Deep learning and partial differential equations

organized by
Lin Lin, Jianfeng Lu, and Lexing Ying

Workshop Summary

The aim of this workshop was to bring together experts in deep learning and partial differential equations. This is a fast moving field with significant interests from multiple disciplines, such as physics, chemistry, materials science, data sciences, finance etc. The workshop participants had varied backgrounds, and included experts in applied mathematics, physics, chemistry and computer science. The workshop followed the usual AIM format. There were two talks each morning. We held an open problem session on the Monday afternoon, generating around 20 problems on wide-ranging topics. In all cases the problems centered on understanding the mathematical understanding of deep neural networks, in particular the error analysis, and the neural network architecture suitable for scientific computing applications. In the afternoons of Tuesday to Friday we broke up into five groups to discuss some of these problems. Each group consists of two to eight people. The groups were fluid, with people moving and creating new groups from one day to the next. Below are brief reports on the problems that were discussed.

(1) **Hamilton-Jacobi-Bellman equation**. This subgroup focused on solving Hamilton-Jacobi-Bellman equations of the form

$$\begin{cases} u_t + H(x, \nabla u) = 0 & \text{in } \mathbb{R}^n \times \{t > 0\} \\ u = g & \text{on } \mathbb{R}^n \times \{t = 0\}. \end{cases} \tag{1}$$

The group started by investigating the current work [nakamura2019adaptive] for solving high dimensional Hamilton-Jacobi equations. This solves the characteristic equations

$$\begin{cases} \dot{x}(t) = \nabla_p H(x(t), p(t)) \\ \dot{p}(t) = -\nabla_x H(x(t), p(t)) \\ \dot{z}(t) = \nabla_p H(x(t), p(t)) \cdot p(t) - H(x(t), p(t)), \end{cases} \tag{2}$$

forwards in time from random initial conditions, and treats the problem as an interpolation problem, interpolation from the solution values obtained along the characteristics. The interpretation of the characteristics is that $z(t) = u(x(t), t)$ and $p(t) = \nabla u(x(t), t)$ for short time (i.e., when u is smooth and characteristics do not cross). Once characteristics cross, which can happen in relatively short time, the values of $z(t)$ and $p(t)$ are incorrect, and cannot be used for interpolation.

The group decided to propose and investigate a new method that can handle the case of crossing characteristics, which is fundamental for Hamilton-Jacobi equations, where solutions are in general not continuously differentiable and must be interpreted in the viscosity sense. The method is based on tracing characteristics backwards in time from a point (x, t) to the initial time, which can be done for Hamiltonians

that are *convex* in the gradient, which is a wide range of problems including state-dependent optimal control. Other recent work [chow2019algorithm] is also based on solving characteristics backwards in time, but [chow2019algorithm] computes the solutions at individual points (x, t) and does not use deep leaning to learn the solution function $u(x, t; \omega)$.

To describe the proposed approach, we parameterize $u = u(x, t; \omega)$ with a deep neural network with weights ω and repeat the following steps until convergence.

(a) Sample $(x_1, t_1), \ldots, (x_n, t_n) \in \mathbb{R}^n \times (0, T]$ and $y_1, \ldots, y_m \in \mathbb{R}^n$ at random.

(b) For each $i = 1, \ldots, n$, solve the characteristic equations (2) backwards in time to compute the triple $(x^i(t), p^i(t), z^i(t))$ with boundary conditions $x^i(t_i) = x_i$, $p^i(t_i) = \nabla u(x_i, t_i; \omega)$, and $z^i(0) = g(x^i(0))$.

(c) Form the loss function

$$L(\omega) = \sum_{i=1}^{n} \left[\int_0^{t_i} (u(x^i(t), t; \omega) - z^i(t))^2 \, dt + [u_t(x_i, t_i; \omega) + H(x_i, \nabla u(x_i, t_i; \omega))]^2 \right] + \sum_{i=1}^{m} (u(y_i, 0) - g(y_i))^2.$$

(d) Take one (or several) steps of gradient descent

$$\omega^{k+1} = \omega^k - \alpha \nabla_\omega L(\omega^k)$$

where $\omega^0 = \omega$.

The loss function includes the residual of the PDE and boundary condition, as in the Deep Galerkin Method (DGM) [sirignano2018dgm], but also includes a term that encourages the solution to respect the backwards characteristics, whose purpose is to select the viscosity solution from among the many non-unique Lipschitz almost everywhere solutions that could be selected by a naive application of DGM.

The group found some Python code using pytorch to implement the DGM method and began to experiment with modifying the code to implement the algorithm above. We expect that some changes may be necessary to the algorithm once we are able to implement and test it. This is where we left off at the end of the week, and we all agreed to experiment with the algorithm and have a skype meeting in about 1 month to update.

