RESEARCH PROPOSAL OF GEOFF PLEISS

The next decade will increasingly rely on predictive modeling for scientific discovery, decision making, and safety-critical forecasting. Future success in these settings will require predictive models that are not only *powerful*—capable of processing large quantities of complex data—but also *reliable*—capable of codifying expert knowledge (e.g. how to extrapolate beyond observed data) and quantifying what is unknown due to limited observations. Unfortunately, today's machine learning models are often incapable of satisfying both desiderata. For example, the predictive power of *neural networks* often comes at the cost of erratic extrapolation and spurious correlation. Conversely, the inductive biases and uncertainty quantification afforded by *probabilistic models* are rarely suited for the scale and complexity of modern datasets. My research aims to eliminate this tradeoff; creating *powerful and reliable* predictive models. Specifically, I will develop models for three key problems: 1) large-scale spatio-temporal modeling, for inferences in the physical sciences; 2) uncertainty-aware blackbox optimization, for decision making in engineering; and 3) robust neural network ensembling, for forecasts in safety-critical settings.

1) LARGE-SCALE SPATIO-TEMPORAL MODELING

Many problems throughout the physical sciences involve inferring some latent quantity from collections of spatially-correlated signals or time series. A illustration from astronomy is locating "dust" in the Milky Way from observations of starlight extinctions [41]. Probabilistic models, especially Gaussian processes (GPs), have historically been used for such inferences in geostatistics [22], neuroscience [8], and epidemiology [34], in large part because of their formalism for incorporating prior domain knowledge and quantifying uncertainty. Unfortunately, GPs are often ill-suited for modern datasets: lacking both 1) the ability to scale to large datasets and 2) the ability to readily model non-stationary and high-dimensional data. My prior research has addressed the first challenge. In a series of papers [15, 25, 26, 38], I revamped the numerical algorithms underpinning GP inference, utilizing memory-efficient iterative methods that are extremely amenable to GPU acceleration. My approach enabled exact GP inference on millions of data points [36]—a significant two-orders-of-magnitude improvement over prior work [9]. With this scalability in place, my current and future research aims to address the second challenge of modeling complex phenomena. In pursuit of this goal, I am focusing on two key directions:

Relaxing invariance assumptions. Many spatio-temporal GPs assume invariances, such as stationarity, which are often violated by naturally-occurring discontinuities in large-scale and high-dimensional datasets. Rather than completely abandoning such invariance assumptions (as proposed in [e.g. 29]), I propose the development of *approximately-invariant GPs*: models biased towards invariances but flexible enough to model discontinuities. In recent theoretical work [24], I demonstrate that deep (hierarchical) GP models encode a relaxed notion of stationarity, with the width of hidden layers controlling the degree of relaxation. The next major undertaking is to make these models practical, using an algorithmic approach for designing approximately-invariant versions of simple GPs, alongside further theoretical work to characterize their functional properties.

Simplifying optimization dynamics. Model selection and approximate inference with spatio-temporal GPs is typically cast as gradient-based optimization, enabling highly parametric priors (e.g. spectral kernels [39] or deep kernels [40]) capable of extrapolating on complex data modalities. However, the optimization dynamics associated with these tasks are often quite unfavorable. Pathologies even emerge for model selection in Bayesian linear regression [32], and the optimization

dynamics become increasingly worse for more complex GP models [21]. Alleviating these barriers is necessary to make these powerful models simpler and more accessible to non-expert practitioners. To that end, I have begun developing theoretical analyses of gradient-based model selection for GPs [28, 37]. Utilizing these findings, I aim to redesign spectral kernels and deep kernels to improve optimization dynamics, reducing the need for careful initializations and hand-tuned procedures. Further extending this theory to hierarchical GP models—in which model selection and approximate inference are often performed simultaneously—will greatly extend the predictive capabilities and usability of spatio-temporal models.

2) UNCERTAINTY-AWARE BLACKBOX OPTIMIZATION

Design problems throughout science and engineering often cannot be translated into mathematical models, making them ill-suited for traditional optimization techniques. Bayesian optimization (BO) is a powerful method whereby these blackbox functions are modelled by probabilistic surrogates (typically GPs) and the surrogates' *posterior uncertainty* guides the exploration-vs-exploitation tradeoff. With recent advances in high-dimensional techniques [e.g. 11, 13] and GP scalability (see Sec. 1 for examples from my prior research), there is growing interest in applying BO in challenging applications like drug discovery [31] and nuclear fusion control [7]. However, existing BO pipelines neglect several sources of modeling error, which—although relatively inconsequential in small-scale BO problems—significantly hinder optimization in complex and large-scale settings [19, 25]. Building upon my ongoing work [26, 38], I aim to represent these modeling errors as *additional sources of uncertainty* that inform the exploration-vs-exploitation tradeoff:

Optimizing with computational uncertainty. The BO workflow often relies numerous computational shortcuts to maintain tractability, such as 1) ignoring the influence of GP hyperparameters, 2) restricting the search space to a small discrete set of candidate points, and/or 3) utilizing sparse approximations of the GP posterior [20]. As problem complexity increases, these shortcuts become more necessary but can lead to erroneous solutions if not accounted for in the modeling process [25, 26]. To mitigate this issues, I have begun developing a framework of *computational uncertainty* that captures approximation error resulting from limited computation (just as *posterior uncertainty* captures modeling error due to limited data). In preliminary work, I introduce a family of GP approximations where the resulting predictive distribution incorporates both sources of uncertainty [38]. I aim to extend this notion of computational uncertainty to all components of the BO pipeline, including hyperparameter optimization and candidate point selection. Of course, computational uncertainty is only beneficial if it is incorporated into the optimization procedure; therefore, I will pursue *computation-aware search policies* where this uncertainty is used to enhance optimization.

