainly due to the fact that, given enough training data, the resulting parameters θ^* lead to a network $\mathcal{N}_{\theta^*}$ that tends to behave well also in other points outside the training set, that is, they have good approximation and interpolation capabilities. Mathematically, this fact is also corroborated by the *universal approximation theorem for*

neural networks [3], which states that if the activation function σ is a continuous non-polynomial function, then every family of neural network functions of a fixed number of layers is dense to the set of continuous functions $C(\mathbb{R}^r, \mathbb{R}^s)$, in the topology of uniform convergence in compact sets.

Main framework of learning-informed optimal control

Inspired by key applications on optimization models with physical laws constraints, such as quantitative magnetic resonance imaging (qMRI), a versatile data-driven framework was proposed and mathematically analyzed in [1]. The starting point is a general optimal control problem of the form

$$\begin{aligned} &\underset{(y,u)}{\text{minimize}} \quad \frac{1}{2}\|Ay - g\|_H^2 + \frac{\alpha}{2}\|u\|_U^2, \quad \text{over } (y,u) \in (Y \times U), \\ &\text{subject to} \quad e(y,u) = 0, \quad \text{and} \quad u \in \mathcal{C}_{ad}. \end{aligned} \tag{2}$$

Here, Y, U are some appropriate function spaces, $\alpha > 0$, $\mathcal{C}_{ad}$ is a constraint set for the control u, A is a linear operator — for instance, in the case of inverse problems, it can be regarded as the forward operator of the problem —, and g denotes some given data. The term of focus in (2) is the equation $e(y,u) = 0$, a differential equation describing a physical process, with y being the solution variable (state). Assuming uniqueness of solutions for (2), we write $y = \Pi(u)$ to define the well-defined control-to-state map. We focus on the case where the precise form of the physical process e is unknown, (i) either as a whole or (ii) with respect to a specific constituent. For the latter, consider, for example, the following semilinear partial differential equation:

$$-\Delta y + f(x,y) = u, \quad \text{in } \Omega \subseteq \mathbb{R}^d, \tag{3}$$

where f is a completely unknown nonlinear function. In that case, any calculation of $y = \Pi(u)$ is out of reach. Nevertheless, given the availability of a data pair set $\{(u_i, y_i) : i = 1, \dots, N\}$, such that $y_i \sim \Pi(u_i)$, one can train a neural network $\mathcal{N}$ and use it to approximate the overall unknown control-to-state map. In the case of (3), the neural network $\mathcal{N}$ aims to approximate only the unknown constituent f arriving in the following *learning-informed* PDE

$$-\Delta y + \mathcal{N}(x,y) = u, \quad \text{in } \Omega, \tag{4}$$

In general, we end up with a learning-informed control-to-state map, denoted by Π_N, that can then be embedded into the optimal control problem. A schematic illustration of this framework is shown in the diagram of Figure 3.

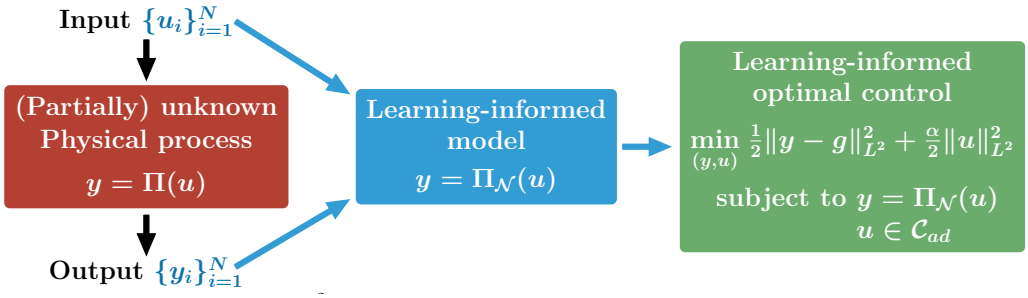


Fig. 3: *Schematic illustration of the learning-informed optimal control framework*

