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

Xilie Xu [*] [1] Jingfeng Zhang [*] [2] Feng Liu [3] Masashi Sugiyama [2] [4] Mohan Kankanhalli [1]

Abstract

Non-parametric two-sample tests (TSTs) that judge whether two sets of samples are drawn from the same distribution, have been widely used in the analysis of critical data. People tend to employ TSTs as trusted basic tools and rarely have any doubt about their reliability. This paper systematically uncovers the failure mode of non-parametric TSTs through adversarial attacks and then proposes corresponding defense strategies. First, we theoretically show that an adversary can upper-bound the distributional shift which guarantees the *attack's invisibility*. Furthermore, we theoretically find that the adversary can also degrade the lower bound of a TST's *test power*, which enables us to iteratively minimize the *test criterion* in order to search for adversarial pairs. To enable TST-agnostic attacks, we propose an *ensemble attack* (EA) framework that jointly minimizes the different types of test criteria. Second, to robustify TSTs, we propose a *max-min optimization* that iteratively generates adversarial pairs to train the deep kernels. Extensive experiments on both simulated and real-world datasets validate the adversarial vulnerabilities of non-parametric TSTs and the effectiveness of our proposed defense. Source code is available at https://github.com/GodXuxilie/Robust-TST.git.

1. Introduction

Non-parametric two-sample tests (TSTs) that judge whether two sets of samples drawn from the same distribution have been widely used to analyze critical data in physics (Baldi et al., 2014), neurophysiology (Rasch et al., 2008), biology (Borgwardt et al., 2006), etc. Compared with traditional methods (such as the *t*-test), non-parametric TSTs can relax the strong parametric assumption about the distributions being studied and are effective in complex domains (Gretton et al., 2009; 2012; Chwialkowski et al., 2015; Jitkrittum et al., 2016; Sutherland et al., 2017; Lopez-Paz & Oquab, 2016; Cheng & Cloninger, 2019; Liu et al., 2020a; 2021). Notably, the use of deep kernels (Liu et al., 2020a) flexibly empowers the non-parametric TSTs to learn even more complex distributions.

However, the adversarial robustness of non-parametric TSTs is rarely studied, despite its extensive studies for deep neural networks (DNNs). Studies of DNNs' adversarial robustness (Madry et al., 2018) have enabled significant advances in defending against adversarial attacks (Szegedy et al., 2014), which can help enhance the security in various domains such as computer vision (Xie et al., 2017; Mahmood et al., 2021), natural language processing (Zhu et al., 2020; Yoo & Qi, 2021), recommendation system (Peng & Mine, 2020), etc. We therefore undertake this pioneer study on adversarial robustness of non-parametric TSTs, which uncovers the failure mode of non-parametric TSTs through adversarial attacks and facilitate an effective strategy for making TSTs reliable in critical applications (Baldi et al., 2014; Rasch et al., 2008; Borgwardt et al., 2006).

First, we theoretically show the adversary could upper-bound the distributional shift and degrade the lower bound of a TST's test power (details in Section 3.1). Given a benign pair $(S_{\mathbb{P}}, S_{\mathbb{Q}})$, in which $S_{\mathbb{P}} = \{x_i\}_{i=1}^m \sim \mathbb{P}^m$ and $S_{\mathbb{Q}} = \{y_j\}_{j=1}^n \sim \mathbb{Q}^n$, an ℓ_∞-bounded adversary could generate the adversarial pair $(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}})$. We will show in Proposition 1 that the maximum mean discrepancy (MMD) (Gretton et al., 2012) between the benign and adversarial pairs is upper-bounded, which guarantees imperceptible adversarial perturbations (Szegedy et al., 2014). Furthermore, we will show in Theorem 2 that the adversary can degrade the lower bound of a TST's test power, which implies that a TST could wrongly determine $\mathbb{P} = \mathbb{Q}$ with a larger probability under adversarial attacks when $\mathbb{P} \neq \mathbb{Q}$ holds.

Then, we realize effective adversarial attacks against non-parametric TSTs (details in Section 3.2). We formulate an attack as a constraint optimization problem that minimizes a TST's test criterion (Liu et al., 2020a) within the

[*]Equal contribution [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. Correspondence to: Jingfeng Zhang <jingfeng.zhang@riken.jp>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

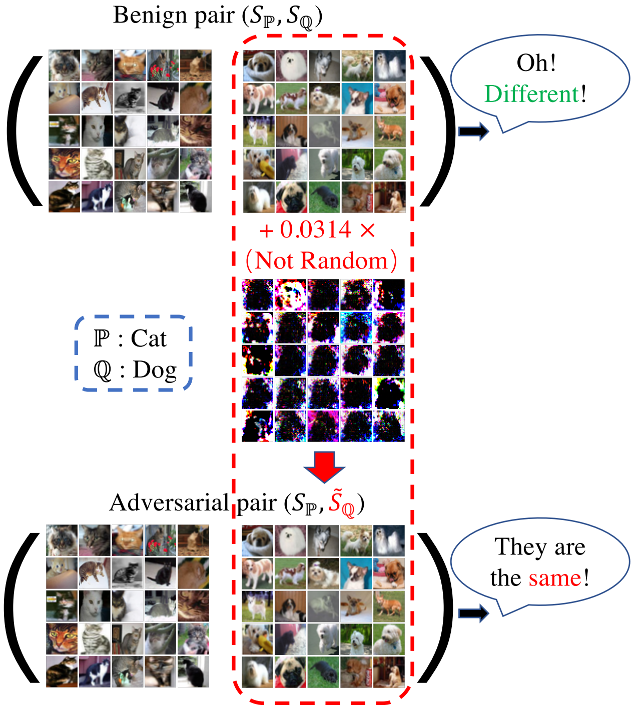

Benign pair $(S_{\mathbb{P}}, S_{\mathbb{Q}})$
Oh! Different!
+ 0.0314 ×
(Not Random)
$\mathbb{P}$: Cat
$\mathbb{Q}$: Dog
Adversarial pair $(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}})$
They are the same!

Figure 1. 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 $(S_{\mathbb{P}}, S_{\mathbb{Q}})$. Experimental details are in Section 5.1.

ℓ_∞-bound of size ϵ on $S_{\mathbb{Q}}$. We utilize projected gradient descent (PGD) (Madry et al., 2018) to efficiently search the adversarial set $\tilde{S}_{\mathbb{Q}}$ and incorporate automatic schedule of the step size (Croce & Hein, 2020) to improve the optimization convergence. Moreover, we extend the attack beyond a specific TST to a generic TST-agnostic attack, namely, ensemble attack (EA). EA jointly minimizes a weighted sum of different test criteria, which can simultaneously fool various TSTs. For example, Figure 1 shows non-parametric TSTs can correctly differentiate the benign pair of "cats" and "dogs" (top) coming from the different distributions, but wrongly judge adversarial pairs (bottom) as belonging to the same distribution.

Second, to robustify the non-parametric TSTs, we study the corresponding defense approaches (details in Section 4). A straightforward defense seems to use an ensemble of TSTs. We find an ensemble of TSTs is sometimes effective against a specific attack targeting a certain type of TSTs but almost always fails under EA (see experiments in Section 5.1). Therefore, to effectively defend against adversarial attacks, we propose to adversarially learn the robust kernels. The defense is formulated as a *max-min* optimization that is similar in flavor to the adversarial training's *min-max* formulation (Madry et al., 2018). For its realization, we iteratively generate adversarial pairs by minimizing the test criterion in the *inner minimization* and update kernel parameters by maximizing the test criterion on the adversarial pairs in the *outer maximization*. We realize our defense using deep kernels that have achieved the state-of-the-art (SOTA) performance in non-parametric TSTs (Liu et al., 2020a).

Lastly, we empirically justify the proposed attacks and defenses (in Section 5). We evaluate the test power of many existing non-parametric TSTs (non-robust) and the robust-kernel TST (robust) under the EA on simulated and real-world datasets, including complex synthetic distributions, high-energy physics data, and challenging images. Comprehensive experimental results validate that the existing non-parametric TSTs lack adversarial robustness; we can significantly improve the adversarial robustness of non-parametric TSTs through adversarially learning the deep kernels.

2. Non-Parametric Two-Sample Tests

In this section, we provide the preliminaries of non-parametric TSTs and provide discussions with the related studies in Appendix C.

2.1. Problem Formulation

Let $\mathcal{X} \subset \mathbb{R}^d$ and $\mathbb{P}, \mathbb{Q}$ be Borel probability measures on $\mathcal{X}$. A non-parametric TST $\mathcal{J}(S_{\mathbb{P}}, S_{\mathbb{Q}}) : \mathcal{X}^m \times \mathcal{X}^n \mapsto \{0, 1\}$ is used to distinguish between the null hypothesis $\mathcal{H}_0 : \mathbb{P} = \mathbb{Q}$ and the alternative hypothesis $\mathcal{H}_1 : \mathbb{P} \neq \mathbb{Q}$, where $S_{\mathbb{P}}$ and $S_{\mathbb{Q}}$ are independent identically distributed (IID) samples of size m and n drawn from $\mathbb{P}$ and $\mathbb{Q}$, respectively. A non-parametric TST constructs a mean embedding based on a kernel parameterized with θ for each distribution, and utilizes the differences in these embeddings as the *test statistic* for the hypothesis test. The judgement is made by comparing the test statistic $\mathcal{D}(S_{\mathbb{P}}, S_{\mathbb{Q}})$ with a particular threshold r: if the threshold is exceeded, then the test rejects $\mathcal{H}_0$. The *test power* (TP) of a non-parametric TST $\mathcal{J}$ is measured by the probability of correctly rejecting $\mathcal{H}_0$ when the alternative hypothesis is true, i.e., $\mathrm{TP}(\mathcal{J}) = \mathbb{E}_{S_{\mathbb{P}} \sim \mathbb{P}^m, S_{\mathbb{Q}} \sim \mathbb{Q}^n}[\mathbb{1}(\mathcal{J}(S_{\mathbb{P}}, S_{\mathbb{Q}}) = 1)]$ for a particular $\mathbb{P} \neq \mathbb{Q}$. A non-parametric TST optimizes its learnable parameters θ via maximizing its *test criterion*, thus approximately maximizing its test power.

2.2. Test Statistics

Here, we introduce a typical test statistic, maximum mean discrepancy (MMD) (Gretton et al., 2012), and leave other test statistics in Appendix D, such as tests based on Gaussian kernel mean embeddings at specific positions (Chwialkowski et al., 2015; Jitkrittum et al., 2016) and classifier two-sample tests (C2ST) (Lopez-Paz & Oquab, 2016; Cheng & Cloninger, 2019).

Definition 1 (Gretton et al. (2012))**.** Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel of a reproducing kernel Hilbert space $\mathcal{H}_k$, with feature maps $k(\cdot, x) \in \mathcal{H}_k$. Let $X \sim \mathbb{P}$ and $Y \sim \mathbb{Q}$, and define the kernel mean embeddings $\mu_{\mathbb{P}} = \mathbb{E}[k(\cdot, X)]$ and $\mu_{\mathbb{Q}} = \mathbb{E}[k(\cdot, Y)]$. Under mild integrability conditions,

$$\begin{aligned}\mathrm{MMD}(\mathbb{P}, \mathbb{Q}; \mathcal{H}_k) &= \sup_{f \in \mathcal{H}, \|f\|_{\mathcal{H}_k} \leq 1} |\mathbb{E}[f(X)] - \mathbb{E}[f(Y)]| \\ &= \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}_k}. \end{aligned} \tag{1}$$

For characteristic kernels, $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}; \mathcal{H}_k) = 0$ if and only if $\mathbb{P} = \mathbb{Q}$. Assuming $n = m$, we can estimate MMD (Eq. (1)) using the U-statistic estimator, which is unbiased for MMD^2 and has nearly minimal variance among all unbiased estimators (Gretton et al., 2012):

$$\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k) = \frac{1}{n(n-1)} \sum_{i \neq j} H_{ij}, \tag{2}$$

$$H_{ij} = k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i),$$

where $x_i, x_j \in S_{\mathbb{P}}$ and $y_i, y_j \in S_{\mathbb{Q}}$.

In this paper, we investigate six types of non-parametric TSTs as follows since Liu et al. (2020a; 2021) have shown they are powerful on complex data.

- $\mathcal{D}^{(\mathrm{G})}(\cdot, \cdot; k^{(\mathrm{G})}) = \widehat{\mathrm{MMD}}^2(\cdot, \cdot; k^{(\mathrm{G})})$ for tests based on MMD with Gaussian kernels (MMD-G) (Sutherland et al., 2017) with the learnable lengthscale σ_ϕ, in which $k^{(\mathrm{G})}(x, y) = \exp(-\frac{1}{2\sigma_\phi}\|x - y\|^2)$.
- $\mathcal{D}^{(\mathrm{D})}(\cdot, \cdot; k^{(\mathrm{D})}) = \widehat{\mathrm{MMD}}^2(\cdot, \cdot; k^{(\mathrm{D})})$ for tests based on MMD with deep kernels (MMD-D) (Liu et al., 2020a). Note that $k^{(\mathrm{D})}(x, y) = [(1 - \gamma)\exp(-\frac{1}{2\sigma_\phi}\|\phi(x) - \phi(y)\|^2) + \gamma]\exp(-\frac{1}{2\sigma_q}\|x - y\|^2)$ where $\gamma, \sigma_\phi, \sigma_q$ are the learnable parameters and $\phi(\cdot)$ is a parameterized deep network to extract the features.
- $\mathcal{D}^{(\mathrm{S})}(\cdot, \cdot)$ (Eq. (10)) for C2ST based on Sign (C2ST-S) (Lopez-Paz & Oquab, 2016). A classifier $f : \mathcal{X} \to \mathbb{R}$ that outputs the classification probabilities is utilized by C2ST. Liu et al. (2020a) pointed out that the test statistic of C2ST-S is equivalent to MMD with kernel $k^{(\mathrm{S})}$, i.e., $\mathcal{D}^{(\mathrm{S})}(\cdot, \cdot) = \widehat{\mathrm{MMD}}^2(\cdot, \cdot; k^{(\mathrm{S})})$ where $k^{(\mathrm{S})}(x, y) = \frac{1}{4}\mathbb{1}(f(x) > 0)\mathbb{1}(f(y) > 0)$.
- $\mathcal{D}^{(\mathrm{L})}(\cdot, \cdot)$ (Eq. (11)) for C2ST-L (Cheng & Cloninger, 2019) that utilizes the discriminator's measure of confidence. Its test statistic is also equivalent to MMD with kernel $k^{(\mathrm{L})}$ (Liu et al., 2020a), i.e., $\mathcal{D}^{(\mathrm{L})}(\cdot, \cdot) = \widehat{\mathrm{MMD}}^2(\cdot, \cdot; k^{(\mathrm{L})})$ where $k^{(\mathrm{L})}(x, y) = f(x)f(y)$.
- $D^{(\mathrm{ME})}(\cdot, \cdot)$ (Eq. (12)) for tests based on differences in Gaussian kernel mean embeddings at specific locations (Chwialkowski et al., 2015; Jitkrittum et al., 2016), namely Mean Embedding (ME).
- $D^{(\mathrm{SCF})}(\cdot, \cdot)$ (Eq. (13)) for tests based on Gaussian kernel mean embeddings at a set of optimized frequency (Chwialkowski et al., 2015; Jitkrittum et al., 2016), namely Smooth Characteristic Functions (SCF).

2.3. Test Criterion

In this subsection, we introduce the test criteria for non-parametric TSTs based on MMD (Sutherland et al., 2017; Liu et al., 2020a; Lopez-Paz & Oquab, 2016; Cheng & Cloninger, 2019).

Theorem 1 (Asymptotics of MMD under $\mathcal{H}_1$ (Serfling, 2009))**.** *Under the alternative, $\mathcal{H}_1 : \mathbb{P} \neq \mathbb{Q}$, the standard central limit theorem holds:*

$$\sqrt{n}(\widehat{\mathrm{MMD}}^2 - \mathrm{MMD}^2) \to \mathcal{N}(0, \sigma^2_{\mathcal{H}_1}),$$
$$\sigma^2_{\mathcal{H}_1} = 4(\mathbb{E}[H_{12}H_{13}] - \mathbb{E}[H_{12}]^2),$$

where H_{12}, H_{13} refer to H_{ij} in Eq. (2).

Guided by the asymptotics of MMD (Theorem 1), the test power is estimated as follows:

$$\Pr(n\widehat{\mathrm{MMD}}^2 > r) \to \Phi(\frac{\sqrt{n}\mathrm{MMD}^2}{\sigma_{\mathcal{H}_1}} - \frac{r}{\sqrt{n}\sigma_{\mathcal{H}_1}}), \tag{3}$$

where Φ is the cumulative distribution function (CDF) of standard normal distribution and r is the rejection threshold approximately found via permutation testing (Dwass, 1957; Fernández et al., 2008). This general method is usually considered best to estimate the null hypothesis: under $\mathcal{H}_0$, samples from $\mathbb{P}$ and $\mathbb{Q}$ are interchangeable, and repeatedly re-computing the test statistic with samples randomly shuffled between $S_{\mathbb{P}}$ and $S_{\mathbb{Q}}$ estimates its null distribution.