Optimizing with objective uncertainty. In complex design problems (e.g. drug discovery), it is often necessary to use imperfect proxies or surrogate metrics (e.g. computer simulations of molecular binding affinity) as substitute measures for the true desired behavior (e.g. efficacy of the drug). This discrepancy introduces numerous pathologies, such as adversarial solutions that exploit flaws in the proxy metric or solutions that fail to trade off multiple desired objectives. While prior works have proposed one-off strategies to address these issues [e.g. 10, 19], I propose an overarching framework for *optimizing under objective uncertainty*. The framework will treat the true desired outcomes as latent variables under a biased observation model, with search policies that effectively use posterior uncertainty and objective uncertainty (and also computational uncertainty). The mechanisms to quantify objective uncertainty will depend on the available side information, ranging from human-in-the-loop feedback to generative models from related problem areas.

3) ROBUST NEURAL NETWORK ENSEMBLES

As neural networks become prominent in high-stakes applications like autonomous driving, precision medicine, and automated finance, it is crucial to mitigate against catastrophic erroneous predictions. A significant portion of my prior work has focused on addressing the risks of overconfidence [16], observational noise [27], and covariate shift [2]. Of the many proposed approaches, ensembles of neural networks (or approximations thereof) have become the prevailing method to reduce these risks [18, 23, 33], based on long-standing intuitions that the diversity amongst independently-trained models prevents correlated errors [14]. However, my recent research demonstrates that these intuitions—largely derived from ensembles of low-capacity models (e.g. random forests [6])—can be entirely misleading for ensembles of high-capacity neural networks. In particular, ensembles of neural networks improve when *predictive diversity is minimized* [1, 3] and do not provide additional uncertainty/robustness beyond what can be achieved with a standalone (but larger) neural network [2]. These findings expose a chasm between intuition and practice, with immediate implications for uncertainty quantification, robustness, and safety considerations. I am thus pursuing two directions to improve the foundations of neural network ensembles:

Understanding overparameterized ensembles. The success of ensemble methods essentially boils down to variance reduction [5, 12]. At the same time, recent theoretical work on overparameterized neural networks suggests that increasing capacity (e.g. width or depth) also amounts to variance reduction [4] (in contrast to underparameterized models, where additional capacity increases variance). My recent empirical results suggest that both strategies—ensembling and increasing capacity—yield predictions that are almost indistinguishable [2]. This functional similarity between ensembles and single (larger) neural networks—the latter of which are considered unreliable, brittle, and overconfident [23]—suggests that current ensembles are also fundamentally insufficient for safety-critical settings. To support these empirical findings, I am developing theoretical characterizations of ensembling in overparameterized random feature models. This analysis investigates 1) the degree of functional similarity between ensembles and single models and 2) how architectural choices (e.g. width, non-linearity, etc.) influence this similarity. A deeper understanding of the relationship between ensembles and single (larger) models will illuminate to what extent existing ensembles can or cannot offer meaningful uncertainty quantification/robustness.

Ensembling beyond arithmetic averaging. Although current ensembles may be functionally similar to standalone neural networks, there is reason to believe that simple modifications could yield vastly improved predictive capabilities. One underexplored area is the flexibility afforded by the "voting mechanism"—i.e. the reduction that combines multiple models into a single prediction. Currently, it is common to use the arithmetic or geometric mean of component model predictions [17]. However, there exists a wide range of other possible reductions, each with distinct properties. I propose an *objective aware ensembling framework* where, given a target risk function, the optimal reduction for a set of existing pretrained models is selected as a postprocessing step. This framework allows pretrained neural networks to be reused to achieve different desired outcomes (e.g. accuracy, calibration, robustness to outliers, etc.) simply by adjusting the reduction. As an initial step, one could consider a simple parametric family of reductions—e.g. averages defined by simple one-dimensional polynomials—learned through a small hold-out validation set (in the same vein as [16]). To further enhance predictive capabilities, this approach can be extended to *input-dependent* reduction algorithms, drawing from existing work on mixture-of-expert and attention mechanisms [30, 35]. This approach holds the potential for many new ensemble methods that could offer meaningful reliability improvements necessary for safety-critical settings.

REFERENCES

[1] Taiga Abe, E. Kelly Buchanan, **Geoff Pleiss**, and John P. Cunningham. The best deep ensembles sacrifice predictive diversity. In *NeurIPS "I Can't Believe It's Not Better" Workshop*, 2022.

