Journal of Machine Learning Research 26 (2025) 1-72 Submitted 8/24; Published 10/25

On the Robustness of Kernel Goodness-of-Fit Tests

Xing Liu XINGLIU97@OUTLOOK.COM
QuantCo

François-Xavier Briol F.BRIOL@UCL.AC.UK
Department of Statistical Science
University College London

Editor: Bharath Sriperumbudur

Abstract

Goodness-of-fit testing is often criticized for its lack of practical relevance: since "all models are wrong", the null hypothesis that the data conform to our model is ultimately always rejected as the sample size grows. Despite this, probabilistic models are still used extensively, raising the more pertinent question of whether the model is *good enough* for the task at hand. This question can be formalized as a robust goodness-of-fit testing problem by asking whether the data were generated from a distribution that is a mild perturbation of the model. In this paper, we show that existing kernel goodness-of-fit tests are not robust under common notions of robustness including both qualitative and quantitative robustness. We further show that robustification techniques using tilted kernels, while effective in the parameter estimation literature, are not sufficient to ensure both types of robustness in the testing setting. To address this, we propose the first robust kernel goodness-of-fit test, which resolves this open problem by using kernel Stein discrepancy (KSD) balls. This framework encompasses many well-known perturbation models, such as Huber's contamination and density-band models.

Keywords: robustness, hypothesis testing, kernel methods, Stein's method

1. Introduction

Goodness-of-fit (GOF) testing (D'Agostino, 1986; Lehmann and Romano, 2022, Chapter 16) tackles the question of how well a given probabilistic model describes some observed data. More formally, given a model P and observations drawn independently from some distribution Q, GOF testing compares the null hypothesis $H_0 : Q = P$ against the alternative hypothesis $H_1 : Q \neq P$. This process is fundamental to validating a model before it is used for predictions, decision-making, or providing probabilistic guarantees on important quantities of interest. Indeed, GOF testing is closely linked to the concept of statistical significance, which now permeates almost all areas of science and engineering.

In this paper, we focus exclusively on kernel-based GOF tests (Chwialkowski et al., 2016; Liu et al., 2016). Unlike most classical alternatives such as likelihood-ratio tests (Hogg et al., 1977, Chapter 8.7) and Kolmogorov-Smirnov tests (Kolmogorov, 1933), kernel GOF tests can be used for models whose density function is known only up to a multiplicative constant. This is a significant advantage since it enables their use for models which are beyond the reach of other tests, including many modern flexible density estimation models, energy-based models, large graphical models, and Bayesian posteriors. Kernel GOF tests achieve this through a tractable test statistic based on the *score function* (i.e., the gradient of the log-density), which is widely available since it can be evaluated without knowledge of the normalization constant of the density. This property has made kernel GOF tests popular and has led to numerous extensions specializing the approach to problems involving time-to-event data (Fernandez et al., 2020), discrete data (Yang et al., 2018), point processes (Yang et al., 2019), manifold data (Xu and Matsuda, 2020), graphs (Xu and Reinert, 2021), protein structures (Amin et al., 2023), and text documents of variable lengths (Baum et al., 2023).

A significant drawback of GOF testing is the following conundrum: in almost all real-world applications, models are wrong, causing GOF tests to ultimately reject the null hypothesis as the sample size grows. However, there is often a difference between statistical significance from a GOF perspective, and the practical relevance of the model. For instance, *data corruption* may arise in signal processing due to sensor failures (Rizzoni and Min, 1991; Sharma et al., 2010), in classification tasks due to mislabelled entities (Frénay and Verleysen, 2013), and in radio systems due to impulsive noise (Blackard et al., 1993). In these cases, H_0 may be rejected even though P is *almost* correct up to these mild perturbations, and thus it could still be useful for downstream tasks such as prediction or probabilistic inference.

This conundrum has led to the development of *robust tests* (Huber, 1965; LeCam, 1973; Dabak and Johnson, 1994; Fauß et al., 2021), which aim to control the Type-I error rate (the probability of falsely rejecting the null hypothesis) under a *composite* null hypothesis $H_0^{\mathrm{C}} : Q \in \mathcal{P}_0$. The set $\mathcal{P}_0$ is a family of probability distributions, called the uncertainty set (Fauß et al., 2021), which is constructed to include P and distributions similar to P in some appropriate notion. By designing $\mathcal{P}_0$ to contain potential contaminations, such as adversarial contaminations or outliers, the composite null hypothesis H_0^{C} can now hold in realistic scenarios, thus avoiding the aforementioned conundrum and aligning statistical and practical significance.

As a concrete example, consider a scenario where a practitioner models a data set of sensor signals that are potentially contaminated by outliers. Suppose the fitted model accurately captures the true signal in the data set. A standard GOF test might still reject the standard null hypothesis $H_0 : Q = P$ due to the presence of outliers. Instead, one can adopt a composite null hypothesis $H_0^{\mathrm{C}} : Q \in \mathcal{P}_0$, where $\mathcal{P}_0$ is constructed to include "contaminated versions" of P within a certain tolerance. A GOF test designed to be calibrated against H_0^{C} can now reject the model with the correct test level.

Unfortunately, robust kernel-based GOF tests have only received limited attention in the existing literature. The closest work is that of Sun and Zou (2023); Schrab and Kim (2024), who proposed robust tests using uncertainty sets defined by the Maximum Mean Discrepancy (MMD; Müller 1997; Gretton et al. 2012). These tests require samples from *both* Q and P. However, generating samples from P may be infeasible or computationally expensive, and approximating P by finite samples can also reduce statistical efficiency.

We propose a novel class of robust kernel GOF tests based on the *kernel Stein discrepancy* (KSD). These tests are appropriate in the *one-sample* setting, where we assume samples are available only from Q, and an unnormalized density is available for P. To guarantee robustness against mild perturbations, we use *tilted* kernels proposed in Barp et al. (2019); Matsubara et al. (2022), in a related construction for robust parameter estimation. We then study *qualitative* and *quantitative* robustness, two notions of robustness which capture a

Goodness-of-Fit Test	Qualitative robustness	Quantitative Robustness	Result
Existing KSD test + stationary kernels	✗	✗	Theorem 1
Existing KSD test + tilted kernels	✓	✗	Theorem 3
Novel KSD tests proposed in this paper	✓	✓	Theorem 4

Table 1: Summary of theoretical contributions to the robustness of kernel GOF tests provided in this paper.

test's insensitivity to perturbations around P, as formally introduced in Section 2.3. Our contributions are summarized in Table 1 and detailed below:

1. In Section 3, we study the robustness properties of existing KSD tests. We show that KSD tests with stationary kernels are *not necessarily* robust under infinitesimal model deviation (which we call *qualitative robustness*; see Definition 1), and *never* robust against fixed classes of misspecified models (called *quantitative robustness*; see Definition 2). We then show that qualitative robustness can be guaranteed by using appropriately tilted kernels, but this is *not* sufficient to guarantee quantitative robustness.

2. In Section 4, we propose a kernel GOF test that controls the Type-I error rate against all distributions in a KSD-ball of radius $\theta > 0$ centered at P. This test is straightforward to implement since it is identical to existing KSD tests up to a shift of the test statistic by θ. We then make recommendation on how to select θ in practice to ensure robustness to common perturbations such as Huber's contaminations (which includes outliers) or density-band contaminations.

2. Background

We now review the necessary background for this work, covering kernel Stein discrepancy and robust GOF testing. We begin by briefly summarizing our notation.

Let $\mathcal{P}(\mathbb{R}^d)$ denote the set of probability measures on $\mathbb{R}^d$. Given an integer $r > 0$, $\mathcal{C}^r$ is the set of functions $f : \mathbb{R}^d \to \mathbb{R}$ that are r-times continuously differentiable. The gradient of $f \in \mathcal{C}^1$ is denoted $\nabla f(\mathbf{x}) = (\partial_1 f(\mathbf{x}), \ldots, \partial_d f(\mathbf{x}))^\top$. The set $\mathcal{C}_b^1$ contains functions in $\mathcal{C}^1$ that are bounded and with bounded gradient, and the set $\mathcal{C}_b^2$ contains $f \in \mathcal{C}_b^1$ for which $\sup_{\mathbf{x}\in\mathbb{R}^d} |\nabla^\top \nabla f(\mathbf{x})| < \infty$, where $\nabla^\top \nabla f(\mathbf{x}) = \sum_{j=1}^d \partial_j \partial_j f(\mathbf{x})$. For a function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with two arguments, $\nabla_i k(\mathbf{x}, \mathbf{x}')$ will be used to denote its gradient with respect to the i-th argument respectively for $i = 1, 2$, and we define $\nabla_1^\top \nabla_2 k(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^d \partial_{1j} \partial_{2j} k(\mathbf{x}, \mathbf{x}')$, where $\partial_{ij} k(\mathbf{x}, \mathbf{x}') = [\nabla_i k(\mathbf{x}, \mathbf{x}')]_j$ for $i = 1, 2$ and $j = 1, \ldots, d$. The set $\mathcal{C}^{(1,1)}$ (resp., $\mathcal{C}_b^{(1,1)}$) denotes all functions $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with $(\mathbf{x}, \mathbf{x}') \mapsto \partial_{1j} \partial_{2j} k(\mathbf{x}, \mathbf{x}')$ continuous (resp., continuous and bounded) for all $j = 1, \ldots, d$.

2.1 The Kernel Stein Discrepancy

A key ingredient for kernel GOF tests is the kernel Stein discrepancy (KSD, Chwialkowski et al., 2016; Liu et al., 2016; Gorham and Mackey, 2017; Oates et al., 2017). Throughout, we assume $P, Q \in \mathcal{P}(\mathbb{R}^d)$, and P admits a Lebesgue density function $p \in \mathcal{C}^1$ such that $p > 0$ on $\mathbb{R}^d$. Let $k \in \mathcal{C}^{(1,1)}$ be a scalar-valued reproducing kernel with associated reproducing kernel Hilbert space (RKHS, Berlinet and Thomas-Agnan, 2004) denoted by $\mathcal{H}_k$. The *Langevin KSD* gives a notion of discrepancy between Q and P and is defined as

$$D(Q, P) \coloneqq \sup_{f \in \mathcal{B}} \left| \mathbb{E}_{\mathbf{X} \sim Q}[(\mathcal{A}_p f)(\mathbf{X})] \right| = \sup_{f \in \mathcal{B}} \left| \int_{\mathbb{R}^d} (\mathcal{A}_p f)(\mathbf{x}) Q(\mathrm{d}\mathbf{x}) \right| ,$$

where $\mathcal{B} \coloneqq \{h = (h_1, \ldots, h_d) : \ h_j \in \mathcal{H}_k \text{ and } \sum_{j=1}^d \|h_j\|^2_{\mathcal{H}_k} \leq 1\}$ is the unit ball in the d-times Cartesian product $\mathcal{H}_k^d$ of the RKHS $\mathcal{H}_k$. The operator $\mathcal{A}_p$ is a linear operator called the *Langevin Stein operator* and it maps (suitably regular) vector-valued functions f to scalar-valued ones via $(\mathcal{A}_p f)(\mathbf{x}) \coloneqq \mathbf{s}_p(\mathbf{x})^\top f(\mathbf{x}) + \nabla^\top f(\mathbf{x})$, where $\mathbf{s}_p(\mathbf{x}) \coloneqq \nabla_{\mathbf{x}} \log p(\mathbf{x})$ is known as the *(Hyvärinen) score function* of P (Hyvärinen, 2005). Throughout this paper, we will shorten Langevin KSD to KSD for brevity, but note that other Stein operators can also be used to construct KSDs; see Anastasiou et al. (2023) for a review. When $\mathbb{E}_{\mathbf{X} \sim P}[\|\mathbf{s}_p\|_2] < \infty$, the *squared* KSD admits the following double-integral form (Barp et al., 2024, Corollary 1; Gorham and Mackey, 2017)

$$\begin{aligned} D^2(Q, P) \ &= \ \mathbb{E}_{\mathbf{X}, \mathbf{X}' \sim Q}[u_p(\mathbf{X}, \mathbf{X}')] = \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} u_p(\mathbf{x}, \mathbf{x}') Q(\mathrm{d}\mathbf{x}) Q(\mathrm{d}\mathbf{x}') , \\ u_p(\mathbf{x}, \mathbf{x}') \ &\coloneqq \ \mathbf{s}_p(\mathbf{x})^\top \mathbf{s}_p(\mathbf{x}') k(\mathbf{x}, \mathbf{x}') + \mathbf{s}_p(\mathbf{x})^\top \nabla_2 k(\mathbf{x}, \mathbf{x}') \qquad (1a) \\ &\quad + \nabla_1 k(\mathbf{x}, \mathbf{x}')^\top \mathbf{s}_p(\mathbf{x}') + \nabla_1^\top \nabla_2 k(\mathbf{x}, \mathbf{x}') , \qquad (1b) \end{aligned}$$

whenever $D^2(Q, P)$ is well-defined. The function u_p is also a reproducing kernel, known as the *Stein (reproducing) kernel.* The main advantage of the KSD is that it is computable even if p is unnormalized. Indeed, u_p depends on p only through $\mathbf{s}_p$, which does not depend on the normalizing constant of p (as the constant is cancelled due to the differentiation). The KSD can be straightforwardly estimated at $\mathcal{O}(n^2)$ cost using a V-statistic estimator. It is formed by replacing Q with its empirical version Q_n based on independent observations $\mathbb{X}_n \coloneqq \{\mathbf{X}_i\}_{i=1}^n$ drawn from Q:

$$D^2(\mathbb{X}_n) \ \coloneqq \ D^2(Q_n, P) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n u_p(\mathbf{X}_i, \mathbf{X}_j) , \qquad (2)$$

where we have overloaded the notation by writing $D^2 : (\mathbb{R}^d)^n \to [0, \infty)$ as a test statistic. The V-statistic is non-negative, biased but consistent.

An alternative estimator is a U-statistic, which differs from (2) by summing only over *disjoint* index pairs $i \neq j$. Although a U-statistic is unbiased, it can take negative values. Therefore, we focus on V-statistics in this paper, and leave a discussion on how to extend our results to U-statistics to Section C.

When the kernel k is *characteristic*, the KSD is *P-separating*, meaning that $D(Q, P) \geq 0$, with equality if and only if $Q = P$ for all Q that finitely integrates $\|\mathbf{s}_p\|_2$; see, e.g., Barp et al. (2024, Theorem 3). Most characteristic kernels used in kernel GOF tests are sufficiently smooth stationary kernels of the form $k(\mathbf{x}, \mathbf{x}') = h(\mathbf{x} - \mathbf{x}')$, where $h \in \mathcal{C}_b^2$ and $h(0) > 0$. These include the squared-*exponential kernel* $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2 / (2\lambda^2))$ and the *inverse* *multi-quadric* (IMQ) kernel $k(\mathbf{x}, \mathbf{x}') = (1 + \|\mathbf{x} - \mathbf{x}'\|_2^2 / \lambda^2)^{-b}$, where $b > 0$, and $\lambda > 0$ is a hyperparameter known as the *bandwidth.* The IMQ kernel is often preferred in practice since the resulting KSD has the desirable property of *P-convergence control* (Gorham and Mackey, 2017; Barp et al., 2024). The bandwidth λ also plays a crucial role in the performance of kernel-based tests (Reddi et al., 2015; Ramdas et al., 2015; Huang et al., 2023), and a *median-heuristic* (Fukumizu et al., 2009, Section 5) is often used to select its value in practice (Liu et al., 2016; Chwialkowski et al., 2016).

