RESEARCH STATEMENT

Vedant Sachdeva (sachdved@gmail.com)

Abstract

Both artificial and biological systems leverage the data they gather from external environments to learn signals and strategies that optimize particular objective functions. My research has revolved around understanding how such learning in biological systems occurs when the external environment is time-varying at timescales equal to or faster than the rate of evolutionary adaptation. The key result of my work has been that when external environments vary in predictable ways, evolution drives organisms towards predictive strategies, as such strategies increase the probability of survival of an organism. I demonstrated this evolved predictive capacity on both theoretical models of the visual and immune system as well as on real data from HIV sequences. After my doctorate work, I worked on developing generative models in protein biology to enable the production of synthetic biologic compounds. In my work, I explored diffusion models, variational autoencoders, and transformers as generative models. In particular, I explored how such models represented sequences, when statistical signals of different scales were present. This work offered me insight into the emergence of versions of Simpson's Paradox in protein data. Together, these works build towards a theory for how to learn both long-range global variation, by training on time-varying data, and short-range local variation, by fine-tuning a globally-trained model.

Going forward, I am interested in building theories of biological learning, with a particular focus on continual training and rapid adaptation to novel stimuli. The end goal of my work would be to develop both theory and practical machine learning advances that can be rapidly trained on small amounts of data and demonstrate how local, tractable learning rules can achieve complex computations. I would like to pursue this work as a joint fellow between the Center for Computational Mathematics and the Center for Computational Neuroscience. In particular, I would like to work with Dr. Dimitri Chklovskii to explore biologically plausible models of variational inference, and Dr. Alberto Bietti, to explore theoretical foundations of deep learning.

Previous and Current Work

Predictive computation has been observed in both molecular and neural systems, usually emerging as a consequence of evolutionary pressures that change at or more rapidly than the timescale of evolution. Such predictive computation can emerge in the context of a prediction in a dynamical system but also in the context of generalization, as in the case of broadly neutralizing antibodies. Prior studies of evolution focused more heavily on contexts where environmental pressures changed slower than adaptation timescales, so populations were capable of becoming extremely well-adapted to their current environments. However, adapting well to current environments may come at the cost of handling sharp changes. In my work, we explored the tension between specialist and generalist strategies as a function of the rate of environmental variation. We hypothesized that while task specific specialist strategies offer the highest possible fitness at a given point in time, there may exist generalist strategies that, despite offering a lower fitness at a given point in time, when integrated over longer timescales, provide a larger overall fitness. Examples of the emergence of such strategies can be seen in the adaptive immune system, where broadly neutralizing antibodies against a class of viruses offer long-term fitness benefits, despite not binding any individual strain of said virus particularly well [1, 2], and in the visual system, where retinal ganglion cells encode information about the external environment in a manner that is optimal for predictive computation[3, 4].

Generalization in the Adaptive Immune System

The work described here can be found in its entirety at [5].

We consider affinity maturation in the adaptive immune system as a model system for evolution. In affinity maturation, germline B-cells are recruited to a germinal center and induced to undergo somatic hypermutation, where they rapidly mutate and reproduce, exploring the space of antigen binders. They are then exposed to antigens via a folicular dendritic cells, and depending upon the B-cells' ability to bind the presented antigen, they internalize it. A selective signal is then sent to all B-cells that induces apoptosis in B-cells that have not internalized enough antigen. This cycle of mutation, reproduction, and selection is repeated until either the

germinal center collapses or the population of B-cells is sufficiently saturated. When affinity maturing against rapidly varying viruses, germinal center collapse occurs often, but in some cases, B-cells are able to evolve broadly neutralizing antibodies that generalize against the full virus family.

To explore this generalization in this system, we developed an agent-based model in which B-cells are described by Ising spin vectors, and their ability to bind antigen is described by a fitness "seascape" which fluctuates over time. To construct generalist and specialist strategies, we impose that some aspects of this seascape are fixed in time, while others vary. High fitness states coming as a consequence of the time-varying aspects of the fitness seascape can be thought of as specialist strategies, while high fitness states emerging from binding to the non-time-varying aspect of the seascape are generalist strategies. By rule, the fitnesses of the generalist strategies are lower than the fitnesses of the specialist strategies at any given moment in time. This is to mirror phenomenological observations in the antibody space. We demonstrate that B-cells stabilize to generalists when the fitness seascape varies at timescales near the rate of adaptation of the B-cells. If the fitness seascape varies too rapidly, we observe population extinction - often referred to as germinal center collapse or frustration - and if it varies too slowly, specialist strategies become dominant, as the population has time to fully equilibrate.