For reasonably large n, the test power is dominated by the first term of Eq. (3), and thus the TST yields the most powerful test by approximately maximizing the test criterion (Liu et al., 2020a)

$$\mathcal{F}(\mathbb{P}, \mathbb{Q}; k) = \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}; k)/\sigma_{\mathcal{H}_1}(\mathbb{P}, \mathbb{Q}; k). \tag{4}$$

Further, $\mathcal{F}(\mathbb{P}, \mathbb{Q}; k)$ can be empirically estimated with

$$\hat{\mathcal{F}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k) = \frac{\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k)}{\hat{\sigma}_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k)}, \tag{5}$$

where $\hat{\sigma}^2_{\mathcal{H}_1,\lambda}$ is a regularized estimator of $\sigma^2_{\mathcal{H}_1}$:

$$\hat{\sigma}^2_{\mathcal{H}_1,\lambda} = \frac{4}{n^3}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} H_{ij}\right)^2 - \frac{4}{n^4}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} H_{ij}\right)^2 + \lambda,$$

where λ is a positive constant. The test criterion of the MMD test (e.g., MMD-G, MMD-D, C2ST-S and C2ST-L) is calculated based on its corresponding kernel. We let σ^2_θ denote $\sigma^2_{\mathcal{H}_1}(\mathbb{P}, \mathbb{Q}; k_\theta)$, and analogously $\hat{\sigma}^2_\theta$ denote $\hat{\sigma}^2_{\mathcal{H}_1,\lambda}(\mathbb{P}, \mathbb{Q}; k_\theta)$, for simplicity.

In addition, Chwialkowski et al. (2015) and Jitkrittum et al. (2016) analyzed that the test power of ME tests, and SCF tests can be approximately maximized by maximizing the corresponding test criterion as well, i.e., $\hat{\mathcal{F}}^{(\mathrm{ME})}(S_{\mathbb{P}}, S_{\mathbb{Q}})$ and $\hat{\mathcal{F}}^{(\mathrm{SCF})}(S_{\mathbb{P}}, S_{\mathbb{Q}})$ (details in Appendix D).

To avoid notation clutter, we simply let θ represent all of the learnable parameters in a non-parametric TST. The optimized parameters of a TST are obtained as follows.

$$\hat{\theta} \approx \arg\max_{\theta} \hat{\mathcal{F}}(S_{\mathbb{P}}^{\mathrm{tr}}, S_{\mathbb{Q}}^{\mathrm{tr}}; k_{\theta}), \tag{6}$$

where $(S_{\mathbb{P}}^{\mathrm{tr}}, S_{\mathbb{Q}}^{\mathrm{tr}})$ is the training pair. Then, we conduct a hypothesis test based on $\mathcal{D}(S_{\mathbb{P}}^{\mathrm{te}}, S_{\mathbb{Q}}^{\mathrm{te}}; k_{\hat{\theta}})$, where $(S_{\mathbb{P}}^{\mathrm{te}}, S_{\mathbb{Q}}^{\mathrm{te}})$ is the test pair.

3. Adversarial Attacks Against Non-Parametric TSTs

In this section, we first show the possible existence of adversarial attacks against a non-parametric TST. Then, we propose a method to generate adversarial test pairs that can fool a TST. To enable TST-agnostic attacks, we propose a unified attack framework, i.e., ensemble attack.

3.1. Theoretical Analysis

This section theoretically shows that there could exist adversarial attacks that can invisibly undermine a TST. We first lay out the needed assumptions on kernel functions.

Assumption 1. The possible kernel parameterized with $\theta \in \mathbb{R}^{\kappa}$ lies in Banach space. The set of possible kernel parameters Θ is bounded by R_{Θ}, i.e., $\Theta \subseteq \{\theta \mid \|\theta\| \leq \mathrm{R}_{\Theta}\}$. We let $\bar{\Theta}_s = \{\theta \in \Theta \mid \sigma_{\theta}^2 \geq s^2 > 0\}$ in which s is a positive constant.

Assumption 2. The kernel function k_{θ} is uniformly bounded, i.e., $\sup_{\theta \in \Theta} \sup_{x \in \mathcal{X}} k_{\theta}(x, x) \leq \nu$. We treat ν as a constant.

Assumption 3. The kernel function $k_{\theta}(x, y)$ satisfies the Lipschitz conditions as follows.

$$|k_{\theta}(x, y) - k_{\theta'}(x, y)| \leq L_1 \|\theta - \theta'\|;$$
$$|k_{\theta}(x, y) - k_{\theta}(x', y')| \leq L_2(\|x - x'\| + \|y - y'\|),$$

where L_1 and L_2 are positive constants.

We consider a potential risk that causes a malfunction of a non-parametric TST: an adversarial attacker that aims to deteriorate the TST's test power, can craft an adversarial pair $(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}})$ as the input to the TST during the testing procedure, in which the two sets $\tilde{S}_{\mathbb{Q}}$ and $S_{\mathbb{Q}}$ are nearly indistinguishable. We provide a detailed description of the attacker against non-parametric TSTs in Appendix F.

We define the ϵ-ball centered at $x \in \mathcal{X}$ as follows:

$$\mathcal{B}_{\epsilon}[x] = \{\tilde{x} \in \mathcal{X} \mid \|x - \tilde{x}\|_{\infty} \leq \epsilon\}.$$

Further, an ℓ_{∞}-bound of size ϵ on the set $S_{\mathbb{Q}}$ is defined as

$$\mathcal{B}_{\epsilon}[S_{\mathbb{Q}}] = \{\tilde{S}_{\mathbb{Q}} = \{\tilde{x}_i \in \mathcal{X}\}_{i=1}^{n} \mid$$
$$\tilde{x}_i \in \mathcal{B}_{\epsilon}[x_i], \quad \forall x_i \in S_{\mathbb{Q}}, \tilde{x}_i \in \tilde{S}_{\mathbb{Q}}\}.$$

Without loss of generality, we assume that the adversarial perturbation is ℓ_{∞}-bounded of size ϵ, i.e., $\tilde{S}_{\mathbb{Q}} \in \mathcal{B}_{\epsilon}[S_{\mathbb{Q}}]$. We leave exploring the effects of other constraints that can bound the "human imperception" as the future work, such as Wasserstein-distance constraints (Wong et al., 2019).

Under ℓ_{∞}-bounded attacks, we conduct our theoretical analysis of distributional shift in the test pairs as follows.

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*

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

The proof is in Appendix B.1.

Remark 1. Proposition 1 shows that ϵ can control the upper bound of distributional shift measured by MMD between samples in the test pair. In other words, a small ϵ can ensure the difference between $\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\theta})$ and $\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_{\theta})$ is numerically small. Therefore, an ℓ_{∞}-bounded adversary can make the adversarial perturbation imperceptible, thus guaranteeing the attack's *invisibility*.

Next, we provide a lemma that theoretically analyzes the adversary's influence on the estimated test criterion.

Lemma 1. *In the setup of Proposition 1, with probability at least $1 - \delta$, we have*

$$\sup_{\theta \in \bar{\Theta}_s} |\hat{\mathcal{F}}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\theta}) - \hat{\mathcal{F}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_{\theta})|$$
$$= \mathcal{O}\left(\frac{\epsilon L_2\sqrt{d\left(\log\frac{1}{\delta} + \kappa\log(\mathrm{R}_{\Theta}\sqrt{n_{\mathrm{te}}})\right)} + L_1}{s\sqrt{n_{\mathrm{te}}}}\right).$$

The proof is in Appendix B.2.

Remark 2. Lemma 1 shows that, when $\epsilon > 0$, a TST needs a larger number of test samples to facilitate the estimated test criterion on the adversarial test pair to converge to the estimated test criterion on the benign test pair. In other words, the estimated test criterion in adversarial settings ($\epsilon > 0$) could be lower than the estimated test criterion in benign settings for a particular n_{te}.

Since the test criterion dominates the test power, Lemma 1 motivates us to further theoretically analyze the adversary's effects on the lower bound of a TST's test power as follows.

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*

$$\Pr(n_{\mathrm{te}}\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\hat{\theta}_{n_{\mathrm{tr}}}}) > r^{(n_{\mathrm{te}})}) \geq \Phi\Bigg[\sqrt{n_{\mathrm{te}}}\Bigg(\mathcal{F}^* - \frac{C_1}{\sqrt{n_{\mathrm{tr}}}}\sqrt{\log\frac{\sqrt{n_{\mathrm{tr}}}}{\delta}} - \frac{C_2 L_2 \epsilon \sqrt{d}}{\sqrt{n_{\mathrm{te}}}}\sqrt{\log\frac{\sqrt{n_{\mathrm{te}}}}{\delta}}\Bigg) - C_3\sqrt{\log\frac{1}{\alpha}}\Bigg].$$

The proof is in Appendix B.3.

Remark 3. Theorem 2 indicates that the lower bound of test power can become lower with the increase of the adversarial budget ϵ, the dimensionality of data d and Lipschitz constant L_2 of the kernel function, which implies that the test power of a TST could be further degraded in the adversarial setting. In other words, a non-parametric TST could wrongly accept $\mathcal{H}_0$ with a larger probability in the adversarial setting when $\mathbb{P} \neq \mathbb{Q}$ holds. Therefore, with the $\epsilon > 0$ being constrained within a reasonable range, there could exist an adversarial attack that can invisibly fool a non-parametric TST.

3.2. Generation of Adversarial Pairs

Formulation. Motivated by Theorem 1, a TST could output a wrong judgement on an adversarial pair $(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}})$ with a larger probability when the test criterion $\hat{\mathcal{F}}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}})$ becomes smaller. Therefore, to generate an adversarial pair against a non-parametric TST, we update $\tilde{S}_{\mathbb{Q}}$ via minimizing the test criterion $\hat{\mathcal{F}}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}})$. We formulate adversarial attacks against a non-parametric TST $\mathcal{J}$ in the following:

$$\tilde{S}_{\mathbb{Q}} = \underset{\tilde{S}_{\mathbb{Q}} \in \mathcal{B}_\epsilon[S_{\mathbb{Q}}]}{\arg\min}\ \hat{\mathcal{F}}^{(\mathcal{J})}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}), \tag{7}$$

where $\hat{\mathcal{F}}^{(\mathcal{J})}(\cdot,\cdot)$ is the test criterion, $\tilde{S}_{\mathbb{Q}}$ is constrained in an ϵ-ball centered at $S_{\mathbb{Q}}$.

Realization. We utilize PGD (Madry et al., 2018) to approximately solve the minimization problem of Eq. (7). Given a starting point $S_{\mathbb{Q}}^{(0)}$, step size $\rho > 0$, iteration number $t \in \mathbb{N}$, and the size of adversarial budget $\epsilon \geq 0$, PGD works as follows:

$$S_{\mathbb{Q}}^{(t+1)} = \{\Pi_{\mathcal{B}_\epsilon[x_i^{(0)}]}\big(x_i^{(t)} - \rho\,\mathrm{sign}(\nabla_{x_i^{(t)}}\hat{\mathcal{F}}(S_{\mathbb{P}}, S_{\mathbb{Q}}^{(t)}))\big)\}_{i=1}^n,$$

where $x_i^{(0)} \in S_{\mathbb{Q}}^{(0)}$, $x_i^{(t)} \in S_{\mathbb{Q}}^{(t)}$, $\Pi_{\mathcal{B}_\epsilon[x^{(0)}]}(\cdot)$ is the projection function that projects the adversarial data back into the ϵ-ball centered at $x^{(0)}$, and $\hat{\mathcal{F}}(\cdot,\cdot)$ is a differentiable function.

Algorithm 1 Ensemble Attack (EA)

```
Input: benign pair $(S_{\mathbb{P}}, S_{\mathbb{Q}})$, maximum PGD step $T$, adversarial budget $\epsilon$, test criterion function set $\hat{\mathbb{F}}$, weight set $\mathbb{W}$, checkpoint $\mathbb{C} = \{c_0, \ldots, c_n\}$
Output: adversarial pair $(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}})$
$S_{\mathbb{Q}}^{(0)} \leftarrow S_{\mathbb{Q}}$ and $\rho \leftarrow \epsilon$
$S_{\mathbb{Q}}^{(1)} \leftarrow \{\Pi_{\mathcal{B}_\epsilon[x_i^{(0)}]}\big(x_i^{(0)} - \rho\,\mathrm{sign}(\nabla_{x_i^{(0)}}\ell(S_{\mathbb{P}}, S_{\mathbb{Q}}^{(0)}))\big)\}_{i=1}^n$
$\ell_{\min} \leftarrow \min\{\ell(S_{\mathbb{P}}, S_{\mathbb{Q}}^{(0)}), \ell(S_{\mathbb{P}}, S_{\mathbb{Q}}^{(1)})\}$
$\tilde{S}_{\mathbb{Q}} \leftarrow S_{\mathbb{Q}}^{(0)}$ if $\ell_{\min} \equiv \ell(S_{\mathbb{P}}, S_{\mathbb{Q}}^{(0)})$ else $\tilde{S}_{\mathbb{Q}} \leftarrow S_{\mathbb{Q}}^{(1)}$
for $t = 1$ to $T-1$ do
    $S_{\mathbb{Q}}^{(t+1)} \leftarrow \{\Pi_{\mathcal{B}_\epsilon[x_i^{(0)}]}\big(x_i^{(t)} - \rho\mathrm{sign}(\nabla_{x_i^{(t)}}\ell(S_{\mathbb{P}}, S_{\mathbb{Q}}^{(t)}))\big)\}_{i=1}^n$
    if $\ell_{\min} > \ell(S_{\mathbb{P}}, S_{\mathbb{Q}}^{(t+1)})$ then
        $\tilde{S}_{\mathbb{Q}} \leftarrow S_{\mathbb{Q}}^{(t+1)}$ and $\ell_{\min} \leftarrow \ell(S_{\mathbb{P}}, S_{\mathbb{Q}}^{(t+1)})$
    end if
    if $t \in \mathbb{C}$ then
        if Condition 1 or Condition 2 then
            $\rho \leftarrow \rho/2$ and $S_{\mathbb{Q}}^{(t+1)} \leftarrow \tilde{S}_{\mathbb{Q}}$
        end if
    end if
end for
```

Further, we introduce a strategy that automatically schedules the step size ρ, which can improve the convergence of PGD (Croce & Hein, 2020). We start with step size $\rho^{(0)} = \epsilon$ at iteration 0 and identify whether it is necessary to halve the current step size at checkpoints $c_0, c_1, \ldots, c_n$. We set two conditions:

1. $\sum_{i=c_{j-1}}^{c_j-1} \mathbb{1}_{\hat{\mathcal{F}}(S_{\mathbb{P}}, S_{\mathbb{Q}}^{(i+1)}) < \hat{\mathcal{F}}(S_{\mathbb{P}}, S_{\mathbb{Q}}^{(i)})} < 0.75 \cdot (c_j - c_{j-1})$;
2. $\rho^{(c_{j-1})} \equiv \rho^{(c_j)}$ and $\hat{\mathcal{F}}_{\min}^{(c_{j-1})} \equiv \hat{\mathcal{F}}_{\min}^{(c_j)}$,

where $\hat{\mathcal{F}}_{\min}^{(t)}$ is the lowest value of the test criterion found in the first t iterations. If one of the conditions is triggered, then the step size at iteration $t = c_j$ is halved and $\rho^{(t)} = \rho^{(c_j)}/2$ for every $t \in \{c_j+1, \ldots, c_{j+1}\}$. If at a checkpoint c, the step size gets halved, then we set $S_{\mathbb{Q}}^{(c+1)}$ to the current $\tilde{S}_{\mathbb{Q}}$.

3.3. TST-Agnostic Ensemble Attack

In practice, different TSTs have different formulations of the test criteria. To provide a generic TST-agnostic attack framework, we propose the ensemble attack (EA) that finds the adversarial set $\tilde{S}_{\mathbb{Q}}$ as follows.

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

where $\mathbb{J} = \{\mathcal{J}_1, \mathcal{J}_2, \ldots, \mathcal{J}_n\}$ a set of non-parametric TSTs, $\hat{\mathbb{F}} = \{\hat{\mathcal{F}}^{(\mathcal{J}_1)}, \hat{\mathcal{F}}^{(\mathcal{J}_2)}, \ldots, \hat{\mathcal{F}}^{(\mathcal{J}_n)}\}$ is a set composed of the test criterion for each TST $\mathcal{J}_i \in \mathbb{J}$,

$\mathbb{W} = \{w^{(\mathcal{J}_1)}, w^{(\mathcal{J}_2)}, \ldots, w^{(\mathcal{J}_n)}\}$ is a weight set, and $\sum_{w^{(\mathcal{J}_i)} \in \mathbb{W}} w^{(\mathcal{J}_i)} = 1$. For notational simplicity, we let

$$\ell(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}) = \sum_{w^{(\mathcal{J}_i)} \in \mathbb{W}, \hat{\mathcal{F}}^{(\mathcal{J}_i)} \in \hat{\mathbb{F}}} w^{(\mathcal{J}_i)} \hat{\mathcal{F}}^{(\mathcal{J}_i)}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}).$$

We utilize "PGD with a dynamic schedule of step size ρ" (see above) to realize EA. We summarize the realization of EA in Algorithm 1. Note that an adversarial attack against a TST $\mathcal{J}$ is the special case of EA when we set $w^{(\mathcal{J})} = 1$.