2.2 Kernel Goodness-of-Fit Testing

KSD is a natural choice of test statistic for GOF testing, since, by the P-separation property of KSD, testing $H_0 : Q = P$ is equivalent to testing whether $D(Q, P) = 0$. As KSD is non-negative, A GOF test can therefore be constructed whereby H_0 is rejected for large values of the KSD estimate (Chwialkowski et al., 2016; Liu et al., 2016). KSD tests have been used for GOF testing with unnormalized models in a wide range of applications and data structures (Yang et al., 2018; Fernandez et al., 2020; Xu and Reinert, 2021; Xu and Matsuda, 2021; Amin et al., 2023). Extensions have also been developed to address its limitations, such as high computational cost (Jitkrittum et al., 2017; Huggins and Mackey, 2018), difficulty in bandwidth selection (Schrab et al., 2022) and lack of test power against certain alternatives (Liu et al., 2023).

A significant practical challenge with KSD tests is determining an appropriate decision threshold. A valid threshold can be obtained by considering the quantiles of the null distribution of $D^2(\mathbb{X}_n)$, but since this distribution is not usually tractable, bootstrapping is often employed to estimate it. One approach is to compute the empirical quantiles of bootstrap samples of the form

$$D^2_{\mathbf{W}}(\mathbb{X}_n) \;=\; \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (W_i - 1)(W_j - 1) u_p(\mathbf{X}_i, \mathbf{X}_j) \; , \tag{3}$$

where $\mathbf{W} \coloneqq (W_1, \ldots, W_n) \sim \text{Multinomial}(n; 1/n, \ldots, 1/n)$. This procedure, called *Efron's bootstrap* or *weighted bootstrap* (Arcones and Gine, 1992; Janssen, 1994), assigns a multinomial weight to each observation, which mimics recomputing KSD using data sampled with replacement from $\mathbb{X}_n$. Other bootstrap methods such as wild bootstraps (Leucht and Neumann, 2013; Shao, 2010) are also viable, particularly when the samples are potentially correlated (Chwialkowski et al., 2016), but we focus on the weighted bootstrap since we find that they perform similarly in our setting with independent samples; see Section B.8.

The test threshold is selected as the $(1 - \alpha)$-quantile of the distribution of $D^2_{\mathbf{W}}(\mathbb{X}_n)$ conditional on $\mathbb{X}_n$, i.e.,

$$q^2_{\infty, 1-\alpha}(\mathbb{X}_n) \;\coloneqq\; \inf \left\{ u \in \mathbb{R} : \; 1 - \alpha \leq \Pr_{\mathbf{W}} \left(D^2_{\mathbf{W}}(\mathbb{X}_n) \leq u \mid \mathbb{X}_n \right) \right\} . \tag{4}$$

In practice, this quantile is approximated with Monte-Carlo estimation by first drawing B independent copies $\{\mathbf{W}^b\}_{b=1}^{B}$ of $\mathbf{W}$, and then computing

$$q^2_{B, 1-\alpha}(\mathbb{X}_n) \;\coloneqq\; \inf \left\{ u \in \mathbb{R} : \; 1 - \alpha \leq \frac{1}{B+1} \left(\mathbb{1}\{D^2(\mathbb{X}_n) \leq u\} + \sum_{b=1}^{B} \mathbb{1}\{D^2_{\mathbf{W}^b}(\mathbb{X}_n) \leq u\} \right) \right\} , \tag{5}$$

where $\mathbb{1}\{\mathcal{A}\}$ denotes the indicator function for the event $\mathcal{A}$. The KSD test then rejects H_0 if $D^2(\mathbb{X}_n) > q^2_{B,1-\alpha}(\mathbb{X}_n)$. This test is asymptotically well-calibrated for $H_0 : Q = P$ (Liu et al., 2016, Theorem 4.3). When KSD is P-separating, this test is also *consistent*, meaning that, whenever $Q \neq P$, the probability of rejection approaches one as the sample size $n \to \infty$ (Chwialkowski et al., 2016; Liu et al., 2016).

2.3 Robustness for Goodness-of-Fit Testing

Robustness of GOF tests refers to the lack of sensitivity of the test outcome to small model deviations (Rieder, 1982; Lambert, 1982). Model deviations can be formalized as a neighborhood $\mathcal{P}_0$ around a nominal distribution P. The neighborhood $\mathcal{P}_0$ is often called an *uncertainty set*, and encodes the practitioner's uncertainty on P. One popular example for $\mathcal{P}_0$ is *Huber's contamination model* (Huber, 1964, 1965)

$$\mathcal{P}(P; \epsilon) \;:=\; \{(1-\epsilon')P + \epsilon' R :\; 0 \leq \epsilon' \leq \epsilon,\; R \in \mathcal{P}(\mathbb{R}^d)\}\;, \tag{6}$$

where $\epsilon \in [0,1]$ is the maximal contamination ratio, and the probability measure R acts as arbitrary contamination. This model is appropriate when practitioners believe that P accurately describes all but a small proportion of the data.

Huber's models have been studied in the context of robust GOF testing including Huber (1965); Qin and Priebe (2017), as well as in robust estimation (Hampel, 1974; Huber and Ronchetti, 2011). Special interests lie in the case when R is restricted to point masses, i.e., $R = \delta_{\mathbf{z}}$ for some $\mathbf{z} \in \mathbb{R}^d$, where robustness is often called *bias-robustness* (Huber and Ronchetti, 2011). Beyond Huber's models, $\mathcal{P}_0$ can also be chosen as *density-band models* (Kassam, 1981), which assume the density of the data-generating distribution is close to the model density up to a small error (see Section 4.2). Moreover, $\mathcal{P}_0$ can be set to a ball defined via a statistical divergence or metric, such as Hellinger distance (LeCam, 1973), Wasserstein metric (Gao et al., 2018), and Maximum Mean Discrepancy (Sun and Zou, 2023).

Uncertainty sets $\mathcal{P}_0$ provide a framework for assessing the robustness of a GOF test, by studying the rejection probability when Q deviates from the nominal distribution P but remains within $\mathcal{P}_0$. In the robust testing literature, a common approach is to consider a sequence $\{\mathcal{P}_0^n\}_{n=1}^\infty$ of uncertainty sets with *shrinking* size as the sample size n increases. This is formalized in the following notion of *qualitative robustness*, inspired by Rieder (1982).

Definition 1 (Qualitative robustness to a sequence of neighborhood)**.** *Let $\mathcal{P}_0^1 \supseteq \mathcal{P}_0^2 \supseteq \ldots$ be a sequence of subsets of $\mathcal{P}(\mathbb{R}^d)$ that contain P and such that $\cap_{n=1}^\infty \mathcal{P}_0^n = \{P\}$. For any positive integer n, let $T_n : \left(\mathbb{R}^d\right)^n \to \mathbb{R}$ be a test statistic where a large value of T_n suggests deviation from the null $H_0 : Q = P$, and let $\gamma_n : \left(\mathbb{R}^d\right)^n \to \mathbb{R}$ be a function that computes the decision threshold. A sequence of hypothesis tests (indexed by n) that reject H_0 when $T_n > \gamma_n$ is* qualitatively robust *to $\{\mathcal{P}_0^n\}_{n=1}^\infty$ if, as $n \to \infty$,*

$$\sup\nolimits_{Q \in \mathcal{P}_0^n} \left| \Pr\nolimits_{\mathbb{X}_n \sim Q}\left(T_n(\mathbb{X}_n) > \gamma_n(\mathbb{X}_n)\right) - \Pr\nolimits_{\mathbb{X}_n^* \sim P}\left(T_n(\mathbb{X}_n^*) > \gamma_n(\mathbb{X}_n^*)\right)\right| \;\to\; 0\;. \tag{7}$$

Intuitively, this notion of robustness asks how sensitive the test outcome is under infinitesimally small model deviation. The shrinking-size condition ensures that the rejection probability under any $Q \neq P$ does not trivially approach one due to the consistency of the test. By choosing the $\mathcal{P}_0^n$ in Definition 1 to be Prokhorov balls (Prokhorov, 1956), we would recover the qualitative robustness introduced in Rieder (1982, Definition 2.1), which parallels the conventional notion of qualitative robustness for estimators (Rieder, 1982, Remark 2). However, distributions in Prokhorov balls do not have a simple form, posing challenges to the analysis. Our definition extends it to a general sequence of neighborhood. In Section 3, we will choose $\mathcal{P}_0^n$ to be Huber's contamination models (6), which both significantly simplifies the analysis and encompasses a wide range of relevant scenarios. We will show that, within Huber's neighborhood, the standard KSD test is *not necessarily* qualitatively robust with stationary kernels, but is qualitatively robust with appropriately *tilted* kernels.

Moreover, our definition of qualitative robustness is tied to a *sequence* of shrinking neighborhood $\{P_0^n\}_{n=1}^{\infty}$. Clearly, if a test is qualitatively robust to a sequence $\{P_0^n\}_{n=1}^{\infty}$, then it is also qualitatively robust to any sequences that decay faster. Therefore, loosely speaking, the rate of decay of $\{P_0^n\}_{n=1}^{\infty}$ characterizes the *degree* of qualitative robustness. In Section 3, we will explicitly derive the rate required for the standard KSD test to retain qualitative robustness.

However, qualitative robustness has its own limitations. It only concerns the insensitivity of a test to *sufficiently small* model deviations, but offers no guarantees for deviations of a *fixed* size. The latter scenario is more practically pertinent, as practitioners often need to account for a specific form of model misspecification and require the test to remain well-calibrated under that level of uncertainty. This can be formalized by relaxing the point null hypothesis H_0 to a composite hypothesis $H_0^{\mathrm{C}} : Q \in \mathcal{P}_0$, and requiring calibration under H_0^{C}. This motivates the following notion of *quantitative robustness.*

Definition 2 (Quantitative robustness to a single neighborhood)**.** *Given $\alpha \in (0,1)$ and $\mathcal{P}_0 \subset \mathcal{P}(\mathbb{R}^d)$, a test is* quantitatively robust to $\mathcal{P}_0$ *at level α if its rejection probability under* any $Q \in \mathcal{P}_0$ *does not exceeds α.*

Quantitatively robust GOF tests have been developed for various types of uncertainty sets $\mathcal{P}_0$, including Huber's contamination model (Huber, 1965) and neighborhoods defined by Kullback-Leibler divergence (Levy, 2008; Yang and Chen, 2018) or α-divergence (Gül and Zoubir, 2016). These tests enjoy minimax optimality but require the normalizing constant of P to be known, thus not applicable to unnormalized models. Two-sample tests that are quantitatively robust to Hellinger distance (LeCam, 1973), Wasserstein distance (Gao et al., 2018), and Maximum Mean Discrepancy (Sun and Zou, 2023; Gao et al., 2021) have also been proposed. However, to be used for GOF testing, they require approximation of P by finite samples, which incurs extra approximation error and can be computationally demanding due to the non-trivial task of sampling from P.

In Section 4, we will choose $\mathcal{P}_0$ to be a *KSD ball* centered at P. We choose KSD balls over other types of neighborhood because *(i)* this choice naturally lends itself to a GOF test that is both easy to implement and guarantees robustness at little extra computational cost, and *(ii)* it is easy to select the radius of such balls so that our proposed test is quantitatively robust to various types of contamination of interest, such as Huber's contaminations or density-band contaminations.

We conclude this section with a brief comparison between qualitative and quantitative robustness. Qualitative robustness concerns the limiting behavior of a test against some specific class of local alternatives. Thus, it is closely tied to the minimax separation boundary of a test (Ingster, 1987, 1993). Moreover, it offers a notion of "degree of robustness" of a given test via the decay rate of the uncertainty neighbouhood sequence $\{\mathcal{P}_0^n\}_{n=1}^\infty$. In contrast, quantitative robustness cannot be used for this purpose, because any consistent test is not quantitatively robust by definition. However, quantitative robustness has more *practical* relevance, as it concerns a *fixed* composite null set, which can be explicitly constructed to enforce robustness. This is why we have introduced two different notions of robustness.

3. The (Lack of) Robustness of Existing Kernel Goodness-of-fit Tests

We will now study the robustness of standard KSD GOF tests using stationary kernels in Section 3.1, and then their tilted counterparts which are popular in the parameter estimation literature in Section 3.2.

3.1 Existing KSD Tests with Stationary Kernels are not Qualitatively Robust

Our first result states that contamination of Huber's type can considerably affect the probability of the standard KSD test rejecting the null hypothesis $H_0 : Q = P$. Our result holds with the bootstrap threshold defined in (4), and all probabilities are taken over the randomness of both the sample and the bootstrap weights $\mathbf{W} \sim \text{Multinomial}(n; 1/n, \dots, 1/n)$. The proof is in Section A.1.

Theorem 1. *Assume* $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2^4] < \infty$*, the function* $\mathbf{x} \mapsto \|\mathbf{s}_p(\mathbf{x})\|_2$ *is unbounded,* $k(\mathbf{x}, \mathbf{x}') = h(\mathbf{x} - \mathbf{x}')$ *with* $h \in \mathcal{C}_b^2$ *and* $h(0) > 0$*, and assume the integrability conditions*

$$\sup_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{s}_p(\mathbf{z})\|_2^4 \mathbb{E}_{\mathbf{X}\sim P}\big[h(\mathbf{X}-\mathbf{z})^4\big] \ < \ \infty \quad \textit{and} \quad \sup_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{s}_p(\mathbf{z})\|_2^2 \mathbb{E}_{\mathbf{X}\sim P}\big[\|\nabla h(\mathbf{X}-\mathbf{z})\|_2^2\big] \ < \ \infty \ .$$

Then, for any test level $\alpha \in (0, 1)$ *and any sequence* $\{\epsilon_n\}_{n=1}^\infty$ *with* $\epsilon_n = o(n)^{-1}$*, the following holds as* $n \to \infty$*,*

$$\sup_{Q\in\mathcal{P}(P;\epsilon_n)} \left| \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}\left(D^2(\mathbb{X}_n) > q^2_{\infty,1-\alpha}(\mathbb{X}_n)\right) - \Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}}\left(D^2(\mathbb{X}_n^*) > q^2_{\infty,1-\alpha}(\mathbb{X}_n^*)\right)\right|$$
$$\to \ 1 - \alpha \ ,$$

where $\mathcal{P}(P; \epsilon_n)$ *is the Huber's contamination model defined in* (6).

Theorem 1 immediately implies that the standard KSD test is *not* qualitatively robust to any sequence of Huber's models that satisfies the rate condition $\epsilon_n = o(n)^{-1}$. At first glance, the lack of qualitative robustness might seem trivial given that the KSD test is consistent, meaning that the rejection probability under any $Q \neq P$ approaches one as $n \to \infty$. However, one subtlety is that the set $\mathcal{P}(P; \epsilon_n)$ can also shrink in size, since ϵ_n is allowed to decay. The condition $\epsilon_n = o(n)^{-1}$ imposes a lower bound on this decay rate. Intuitively, this rate condition requires $\epsilon_n n$, the expected number of contamination in the sample, to grow with n. In particular, this excludes the case where $\epsilon_n n$ is a constant. Moreover, since Huber's models $\mathcal{P}(P; \epsilon_n)$ are contained within Prokhorov balls, Theorem 1 immediately implies non-qualitative robustness to Prokhorov neighborhoods.