[2] Taiga Abe, E. Kelly Buchanan, **Geoff Pleiss**, Richard Zemel, and John P. Cunningham. Deep ensembles work, but are they necessary? In *Neural Information Processing Systems*, 2022.

[3] Taiga Abe, E. Kelly Buchanan, **Geoff Pleiss**, and John P. Cunningham. Pathologies of predictive diversity in deep ensembles. *Under submission*, 2023.

[4] Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. In *Neural Information Processing Systems*, 2020.

[5] Peter Bartlett, Yoav Freund, Wee Sun Lee, and Robert E. Schapire. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651–1686, 1998.

[6] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[7] Youngseog Chung, Ian Char, Willie Neiswanger, Kirthevasan Kandasamy, Andrew Oakleigh Nelson, Mark D. Boyer, Egemen Kolemen, and Jeff Schneider. Offline contextual Bayesian optimization for nuclear fusion. In *NeurIPS Workshop on Machine Learning and the Physical Sciences*, 2019.

[8] John P Cunningham, Byron M Yu, Krishna V Shenoy, and Maneesh Sahani. Inferring neural firing rates from spike trains using Gaussian processes. In *Neural Information Processing Systems*, 2007.

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

[10] Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Parallel Bayesian optimization of multiple noisy objectives with expected hypervolume improvement. In *Neural Information Processing Systems*, 2021.

[11] Aryan Deshwal and Jana Doppa. Combining latent space and structured kernels for Bayesian optimization over combinatorial spaces. In *Neural Information Processing Systems*, 2021.

[12] Thomas G. Dietterich. Ensemble methods in machine learning. In *Workshop on Multiple Classifier Systems*, 2000.

[13] David Eriksson, Michael Pearce, Jacob R. Gardner, Ryan D. Turner, and Matthias Poloczek. Scalable global optimization via local Bayesian optimization. In *Neural Information Processing Systems*, 2019.

[14] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[15] 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.

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

[17] Neha Gupta, Jamie Smith, Ben Adlam, and Zelda Mariet. Ensembling over classifiers: a bias-variance perspective. *arXiv preprint arXiv:2206.10566*, 2022.

[18] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Neural Information Processing Systems*, 2017.

[19] Natalie Maus, Kaiwen Wu, David Eriksson, and Jacob R. Gardner. Discovering many diverse solutions with Bayesian optimization. In *Artificial Intelligence and Statistics*, 2023.

[20] Henry B. Moss, Sebastian W. Ober, and Victor Picheny. Information-theoretic inducing point placement for high-throughput Bayesian optimisation. In *Artificial Intelligence and Statistics*, 2022.

[21] Sebastian W. Ober, Carl E. Rasmussen, and Mark van der Wilk. The promises and pitfalls of deep kernel learning. In *Uncertainty in Artificial Intelligence*, 2021.

[22] Margaret A Oliver and Richard Webster. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System*, 4(3):313–332, 1990.

[23] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Neural Information Processing Systems*, 2019.

[24] **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.

[25] **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.

[26] **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.

[27] **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.

[28] 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.

[29] Sami Remes, Markus Heinonen, and Samuel Kaski. Non-stationary spectral kernels. In *Neural Information Processing Systems*, 2017.

[30] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

[31] Samuel Stanton, Wesley Maddox, Nate Gruver, Phillip Maffettone, Emily Delaney, Peyton Greenside, and Andrew Gordon Wilson. Accelerating Bayesian optimization for biological sequence design with denoising autoencoders. In *International Conference on Machine Learning*, 2022.

[32] Will Stephenson, Zachary Frangella, Madeleine Udell, and Tamara Broderick. Can we globally optimize cross-validation loss? Quasiconvexity in ridge regression. In *Neural Information Processing Systems*, 2021.

[33] Dustin Tran, Jeremiah Liu, Michael W. Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, and others. Plex: Towards reliability using pretrained large model extensions. *arXiv preprint arXiv:2207.07411*, 2022.

[34] Jarno Vanhatalo and Aki Vehtari. Sparse log Gaussian processes via mcmc for spatial epidemiology. In *Gaussian processes in practice*, 2007.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[36] 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.

[37] Jonathan Wenger, **Geoff Pleiss**, Philipp Hennig, John P. Cunningham, and Jacob R. Gardner. Preconditioning for scalable Gaussian process hyperparameter optimization. In *International Conference on Machine Learning*, 2022.

[38] Jonathan Wenger, **Geoff Pleiss**, Marvin Pfortner, Philipp Hennig, and John P. Cunningham. Posterior and computational uncertainty in Gaussian processes. In *Neural Information Processing Systems*, 2022.

[39] Andrew Gordon Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *International Conference on Machine Learning*, 2013.

[40] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P. Xing. Deep kernel learning. In *Artificial Intelligence and Statistics*, 2016.

[41] Luhuan Wu, Andrew Miller, Lauren Anderson, **Geoff Pleiss**, David Blei, and John P. Cunningham. Hierarchical inducing point Gaussian process for inter-domian observations. In *Artificial Intelligence and Statistics*, 2021.