4. Defending Non-Parametric TSTs

In this section, to counteract the threats incurred by adversarial attacks, we propose defensive strategies to enhance the test power of non-parametric TSTs under attacks.

4.1. A Simple Ensemble as A Vanilla Defense

In machine learning, ensemble methods leverage various learning algorithms together to obtain better performance than could be obtained from any of the individual learning algorithms alone (Opitz & Maclin, 1999; Rokach, 2010). Therefore, a simple ensemble of different non-parametric TSTs could be a vanilla defense. Correspondingly, we let the test power of an ensemble of TSTs measure the probability of any non-parametric TST $\mathcal{J}_i \in \mathbb{J}$ correctly rejecting $\mathcal{H}_0$ when $\mathcal{H}_1$ is true, i.e., for a particular $\mathbb{P} \neq \mathbb{Q}$,

$$\mathrm{TP}(\mathbb{J}) = \mathbb{E}_{S_{\mathbb{P}} \sim \mathbb{P}^m, S_{\mathbb{Q}} \sim \mathbb{Q}^n}[\vee_{\mathcal{J}_i \in \mathbb{J}} \mathbb{1}(\mathcal{J}_i(S_{\mathbb{P}}, S_{\mathbb{Q}}) = 1)].$$

However, this simple defense cannot effectively improve the test power of TSTs under EA. We empirically find that EA can significantly degrade the test power of an ensemble of different TSTs (see Table 1). Therefore, the ensemble of TSTs is no longer an effective defensive strategy.

4.2. Adversarially Learning Kernels for TSTs

To effectively enhance the robustness of non-parametric TSTs, we propose a general defense which employs adversarial learning (Madry et al., 2018) to obtain robust kernels for non-parametric TSTs. The learning objective of robust kernels is formulated as a max-min optimization:

$$\hat{\theta} \approx \arg\max_{\theta} \min_{\tilde{S}_{\mathbb{Q}} \in \mathcal{B}_{\epsilon}[S_{\mathbb{Q}}]} \hat{\mathcal{F}}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\theta}). \tag{8}$$

Eq. (8) is equivalent to a minimax optimization problem by simply flipping its inner minimization term and its outer maximization term simultaneously. Then, Danskin's theorem (Danskin, 1966) can apply (Madry et al., 2018). Therefore, we can adversarially learn the deep kernels with one step minimizing the test criterion to find an adversarial pair and one step maximizing the test criterion on the adversarial pair w.r.t. the parameters θ.

Algorithm 2 Adversarially Learning Deep Kernels

```
**Input:** benign pair $(S_{\mathbb{P}}, S_{\mathbb{Q}})$, maximum PGD step $T$, adversarial budget $\epsilon$, checkpoint $\mathbb{C} = \{c_0, \ldots, c_n\}$, deep kernel $k_{\theta}^{(\mathrm{RoD})}$ parameterized by $\theta$, training epochs $E$, learning rate $\eta$
**Output:** parameters of robust deep kernel $\theta$
**for** $e = 1$ **to** $E$ **do**
    $X \leftarrow$ minibatch from $S_{\mathbb{P}}$; $Y \leftarrow$ minibatch from $S_{\mathbb{Q}}$
    Generate an adversarial pair $(X, \tilde{Y})$ by Algorithm 1 with setting $\hat{\mathbb{F}} = \{\hat{\mathcal{F}}^{(\mathrm{RoD})}(\cdot, \cdot; k_{\theta}^{(\mathrm{RoD})})\}$
    $\theta \leftarrow \theta + \eta \nabla_{\theta} \hat{\mathcal{F}}^{(\mathrm{RoD})}(X, \tilde{Y}; k_{\theta}^{(\mathrm{RoD})})$
**end for**
```

Robust deep kernels for TSTs (MMD-RoD). Since MMD-D (Liu et al., 2020a) has been validated as a superior non-parametric TST, our defense is based on the deep kernels, i.e., $\hat{\mathcal{F}}^{(\mathrm{RoD})}(\cdot, \cdot; k_{\theta}^{(\mathrm{RoD})}) = \hat{\mathcal{F}}^{(\mathrm{D})}(\cdot, \cdot; k_{\theta}^{(\mathrm{RoD})})$ where $k_{\theta}^{(\mathrm{RoD})} = k_{\theta}^{(\mathrm{D})}$. We let θ denote all the learnable parameters ($\gamma, \sigma_{\phi}, \sigma_q$ and the parameters of the DNN $\phi(\cdot)$) for a robust deep kernel. We summarize the training procedure of adversarially learning deep kernels in Algorithm 2. The testing procedure of MMD-RoD exactly follows MMD-D (Liu et al., 2020a) and is introduced in Appendix E.3.

5. Experiments

In this section, we empirically uncover the adversarial vulnerabilities of non-parametric TSTs and demonstrate the efficacy of our proposed MMD-RoD in enhancing adversarial robustness of non-parametric TSTs.

5.1. Test Power Evaluated under Ensemble Attacks

We conduct six typical non-parametric TSTs (MMD-D, MMD-G, C2ST-S, C2ST-L, ME and SCF) under EA on five benchmark datasets—Blob (Gretton et al., 2012; Jitkrittum et al., 2016; Sutherland et al., 2017), high-dimensional Gaussian mixture (HDGM) (Liu et al., 2020a), Higgs (Chwialkowski et al., 2015), MNIST (LeCun et al., 1998; Radford et al., 2015) and CIFAR-10 (Krizhevsky, 2009). $\mathbb{P}$ and $\mathbb{Q}$ of each dataset are illustrated in Appendix E.1. Note that $\mathbb{P} \neq \mathbb{Q}$ in each dataset. For Blob, HDGM and Higgs, we randomly sample a training pair ($S_{\mathbb{P}}^{\mathrm{tr}}$, $S_{\mathbb{Q}}^{\mathrm{tr}}$) for learning a kernel once for each non-parametric TST. For MNIST and CIFAR-10, we select a subset of the available data as training data $S_{\mathbb{P}}^{\mathrm{tr}}$ and $S_{\mathbb{Q}}^{\mathrm{tr}}$. The training settings (e.g., the structure of neural network and the optimizer) follow Liu et al. (2020a) and are illustrated in detail in Appendix E.2.

During the testing procedure, we randomly sample 100 new pairs ($S_{\mathbb{P}}^{\mathrm{te}}$, $S_{\mathbb{Q}}^{\mathrm{te}}$), disjoint from the training data, as the benign test pairs. We let n_{tr} and n_{te} be large enough to

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

ensure TSTs can achieve a high test power in benign settings. EA is implemented on each benign test pair and generates the corresponding adversarial test pair as the input for TSTs. We illustrate experimental settings of permutation test in Appendix E.3. Note that we utilize the wild bootstrap process (Chwialkowski et al., 2014) (introduced in Appendix E.3) to resample the value of MMD for MMD-D and MMD-G (as well as MMD-RoD) since adversarial data are probably not IID. Wild bootstrap process guarantees that we can get correct p-values in non-IID/IID scenarios. We repeat the full process 10 times, and report the average test power (comparing $\mathbb{P}$ to $\mathbb{Q}$) of each non-parametric TST as well as an ensemble of these six typical TSTs (denoted as "Ensemble") in Table 1. In addition, we confirm that these TSTs have reasonable Type I errors (comparing $\mathbb{P}$ to $\mathbb{P}$) in Appendix E.5.

EA minimizes a weighted sum of test criteria of six typical TSTs, i.e., $\hat{\mathbb{F}} = \{\hat{\mathcal{F}}^{(\mathrm{D})}, \hat{\mathcal{F}}^{(\mathrm{G})}, \hat{\mathcal{F}}^{(\mathrm{S})}, \hat{\mathcal{F}}^{(\mathrm{L})}, \hat{\mathcal{F}}^{(\mathrm{ME})}, \hat{\mathcal{F}}^{(\mathrm{SCF})}\}$. Weight set $\mathbb{W}$ is manually set for each dataset and is summarized in Table 7 (Appendix E.4). For all datasets, $T = 50$. ϵ for each dataset is summarized in Table 1.

In Table 1, we implement EA in the white-box setting where we can obtain the non-parametric TST's all information (e.g., the kernel parameters). Table 1 demonstrates that the test power of each particular non-parametric TST and even Ensemble are significantly deteriorated among all datasets. It empirically validates that many existing non-parametric TSTs suffer from severe adversarial vulnerabilities.

In addition, we surprisingly find that $\epsilon = 0.05$ is large enough to significantly degrade the test power on MNIST. In contrast, conventional adversarial attacks that aim to fool DNNs on MNIST need a larger adversarial budget ϵ which is up to 0.3 (Madry et al., 2018). It seems that non-parametric TSTs are more adversarially vulnerable than classifiers. However, this claim could be inaccurate for two reasons. First, attack target is different. We target to fool non-parametric TSTs that belong to hypothesis tests, while previous works aim to attack DNN-based classifiers. Second, measurement is different. We cannot fairly compare the non-parametric TST's test power to the classifier's classification accuracy.

5.2. Adversarial Robustness of MMD-RoD

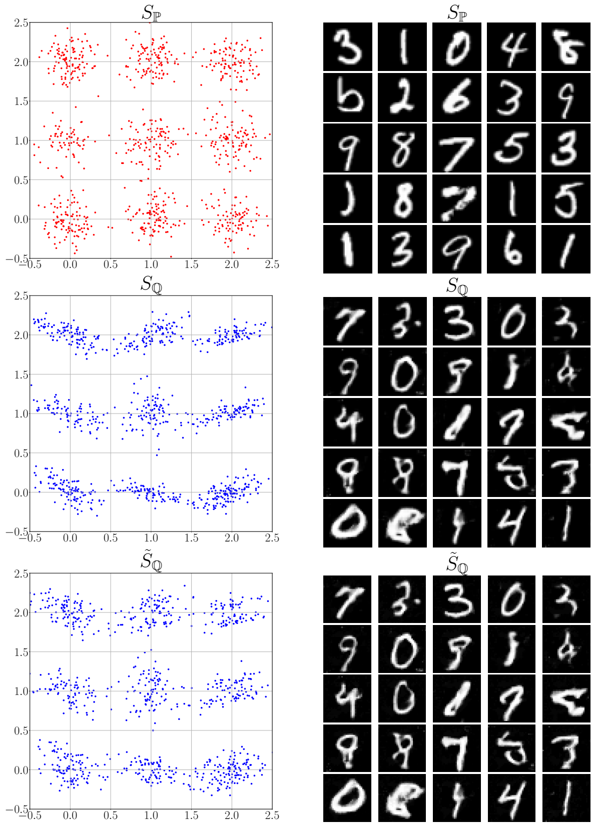
For hyperparameters of adversarially learning kernels, we keep ϵ same as the dataset-corresponding adversarial budget in Table 1, and set $T = 1$ for all datasets. Other training settings such as the structure of the neural network and the optimizer as well as the testing procedure of MMD-RoD exactly follow MMD-D (Liu et al., 2020a). We call an ensemble of six typical TSTs and MMD-RoD as "Ensemble$^+$". Here, EA is conducted based on the test criteria of TSTs in Ensemble$^+$. As for $\mathbb{W}$, we let $w^{(\mathrm{RoD})}$ and $w^{(\mathrm{D})}$ in this section be half of $w^{(\mathrm{D})}$ in Section 5.1. Other attack settings (e.g., $n_{\mathrm{te}}, T, \epsilon$) for each dataset follow Section 5.1. The Type I error of MMD-RoD is reported in Appendix E.5.

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

Table 2 reports the test power of MMD-RoD and Ensemble$^+$ in benign and adversarial settings. Table 2 shows that the test power of MMD-RoD and Ensemble$^+$ under EA are significantly enhanced on most datasets such as MNIST and CIFAR-10, even without sacrificing test power in the benign setting. It validates robust deep kernels can improve adversarial robustness of non-parametric TSTs.

We surprisingly observe in Table 2 that benign test power of MMD-RoD on MNIST and CIFAR-10 remains high while the test power under attacks is significantly improved. This seems to conflict with the robustness-accuracy trade-off in conventional adversarial training (Zhang et al., 2019b). The main reason could be that the metric is different, i.e., test power for non-parametric TSTs v.s. classification accuracy for classifiers. Due to this difference, the trade-off between benign test power and adversarial robustness may not hold in the case of non-parametric TSTs. In addition, there are published papers (Yang et al., 2020a; Pang et al., 2022)



(a) Blob (b) MNIST

Figure 2. Visualization of adversarial test sets.

that claimed there should be no trade-off between benign accuracy and adversarial robustness.

MMD-RoD unexpectedly performs poorly on HDGM and Higgs, which has low test power in both benign and adversarial settings. The poor performance in the benign setting could be attributed to that the most adversarial training pairs can lead to the cross-over mixture problem (Zhang et al., 2020a), thus making the learning extremely difficult and even fail. The reason for the poor robustness could be that the number of training data is small since enhancing adversarial robustness needs more training data (Schmidt et al., 2018). Therefore, we believe that utilizing the style of friendly adversarial training (Zhang et al., 2020a) for learning kernels along with sampling more training data can further enhance the performance of MMD-RoD. We leave further improving MMD-RoD as future work.

5.3. Visualization of Adversarial Test Sets

We visualize benign test set $S_{\mathbb{Q}}$ (middle) and the corresponding adversarial test set $\tilde{S}_{\mathbb{Q}}$ (bottom) on Blob and MNIST in Figure 2 as well as CIFAR-10 in Figure 1. The adversarial data are generated in the experiments illustrated in Section 5.1. Note that the benign test pair we choose to visualize can be correctly judged as samples drawn from different distributions by each TST in Ensemble, and its corresponding adversarial test pair can successfully fool Ensemble. Due to limited space, we visualize only a part of samples from each set. Figure 1-2 verify that the differences between $S_{\mathbb{Q}}$ and $\tilde{S}_{\mathbb{Q}}$ is almost visually indistinguishable to humans, and meanwhile the distribution of $S_{\mathbb{P}}$ is explicitly different from that of $\tilde{S}_{\mathbb{Q}}$. Therefore, Figure 1-2 empirically validate that an ℓ_∞-bound can guarantee the invisibility of adversarial attacks.

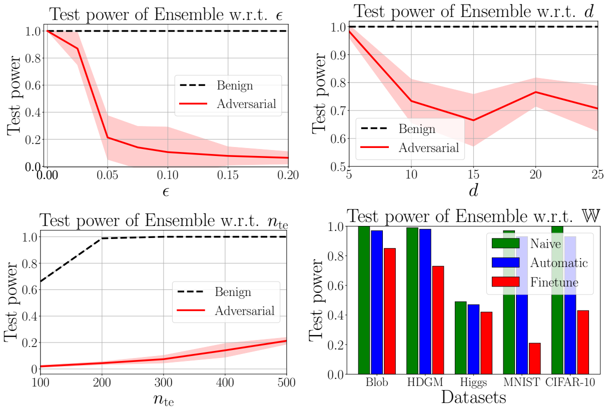


Figure 3. Ablation studies on important hyperparameters.

5.4. Ablation Studies on Important Hyperparameters

In this subsection, we conduct ablation studies on important hyperparameters, including ϵ, d, n_{te} and $\mathbb{W}$. Comprehensive results further validate non-parametric TSTs lack adversarial robustness.

Evaluation with different ϵ. We report the average test power of Ensemble under EA with $\epsilon \in \{0.00, 0.025, 0.05, 0.075, 0.1, 0.15, 0.2\}$ on MNIST. Other settings keep same as Section 5.1. The upper left panel of Figure 3 shows that the test power of Ensemble under EA (red solid line) becomes lower as ϵ increases, and is significantly lower than the test power evaluated in the benign setting (black dash line) over different ϵ, which is in line with the conclusion of Theorem 2.

Evaluation with different d. We evaluate the test power of Ensemble under EA on HDGM with different $d \in \{5, 10, 15, 20, 25\}$. The settings follow Section 5.1 except the dimensionality of Gaussian mixture. The upper right panel of Figure 3 shows that the test power in the adversarial setting (red solid line) decreases as d rises and remains lower than benign test power (black dash line). However, with larger d (e.g., $d > 15$) the test power under EA does not keep degrading and even rises. We believe it is due to that the weight set for EA with larger d is set inappropriately. We discuss the reasons in detail in Appendix E.7.

