Neural Network Two-Sample Testing: Detecting Distribution Differences

Varun Khurana

Brown University

Neural Network Two-Sample Testing: A Poster Overview

Introduction

Problem: Given two distributions, p and q, determine if they are the same using a hypothesis test:

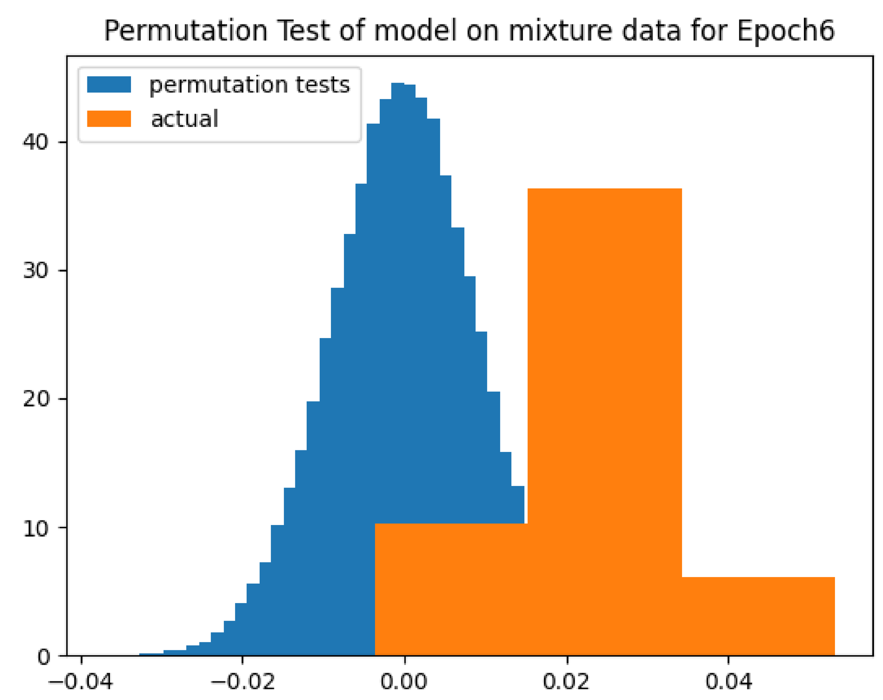
$$H_0 : p = q \quad \text{vs.} \quad H_1 : p \neq q$$

Methods: Kernel MMD, Optimal Transport, Kolmogorov-Smirnov test, **Classifier two-sample tests**

$$T(\theta; \widehat{p}, \widehat{q}) = \int_{\mathbb{R}^d} f(x, \theta) d(\widehat{p} - \widehat{q})(x)$$

Main Idea and Motivation

- Cheng et al (2022) *large* theoretical bounds for classifier neural network two-sample tests do not match experiments.
- Small networks can detect distribution differences quickly.
- Statistical power: percent of (orange) two-sample statistics lie past 95th percentile of associated permutation test curve (blue)

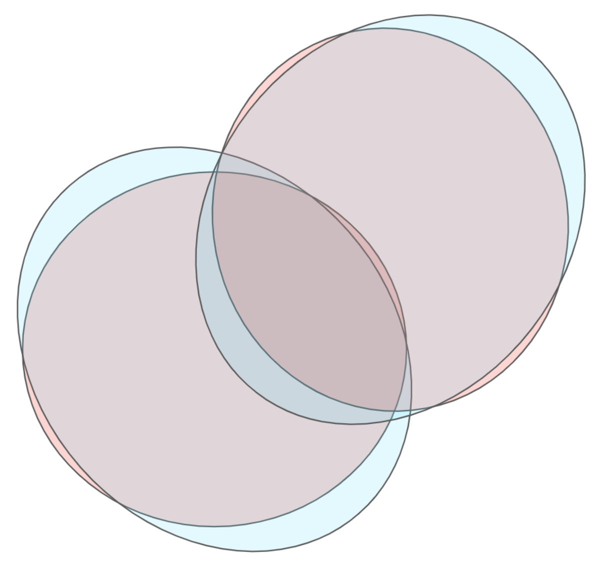


Figure

Two-layer ReLU network | 6000 training samples/distribution | 1000 test samples/distribution | $d = 20$ | 1000 tests/cell | 100 permutation tests/test | x-axis: two-sample statistic

