Learning Energy Minimization on Moduli Space of Connections

L. J. Pereira

1 Overview

In gauge theory, the moduli space of connections results from quotienting the space of principal connections of a fiber or vector bundle by the structure group, generating a gauge equivariant parameter space. A moduli space of connections can be used to cover the activity of a collection of deep neural networks, which are represented as trajectories on vector bundles. On this upper bounding space, a top-down energy-based attention mechanism can be trained from sampled activity of the bottom-up trajectories of the underlying networks through a composition of energies. Dynamics of attention are computed by minimizing the energy functional of the space. In particular, the Yang-Mills moduli space, a subset of the total connection space, can be constructed to be a smooth, compact, and oriented manifold in 4 dimensions with critical points known as Yang-Mills connections (or instantons). These connections minimize curvature between bundles and, from an information geometric perspective, they minimize relative entropy or KL divergence between two manifolds that are gauge equivariant. Allowing these connections to function as sources of variational noise, we aim to minimize the total energy of the moduli space. Associating the physics of unifed field theories to findings in neuroscience and machine learning theory may justify further introspection into cosmological fine-tuning and the Anthropic Principle.

1.1 Prerequisite Review

A fiber bundle serves as a useful mathematical object to analyse both the recursive construction of artificial and biological neurons and neural networks, as well as their application in the geometry of latent information, which uses manifold representations for information processing and statistical learning. Finer details can be found in my notebooks on differential geometry, Lie groups and algebras, and gauge theory. A fiber bundle makes precise the idea of one topological space (called a fiber) being parameterized by another topological space (called a base). The bundle also comes with a group action on the fiber that represents the different ways the fiber can be viewed as equivalent. The fiber bundle has a property, known as local trivialization, that allows neighborhoods of the bundle to be computed as simple, oriented product spaces, despite the global space possibly being unoriented or twisted.

A family of fibers associated to a base can be described by defining a standard (or template) fiber which all other fibers are isomporphic to. This is formalized by defining a projection mapping, which may be diffeomorphic or homotopic, that connects positional data from the entirety of the space of fibers to a base and implicitly from one fiber to another. When the template fiber is a vector space, the bundle is called a vector bundle. Similarly, the standard connections between fibers, known as a principal Ehresmann connection, can intuitively be understood as a covariant directional derivative on the tangent spaces of the manifolds and will always exist. An interesting type of connection occurs when using a fiber bundle equipped with a Lie group action. Lie groups have a unique recursive nature as a result of the group being itself a differentiable manifold. When equipped with a Lie group action, the bundle structure can be used to represent both the original fiber or vector bundle as well as a higher level collection of tangent spaces connecting bundles, in what's known as a bundle of connections. As will be explained in the section on moduli spaces, The Yang-Mills or instanton connections within this bundle of connections correspond to those that minimize their curvature.

1.2 A Priori Structure Groups

A biological first principle of covariance arises naturally from analysis of neuronal activity, which appears to favour functional localization and Hebbian learning. Moreover, brain networks appear to flow in connectome-specific smooth diffusive waves along gyrification paths which are theorized to be caused by differential tangential growth. Recall, covariance is a measure of the joint variability of two random variables and is increasingly positive when a pair show similar behavior and is negative when dissimilar. Covariance finds a direct description in differential geometry as a principal Ehressmann connection, which represents the covariant derivative between fibers of a bundle. To allow dynamics on a bundle using gauge theory, it becomes necessary to impose a generalized a priori covariance principle to maintain integrity of information being transported in bilateral and hierarchical directions. Yet for a standard learning model, covariance of functionality is an a posteriori feature since the joint variability is unknown until individual modules are fully trained. Covariance of fibers can be achieved by imposing the structure group to be a Lie Group, but can also be achieved by imposing restrictions on the projection map of Riemannian manifolds without explicitly defining the structure group beforehand.

1.3 Moduli Spaces

With covariance established on connections, it becomes possible to perform inference using higher levels of abstraction on gauge fields. This is done by constructing a gauge equivariant bundle of connections known as a moduli space of connections which can be further reduced into a Yang-Mills moduli space defined to be a finite dimensional manifold. This reduced space has local and global minima being connections with minimized energy known as Yang-Mills connections or instantons. Yang-Mills connections serve as a natural choice of connection on principal and vector bundles since they minimize their curvature. From an information geometric perspective, this can be thought of as minimizing relative entropy between sampled trajectories of two manifolds which happen to be gauge equivariant. The gauge field strength is the curvature F_A of the connection, and the energy of the gauge field is given by the Yang–Mills action functional:

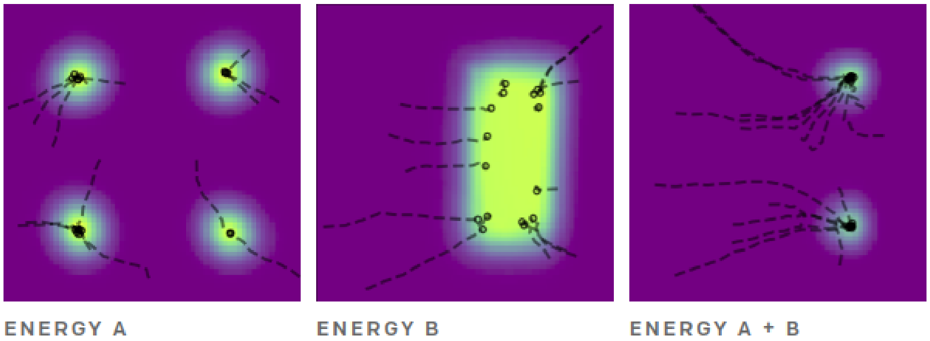
$$\mathrm{YM}(A) = \int_X \|F_A\|^2 \, d\mathrm{vol}_g.$$

With the aim of having zero or vanishing curvature, we search for a connection with curvature as small as possible. The Yang–Mills action functional corresponds to the L^2-norm of the curvature, and its Euler–Lagrange equations describe the critical points of this functional, either the absolute minima or local minima. That is, Yang–Mills connections are those that minimize their curvature.

1.4 Variational Inference and Energy Composition

As underlying neural networks perform inference, their trajectories pass through hierarchies of weighted hidden layers in a bottom-up manner. At the same time, a top-down variational noise is produced around the instantons that best minimize energy of total activity on the moduli space of connections. Sampling through the joint distribution on all latent variables is represented by generation on an energy-based model (EBM) that is the sum of each conditional EBM and corresponds to a product of experts model. This energy minimization forms an attention mechanism that is learned on the top-down manifold as well as having a generative effect on underlying neural networks through variational inference. The main idea behind variational methods is to pick a family of distributions over the latent variables with its own variational parameters, $q(z_{1:m}|v)$, and then attempt to find the setting of the parameters that makes q close to the posterior of interest (the instantion). Closeness of the two distributions in variational inference is measured with the Kullback-Leibler (KL) divergence,

$$KL(q||p) = E_q\left[\log \frac{q(z)}{p(z|x)}\right].$$



A 2D example of combining energy functions through their summation and the resulting sampling trajectories.

2 Exploratory

Recent research in ML, Neuroscience, and Physics strengthens this direction of research; including the GLOM system for representing part-whole hierarchies in neural networks (Hinton 2021), similar part-whole investigations of biological neural trajectories using geometric analysis (Russo et al, 2020), as well as demonstrations of soliton generation from Bose-Einstein condensate. Examining pure mathematics research in ergodic theory and hyperbolic dynamics on moduli spaces enforces the ergodic free-energy minimizing model described in previous notebooks. An encoding mechanism using the cobordism property of the Yang-mills moduli space is also considered.

2.1 Part-Whole Hierarchies

In Artificial Neural Networks (Hinton 2021): `https://arxiv.org/abs/2102.12627`.

In the brain, using geometric analysis (Russo et al, 2020):

`https://www.sciencedirect.com/science/article/abs/pii/S0896627320303664`.

`https://www.sciencedirect.com/science/article/abs/pii/S0092867420312289`

2.2 Cobordism

Moduli of Yang–Mills connections have been most studied when the dimension of the base manifold X is four. Here the Yang–Mills equations admit a simplification from a second-order PDE to a first-order PDE, the anti-self-duality equations. Additionally, these manifolds demonstrate a cobordism property where in specific circumstances (when the intersection form is definite) the moduli space of ASD instantons on a smooth, compact, oriented, simply-connected four-manifold X gives a cobordism between a copy of the manifold itself, and a disjoint union of copies of the complex projective plane $\mathbb{CP}^2$. This might provide a natural and gauge invariant method of dimensionality reduction to be applied to the hierarchical encoding and decoding of an upper bounding layer.

2.3 Solitons and Instantons

A soliton is a localized, non-dispersive solution of a nonlinear theory in Euclidean space and is a real object. Conversely, instantons are not real and only exist as solutions to the equations of motion of a quantum field theory after a Wick rotation, in which time is made imaginary. Note, a Wick rotation is a transformation that substitutes an imaginary-number variable for a real-number variable in order to solve a problem of (complex) Minkowski space in Euclidean space. Therefore, instantons are not observable, but are used to calculate and explain quantum mechanical effects that can be observed, such as tunneling. In quantum chromodynamics (QCD), instantons are believed to tunnel between the topologically different color vacua.