Remark 1 (Unbounded Stein kernel)**.** *Theorem 1 is a consequence of the fact that the Stein kernel evaluated at* $\mathbf{x}$*, i.e.,* $\mathbf{x} \mapsto u_p(\mathbf{x}, \mathbf{x})$*, is* unbounded *when* P *has an exploding* *score function. Exploding score functions are often associated with light tails. Examples of such distributions include those with a density of the form $p(\mathbf{x}) \propto \exp(-\|\mathbf{x}\|_2^r)$ with $r \geq 2$. These distributions have sub-Gaussian tails, and their score function has the form $\mathbf{s}_p(\mathbf{x}) = -r\mathbf{x}\|\mathbf{x}\|_2^{r-2}$, which is unbounded for $r \geq 2$. On the other hand, this exploding-score condition excludes heavy-tailed models with bounded score functions, such as super-Laplacian or t-distributions (Gorham and Mackey, 2017; Barp et al., 2024).*

Remark 2 (Moment conditions)**.** *The moment conditions in Theorem 1 are mild. For the example from Remark 1, namely $p(\mathbf{x}) \propto \exp(-\|\mathbf{x}\|_2^r)$ with $r \geq 2$, direct computation shows that these conditions hold if $h(\mathbf{x})$ and $\nabla h(\mathbf{x})$ decay at least as fast as $\|\|\mathbf{x}\|_2^{-(r-1)}$ so as to cancel the growth of the score $\|\mathbf{s}_p(\mathbf{x})\|_2 = r\|\mathbf{x}\|_2^{r-1}$. Such kernels include squared-exponential kernels, IMQ kernels $h(u) = (1 + \|u\|_2^2)^{-s/2}$ with $s \geq r - 1$, and Matérn kernels with sufficient smoothness. Section 5.1 will provide numerical evidence using a Gaussian model.*

Remark 3 (Connection to separation boundaries)**.** *Let $\mathcal{Q}_\delta \coloneqq \{Q \in \mathcal{P}(\mathbb{R}^d) : S(Q, P) \geq \delta\}$, where S is some statistical divergence or metric. The* separation boundary *(Ingster, 1987, 1993) of a test is the fastest decaying sequence $\{\delta_n\}_n$ such that the test power under any $Q \in \mathcal{Q}_{\delta_n}$ still converges to 1 as $n \to \infty$. From this perspective, Theorem 1 implies that the separation boundary of the standard KSD test must decay at least with rate $\epsilon_n = o(n)^{-1}$, whenever Q_{δ_n} contains Huber-contamination models. This complements existing results on the separation boundary of KSD tests. The most relevant works are Schrab et al. (2022); Hagrass et al. (2025), where they consider alternative distributions Q that have a density with respect to either the Lebesgue measure or the target model P. In particular, their results cannot be used to derive Theorem 1, since their density assumption on Q is violated in our case. Indeed, we consider alternatives of the form $Q = (1 - \epsilon_n)P + \epsilon_n\delta_{\mathbf{z}}$, which involve a Dirac delta measure and thus has no density with respect to the Lebesgue measure nor any continuous measures, including P.*

3.2 Existing KSD Tests with Tilted Kernels are Qualitatively Robust

Since unbounded Stein kernels are the main cause of the lack of robustness of existing KSD tests, a natural approach to enforce robustness is to choose a suitable k so that the Stein kernel u_p becomes bounded. We now show that this can be achieved using *tilted* kernels (Barp et al., 2019; Matsubara et al., 2022). Specifically, we will show in Theorem 3 that, with a tilted kernel, a small proportion of contamination in the data has negligible impact on the outcome of the standard KSD test.

We first prove that tilted kernels give rise to bounded Stein kernels. The proof is deferred to Section A.2.

Lemma 2 (Bounded Stein kernel)**.** *Suppose $k(\mathbf{x}, \mathbf{x}') = w(\mathbf{x})h(\mathbf{x} - \mathbf{x}')w(\mathbf{x}')$, where $h \in \mathcal{C}_b^2$ is a stationary reproducing kernel, and the weighting function $w \in \mathcal{C}_b^1$ satisfies the condition $\sup_{\mathbf{x} \in \mathbb{R}^d} \|w(\mathbf{x})\mathbf{s}_p(\mathbf{x})\|_2 < \infty$. Then*

$$\sup_{\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d} |u_p(\mathbf{x}, \mathbf{x}')| \;\leq\; \sup_{\mathbf{x} \in \mathbb{R}^d} u_p(\mathbf{x}, \mathbf{x}) \;=\; \tau_\infty \;<\; \infty.$$

Notably, the function $\mathbf{x} \mapsto u_p(\mathbf{x}, \mathbf{x})$ is non-negative since u_p is a reproducing kernel. Barp et al. (2024, Theorem 9) showed that KSD with such tilted kernels still satisfies P-separation, namely $D(Q, P) = 0 \iff Q = P$, provided that $\|\mathbf{s}_p\|_2$ grows at most root-exponentially and the spectral density of the translation-invariant kernel k—which exists by Bochner's Theorem (Berlinet and Thomas-Agnan, 2004, Theorem 20)—is bounded away from zero on compact sets.

Most stationary kernels used in KSD tests (such as the IMQ or squared-exponential kernels) satisfy the conditions in Lemma 2. Intuitively, the weighting function w is used to counteract the growth of the score function $\mathbf{s}_p$. When the score grows polynomially like $\|\mathbf{s}_p(\mathbf{x})\|_2 = \mathcal{O}(\|\mathbf{x}\|_2^r)$ for some $r > 0$, it suffices to choose $w(\mathbf{x}) = (1 + a^2\|\mathbf{x}\|_2^2)^{-b}$ for any $a > 0$ and $b \geq 1/(2r)$. This weight function is common in the frequentist parameter estimation literature (Barp et al., 2019) and the generalized Bayesian inference literature (Matsubara et al., 2022; Altamirano et al., 2023, 2024; Duran-Martin et al., 2024).

The following result states that, with a tilted kernel, the outcome of the standard KSD test will not be significantly affected by small proportions of contamination in the data. The proof is provided in Section A.3.

Theorem 3. *Assume $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$ and that k is a tilted kernel satisfying the conditions in Lemma 2. Then, for any test level $\alpha \in (0, 1)$ and any sequence $\epsilon_n = o(n^{-1/2})$, the following holds as $n \to \infty$,*

$$\sup_{Q\in\mathcal{P}(P;\epsilon_n)} \left| \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}\left(D^2(\mathbb{X}_n) > q^2_{\infty,1-\alpha}(\mathbb{X}_n)\right) - \Pr_{\mathbb{X}_n^*\sim P,\mathbf{W}}\left(D^2(\mathbb{X}_n^*) > q^2_{\infty,1-\alpha}(\mathbb{X}_n^*)\right)\right| \to 0 \ ,$$

where $\mathcal{P}(P; \epsilon_n)$ is the Huber's contamination model defined in (6)*.*

The LHS of the above convergence is the worst-case difference between the rejection probabilities of the standard KSD test with and without data corruptions. This result thus immediately implies that the standard KSD test is qualitative robust to any sequence of Huber's models so long as the contamination ratio decays sufficiently fast as $\epsilon_n = o(n^{-1/2})$. The condition on ϵ_n allows the expected number of contaminated data $n\epsilon_n$ to grow with n but requires the growth rate to be slow. In particular, it is met when the expected number of contaminated data is bounded, i.e., $\epsilon_n = \mathcal{O}(n^{-1})$. This condition is not an artifact of the proof: in Section B.1, we show empirically that this rate is tight, i.e., the tilted-KSD test is no longer qualitatively robust when $\epsilon_n = n^{-r}$ for any $r \leq 1/2$.

Remark 4. *The intuition behind Theorem 3 is that, when the conditions in Lemma 2 are met so that the Stein kernel u_p is bounded, the impact of any single outlier on the test statistic $D^2(\mathbb{X}_n)$ can be bounded. In contrast, under the setting of Theorem 1, where k is translation-invariant and thus u_p is unbounded, even a single outlier can drive $D^2(\mathbb{X}_n)$ to infinity. This is the key motivation for using a tilted kernel to guarantee robustness: by choosing a suitable weighting function w, the Stein kernel can be made bounded, thereby mitigating the effect of outliers.*

Remark 5. *The boundedness condition on $\|w(\mathbf{x})\mathbf{s}_p(\mathbf{x})\|_2$ required of Lemma 2 and Theorem 3 bears similarity with the ones imposed for KSD-based robust estimation methods (Barp et al., 2024; Matsubara et al., 2022). For example, Barp et al. (2019, Proposition 7) showed that their KSD-based estimator is* globally bias-robust *assuming $\mathbf{x} \mapsto \|\mathbf{s}_p(\mathbf{x})\|_2 \int_{\mathbb{R}^d} \|k(\mathbf{x}, \mathbf{x}')\mathbf{s}_p(\mathbf{x}')\|_2 Q(\mathrm{d}\mathbf{x}')$ is bounded. This is slightly weaker than those we assume in Lemma 2 and Theorem 3, but also harder to verify in practice.*

Algorithm 1 Robust-KSD (R-KSD) test for goodness-of-fit evaluation.

```
Input: Data $\mathbb{X}_n = \{\mathbf{x}_i\}_{i=1}^n$; target distribution $P$; uncertainty radius $\theta$; bootstrap sample size $B$; test level $\alpha$.
Compute test statistic $\Delta_\theta(\mathbb{X}_n)$ as defined in (9).
for $b = 1, \ldots, B$ do
    Draw $\mathbf{W}^b \sim$ Multinomial$(n; 1/n, \ldots, 1/n)$ and compute bootstrapped sample $D^2_{\mathbf{W}^b}(\mathbb{X}_n)$ by (3).
end for
Compute the (non-squared) bootstrapped quantile $q_{B,1-\alpha}(\mathbb{X}_n)$ by (5).
Reject $H_0^{\mathrm{C}} : Q \in \mathcal{B}^{\mathrm{KSD}}(P;\theta)$ if $\Delta_\theta(\mathbb{X}_n) > q_{B,1-\alpha}(\mathbb{X}_n)$.
```

Remark 6 (Connection to separation boundaries, continued)**.** *Theorem 3 implies that the separation boundary of the KSD tests cannot decay faster than $n^{-1/2}$ when considering alternatives that include Huber-contamination models (see also Remark 3). This can be compared to existing results on the separation boundaries of tests based on* Maximum Mean Discrepancy *(MMD, Müller, 1997; Gretton et al., 2012). MMD is a broader family of discrepancies that recovers KSD when the kernel is chosen to be the Stein kernel u_p (Barp et al., 2024, Theorem 1). The separation rates of MMD tests have been studied extensively (Balasubramanian et al., 2021; Hagrass et al., 2024b), but these works assume the alternative Q has a density with respect to either the Lebesgue measure or P, which does not hold in our contamination setting. An exception is Hagrass et al. (2024a), which does not make this assumption. However, their results are limited to the two-sample setting, where the target distribution P is unknown and approximated by finite samples, again differing from our setup.*

Although tilted kernels enforce *qualitative* robustness, they are *not* enough to guarantee *quantitative* robustness. Indeed, so long as k is characteristic, the consistency of the standard KSD test implies that $\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(D^2(\mathbb{X}_n) > q^2_{\infty,1-\alpha}(\mathbb{X}_n)) \to 1$ whenever $Q \neq P$ (Liu et al. 2016, Proposition 4.2; Schrab et al. 2022, Theorem 3.3), so KSD tests with either stationary or tilted kernels are *not* quantitatively robust to any neighborhoods *strictly* larger than the singleton set $\{P\}$. We provide numerical evidence in Section 5.1.

4. Robust Kernel Goodness-of-fit Tests for KSD-Ball Uncertainty Sets

We now propose a novel robust KSD test for the setting where the uncertainty set is a KSD-ball. Given $P \in \mathcal{P}(\mathbb{R}^d)$ with density $p \in \mathcal{C}^1$, we consider the composite hypotheses

$$H_0^{\mathrm{C}} : \; Q \in \mathcal{B}^{\mathrm{KSD}}(P;\theta) \;, \qquad H_1^{\mathrm{C}} : \; Q \notin \mathcal{B}^{\mathrm{KSD}}(P;\theta) \;, \tag{8}$$

where $\theta \geq 0$ and $\mathcal{B}^{\mathrm{KSD}}(P;\theta) \coloneqq \{Q : \; D(Q,P) \leq \theta\}$. The advantage of using a KSD-ball as the uncertainty set is that it lends naturally to a simple, tractable test that is robust to the contamination models we reviewed in Section 2.3. We first describe our novel test, which we call *robust-KSD* test, and show that it is quantitatively robust to KSD balls in the sense of Definition 2. We then discuss in Section 4.2 how to choose the uncertainty radius θ to incorporate common contamination models, such as Huber's contamination models and density-band models.

4.1 A Robust KSD Test

Given a prescribed test level $\alpha \in (0, 1)$, our robust KSD test uses the following test statistic

$$\Delta_\theta(\mathbb{X}_n) \;\coloneqq\; \max\big(0, D(\mathbb{X}_n) - \theta\big) \;, \tag{9}$$

where $D(\mathbb{X}_n)$ is the square-root of the V-statistic (2). The test rejects H_0^{C} for large values of $\Delta_\theta(\mathbb{X}_n)$. The decision threshold should be chosen to control the Type-I error rate for all possible $Q \in \mathcal{B}^{\mathrm{KSD}}(P;\theta)$, i.e., we require $\gamma = \gamma(\mathbb{X}_n)$ so that $\sup_{Q\in\mathcal{B}^{\mathrm{KSD}}(P;\theta)} \mathrm{Pr}_{\mathbb{X}_n\sim Q,\mathbf{W}}(\Delta_n(\mathbb{X}_n) > \gamma) \leq \alpha$. One approach to construct γ is to use deviation inequalities for bounded functions as done in Gretton et al. (2012, Corollary 11). However, such approach tends to be overly conservative (Gretton et al., 2012). Instead, we propose a bootstrap procedure to construct a decision threshold. We will show that choosing $\gamma = q_{\infty,1-\alpha}(\mathbb{X}_n)$, the square-root of the (population) bootstrap quantile defined in (4), gives the desired Type-I error control asymptotically. Our robust-KSD test therefore rejects $H_0^{\mathrm{C}} : Q \in \mathcal{B}^{\mathrm{KSD}}(P;\theta)$ if $\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n)$. This is summarized in Algorithm 1.

