Adversarial Attack and Defense for Non-Parametric Two-Sample Tests

Xilie Xu[1*] Jingfeng Zhang[2*] Feng Liu[3] Masashi Sugiyama[24] Mohan Kankanhalli[1]

[1]School of Computing, National University of Singapore
[2]RIKEN Center for Advanced Intelligence Project (AIP)
[3]School of Mathematics and Statistics, The University of Melbourne
[4]Graduate School of Frontier Sciences, The University of Tokyo
*Equal contribution

ICML 2022

Introduction

Motivation: Non-parametric two-sample tests (TSTs) have been widely applied to analysing critical data in physics[1], neurophysiology[2], biology[3], etc. Adversarial robustness of non-parametric TSTs has not been studied so far, despite its extensive studies for deep neural networks.

Our contribution: We undertake the pioneer study on adversarial robustness of non-parametric TSTs.

- We propose a generic **ensemble attack** framework which uncovers the failure mode of non-parametric TSTs and reveals non-parametric TSTs are adversarially vulnerable.
- To counteract the threats incurred by adversarial attacks, we propose to **adversarially learn kernels** for non-parametric TSTs, which makes TSTs more reliable in critical applications.

Adversarial Attacks Against Non-Parametric TSTs

Potential risk that causes a malfunction of a non-parametric TST:

1. The attacker aims to deteriorate the test's test power.
2. The attacker can craft an adversarial pair $(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}})$ as the input to the test during the testing procedure.
3. The two sets $\tilde{S}_{\mathbb{Q}}$ and $S_{\mathbb{Q}}$ should be nearly indistinguishable --- we assume the adversarial perturbation is l_∞-bounded.

Theoretical analysis

- (Proposition 1) An l_∞-bounded adversary can make adversarial perturbation imperceptible, thus guaranteeing the attack's *invisibility.*
- (Theorem 2) The test power of a non-parametric TST could be further degraded in the adversarial setting.

Proposition 1. *Under Assumptions 1 to 3, we use n_{tr} samples to train a kernel k_θ parameterized with θ and n_{te} samples to run a test of significance level α. Given the adversarial budget $\epsilon \geq 0$, the benign pair $(S_{\mathbb{P}}, S_{\mathbb{Q}})$ and the corresponding adversarial pair $(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}})$ where $\tilde{S}_{\mathbb{Q}} \in \mathcal{B}_\epsilon[S_{\mathbb{Q}}]$, with the probability at least $1-\delta$, we have*

$$\sup_\theta |\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) - \widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)| \leq \frac{8L_2\epsilon\sqrt{d}}{\sqrt{n_{\mathrm{te}}}}\sqrt{2\log\frac{2}{\delta} + 2\kappa\log(4\mathrm{R}_\Theta\sqrt{n_{\mathrm{te}}})} + \frac{8L_1}{\sqrt{n_{\mathrm{te}}}}.$$

Theorem 2. *In the setup of Proposition 1, given $\hat{\theta}_{n_{\mathrm{tr}}} = \arg\max_{\theta\in\bar{\Theta}_s}\hat{\mathcal{F}}(k_\theta)$, $r^{(n_{\mathrm{te}})}$ denoting the rejection threshold, $\mathcal{F}^* = \sup_{\theta\in\bar{\Theta}_s}\mathcal{F}(k_\theta)$, and constants C_1, C_2, C_3 depending on $\nu, L_1, \lambda, s, \mathrm{R}_\Theta$ and κ, with probability at least $1-\delta$, the test under adversarial attack has power*