Further work in this vein explores the idea of evolvability, where we attempt to experimentally demonstrate regimes where it may be favorable to pay a short-term cost in order to increase the probability of having offspring who have a fitness benefit. We explored this in the context of polymerases, asking when it was favorable to prefer higher mutation rates, even at the cost of lower copying efficiency. In our work, to be published, we demonstrate experimentally that there are regimes where certain selective pressures favor higher mutation rates. Intuitively, this is described as a regime where selection pressure is not so high that the short-term cost dominates but not so low that the long-term benefit is not useful. We further demonstrate that for some three polymerases, there exist selection presssures for which no single polymerase is consistently likely to dominate. This challenges conventional notions of fitness.

Efficient Sensory Encoding Schemes

Part of the work described here can be found at [6]. Another part is described in Chapter 5 of my thesis, found at [7].

The efficient coding hypothesis proposes that neural systems encode information in a manner that maximizes the amount of information about a particular task. It was later proposed that the particular task ought to be prediction, in order to enable biological organisms to respond to the environment optimally, in spite of a delay between sensory processing and motor response. In this framework, we hypothesize that biological organisms sensory evolve encoding schemes, Z_t, to describe a dynamical system, X_t at time t such that Z_t is maximally informative of $X_{t+\Delta t}$ for a given amount of information about X_t. This problem can be formalized as the optimization of the following expression:

$$\max_{p(z_t|x_t)} I(X_{t+\Delta t}; Z_t) - \beta I(X_t; Z_t). \tag{1}$$

This is the prototypical information bottleneck (IB) objective.

Using this objective function, I provide mathematical predictions for the optimal sensory encoding scheme for prediction in both visual and immune systems. For the visual system, we propose that the visual world evolves according to generalized Langevin dynamics. Such dynamics offer us the ability to analytically calculate the optimal sensory encoding scheme. We predict the optimal sensory encoding scheme for a range of different outcomes in these Langevin dynamics, including extending the results to Langevin dynamics with power law correlations in the data. This extends the results of these predictions to include more realistic stimuli, as natural stimuli also have power law correlations.

We then extend our results to consider feedback loops. While we do not consider the optimal control problem, we do ask the question - how should sensory encoding schemes evolve to accommodate for the fact that we are making repeated observations, and updating our observations with both our new observation and knowledge of the dynamical system. This is tantamount to asking the question, given a Kalman Filter with a measurement model of fixed power and a known dynamical process, what is the optimal measurement model? For Gaussian processes, this question can be formulated using a modification to the information bottleneck objective function.

$$\max_{f(\hat{x}_{t-\Delta t}, z_t), p(z_t|x_t} I(X_t; \hat{X}_t) - \beta I(X_t; Z_t) \tag{2}$$

This follows from a mutual information-based formulation of Kalman filters[8]. Here, $f(\hat{x}_{t-\Delta t}, z_t)$ is the function that implements the Kalman innovation, and $\hat{X}_t$ is the estimate of the system. Under this method, we can demonstrate the optimal encoding scheme for a given cost of sensory encoding and the amount of noise in the dynamical process. In each case we assume the dynamical system is known to the biological system. A question for how the dynamical system can be learned has been addressed in [9].

Simpson's Paradox in Generative Models

The notes associated with the work described here can be found at [12].

Generative models, such as variational auto-encoders and diffusion models, can be trained on protein sequence data to generate new examples of given proteins. Such models are currently being used in the construction of a number of biological compounds, ranging from antibodies to enzymes. A core challenge, however, is explicitly tuning the phenotype of a protein sequence in the generation process of these models. Some proposals involve using latent space representations of protein sequence space and performing Bayesian optimization against models of phenotype as a means of sampling the space. Other methods involve training with explicit labels, to drive conditional generation of sequences later. In either case, both revolve around the idea that latent space can be correlated to the phenotype of a protein, either explicitly or implicitly.