Evaluation with different n_{te}. We evaluate the test power of Ensemble under EA on MNIST with different

$n_{\mathrm{te}} \in \{100, 200, 300, 400, 500\}$, and use the same settings as Section 5.1. The lower left panel of Figure 3 shows that the test power in the adversarial setting (red solid line) increases as n_{te} becomes larger, but the test power under EA is always severely deteriorated compared to benign test power (black dash line), which reflects that non-parametric TSTs lack adversarial robustness.

Evaluation with different $\mathbb{W}$. We report the test power of Ensemble under EA with three weight strategies: 1) “Naive” (green pillar) denotes that we set $\mathbb{W} = \{1/6, 1/6, 1/6, 1/6, 1/6, 1/6\}$; 2) “Automatic” (blue pillar) denotes that we use the softmax of test criterion for each test as $\mathbb{W}$ at each PGD iteration, i.e., $w^{(\mathcal{J}_i)} = \frac{\exp(\hat{\mathcal{F}}^{(\mathcal{J}_i)})}{\sum_{j=1}^{n} \exp(\hat{\mathcal{F}}^{(\mathcal{J}_j)})}$; 3) “Finetune” (red pillar) denotes that we set manually-finetuned $\mathbb{W}$ for each dataset. The finetuned weight set is summarized in Appendix E.2. Other settings follow Section 5.1. The lower right panel of Figure 3 shows that the test power of Ensemble under EA can be severely deteriorated with an appropriate weight strategy.

5.5. Transferability of Adversarial Attacks

Further, we empirically demonstrate that our proposed EA against non-parametric TSTs has transferability.

Transferability between different types of non-parametric TSTs. We report test power of non-parametric TSTs under the adversarial attack against a certain type of TSTs on MNIST in Figure 4(a) and the test power of non-parametric TSTs under EA against a TST ensemble composed by leaving one TST out of Ensemble on MNIST in Figure 4(b). The experimental details and results are in Appendix E.6. Figure 4(a) shows that attacks against a certain type of TST sometimes can fool other types of TSTs. Figure 4(b) demonstrates that attacks against an ensemble of TSTs sometimes can successfully fool TSTs that are not included in the attack ensemble. Therefore, Figure 4 validates our proposed EA has transferability between different types of non-parametric TSTs.

Transferability between target and surrogate non-parametric TSTs. Here, we assume that the attacker cannot obtain the target non-parametric TST’s kernel parameters and training data, and it only knows the target non-parametric TST’s test criterion (including its kernel function). We generate adversarial pairs via EA based on an ensemble of surrogate non-parametric TSTs on MNIST (other attack configurations follow Section 5.1) and then report the average test power of target tests on these adversarial pairs in Table 3. Surrogate tests are trained on the training data with different random seeds. Table 3 shows that the test power of each target non-parametric TST and Ensemble are deteriorated under EA based on surrogate non-parametric TSTs, which further validates that existing non-parametric TSTs are adversarially vulnerable.

Table 3. Transferability between target and surrogate non-parametric TSTs.

MMD-D	MMD-G	C2ST-S	C2ST-L	ME	SCF	Ensemble
0.564±0.09	0.149±0.00	0.418±0.03	0.471±0.04	0.064±0.01	0.001±0.00	0.751±0.01

Transferability between different test sets drawn from $\mathbb{P}$. We replace the set $S_{\mathbb{P}}$ with $S'_{\mathbb{P}}$ where $S'_{\mathbb{P}}$ is drawn from the distribution $\mathbb{P}$ with different random seeds (i.e., $S_{\mathbb{P}} \neq S'_{\mathbb{P}}$). $\tilde{S}_{\mathbb{Q}}$ is generated by EA on the benign test pair $(S_{\mathbb{P}}, S_{\mathbb{Q}})$. We report the average test power of non-parametric TSTs on $(S'_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}})$ under EA on MNIST (details follow Section 5.1) in Table 4. Table 4 shows that EA still hurts the test power of non-parametric TSTs on $(S'_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}})$, and implies that EA has a good transferability property between different test sets drawn from $\mathbb{P}$.

Table 4. Transferability between different test sets drawn form $\mathbb{P}$.

MMD-D	MMD-G	C2ST-S	C2ST-L	ME	SCF	Ensemble
0.166±0.05	0.201±0.00	0.013±0.00	0.018±0.00	0.270±0.03	0.017±0.01	0.486±0.04

6. Conclusions

This paper systematically studies adversarial robustness of non-parametric TSTs. We propose a generic ensemble attack framework which reveals non-parametric TSTs are adversarially vulnerable.To counteract these risks, we propose to adversarially learn kernels for non-parametric TSTs. We empirically show that SOTA non-parametric TSTs can fail catastrophically under adversarial attacks, and our proposed MMD-RoD can substantially enhance the adversarial robustness of non-parametric TSTs. We believe our work makes people aware of potential risks when they apply non-parametric TSTs to critical applications.

One of the limitations of our current work is that our proposed attack method is computationally heavy and user-dependent, in that it needs very large GPU memory when n_{te} is too large and the weight set needs to be manually finetuned. Future research includes (a) how to fool non-parametric TSTs by perturbing fewer samples, (b) how to adaptively adjust the weight set at each PGD iteration.

Acknowledgements

Jingfeng Zhang was supported by JST, ACT-X Grant Number JPMJAX21AF. Masashi Sugiyama was supported by JST AIP Acceleration Research Grant Number JPMJCR20U3 and the Institute for AI and Beyond, UTokyo. Mohan Kankanhalli’s research is supported by the National Research Foundation, Singapore under its Strategic Capability 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.

References

Alaifari, R., Alberti, G. S., and Gauksson, T. Adef: an iterative algorithm to construct adversarial deformations. In *ICLR*, 2019.

Amsaleg, L., Bailey, J., Barbe, D., Erfani, S., Houle, M. E., Nguyen, V., and Radovanović, M. The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality. In *2017 IEEE Workshop on Information Forensics and Security (WIFS)*, pp. 1–6. IEEE, 2017.

Andriushchenko, M., Croce, F., Flammarion, N., and Hein, M. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pp. 484–501. Springer, 2020.

Athalye, A., Carlini, N., and Wagner, D. A. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018.

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.

Bhattacharjee, R. and Chaudhuri, K. When are non-parametric methods robust? In *International Conference on Machine Learning*, pp. 832–841. PMLR, 2020.

Bhattacharjee, R. and Chaudhuri, K. Consistent non-parametric methods for maximizing robustness. *Advances in Neural Information Processing Systems*, 34, 2021.

Biggio, B. and Roli, F. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

Borgwardt, K. M., Gretton, A., Rasch, M. J., Kriegel, H.-P., Schölkopf, B., and Smola, A. J. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

Cai, Q., Liu, C., and Song, D. Curriculum adversarial training. In *IJCAI*, 2018.

Carlini, N. and Wagner, D. A. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy (SP)*, 2017.

Carmon, Y., Raghunathan, A., Schmidt, L., Liang, P., and Duchi, J. C. Unlabeled data improves adversarial robustness. In *NeurIPS*, 2019.

Chen, H. and Friedman, J. H. A new graph-based two-sample test for multivariate and object data. *Journal of the American statistical association*, 112(517):397–409, 2017.

Chen, H., Zhang, H., Boning, D. S., and Hsieh, C.-J. Robust decision trees against adversarial examples. In *ICML*, pp. 1122–1131, 2019.

Chen, J., Jordan, M. I., and Wainwright, M. J. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 ieee symposium on security and privacy (sp)*, pp. 1277–1294. IEEE, 2020.

Chen, P.-Y., Sharma, Y., Zhang, H., Yi, J., and Hsieh, C.-J. Ead: elastic-net attacks to deep neural networks via adversarial examples. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Chen, T., Zhang, Z., Liu, S., Chang, S., and Wang, Z. Robust overfitting may be mitigated by properly learned smoothening. In *ICLR*, 2021.

Cheng, M., Le, T., Chen, P.-Y., Zhang, H., Yi, J., and Hsieh, C.-J. Query-efficient hard-label black-box attack: An optimization-based approach. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJlk6iRqKX.

Cheng, M., Singh, S., Chen, P. H., Chen, P.-Y., Liu, S., and Hsieh, C.-J. Sign-opt: A query-efficient hard-label adversarial attack. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SklTQCNtvS.

Cheng, X. and Cloninger, A. Classification logit two-sample testing by neural networks. *arXiv preprint arXiv:1909.11298*, 2019.

Chwialkowski, K. P., Sejdinovic, D., and Gretton, A. A wild bootstrap for degenerate kernel tests. In *Advances in neural information processing systems*, pp. 3608–3616, 2014.

Chwialkowski, K. P., Ramdas, A., Sejdinovic, D., and Gretton, A. Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, 28:1981–1989, 2015.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020.

Croce, F., Andriushchenko, M., Sehwag, V., Debenedetti, E., Flammarion, N., Chiang, M., Mittal, P., and Hein, M. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Danskin, J. M. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, 14(4):641–664, 1966.

Ding, G. W., Sharma, Y., Lui, K. Y. C., and Huang, R. Mma training: Direct input space margin maximization through adversarial training. In *ICLR*, 2020.

Dong, M., Li, Y., Wang, Y., and Xu, C. Adversarially robust neural architectures. *arXiv preprint arXiv:2009.00902*, 2020.

Dwass, M. Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, pp. 181–187, 1957.

Erdemir, E., Bickford, J., Melis, L., and Aydore, S. Adversarial robustness with non-uniform perturbations. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=oi08QWKs84.

Fang, T., Lu, N., Niu, G., and Sugiyama, M. Rethinking importance weighting for deep learning under distribution shift. In *NeurIPS*, 2020a.

Fang, Z., Lu, J., Liu, F., Xuan, J., and Zhang, G. Open set domain adaptation: Theoretical bound and algorithm. *IEEE transactions on neural networks and learning systems*, 2020b.

Feinman, R., Curtin, R. R., Shintre, S., and Gardner, A. B. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

Fernández, V. A., Gamero, M. J., and Garcia, J. M. A test for the two-sample problem based on empirical characteristic functions. *Computational statistics & data analysis*, 52 (7):3730–3748, 2008.

Gao, R., Xie, L., Xie, Y., and Xu, H. Robust hypothesis testing using wasserstein uncertainty sets. In *NeurIPS*, pp. 7913–7923, 2018.

Gao, R., Liu, F., Zhang, J., Han, B., Liu, T., Niu, G., and Sugiyama, M. Maximum mean discrepancy test is aware of adversarial attacks. In *International Conference on Machine Learning*, pp. 3564–3575. PMLR, 2021.

Ghoshdastidar, D., Gutzeit, M., Carpentier, A., and von Luxburg, U. Two-sample tests for large random graphs using network statistics. In *Conference on Learning Theory*, pp. 954–977. PMLR, 2017.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., and Schölkopf, B. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pp. 2839–2848. PMLR, 2016.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

Gowal, S., Rebuffi, S.-A., Wiles, O., Stimberg, F., Calian, D. A., and Mann, T. A. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 34, 2021.

Gretton, A., Fukumizu, K., Harchaoui, Z., and Sriperumbudur, B. K. A fast, consistent kernel two-sample test. In *NIPS*, volume 23, pp. 673–681, 2009.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Grosse, K., Manoharan, P., Papernot, N., Backes, M., and McDaniel, P. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.

Gül, G. and Zoubir, A. M. Minimax robust hypothesis testing. *IEEE Transactions on Information Theory*, 63(9): 5572–5587, 2017.

Hein, M. and Andriushchenko, M. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *NIPS*, pp. 2263–2273, 2017.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021.

Huber, P. J. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

Ilyas, A., Engstrom, L., Athalye, A., and Lin, J. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pp. 2137–2146. PMLR, 2018.

Jitkrittum, W., Szabó, Z., Chwialkowski, K. P., and Gretton, A. Interpretable distribution features with maximum testing power. *Advances in Neural Information Processing Systems*, 29:181–189, 2016.

Kanth Nakka, K. and Salzmann, M. Learning transferable adversarial perturbations. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

Kim, J., Lee, B.-K., and Ro, Y. M. Distilling robust and non-robust features in adversarial examples by information bottleneck. *Advances in Neural Information Processing Systems*, 34, 2021.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. URL http://arxiv.org/abs/1412.6980.

Kirchler, M., Khorasani, S., Kloft, M., and Lippert, C. Two-sample testing using deep learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1387–1398. PMLR, 2020.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Leucht, A. and Neumann, M. H. Dependent wild bootstrap for degenerate u-and v-statistics. *Journal of Multivariate Analysis*, 117:257–280, 2013.

Levy, B. C. Robust hypothesis testing with a relative entropy tolerance. *IEEE Transactions on Information Theory*, 55 (1):413–421, 2008.

Li, S. and Wang, X. Fully distributed sequential hypothesis testing: Algorithms and asymptotic analyses. *IEEE Transactions on Information Theory*, 64(4):2742–2758, 2018.

Liu, F., Lu, J., Han, B., Niu, G., Zhang, G., and Sugiyama, M. Butterfly: A panacea for all difficulties in wildly unsupervised domain adaptation. In *NeurIPS LTS Workshop*, 2019.

Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning*, pp. 6316–6326. PMLR, 2020a.

Liu, F., Zhang, G., and Lu, J. Multi-source heterogeneous unsupervised domain adaptation via fuzzy-relation neural networks. *IEEE Transactions on Fuzzy Systems*, 2020b.

Liu, F., Xu, W., Lu, J., and Sutherland, D. J. Meta two-sample testing: Learning kernels for testing with limited data. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=EUlAerrk47Y.

Lopez-Paz, D. and Oquab, M. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*, 2016.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Mahmood, K., Mahmood, R., and Van Dijk, M. On the robustness of vision transformers to adversarial examples. *arXiv preprint arXiv:2104.02610*, 2021.

Metzen, J. H., Genewein, T., Fischer, V., and Bischoff, B. On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Mopuri, K. R., Ganeshan, A., and Babu, R. V. Generalizable data-free objective for crafting universal adversarial perturbations. *IEEE transactions on pattern analysis and machine intelligence*, 41(10):2452–2465, 2018.

Oneto, L., Donini, M., Luise, G., Ciliberto, C., Maurer, A., and Pontil, M. Exploiting mmd and sinkhorn divergences for fair and transferable representation learning. In *NeurIPS*, 2020.

Opitz, D. and Maclin, R. Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, 11:169–198, 1999.

Pang, T., Xu, K., Du, C., Chen, N., and Zhu, J. Improving adversarial robustness via promoting ensemble diversity. In *ICML*, 2019.

Pang, T., Yang, X., Dong, Y., Su, H., and Zhu, J. Bag of tricks for adversarial training. *ICLR*, 2021.

Pang, T., Lin, M., Yang, X., Zhu, J., and Yan, S. Robustness and accuracy could be reconcilable by (proper) definition. *arXiv preprint arXiv:2202.10103*, 2022.

Papernot, N., McDaniel, P., Sinha, A., and Wellman, M. Towards the science of security and privacy in machine learning. *arXiv:1611.03814*, 2016.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

Peng, S. and Mine, T. A robust hierarchical graph convolutional network model for collaborative filtering. *arXiv preprint arXiv:2004.14734*, 2020.

Poggio, T. and Shelton, C. R. On the mathematical foundations of learning. *American Mathematical Society*, 39(1): 1–49, 2002.

Qin, C., Martens, J., Gowal, S., Krishnan, D., Dvijotham, K., Fawzi, A., De, S., Stanforth, R., and Kohli, P. Adversarial robustness through local linearization. In *NeurIPS*, 2019.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Rahmati, A., Moosavi-Dezfooli, S.-M., Frossard, P., and Dai, H. Geoda: a geometric framework for black-box adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8446–8455, 2020.

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.

Rebuffi, S.-A., Gowal, S., Calian, D. A., Stimberg, F., Wiles, O., and Mann, T. A. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34, 2021.

Robey, A., Chamon, L., Pappas, G., Hassani, H., and Ribeiro, A. Adversarial robustness with semi-infinite constrained learning. *Advances in Neural Information Processing Systems*, 34, 2021.

Rokach, L. Ensemble-based classifiers. *Artificial intelligence review*, 33(1):1–39, 2010.

Sarkar, A., Sarkar, A., Gali, S., and Balasubramanian, V. N. Adversarial robustness without adversarial training: A teacher-guided curriculum learning approach. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=MqCzSKCQ1QB.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *NeurIPS*, 2018.

Sehwag, V., Wang, S., Mittal, P., and Jana, S. Hydra: Pruning adversarially robust neural networks. *NeurIPS*, 2020.

Serfling, R. J. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 2009.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! In *NeurIPS*, 2019.

Sitawarin, C. and Wagner, D. On the robustness of deep k-nearest neighbors. In *2019 IEEE Security and Privacy Workshops (SPW)*, pp. 1–7. IEEE, 2019.

Song, C., He, K., Wang, L., and Hopcroft, J. E. Improving the generalization of adversarial training with domain adaptation. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyfIfnC5Ym.

Sriramanan, G., Addepalli, S., Baburaj, A., and Babu, R. V. Guided adversarial attack for evaluating and enhancing adversarial defenses. *arXiv preprint arXiv:2011.14969*, 2020.