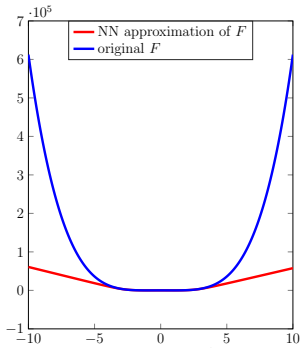


Fig. 4: *Approximation of a double-well potential F by a neural network*

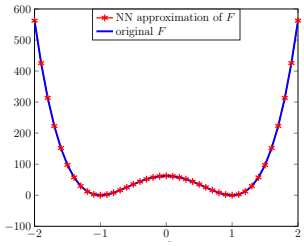


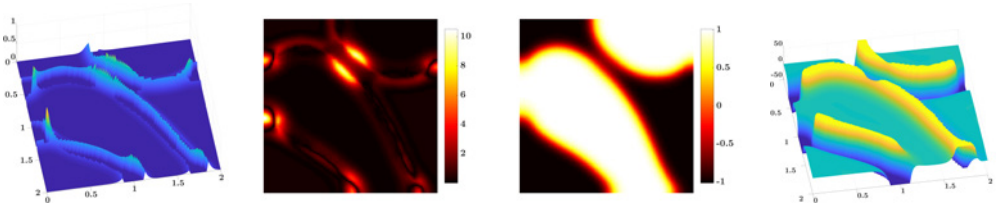
Fig. 5: *Detail of Figure 4*

The versatility and applicability of this approach was established in [1] by its validation in two key applications, discussed next.

Optimal control of semilinear partial differential equations. The following general optimal control problem of learning-informed semilinear PDEs was studied in [1]

$$\begin{aligned} \underset{(y,u)}{\text{minimize}} \quad & \frac{1}{2}\|y-g\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{L^2(\Omega)}, \quad \text{over } (y,u) \in (H^1(\Omega)\times L^2(\Omega)), \\ \text{subject to} \quad & -\Delta y + \mathcal{N}(x,y) = u, \quad \text{in } \Omega, \quad \partial_\nu y = 0 \ \text{ on } \partial\Omega, \\ \text{and} \quad & \underline{u}(x) \le u(x) \le \overline{u}(x), \quad \text{for a.e. } x \in \Omega. \end{aligned} \tag{5}$$

Here, $\mathcal{N}:\Omega\times\mathbb{R}\to\mathbb{R}$ is an ANN that has been sufficiently trained offline in order to approximate an unknown function f in its domain. By formulating the PDE as a first-order optimality condition of a variational problem, existence of solutions to the PDE and further to the optimal control problem were shown. Note that uniqueness for the PDE solution can only be guaranteed if $\mathcal{N}$ is strictly monotone in the second variable, which cannot be reasonably assumed if f is not strictly monotone. In [1], a particular example of a stationary Allen–Cahn equation has been tested as a benchmark problem. There, the nonlinearity f is associated to the derivative of a double-well potential function F, which models the separation of a fluid into two states and whose precise form has traditionally been a matter of modeling choice rather than data driven. In Figures 4 and 5, we show an example where the (derivative of the) ground potential F is learned by a neural network using some local data. Despite the fact that the learned potential looks globally quite different — recalling that the network approximation is good only in a compact set — the important double-well part is well approximated. Indeed, the solution $(y_{\mathcal{N}}, u_{\mathcal{N}})$ of (5) ends up being close to the corresponding solution of the ground-truth model $(\overline{y},\overline{u})$, Figure 6.



Control difference $|u_{\mathcal{N}} - \overline{u}|$ | State difference $|y_{\mathcal{N}} - \overline{y}|$ | Learning-informed state $y_{\mathcal{N}}$ and control $u_{\mathcal{N}}$

Fig. 6: *Comparison of the solutions of the learning-informed and ground-truth optimal control problems of the stationary Allen–Cahn equation with the double-well potentials shown in Figure 4*

