

Uncertainty of Network Topology with Applications to Out-of-Distribution Detection

Sing-Yuan Yeh

Supervised by Prof. Chun-Hao Yang

National Taiwan University

June 26, 2026 @ STSC

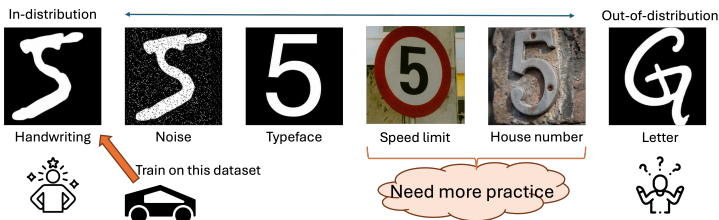
Outline

- 1 Preliminary
- 2 Main Result
- 3 Experiment

Uncertainty quantification (UQ) for deep learning

Out-of-Distribution: Out-of-distribution (OOD) samples are inputs that differ from the data distribution seen during training.

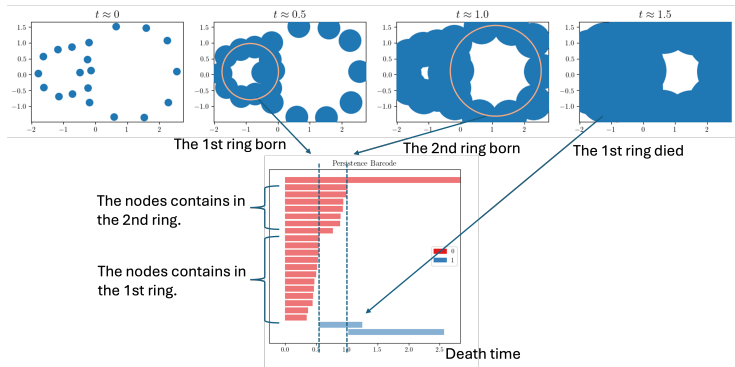
- We need uncertainty estimates to determine when a model should not be trusted.
- High uncertainty often indicates the need for additional training data.



A reliable model should know when it is outside its training distribution.

Persistent homology (Topology data analysis)

- **Filtration:** Grow balls around data points to track topological changes.
- **Life Cycle:** Record when a feature is *born* (appears) and *dies* (filled).
- **Persistence Diagram:** Plot (birth, death) pairs; longer lifespan

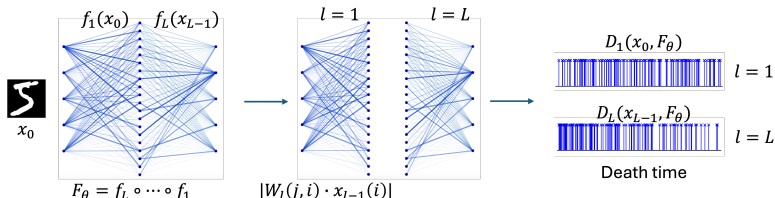


From network activations to topology [Lacombe et al., IJCAI, 2021]

- For an input x , each layer induces a weighted bipartite graph.
- Edge weights are defined by the neuron activations and network weights:

$$|W_\ell(j, i)x_{\ell-1}(i)|.$$

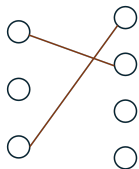
- Persistent homology summarizes each activation graph as a persistence diagram.



Advantages of topological uncertainty [Rieck et al., ICLR, 2019]

- **Global Structure Tracking:** Captures multi-scale topological features across all neuron activation scales without relying on manual activation thresholds.
- **Deep Structural Inspection:** Moves beyond output-level variance (e.g., MC Dropout) to expose topological shifts inside hidden layers.