Sriramanan, G., Addepalli, S., Baburaj, A., et al. Towards efficient and effective adversarial training. *Advances in Neural Information Processing Systems*, 34, 2021.

Stojanov, P., Gong, M., Carbonell, J., and Zhang, K. Data-driven approach to multiple-source domain adaptation. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3487–3496. PMLR, 2019.

Sugiyama, M., Suzuki, T., Itoh, Y., Kanamori, T., and Kimura, M. Least-squares two-sample test. *Neural networks*, 24(7):735–751, 2011.

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.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *ICLR*, 2014.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I. J., Boneh, D., and McDaniel, P. D. Ensemble adversarial training: Attacks and defenses. In *ICLR*, 2018.

Wang, Q., Liu, F., Han, B., Liu, T., Gong, C., Niu, G., Zhou, M., and Sugiyama, M. Probabilistic margins for instance reweighting in adversarial training. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=rg8gNkvs3u.

Wang, Y., Jha, S., and Chaudhuri, K. Analyzing the robustness of nearest neighbors to adversarial examples. In *International Conference on Machine Learning*, pp. 5133–5142. PMLR, 2018.

Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., and Gu, Q. On the convergence and robustness of adversarial training. In *ICML*, 2019.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., and Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2020.

Wong, E. and Kolter, J. Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *ICML*, 2018.

Wong, E., Schmidt, F., and Kolter, Z. Wasserstein adversarial examples via projected sinkhorn iterations. In *International Conference on Machine Learning*, pp. 6808–6817. PMLR, 2019.

Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. In *ICLR*, 2020.

Wu, D., Wang, Y., Xia, S.-T., Bailey, J., and Ma, X. Skip connections matter: On the transferability of adversarial examples generated with resnets. In *International Conference on Learning Representations*, 2020a.

Wu, D., Xia, S.-T., and Wang, Y. Adversarial weight perturbation helps robust generalization. *NeurIPS*, 33, 2020b.

Xiao, C., Zhu, J.-Y., Li, B., He, W., Liu, M., and Song, D. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HyydRMZC-.

Xie, C., Wang, J., Zhang, Z., Zhou, Y., Xie, L., and Yuille, A. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1369–1378, 2017.

Xie, L., Gao, R., and Xie, Y. Robust hypothesis testing with wasserstein uncertainty sets. *arXiv preprint arXiv:2105.14348*, 2021.

Yan, Z., Guo, Y., and Zhang, C. Deep defense: Training dnns with improved adversarial robustness. In *NeurIPS*, pp. 417–426, 2018.

Yan, Z., Guo, Y., Liang, J., and Zhang, C. Policy-driven attack: Learning to query for hard-label black-box adversarial examples. In *International Conference on Learning Representations*, 2020.

Yang, Y., Rashtchian, C., Wang, Y., and Chaudhuri, K. Adversarial examples for non-parametric methods: Attacks, defenses and large sample limits. *arXiv preprint arXiv:1906.03310*, 2019.

Yang, Y., Rashtchian, C., Zhang, H., Salakhutdinov, R. R., and Chaudhuri, K. A closer look at accuracy vs. robustness. In *NeurIPS*, 2020a.

Yang, Y.-Y., Rashtchian, C., Wang, Y., and Chaudhuri, K. Robustness for non-parametric classification: A generic attack and defense. In *International Conference on Artificial Intelligence and Statistics*, pp. 941–951. PMLR, 2020b.

Yao, C., Bielik, P., TSANKOV, P., and Vechev, M. Automated discovery of adaptive attacks on adversarial defenses. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=nWz-Si-uTzt.

Yoo, J. Y. and Qi, Y. Towards improving adversarial training of nlp models. *arXiv preprint arXiv:2109.00544*, 2021.

Yu, Y., Gao, X., and Xu, C.-Z. Lafeat: Piercing through adversarial defenses with latent features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5735–5745, 2021.

Zhang, D., Zhang, T., Lu, Y., Zhu, Z., and Dong, B. You only propagate once: Accelerating adversarial training via maximal principle. In *NeurIPS*, 2019a.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019b.

Zhang, J., Xu, X., Han, B., Niu, G., Cui, L., Sugiyama, M., and Kankanhalli, M. Attacks which do not kill training make adversarial learning stronger. In *ICML*, 2020a.

Zhang, J., Zhu, J., Niu, G., Han, B., Sugiyama, M., and Kankanhalli, M. Geometry-aware instance-reweighted adversarial training. In *ICLR*, 2021.

Zhang, T., Yamane, I., Lu, N., and Sugiyama, M. A one-step approach to covariate shift adaptation. In *Asian Conference on Machine Learning*, pp. 65–80. PMLR, 2020b.

Zhang, Y., Liu, F., Fang, Z., Yuan, B., Zhang, G., and Lu, J. Clarinet: A one-step approach towards budget-friendly unsupervised domain adaptation. In *IJCAI*, pp. 2526–2532, 2020c. URL https://doi.org/10.24963/ijcai.2020/350.

Zheng, T., Chen, C., and Ren, K. Distributionally adversarial attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2253–2260, 2019.

Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., and Liu, J. Freelb: Enhanced adversarial training for natural language understanding. In *ICLR*, 2020.

Zou, D., Frei, S., and Gu, Q. Provable robustness of adversarial training for learning halfspaces with noise. In *ICML*, pp. 13002–13011, 2021. URL http://proceedings.mlr.press/v139/zou21a.html.

A. Notation Table

Table 5. A notation table in convenience for viewing.

Notation	Description
$\mathcal{J}$	The non-parametric TST
$\mathbb{J}$	The set of non-parametric TSTs
$\mathcal{H}_0$	The null hypothesis
$\mathcal{H}_1$	The alternative hypothesis
α	The significance level
r	The rejection threshold
TP	The measurement function for the test power
$\mathcal{D}$	The test statistic function
$\hat{\mathcal{F}}$	The test criterion function
$\hat{\mathbb{F}}$	The set of test criterion functions
d	The dimensionality of data
$\mathcal{X}$	The data feature space $\subset \mathbb{R}^d$
$\mathbb{P}$	The Borel probability measure on $\mathcal{X}$
$\mathbb{Q}$	The Borel probability measure on $\mathcal{X}$
$\mathbb{P}^m$	The joint probability distribution $\mathbb{P}^m = \overbrace{\mathbb{P} \times \mathbb{P} \times \ldots \times \mathbb{P}}^{m}$
$\mathbb{Q}^n$	The joint probability distribution $\mathbb{Q}^n = \overbrace{\mathbb{Q} \times \mathbb{Q} \times \ldots \times \mathbb{Q}}^{n}$
$S_{\mathbb{P}}$	The set $S_{\mathbb{P}} = \{x_i\}_{i=1}^m \sim \mathbb{P}^m$
$S_{\mathbb{Q}}$	The set $S_{\mathbb{Q}} = \{y_i\}_{i=1}^n \sim \mathbb{Q}^n$
n_{tr}	The number of training samples drawn from a particular distribution
n_{te}	The number of testing samples drawn from a particular distribution
k	The kernel function
θ	The kernel parameter
κ	The dimensionality of the kernel parameter
Θ	The set of kernel parameters
R_{Θ}	A positive constant that bounds the kernel parameter $\theta \in \Theta$
s	A positive constant used in defining $\bar{\Theta}_s$
$\bar{\Theta}_s$	A set of kernel parameters $\bar{\Theta}_s = \{\theta \in \Theta \mid \sigma_\theta^2 \geq s^2 > 0\}$
ν	A constant that uniformly bounds the kernel function
L_1	Lipschitz constant of the kernel function
L_2	Lipschitz constant of the kernel function
λ	A constant $\in (0, 1)$ used in calculating $\hat{\sigma}_{\mathcal{H}_1,\lambda}$ (in Eq. (5))
$\tilde{S}_{\mathbb{Q}}$	The adversarial data corresponding to $S_{\mathbb{Q}}$
ϵ	The size of adversarial budget
T	The maximum PGD step
ρ	The step size
w	The weight for the test criterion function
$\mathbb{W}$	The weight set
$\mathbb{C}$	The checkpoint set
E	The number of training epoch
η	The learning rate of the optimizer
f	A classifier that outputs classification probabilities
ϕ	A neural network
γ	A learnable parameter in the deep kernel
σ_ϕ	A learnable parameter in the deep kernel
G	The number of test locations
$\mathcal{V}$	The set of test locations

B. Theoretical Analysis

All the proofs are inspired by Liu et al. (2020a).

B.1. Uniform Convergence Results

These results, on the uniform convergence of $\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta)$ and $\hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta)$, were used in the proof of Theorem 2.

Proposition 1 (Restated). 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 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

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

Proof of Proposition 1. We study the random error function

$$\Delta(\theta) = \widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) - \widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta).$$

First, we choose P points $\{\theta_i\}_{i=1}^P$ such that any $\theta_i \in \Theta$ and $\min_i \|\theta - \theta_i\| \leq q$; Assumption 1 ensures this is possible with at most $P = (4\mathrm{R}_\Theta/q)^\kappa$ points (Poggio & Shelton, 2002).

We define $\tilde{H}_{ij} = k(x_i, x_j) + k(\tilde{y}_i, \tilde{y}_j) - k(x_i, \tilde{y}_j) - k(x_j, \tilde{y}_i)$ where $x_i, x_j \in S_{\mathbb{P}}$ and $\tilde{y}_i, \tilde{y}_j \in \tilde{S}_{\mathbb{Q}}$. Note that $\tilde{y}_i = y_i + \zeta_i$ for any $y_i \in S_{\mathbb{Q}}$ and $\tilde{y}_i \in \tilde{S}_{\mathbb{Q}}$ where ζ_i is an adversarial perturbation under an ℓ_∞-bound of size ϵ. Correspondingly, $\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) = \frac{1}{n(n-1)}\sum_{i\neq j}\tilde{H}_{ij}$. Via Assumption 3 we know that $|\tilde{H}_{ij} - H_{ij}| \leq 4L_2\epsilon\sqrt{d}$.

Because $\tilde{S}_{\mathbb{Q}} \in \mathcal{B}_\epsilon[S_{\mathbb{Q}}]$, it holds that $|\tilde{H}_{ij} - H_{ij}| \to 0$ when $\epsilon \to 0$. Therefore, we have $\mathbb{E}\Delta \to 0$. Recall that $\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta) = \frac{1}{n(n-1)}\sum_{i\neq j}H_{ij}$. If we replace (x_1, y_1) with $(x_1{}', y_1{}')$, we can obtain $\widehat{\mathrm{MMD}'}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta) = \frac{1}{n(n-1)}\sum_{i\neq j}F_{ij}$ and $\widehat{\mathrm{MMD}'}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) = \frac{1}{n(n-1)}\sum_{i\neq j}\tilde{F}_{ij}$, where F (or $\tilde{F}$) agrees with H (or $\tilde{H}$) except when i or j is 1. Then, we have

$$\begin{aligned}
&|\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) - \widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta) - (\widehat{\mathrm{MMD}'}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) - \widehat{\mathrm{MMD}'}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta))| \\
&\leq \frac{1}{n(n-1)}|\sum_{i>1}(\tilde{H}_{i1} - H_{i1} - (\tilde{F}_{i1} - F_{i1})) + \sum_{j>1}(\tilde{H}_{1j} - H_{1j} - (\tilde{F}_{1j} - F_{1j}))| \\
&\leq \frac{1}{n(n-1)}\left(\sum_{i>1}|\tilde{H}_{i1} - H_{i1}| + \sum_{i>1}|(\tilde{F}_{i1} - F_{i1})| + \sum_{j>1}|\tilde{H}_{1j} - H_{1j}| + \sum_{j>1}|\tilde{F}_{1j} - F_{1j}|\right) \\
&\leq \frac{16L_2\epsilon\sqrt{d}}{n}.
\end{aligned}$$

Using McDiarmid's inequality for each $\Delta(\theta_i)$ and a union bound, we then obtain that with probability at least $1-\delta$,

$$\max_{i\in\{1,2,\ldots,P\}}\Delta(\theta) \leq \frac{16L_2\epsilon\sqrt{d}}{\sqrt{2n}}\sqrt{\log\frac{2P}{\delta}} \leq \frac{8L_2\epsilon\sqrt{d}}{\sqrt{n}}\sqrt{2\log\frac{2}{\delta} + 2\kappa\log\frac{4\mathrm{R}_\Theta}{q}}.$$

Via Assumption 3, for any two $\theta, \theta' \in \Theta$, we also have

$$\begin{aligned}
|\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) - \widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\theta'})| &\leq \frac{1}{n(n-1)}\sum_{i\neq j}|\tilde{H}_{ij}^{(\theta)} - \tilde{H}_{ij}^{(\theta')}| \\
&\leq \frac{1}{n(n-1)}\sum_{i\neq j}4L_1\|\theta - \theta'\| = 4L_1\|\theta - \theta'\| \leq 4L_1 q.
\end{aligned}$$

Similarly, $|\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta) - \widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_{\theta'})| \le 4L_1 q$.

Combining these two results, we know that with probability at least $1-\delta$,

$$\sup_\theta |\Delta(\theta)| \le \max_{i\in\{1,2,\dots,P\}} \Delta(\theta) + 8L_1 q \le \frac{8L_2\epsilon\sqrt{d}}{\sqrt{n}}\sqrt{2\log\frac{2}{\delta} + 2\kappa\log\frac{4\mathrm{R}_\Theta}{q}} + 8L_1 q.$$

Since the adversary perturbs benign test pairs, we let $n = n_{\mathrm{te}}$ and $q = \frac{1}{\sqrt{n_{\mathrm{te}}}}$, thus yielding the desired results. □

Proposition 2. *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 adversarial budget $\epsilon \ge 0$, 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*

$$\sup_\theta |\hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) - \hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)| \le \frac{1024\nu L_2\epsilon\sqrt{d}}{\sqrt{n_{\mathrm{te}}}}\sqrt{2\log\frac{2}{\delta} + 2\kappa\log(4\mathrm{R}_\Theta\sqrt{n_{\mathrm{te}}})} + \frac{512L_1\nu}{\sqrt{n_{\mathrm{te}}}}.$$

Proof of Propositon 2. We study the random error function

$$\Delta(\theta) = \hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) - \hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta).$$

Note that $\hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta) = \frac{4}{n^3}\sum_{i=1}^n\Big(\sum_{j=1}^n H_{ij}\Big)^2 - \frac{4}{n^4}\Big(\sum_{i=1}^n\sum_{j=1}^n H_{ij}\Big)^2 + \lambda$, and $\hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) = \frac{4}{n^3}\sum_{i=1}^n\Big(\sum_{j=1}^n \tilde{H}_{ij}\Big)^2 - \frac{4}{n^4}\Big(\sum_{i=1}^n\sum_{j=1}^n \tilde{H}_{ij}\Big)^2 + \lambda$.

Because $\tilde{S}_{\mathbb{Q}} \in \mathcal{B}_\epsilon[S_{\mathbb{Q}}]$, it holds that $|\tilde{H}_{ij} - H_{ij}| \to 0$ when $\epsilon \to 0$. Therefore, we have $\mathbb{E}\Delta \to 0$.