Inverse problems on quantitative imaging. It turns out that the task of quantitative MRI, a high-level description of which we have already given, can be formulated as a special case of the general optimal control problem (2). This formulation reads as follows:

$$\begin{aligned} \underset{(y,u)}{\text{minimize}} \quad & \frac{1}{2}\|P\mathcal{F}(y) - g^\delta\|^2_{L^2(\Omega)} + \frac{\alpha}{2}\|u\|^2_{H^1(\Omega)}, \quad \text{over } (y,u) \in [L^2(\Omega)]^{3L}\times[H^1(\Omega)]^3 \\ \text{s.t.} \quad & \frac{\partial y}{\partial t}(t) = y(t)\times\gamma B(t) - \left(\frac{y_1(t)}{T_2}, \frac{y_2(t)}{T_2}, \frac{y_3(t)-\rho m_e}{T_1}\right), \ t = t_1,\dots,t_L, \\ \text{and } & y(0) = \rho m_0, \quad u := (\rho, T_1, T_2) \in \mathcal{C}_{ad} \subset ([L^\infty(\Omega)]^+)^3. \end{aligned} \tag{6}$$

The goal is to estimate the physical unit values of (T_1, T_2), the tissue-dependent magnetic relaxation parameters, as well as the proton spin density ρ, with the ultimate target being the classification of a given tissue slice Ω. These biophysical quantities are inserted into the physical dynamics, via the *Bloch equations*, the above ODE system, which describes the evolution of the magnetization y in a tissue volume unit (voxel). In an MRI experiment, subsamples of the Fourier coefficients $(P\mathcal{F})$ of y are measured at specific times $t_1, \dots, t_L$, resulting in possibly noisy data g^δ. This ill-posed *inverse problem* is modeled in the first line of (6), where additional H^1 regularization is imposed on the unknown parameter maps $T_1, T_2 : \Omega \to \mathbb{R}$ as well as ρ. If the parameter-to-solution map of the Bloch equations $y = \Pi(T_1, T_2)$ is explicitly known, then it can be embedded into the minimization problem (6), resulting in the following least-squares formulation:

$$\underset{(\rho,T_1,T_2)}{\text{minimize}} \quad \frac{1}{2}\|P\mathcal{F}(\rho\Pi(T_1,T_2)) - g^\delta\|^2_{L^2(\Omega)}, \qquad \text{s.t. } (\rho, T_1, T_2) \in \mathcal{C}_{ad}. \tag{7}$$

This approach was introduced in our previous work [2] where (7) was solved with a Levenberg–Marquadt method, giving superior results compared to some of the current state-of-the-art methods in qMRI [4]. Nevertheless, explicit formulas of the Bloch map are only available in certain special choices of the external magnetic field B. However, numerical methods or some elaborate targeted experiments can provide data that facilitate a neural network approximation $\Pi_{\mathcal{N}}$ that can take the role of Π in (7). It was shown in [1] that this learning-informed model can achieve similar results to the "ground-truth" one; see Figure 7. Furthermore, the approach is more flexible and it has the capability to learn some potential perturbation of the initially believed to be accurate model. Finally, there is a significant reduction in the computational load, since a repetitive solution of the exact physical model is avoided.

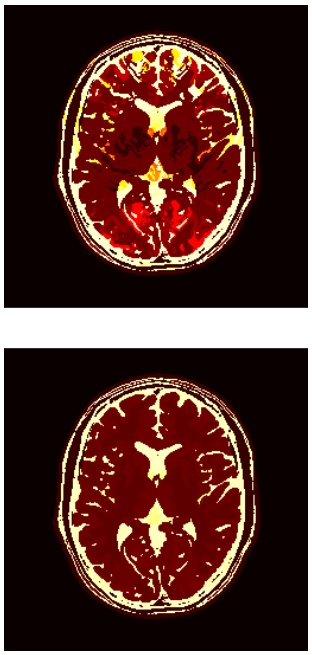


Fig. 7: *T2 reconstructed by a dictionary-based method [4] (top) and by the proposed learning-informed method [1] (bottom)*