The validity and consistency of our robust KSD test is stated in the following result, proved in Section A.5.2. In particular, it implies that our robust-KSD test is quantitatively robust to KSD balls *in the infinite-sample limit.*

Theorem 4. *Suppose $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$ and $k(\mathbf{x},\mathbf{x}') = w(\mathbf{x})h(\mathbf{x}-\mathbf{x}')w(\mathbf{x}')$, with stationary reproducing kernel $h \in \mathcal{C}_b^2$ and weighting function $w \in \mathcal{C}_b^1$. Define the set of distributions $\mathcal{P}(\mathbb{R}^d; w) \coloneqq \{Q \in \mathcal{P}(\mathbb{R}^d) : \mathbb{E}_{\mathbf{X}\sim Q}[\|w(\mathbf{X})\mathbf{s}_p(\mathbf{X})\|_2^4] < \infty\}$.*

1. *(Calibration) It holds that*

$$\sup_{Q\in\mathcal{B}^{\mathrm{KSD}}(P;\theta)\cap\mathcal{P}(\mathbb{R}^d;w)} \limsup_{n\to\infty} \mathrm{Pr}_{\mathbb{X}_n\sim Q,\mathbf{W}}\left(\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n)\right) \;\leq\; \alpha \;.$$

2. *(Consistency) For any $Q \in \mathcal{P}(\mathbb{R}^d; w)\backslash\mathcal{B}^{\mathrm{KSD}}(P;\theta)$, it holds that*

$$\lim_{n\to\infty} \mathrm{Pr}_{\mathbb{X}_n\sim Q,\mathbf{W}}\left(\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n)\right) \;=\; 1 \;.$$

The condition $w \in \mathcal{C}_b^1$ requires the weighting function w and its gradient to be bounded. In particular, this holds for $w(x) = 1$, in which case $k(\mathbf{x},\mathbf{x}') = h(\mathbf{x}-\mathbf{x}')$ reduces to an *untilted* stationary kernel such as IMQ or squared-exponential kernels. Theorem 4 also assumes the data-generating distribution Q finitely integrates the fourth moment of the weighted score, i.e., $\|w(\mathbf{x})\mathbf{s}_p(\mathbf{x})\|_2^4$. This automatically holds for any $Q \in \mathcal{P}(\mathbb{R}^d)$ when k is a tilted kernel satisfying the conditions in Lemma 2, in which case $\|w(\mathbf{x})\mathbf{s}_p(\mathbf{x})\|_2$ becomes bounded.

In practice, the decision threshold $q_{\infty,1-\alpha}$ is intractable, and we again use the Monte Carlo estimate $q_{B,1-\alpha}$, defined as the squared root of (5), as an approximation. Compared with the standard KSD test targeting the point null $H_0 : Q = P$, our robust-KSD test uses the same bootstrap procedure to compute the decision threshold, but has a slightly different test statistic $\Delta_\theta(\mathbb{X}_n)$ instead of $D^2(\mathbb{X}_n)$, given by (2), to account for the composite null. Moreover, as $\theta \to 0$, the test statistic Δ_θ approaches $D(\mathbb{X}_n)$, the test statistic of the standard test, and the KSD-ball $\mathcal{B}^{\mathrm{KSD}}(P;\theta)$ falls to the singleton $\{P\}$, so the robust-KSD test reduces to the standard test. Our robust KSD test can hence be viewed as a generalization of the standard KSD test to the composite hypotheses in (8). In particular, the robust KSD test is also qualitatively robust whenever the conditions on k and s_p from Theorem 3 hold; see Proposition 14 in Section A.3 for a formal statement and proof.

Another advantage of our robust KSD test being a direct generalization of the standard test is that no extra computation is required to guarantee robustness. For a given $\theta > 0$, our robust test only requires a minor transformation of the test statistic of the standard KSD test. It hence has the same computational cost as the standard test, namely $\mathcal{O}(n^2 d)$. However, extra computation is potentially needed in determining θ, as it might require optimizing the Stein kernel. This will be discussed in detail in Section 4.2.

Remark 7 (Pointwise and uniform controls)**.** *The Type-I error control shown in Theorem 4 is* pointwise *over the null distributions Q. For robust tests, a more desirable control is often a* uniform *one, whereby the supremum over Q is taken* before *the limit over n (Lehmann and Romano, 2022, Chapter 11). For the proposed test, a uniform control can be shown if the bootstrapped quantile is replaced by the ground-truth quantile, say $q^*_{1-\alpha}(\mathbb{X}_n)$. That is (see also Remark 9 in the appendix),*

$$\limsup_{n\to\infty} \sup_{Q \in \mathcal{B}^{\mathrm{KSD}}(P;\theta)\cap\mathcal{P}(\mathbb{R}^d;w)} \Pr_{\mathbb{X}_n\sim Q,\mathbf{W}}\left(\Delta_\theta(\mathbb{X}_n) > q^*_{1-\alpha}(\mathbb{X}_n)\right) \;\leq\; \alpha \;.$$

However, extending this to the bootstrapped quantile requires uniformly bounding the bootstrap approximation errors, which is non-trivial. We leave this as an open question for future research. In the literature, most kernel-based robust tests that offer uniform controls are two-sample tests using either permutation (Schrab and Kim, 2024) or deviation bounds (Sun and Zou, 2023). Yet, these approaches are either not applicable to our one-sample setting or too conservative. This is why our test uses bootstrapping. Another kernel robust test that uses bootstrapping is Key et al. (2025), but their Type-I error control is also pointwise. In Section D, we propose an alternative robust test that leverages deviation bounds in a manner similar to Sun and Zou (2023) to achieve uniform controls, but at the expense of a lower test power.

4.2 Choosing the Uncertainty Radius

A crucial design choice in our robust tests is the uncertainty radius θ. This should be guided by the types of contamination that the practitioner is willing to tolerate. In this section, we discuss how to choose θ for Huber's contamination models and density-band models when using KSD-balls based on tilted kernels.

Firstly, suppose we are considering Huber's contamination model $\mathcal{P}(P;\epsilon_0)$ for some $\epsilon_0 \in [0,1]$. Then θ should be chosen such that the KSD-ball $\mathcal{B}^{\mathrm{KSD}}(P;\theta)$ contains $\mathcal{P}(P;\epsilon_0)$. The following result, proved in Section A.6, shows how to achieve this given an upper bound on the Stein kernel $\tau_\infty = \sup_{\mathbf{x}\in\mathbb{R}^d} u_p(\mathbf{x},\mathbf{x})$.

Proposition 5. *Suppose k satisfies the conditions in Lemma 2 and $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$. Let $\epsilon_0 \in [0,1]$. Then $\mathcal{P}(P;\epsilon_0) \subseteq \mathcal{B}^{\mathrm{KSD}}(P;\theta)$ if $\theta = \epsilon_0\tau_\infty^{1/2}$ and this bound is tight, i.e., $\sup_{Q\in\mathcal{P}(P;\epsilon_0)} D(Q,P) = \epsilon_0\tau_\infty^{1/2}$.*

This result suggests that, when at most a proportion ϵ_0 of data is corrupted, setting $\theta = \epsilon_0\tau_\infty^{1/2}$ will ensure quantitative robustness to $\mathcal{P}(P;\epsilon_0)$. Note that using a similar proof, Huber's contamination models can also be related to the MMD balls; see Lemma 23 in Section B.2.

One practical issue is that τ_∞, the supremum of the Stein kernel, does not always admit a closed-form expression. One way to compute it is by numerical optimization of $\mathbf{x} \mapsto u_p(\mathbf{x}, \mathbf{x})$, but this assumes the optimizer converges and requires extra computation, thus not suitable in high dimension or when evaluation of the score function is costly. We propose an alternative approach, where τ_∞ is approximated by the maximum of the Stein kernel evaluated at the observed data, i.e., $\max_{i=1,\ldots,n} u_p(\mathbf{X}_i, \mathbf{X}_i)$. This approach requires no extra computation and gives a reasonable estimate with moderate or large n. In Section 5, we use this approach and show empirically that the resulting test is still well-calibrated despite this approximation. Further discussions on how this approximation affects the test performance can be found in Section B.3, where we demonstrate that this approach still controls the Type-I error rate, even with a small sample size.

Remark 8. *Proposition 5 holds for Huber's contamination models of the form $Q = (1-\epsilon)P + \epsilon R$ with $\epsilon \leq \epsilon_0$. In particular, it holds for* any *contamination distribution R. This agnosticism can be useful in practice, because the exact form of contamination is often unknown. On the other hand, when prior knowledge about R* is *available, it is also possible to incorporate it into the proposed test. For example, if R is known to have a support bounded by some $B > 0$, then the bound in Proposition 5 can be tightened by replacing the worst-case bound $\tau_\infty = \sup_{\mathbf{x} \in \mathbb{R}^d} u_p(\mathbf{x}, \mathbf{x})$ with the localized version $\tau_\infty = \sup_{\|\mathbf{x}\|_2 \leq B} u_p(\mathbf{x}, \mathbf{x})$. As discussed in Fauß et al. (2021), such assumptions are realistic in, e.g., highly regulated experimental environments, where any outliers are known to lie within bounded regions.*

Another uncertainty model commonly studied in the literature is the *density-band* model (Kassam, 1981; Hafner, 1993), defined as distributions whose density function lies within an error band of a nominal model with density p, i.e., $\{Q \in \mathcal{P}(\mathbb{R}^d) : \ Q$ has density q with $|q(\mathbf{x}) - p(\mathbf{x})| \leq \delta(\mathbf{x})$ for all $\mathbf{x}\}$, for some function $\delta : \mathbb{R}^d \to [0, \infty)$. It can be shown that density-band models can be rewritten as Huber's contamination models (6) with additional constraints on the outlier distribution (Fauß and Zoubir, 2016). However, compared with Huber's models, density-band models have the advantage of being more interpretable and more natural for certain forms of model disparity such as heavy tails. The following result suggests how the uncertainty radius θ should be chosen for such models and is proved in Section A.6.3.

Proposition 6. *Suppose k satisfies the conditions in Lemma 2 and $\mathbb{E}_{\mathbf{X} \sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$. Furthermore, assume $Q, P \in \mathcal{P}(\mathbb{R}^d)$ admit positive densities q, p on $\mathbb{R}^d$ and $p \in \mathcal{C}^1$. If $|q(\mathbf{x}) - p(\mathbf{x})| \leq \delta(\mathbf{x})$ for some function $\delta : \mathbb{R}^d \to [0, \infty)$ such that $\delta_0 := \int_{\mathbb{R}^d} \delta(\mathbf{x}) \, \mathrm{d}\mathbf{x} < \infty$, then $D(Q, P) \leq \delta_0 \tau_\infty^{1/2}$.*

The bound in Proposition 6 is not tight, as its proof involves bounding an integrand that can take negative values by its absolute value. Nevertheless, as we will show in Section 5.1.4, this bound is not overly loose and can still be useful. In particular, it allows us to design a KSD test that is robust to tail misspecification while still maintaining non-trivial power.

Proposition 5 and 6 also reveal a limitation of our robust KSD test—it can be overly conservative to alternatives that are not the intended contamination but still lie within the chosen KSD-ball uncertainty set. For example, to ensure robustness to Huber's contamination models (6) with tolerance ϵ_0, Proposition 5 suggests setting the uncertainty radius to

$\theta = \epsilon_0 \tau_\infty^{1/2}$. However, the resulting KSD-ball $\mathcal{B}^{\mathrm{KSD}}(P;\theta)$ contains not only Huber's models, but also other distributions such as density-band models. As a result, the robust KSD test will control Type-I errors for all these distributions, regardless of whether they are the intended contamination type. This limitation is not unique to our method; it is a generic drawback of all robust tests based on uncertainty sets (Fauß et al., 2021).

5. Numerical Experiments

We will now evaluate the proposed GOF tests using both synthetic and real data. Unless otherwise mentioned, all standard KSD tests are based on an IMQ kernel $k(\mathbf{x}, \mathbf{x}') = h_{\mathrm{IMQ}}(\mathbf{x} - \mathbf{x}')$ where $h_{\mathrm{IMQ}}(\mathbf{u}) = (1 + \|\mathbf{u}\|_2^2/\lambda^2)^{-1/2}$ with a bandwidth $\lambda^2 > 0$ selected via the median heuristic, i.e., $\lambda_{\mathrm{med}} = \mathrm{Median}\{\|\mathbf{X}_i - \mathbf{X}_j\|_2 : 1 \leq i < j \leq n\}$. All tilted-KSD and robust-KSD tests are based on a *tilted* IMQ kernel with weight $w(\mathbf{x}) = (1 + \|\mathbf{x} - \mathbf{a}\|_2^2/c)^{-b}$, where $\mathbf{a} \in \mathbb{R}^d$ and $c > 0$. Intuitively, $\mathbf{a}$ and c respectively centers and scales the input. We fix $\mathbf{a} = 0$ and $c = 1$ in all experiments, as all data will always be centered and on a suitable scale. More generally, we could replace $\|\mathbf{x} - \mathbf{a}\|_2^2/c$ by a weighted norm of the form $(\mathbf{x} - \mathbf{a})^\top C(\mathbf{x} - \mathbf{a})$, where $C \in \mathbb{R}^{d \times d}$ is a pre-conditioning matrix, chosen possibly as the empirical covariance matrix or robust estimates of it. Since our experiments will focus on sub-Gaussian models, we choose $b = 1/2$. This ensures the Stein kernel is bounded.

All tests have nominal level $\alpha = 0.05$. The probability of rejection is computed by averaging over 100 repetitions, and the 95% confidence intervals are reported. Our robust-KSD test will be shortened as *R-KSD*. Code for reproducing all experiments can be found at `github.com/XingLLiu/robust-kernel-test`.

5.1 Toy Gaussian Model

We first consider a Gaussian model $P = \mathcal{N}(0, 1)$, in which case the score function of the model is $\mathbf{s}_p(x) = -x$ and is hence unbounded. This toy example is simplistic and not representative of the unnormalized models our test is most suited for, but it will nonetheless be helpful to study our algorithmic choices, to verify Theorem 1 and Theorem 3 numerically, and to compare against alternative robust tests.

5.1.1 Decay Rate of Weighting Function

We first draw random samples $\mathbb{X}_n$ of size $n = 500$ from $Q = (1-\epsilon)P + \epsilon\delta_z$, for different values of $\epsilon \in [0, 1]$ and $z \in \mathbb{R}$. This setting mimics the presence of outliers at z. Given a stationary kernel h, a natural question is "how does the choice of weight w affect the Stein kernel and the test power?". On the left-hand side of Figure 1, we plot the Stein kernel $x \mapsto u_p(x, x)$ for different values of b. As b grows, the tails of the Stein kernel are progressively down-weighted. We then plot the rejection probability of the standard KSD tests using these Stein kernels when the outlier is $z = 10$ on the right-hand side of Figure 1. As b increases, the rejection probability decreases, suggesting that the test power is lower when the tails of the Stein kernel are overly down-weighted. This is not surprising since the Stein kernel decays faster for larger values of b, making the test insensitive to model deviations at the tails. Ideally, w should decay just enough so that u_p remains bounded, but not too fast as it would lose power. This highlights the trade-off between robustness and power in the choice of w. A

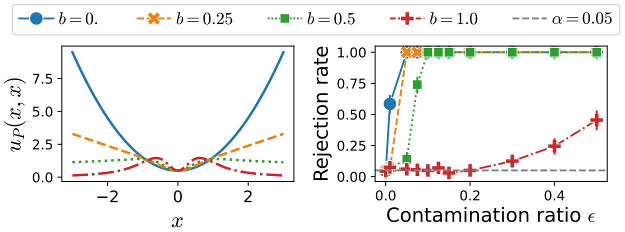


Figure 1: *Left.* Stein kernel for $P = \mathcal{N}(0, 1)$ and an IMQ kernel tilted by $w(x) = (1 + x^2)^{-b}$. The larger b is, the more the tails of the function $x \mapsto u_p(x, x)$ are down-weighted. The choice $b = 0$ corresponds to no weighting, reducing to an IMQ kernel. *Right.* The rejection probability under contamination by $R = \delta_z$ with $z = 10$.