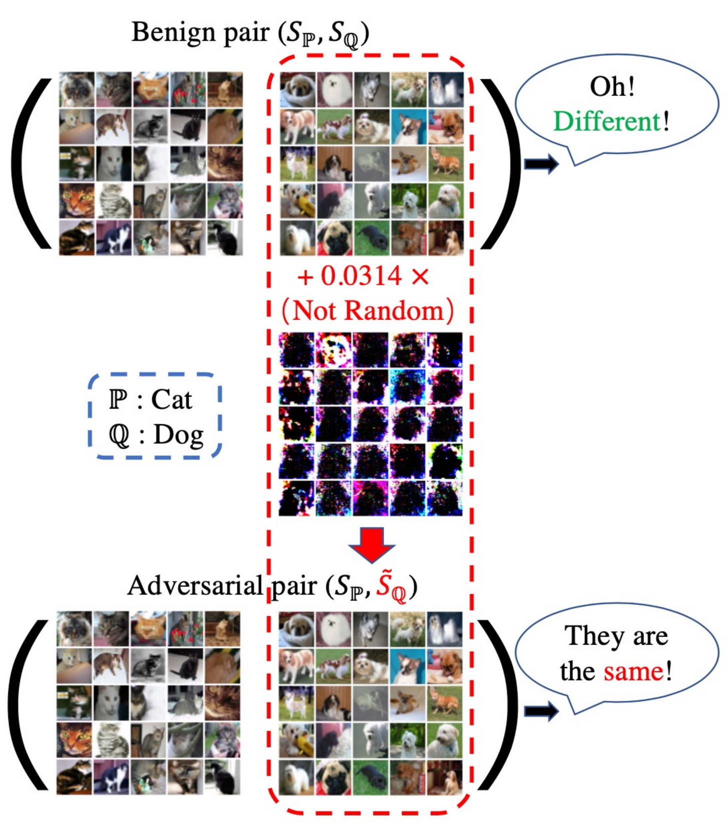
$$\Pr(n_{\mathrm{te}}\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\hat{\theta}_{n_{\mathrm{tr}}}}) > r^{(n_{\mathrm{te}})}) \geq \Phi\left[\sqrt{n_{\mathrm{te}}}\left(\mathcal{F}^* - \frac{C_1}{\sqrt{n_{\mathrm{tr}}}}\sqrt{\log\frac{\sqrt{n_{\mathrm{tr}}}}{\delta}} - \frac{C_2L_2\epsilon\sqrt{d}}{\sqrt{n_{\mathrm{te}}}}\sqrt{\log\frac{\sqrt{n_{\mathrm{te}}}}{\delta}}\right) - C_3\sqrt{\log\frac{1}{\alpha}}\right].$$

Generation of adversarial pairs

We propose TST-agnostic ensemble attack --- search for the adversarial set $\tilde{S}_{\mathbb{Q}}$ via minimizing a weighted sum of test criteria.

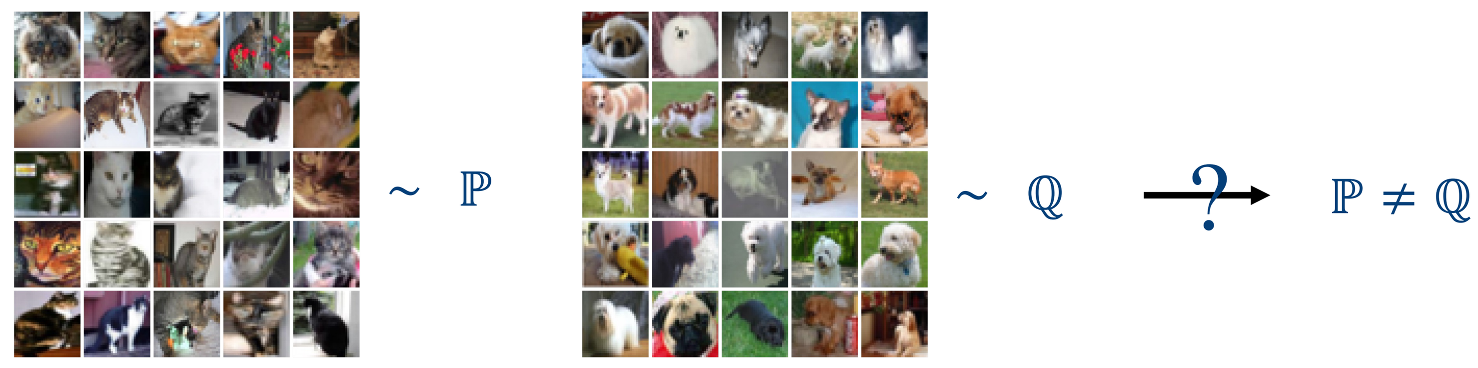
$$\tilde{S}_{\mathbb{Q}} = \arg\min_{\tilde{S}_{\mathbb{Q}}\in\mathcal{B}_\epsilon[S_{\mathbb{Q}}]} \sum_{w^{(\mathcal{J}_i)}\in\mathbb{W}, \hat{\mathcal{F}}^{(\mathcal{J}_i)}\in\hat{\mathbb{F}}} w^{(\mathcal{J}_i)}\hat{\mathcal{F}}^{(\mathcal{J}_i)}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}})$$