In our work, we demonstrate that a potential pitfall of training on many sequences is that the sequence space can become locally distorted. Taking SH3, a ligand-binding domain, as an example, we observe that training a variational autoencoder (VAE) results in the sequences naturally becoming clustered by the ligand they bind. Such a result readily enables conditional generation of examples of a given ligand binder, but does not readily afford resolution into the reaction rate of the ligand binding. This is to say that by training on the larger family of SH3 domains, we can coarsely classify ligand binders, we cannot control the quality of the ligand binder. We hypothesize that while the features that differentiate the different classes of ligand binders were global, each class evolved under its own selection functions. Thus, the scale at which different classes of ligand binders can be distinguished differs from the scale that describes the variation within a given class of ligand binder.

This hypothesis naturally suggests that a fine-tuned model, narrowly focused on a single class of ligand binders will outperform the globally trained model, implying a sort of Simpson's paradox in the data. To explicitly test the emergence of this paradox, we trained a model on a single class of ligand-binders, as identified by locality in VAE latent space. We then determined that a regressor trained to predict binding affinity for a particular ligand is better when using a locally selected sequence set, rather than the full family. The data for this test emerges from an experimental assay that links the growth rate of yeast cells to the ability of designed SH3 domains to bind a peptide ligand, PBS2. In wild-type, SH3 binding (of the protein SHO1) to PBS2 activates osmotic stress pathway which enables yeast to grow in high osmotic environments. This result challenges notions that large corpuses of data is necessarily better for training deep learning models for generative tasks.

Proposed Work

In order to explore how learning can occur under constraints such as limited data or local learning rules, I am proposing two projects. The first project explores how a variational free energy function based on information theoretic quantities could be learned under biologically plausible learning rules. This project would expand previous theories of how the brain implements optimal control under feedback and would connect ideas from the efficient coding hypothesis to biologically plausible learning paradigms. This provides a hypothesis for how the brain can both process and enact behavior in response to environmental stimuli. The second project explores which statistical moments of data are captured by deep learning models. This result could demonstrate which moments in a model change when fine-tuned, show how a large pretrained model can be adapted to specific tasks with just a few examples. It could also offer an approach to knowledge distillation from a large model into a small model without training.

Biologically Plausible Models for Non-Linear Optimal Control

A significant fraction of my work has explored the conditions under which predictive computation can be evolved by biological systems. This work has revolved around the idea that the brain learns encoding schemes that maximize mutual information between its representation and a particular target - usually the future - subject

to constraints. Organisms use these sensory encoding schemes to drive decision making in a continuous online fashion, where they sense an object, enact a behavior, and then sense the object again to decide the next behavior. This cycle of control and feedback has been explored under the context of a Kalman filter[9], demonstrating how both a sensory encoding scheme and a behavioral strategy can be leveraged to maximize reward. However, this work presents a linear theory, which is an information theoretic-optimal approach for only Gaussian statistics. I would like to generalize this work by formulating an optimal control objective function with information theoretic quantities that can be made biologically plausible. I would demonstrate my results on the source-search problem and attempt to implement infotaxis[11] in this setting.

Biologically plausible learning rules are, however, not easily achieved. In order to achieve these rules, I will begin by formulating the objective function as a variational free energy function and reformulating the architecture of the learning as a Markov Decision Process[10]. Previous work has demonstrated that using this approach, learning can be made to be biologically plausible. We could how these rules work on the source-search problem. In particular, we could demonstrate how the prior estimate on the location of a source changes as the organism enacts behavior to go to a new state and senses its new environment. A key result from this study would first be a theoretical prediction that could be compared to activation of neurons in moths as moths execute their searches for odors. A second key result would be the demonstration of how a network can, in an online manner, estimate entropy and potentially, mutual information with relevant targets, even in cases where the environmental statistics are non-Gaussian.

A part of the results given here would also be the demonstration of how objective functions with information theoretic quantities can be optimized in a biologically plausible. This expands the family of possible objective functions significantly, and could be used to explore other problems, such as the decomposition of non-linear mixed sources may become tractable in this setting, when cast as a problem of min-maxing mutual information between different dimensions of some latent representation. In addition, achieving such a representation could also enable me to explain how the complex computations and learned features achieved by modern AI architectures can be achieved by the brain, despite its constraints.

