Success of Uncertainty-Aware Deep Models Depends on Data Manifold Geometry

Mark Penrod, Harrison Termotto, Varshini Reddy, Jiayu Yao, Finale Doshi-Velez, Weiwei Pan

HARVARD School of Engineering and Applied Sciences

IACS INSTITUTE FOR APPLIED COMPUTATIONAL SCIENCE AT HARVARD UNIVERSITY

MOTIVATION

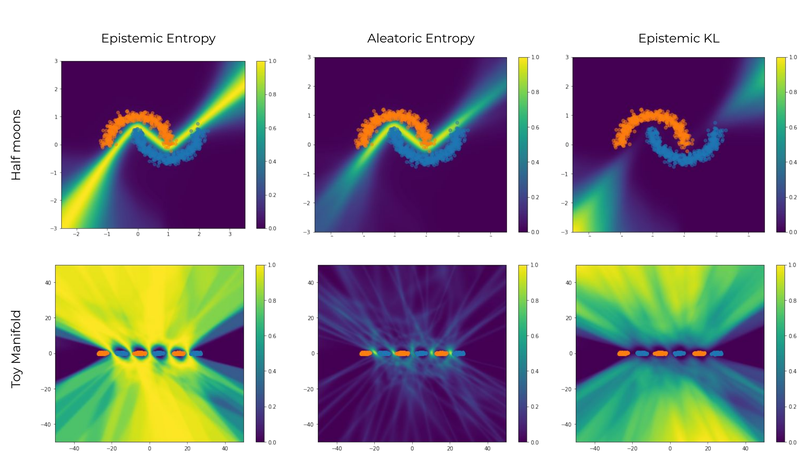
- In risk-sensitive applications, deep learning models must effectively handle edge-case data (i.e. adversaries and OoD data)
- Predictive uncertainty is useful for these tasks, but the choice of uncertainty-aware deep learning model is often unclear
- **Goal**: Compare and analyze the performance of uncertainty-aware models on adversarial robustness and OoD/adversarial detection, considering also varying datasets and uncertainty metrics

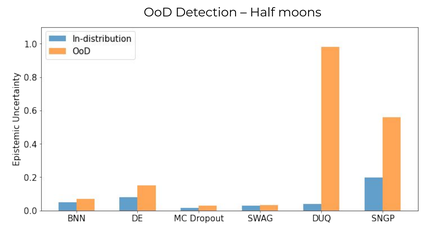
METHODS

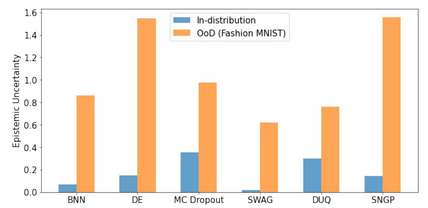
- Set of **Bayesian**, **Frequentist**, and **Deterministic** approaches to uncertainty quantification
- Test with **half moons**, **MNIST**, and **"toy manifold"** data on classification task
- Evaluate uncertainty via **epistemic** and **aleatoric entropy**, as well as **KL-divergence-based metric** (epistemic)

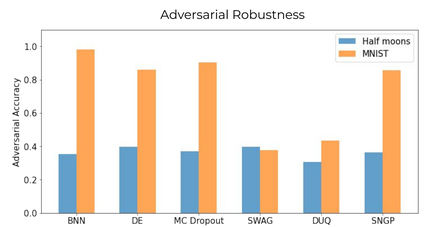
PERFORMANCE DEPENDS ON DATA SUB-MANIFOLD

- In half moons, all models adversarially robust, only some models consistently able to detect OoD; this pattern reverses in MNIST
- Toy manifold dataset mimics natural image manifold and recreates the discrepancies identified in MNIST









UNCERTAINTY METRICS

- For full dimensional data, aleatoric entropy reliably catches adversaries, epistemic measures catch OoD
- For data on a sub-manifold, epistemic measures catch OoD *and* adversaries, aleatoric uncertainty concentrates on manifold

TAKEAWAY

- Model, dataset, uncertainty metric, and task-related desiderata interact in complex ways, demanding nuanced and holistic of all factors during deployment