If we replace (x_1, y_1) with (x_1', y_1'), we can obtain $\hat{\sigma'}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta) = \frac{4}{n^3}\sum_{i=1}^n\Big(\sum_{j=1}^n F_{ij}\Big)^2 - \frac{4}{n^4}\Big(\sum_{i=1}^n\sum_{j=1}^n F_{ij}\Big)^2 + \lambda$ and $\hat{\sigma'}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) = \frac{4}{n^3}\sum_{i=1}^n\Big(\sum_{j=1}^n \tilde{F}_{ij}\Big)^2 - \frac{4}{n^4}\Big(\sum_{i=1}^n\sum_{j=1}^n \tilde{F}_{ij}\Big)^2 + \lambda$, where F (or $\tilde{F}$) agrees with H (or $\tilde{H}$) except when i or j is 1. Via Assumption 2, we have $|H_{ij}| \le 4\nu$. Then, we have

$$\begin{aligned}
&|\hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) - \hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta) - (\hat{\sigma'}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) - \hat{\sigma'}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta))| \\
&\le \frac{4}{n^3}\Bigg|\sum_{i=1}^n\Bigg[\Big(\sum_{j=1}^n \tilde{H}_{ij}\Big)^2 - \Big(\sum_{j=1}^n H_{ij}\Big)^2 - \Big(\sum_{j=1}^n \tilde{F}_{ij}\Big)^2 + \Big(\sum_{j=1}^n F_{ij}\Big)^2\Bigg]\Bigg| \\
&\quad + \frac{4}{n^4}\Bigg|\Bigg[\Big(\sum_{i=1}^n\sum_{j=1}^n \tilde{H}_{ij}\Big)^2 - \Big(\sum_{i=1}^n\sum_{j=1}^n H_{ij}\Big)^2 - \Big(\sum_{i=1}^n\sum_{j=1}^n \tilde{F}_{ij}\Big)^2 + \Big(\sum_{i=1}^n\sum_{j=1}^n F_{ij}\Big)^2\Bigg]\Bigg| \\
&\le \frac{4}{n^3}\Bigg|\sum_{i=1}^n\Bigg[\Big(\sum_{j=1}^n \tilde{H}_{ij} - \sum_{j=1}^n \tilde{F}_{ij}\Big)\Big(\sum_{j=1}^n \tilde{H}_{ij} + \sum_{j=1}^n \tilde{F}_{ij}\Big) - \Big(\sum_{j=1}^n H_{ij} - \sum_{j=1}^n F_{ij}\Big)\Big(\sum_{j=1}^n H_{ij} + \sum_{j=1}^n F_{ij}\Big)\Bigg]\Bigg| \\
&\quad + \frac{4}{n^4}\Bigg|\Big(\sum_{ij} \tilde{H}_{ij} - \sum_{ij} \tilde{F}_{ij}\Big)\Big(\sum_{ij} \tilde{H}_{ij} + \sum_{ij} \tilde{F}_{ij}\Big) - \Big(\sum_{ij} H_{ij} - \sum_{ij} F_{ij}\Big)\Big(\sum_{ij} H_{ij} + \sum_{ij} F_{ij}\Big)\Bigg| \\
&\le \frac{4}{n^3}\Bigg|\Big(\sum_{j=1}^n \tilde{H}_{1j} - \sum_{j=1}^n \tilde{F}_{1j}\Big)\Big(\sum_{j=1}^n \tilde{H}_{1j} + \sum_{j=1}^n \tilde{F}_{1j}\Big) + \sum_{i>1}\Big(\tilde{H}_{i1} - \tilde{F}_{i1}\Big)\Big(\sum_{j=1}^n \tilde{H}_{ij} + \sum_{j=1}^n \tilde{F}_{ij}\Big) \\
&\quad - \Big(\sum_{j=1}^n H_{1j} - \sum_{j=1}^n F_{1j}\Big)\Big(\sum_{j=1}^n H_{1j} + \sum_{j=1}^n F_{1j}\Big) - \sum_{i>1}\Big(H_{i1} - F_{i1}\Big)\Big(\sum_{j=1}^n H_{ij} + \sum_{j=1}^n F_{ij}\Big)\Bigg| \\
&\quad + \frac{4}{n^4}\Bigg|\sum_{ij} \tilde{H}_{ij} - \sum_{ij} \tilde{F}_{ij}\Bigg| \cdot \Bigg|\sum_{ij} \tilde{H}_{ij} + \sum_{ij} \tilde{F}_{ij} - \Big(\sum_{ij} H_{ij} + \sum_{ij} F_{ij}\Big)\Bigg|
\end{aligned}$$

$$
\begin{aligned}
&+\frac{4}{n^4}\Big|\sum_{ij}H_{ij}+\sum_{ij}F_{ij}\Big|\cdot\Big|\sum_{ij}\tilde{H}_{ij}-\sum_{ij}\tilde{F}_{ij}-\Big(\sum_{ij}H_{ij}-\sum_{ij}F_{ij}\Big)\Big|\\
\leq\ &\frac{4}{n^3}\Bigg(\Big|\sum_{j=1}^{n}\tilde{H}_{1j}-\sum_{j=1}^{n}\tilde{F}_{1j}\Big|\cdot\Big|\sum_{j=1}^{n}\tilde{H}_{1j}+\sum_{j=1}^{n}\tilde{F}_{1j}-\Big(\sum_{j=1}^{n}H_{1j}+\sum_{j=1}^{n}F_{1j}\Big)\Big|\\
&+\Big|\sum_{j=1}^{n}\tilde{H}_{1j}+\sum_{j=1}^{n}\tilde{F}_{1j}\Big|\cdot\Big|\sum_{j=1}^{n}\tilde{H}_{1j}-\sum_{j=1}^{n}\tilde{F}_{1j}-\Big(\sum_{j=1}^{n}H_{1j}-\sum_{j=1}^{n}F_{1j}\Big)\Big|\Bigg)\\
&+\frac{4}{n^3}\sum_{i>1}\Bigg(\Big|\tilde{H}_{i1}-\tilde{F}_{i1}\Big|\cdot\Big|\sum_{j=1}^{n}\tilde{H}_{ij}+\sum_{j=1}^{n}\tilde{F}_{ij}-\Big(\sum_{j=1}^{n}H_{ij}+\sum_{j=1}^{n}F_{ij}\Big)\Big|\\
&+\Big|\sum_{j=1}^{n}H_{ij}+\sum_{j=1}^{n}F_{ij}\Big|\cdot\Big|\tilde{H}_{i1}-\tilde{F}_{i1}-\Big(H_{i1}-F_{i1}\Big)\Big|\Bigg)\\
&+\frac{4}{n^4}\cdot 2(2n-1)\cdot 4\nu\cdot(n^2\cdot 4L_2\epsilon\sqrt{d}+n^2\cdot 4L_2\epsilon\sqrt{d})\\
&+\frac{4}{n^4}\cdot(n^2\cdot 4\nu+n^2\cdot 4\nu)\cdot((2n-1)\cdot 4L_2\epsilon\sqrt{d}+(2n-1)\cdot 4L_2\epsilon\sqrt{d})\\
\leq\ &\frac{4}{n^3}\cdot(n\cdot 4\nu+n\cdot 4\nu)\cdot(n\cdot 4L_2\epsilon\sqrt{d}+n\cdot 4L_2\epsilon\sqrt{d})+\frac{4}{n^3}\cdot(n\cdot 4\nu+n\cdot 4\nu)\cdot(n\cdot 4L_2\epsilon\sqrt{d}+n\cdot 4L_2\epsilon\sqrt{d})\\
&+\frac{4}{n^3}\cdot(n-1)\cdot(8\nu\cdot(n\cdot 4L_2\epsilon\sqrt{d}+n\cdot 4L_2\epsilon\sqrt{d})+(n\cdot 4\nu+n\cdot 4\nu)\cdot(4L_2\epsilon\sqrt{d}+4L_2\epsilon\sqrt{d}))\\
&+\frac{512(2n-1)}{n^2}\nu L_2\epsilon\sqrt{d}\\
\leq\ &\frac{1024(2n-1)}{n^2}\nu L_2\epsilon\sqrt{d}\leq\frac{2048\nu L_2\epsilon\sqrt{d}}{n}.
\end{aligned}
$$

Using McDiarmid's inequality for each $\Delta(\theta_i)$ and a union bound, we then obtain that with probability at least $1-\delta$,

$$
\max_{i\in\{1,2,\ldots,P\}}\Delta(\theta)\leq\frac{2048\nu L_2\epsilon\sqrt{d}}{\sqrt{2n}}\sqrt{\log\frac{2P}{\delta}}\leq\frac{1024\nu L_2\epsilon\sqrt{d}}{\sqrt{n}}\sqrt{2\log\frac{2}{\delta}+2\kappa\log\frac{4\mathrm{R}_\Theta}{q}}.
$$

According to Lemma 19 in Liu et al. (2020a), for any two $\theta,\theta'\in\Theta$, we have

$$
|\hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_\mathbb{P},S_\mathbb{Q};k_\theta)-\hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_\mathbb{P},S_\mathbb{Q};k_{\theta'})|\leq 256L_1\nu\|\theta-\theta'\|\leq 256L_1\nu q.
$$

Similarly, $|\hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_\mathbb{P},\tilde{S}_\mathbb{Q};k_\theta)-\hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_\mathbb{P},\tilde{S}_\mathbb{Q};k_{\theta'})|\leq 256L_1\nu q$.

Combining these two results, we know that with probability at least $1-\delta$,

$$
\sup_\theta|\Delta(\theta)|\leq\max_{i\in\{1,2,\ldots,P\}}\Delta(\theta)+512L_1\nu q\leq\frac{1024\nu L_2\epsilon\sqrt{d}}{\sqrt{n}}\sqrt{2\log\frac{2}{\delta}+2\kappa\log\frac{4\mathrm{R}_\Theta}{q}}+512L_1\nu q.
$$

Since the adversary perturbs benign test pairs, we let $n=n_{\mathrm{te}}$ and $q=\frac{1}{\sqrt{n_{\mathrm{te}}}}$, thus yielding the desired results. □

B.2. Proof of Lemma 1

Lemma 1 (Restated). 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 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

$$
\begin{aligned}
&\sup_{\theta\in\bar{\Theta}_s}|\hat{\mathcal{F}}(S_\mathbb{P},\tilde{S}_\mathbb{Q};k_\theta)-\hat{\mathcal{F}}(S_\mathbb{P},S_\mathbb{Q};k_\theta)|\\
&\leq\frac{L_2\epsilon\sqrt{d}}{\sqrt{n_{\mathrm{te}}}}\left[\frac{8}{s}+\frac{2048\nu^2}{s^3}\right]\sqrt{2\log\frac{2}{\delta}+2\kappa\log(4\mathrm{R}_\Theta\sqrt{n_{\mathrm{te}}})}+\left[\frac{8L_1}{s\sqrt{n_{\mathrm{te}}}}+\frac{1024L_1\nu}{s^3\sqrt{n_{\mathrm{te}}}}\right]:=\tilde{\xi},
\end{aligned}
$$

and by treating ν as a constant, we have

$$\sup_{\theta\in\bar{\Theta}_s} |\hat{\mathcal{F}}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) - \hat{\mathcal{F}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)| = \mathcal{O}\bigg(\frac{\epsilon L_2 \sqrt{d\big(\log\frac{1}{\delta} + \kappa\log(\mathrm{R}_\Theta\sqrt{n_{\mathrm{te}}})\big)} + L_1}{s\sqrt{n_{\mathrm{te}}}}\bigg).$$

Proof of Lemma 1. Using Proposition 1 and 2, we have

$$\begin{aligned}
&\sup_{\theta\in\bar{\Theta}_s} |\hat{\mathcal{F}}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) - \hat{\mathcal{F}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)| \\
&= \sup_{\theta\in\bar{\Theta}_s} \Big|\frac{\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta)}{\hat{\sigma}_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta)} - \frac{\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)}{\hat{\sigma}_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)}\Big| \\
&\leq \left|\frac{\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta)}{\hat{\sigma}_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta)} - \frac{\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)}{\hat{\sigma}_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta)}\right| + \left|\frac{\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)}{\hat{\sigma}_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta)} - \frac{\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)}{\hat{\sigma}_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)}\right| \\
&= \frac{1}{|\hat{\sigma}_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta)|} \cdot |\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)| \\
&\quad + \frac{|\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)| \cdot |\hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) - \hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)|}{|\hat{\sigma}_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)| \cdot |\hat{\sigma}_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta)| \cdot |(\hat{\sigma}_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) + \hat{\sigma}_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta))|} \\
&\leq \frac{1}{s}|\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)| + \frac{4\nu}{2s^3}|\hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta) - \hat{\sigma}^2_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)| \\
&= \frac{1}{s}\Big(\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}}}}\Big) + \frac{2\nu}{s^3}\Big(\frac{1024\nu L_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{512L_1\nu}{\sqrt{n_{\mathrm{te}}}}\Big) \\
&= \left[\frac{8L_2\epsilon\sqrt{d}}{s\sqrt{n_{\mathrm{te}}}} + \frac{2048\nu^2L_2\epsilon\sqrt{d}}{s^3\sqrt{n_{\mathrm{te}}}}\right]\sqrt{2\log\frac{2}{\delta} + 2\kappa\log(4\mathrm{R}_\Theta\sqrt{n_{\mathrm{te}}})} + \left[\frac{8L_1}{s\sqrt{n_{\mathrm{te}}}} + \frac{1024L_1\nu}{s^3\sqrt{n_{\mathrm{te}}}}\right].
\end{aligned}$$

□

B.3. Proof of Theorem 2

Before providing the proof of Theorem 2, we need the following lemma. We let $\mathcal{F}(k_\theta)$ refer to $\mathcal{F}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)$, and analogously $\hat{\mathcal{F}}(k_\theta)$ refer to $\hat{\mathcal{F}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)$, for simplicity.

Lemma 2 (Liu et al. (2020a))**.** *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. With probability at least $1-\delta$, we have*

$$\begin{aligned}
&\sup_{\theta\in\bar{\Theta}_s} |\hat{\mathcal{F}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta) - \mathcal{F}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta)| \\
&\leq \frac{2\nu}{s^3}\lambda + \frac{1}{\sqrt{n_{\mathrm{tr}}}}\left[\frac{8\nu}{s} + \frac{1792\nu}{s^2 s}\right]\sqrt{2\log\frac{2}{\delta} + 2\kappa\log(4\mathrm{R}_\Theta\sqrt{n_{\mathrm{tr}}})} + \left[\frac{8}{s\sqrt{n_{\mathrm{tr}}}} + \frac{2048\nu^2}{\sqrt{n_{\mathrm{tr}}}s^2 s}\right]L_1 + \frac{4608\nu^3}{s^2 n_{\mathrm{tr}} s} := \xi.
\end{aligned}$$

Then, we provide the proof of Theorem 2.

Theorem 2 (Restated). 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

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

Proof of Theorem 2. Letting $\theta^* = \arg\max\mathcal{F}(k_\theta)$, we know that $\hat{\mathcal{F}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_{\hat{\theta}_{n_{\mathrm{tr}}}}) \geq \hat{\mathcal{F}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_{\theta^*})$ because $\hat{\theta}_{n_{\mathrm{tr}}}$ maximizes $\hat{\mathcal{F}}$. Using Lemma 1 and 2 , in the adversarial setting, we can obtain

$$\mathcal{F}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\hat{\theta}_{n_{\mathrm{tr}}}}) \geq \hat{\mathcal{F}}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\hat{\theta}_{n_{\mathrm{tr}}}}) - \xi \geq (\hat{\mathcal{F}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_{\hat{\theta}_{n_{\mathrm{tr}}}}) - \tilde{\xi}) - \xi \geq \hat{\mathcal{F}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_{\theta^*}) - \xi - \tilde{\xi}$$

$$\geq (\mathcal{F}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_{\theta^*}) - \xi) - \xi - \tilde{\xi} = \mathcal{F}^* - 2\xi - \tilde{\xi}. \tag{9}$$

Corollary 11 of Gretton et al. (2012) implies that $r^{(n_{\mathrm{te}})} \leq 4\nu\sqrt{\log(\alpha^{-1})n_{\mathrm{te}}}$ no matter the choice of θ. According to Theorem 1 and Eq. (9), with probability at least $1-\delta$, the test in adversarial settings has power

$$\begin{aligned}
&\Pr\left[n_{\mathrm{te}}\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\hat{\theta}_{n_{\mathrm{tr}}}}) > r^{(n_{\mathrm{te}})}\right] \\
&= \Pr\left[n_{\mathrm{tr}}\frac{\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\hat{\theta}_{n_{\mathrm{tr}}}}) - \mathrm{MMD}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\hat{\theta}_{n_{\mathrm{tr}}}})}{\sigma_{\mathcal{H}_1}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\hat{\theta}_{n_{\mathrm{tr}}}})} > \frac{r^{(n_{\mathrm{te}})}}{\sqrt{n_{\mathrm{te}}}\sigma_{\mathcal{H}_1}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\hat{\theta}_{n_{\mathrm{tr}}}})} - \frac{\sqrt{n_{\mathrm{te}}}\mathrm{MMD}^2(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\hat{\theta}_{n_{\mathrm{tr}}}})}{\sigma_{\mathcal{H}_1}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\hat{\theta}_{n_{\mathrm{tr}}}})}\right] \\
&\to \Phi\left[\sqrt{n_{\mathrm{te}}}\mathcal{F}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\hat{\theta}_{n_{\mathrm{te}}}}) - \frac{r^{(n_{\mathrm{te}})}}{\sqrt{n_{\mathrm{te}}}\sigma_{\mathcal{H}_1}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_{\hat{\theta}_{n_{\mathrm{tr}}}})}\right] \\
&\geq \Phi\left[\sqrt{n_{\mathrm{te}}}(\mathcal{F}^* - 2\xi - \tilde{\xi}) - \frac{r^{(n_{\mathrm{te}})}}{s\sqrt{n_{\mathrm{te}}}}\right] \\
&\geq \Phi\left[\sqrt{n_{\mathrm{te}}}\left(\mathcal{F}^* - \frac{C_1}{\sqrt{n_{\mathrm{tr}}}}\sqrt{\log\frac{\sqrt{n_{\mathrm{tr}}}}{\delta}} - \frac{C_2 L_2 \epsilon\sqrt{d}}{\sqrt{n_{\mathrm{te}}}}\sqrt{\log\frac{\sqrt{n_{\mathrm{te}}}}{\delta}}\right) - C_3\sqrt{\log\frac{1}{\alpha}}\right],
\end{aligned}$$

where C_1, C_2, C_3 are constants depending on $\nu, L_1, \kappa, \mathrm{R}_\Theta, \lambda$ and s. □

C. Related Works

In this section, we discuss the differences between our work and the related studies.

