RESEARCH STATEMENT OF GEOFF PLEISS

Machine learning has produced extraordinary predictive capabilities, but the success of these **powerful** models is often limited to accuracy on stationary data. The next decade will increasingly require predictive models as building blocks of larger decision-making pipelines, where they will be exposed to nonstationary data, and errors from models will propagate downstream. Such settings require the capacity for **general reasoning**, including 1) *inductive biases* that capture what is known beyond observed data (i.e. how to extrapolate to unseen data), and 2) *uncertainty estimates* that quantify what is unknown due to limited observations. Combining today's modeling power—typically the domain of neural networks—with suitable inductive biases and uncertainty quantification—typically the domain of probabilistic models—is both timely and necessary for the next evolution of machine learning capabilities. The primary technical challenges are adapting the reasoning mechanisms of probabilistic models to work with neural networks, while reducing their computational overhead so such mechanisms are practically viable.

My research places me at the nexus of *deep learning*, *probabilistic modeling*, and *numerical linear algebra*, enabling me to address both of these challenges. One line of my work focuses directly on neural networks, improving their uncertainty estimates while understanding their predictive capabilities through the lens of probabilistic models. Another line focuses on the inductive biases of Gaussian processes (GP), improving their computational efficiency and ultimately replicating their desirable properties in neural networks. This research profile ideally situates me to unite these paradigms, transforming today's powerful models into general reasoning models. In addition, I have a proven record of coupling my findings with *performant and easy-to-use software* used widely throughout research and industry, facilitating adoption and innovation in this area.

PRIOR AND CURRENT WORK

Calibrating the uncertainty estimates of neural networks. One necessary prerequisite of general reasoning is the ability to communicate predictive uncertainty. While the representational capacity of neural networks makes them ideal for complex predictive tasks, this power also makes them susceptible to learning spurious features that yield wildly overconfident predictions unrepresentative of true probabilities. A large focus of my Ph.D. research was reducing spurious extrapolations and improving uncertainty estimates, laying the groundwork for future reasoning. My work showed that existing calibration techniques (e.g. Platt scaling [9] or isotonic regression [20]) essentially perform a weighted average of the prediction against the uniform distribution. Distilling this idea to its logical extreme leads to the temperature scaling method [4], which accomplishes this averaging through a constant scaling of network outputs. Despite (or perhaps because of) its simplicity, this method often achieves near perfect calibration, enabling neural networks to output true probabilities. My work also addresses identifying corrupted data and spurious features responsible for poor extrapolation. Exploiting the "simplicity bias" of stochastic optimizers [1], I demonstrate that a running average of the sample margin during training—the "area under the margin" statistic [14]—identifies low-likelihood data with high precision. Both methods create *a basis for future reasoning capabilities*: equipping neural networks with calibrated uncertainty estimates and the ability to identify sources of spurious extrapolation.

Scaling the inductive biases of Gaussian processes. Beyond calibrated outputs, general reasoning requires explicitly dictating how a model should extrapolate in regions with limited observations. Gaussian processes offer an inductive biases with this desirable property: extrapolation is specified

through an interpretable language of prior covariance functions. However, their asymptotic computational complexity has led to a common belief that GP cannot scale beyond a few thousand data points. My dissertation work demonstrates that GP can be *viable on millions of data points* as long as the underlying computations effectively utilize modern compute hardware. In line with the key tenet of my research, the solution lies *at the nexus of GP and neural networks*, as the latter are similarly computationally intensive but use modern hardware to great effect. Building on prior work that investigates similar ideas [e.g. 2], I comprehensively overhaul all components of GP inference in a series of papers [3, 12, 13, 15, 19], replacing standard computations with algorithms that solely rely on matrix-vector multiplication (the computational primitive of neural networks). This approach—based on numerical techniques like Krylov subspace methods—improves wall-clock time by orders of magnitude, as matrix multiplication is extremely amenable to GPU acceleration.

Of course, unlocking the potential of Gaussian processes requires more than computational efficiency. Again, there is opportunity to apply insights from deep learning, recognizing that software frameworks like TensorFlow and PyTorch underpin the innovation and widespread adoption of neural networks. To that end, I cofounded and maintain the GPyTorch project, with the mission of translating these often inaccessible GP inference algorithms into professional easy-to-use software. Thanks to our work and numerous contributors, it has become the de facto standard for GP research and industry, used in over 300 scientific papers and in industrial applications at companies like Facebook, JP Morgan, and Mars.