2.4 Bose-Einstein Condesate

A Bose–Einstein condensate (BEC) is a state of matter which is typically formed when a gas of bosons at low densities is cooled to temperatures very close to absolute zero causing a large fraction of the bosons to occupy the lowest quantum state, at which point microscopic quantum mechanical phenomena, particularly wavefunction interference, become macroscopic. This transition to BEC occurs below a critical temperature that is given by:

$$T_{\mathrm{c}} = \left(\frac{n}{\zeta(3/2)}\right)^{2/3} \frac{2\pi\hbar^2}{mk_{\mathrm{B}}}$$

T_{c} is the critical temperature,
n the particle density,
m the mass per boson,
$\hbar$ the reduced Planck constant,
k_{B} the Boltzmann constant and
ζ the Riemann zeta function;

Using quantum gases made from atoms, it was demonstrated to be possible to create magnetic solitons in a (dipolar) BEC made from atoms with different spins (Farolfi, 2020). These quantum solitons are density waves, meaning they are local waves of particles. Spatial phase distributions can be optically imprinted onto a BEC of atoms and can also be shown to create solitons (Denschlag, 2000).

2.5 Variational Methods in EBMs

An energy function E in an EBM can be thought of as an unnormalized negative log probability. To convert an energy function to its equivalent probabilistic representation after normalization, $P(y \mid x)$, we apply the Gibbs-Boltzmann formula with latent variables z being marginalized implicitly through integration, i.e. $P(y \mid x) = \int_z P(y, z|x)$. Then,

$$P(y \mid x) = \frac{\int_z \exp(-\beta E(x,y,z))}{\int_y \int_z \exp(-\beta E(x,y,z))}$$

The derivation introduces a β term which is the inverse of temperature T, so as $\beta \to \infty$ the temperature goes to zero, and we see that $\check{y} = \mathrm{argmin}_y E(x,y)$. This inverse temperature limit appears similar to the critical temperature in BEC. We can redefine our energy function as an equivalent function with free energy F_β,

$$\begin{aligned} F_\infty(x,y) &= \mathrm{argmin}_z E(x,y,z) \\ F_\beta(x,y) &= -\frac{1}{\beta} \log \int_z \exp(-\beta E(x,y,z)). \end{aligned}$$

If we have a latent variable model and want to eliminate the latent variable z in a probabilistically correct way, we just need to redefine the energy function in terms of F_β,

$$P(y \mid x) = \frac{\exp(-\beta F_\beta(x,y,z))}{\int_y \exp(-\beta F_\beta(x,y,z))}.$$

With variational methods, instead of only minimizing the energy function with respect to z we prevent the energy function from being 0 everywhere by constraining the flexibility of the latent variable z. The energy function is defined as sampling z randomly according to a distribution whose logarithm is the cost that links it to z. This distribution is commonly chosen to be a Gaussian with mean $\bar{z}$ which results in Gaussian noise being added to $\bar{z}$. The reparameterization trick is often used to allow for backpropagation during training despite the random sampling.

Work in progress

References

[1] Gao, Tingran. "The diffusion geometry of fibre bundles: Horizontal diffusion maps." Applied and Computational Harmonic Analysis 50 (2021): 147-215.

[2] Eckhard Meinrenken, "Principal bundles and connections", Lecture Notes.

[3] Tao, T. "What is a gauge?" (2008) https://terrytao.wordpress.com/2008/09/27/what-is-a-gauge/

[4] Wetzel, Sebastian J., and Manuel Scherzer. "Machine learning of explicit order parameters: From the Ising model to SU (2) lattice gauge theory." Physical Review B 96.18 (2017): 184410.

[5] Du, Yilun, and Igor Mordatch. "Implicit generation and generalization in energy-based models." arXiv preprint arXiv:1903.08689 (2019).

[6] Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." Journal of the American statistical Association 112.518 (2017): 859-877.

[7] Hinton, Geoffrey. "How to represent part-whole hierarchies in a neural network." arXiv preprint arXiv:2102.12627 (2021).

[8] Yann LeCun, DS-GA 1008, DEEP LEARNING, NYU CENTER FOR DATA SCIENCE

[9] Russo, Abigail A., et al. "Neural trajectories in the supplementary motor area and motor cortex exhibit distinct geometries, compatible with different classes of computation." Neuron 107.4 (2020): 745-758.

[10] Farolfi, A., et al. "Observation of magnetic solitons in two-component Bose-Einstein condensates." Physical Review Letters 125.3 (2020): 030401.

[11] Denschlag, J., et al. "Generating solitons by phase engineering of a Bose-Einstein condensate." Science 287.5450 (2000): 97-101.