(2) **Symmetry and anti-symmetry**: This subgroup focused on understanding the interplay between permutation symmetry and generalization error, and the efficient neural network representation of permutation-invariant function, permutation-equivariant mapping, and anti-symmetric functions. One concrete outcome of the discussion was a new proof generalizing the neural network representation of permutation invariant functions to dimension $d > 1$. The group also identified another universal representation for anti-symmetric functions for any $d \geq 1$, which we expect will be useful for variational Monte Carlo simulation of quantum-many body systems. A preprint documenting these progresses is currently under preparation.

(3) **Mean field limit of CNN and ResNet:** This subgroup focused on understanding the mean field limit of existing CNNs and ResNets as well as designing new architectures that give rise to the compositional function class presented in Weinan E's talk. One concrete outcome was a mean field theory for two layer CNN. The subgroup also discussed the mean field models for a few ResNet-type models, such as ResNeXt.

Towards the end, the subgroup produced a few proposals of residual network architectures for the compositional function class, by exploiting ideas including time-scale separation, generative models, meta-learning, and paramter reordering.

(4) **Phase transitions and Neural Tangent Kernel:** This subgroup concentrated on the theoretical study of the training dynamics of neural networks in both parameter and function spaces. More specifically, recent studies have highlighted two possible approaches to the study of convergence of wide single layer neural networks while trained for supervised learning. The first approach concentrates on neural networks parametrized as $\hat{f}(x,\theta) := \frac{1}{n}\sum_{i=1}^{n} c_i\sigma(x,\theta_i)$, where $(c_i,\theta_i)_i$ are the parameters of the network. By the structure of the approximating function this approach is called *mean field* approach. Convergence of the nonlinear gradient descent dynamics in parameter space for $n \to \infty$ can be proven in this setting by leveraging the variational structure of the supervised learning problem. On the other hand, when the weights of the neural network are initialized under a different scaling, the network behaves as a linear model (a constant kernel method) under training when $n \to \infty$. This regime is commonly referred to as the *Neural Tangent Kernel* (NTK) regime. In this setting convergence to the global minimum, at least in the overparametrized framework, can be proven much more easily.

This subgroup has focused on the relation between the two models introduced above. In particular, it is possible to see the NTK regime as emerging from a linear scaling of the mean field model under a parameter α in the limit $\alpha \to \infty$. The subgroup has worked on the question of whether it is possible to draw a connection between the two models by studying the behavior of the rescaled model for finite values of α. The first step in this direction consists in carrying out a perturbative expansion of the model around $\alpha = \infty$. A similar asymptotic expansion is the object of two recent papers (Dyer & Gur-Ari, arXiv:1909.11304; Huang & Yau, arXiv:1909.08156). In the subgroup we have worked on linking the results of these two papers and the scaling transformation introduced above. We have further started to discuss in how to measure the evolution of the kernel in the expansion around $\alpha = \infty$ and its approach to the evolution of the kernel trained in the mean field regime. In this sense, it would be interesting to investigate about the presence or not of a phase transition between the two regimes for a certain value of α. Discussions in this sense are still ongoing and will hopefully lead to interesting results and fruitful collaborations.

(5) **Group symmetry in deep learning and image inverse problems:** The group symmetry working group firstly discussed the possibility of using higher-order moments as feature representations for the task of using a neural network to regress potential energy surface (PES) from atomic locations and bond information. The difficulty lies in the efficiently computation of rotation equivariant representation, which involves Fourier transform on the sphere (by spherical harmonics). For image inverse problem, we started by reviewing various image restoration applications in the field, including image denoising, deblurring, de-flection, de-convolution, super-resolution, and the other inverse problems mainly image inpainting and post-electron tomography (PET). The popular methods in the field fall into two categories, the pixel-to-pixel networks (or U network), and ResNet methods which resemble traditional PDE methods (unrolling dynamics), represented by LISTA and TRD. The group then discussed how to combine prior information in images and data-driven

methods in image restoration deep networks, which can be done in two methods: first, early stopping in unrolling dynamics and removing the fidelity term, and second, keeping the fidelity term while leaving hyperparameters trainable from data, such as the step size.

Overall the organizers concluded that the workshop was very successful. It brought together applied mathematicians and domain scientists working in different but related fields, exposed everyone to new ideas, and forged new collaborations. We are optimistic that workshop participants will continue to work on these questions, and that the list of problems, which is available on AIM's website, will be a valuable resource for the field.

The organizers sincerely thank the AIM staff for their help and support!

Bibliography

[chow2019algorithm] Y. T. Chow, J. Darbon, S. Osher, and W. Yin. Algorithm for overcoming the curse of dimensionality for state-dependent hamilton-jacobi equations. *Journal of Computational Physics*, 387:376–409, 2019.

[nakamura2019adaptive] T. Nakamura-Zimmerer, Q. Gong, and W. Kang. Adaptive deep learning for high dimensional hamilton-jacobi-bellman equations. *arXiv preprint arXiv:1907.05317*, 2019.

[sirignano2018dgm] J. Sirignano and K. Spiliopoulos. Dgm: A deep learning algorithm for solving partial differential equations. *Journal of Computational Physics*, 375:1339–1364, 2018.