Improving neural network inductive biases via Gaussian processes. With this new-found computational efficiency, I aim to adapt the desirable properties of GP inductive biases to neural networks (see "Future Directions"). However, improving neural networks first requires *understanding the limitations of their current inductive biases*, and again my work demonstrates that key insights lie at the intersection of neural networks and GP. One example is my research on how width impacts the neural network inductive bias [10]. The insight of this work was recognizing that Deep Gaussian Processes (DGP)—hierarchical models where layers are given by vector-valued GP—are ideal models for analyzing neural networks. DGP are in fact a superclass of neural networks, yet they are often easier to analyze since their building blocks (GP) have well understood inductive biases. My work uncovers pathological behavior of wide DGP, which collapse to shallower models as width increases. DGP also offer a new formalized interpretation of "representation learning," and I theoretically demonstrate that this ability to learn representations diminishes as width increases. Importantly, these findings are strongly predictive of similar trends in neural networks, offering a new understanding about the properties of their inductive biases.

FUTURE DIRECTIONS

My prior work addresses prerequisites for general reasoning: ensuring that neural networks can communicate uncertainty estimates, and that the inductive biases of Gaussian processes scale to large datasets. With this foundation, my future work will further combine the abilities of neural networks and probabilistic methods towards the goal of powerful general reasoning models.

Inductive biases of neural networks. Having identified limitations of neural network inductive biases [e.g. 6], I aim to apply ideas and mechanisms from Gaussian processes to make these models more capable of general reasoning. Specifically, my research will focus on implementing three properties of GP inductive biases: "*failing gracefully*" in regions of limited data, *identifying out-of-distribution* (OOD) data, and *adapting to different data modalities*. Because it is challenging to extrapolate from limited observations, the fist criterion stipulates that a model should "fail

gracefully" by falling back to some sensible default prediction. Here I draw direct inspiration from GP, which revert to average-case predictions when data are weakly correlated with other observations (as specified by the covariance prior). I am currently developing neural network parameterizations that mimic this behavior, modifying the initialization and regularization of each layer to produce piecewise linear approximations of GP with specified covariances. The goal is an architecture-agnostic framework where this mean-reverting mechanism can be applied to convolutional networks, transformers, and other networks. The second criterion aims to prevent models from making predictions on nonsensical or sufficiently OOD data. In preliminary work, I demonstrate that a model must be identifiable to accurately detect OOD data. Since neural networks are generally non-identifiable, I am developing a locally linear approximation that characterizes the set of observationally equivalent neural networks, from which it is also possible to detect OOD data. A more general solution will build upon recent developments in identifiable overparameterized models [16]. The third criterion ensures that inductive biases can be specialized to a given data domain, just as GP covariance priors can be specialized through hyperparameter optimization. In the spirit of empirical Bayes, I envision that architectural components (e.g. the activation function) can be optimized using randomly-initialized models, yielding maximum likelihood architectures for a given dataset. Success in these directions will lead to inductive biases that encode knowledge beyond observed data, bringing us closer to the goal of general reasoning in neural networks.

Uncertainty quantification in deep learning. General reasoning also requires that models *communicate lack of knowledge* through estimates of uncertainty. My prior work [4] was a necessary first step towards ensuring that neural networks output semantically meaningful probabilities. The next challenge is *accurately quantifying all possible sources of uncertainty*—ranging from a lack of data to irreducible "noise" in the features—while ensuring that this uncertainty quantification (UQ) is *computationally efficient*. Current UQ methods often require ensembles of multiple neural networks, which are highly impractical for many large-scale problems. My prior work addresses the training-time cost of ensembles, generating so-called "Snapshot Ensembles" from the optimization trajectory of a single network [7]. The next step is to develop approximations of ensembles that only require a single forward pass during evaluation, building on the work of [5, 8]. Another promising direction is dynamically trading off UQ and computation, drawing upon techniques from the cascade literature [18]. Additionally, my research will focus on *characterizing differences between various UQ mechanisms*. Ensembles, variational methods, and other UQ mechanisms produce different approximations of predictive uncertainty, and each are susceptible to miscalibration in different ways. From a theoretical perspective, I aim to derive the "effective priors" that produce the predictive distributions of each UQ mechanism. In conjunction with my other line of work, I will also investigate how each UQ method responds to misspecified inductive biases, building on techniques from the robust statistics literature. Success in these directions will yield general reasoning models which can be composed into pipelines that reliably propagate uncertainty.