Hard Problem

Consider Gaussian mixture models P and Q given by



$$P = \sum_{i=1}^{2} \frac{1}{2} \mathcal{N}(\mu_i^h, I_d)$$

$$Q = \sum_{i=1}^{2} \frac{1}{2} \mathcal{N}\left(\mu_i^h, \begin{bmatrix} 1 & \Delta_i^h & 0_{d-2} \\ \Delta_i^h & 1 & 0_{d-2} \\ 0_{d-2}^\top & 0_{d-2}^\top & I_{d-2} \end{bmatrix}\right),$$

where $\mu_1^h = 0_d$, $\mu_2^h = 0.5 * \mathbf{1}_d$, $\Delta_1^h = 0.5$, and $\Delta_2^h = -0.5$.

Setup

Training Setup: Train a classifier $f : \mathbb{R}^d \times \Theta \to \mathbb{R}$ to distinguish between empirical measures $\widehat{p}$ and $\widehat{q}$ from samples of p and q.

- Classifier two-sample tests use a loss function:

$$\widehat{L}(\theta) = \frac{1}{2}\left(\int_{\mathbb{R}^d} (f(x, \theta) - 1)^2 \widehat{p}(x) dx + \int_{\mathbb{R}^d} (f(x, \theta) + 1)^2 \widehat{q}(x) dx\right).$$

- In population limit, loss function becomes

$$L(\theta) = \frac{1}{2}\left\| f(\cdot, \theta) - \underbrace{\frac{p - q}{p + q}(\cdot)}_{\equiv f^*(\cdot)} \right\|_{L^2(p+q)}^2.$$

Training Dynamics:

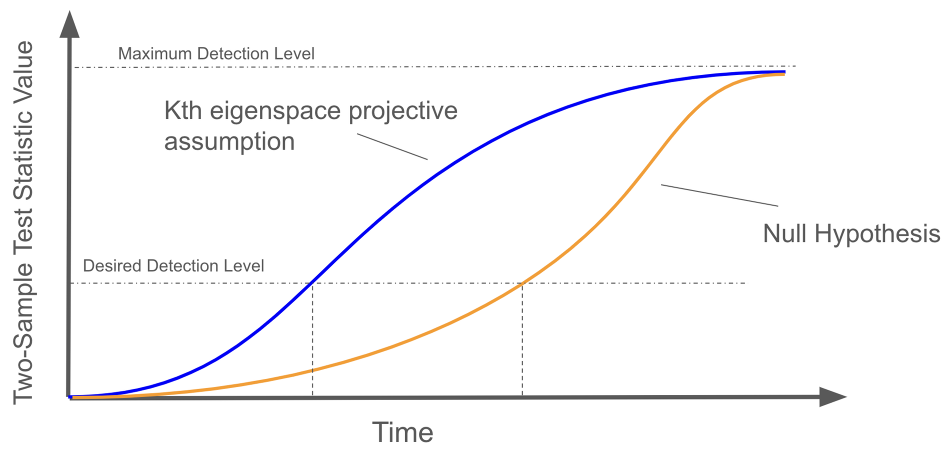
$$\partial_t \widehat{u}(x, t) = -\frac{1}{2}\left(\mathbb{E}_{x' \sim \widehat{p}} \widehat{K}_t(x, x') \widehat{e}_p(x', t) + \mathbb{E}_{x' \sim \widehat{q}} \widehat{K}_t(x, x') \widehat{e}_q(x', t)\right)$$

Training dynamics depend on neural tangent kernel (NTK) matrix $K_t(x, x') = \langle \nabla_\theta f(x, \theta(t)), \nabla_\theta f(x', \theta(t)) \rangle_\Theta$ and the errors $\widehat{e_p}$ and $\widehat{e_q}$ from p and q distributions, respectively.

