Efficient Statistical Tests: A Neural Tangent Kernel Approach

Sheng Jia, Ehsan Nezhadarya, Yuhuai Wu, Jimmy Ba
University of Toronto, Vector Institute, LG Electronics

ICML 2021

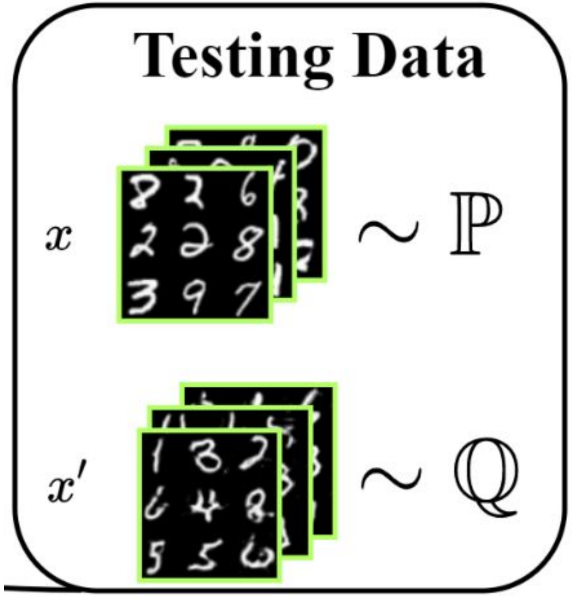
Outline

1. **Motivation**
2. Background
3. Contributions
4. Method: SCNTK for statistical tests
5. Experiments

Motivation

- **Two-sample tests:**
  Given two sets of samples, we determine whether they come from the same distribution.

- **Why** do we care about statistical tests?
  - Standard ML algorithms should only be applied in deployment if the test and training data share the same underlying distribution.



Motivation

- **Challenges with optimized kernel methods for statistical tests:**
  These methods use a portion of test data to maximize the test power, and use the rest for testing the hypothesis.

❖ There will be more computations involved from the training phase.

❖ If the sample size is much smaller than the data dimension, a fixed kernel method that uses all the available data for testing could outperform these optimized methods if the kernel is expressive enough.

Outline

1. Motivation
2. **Background**
3. Contributions
4. Method: SCNTK for statistical tests
5. Experiments

Maximum Mean Discrepancy (MMD)

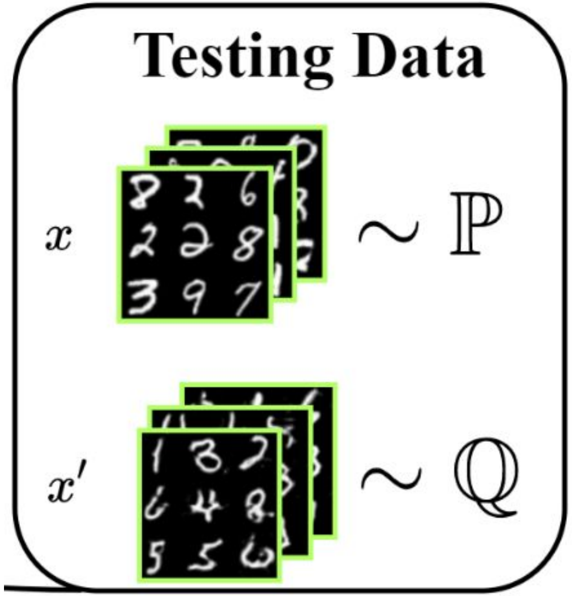
- MMD measures the distance between two distributions.

Given samples and a kernel, we can empirically estimate it

$$\widehat{\mathrm{MMD}}_u^2 = \frac{1}{m^2 - m} a + \frac{1}{n^2 - n} b - \frac{2}{m(n-1)} c$$