Two-sample tests. TST is a premier statistical method to judge whether two sets of data come from the same distribution. Classical TSTs such as *t*-test and Kolmogorov-Smirnov test require strong assumptions on the distributions being studied and are only efficient when applied to one-dimensional data. Non-parametric TSTs, relaxing the distributional assumptions and being able to handling complex distributions, have been applied to a wide of real-world domains (Gretton et al., 2009; Sugiyama et al., 2011; Gretton et al., 2012; Sutherland et al., 2017; Chen & Friedman, 2017; Ghoshdastidar et al., 2017; Li & Wang, 2018; Kirchler et al., 2020; Chwialkowski et al., 2015; Jitkrittum et al., 2016; Lopez-Paz & Oquab, 2016; Cheng & Cloninger, 2019; Liu et al., 2020a; 2021). These tests have also allowed applications in various machine learning problems such as domain adaptation, covariate shift, label-noise learning, generative modeling, fairness and causal discovery (Bińkowski et al., 2018; Zhang et al., 2020b; Fang et al., 2020a; Gong et al., 2016; Fang et al., 2020b; Liu et al., 2019; Zhang et al., 2020c; Liu et al., 2020b; Stojanov et al., 2019; Lopez-Paz & Oquab, 2016; Oneto et al., 2020). However, people rarely doubt the reliability of non-parametric TSTs. In other words, adversarial robustness of non-parametric TSTs is barely studied. In this paper, we leverage our proposed adversarial attack to disclose the failure mode of non-parametric TSTs and propose an effective strategy to make TSTs reliable in analyzing critical data.

Robust hypothesis tests. Previous robust hypothesis tests are composite tests where the null and the alternative hypotheses include a family of distributions, to obtain the reliable estimation of the underlying distributions when there exists outliers in training dataset. These robust tests introduce various uncertainty sets for the distributions under the null and the alternative hypotheses such as ϵ-contamination sets (Huber, 2004) and sets centered around the empirical distribution defined via Kullback-Leibler divergence (Levy, 2008; Gül & Zoubir, 2017) or Wasserstein metric (Gao et al., 2018; Xie et al., 2021). In comparison, our study discloses a premier hypothesis testing method (i.e., non-parametric TSTs) is non-robust against adversarial attacks during the testing procedure. Further, we develop a novel defense—robust deep kernels for TSTs, to enhance adversarial robustness of non-parametric TSTs at the testing time.

Adversarial attacks and defenses. There is a bunch of studies on adversarial attacks (Szegedy et al., 2014; Goodfellow et al., 2015; Moosavi-Dezfooli et al., 2016; Papernot et al., 2016; Carlini & Wagner, 2017; Chen et al., 2018; Ilyas et al., 2018; Athalye et al., 2018; Cheng et al., 2019; Xiao et al., 2018; Zheng et al., 2019; Wong et al., 2019; Mopuri et al., 2018; Alaifari et al., 2019; Sriramanan et al., 2020; Cheng et al., 2020; Chen et al., 2020; Rahmati et al., 2020; Yan et al., 2020; Croce & Hein, 2020; Wu et al., 2020b;a; Andriushchenko et al., 2020; Croce et al., 2020; Yu et al., 2021; Yao et al., 2021;

Hendrycks et al., 2021; Kanth Nakka & Salzmann, 2021) and defenses (Madry et al., 2018; Cai et al., 2018; Yan et al., 2018; Wang et al., 2019; Song et al., 2019; Tramèr et al., 2018; Wong & Kolter, 2018; Shafahi et al., 2019; Pang et al., 2019; Carmon et al., 2019; Wang et al., 2020; Ding et al., 2020; Wu et al., 2020b; Dong et al., 2020; Wong et al., 2020; Sehwag et al., 2020; Zhang et al., 2019a; Qin et al., 2019; Zhang et al., 2019b; 2020a; 2021; Sriramanan et al., 2020; 2021; Robey et al., 2021; Zou et al., 2021; Kim et al., 2021; Wang et al., 2021; Sarkar et al., 2021; Pang et al., 2021; Chen et al., 2021; Erdemir et al., 2021; Gowal et al., 2021; Rebuffi et al., 2021) in the parametric settings, especially focusing on DNNs. On the other hand, studies on robustness of non-parametric classifiers (e.g., nearest neighbors, decision trees, random forests and kernel classifiers) are gaining a growing attention (Amsaleg et al., 2017; Hein & Andriushchenko, 2017; Wang et al., 2018; Chen et al., 2019; Sitawarin & Wagner, 2019; Yang et al., 2019; 2020b; Bhattacharjee & Chaudhuri, 2020; 2021) as well. In contrast, our study focuses on adversarial robustness of non-parametric TSTs, which belongs to the field of hypothesis test rather than classification problems.

Statistical adversarial data detection. Non-parametric TSTs have been applied to judge if upcoming data contains adversarial data that is statistically different from benign data distribution (Metzen et al., 2017; Feinman et al., 2017; Grosse et al., 2017; Gao et al., 2021). These works focus on utilizing statistical methods (e.g., TSTs) to distinguish adversarial data against DNNs from benign data. Compared to these works, our work investigates TST itself. We disclose the adversarial vulnerabilities of non-parametric TSTs through adversarial attacks and further propose effective defensive strategies to make non-parametric TSTs reliable.

D. Non-Parametric Two-Sample Tests

We provide an introduction to the typical non-parametric TSTs in this section.

D.1. Test Statistics

C2ST-S (Lopez-Paz & Oquab, 2016) Classifier-based two-sample test (C2ST) utilizes a classifier $f : \mathcal{X} \to \mathbb{R}$ that outputs the classification probabilities. C2ST trains f via maximizing the classification accuracy, and then makes judgements on the test pairs. C2ST-S is based on the sign of classification probabilities. The test statistic of C2ST-S proposed in Lopez-Paz & Oquab (2016) is

$$\mathcal{D}^{(\mathrm{S})}(S_{\mathbb{P}}, S_{\mathbb{Q}}) = \frac{1}{2n} \sum_{x_i \in S_{\mathbb{P}}} \mathbb{1}(f(x_i) > 0) + \frac{1}{2n} \sum_{y_i \in S_{\mathbb{Q}}} \mathbb{1}(f(y_i) < 0). \tag{10}$$

Further, Liu et al. (2020a) pointed out that $\mathcal{D}^{(\mathrm{S})}(S_{\mathbb{P}}, S_{\mathbb{Q}})$ is equivalent to $\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k^{(\mathrm{S})})$.

C2ST-L (Cheng & Cloninger, 2019) C2ST-L utilizes the classification confidence given by f instead of only accessing the sign of f's output. Letting $f : \mathcal{X} \to \mathbb{R}$ be a classifier that outputs classification probabilities, the test statistic of C2ST-L proposed in Cheng & Cloninger (2019) is

$$\mathcal{D}^{(\mathrm{L})}(S_{\mathbb{P}}, S_{\mathbb{Q}}) = \frac{1}{n} \sum_{x_i \in S_{\mathbb{P}}} f(x_i) - \frac{1}{n} \sum_{y_i \in S_{\mathbb{Q}}} f(y_i). \tag{11}$$

Similar to C2ST-S, Liu et al. (2020a) also pointed out that $\mathcal{D}^{(\mathrm{L})}(S_{\mathbb{P}}, S_{\mathbb{Q}})$ is equivalent to $\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k^{(\mathrm{L})})$.

ME (Chwialkowski et al., 2015; Jitkrittum et al., 2016). Given a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a set of G test locations $\mathcal{V} = \{v_i\}_{i=1}^G$, the test statistic of ME is

$$\mathcal{D}^{(\mathrm{ME})}(S_{\mathbb{P}}, S_{\mathbb{Q}}) = n \bar{z}_n^\top S_n^{-1} \bar{z}_n, \tag{12}$$

where $\bar{z}_n = \frac{1}{n} \sum_{i=1}^n z_i$, $S_n = \frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z}_n)(z_i - \bar{z}_n)^\top$, and $z_i = (k(x_i, v_j) - k(y_i, v_j))_{j=1}^G \in \mathbb{R}^G$.

SCF (Chwialkowski et al., 2015; Jitkrittum et al., 2016). Given a positive definite kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and a set of G test locations $\mathcal{V} = \{v_i\}_{i=1}^G$, the test statistic of SCF is

$$\mathcal{D}^{(\mathrm{SCF})}(S_{\mathbb{P}}, S_{\mathbb{Q}}) = n \bar{z}_n^\top S_n^{-1} \bar{z}_n, \tag{13}$$

where $\bar{z}_n = \frac{1}{n}\sum_{i=1}^{n} z_i$, $S_n = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z}_n)(z_i - \bar{z}_n)^\top$, and $z_i = [\hat{h}(x_i)\sin(x_i^\top v_j) - \hat{h}(y_i)\sin(y_i^\top v_j), \hat{h}(x_i)\cos(x_i^\top v_j) - \hat{h}(y_i)\cos(y_i^\top v_j)]_{j=1}^{G}$. $\hat{h}(x) = \int_{\mathbb{R}^d} \exp(-iux)l(u)du$ is the Fourier transform of $l(x)$, and $h : \mathbb{R}^d \to \mathbb{R}$ is an analytic translation-invariant kernel.

D.2. Test Criterion

For C2ST-S and C2ST-L, Lopez-Paz & Oquab (2016) and Cheng & Cloninger (2019) proposed to maximize f's classification accuracy, but it cannot directly maximize the test power (Liu et al., 2020a). In this paper, therefore, we take $\hat{\mathcal{F}}^{(\mathrm{S})}(S_{\mathbb{P}}, S_{\mathbb{Q}}) = \frac{\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k^{(\mathrm{S})})}{\hat{\sigma}_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k^{(\mathrm{S})})}$ and $\hat{\mathcal{F}}^{(\mathrm{L})}(S_{\mathbb{P}}, S_{\mathbb{Q}}) = \frac{\widehat{\mathrm{MMD}}^2(S_{\mathbb{P}}, S_{\mathbb{Q}}; k^{(\mathrm{L})})}{\hat{\sigma}_{\mathcal{H}_1,\lambda}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k^{(\mathrm{L})})}$ as the test criterion for C2ST-S and C2ST-L, respectively. To make $\hat{\mathcal{F}}^{(\mathrm{S})}$ differentiable, we modify the kernel for C2ST-S as follows:

$$k^{(\mathrm{S})}(x, y) = \frac{1}{16}(\frac{f(x)}{|f(x)|} + 1)(\frac{f(y)}{|f(y)|} + 1). \tag{14}$$

For ME and SCF tests, Chwialkowski et al. (2015) and Jitkrittum et al. (2016) theoretically pointed out that maximizing $\mathcal{D}^{(\mathrm{ME})}(S_{\mathbb{P}}, S_{\mathbb{Q}})$ and $\mathcal{D}^{(\mathrm{SCF})}(S_{\mathbb{P}}, S_{\mathbb{Q}})$ can maximize the test power of ME and SCF, respectively. Therefore, $\hat{\mathcal{F}}^{(\mathrm{ME})}(\cdot,\cdot) = \mathcal{D}^{(\mathrm{ME})}(\cdot,\cdot)$ and $\hat{\mathcal{F}}^{(\mathrm{SCF})}(\cdot,\cdot) = \mathcal{D}^{(\mathrm{SCF})}(\cdot,\cdot)$.

E. Experimental Details and Results

E.1. Datasets

In this section, we introduce the distribution $\mathbb{P}$ and $\mathbb{Q}$ of each dataset.

Blob. Blob is often used to validate two-sample test methods (Gretton et al., 2012; Jitkrittum et al., 2016; Sutherland et al., 2017). We show the specifications of $\mathbb{P}$ and $\mathbb{Q}$ of Blob in Table 6.

High-dimensional Gaussian mixture. High-dimensional Gaussian mixture (HDGM) was utilized as a benchmark dataset in Liu et al. (2020a). HDGM can be regarded as high-dimensional Blob which contains two modes with the same variance and different covariance. We show the specifications of $\mathbb{P}$ and $\mathbb{Q}$ of HDGM in Table 6.

We set $d = 10$ for experiments on HDGM in Section 5.1 and 5.2. In section 5.4, we conduct experiments on HDGM with different $d \in \{5, 10, 15, 20, 25\}$. In practice, the scale of data from HDGM is roughly betwen -4.37 and 4.70. The adversarial budget we set in the experiments ($\epsilon = 0.05$) on HDGM is small enough.

Table 6. Specifications of $\mathbb{P}$ and $\mathbb{Q}$ of synthetic datasets. $\mu_1^b = [0,0], \mu_2^b = [0,1], \mu_3^b = [0,2], ..., \mu_8^b = [2,1], \mu_9^b = [2,2]$. $\mu_1^h = \mathbf{0}_d, \mu_2^h = 0.5 \times \mathbf{1}_d$ where $\mathbf{1}_d$ is an identity matrix with size d. $\Delta_i^b = -0.02 - 0.002 \times (i-1)$ if $i < 5$ and $\Delta_i^b = 0.02 + 0.002 \times (i-6)$ if $i > 5$. if $i = 5$, $\Delta_i^b = 0$. Δ_1^h and Δ_2^h are set to 0.5 and -0.5, respectively.

Datasets	$\mathbb{P}$	$\mathbb{Q}$
Blob	$\sum_{i=1}^{9}\frac{1}{9}\mathcal{N}(\mu_i^b, 0.03 \times I_2)$	$\sum_{i=1}^{9}\frac{1}{9}\mathcal{N}\left(\mu_i^b, \begin{bmatrix} 0.03 & \Delta_i^b \\ \Delta_i^b & 0.03 \end{bmatrix}\right)$
HDGM	$\sum_{i=1}^{2}\frac{1}{2}\mathcal{N}(\mu_i^h, I_d)$	$\sum_{i=1}^{2}\frac{1}{2}\mathcal{N}\left(\mu_i^h, \begin{bmatrix} 1 & \Delta_i^h & \mathbf{0}_{d-2} \\ \Delta_i^h & 1 & \mathbf{0}_{d-2} \\ \mathbf{0}_{d-2}^T & \mathbf{0}_{d-2}^T & I_{d-2} \end{bmatrix}\right)$

Higgs. For the experiments on Higgs, we compare the jet Φ-momenta distribution ($d = 4$) of the background process, $\mathbb{P}$, which lacks Higgs bosons, to the corresponding distribution $\mathbb{Q}$ for the process that produces Higgs bosons, following Chwialkowski et al. (2015). Higgs dataset can be downloaded from UCI Machine Learning Repository. In practice, the scale of data from Higgs is betwen -1.74 and 1.74. The adversarial budget we set in the experiments ($\epsilon = 0.05$) on Higgs is small enough.

MNIST. For the experiments on MNIST, we compare true MNIST images drawn from MNIST dataset (LeCun et al., 1998) (regarded as the distribution $\mathbb{P}$) to fake MNIST images generated from a pretrained deep convolutional generative adversarial network (DCGAN) (Radford et al., 2015) (regarded as the distribution $\mathbb{Q}$). Samples drawn from $\mathbb{Q}$ can be generated by implementing dcgan.py.

CIFAR-10. For the experiments on CIFAR-10, we compare samples drawn from the class "cat" (regarded as the distribution $\mathbb{P}$) to samples drawn from the class "dog" (regarded as the distribution $\mathbb{Q}$) in CIFAR-10 dataset (Krizhevsky, 2009). CIFAR-10 dataset can be downloaded via PyTorch (Paszke et al., 2019).

E.2. Training Settings

We conduct all experiments on Python 3.8 (PyTorch 1.1) with NVIDIA RTX A50000 GPUs. We run MMD-D, MMD-G, C2ST-S, C2ST-L, ME and SCF using the GitHub code provided by Liu et al. (2020a) and implement MMD-RoD by ourselves. Following Lopez-Paz & Oquab (2016), we use a deep neural network f as the classifier in C2ST-S and C2ST-L, and train f by minimizing cross-entropy loss. The neural network structure ϕ in MMD-D and MMD-RoD has the same architecture with feature extractor in f, i.e., $f = g \circ \phi$ where g is composed of two fully-connected layers and outputs the classification probabilities. For MNIST and CIFAR-10, we normalize the raw data into the scale $[-1, 1]$.

For Blob, HDGM and Higgs, ϕ is a five-layer fully-connected neural network. The number of neurons in hidden and output layers of ϕ are set to 50 for Blob, $3 \times d$ for HDGM and 20 for Higgs, where d is the dimensionality of samples. For MNIST and CIFAR-10, ϕ is a convolutional neural network (CNN) that contains four convolutional layers and one fully-connected layer. The structure of the CNN exactly follows Liu et al. (2020a).

We use Adam optimizer (Kingma & Ba, 2015) to optimize (1) parameters of f in C2ST-S and C2ST-L, (2) parameters of ϕ in MMD-D and MMD-RoD and (3) kernel lengthscale in MMD-G. We set drop-out rate to zero when training C2ST-S, C2ST-L, MMD-D and MMD-RoD on all datasets. We set the number of training samples n_{tr} to 100 for Blob, $3,000$ for HDGM, $5,000$ for Higgs, 500 for MNIST and CIFAR-10.

