A Higher Gauge Theory for Machine Learning and Inference

Dalton A R Sakthivadivel [1 2]

[1]Department of Mathematics, Stony Brook University [2]Department of Physics and Astronomy, Stony Brook University

Introduction

- There is currently no rigorous theory of the effectiveness of deep and machine learning [1]
- Inference by machine learning algorithms is a particular variational problem on the dynamics of a data-generating process [2]
- A higher gauge theory describing the dynamics of string-like objects has been developed by Baez and Schreiber to generalise Yang-Mills theory [3]
- On this basis, we develop a formal, geometric theory of inference as computing the solution to an arbitrary dynamical system

This provides an explanation of the effectiveness of deep learning, by showing it computes solution sets of arbitrary dynamical systems as data-generating processes.

Preliminaries

- A **field theory** is a field of entities (scalars, vectors, tensors) X over a space-time Σ.
- Consider the $U(1)$ gauge group in electromagnetism. A **gauge theory** describes a field theory for which the Lagrangian $\mathcal{L}(\phi)$ of field configurations is invariant under differing choices of some set of group elements $g(x) \in G$.
- A **principal bundle** is the geometric framework in which a field theory and a gauge theory are 'packaged' together, as a fibration $X \to \Sigma$ with invariance $P \times_G F \to \Sigma$.
- An **associated bundle** provides a matter field given by a representation of G, such as an electric field transforming under $\rho : U(1) \to GL(1, \mathbb{C})$.
- A **connection one-form** is, loosely, a map ω to the Lie algebra $\mathfrak{g}$ of G. The integration of a differential form along a particle's path is a term in the action of the particle (cf. Chern-Simons theory). In a $U(1)$ gauge theory, $\mathfrak{g}$ is the electromagnetic four-potential.
- A **section** γ is a path that maps inputs to outputs such that $\gamma : \Sigma \to X$.
- The **parallel transportation** of a point-like object ϵ, $\mathrm{tra}(\epsilon)$, is the translation along a path in the fibres in a manner that follows the connection.
- A **dynamical system** or **data-generating process** is a function that produces states for inputs, so that a dynamical system creates a path in the configuration space X.

Higher Gauge Theory in Bundle Gerbes

- Let X be a field and ϵ be a particle in the field, such that X is the configuration space of ϵ. Clearly, $\mathrm{tra}(\epsilon)$ describes the solution to a dynamical system, such that for
$$\dot{x} = f(x), \tag{1}$$
parallel transportation of an ϵ gives us a trajectory
$$\gamma = \{x(\sigma_0), \dots, x(\sigma_f)\}$$
satisfying everywhere tangency to (1) for $x \in X$ and $\sigma \in \Sigma$. Each trajectory has Lie group elements g attached to it, so that in parallel transportation on a principal bundle, interaction with a gauge field is measured. Each path obeys least action, such that the path follows the Lie algebra-valued connection one-form on which it is defined.
- Inference is a variational problem over an ensemble of paths, so that one considers not the optimal single trajectory, but a probability density $p(\gamma)$ as the optimal assignment of probabilities to multiple trajectories (given a constraint J). Can we define parallel transportation in the space of paths, $\gamma \in \Gamma$?
- Following Baez and Schreiber, we generalise parallel transportation of particles on paths, to parallel transportation of paths, by taking the fibration of the path space $\prod_\Sigma(X) \mapsto \Gamma$ and giving another gauge group H to the total space Γ.
- Parallel transportation of paths on a parameterised surface (a moduli space of paths), such that we trace out a worldsheet, now follows the connection between path-wise connections, $\mathfrak{h}$, as well as the path-wise connection $\mathfrak{g}$. This higher order interaction between gauge fields, with the higher dimensional structure of path transport, is encapsulated in a two-bundle and related structures.
- The parallel transportation of paths is given by the parallel transportation on a two-bundle with connection,
$$\mathrm{tra}(\gamma) = \exp\left\{-\int_\Sigma \kappa(\gamma, \omega)\right\},$$
for κ a shift operation [4] between two pullback-connections $\gamma_0^* \omega_0$ and $\gamma_1^* \omega_1$, creating a worldsheet of lifts along $\mathfrak{h}$. Like one-transport, this satisfies an ODE for horizontal lifts.