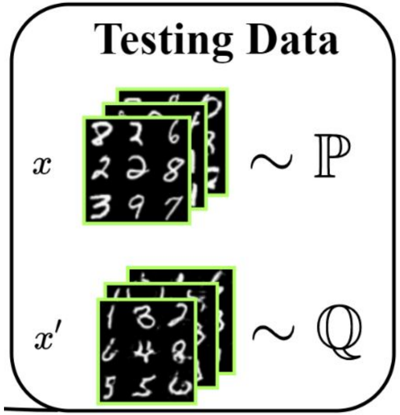
$$a = \sum_{i=1}^{m} \sum_{j \neq i}^{m} K(\mathbf{x}_i, \mathbf{x}_j) \quad b = \sum_{i=1}^{n} \sum_{j \neq i}^{n} K(\mathbf{y}_i, \mathbf{y}_j)$$

$$c = \sum_{i=1}^{m} \sum_{j \neq i}^{n} K(\mathbf{x}_i, \mathbf{y}_j)$$



Two-sample hypothesis testing

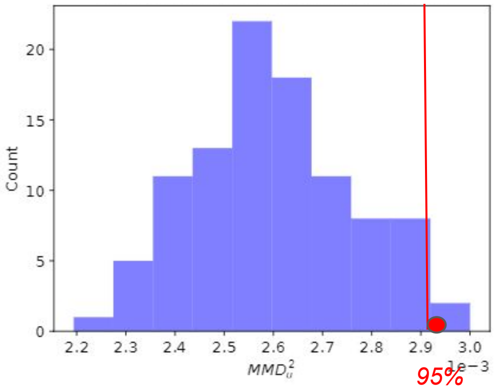
- **Null hypothesis** $h_0 : \mathbb{P} = \mathbb{Q}$
- **Alternative hypothesis** $h_1 : \mathbb{P} \neq \mathbb{Q}$



We use **permutation tests.**

❖ Under the null hypothesis, we shuffle the samples between two sets to recompute MMD test statistics, and estimate the sampling distribution.

❖ If MMD computed with the unshuffled samples is outside the 0.95 quantile, null hypothesis is rejected.



Maximum Mean Discrepancy (MMD)

- What kernel can be used?
  - ❖ Simple fixed kernels such as Gaussian and Laplace kernels.
  - ❖ Deep kernels that apply a gaussian kernel to the learned features that maximize the test power [Liu et al., 2020].
  - ❖ In this work, we apply Neural Tangent Kernel (NTK) [Jacot et al., 2018].

Outline

1. Motivation
2. Background
3. **Contributions**
4. Method: SCNTK for statistical tests
5. Experiments

Our Contributions

- Show conditions under which our **simple modifications** to Neural Tangent Kernels for MLP and CNN make them **shift-invariant** and **characteristic**.

- Demonstrate that our NTK-based statistical tests provide a **competitive and efficient** alternative to current state-of-the-art methods that require a training phase.

Outline

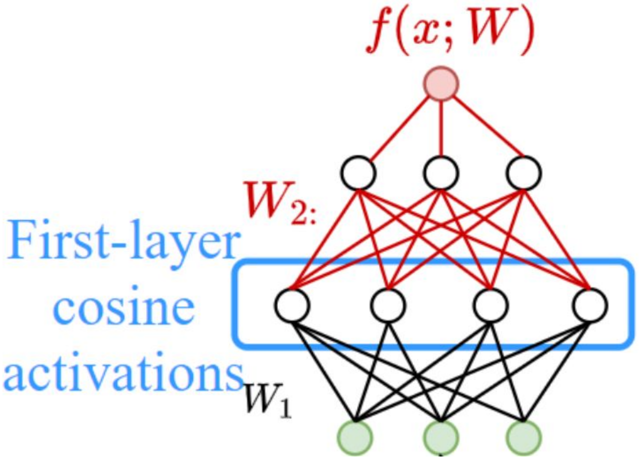
1. Motivation
2. Background
3. Contributions
4. **Method: SCNTK for statistical tests**
5. Experiments

Method: SCNTK for statistical tests

- Our kernel is the inner products of the gradients excluding the first-layer weights.

$$K_{sc}(\mathbf{x},\mathbf{x}') = \sum_{l=2}^{L}\sum_{\beta=1}^{C^{(\beta)}}\left\langle \frac{\partial f(\mathbf{x},\boldsymbol{\theta}_0)}{\partial \mathbf{W}^{(l)}_{(\beta)}}, \frac{\partial f(\mathbf{x}',\boldsymbol{\theta}_0)}{\partial \mathbf{W}^{(l)}_{(\beta)}}\right\rangle$$