For ME and SCF, we follow (Chwialkowski et al., 2015) and set $J = 10$ for Higgs. For other datasets, we set $J = 5$.

For C2ST-S and C2ST-L, we set batchsize to 128 for Blob, HDGM and Higgs, and 100 for MNIST and CIFAR-10. We set the number of training epochs to $9000 \times n^{te}/$batchsize for Blob, $1,000$ for HDGM and Higgs, $2,000$ for MNIST and CIFAR-10. We set learning rate to 0.001 for Blob, HDGM and Higgs, and 0.0002 for MNIST and CIFAR-10.

For MMD-D, we use full batch (i.e., all samples) to train MMD-D and MMD-RoD for Blob, HDGM and Higgs. We use mini-batch (batchsize is 100) to train MMD-D and MMD-RoD for MNIST and CIFAR-10. We set the number of training epochs to $2,000$ for Blob, HDGM, Higgs and MNIST, and $1,000$ for CIFAR-10. We set learning rate to 0.0005 for Blob and Higgs, 0.00001 for HDGM, 0.001 for MNIST and 0.0002 for CIFAR-10.

For MMD-RoD, we keep ϵ for each dataset same as that in Table 1 and set T to 1 for all datasets. We set learning rate to 0.0005 for MNIST. Other training settings of MMD-RoD keep same as that of MMD-D.

E.3. Testing Procedure

We use permutation test to compute p-values of MMD-D, MMD-G, C2ST-S, C2ST-L and MMD-RoD. We set α to 0.05 and the iteration number of permutation test to 100 for all experiments. In addition, we utilize the wild bootstrap process (Chwialkowski et al., 2014) to resample the value of MMD for MMD-D, MMD-G and MMD-RoD since the adversarial data are probably not IID. The wild bootstrap can ensure that we obtain correct p-values in non-IID/IID scenarios (Chwialkowski et al., 2014).

Wild bootstrap process. Following Leucht & Neumann (2013) and Chwialkowski et al. (2014), we utilize the following wild bootstrap process:

$$W_t = e^{-1/l} W_{t-1} + \sqrt{1 - e^{-2/l}} \tau_t, \tag{15}$$

where $W_0, \tau_0, ..., \tau_t$ are independent standard normal random variables. In all experiments, we set $l = 0.5$.

We summarize the permutation test with wild bootstrap process for non-parametric TSTs based on MMD in Algorithm 3.

Algorithm 3 Testing with k_θ on $S_\mathbb{P}$ and $S_\mathbb{Q}$

```
Input: input pair $(S_\mathbb{P}, S_\mathbb{Q})$, kernel $k_\theta$ parameterized with $\theta$, iteration number of permutation test $n_{\mathrm{perm}}$
Output: $est$, p-value: $\frac{1}{n_{\mathrm{perm}}}\sum_{i=1}^{n_{\mathrm{perm}}}\mathbb{1}(perm_i > est)$
$est \leftarrow \widehat{\mathrm{MMD}}^2(S_\mathbb{P}, S_\mathbb{Q}; k_\theta)$
for $i = 1$ to $n_{\mathrm{perm}}$ do
    Generate $\{W_i^\mathbb{P}\}_{i=1}^{n_{\mathrm{te}}}$ and $\{W_i^\mathbb{Q}\}_{i=1}^{n_{\mathrm{te}}}$ using Eq. (15)
    $\{\tilde{W}_i^\mathbb{P}\}_{i=1}^{n_{\mathrm{te}}} \leftarrow \{W_i^\mathbb{P}\}_{i=1}^{n_{\mathrm{te}}} - \frac{1}{n_{\mathrm{te}}}\sum_{i=1}^{n_{\mathrm{te}}} W_i^\mathbb{P}$
    $\{\tilde{W}_i^\mathbb{Q}\}_{i=1}^{n_{\mathrm{te}}} \leftarrow \{W_i^\mathbb{Q}\}_{i=1}^{n_{\mathrm{te}}} - \frac{1}{n_{\mathrm{te}}}\sum_{i=1}^{n_{\mathrm{te}}} W_i^\mathbb{Q}$
    $perm_i \leftarrow \frac{1}{n_{\mathrm{te}}(n_{\mathrm{te}}-1)}\sum_{i,j} H_{ij}\tilde{W}_i^\mathbb{P}\tilde{W}_j^\mathbb{Q}$
end for
```

E.4. Weight Set Configurations

Observed from the lower right panel of Figure 3, we empirically find that an appropriate weight set is critical to the performance of EA. We finetune the weight set by increasing the weight of the TST that is difficult to be successfully fooled. Table 7 summarizes the manually-finetuned weight of MMD-D, MMD-G, C2ST-S, C2ST-L, ME and SCF for each dataset.

Table 7. The manually-finetuned weight set of EA for each dataset.

Datasets	$\mathbb{W}$
Blob	$\{\frac{5}{29}, \frac{1}{29}, \frac{1}{29}, \frac{20}{29}, \frac{1}{29}, \frac{1}{29}\}$
HDGM	$\{\frac{25}{79}, \frac{1}{79}, \frac{1}{79}, \frac{50}{79}, \frac{1}{79}, \frac{1}{79}\}$
Higgs	$\{\frac{3}{98}, \frac{45}{98}, \frac{4}{98}, \frac{3}{98}, \frac{40}{98}, \frac{3}{98}\}$
MNIST	$\{\frac{1}{109}, \frac{45}{109}, \frac{1}{109}, \frac{1}{109}, \frac{60}{109}, \frac{1}{109}\}$
CIFAR-10	$\{\frac{1}{80}, \frac{50}{80}, \frac{4}{80}, \frac{4}{80}, \frac{20}{80}, \frac{1}{80}\}$

E.5. Type I Errors

The Type I error of a TST measures the probability of rejecting $\mathcal{H}_0$ when $\mathcal{H}_0$ is true. If the Type I error was much higher than α, this TST would always reject the null hypothesis, which invalidates this TST (Chwialkowski et al., 2014). Therefore, a reasonable Type I error of a TST should not be much higher than α.

Type I Errors of six typical non-parametric TSTs. We report the Type I error of typical non-parametric TSTs on each dataset in Table 8. As for the experimental configurations, the only difference from settings in Section 5.1 is that the training pairs and test pairs are composed of samples drawn from the same distribution $\mathbb{P}$. Table 8 shows that these six typical non-parametric TSTs have reasonable Type I errors in benign settings.

Table 8. We report the Type I error of six typical non-parametric TSTs ($\alpha = 0.05$) on five benchmark datasets.

Datasets	n_{te}	MMD-D	MMD-G	C2ST-S	C2ST-L	ME	SCF
Blob	100	0.056±0.000	0.056±0.000	0.049±0.000	0.051±0.000	0.051±0.000	0.042±0.000
HDGM	3000	0.057±0.000	0.048±0.000	0.056±0.000	0.040±0.000	0.050±0.000	0.041±0.000
Higgs	5000	0.058±0.000	0.043±0.000	0.040±0.001	0.045±0.001	0.043±0.000	0.029±0.000
MNIST	500	0.026±0.000	0.009±0.000	0.030±0.000	0.038±0.000	0.026±0.000	0.010±0.000
CIFAR-10	500	0.032±0.000	0.001±0.000	0.000±0.000	0.003±0.000	0.001±0.000	0.000±0.000

Type I error of MMD-RoD. We report the Type I error of MMD-RoD in Table 9. The training pairs and test pairs are composed of samples drawn from the same distribution $\mathbb{P}$. The training settings and testing procedure of MMD-RoD exactly follow Section 5.2. Table 9 shows that the Type I error of MMD-RoD maintains reasonable in benign settings.

Table 9. The Type I error of MMD-RoD.

Blob	HDGM	Higgs	MNIST	CIFAR-10
0.049±0.004	0.056±0.000	0.030±0.001	0.002±0.000	0.000±0.000

E.6. Transferability between Different Types of Non-Parametric TSTs

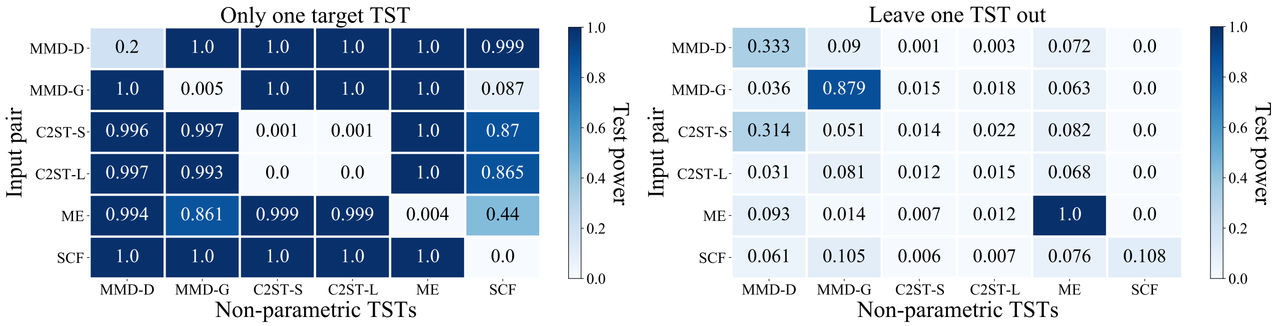
We report the test power of non-parametric TSTs under the adversarial attack against a certain type of TSTs on MNIST in Figure 4(a). The experimental settings are kept the same as in Section 5.1 except $\mathbb{W}$. We set $w^{(\mathcal{J})} = 1$ for the attack implemented on benign test pairs in each row of Figure 4(a) where $\mathcal{J}$ is the target non-parametric TST (corresponding to the ordinate). Figure 4(a) shows that a specific attack against a certain type of TST sometimes can fool other types of TSTs.

Therefore, an ensemble of TSTs is sometimes effective against a specific attack against a certain type of TST. For example, an ensemble of C2ST-S and C2ST-L could still be vulnerable against the attack against C2ST-S since the test power of C2ST-S and C2ST-L are simultaneously degraded under the attack against C2ST-S (see the third row of Figure 4(a)). However, an ensemble of those six typical non-parametric TSTs can defend the attack against C2ST-S since MMD-D, MMD-G and ME all have a high test power under the attack against C2ST-S (see the third row of Figure 4(a)).

However, an ensemble of TSTs is no longer an effective defense under EA. Compared to the attack against a particular type of TST, our proposed EA that jointly minimizes a weighted sum of different test criteria can significantly degrades the test power of different TSTs simultaneously (empirically validated in Section 5.1).

In addition, we further show the test power of non-parametric TSTs under EA against a TST ensemble composed by leaving one TST (corresponding to the ordinate) out of Ensemble on MNIST in Figure 4(b). The experimental settings follow Section 5.1 except $\mathbb{W}$. In each row of Figure 4(b), we set the weight of the TST (corresponding to the ordinate) that is needed to be left out to 0; we then normalize the weights of leftover TSTs in Ensemble to $[0, 1]$ according to the original weight set summarized in Table 7, so that the weight sum is 1. Figure 4(b) demonstrates that attacks against an ensemble of TSTs sometimes can successfully fool TSTs that are not included in the attack ensemble.

All in all, Figure 4 validates that our proposed EA has transferability between different types of non-parametric TSTs, and it further validates that existing non-parametric TSTs lack adversarial robustness.



(a) (b)

Figure 4. In Figure 4(a), each number represents the test power of the non-parametric TST corresponding to its abscissa on adversarial test pairs generated by the attack against the target TST corresponding to its ordinate. In Figure 4(b), each number represents the test power of the non-parametric TST corresponding to the abscissa on adversarial test pairs generated by the attack against Ensemble except the TST corresponding to its ordinate.

E.7. Discussions about the Situation When d is Larger

In this section, we discuss the reason for the phenomenon where the test power of Ensemble under EA does not continue to decrease with larger d (e.g., $d > 15$), which is shown in the upper right of Figure 3. We demonstrate the test power of each particular non-parametric TST and Ensemble under EA with different d in Table 10. Table 10 shows that, with the increasing of d, the test power of most TSTs (e.g., MMD-D, MMD-G) becomes lower. However, the ME test seems to be difficult to be successfully fooled with larger d, especially $d > 15$. We believe that upweighting the test criterion of ME (i.e., enlarging

$w^{(ME)}$) during conducting EA on HDGM with larger d could make EA further hurt the test power of ME and Ensemble.

Table 10. We report the average test power of six typical non-parametric TSTs ($\alpha = 0.05$) as well as Ensemble under EA on HDGM with different $d \in \{5, 10, 15, 20, 25\}$.

d	MMD-D	MMD-G	C2ST-S	C2ST-L	ME	SCF	Ensemble
5	0.289±0.019	0.613±0.029	0.123±0.017	0.597±0.137	0.885±0.080	0.297±0.003	0.983±0.023
10	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
15	0.094±0.002	0.063±0.000	0.079±0.000	0.086±0.000	0.655±0.000	0.003±0.000	0.665±0.093
20	0.008±0.000	0.014±0.000	0.067±0.000	0.051±0.000	0.696±0.000	0.006±0.000	0.765±0.051
25	0.000±0.000	0.000±0.000	0.009±0.000	0.000±0.000	0.762±0.000	0.000±0.000	0.707±0.081

E.8. Extensive Experiments about Adversarially Learning Kernels for TSTs

Here, we study a different adversarial learning objective for obtaining robust kernels that minimizes a weighted sum of benign and adversarial loss (Goodfellow et al., 2015; Zhang et al., 2019b), which is formulated as follows.

$$\hat{\theta} \approx \arg\max_{\theta}(\beta \cdot \hat{\mathcal{F}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta) + (1-\beta) \cdot \hat{\mathcal{F}}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta)), \tag{16}$$

where the adversarial set $\tilde{S}_{\mathbb{Q}}$ is generated using Eq. (7) and $0 \leq \beta \leq 1$ is a constant. Note that Eq. (8) is a special case of Eq. (16) when we set $\beta = 0$.

We call non-parametric TSTs with robust deep kernels obtained by Eq. (16) as "MMD-RoD*". The training algorithm of MMD-RoD* is almost same as Algorithm 2 expect that the Line 6 in Algorithm 2 is replaced with $\theta \leftarrow \theta + \eta\nabla_\theta(\beta \cdot \hat{\mathcal{F}}(S_{\mathbb{P}}, S_{\mathbb{Q}}; k_\theta) + (1-\beta) \cdot \hat{\mathcal{F}}(S_{\mathbb{P}}, \tilde{S}_{\mathbb{Q}}; k_\theta))$.

We conduct experiments to evaluate the adversarial robustness of MMD-RoD*. We set $\beta = 0.5$ and denote the ensemble of six typical non-parametric TSTs and MMD-RoD* as "Ensemble*". Other settings of training, attack and testing procedure exactly follow Section 5.2. We report the test power of MMD-RoD* and Ensemble* in Table 11.

Table 11. Test power of MMD-RoD* and Ensemble*.

	EA	Blob	HDGM	Higgs	MNIST	CIFAR-10
MMD-RoD*	×	1.00±0.04	1.00±0.02	0.52±0.00	**1.00**±0.12	**1.00**±0.00
	√	0.13±0.06	0.01±0.00	0.19±0.02	**0.86**±0.00	**0.84**±0.01
Ensemble*	×	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00	1.00±0.00
	√	0.85±0.01	0.74±0.02	0.54±0.04	**0.89**±0.00	**0.88**±0.00

Compared to MMD-RoD (in Table 2), we find MMD-RoD* that incorporates benign training pairs into adversarially learning kernels improves the test power in benign settings (especially on HDGM), but obtains the lower test power in adversarial settings among all datasets. Therefore, we recommend utilizing only adversarial training pairs for adversarially learning deep kernels.

F. Description of Attackers against Non-Parametric TSTs

In this section, we provide a detailed description of the attacker against non-parametric TSTs from four perspectives "goal, knowledge, capability, strategy" (Biggio & Roli, 2018).

- **Goal.** The attacker aims to make a target non-parametric TST incorrectly judge two sets of data are drawn from the same distribution during the test procedure, when in reality these two sets of data are drawn from different distributions.
- **Knowledge.** Depending on the assumptions made on the attacker's knowledge, we have different attack scenarios.
  - Perfect-knowledge white-box attacks. The attacker is assumed to know everything about the target non-parametric TST, such as the target non-parametric TST's test criterion function and kernel parameters.

  - Limited-knowledge gray-box attacks. The attacker has part of the target non-parametric TST's knowledge. For example, the attacker knows the target non-parametric TST's test criterion function, but does not know its kernel parameters and training data.
  - Zero-knowledge black-box attacks. The attacker does not have any knowledge about the target non-parametric TST. The attacker can only query the non-parametric TST in a black-box manner and then obtain the judgement on the test pairs.

- **Capability.** The attacker can only manipulate test data, and the malicious perturbations should be human-imperceptible.

- **Strategy.** The attacker searches for adversarial sets via minimizing the target non-parametric TST's test criterion under data manipulation constraints.