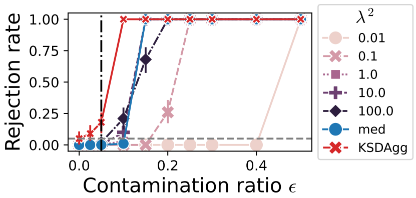


Figure 2: Rejection probability of robust-KSD with different bandwidths λ. "med" is the median heuristic. "KSDAgg" is the test of Schrab et al. (2022). The dashed line is $\alpha = 0.05$. The vertical line is the maximal proportion of contamination $\epsilon_0 = 0.05$ controlled by robust-KSD.

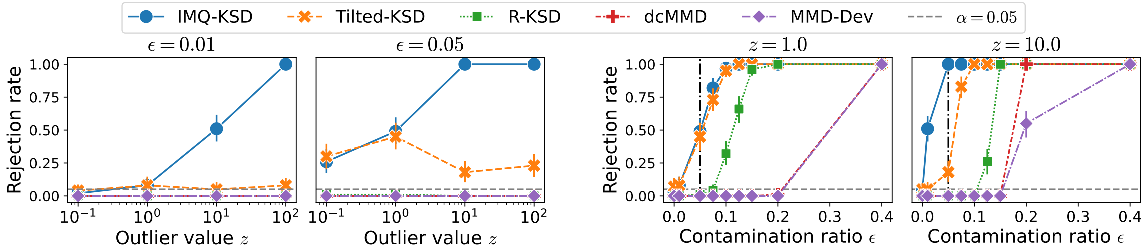


Figure 3: Rejection probability under an outlier-contaminated Gaussian model with different outlier values z and contamination ratios ϵ. The grey dotted horizontal line is the test level $\alpha = 0.05$, and the black dash-dot vertical line corresponds to $\epsilon_0 = 0.05$. The KSD tests with IMQ kernel lack both qualitative and quantitative robustness since they reject even for small z or ϵ. The tilted-KSD test is more robust in cases where z or ϵ are larger, but ultimately still reject the null due to their lack of quantitative robustness.

similar trade-off has been observed in robust estimation, where an overly decaying w can impede the estimation accuracy (Barp et al., 2019, Appendix F.2). Henceforth, we choose $b = 1/2$ to balance this trade-off, as it is the smallest value that ensures u_p is bounded for sub-Gaussian models.

5.1.2 Kernel Bandwidth

We now use the same data set to investigate how the choice of the kernel bandwidth λ affects the performance of the robust-KSD test. It is well-known that the bandwidth plays a crucial role in kernel-based tests and can considerably affect the ability of the test to detect model disparities (Gretton et al., 2012; Ramdas et al., 2015; Reddi et al., 2015; Schrab et al., 2022, 2023). In general, a smaller bandwidth is better at detecting local differences, while a larger bandwidth is more suited for global deviations (Schrab et al., 2023). A common strategy for bandwidth selection in kernel-based testing is the median heuristic. Schrab et al. (2022, 2023) also proposed testing frameworks that aggregate the test result with multiple bandwidths to give the final decision, thereby avoiding bandwidth selection and achieving higher test power. A natural question is therefore: "how does the bandwidth affects the robustness of the proposed test?".

We run the robust-KSD test with various choices of bandwidths $\lambda^2 \in \Lambda \cup \{\lambda^2_{\mathrm{med}}\}$, where $\Lambda = \{0.01, 0.1, 1, 10, 100\}$ and the uncertainty radius θ is set to control at most $\epsilon = 0.05$ contaminations in the sample. We also compare it with the KSDAgg test of Schrab et al. (2022) using the same tilted kernel and aggregating over the collection of bandwidths Λ. Other configurations of KSDAgg follow the recommendations in Schrab et al. (2022, Section 4.2). This experiment is repeated 200 times to reduce randomness, and the results are reported in Figure 2. For any choice of λ, the robust-KSD test is able to control the Type-I error when $\epsilon \leq \epsilon_0$, where ϵ is the proportion of contamination in the data, suggesting that robust-KSD remains well-calibrated for all bandwidths. If instead $\epsilon > \epsilon_0$, all robust-KSD tests eventually reject H_0^{C}, with λ chosen by median heuristic achieving the highest power. The KSDAgg test is not robust and more sensitive to contamination than all robust-KSD tests at all contamination levels ϵ. This is not surprising given that KSDAgg is designed to *maximize* the test power over multiple bandwidths, rendering this test more sensitive to a wide range of alternatives, including contaminated models. As a result, we henceforth do not include KSDAgg in our experiments. A similar study for the *standard* KSD test is given in Section B.4.

5.1.3 Standard versus Robust KSD Tests

Using the same data set, we then compare KSD tests with a stationary kernel to tilted-KSD tests and our proposed robust-KSD test in Figure 3. As shown in the left two plots, with a contamination ratio ϵ of 0.01 or 0.05, the standard IMQ-KSD test rejects the point null with high probability for large outlier values z. This aligns with the lack of qualitative robustness of tests based on stationary kernels proved in Theorem 1. In contrast, the standard Tilted-KSD test rejects with lower probability for *all* z values, aligning with the qualitative robustness result with tilted kernels in Theorem 3. Notably, our theoretical results in both Section 3 and Section 4 apply only to *fixed* kernels, thus excluding kernels based on observed data such as those selected via median heuristic. However, our numerical results suggest that the conclusion of our theory may hold more broadly.

The Type-I error of the tilted-KSD test still exceeds the nominal level $\alpha = 0.05$, suggesting that a tilted kernel alone is not enough to enforce *quantitative* robustness. Notably, the rejection probability of Tilted-KSD is highest when $z = 1$ but declines for larger z. This is because $u_p(z, z)$ with the tilted kernel peaks at around $|z| \approx 1$ and then decreases with z (see Figure 1), making the test more sensitive when $|z| \approx 1$ but less sensitive to large outliers. In contrast, running our robust-KSD test with a pre-set contamination control $\epsilon_0 = 0.05$, we can see from the left two plots of Figure 3 that it remains well-calibrated for all values of outlier z. The results also suggest that R-KSD is conservative since its rejection probability is close to 0; we attribute this to the fact that the robust tests are designed to control *all* types of contamination, of which the injected outlier is only one specific type. To demonstrate the non-trivial power of the robust tests, we show in the right two plots the rejection probability against different contamination ratios ϵ. As ϵ grows, R-KSD eventually is capable of rejecting the null hypothesis $\mathcal{H}_0 : Q \in \mathcal{B}^{\mathrm{KSD}}(P; \theta)$ with test power approaching one.

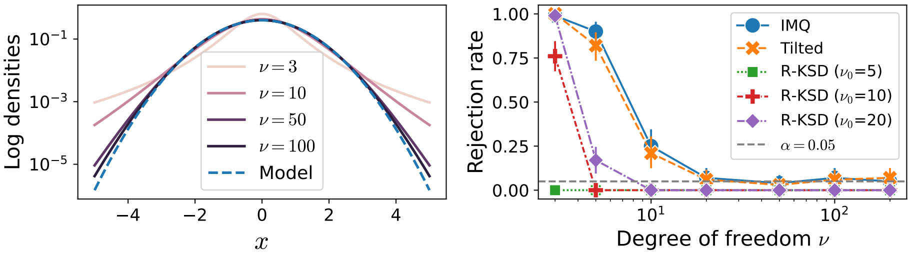


Figure 4: Heavy-tailed experiment. *Left.* Log densities of a standard Gaussian model and scaled t-distributions with different degree-of-freedom (dof) and moment-matched to Gaussian. *Right.* Rejection probability of the standard and robust tests with different θ set to control the cases $\nu \geq \nu_0$ for different values of ν_0.

We also compared our tests with two existing robust kernel tests, namely the robust MMD tests of Schrab and Kim (2024) (denoted *dcMMD*) and of Sun and Zou (2023, Eq. 38) (denoted *MMD-Dev*). Both tests are provably quantitatively robust to Huber's contamination models, but they are two-sample tests that require extra samples from the model P. When the same number of samples from Q and from P are used and the cost of simulation is negligible, the cost of a single evaluation of MMD comparable to that of KSD. However, this cost could be prohibitively expensive when P is a more complex model. Implementation details of these tests are deferred to Section B.2. As shown in Figure 3, both tests are more conservative than our proposed KSD tests. The conservativeness of dcMMD is because it is designed for a contamination model that is *stronger* than Huber's model, while the conservativeness of MMD-Dev is because it uses a deviation inequality to construct the decision threshold, which is known to be conservative; see, e.g., Gretton et al. (2012).

5.1.4 Misspecified Tails

Finally, we now change the data-generating mechanism to study the impact of misspecified tails. We sample from $Q_\nu = t_\nu \sqrt{(\nu-2)/\nu}$, where t_ν is the Student's t-distribution with degree-of-freedom (dof) $\nu > 2$, and the scaling factor $\sqrt{(\nu-2)/\nu}$ ensures its second moment matches that of P. We set the uncertainty radius so that the R-KSD test is robust to Q_ν with $\nu \geq \nu_0$ for different values of ν_0. This is achieved using Proposition 6 to derive a bound on $D(Q_\nu, P)$, which is discussed in detail in Proposition 22 of Section A.6.4. The results are reported in Figure 4. For small values of ν, the tail of the model is misspecified, leading the robust tests with $\nu_0 = 10$ or 20 to reject with high probability because the KSD between Q_ν and P is large. As ν grows, Q_ν converges weakly to the standard Gaussian, making the model well-specified in the limit $\nu \to \infty$. Consequently, neither the standard tests or the robust tests reject the null hypothesis for large ν. The robust test with $\nu_0 = 5$ shows no power because, to be robust to tails that are so misspecified, the uncertainty radius θ computed using Proposition 6 must be very large, thus resulting in a small test statistic (cf. (9)) and low test power. Overall, the robust tests are more conservative than the standard one. This is because the robust tests control *all* possible contaminations within a KSD-ball, while misspecified tails are only one of the many possible such forms of contamination.

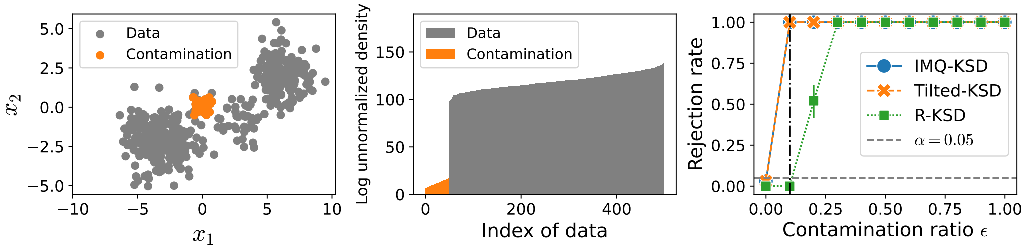


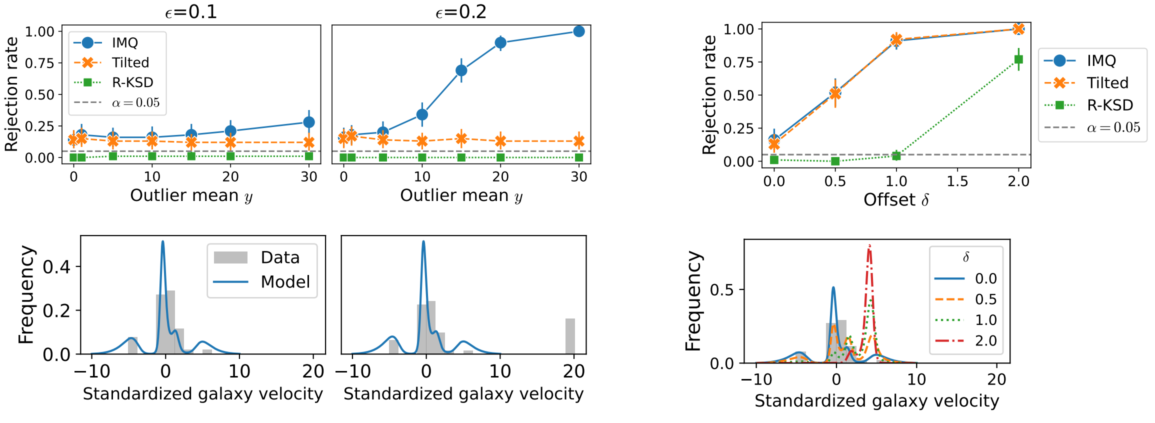
Figure 5: Gaussian-Bernoulli RBM experiment. *Left.* Data generated from P and injected contamination in the first two dimensions. *Middle.* Log unnormalized densities of the data and contamination ordered from small to large; the injected contamination is indeed abnormal since they have much lower densities. *Right.* Probability of rejection with a Gaussian-Bernoulli RBM against the contamination ratio ϵ. The robust tests are calibrated to control the Type-I error with no more than $\epsilon_0 = 0.1$ proportion of contamination.

5.2 Gaussian-Bernoulli Restricted Boltzmann Machines

Our next experiment is an example of energy-based model called a Gaussian-Bernoulli Restricted Boltzmann Machine (RBM) model (Cho et al., 2013). RBMs have been used in a wide range of scientific applications (see Fischer and Igel, 2014; Zhang et al., 2019, for a review), and they are a common benchmarks for assessing kernel GOF tests (Liu et al., 2016; Jitkrittum et al., 2017; Schrab et al., 2022). RBMs are latent-variable models with joint density $p(\mathbf{x}, \mathbf{h}) \propto \exp(\frac{1}{2}\mathbf{x}^\top B\mathbf{h} + \mathbf{b}^\top \mathbf{x} + \mathbf{c}^\top \mathbf{h} - \frac{1}{2}\|\mathbf{x}\|_2^2)$, where $\mathbf{x} \in \mathbb{R}^d$ is an observable variable, $\mathbf{h} \in \{\pm 1\}^{d'}$ is a binary hidden variable with latent dimension d', and $B \in \mathbb{R}^{d \times d'}, \mathbf{b} \in \mathbb{R}^d$ and $\mathbf{c} \in \mathbb{R}^{d'}$ are model parameters. Computing the normalizing constant of the marginal $p(\mathbf{x})$ requires summing over $2^{d'}$ terms, so it becomes intractable when d' is large. However, its score function has a closed-form expression $\mathbf{s}_p(\mathbf{x}) = b - \mathbf{x} + B\tanh(B^\top \mathbf{x} + \mathbf{c})$, where tanh is applied entry-wise. We set $d = 50$ and $d' = 10$, and randomly initialize $B, \mathbf{b}, \mathbf{c}$ by sampling each entry independently from a $\mathcal{N}(0, 1)$. We then generate $n = 500$ samples from the model by block Gibbs sampling following Cho et al. (2013); Jitkrittum et al. (2017), and randomly replace $\epsilon \in [0, 1]$ proportions of the data with outliers drawn from $R = \mathcal{N}(0, 0.1^2 I_d)$; see the left and middle of Figure 5. As shown by the plots, the outliers are clearly abnormal since they have low density values.

We plot the rejection probabilities of the tests on the right-hand side of Figure 5. The standard IMQ-KSD and Tilted-KSD tests are not well-calibrated under the composite null $H_0^{\mathrm{C}} : Q \in \mathcal{B}^{\mathrm{KSD}}(P; \theta)$, while robust-KSD with θ chosen by setting $\epsilon_0 = 0.1$ controls the rejection probability below the level $\alpha = 0.05$ when the contamination ratio ϵ does not exceed ϵ_0. On the other hand, when ϵ exceeds the maximal allowed proportion ϵ_0, the robust test rejects H_0^{C} with probability approaching 1, thus showing its power.