An example of adversarial pair $(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}})$ generated by embedding an adversarial perturbation in the benign set $S_{\mathbb{Q}}$ of the benign pair $(S_{\mathbb{P}}, S_{\mathbb{Q}})$.



Problem Formulation

Non-parametric TST is a basic tool to judge whether two sets of samples are drawn from the same distribution.



- How can a TST make the judgement?

The test compares the test statistic with a particular threshold: if the threshold is exceeded, then the test accepts the alternative hypothesis ($\mathcal{H}_1$: $\mathbb{P} \neq \mathbb{Q}$); otherwise, accepts the null hypothesis ($\mathcal{H}_0$: $\mathbb{P} = \mathbb{Q}$).

Three key components of non-parametric TSTs:

- **Test statistic** $\mathcal{D}(S_{\mathbb{P}}, S_{\mathbb{Q}})$ --- the differences between the mean embedding based on a parameterized kernel for each distribution, e.g., maximum mean discrepancy[4] (MMD).
- **Test criterion** $\hat{\mathcal{F}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k)$ --- a non-parametric TST optimizes its learnable parameters via maximizing its test criterion, thus approximately maximizing the lower bound of its test power.
- **Test power** --- the probability of correctly rejecting $\mathcal{H}_0$ against a particular number of inputs from $\mathcal{H}_1$.

Defending Non-Parametric TSTs

Adversarially learning kernels for non-parametric TSTs

- The learning objective of robust kernels is formulated as a max-min optimization:

$$\hat{\theta} \approx \arg\max_\theta \min_{\tilde{S}_{\mathbb{Q}}\in\mathcal{B}_\epsilon[S_{\mathbb{Q}}]} \hat{\mathcal{F}}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta)$$

- Our defense is based on deep kernels, i.e., robust deep kernels for non-parametric TSTs (MMD-RoD).

Algorithm 2 Adversarially Learning Deep Kernels

```
1: Input: benign pair (S_P, S_Q), maximum PGD step T,
   adversarial budget ε, checkpoint C = {c_0, ..., c_n},
   deep kernel k_θ^(RoD) parameterized by θ, training epochs
   E, learning rate η
2: Output: parameters of robust deep kernel θ
3: for e = 1 to E do
4:    X ← minibatch from S_P; Y ← minibatch from S_Q
5:    Generate an adversarial pair (X, Ỹ) by Algorithm 1
      with setting F̂ = {F̂^(RoD)(·, ·; k_θ^(RoD))}
6:    θ ← θ + η ∇_θ F̂^(RoD)(X, Ỹ; k_θ^(RoD))
7: end for
```

Experiments

Test power evaluated under ensemble attacks

- Many existing non-parametric TSTs suffer from severe adversarial vulnerabilities.

Table 1. We report the average test power of six typical non-parametric TSTs ($\alpha = 0.05$) as well as Ensemble on five benchmark datasets in benign and adversarial settings, respectively. The lower the test power under attacks is, the more adversarially vulnerable is the TST.

