Graduate Research Plan

Research Title: Symmetry for Understanding Model Calibration

Summary of Graduate Research Plan: As machine learning pipelines become widely adopted in safety-critical and high impact applications, strong reliability measures and uncertainty estimates become exceedingly important. For instance, uncertainty estimates are used in drug discovery as a guard rail against potentially dangerously wrong chemical property predictions. Moreover, some domains such as cosmology use uncertainty measures as a signal for other algorithms like Hamiltonian Monte Carlo (HMC) in order to tighten constraints on physical constants. Beyond these examples, uncertainty quantification shows a pronounced importance in the data-sparse regimes where equivariant neural networks—a class of neural networks that capture symmetry constraints such as rotations and permutations—tend to excel. While we have come to develop a very mature view of the tradeoffs between equivariant and non-equivariant networks [1], most of these tradeoffs are stated for traditional classification and regression tasks where the outputs are deterministic. **Thus, it is paramount to understand the sparsely studied relationship between uncertainty and equivariance.** The objective of my research is to study the loss landscapes and learning behaviors of equivariant neural networks when estimating distributions and to understand if equivariance can make neural networks better calibrated. This work will balance a theoretical case study on small networks that can be characterized completely with an empirical study using epistemic neural networks [2] that fills in the gaps where the theory is less tractable. Staggering the work in this sequential way allows me to first come to an understanding of the important properties of the calibration error loss landscape before studying how these landscapes manifest in observed training dynamics.

Thrust I – Counting the Connected Components under ECE Loss: The first thrust of my research is to characterize the effect of equivariance on the overall loss landscape when trying to learn a categorical distribution. In particular, I will study the loss landscape of the expected calibration error (ECE) [3], which examines the discrepancy between a model's accuracy and predicted confidence. I will take advantage of the work of [4], who offer a rich framework for describing the topology of a neural network's loss landscape. In particular, they characterize the level set of a loss function as a set of disjoint connected components. Connected components are of interest, because it is theorized that having fewer connected components leads to better training dynamics [4]. The main contribution of this thrust will be a theorem that compares the number of connected components in the minimum of the ECE loss between equivariant and non-equivariant feed forward networks (FFNs), and I hypothesize that the equivariant network will have fewer. For the equivariant case, the work will specifically consider FFNs constructed using vector neurons [5] which have SO(3) equivariance under the standard representation. **Evaluation Criterion:** Following [4], I can characterize completely the level sets of small FFNs, enabling me to verify the correctness of my theorem. Even if my hypothesis is incorrect, knowing that equivariance does not aid in taming the loss landscape is still an illuminating result. It might suggest that using equivariant models for uncertainty quantification has no adverse side effects, or even that equivariance can be *harmful* in cases of model misspecification.

Thrust II – Empirical Validation with Epistemic NN's: The goal of this thrust is to find evidence that the difference in connected components between equivariant and unconstrained models established in the previous thrust has a measurable effect on the observed calibration error. To do so, I will use epistemic neural networks as classifiers on challenging image classification benchmarks like ImageNet. These tasks are naturally invariant to rotations of the input data, leading to invariance under SO(2) and the ability to leverage insights gained from the theory of SO(3) from the previous stage. In addition to the

aforementioned benchmarks, I will also collect a series of illuminating and simple adversarial examples. For instance, an SO(2)-invariant model trained on MNIST digit classification would be unable to distinguish between a 6 and a 9 and the network's uncertainty estimate should diverge. Should this fail to happen, it would show that an invariant model can not signal to practitioners that it is over-constrained via the uncertainties in its categorical probabilities. **Evaluation Criterion:** The evaluation of SO(2)-invariant epistemic neural networks naturally extends my previous work [**6**], which placed lower and upper bounds on calibration error for invariant models. Using these bounds will allow me to better interpret the observed ECE and confirm that the conclusions implied by the theorem in the first thrust support the experimental results in this thrust. If the theory and experiments do not agree, this would still open up future work diagnosing the true utility of connected components as a way of understanding loss landscapes.

Intellectual Merit

Symmetry provides for us an uncharted lens under which we can study uncertainty calibration. Calibration loss landscapes are difficult to characterize because they have higher degeneracy, that is, there exist many local minima where a model can be well calibrated even if it performs poorly at prediction in general. Through symmetry, we can potentially reduce these degeneracies and in doing so avoid falling into the local minima in the calibration error landscape when training. In the main, my proposed plan provides a new axis under which these calibration loss landscapes can be understood. *This paves the way for new architectures and training paradigms that can better meet the calibration objectives.* The plan will also clarify when symmetry is overconstraining to the detriment of model calibration and underscore when models with relaxed equivariance constraints are more useful to practitioners. My work is also of independent interest to the mathematics community. My work in **[6]** established new proof strategies for deriving approximation error bounds, and I similarly expect that the strategies used in the first thrust of my research will provide use to mathematicians working in topology, geometry, numerical analysis, and many other fields.

Broader Impacts

This research has the potential to improve the adoption of machine learning in industry by providing practitioners the promise of reliability *before* testing model predictions in real lab settings such as pharmaceutical labs, robotics test beds, and chemical plants. This is because a well calibrated uncertainty estimate can provide a signal that a neural network's output is untrustworthy and therefore not worth the time and money to test by performing a measurement on real data. The proposed research will inform when symmetry serves as a helpful bias for producing well calibrated uncertainties, which in turn, allows for practitioners to use these aforementioned warning signals. Understanding if symmetry provides a useful bias for calibration is also especially important to robotic systems where symmetry is commonly employed. This research will clarify when symmetry can be used to help identify unreliable predictions and thereby help prevent accidents or unsafe decisions.

References: [1] Wang, Dian, et al. "A general theory of correct, incorrect, and extrinsic equivariance." *NeurIPS* 2023. [2] Osband, Ian, et al. "Epistemic neural networks." *NeurIPS* 2023. [3] Guo et al. "On calibration of modern neural networks." *ICML* 2017. [4] Zhao, Bo, et al. "Understanding Mode Connectivity via Parameter Space Symmetry." *NeurIPS* 2023 [5] Deng, Congyue, et al. "Vector neurons: A general framework for SO(3)-equivariant networks." *ICCV* 2021. [**6**] **Berman, Edward**. et al "On Uncertainty Calibration for Equivariant Functions." *Under Review at Transactions on Machine Learning Research (TMLR), NeurReps workshop 2025 at NeurIPS. Preprint on arxiv 2510.21691*