Activation Functions as Moment-Generating Functions

I have begun to work on this using a slightly different approach then the one outlined below. See [13] for details. The appeal of non-linear neural network models is their ability to approximate any function, enabling them to flexibly explain any data set on which they are trained. While the Universal Approximation Theorem serves as proof of this statement, I offer that a useful alternative view of explaining this behavior is that neural network models are capable of learning all statistical moments of the data. Intuitively, this view naturally follows from the idea that all functions can be approximated by neural networks and by taking this perspective, we can now ask questions about precisely which statistical moments of data are learned. By asking questions about which statistical moments are learned, we can begin to build simple effective theories of data by distilling down the model captured by a neural network to focus solely on the important statistical moments.

I propose that we can explore the learned statistical moments by Taylor expanding the activation functions around their point of highest curvature, enabling us to express a sophisticated network with many nodes and layers as a polynomial. By approximating a neural network as a polynomial, we can then analytically express the optimal parameters of the model in terms of tractable moments of data. Further, by recalculating such parameters on bootstrap aggregated (bagged) samples of the data, we can obtain confidence intervals on the value of each parameter, enabling us to determine the statistical significance of the various parts of our model. This offers us insight into not only the importance of each moment, but the model's confidence in its estimate of each parameter, as a function of the variance of each moment. Using these results, we can now re-express our model with an effective theory that focuses solely on the most important moments, dramatically reducing the size of the network. In addition, by developing an effective theory, we also can rapidly learn and adapt to new datasets drawn from similar distributions, as we no longer need to fit all moments of the data.

By approximating neural networks with polynomials, we also readily can explain how knowledge distillation works in deep learning. Typically, in knowledge distillation, a small, simple model is trained by a larger general purpose model. In the proposed framework here, however, instead of needing to train a simple model, the smaller model can be readily read out from the Taylor approximation of the large model. To test this, we can compare the performance of the effective theory derived from a large model to a knowledge distilled model, and determine

the difference between the effective theory and the inferred moments of the distilled model.

Conclusion

Altogether, the past work and proposed work contribute to a theory of learning in artificial and biological systems. The results here explore how the brain can learn non-linear tasks that involve integrating both feedback from the environment and sensory information, and they explore how a neural network learns the various moments of the data it is presented. These results could be applied to a host of different problems, ranging from understanding what correlations amongst sequence data are used for certain predictive tasks, to what correlation structure about the environment is needed for an organism to effectively learn and navigate it.

References

[1] S. Wang, A. Chakraborty et. al., "Manipulating the Selection Forces during Affinity Maturation to Generate Cross-Reactive HIV Antibodies", *Cell*, Feb. 2015.

[2] D.R. Burton, L. Hangartner, "Broadly Neutralizing Antibodies to HIV and Their Role in Vaccine Design", *Annual Review of Immunology*, May 2016.

[3] M. Chalk, O. Marre, G. Tkacik. "Toward a unified theory of efficient, predictive, and sparse coding" "Routers with a Single Stage of Buffering", *Proc. Nat. Acad. of Sci.*, Dec. 2017.

[4] S.E. Palmer, O. Marre, M.J. Berry II, and W. Bialek, "Predictive information in a sensory population", *Proc. Nat. Acad. of Sci.*, May 2015.

[5] V. Sachdeva, K.B. Husain, S. Wang, A. Murugan, et. al., "Tuning environmental timescales to evolve and maintain generalists", *Proc. Nat. Acad. of Sci.*, Jun. 2020.

[6] V. Sachdeva, T. Mora, A. Walczak, S.E. Palmer, "Optimal prediction with resource constraints using the information bottleneck", *PLOS Computational Biology*, Aug. 2021.

[7] V. Sachdeva "Predictive Strategies in Time-Varying Environments", *University of Chicago*, Jun. 2022.

[8] Y. Tomita, S. Ohmatsu, T. Soeda, "An application of the information theory to estimation problems", *Information and Control*, Oct. 1976.

[9] J. Friedrich, D. Chklovskii et. al. "Neural Optimal Feedback Controls with Local Learning Rules", *NeurIPS*, 2021.

[10] T. Isomura, H. Shimazaki, K.J. Friston, "Canonical Neural Networks Perform Active Inference", *Nature Communications Biology*, Jan. 2022.

[11] M. Vergassola, E. Villermaux, B. I. Shraiman, "'Infotaxis' as a strategy for searching without gradients", *Nature*, Jan. 2007.

[12] U. Pen, V. Sachdeva, M. Mani, "Separation of Scales in the Protein Universe", *Unpublished Notes*. Link to Github here

[13] U. Pen, V. Sachdeva, "Activation Function Analysis", *Unpublished Notes*. Link to Github here