When generating the synthetic data from P, we run a Gibbs sampler for 7000 steps and discard the first 2000 as burn-in and keep every 10-th datum to reduce correlation. This takes 58.40 seconds on average, compared with 0.84 seconds required to perform the robust-KSD test. As a result, it will clearly not be reasonable to use the robust MMD tests from Sun and Zou (2023); Schrab et al. (2023) here as simulating a large enough number of samples from P would cost orders of magnitude more than the cost of the GOF test.



(a) Level

(b) Power

Figure 6: Probability of rejection (top) and data and fitted models (bottom) for the KEF experiment. *(a)* Extra outliers are added to the data; data and fitted models with no contamination (bottom left) and contamination ratio $\epsilon = 0.2$ (bottom right). *(b)* Different offset values δ are added to the fitted parameter to create model deviation. The radius θ is chosen to control contamination of up to a proportion of $\epsilon_0 = 0.2$.

Moreover, the samples generated through Gibbs sampling are no longer i.i.d., thus violating the assumptions in our theoretical results. Extension to the non-independent case could potentially be made following the approach in Chérief-Abdellatif and Alquier (2022), which studied the robustness of estimators (instead of GOF tests) under correlated data.

5.3 A Kernel Exponential Family Model for Density Estimation

Our next example concerns density estimation using a kernel exponential family (KEF) model (Canu and Smola, 2006). Given a reproducing kernel $k_0 : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and a reference density p_0 on $\mathbb{R}^d$, we have $p_\eta(\mathbf{x}) \propto p_0(\mathbf{x}) \exp(-f_\eta(\mathbf{x}))$, where f_η lies in the RKHS $\mathcal{H}_{k_0}$ associated with k_0. We follow the setting in Matsubara et al. (2022) which uses a finite-basis approximation of the RKHS function so that $f_\eta(\mathbf{x}) = \sum_{l=1}^{25} \eta_l \phi_l(\mathbf{x})$, where $\eta = (\eta_1, \ldots, \eta_{25})^\top \in \mathbb{R}^{25}$ and $\phi_l(\mathbf{x}) = (x^l/\sqrt{l!}) \exp(-x^2/2)$ for $l = 1, 2, \ldots, 25$.

We are interested in testing the goodness-of-fit of the KEF model under data contamination. Suppose that the practitioners believe the observed data are subject to contamination, so that a *robust* estimator for η is used. Robust parameter estimation for exponential family models was studied in Barp et al. (2019), which proposed a minimum-distance estimator using KSD with tilted kernels. With a suitable weighting function, the resulting estimator is *bias-robust* (Barp et al., 2019, Proposition 7) and thus not susceptible to contamination. However, how to test the goodness-of-fit of the estimated model in this setting remains an open question. The standard IMQ-KSD test is not suitable since it is not robust and can thus falsely reject a good model due to contamination. Using a tilted kernel is also not enough since it is not quantitatively robust and hence can suffer from uncontrolled Type-I error. We therefore propose to use our robust-KSD test for this task. One challenge with the KEF model is that its density does not have a closed form, and although sampling from

P is feasible in this low-dimensional example ($d = 1$), it would become challenging in high dimensions, making MMD-based tests not well-suited for this model. Our robust-KSD test, however, does not suffer from this limitation.

We use the data set as Matsubara et al. (2022); Key et al. (2025), which is a 1-dimensional data set of 82 galaxy velocities (Postman et al., 1986; Roeder, 1990). To avoid using the same data for model training and testing, we randomly split the data into equal halves, each containing $n_{\text{data}} = 41$ data points. To each half, we then add n_{ol} independent draws from $R = \mathcal{N}(z, 0.1^2)$ for some value of z for the mean contamination, so that the resulting data set has a size of $n = n_{\text{data}} + n_{\text{ol}}$ and a contamination ratio of $\epsilon = n_{\text{ol}}/(n_{\text{data}} + n_{\text{ol}})$. We then fit the KEF model to one half of the data, and perform the tests on the other half. Both the robust minimum KSD estimation and the robust-KSD test use the same tilted kernel $k(x, y) = w(x)h_\Lambda(x - y)w(y)$ with $w(x) = (1 + x^2)^{-1/2}$ and a sum of IMQ kernels $h_\Lambda(x - y) = \sum_{\lambda^2 \in \Lambda}(1 + |x - y|^2/(2\lambda^2))^{-1/2}$ where $\Lambda = \{0.6, 1, 1.2\}$. The choice of the weighting function follows Matsubara et al. (2022) and the use of a sum kernel is recommended in Key et al. (2025). Direct computation shows that the Stein kernel u_p is bounded in this case. We choose $\epsilon_0 = 0.2$ when computing θ for R-KSD.

The results are shown in Figure 6. We first study the level of our tests in Figure 6a. As the outlier mean z increases, the IMQ-KSD test rejects the null $H_0 : Q = P$ with increasing probability, suggesting that this is due to the presence of outliers. Notably, this happens even though the parameter estimator is not susceptible to contamination; see the bottom row of Figure 6a for the fitted model with and without outliers. When the tilted kernel was used, the standard KSD test has a lower rejection probability, although it is still above the nominal level. On the other hand, the robust-KSD test is well-calibrated for all values of z.

We then study the power of KSD tests in Figure 6b. To show the robust test has non-trivial power when no outliers are present, we fit the model parameters $\hat{\eta}$ on half of the data (without contamination) and replace the first component by $\hat{\eta}_1 + \delta$, where $\delta \in \{0, 0.5, 1, 2\}$ is an error offset, so that the resulting model has a poor fit. Other experimental setups including the choice of θ remain the same as before. As expected, the robust test has lower power than the standard tests; however, as the offset value δ increases, the robust test eventually rejects with high probability, thus showing its power under no contamination.

5.4 Limitations for Multimodal Models

We investigate the performance of the robust test when the model has multimodality and the data are drawn from the same model with misspecified mixing ratios, thus revealing a limitation of our robust-KSD test. It is well-known that the standard KSD test performs poorly with mixture models with well-separated modes (Liu et al., 2023). This is intuitively due to the myopia of the score function, making the test blind to disparities in the mixing ratios, and it is a general issue for KSD and other score-based discrepancies (Wenliang and Kanagawa, 2020; Zhang et al., 2022). We demonstrate that our robust test can also suffer from this limitation, which is made even more prominent due to the relaxation of the point null to a KSD-ball. We consider a 2-dimensional mixture of Gaussian model $P = \sum_{j=1}^{5} \pi_j \mathcal{N}(\gamma\mu_j, I_2)$, where $\pi_j \propto u_j$ with $u_1, \ldots, u_5$ drawn independently from Uniform$(0, 1)$, each mean $\mu_j \in \mathbb{R}^d$ has entries drawn independently from Uniform$(-2, 2)$, and $\gamma > 0$ is a scaling factor controlling the mode separation. The data-generating distribution Q has a different set of mixture ratios

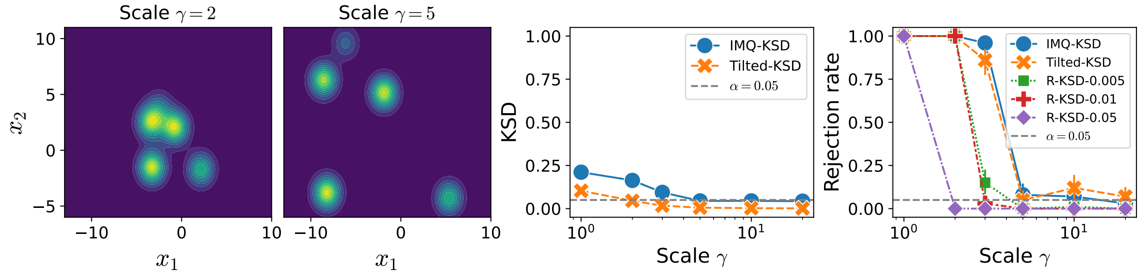


Figure 7: Mixture-of-Gaussian experiment. *Left.* Contour plots of model densities. *Middle.* KSD value with IMQ and tilted kernels. *Right.* Rejection probability of the standard and robust tests. "R-KSD-0.01" refers to the R-KSD test with uncertainty radius θ chosen by setting the maximal contamination tolerance to $\epsilon_0 = 0.01$, and similarly for the others.

that are randomly generated. For R-KSD, the uncertainty radius is chosen to be $\theta = \epsilon_0 \tau_\infty^{1/2}$ with different values of ϵ_0, and we label the corresponding result as R-KSD-ϵ_0.

The results are reported in Figure 7. As the scale γ increases, the modes become more separated, leading to smaller values of $D(Q, P)$, as shown in the left and middle plots. Consequently, Q eventually becomes indistinguishable from P under KSD, and all tests fail to reject with high probability. This is a result of the aforementioned myopia of KSD, which affects both the standard and the robust tests and regardless of whether IMQ or tilted kernels are used. The opposite is observed when γ is small and the modes overlap, where all tests can correctly reject the null with high probability. Compared with the standard tests, the robust one is more conservative because it controls all possible contamination; nevertheless, it still achieves non-trivial power when ϵ_0, thus also the radius θ, is set to a small value.

6. Conclusion and Discussion

This paper studied the robustness of the kernel GOF test of Liu et al. (2016); Chwialkowski et al. (2016). We showed that existing KSD-based tests lack qualitative robustness when stationary kernels are used, but achieve it when employing tilted kernels inspired by robust estimation methods (Barp et al., 2019; Matsubara et al., 2022). Since qualitative robustness alone does not ensure calibration under fixed contamination levels, we proposed a novel robust KSD test that provably controls the Type-I error when the data-generating distribution lies in a KSD-ball around the reference model, while consistently rejecting alternatives outside of it. We then discussed how to select the radius of this ball under different contamination models, namely Huber's contamination model and the density-band model. Empirical results on synthetic and real data sets demonstrate that our robust GOF test is well-calibrated and has non-trivial power. We conclude by discussing promising directions for future work.

Firstly, we established consistency of the proposed robust KSD test against *all* alternative distributions outside the null set, i.e., $H_1^{\mathrm{C}} : Q \notin \mathcal{B}^{\mathrm{KSD}}(P; \theta)$. In the broader robust testing literature, however, it is also common to restrict the alternative set to a neighborhood centered at a specific alternative $P' \neq P$ and study minimax properties against such hypothesis (Huber, 1965; Levy, 2008; Gül and Zoubir, 2016; Sun and Zou, 2023). Understanding whether our robust KSD test achieves minimax optimality in this setting is an interesting future work.

Secondly, as discussed in Section 2.3, our proposed robust KSD test does not require the Stein kernel u_p to be bounded. A bounded Stein kernel is convenient as the resulting KSD-ball $\mathcal{B}^{\mathrm{KSD}}(P;\theta)$ contains many contamination models of interest, such as Huber's model with an *arbitrary* noise distribution. However, this could be too strict in some applications where outliers cannot take an arbitrary value, e.g., when the domain is bounded or the data acquisition procedure is highly regulated to prevent extreme outliers. In these cases, even with an unbounded kernel, the KSD-ball could still capture all contamination models of interest. Unbounded Stein kernels possess appealing theoretical properties such as convergence-control of moments (Kanagawa et al., 2022). Exploring whether such kernels can lead to more powerful robust-KSD tests is another worthwhile future research direction.

Acknowledgments

The authors would like to thank Antonin Schrab, Andrew Duncan, Matias Altamirano, and Charita Dellaporta for helpful discussions. XL was supported by the President's PhD Scholarships of Imperial College London, the EPSRC StatML CDT programme [EP/S023151/1], and the Alan Turing Enrichment Scheme Award. FXB was supported through the EPSRC grant [EP/Y011805/1].

Appendix

Contents

Appendix A. Proofs of Theoretical Results

This section holds proofs of the theoretical results. Throughout the appendix, given i.i.d. random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim Q \in \mathcal{P}(\mathbb{R}^d)$ and an integrable function $f : (\mathbb{R}^d)^n \to \mathbb{R}$, we will write for brevity $\mathbb{E}[f(\mathbf{X}_1, \ldots, \mathbf{X}_n)] = \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} f(\mathbf{x}_1, \ldots, \mathbf{x}_n) Q(\mathrm{d}\mathbf{x}_1) \cdots Q(\mathrm{d}\mathbf{x}_n)$.

A.1 Proof of Theorem 1 and Related Preliminary Results

Overview of proof. For any $\epsilon \in [0, 1]$ and $R \in \mathcal{P}(\mathbb{R}^d)$, a random variable $\mathbf{X}_i$ drawn from $Q = (1-\epsilon)P + \epsilon R$ can be written as $\mathbf{X}_i = (1-\xi_i)\mathbf{X}_i^* + \xi_i \mathbf{Z}_i$, where $\xi_i \sim \text{Bernoulli}(\epsilon)$, $\mathbf{X}_i^* \sim P$ and $\mathbf{Z}_i \sim R$ are independent, and the equality is understood as equality in distribution. Defining the random variable $M' = \sum_{i=1}^n \xi_i$, then $M' \sim \text{Binomial}(n, \epsilon)$ and represents the number of contaminated data in a random sample $\mathbb{X}_n$ drawn from Q. We will first show that the conclusion in Theorem 1 holds *conditional* on the event $\{M' \geq m_0'\}$ for some integer $m_0' \in [0, n]$, i.e., when there are at least m' contaminated data in the sample. This will be sufficient to show Theorem 1 by proving that this event occurs with high probability.

We will use the following notation in the rest of this section: Let $\mathbb{X}_n^*$ be a random sample drawn from Q. For any integer $m' \in [0, n]$ giving the number of corrupted data and any $\mathbf{z} \in \mathbb{R}^d$, we define $\mathbb{X}_{n,\mathbf{z}} = \mathbb{X}_{n-m'}^* \cup \{\mathbf{z}\}^{m'}$.

The rest of this section is organized as follows:

- We first present three preliminary results. Lemma 7 and Lemma 8 respectively bound the test statistic $D^2(\mathbb{X}_{n,\mathbf{z}})$ and $q^2_{\infty,1-\alpha}(\mathbb{X}_{n,\mathbf{z}})$ for a given number of corrupted data, m', by quantities of the form $a_n u_p(\mathbf{z},\mathbf{z}) + b_n$ for some a_n, b_n. These results allow us to compare the relative rate at which $D^2(\mathbb{X}_{n,\mathbf{z}})$ and $q^2_{\infty,1-\alpha}(\mathbb{X}_{n,\mathbf{z}})$ scale with $\mathbf{z}$. Both results will be used to prove Proposition 9, which states that the claim in Theorem 1 holds conditionally on the number of corrupted data.

- Section A.1.1, A.1.2, A.1.3 and A.1.4 provide the proofs for Lemma 7, Lemma 8, Proposition 9 and Theorem 1, respectively.

Assumption 1. *Assume* $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2^4] < \infty$ *and* $k(\mathbf{x},\mathbf{x}') = h(\mathbf{x}-\mathbf{x}')$ *for some* $h \in \mathcal{C}_b^2$ *with* $h(0) > 0$.

Lemma 7. *Assume* $\sigma^2_\infty \coloneqq \sup_{\mathbf{z}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{X}\sim P}[u_p(\mathbf{X},\mathbf{z})^2] < \infty$ *and suppose Assumption 1 holds. Then, for any* $\delta > 0$ *and integers* $0 < m' < n$*, the following event holds with probability at least* $1-\delta$*:*

$$D^2(\mathbb{X}_{n,\mathbf{z}}) \;\geq\; \epsilon^2_{m,n} u_p(\mathbf{z},\mathbf{z}) - \frac{2\sigma_\infty \epsilon_{m,n}}{\sqrt{\delta m}} \;,$$

where $\epsilon_{m,n} = m'/n$ *and* $m = n - m'$.