Mathematical challenges

In terms of the problem (2) and its learned counterpart, many mathematical questions arise. For instance: *Do the learning-informed PDEs admit solutions? Will the optimizers associated to the learning-informed model be close to the one associated to the ground-truth one?* These and similar questions were also addressed in [1]. For instance, focusing on the semilinear PDEs (3) and (4), and under some standard assumption on f (e.g., continuity and certain polynomial growth rates), it was shown that for every $\epsilon > 0$ there exists a neural network $\mathcal{N} \in C^\infty(\mathbb{R}^d \times \mathbb{R})$ such that

$$\sup_{\|y\|_{L^\infty(\Omega)} \le K} \|f(\cdot, y) - \mathcal{N}(\cdot, y)\|_{L^2(\Omega)} < \epsilon \tag{8}$$

with the corresponding learning-informed PDE (4) admitting a weak solution. The constant $K > 0$ is associated to a uniform bound of the type $\|y\|_{H^1(\Omega)} + \|y\|_{C(\overline{\Omega})} \le K$ satisfied for every solution of the original PDE uniformly for all controls $u \in \mathcal{C}_{ad}$, with the solutions of the learning-informed PDE satisfying similar estimates. Indeed, we observed that the uniform boundedness of the range of the input and output data (state variable) played a crucial role in these proofs, stemming from the fact that the density of neural networks holds in the topology of uniform convergence on compact sets. Analogous estimates are shown for the control-to-state maps

$$\|\Pi(u) - \Pi_{\mathcal{N}}(u)\|_{L^2(\Omega)} \le C\epsilon, \text{ for all admissible } u \in L^2(\Omega) \tag{9}$$

as well as their derivatives.

Regarding the corresponding optimal controls, quantitative convergence results were also proven, showing that under certain conditions, the solution of the learning-informed problem $u_{\mathcal{N}}$ will converge to the solution $\overline{u}$ of (2), with specific rates

$$\|u_{\mathcal{N}} - \overline{u}\|_U \leq C\left(L_0\epsilon_1 + \epsilon_1\epsilon_2 + \epsilon_2\|Q(\overline{u}) - g\|_H\right).$$

Here, C is some constant depending on the parameter α, the Lipschitz constant L_0 of the operator $Q := A\Pi$ and its derivative Q', and ϵ_1 and ϵ_2 are error bounds between Q and $Q_{\mathcal{N}} := A\Pi_{\mathcal{N}}$ and their derivatives, respectively.

Finally, a common numerical algorithmic framework using a sequential quadratic programming (SQP) approach combined with semismooth Newton, was used to tackle both problems (5) and (6), while the numerical algorithm for learning was executed in a separate offline phase before the SQP algorithm.

Conclusions and outlook

We introduced a general optimal control framework that incorporates physical processes that are enriched through data-driven components, and we showed its feasibility in two key applications. This idea combines the power of both traditional mathematical modeling with machine learning methods, and is able to deliver more accurate physical models. The latter can finally serve as data-faithful constraints in optimization tasks. In the future, we expect that such approaches will be used to learn small but systematic deviations from previously well-established physical models. Finally, we note that several mathematical challenges arise from this work. For instance, the use of nonsmooth neural networks, stemming from the incorporation of nonsmooth activation functions, has become prevalent due to certain approximation and trainability advantages. For our set-up, this nonsmoothness poses difficulties in establishing rigorous first-order optimality systems for the learning-informed optimal control problems. Future studies should also focus on incorporating the architecture and the training of the networks into the overall minimization process to further robustify the new technique.

References

[1] G. Dong, M. Hintermüller, K. Papafitsoros, *Optimization with learning-informed differential equation constraints and its applications*, WIAS Preprint no. 2754, 2020.

[2] ———, *Quantitative magnetic resonance imaging: From fingerprinting to integrated physics-based models*, SIAM J. Imaging Sci., **12**:2 (2019), pp. 927–971.

[3] M. Leshno, V. Lin, A. Pinkus, S. Schocken, *Multilayer feedforward networks with a nonpolynomial activation function can approximate any function*, Neural Networks, **6**:6 (1993), pp. 861–867.

[4] D. Ma, V. Gulani, N. Seiberlich, K. Liu, J. Sunshine, J.Duerk, M.Griswold, *Magnetic resonance fingerprinting*, Nature, **495** (2013), pp. 187–192.