- With first-layer cosine activations, this allows our kernel to be shift-invariant K(x,x')=K(x-x').



$$h^{(1)}_{(\beta)}(\mathbf{x}) = \mu_0 \cos\left(\sum_{\alpha=1}^{C^{(0)}} \mathbf{W}^{(1)}_{(\alpha),(\beta)} * \mathbf{x} + w_0\right)$$

Shift-invariant property for SNTK

- For a general MLP, we can use the previous work [Arora et al., 2019]

$$K_s(\mathbf{x}, \mathbf{x}') = \sum_{l=2}^{L+1} \left\langle \frac{\partial f(\boldsymbol{\theta}_0, \mathbf{x})}{\partial \boldsymbol{W}^{(l)}}, \frac{\partial f(\boldsymbol{\theta}_0, \mathbf{x}')}{\partial \boldsymbol{W}^{(l)}} \right\rangle = \sum_{l=2}^{L+1} \left(\Sigma^{(l-1)}(\mathbf{x}, \mathbf{x}') \prod_{l'=l}^{L+1} \dot{\Sigma}^{(l')}(\mathbf{x}, \mathbf{x}') \right)$$

where the covariances of pre-activation units are defined recursively.

$$\Sigma^{(0)}(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\top \mathbf{x}'$$

$$\boldsymbol{\Lambda}^{(l)}(\mathbf{x}, \mathbf{x}') = \begin{bmatrix} \Sigma^{(l-1)}(\mathbf{x}, \mathbf{x}) & \Sigma^{(l-1)}(\mathbf{x}, \mathbf{x}') \\ \Sigma^{(l-1)}(\mathbf{x}', \mathbf{x}) & \Sigma^{(l-1)}(\mathbf{x}', \mathbf{x}') \end{bmatrix}$$

$$\Sigma^{(l)}(\mathbf{x}, \mathbf{x}') = c_\sigma \mathop{\mathbb{E}}_{(u,v) \sim \mathsf{N}\left(\mathbf{0}, \boldsymbol{\Lambda}^{(l)}\right)} [\sigma(u)\sigma(v)]$$

With cosine activations, the first covariance will be a gaussian kernel, which is shift-invariant. Hence, the rest of covariances will be shift-invariant.

Characteristic property

Theorem 1 (Sriperumbudur et al. (2010)). *Let K, K_1, K_2 be shift-invariant kernels that can be expressed as $K(\mathbf{x}, \mathbf{y}) = \Psi(\mathbf{x} - \mathbf{y})$ where $\Psi(\cdot)$ is a bounded continuous real-valued positive definite function on $\mathbb{R}^d$. Suppose K is characteristic and $K_2 \neq 0$ Then $K + K_1$ and $K \cdot K_2$ are characteristic.*

- Using the theorem, we can see SNTK is shift-invariant since it is a sum of products of shift-invariant kernels.

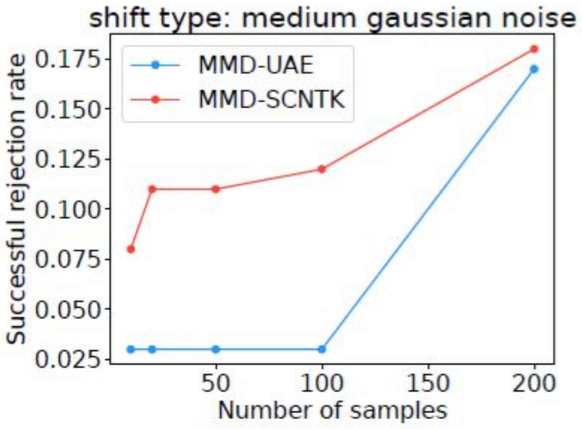
$$K_s = \underbrace{c_\sigma \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|_2^2}{2}\right)}_{①\ \text{characteristic}} \prod_{l'=2}^{L+1} \underbrace{\dot{\Sigma}^{(l')}(\mathbf{x}, \mathbf{x}')}_{②\ \text{shift-inv}} + \sum_{l=3}^{L+1} \left(\underbrace{\Sigma^{(l-1)}(\mathbf{x}, \mathbf{x}')}_{③\ \text{shift-inv}} \prod_{l'=l}^{L+1} \underbrace{\dot{\Sigma}^{(l')}(\mathbf{x}, \mathbf{x}')}_{②\ \text{shift-inv}} \right)$$