Theoretical Insights

Key Theorem: Assume that f^* has a "large enough energy/norm" on the first k eigenfunctions of zero-time NTK K_0. Given a desired detection level $\epsilon > 0$ and time-separation level $C\epsilon \geq \gamma > 0$, with high probability,

$$\underbrace{t^+(\epsilon)}_{\substack{\text{min detection time} \\ \text{under null}}} - \underbrace{t^-(\epsilon)}_{\substack{\text{min detection time} \\ \text{under first } k \\ \text{eigenfunctions assumption}}} \geq \gamma > 0.$$



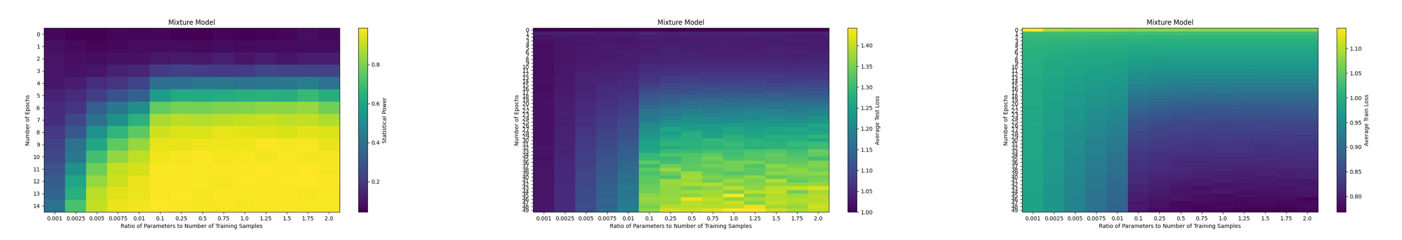
Figure

Separation of two-sample statistics with NTK.

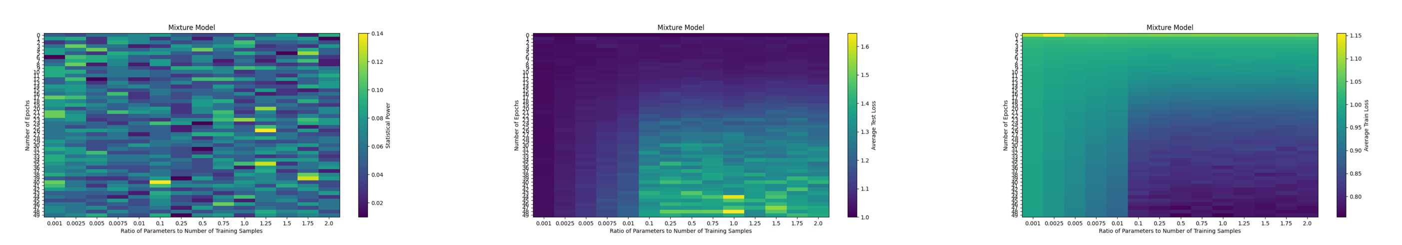
- Result is for neural network, not NTK classifier.
- Form of $t(\epsilon)$ has analytical form but depends on minimizing over subsets of eigenfunctions of K_0.
- Just need a large enough "energy" on "lower" frequency NTK modes!
- $H_0 \implies f^*$ has "high" frequency modes (sampled target function).
  $H_1 \implies f^*$ has "low" frequency modes.

Experimental Results

Statistical Power: Neural networks outperform classical two-sample tests in detecting small differences between distributions.



(a) Heatmap of statistical power under f^* nontrivial projection assumption.
(b) Heatmap of test error under f^* nontrivial projection assumption.
(c) Heatmap of train error under f^* nontrivial projection assumption.



(d) Heatmap of statistical power under null hypothesis.
(e) Heatmap of test error under null hypothesis.
(f) Heatmap of train error under null hypothesis.

- Statistical power of the "alternative" hypothesis case increases much faster than the null hypothesis case.
- The test training error for both cases increase as is expected from initialization.
- Oddly, training error decreases for both cases but this implies that the statistical power.

References

1. V. Khurana, X. Cheng, and A. Cloninger. Training Guarantees of Neural Network Classification Two-Sample Tests by Kernel Analysis, 2024. arXiv:2407.04806.
2. X. Cheng and A. Cloninger. Classification Logit Two-Sample Testing by Neural Networks for Differentiating Near Manifold Densities. IEEE Transactions on Information Theory, 68(10): 6631–6662, October 2022. DOI: 10.1109/tit.2022.3175691.