Lemma 8. *Suppose Assumption 1 holds. Assume* (i) $\sigma^2_\infty \coloneqq \sup_{\mathbf{z}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{X}\sim P}[u_p(\mathbf{X},\mathbf{z})^2] < \infty$ *and* (ii) $\xi^2_P \coloneqq \mathbb{E}_{\mathbf{X}\sim P}[u_p(\mathbf{X},\mathbf{X})^2] < \infty$*. Then there exist constants* $C, C_{P,1}$*, and* $C_{P,2}$ *depending only on* P *such that, for any* $\delta > 0$ *and integers* $0 < m' < n$*, the following event holds with probability at least* $1-\delta$*:*

$$q^2_{\infty,1-\alpha}(\mathbb{X}_{n,\mathbf{z}}) \;\leq\; \frac{C\log(1/\alpha)}{n\sqrt{\delta}}\epsilon_{m,n}u_p(\mathbf{z},\mathbf{z}) + \frac{C_{P,1}}{n\delta} + \frac{C_{P,2}\log(1/\alpha)}{n\sqrt{\delta}} \;,$$

where $\epsilon_{m,n} = m'/n$ *and* $m = n - m'$*, and the probability is over the joint distribution of* $\mathbb{X}_{n,\mathbf{z}}$ *and* $\mathbf{W} \sim$ *Multinomial*$(n; 1/n, \ldots, 1/n)$.

Proposition 9. *There exists a constant* $C > 0$ *such that the following holds. Suppose Assumption 1 holds. Also assume the following integrability conditions*

$$\sup_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{s}_p(\mathbf{z})\|_2^4 \mathbb{E}_{\mathbf{X}\sim P}\big[h(\mathbf{X}-\mathbf{z})^4\big] \;<\; \infty \;, \quad \textit{and} \quad \sup_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{s}_p(\mathbf{z})\|_2^2 \mathbb{E}_{\mathbf{X}\sim P}\big[\|\nabla h(\mathbf{x}-\mathbf{z})\|_2^2\big] \;<\; \infty \;.$$

Further assume that $\mathbf{z} \mapsto \|\mathbf{s}_p(\mathbf{z})\|_2$ *is unbounded. Then for any* $s \in [0,1)$*, there exists* $\mathbf{z} \in \mathbb{R}^d$ *such that, for any sample size* n *and integer* $m' \in \big[\frac{1}{2}n^{1-s}, n\big)$ *giving the number of corrupted data, the condition*

$$n^{\frac{(1-s)}{2}} \;>\; 2\sqrt{2}C\log\left(\frac{4}{\alpha}\right)$$

implies

$$\Pr_{\mathbb{X}^*_{n-m'}\sim P,\mathbf{W}}\left(D^2(\mathbb{X}_{n,\mathbf{z}}) > q^2_{\infty,1-\alpha}(\mathbb{X}_{n,\mathbf{z}})\right) \;\geq\; 1 - n^{-(1-s)}.$$

The assumption $m' \in \big[\frac{1}{2}n^{1-s}, n\big)$ ensures there is a considerable amount of corrupted data in the sample, while excluding the case where all data are contaminated ($m' = n$). The constant factor $1/2$ is arbitrary.

A.1.1 Proof of Lemma 7

Proof The test statistic can be decomposed in three terms: a term with uncorrupted samples, a term with corrupted samples, and a term for interactions between these two groups:

$$\begin{aligned} D^2(\mathbb{X}_{n,\mathbf{z}}) &= \frac{1}{n} \sum_{1 \le i,j \le n} u_P(\mathbf{X}_i, \mathbf{X}_j) \\ &= \frac{1}{n^2} \sum_{1 \le i,j \le m} u_p(\mathbf{X}_i^*, \mathbf{X}_j^*) + \frac{2}{n^2} \sum_{1 \le i \le m < j \le n} u_p(\mathbf{X}_i^*, \mathbf{z}) + \frac{1}{n^2} \sum_{m < i,j \le n} u_p(\mathbf{z}, \mathbf{z}) \\ &= \frac{m^2}{n^2} D^2(\mathbb{X}_m^*) + \frac{2(n-m)}{n^2} \sum_{i=1}^{m} u_p(\mathbf{X}_i^*, \mathbf{z}) + \frac{(n-m)^2}{n^2} u_p(\mathbf{z}, \mathbf{z}) \\ &= (1 - \epsilon_{m,n})^2 D^2(\mathbb{X}_m^*) + 2\epsilon_{m,n}(1 - \epsilon_{m,n}) S_{m,\mathbf{z}} + \epsilon_{m,n}^2 u_p(\mathbf{z}, \mathbf{z}) \ , \end{aligned} \tag{10}$$

where in the last line we have defined $S_{m,\mathbf{z}} \coloneqq m^{-1} \sum_{i=1}^{m} u_p(\mathbf{X}_i, \mathbf{z})$ and $\epsilon_{m,n} \coloneqq (n-m)/n$.

The remainder of the proof proceeds by lower-bounding the first two terms with high probability. The first term can be lower-bounded by 0, since $D^2(\mathbb{X}_m^*)$ is a V-statistic and is hence non-negative. This lower bound will be relatively good as we would expect $D^2((\mathbb{X}_m^*)$ to approach zero at a root-m rate.

Bounding the second term in (10) requires bounding $S_{m,\mathbf{z}}$. Because by assumption $\mathbf{X}_i = \mathbf{X}_i^* \sim P$ are i.i.d. for $i = 1, \ldots, m$, the term $S_{m,\mathbf{z}}$ is a sum of i.i.d. random variables $u_p(\mathbf{X}_i^*, \mathbf{z})$. Moreover, $u_p(\mathbf{X}_i^*, \mathbf{z})$ are zero-mean, because the proof of Gorham and Mackey (2017, Proposition 1) shows that $\mathbb{E}_{\mathbf{X}^* \sim P}[u_p(\mathbf{X}^*, \cdot)] = 0$ whenever $\mathbb{E}_{\mathbf{X}^* \sim P}[\|\mathbf{s}_p(\mathbf{X}^*)\|_2] < \infty$, which holds under Assumption 1. We can therefore compute the variance of $S_{m,\mathbf{z}}$ as

$$\operatorname{Var}(S_{m,\mathbf{z}}) \ = \ \mathbb{E}\left[S_{m,\mathbf{z}}^2\right] \ = \ \frac{1}{m} \mathbb{E}_{\mathbf{X}_1^* \sim P}\left[u_p(\mathbf{X}_1^*, \mathbf{z})^2\right] \ \le \ \frac{\sigma_\infty^2}{m} \ , \tag{11}$$

which is finite since $\sigma_\infty^2 < \infty$ by assumption. We can hence apply Chebyshev's inequality. Choosing

$$t \ \coloneqq \ \left(\frac{\sigma_\infty^2}{\delta m}\right)^{\frac{1}{2}} \ = \ \frac{\sigma_\infty}{\sqrt{\delta m}} \ , \tag{12}$$

the set $\mathcal{A}_1 \coloneqq \left\{|S_{m,\mathbf{z}}| \le t\right\}$ holds with probability at least $1 - \delta$. Furthermore, on $\mathcal{A}_1$, we have by (10) that

$$D^2(\mathbb{X}_n) \ \ge \ (1 - \epsilon_{m,n})^2 D^2(\mathbb{X}_m^*) - 2\epsilon_{m,n}(1 - \epsilon_{m,n}) t + \epsilon_{m,n}^2 u_p(\mathbf{z}, \mathbf{z}) \ \ge \ -2\epsilon_{m,n} t + \epsilon_{m,n}^2 u_p(\mathbf{z}, \mathbf{z}) \ ,$$

where in the last step we have used $\epsilon_{m,n} \in [0, 1]$. This shows the claim. ■

A.1.2 Proof of Lemma 8

To prove this lemma, we will need the following concentration bound for quadratic forms of multinomial random vectors. This is a special instance of the Hanson-Wright inequality due to Adamczak (2015) when the dependent random vectors follow multinomial distributions.

Lemma 10 (Hanson-Wright inequality; Adamczak 2015, Theorem 2.5)**.** *Let* $\mathbf{W} = (W_1, \ldots, W_n) \sim$ *Multinomial*$(n; 1/n, \ldots, 1/n)$*. Then there exists constant* $C > 0$ *such that, for any* $n \times n$ *matrix* A *with entries* $a_{ij} \in \mathbb{R}$ *and any* $\delta > 0$*, we have*

$$\Pr_{\mathbf{W}}\left(\left| \mathbf{W}^\top A \mathbf{W} - \mu \right| \geq C \|A\|_{\mathrm{F}} \log\left(\frac{1}{\delta}\right) \right) \ \leq \ \delta \ ,$$

where $\mu := -n^{-1} \sum_{1 \leq i,j \leq n} a_{ij} + \sum_{1 \leq i \leq n} a_{ii}$*, and* $\|A\|_{\mathrm{F}} = \left(\sum_{1 \leq i,j \leq n} a_{ij}^2 \right)^{1/2}$ *is the Frobenius norm.*

Proof of Lemma 10 Adamczak (2015, Remark 2.2) shows that multinomial random vectors satisfy the convex concentration inequality (Adamczak, 2015, Definition 2.2). Therefore, we can apply Adamczak (2015, Theorem 2.5) to conclude that there exists constants C_1, C_2 such that, for any $n \times n$ matrix A and any $t > 0$,

$$\begin{aligned}
\Pr_{\mathbf{W}}\left(\left| \mathbf{W}^\top A \mathbf{W} - \mu \right| \ \geq \ t \right) \ &\leq \ 2 \exp\left(-\frac{1}{C_1} \min\left(\frac{t^2}{C_2 \|A\|_{\mathrm{F}}^2}, \ \frac{t}{\|A\|_{\mathrm{op}}} \right) \right) \\
&\leq \ 2 \exp\left(-\frac{1}{C_1} \min\left(\frac{t^2}{C_2 \|A\|_{\mathrm{F}}^2}, \ \frac{t}{\|A\|_{\mathrm{F}}} \right) \right) \ ,
\end{aligned}$$

where $\|A\|_{\mathrm{op}} = \sup_{\|\mathbf{x}\|_2 \leq 1} \|A\mathbf{x}\|_2$ is the operator norm, $\mu = \mathbb{E}_{\mathbf{W}}\left[\sum_{1 \leq i,j \leq n} (W_i - 1)(W_j - 1) a_{ij} \right]$, and the last step holds due to the well-known ordering between operator norm and Frobenius norm $\|A\|_{\mathrm{op}} \leq \|A\|_{\mathrm{F}}$ (Golub and van Loan, 2013, Eq. 2.3.7). Moreover, we can compute μ using explicit expressions for the joint central moments of multinomial random variables (Ouimet, 2020, Theorem 2) as follows

$$\begin{aligned}
\mu \ &= \ \sum_{1 \leq i,j \leq n} a_{ij} \mathbb{E}_{\mathbf{W}}\left[(W_i - 1)(W_j - 1) \right] \\
&= \ \sum_{1 \leq i,j \leq n} a_{ij} \mathbb{E}_{\mathbf{W}}\left[(\mathbf{W}_i - 1)(\mathbf{W}_j - 1) \right] + \sum_{1 \leq i \leq n} a_{ii} \mathbb{E}_{\mathbf{W}}\left[(\mathbf{W}_i)^2 \right] \\
&= \ -\frac{1}{n} \sum_{1 \leq i \neq j \leq n} a_{ij} + \left(1 - \frac{1}{n} \right) \sum_{1 \leq i \leq n} a_{ii} \\
&= \ -\frac{1}{n} \sum_{1 \leq i,j \leq n} a_{ij} + \sum_{1 \leq i \leq n} a_{ii} \ .
\end{aligned}$$

For any $\delta > 0$, we claim that the choice

$$t \ = \ \max\left(C_1^{\frac{1}{2}} C_2^{\frac{1}{2}}, \ C_1 \right) \|A\|_{\mathrm{F}} \log\left(\frac{1}{\delta} \right) \tag{13}$$

implies that $\Pr_{\mathbf{W}}(|\mathbf{W}^\top A\mathbf{W} - \mu| \geq t) \leq \delta$, which would complete the proof by setting $C = \max\left(C_1^{1/2}C_2^{1/2}, C_1\right)$. To see this, we note that the the following are equivalent

$$\begin{aligned}
& \delta \;\geq\; 2\exp\left(-\frac{1}{C_1}\min\left(\frac{n^4t^2}{C_2\|A\|_{\mathrm{F}}^2},\, \frac{n^2t}{\|A\|_{\mathrm{F}}}\right)\right) \\
\iff\;& C_1\log\left(\frac{2}{\delta}\right) \;\leq\; \min\left(\frac{n^4t^2}{C_2\|A\|_{\mathrm{F}}^2},\, \frac{n^2t}{\|A\|_{\mathrm{F}}}\right) \\
\iff\;& C_1\log\left(\frac{2}{\delta}\right) \;\leq\; \frac{n^4t^2}{C_2\|A\|_{\mathrm{F}}^2} \qquad\qquad \text{and} \quad C_1\log\left(\frac{2}{\delta}\right) \;\leq\; \frac{n^2t}{\|A\|_{\mathrm{F}}} \\
\iff\;& t \;\geq\; C_1^{1/2}C_2^{1/2}\frac{\|A\|_{\mathrm{F}}}{n^2}\sqrt{\log\left(\frac{2}{\delta}\right)} \qquad \text{and} \quad t \;\geq\; C_1\frac{\|A\|_{\mathrm{F}}}{n^2}\log\left(\frac{2}{\delta}\right) \\
\iff\;& t \;\geq\; \max\left(C_1^{1/2}C_2^{1/2},\, C_1\sqrt{\log\left(\frac{2}{\delta}\right)}\right)\frac{\|A\|_{\mathrm{F}}}{n^2}\sqrt{\log\left(\frac{2}{\delta}\right)} \,.
\end{aligned}$$

Since $\log(2/\delta) > 1$ for $\delta \in (0,1)$, the last inequality is met with t defined in (13). This shows that $\Pr_{\mathbf{W}}\left(\left|\mathbf{W}^\top A\mathbf{W} - \mu\right| \;\geq\; t\right) \leq \delta$, thus finishing the proof. ■

We are now ready to prove Lemma 8.