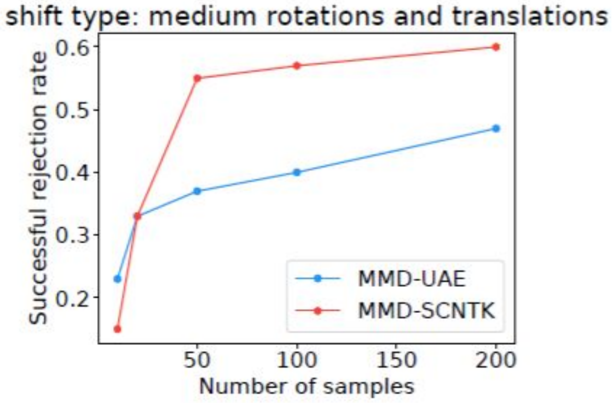
Outline

1. Motivation
2. Background
3. Contributions
4. Method: SCNTK for statistical tests
5. **Experiments**

Comparisons with fixed kernels

- Baseline: A gaussian kernel applied to nonlinear features of the data through a random neural network. **MMD-UAE**
- Dataset: MNIST
- MNIST vs Perturbed/shifted MNIST data. [Rabanser et al., 2019]





Comparisons with optimized kernels

Dataset: ● MNIST vs GAN generated MNIST ● CIFAR10 vs CIFAR10.1

Baselines: [Liu et al., 2020]

- Optimized naive gaussian kernels: ME, SCF, M-O
- Classifier based methods: C2ST-S, C2ST-L
- Deep kernel method: M-D

MNIST	SCNTK	ME	SCF	M-O	C2ST-S	C2ST-L	M-D
200	$0.324_{\pm 0.032}$	$0.414_{\pm 0.050}$	$0.107_{\pm 0.018}$	$0.188_{\pm 0.010}$	$0.193_{\pm 0.037}$	$0.234_{\pm 0.031}$	$0.555_{\pm 0.044}$
400	$0.750_{\pm 0.022}$	$0.921_{\pm 0.032}$	$0.152_{\pm 0.021}$	$0.363_{\pm 0.017}$	$0.65_{\pm 0.039}$	$0.706_{\pm 0.047}$	$0.996_{\pm 0.004}$
600	$0.963_{\pm 0.018}$	$1.000_{\pm 0.000}$	$0.294_{\pm 0.008}$	$0.619_{\pm 0.021}$	$1.000_{\pm 0.000}$	$0.977_{\pm 0.012}$	$1.000_{\pm 0.000}$
800	$1.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.317_{\pm 0.017}$	$0.797_{\pm 0.015}$	$1.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$1.000_{\pm 0.000}$
1000	$1.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$0.346_{\pm 0.019}$	$0.894_{\pm 0.016}$	$1.000_{\pm 0.000}$	$1.000_{\pm 0.000}$	$1.000_{\pm 0.000}$
Avg	0.807	0.867	0.243	0.572	0.768	0.783	0.91
CIFAR	SCNTK	ME	SCF	M-O	C2ST-S	C2ST-L	M-D
2000	**0.805**	0.588	0.171	0.316	0.452	0.529	0.744

SCNTK achieves competitive results without the training phase!

Thanks for your attention!

Reference

- Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Scholkopf, B., and Lanckriet, G. R. Hilbert space embeddings and metrics on probability measures. The Journal of Machine Learning Research, 11:1517–1561, 2010

- Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. On exact computation with an infinitely wide neural net. In Advances in Neural Information Processing Systems, pp. 8141–8150, 2019.

- Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In Advances in neural information processing systems, pp. 8571–8580, 2018

- Rabanser, S., Gunnemann, S., and Lipton, Z. Failing loudly: ¨ An empirical study of methods for detecting dataset shift. In Advances in Neural Information Processing Systems, pp. 1396–1408, 2019.

- Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. Learning deep kernels for non-parametric two sample tests. arXiv preprint arXiv:2002.09116, 2020.