Tradeoffs of machine learning objectives. Beyond general reasoning, the machine learning community has proposed many other criteria for predictive models, ranging from robustness against adversarial attacks to ensuring equitable outcomes for different subpopulations. It is crucial to understand if these objectives are complementary, or—more likely—if they form a Pareto frontier, whereby improving one objective requires sacrificing another [e.g. 17]. For example, I have investigated an *unavoidable tradeoff between uncertainty quantification and various notions of fairness* [11], in which calibrated uncertainty estimates inadvertently create predictive disparities between different groups. I will lead a line of research that aims to identify the tradeoffs inherent to the inductive biases and uncertainties necessary for general reasoning. For example, my work on wide neural networks [10] establishes that neural network posteriors, unlike those of GP, are drawn from a

data-dependent (and thus adaptable) space of functions. It is entirely possible that inductive biases designed for OOD inputs or for "failing gracefully" may limit this adaptability and harm accuracy. Such outcomes, though undesirable, *inform practitioners about the fundamental limitations of machine learning methods*, and enable responsible decisions regarding their practical consequences.

Conclusion. General reasoning combines the capabilities of probabilistic models (inductive biases and uncertainty quantification) with the power of neural networks, all while maintaining computational efficiency. I identify several open directions in this problem space, and my prior work and skill set situate me to make meaningful progress towards this next generation of machine learning.

REFERENCES

[1] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, 2017.

[2] Kurt Cutajar, Michael Osborne, John P. Cunningham, and Maurizio Filippone. Preconditioning kernel matrices. In *International Conference on Machine Learning*, 2016.

[3] Jacob R. Gardner, **Geoff Pleiss**, David Bindel, Kilian Q. Weinberger, and Andrew Gordon Wilson. GPyTorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In *Neural Information Processing Systems*, 2018.

[4] Chuan Guo, **Geoff Pleiss**, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

[5] Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Zhe Liu, Jasper Snoek, Balaji Lakshminarayanan, Andrew M. Dai, and Dustin Tran. Training independent subnetworks for robust prediction. In *International Conference on Learning Representations*, 2021.

[6] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Computer Vision and Pattern Recognition*, 2019.

[7] Gao Huang, Yixuan Li, **Geoff Pleiss**, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get m for free. In *International Conference on Learning Representations*, 2017.

[8] Jeremiah Z. Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In *Neural Information Processing Systems*, 2020.

[9] John Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74, 1999.

[10] **Geoff Pleiss** and John P. Cunningham. The limitations of large width in neural networks: A deep Gaussian process perspective. In *Neural Information Processing Systems*, 2021.

[11] **Geoff Pleiss**, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. In *Neural Information Processing Systems*, 2017.

[12] **Geoff Pleiss**, Jacob R. Gardner, Kilian Q. Weinberger, and Andrew Gordon Wilson. Constant-time predictive distributions for Gaussian processes. In *International Conference on Machine Learning*, 2018.

[13] **Geoff Pleiss**, Martin Jankowiak, David Eriksson, Anil Damle, and Jacob R. Gardner. Fast matrix square roots with applications to Gaussian processes and Bayesian optimization. In *Neural Information Processing Systems*, 2020.

[14] **Geoff Pleiss**, Tianyi Zhang, Ethan R. Elenberg, and Kilian Q. Weinberger. Identifying mislabeled data using the area under the margin ranking. In *Neural Information Processing Systems*, 2020.

[15] Andres Potapczynski, Luhuan Wu, Dan Biderman, **Geoff Pleiss**, and John P. Cunningham. Bias-free scalable Gaussian processes via randomized truncations. In *International Conference on Machine Learning*, 2021.

[16] Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, 2021.

[17] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

[18] Paul Viola and Michael Jones. Robust real-time object detection. *International journal of computer vision*, 4(34-47):4, 2001.

[19] Ke Alexander Wang, **Geoff Pleiss**, Jacob R. Gardner, Stephen Tyree, Kilian Q. Weinberger, and Andrew Gordon Wilson. Exact gaussian processes on a million data points. In *Neural Information Processing Systems*, 2019.

[20] Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Knowledge Discovery and Data Mining*, 2002.