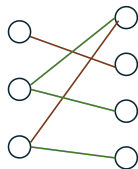
$t = 0.1 (w \leq 0.1)$



(0.03, 0.09)

$\sqsubseteq$

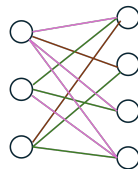
$t = 0.5 (w \leq 0.5)$



(0.03, 0.09, 0.12, 0.3, 0.46)

$\sqsubseteq$

$t = 1. (w \leq 1.)$



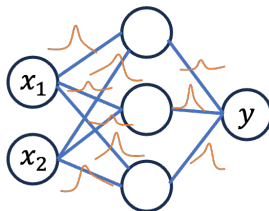
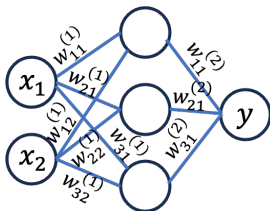
(0.03, 0.09, 0.12, 0.3, 0.46, 0.6)

Outline

- 1 Preliminary
- 2 Main Result
- 3 Experiment

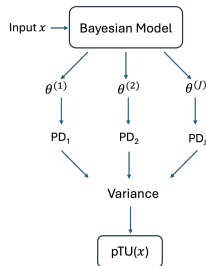
Bayesian neural networks (BNN)

- Instead of learning a single set of params, we infer a posterior distribution $\pi(\theta \mid D_{\text{train}})$ over model parameters.
- We draw multiple params $\theta^{(1)}, \theta^{(2)}, \dots, \theta^{(J)}$ from the posterior to represent models $F_{\theta^{(j)}}$.
- The variation across these posterior samples reflects model uncertainty.



Predictive topological uncertainty (pTU)

- For a input x , draw different params $\theta^{(i)}$.
- Each $\theta^{(j)}$ produces a different PDs.
- For a input x , pTU is mean of “the variance of these PDs across posterior draws”.



- The Frechet variance is defined by

$$\widehat{\text{Var}}_{\theta} [D_{\ell}(x_{\ell-1}, F_{\theta}) \mid \mathcal{D}_{\text{train}}] = \inf_m \sum_{j=1}^J W_p(D_{\ell}(x_{\ell-1}, F_{\theta^{(j)}}), m)^2$$

- For given x , pTU is defined by

$$\mathbf{pTU}(x) = \frac{1}{L} \sum_{\ell=1}^L \text{Var}_{\theta} [D_{\ell}(x_{\ell-1}, F_{\theta}) \mid \mathcal{D}_{\text{train}}] .$$

Main result

Let P_X, P_Y to be distributions of two datasets.

Stability Theorem (Yeh & Yang)

Let $X \sim P_X, Y \sim P_Y$ and $\mathcal{D}_{\text{train}}$ to be training dataset. Then, for $p > 1$, exist a constant $C > 0$ such that

$$W_p(\mathbf{pTU}(X \mid \mathcal{D}_{\text{train}}), \mathbf{pTU}(Y \mid \mathcal{D}_{\text{train}})) \leq CW_p(X, Y).$$

where W_p is p -Wasserstein distance.

For a dataset $\{x_i\}$ sampled from P_X , the empirical distribution $\{\mathbf{pTU}(x_i)\}$ is robust to noise.

Why does the stability theorem matter?

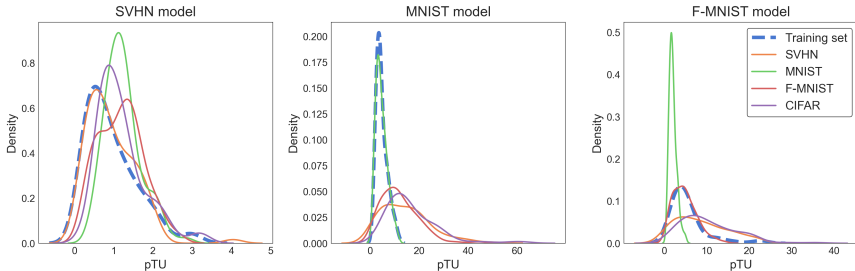
- This means that if two input distributions P_X and P_Y are close, their corresponding pTU distributions will also be close.
- This allows us to compare two datasets by comparing their *one-dimensional* pTU distributions rather than high dim datasets.
- Therefore, OOD detection can be formulated as a two-sample test on pTU values.