Datasets	ϵ	n_{te}	EA	MMD-D	MMD-G	C2ST-S	C2ST-L	ME	SCF	Ensemble
Blob	0.05	100	×	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	0.992±0.002	0.962±0.001	1.000±0.000
			√	**0.131**±0.007	**0.099**±0.003	**0.021**±0.003	**0.715**±0.091	**0.154**±0.011	**0.098**±0.022	**0.846**±0.030
HDGM	0.05	3000	×	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.002	0.942±0.013	1.000±0.000
			√	**0.259**±0.009	**0.081**±0.003	**0.105**±0.000	**0.090**±0.000	**0.500**±0.025	**0.006**±0.000	**0.734**±0.078
Higgs	0.05	5000	×	1.000±0.000	1.000±0.000	0.970±0.002	0.984±0.003	0.830±0.042	0.675±0.071	1.000±0.000
			√	**0.027**±0.001	**0.002**±0.000	**0.065**±0.000	**0.080**±0.006	**0.263**±0.022	**0.058**±0.005	**0.422**±0.013
MNIST	0.05	500	×	1.000±0.000	0.904±0.000	1.000±0.000	1.000±0.000	1.000±0.000	0.386±0.005	1.000±0.000
			√	**0.087**±0.040	**0.102**±0.002	**0.003**±0.000	**0.005**±0.000	**0.062**±0.002	**0.001**±0.000	**0.213**±0.026
CIFAR-10	0.0314	500	×	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	0.033±0.001	1.000±0.000
			√	**0.187**±0.001	**0.279**±0.004	**0.107**±0.017	**0.119**±0.021	**0.079**±0.000	**0.000**±0.000	**0.429**±0.005

Robustness of MMD-RoD

- MMD-RoD can significantly enhance the robustness of non-parametric TSTs without sacrificing the test power in the benign setting on most tasks such as MNIST and CIFAR-10.

Table 2. Test power of MMD-RoD and Ensemble$^+$.

	EA	Blob	HDGM	Higgs	MNIST	CIFAR-10
MMD-RoD	×	**1.00**±0.00	0.61±0.07	0.53±0.00	**1.00**±0.12	**1.00**±0.00
	√	**0.19**±0.06	0.00±0.01	0.23±0.02	**0.98**±0.00	**0.91**±0.00
Ensemble$^+$	×	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
	√	**0.89**±0.01	0.73±0.08	0.54±0.04	**0.98**±0.00	**0.95**±0.00

- Limitation: MMD-RoD unexpectedly perform poorly on HDGM and Higgs datasets, which has low test power in the benign and adversarial settings.
- We leave further improving the adversarial robustness of non-parametric TSTs as future work.

References

[1] Baldi, P., Sadowski, P., and Whiteson, D. Searching for exotic particles in high-energy physics with deep learning. Nature communications, 5(1):1–9, 2014.
[2] Rasch, M., Gretton, A., Murayama, Y., Maass, W., and Logothetis, N. Predicting spiking activity from local field potentials. Journal of Neurophysiology, 99:1461–1476, 2008.
[3] Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Scholkopf, B., and Smola, A. J. Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics, 22(14):e49–e57, 2006.
[4] Gretton, A., Borgwardt, K. M., Rasch, M. J., Scholkopf, B., and Smola, A. A kernel two-sample test. The Journal of Machine Learning Research, 13(1):723–773, 2012.
[5] Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. Learning deep kernels for non-parametric two-sample tests. In ICML, 2020.
[6] Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A. J., and Gretton, A. Generative models and model criticism via optimized maximum mean discrepancy. In ICLR, 2017.
[7] Lopez-Paz, D. and Oquab, M. Revisiting classifier two-sample tests. In ICLR, 2017.
[8] Cheng, X. and Cloninger, A. Classification logit two-sample testing by neural networks. IEEE Transactions on Information Theory, 2022.
[9] Chwialkowski, K. P., Ramdas, A., Sejdinovic, D., and Gretton, A. Fast two-sample testing with analytic representations of probability measures. In NeurIPS, 2015.
[10] Jitkrittum, W., Szabo, Z., Chwialkowski, K. P., and Gretton, A. Interpretable distribution features with maximum testing power. In NeurIPS, 2016.

Acknowledgements

Jingfeng Zhang was supported by JST, ACT-X Grant Number JPMJAX21AF. Masashi Sugiyama was supported by JST AIP Acceleration Research Grant Number JP- MJCR20U3 and the Institute for AI and Beyond, UTokyo. Mohan Kankanhalli's research is supported by the National Research Foundation, Singapore under its Strategic Capa- bility Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.