A Mathematical View of Inference

We consider maximum entropy as a general view of inference, and energy-based learning as its instantiation in ML. Maximum entropy is a general theory of inference given by Jaynes in [5], with a related generalisation to path spaces called maximum calibre [6]. Relatedly, it has been suggested that all types of machine learning are nothing but various limiting cases of energy-based learning [2], which is itself a maximum entropy model. We will now show that

1. Maximum calibre is the parallel transportation over paths
2. This corresponds to the optimal assignment of probabilities over paths
3. This is formally the solution to an equation describing the dynamics of a noisy data-generating process.

Proof of 1. Take maximum calibre as a variational principle over paths, such that we have the maximised action functional
$$p(\gamma) = \arg\max_{p(\gamma)} \left[-\int_\Gamma \ln\{p(\gamma)\} p(\gamma) + J(\gamma) p(\gamma)\, \mathrm{d}\gamma - C\right] \tag{2}$$
for $\mathbb{E}[J] = C$. When maximised with respect to $p(\gamma)$, we have
$$p(\gamma) = \exp\{-J(\gamma)\}. \tag{3}$$
Let changing a value of J' be a functorial shift operation. Then, for J an abelian group forming a product space with X, the parallel transportation of paths is
$$\mathrm{tra}(\gamma) = \exp\left\{-\int_\Sigma \gamma^* j'\right\} = \exp\{-j(\gamma)\}. \tag{4}$$
For a set of values $J = j$, each given by $\int_\Sigma \circ \kappa$, we recover (3). □

Proof of 2. By definition, finding (3) from (2) is a variational problem. Under the parallel transportation with respect to the connection J' on paths, we have for (2) a geodesic on the $\mathfrak{h}$-valued connection, yielding (4) when optimised. Thus, maximum calibre yields a *least action* $p(\gamma)$. □

Proof of 3. Let an SDE be a dynamical system driven by a Wiener process. Any solution to the SDE is described by the probability amplitude as a solution to the SDE's Fokker-Planck equation. Since $p(\gamma)$ is given variationally by (4), parallel transport gives the amplitude of
$$p(x(\sigma_{i+1}) \mid x(\sigma_i), \dots, x(\sigma_0))$$
for any x in the image of γ. Since (4) computes the integration of path-wise differential forms, it is formally equivalent to solving for this density. □

References

[1] Terrence J Sejnowski. The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*, 117(48):30033–30038, 2020.

[2] Yann Lecun, Sumit Chopra, Raia Hadsell, Marc Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning. In G Bakir, T Hofman, B Scholkopt, A Smola, and B Taskar, editors, *Predicting structured data*. MIT Press, 2006.

[3] John C Baez and Urs Schreiber. Higher gauge theory. In Alexei Davydov, Michael Batanin, Michael Johnson, Stephen Lack, and Amnon Neeman, editors, *Categories in Algebra, Geometry and Mathematical Physics: Conference and Workshop in Honor of Ross Street's 60th Birthday, July 11–16/July 18–21, 2005*, number 431 in Contemporary Mathematics, pages 7–30, Providence RI, 2007. American Mathematical Society.

[4] Konrad Waldorf. Parallel transport in principal 2-bundles. *Higher Structures*, 2(1):57–115, 2018.

[5] Edwin Thompson Jaynes. Information theory and statistical mechanics. *Physical Review*, 106:620–630, 1957.

[6] Steve Pressé, Kingshuk Ghosh, Julian Lee, and Ken A Dill. Principles of maximum entropy and maximum caliber in statistical physics. *Reviews of Modern Physics*, 85:1115–1141, Jul 2013.