Outline

- 1 Preliminary
- 2 Main Result
- 3 Experiment

pTU Distributions Reveal Distribution Shifts

- Three training models are trained on SVHN, MNIST, F-MNIST dataset.
- Each dataset induces an empirical distribution of pTU values under different trained model.
- Samples from the same distribution tend to produce similar pTU patterns.
- OOD datasets lead to noticeably shifted pTU distributions.



Two-sample pTU Test

Hypothesis

$$H_0 : P_X = P_Y \quad \text{vs.} \quad H_1 : P_X \neq P_Y.$$

- Let $X_1, \dots, X_n \sim P_X$ and $Y_1, \dots, Y_n \sim P_Y$ be two datasets.
- Compute the empirical pTU values

$$\mathbf{X} = \{\text{pTU}(X_i \mid \mathcal{D}_{\text{train}})\}_{i=1}^n, \quad \mathbf{Y} = \{\text{pTU}(Y_i \mid \mathcal{D}_{\text{train}})\}_{i=1}^n.$$

- Use

$$T_{\text{obs}} = W_p(\mathbf{X}, \mathbf{Y})$$

and a permutation test to obtain the p -value.

OOD Detection via Two-sample pTU Test

- At level $\alpha = 0.05$, most off-diagonal comparisons reject H_0 .
- The pTU distributions are similar on the **diagonal**, and shifted for OOD datasets.

<i>p</i> -value	Test data $\mathcal{X}'$			
Training data $\mathcal{X}$	SVHN	MNIST	F-MNIST	CIFAR
SVHN	0.976	0.001***	0.012**	0.026*
MNIST	0.001***	0.946	0.001***	0.001***
F-MNIST	0.001*	0.001***	0.053	0.001***

Takeaway

- We introduced predictive Topological Uncertainty (pTU), a novel method to quantify uncertainty in BNNs by analyzing the topology of model-input interactions.
- The pTU is stable (w.r.t the Wasserstein distance) and provides a solid theoretical basis for OOD detection.

References

- [1] T. LACOMBE, Y. IKE, M. CARRIERE, F. CHAZAL, M. GLISSE AND Y. UMEDA, *Topological Uncertainty: Monitoring trained neural networks through persistence of activation graphs*, IJCAI, (2021).
- [2] C. BLUNDELL, J. CORNEBISE, K. KAVUKCUOGLU AND D. WIERSTRA, *Weight Uncertainty in Neural Networks*, PMLR, (2015).
- [3] B. RIECK, M. TOGNINALLI, C. BOCK, M. MOOR, M. HORN, T. GUMBSCH, K. BORGWARDT, *Neural Persistence: A Complexity Measure for Deep Neural Networks Using Algebraic Topology*, ICLR, (2019).

Thank You for Your Attention!

Appendix: Permutation Test

- Pool the two pTU samples together.
- Randomly shuffle and split them into two groups of the same sizes.
- Compute the permuted statistic

$$T_i^* = W_p(\text{new_X}, \text{new_Y}).$$

- Repeat for $i = 1, \dots, M$ and estimate

$$p\text{-value} = \frac{1}{M} \sum_{i=1}^M \mathbb{I}(T_i^* \geq T_{\text{obs}}).$$

Appendix: Wasserstein Distance

- Commonly used distance metrics for PDs include the p -Wasserstein distance of the Bottleneck distance.
- For two distributions μ and ν , the p -Wasserstein distance is defined as:

$$W_{p(\mu,\nu)} = \inf_{\gamma \in \Gamma(\mu,\nu)} \left[\mathbb{E}_{(x,y) \sim \gamma} c(x,y)^p \right]^{\frac{1}{p}}$$

where $\Gamma(\mu, \nu)$ is the set of all couplings between μ and ν .

- The function $c(x, y)$ is called the cost function and typically measures the distance between points x and y .
- In our case, we choose the L_∞ norm $c(x, y) = \|x - y\|_\infty$.