Proof of Lemma 8 Denote by F_∞ the cumulative distribution function for the conditional distribution of the bootstrap sample $D^2_{\mathbf{W}}(\mathbb{X}_{n,\mathbf{z}})$ (defined in (3)) given $\mathbb{X}_{n,\mathbf{z}}$, and denote its $(1-\alpha)$-quantile as

$$q^2_{\infty,1-\alpha}(\mathbb{X}_{n,\mathbf{z}}) \;\coloneqq\; \inf\{t \in \mathbb{R} :\; 1-\alpha \leq F_\infty(t)\} \,.$$

To bound $q^2_{\infty,1-\alpha/2}(\mathbb{X}_{n,\mathbf{z}})$, we use a similar argument as in Schrab et al. (2023, Appendix E4) by bounding the exceedance probability of $D^2_{\mathbf{W}}(\mathbb{X}_{n,\mathbf{z}})$. Notably, their proof cannot be applied directly here, because they use a wild bootstrap (Leucht and Neumann, 2013) that corresponds to replacing $\mathbf{W}_i - 1$ with independent Rademacher weights, whereas in our case W_i are correlated and multinomially distributed. To proceed, we first note that we can write $D^2_{\mathbf{W}}(\mathbb{X}_{n,\mathbf{z}})$ as a quadratic form as $D^2_{\mathbf{W}}(\mathbb{X}_{n,\mathbf{z}}) = n^{-2}\mathbf{W}^\top A\mathbf{W}$, where $A = (a_{ij})_{ij}$ with $a_{ij} \coloneqq u_p(\mathbf{X}_i, \mathbf{X}_j)$. We can then use the Hanson-Wright inequality stated in Lemma 10 applied with $\delta = \alpha/2$ to show that there exist positive constant C such that, for any almost sure realizations $\mathbb{X}_{n,\mathbf{z}}$,

$$\begin{aligned}
& \Pr_{\mathbf{W}}\left(D^2_{\mathbf{W}}(\mathbb{X}_{n,\mathbf{z}}) \;\geq\; \mu/n^2 + \frac{C\|A\|_{\mathrm{F}}}{n^2}\log\left(\frac{1}{\alpha}\right) \;\Big|\; \mathbb{X}_{n,\mathbf{z}}\right) \\
&= \Pr_{\mathbf{W}}\left(\mathbf{W}^\top A\mathbf{W} \;\geq\; \mu + C\|A\|_{\mathrm{F}}\log\left(\frac{1}{\alpha}\right) \mid \mathbb{X}_{n,\mathbf{z}}\right) \\
&\leq \Pr_{\mathbf{W}}\left(\left|\mathbf{W}^\top A\mathbf{W} - \mu\right| \;\geq\; C\|A\|_{\mathrm{F}}\log\left(\frac{1}{\alpha}\right) \mid \mathbb{X}_{n,\mathbf{z}}\right) \\
&\leq \alpha \,,
\end{aligned}$$

where $\mu \coloneqq -n^{-1}\sum_{1\leq i,j\leq n} a_{ij} + \sum_{1\leq i\leq n} a_{ii}$. This implies that the $(1-\alpha)$-quantile of $D^2_{\mathbf{W}}(\mathbb{X}_n)$ can be bounded as

$$q^2_{\infty,1-\alpha} \;\leq\; \frac{\mu}{n^2} + t \;=\; \frac{\mu}{n^2} + \frac{\rho_\alpha \|A\|_{\mathrm{F}}}{n^2}\;, \tag{14}$$

where we have defined $\rho_\alpha \coloneqq C\log(1/\alpha)$. We now further simplifies this bound by bounding μ and $\|A\|_{\mathrm{F}}$. The matrix A is the Gram matrix of the Stein kernel u_p, thus positive semi-definite. This implies $\sum_{1\leq i,j\leq n} a_{ij} \geq 0$ and $a_{ii} \geq 0$ for all i, so we have

$$\begin{aligned}\mu \;=\; -\frac{1}{n}\sum_{1\leq i,j\leq n} a_{ij} + \sum_{1\leq i\leq n} a_{ii} \;\leq\; \sum_{1\leq i\leq n} a_{ii} \;&=\; \sum_{1\leq i\leq m} a_{ii} + \sum_{m<i\leq n} a_{ii}\\ &=\; \sum_{1\leq i\leq m} a_{ii} + (n-m)u_p(\mathbf{z},\mathbf{z})\;,\end{aligned} \tag{15}$$

where the last quality holds since by assumption $\mathbf{X}_i = \mathbf{z}$ for $i > m$. Moreover, since $a_{ii} = u_p(\mathbf{X}_i, \mathbf{X}_i) \geq 0$ for any i, we can apply Markov inequality to bound the first term on the RHS. This gives that the following holds with probability at least $1-\delta/4$,

$$\begin{aligned}\sum_{1\leq i\leq m} a_{ii} \;\leq\; \frac{4}{\delta}\sum_{1\leq i\leq m}\mathbb{E}[a_{ii}] \;=\; \frac{4m}{\delta}\mathbb{E}_{\mathbf{X}^*\sim P}[u_p(\mathbf{X}^*,\mathbf{X}^*)] \;&\leq\; \frac{4m}{\delta}\big(\mathbb{E}_{\mathbf{X}^*\sim P}[u_p(\mathbf{X}^*,\mathbf{X}^*)^2]\big)^{\frac{1}{2}}\\ &=\; \frac{4m}{\delta}\xi_P\;,\end{aligned}$$

where the second step holds since by assumption $\mathbf{X}_i = \mathbf{X}_i^* \sim P$ are i.i.d. for $i = 1,\ldots,m$, and where we have substituted $\xi_P^2 = \mathbb{E}_{\mathbf{X}_1\sim P}[u_p(\mathbf{X}_1,\mathbf{X}_1)^2]$. Combining with (15) implies that, with probability at least $1-\delta/4$,

$$\mu \;\leq\; \frac{4m}{\delta}\xi_P + (n-m)u_p(\mathbf{z},\mathbf{z})\;. \tag{16}$$

We now bound $\|A\|^2_{\mathrm{F}}$. We first show that $\Sigma_P^2 \coloneqq \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim P}[u_p(\mathbf{X},\mathbf{X}')^2]$ is finite, where $\mathbf{X},\mathbf{X}' \sim P$ are independent. Since u_p is a reproducing kernel (Barp et al., 2024, Theorem 1), we can use the reproducing property and the Cauchy-Schwarz inequality to yield

$$u_p(\mathbf{x},\mathbf{x}') \;=\; \langle u_p(\cdot,\mathbf{x}), u_p(\cdot,\mathbf{x}')\rangle_{\mathcal{H}_u} \leq \|u_p(\cdot,\mathbf{x})\|_{\mathcal{H}_u}\|u_p(\cdot,\mathbf{x}')\|_{\mathcal{H}_u} \;=\; u_p(\mathbf{x},\mathbf{x})^{\frac{1}{2}}u_p(\mathbf{x}',\mathbf{x}')^{\frac{1}{2}}\;,$$

where $\mathcal{H}_u$ denotes the RKHS associated with u_p. This implies

$$\begin{aligned}\Sigma_P^2 \;=\; \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim P}[u_p(\mathbf{X},\mathbf{X}')^2] \;\leq\; \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim P}[u_p(\mathbf{X},\mathbf{X})u_p(\mathbf{X}',\mathbf{X}')] \;&\overset{(a)}{=}\; \big(\mathbb{E}_{\mathbf{X}\sim P}[u_p(\mathbf{X},\mathbf{X})]\big)^2\\ &\overset{(b)}{\leq}\; \mathbb{E}_{\mathbf{X}\sim P}[u_p(\mathbf{X},\mathbf{X})^2]\\ &=\; \xi_P^2 \;<\; \infty\;,\end{aligned}$$

where (a) holds since $\mathbf{X},\mathbf{X}'$ are i.i.d. and (b) follows from Jensen's inequality. By applying Markov's inequality again, with probability at least $1-\delta/4$, where the probability is taken over the randomness of $\mathbb{X}_m^*$, we have

$$\begin{aligned}
&\|A\|_{\mathrm{F}}^2 \\
&= \sum_{1\le i,j\le n} u_p(\mathbf{X}_i,\mathbf{X}_j)^2 \\
&\le \frac{4}{\delta}\sum_{1\le i,j\le n}\mathbb{E}[u_p(\mathbf{X}_i,\mathbf{X}_j)^2] \\
&= \frac{4}{\delta}\bigg(\sum_{1\le i,j\le m}\mathbb{E}[u_p(\mathbf{X}_i^*,\mathbf{X}_j^*)^2] + 2(n-m)\sum_{1\le i\le m}\mathbb{E}[u_p(\mathbf{X}_i^*,\mathbf{z})^2] + (n-m)^2\mathbb{E}[u_p(\mathbf{z},\mathbf{z})^2]\bigg) \\
&= \frac{4}{\delta}\Big(\big(m(m-1)\Sigma_P^2 + m\xi_P^2\big) + 2m(n-m)\mathbb{E}_{\mathbf{X}_1^*\sim P}[u_p(\mathbf{X}_1^*,\mathbf{z})^2] + (n-m)^2u_p(\mathbf{z},\mathbf{z})^2\Big) \\
&\le \frac{4}{\delta}\Big(\big(m(m-1)\Sigma_P^2 + m\xi_P^2\big) + 2m(n-m)\sigma_\infty^2 + (n-m)^2u_p(\mathbf{z},\mathbf{z})^2\Big)\ , \qquad (17)
\end{aligned}$$

where the second last line holds since

$$\begin{aligned}
\sum_{1\le i,j\le m}\mathbb{E}[u_p(\mathbf{X}_i^*,\mathbf{X}_j^*)^2] &= \sum_{1\le i\ne j\le m}\mathbb{E}[u_p(\mathbf{X}_i^*,\mathbf{X}_j^*)^2] + \sum_{1\le i\le m}\mathbb{E}[u_p(\mathbf{X}_i^*,\mathbf{X}_i^*)^2] \\
&= m(m-1)\mathbb{E}[u_p(\mathbf{X}_i^*,\mathbf{X}_j^*)^2] + m\mathbb{E}[u_p(\mathbf{X}_i^*,\mathbf{X}_i^*)^2] \\
&= m(m-1)\Sigma_P^2 + m\xi_P^2\ ,
\end{aligned}$$

and since we have substituted $\Sigma_P^2 = \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim P}[u_p(\mathbf{X},\mathbf{X}')^2]$, $\xi_P^2 = \mathbb{E}_{\mathbf{X}\sim P}[u_p(\mathbf{X},\mathbf{X})^2]$ and $\sigma_\infty^2 = \sup_{\mathbf{z}\in\mathbb{R}^d}\mathbb{E}_{\mathbf{X}\sim P}[u_p(\mathbf{X},\mathbf{z})^2]$. Using the above inequality and (16) to bound (14), we conclude that the following event holds with probability at least $1-\delta/2-\delta/4-\delta/4 = 1-\delta$, where the randomness is taken over the joint distribution of $\mathbb{X}_m^*$ and $\mathbf{W}$,

$$\begin{aligned}
&q_{\infty,1-\alpha}^2(\mathbb{X}_{n,\mathbf{z}}) \\
&\le \frac{4}{n\delta}\xi_P + \frac{n-m}{n^2}u_p(\mathbf{z},\mathbf{z}) \\
&\quad + \frac{\rho_\alpha}{n^2}\sqrt{\frac{4}{\delta}\big(m(m-1)\Sigma_P^2 + m\xi_P^2 + 2m(n-m)\sigma_\infty^2 + (n-m)^2u_p(\mathbf{z},\mathbf{z})^2\big)} \\
&\le \frac{4}{n\delta}\xi_P + \frac{n-m}{n^2}u_p(\mathbf{z},\mathbf{z}) \\
&\quad + \frac{\rho_\alpha}{n^2\sqrt{\delta}}\Big(2\sqrt{m(m-1)}\Sigma_P + \sqrt{m}\xi_P + \sqrt{2m(n-m)}\sigma_\infty + (n-m)u_p(\mathbf{z},\mathbf{z})\Big)\ , \\
&\le \frac{4}{n\delta}\xi_P + \frac{\epsilon_{m,n}}{n}u_p(\mathbf{z},\mathbf{z}) + \frac{\rho_\alpha}{n\sqrt{\delta}}\Big(2\Sigma_P + \xi_P + \sqrt{2\epsilon_{m,n}}\sigma_\infty + \epsilon_{m,n}u_p(\mathbf{z},\mathbf{z})\Big) \\
&\le \frac{4}{n\delta}\xi_P + \frac{2\rho_\alpha\epsilon_{m,n}}{n\sqrt{\delta}}u_p(\mathbf{z},\mathbf{z}) + \frac{\rho_\alpha}{n\sqrt{\delta}}\big(2\Sigma_P + \xi_P + \sqrt{2}\sigma_\infty\big)\ , \qquad (18)
\end{aligned}$$

where the first inequality holds by (16) and (17) (and using $m\le n$), the second inequality follows from using twice the inequality $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$ for any $a,b\ge 0$, the third inequality holds since $m\le n$ and $\epsilon_{m,n} = (n-m)/n$, and (18) holds since $\epsilon_{m,n}\le 1$ and since the conditions $\delta\in(0,1)$ and $\rho_\alpha\ge 1$ imply

$$\frac{\epsilon_{m,n}}{n}u_p(\mathbf{z},\mathbf{z}) \ \le\ \frac{\epsilon_{m,n}}{n\sqrt{\delta}}u_p(\mathbf{z},\mathbf{z}) \ \le\ \frac{\rho_\alpha\epsilon_{m,n}}{n\sqrt{\delta}}u_p(\mathbf{z},\mathbf{z})\ .$$

Substituting the definition of $\rho_\alpha = C\log(1/\alpha)$, and defining the constants $C' \coloneqq 2C$ and $C_{P,1} \coloneqq 4\xi_P$ and $C_{P,2} \coloneqq C(2\Sigma_P + \xi_P + \sqrt{2}\sigma_\infty)$, we have from (18) that

$$q^2_{\infty,1-\alpha}(\mathbb{X}_{n,\mathbf{z}}) \;\le\; \frac{C'\log(1/\alpha)\epsilon_{m,n}}{n\sqrt{\delta}} u_p(\mathbf{z},\mathbf{z}) + \frac{C_{P,1}}{n\delta} + \frac{C_{P,2}\log(1/\alpha)}{n\sqrt{\delta}} \;.$$

■

A.1.3 Proof of Proposition 9

Proof We first show that the moment conditions (i)-(ii) in Lemma 7 and Lemma 8 are met, and apply these lemmas to conclude the proof.

To verify the moment conditions, we first bound the squared Stein kernel evaluated at any $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$ as

$$\begin{aligned}
&u_p(\mathbf{x},\mathbf{z})^2 \\
&= \left[\mathbf{s}_p(\mathbf{x})^\top \mathbf{s}_p(\mathbf{z}) h(\mathbf{x}-\mathbf{z}) - \mathbf{s}_p(\mathbf{x})^\top \nabla h(\mathbf{x}-\mathbf{z}) + \mathbf{s}_p(\mathbf{z})^\top \nabla h(\mathbf{x}-\mathbf{z}) + \nabla^\top \nabla h(\mathbf{x}-\mathbf{z})\right]^2 \\
&\le 4\Big[\big(\mathbf{s}_p(\mathbf{x})^\top \mathbf{s}_p(\mathbf{z}) h(\mathbf{x}-\mathbf{z})\big)^2 + \big(\mathbf{s}_p(\mathbf{x})^\top \nabla h(\mathbf{x}-\mathbf{z})\big)^2 + \big(\mathbf{s}_p(\mathbf{z})^\top \nabla h(\mathbf{x}-\mathbf{z})\big)^2 \\
&\qquad + \big(\nabla^\top \nabla h(\mathbf{x}-\mathbf{z})\big)^2\Big] \\
&\le 4\Big[\|\mathbf{s}_p(\mathbf{x})\|_2^2 \|\mathbf{s}_p(\mathbf{z})\|_2^2 h(\mathbf{x}-\mathbf{z})^2 + \|\mathbf{s}_p(\mathbf{x})\|_2^2 \|\nabla h(\mathbf{x}-\mathbf{z})\|_2^2 + \|\mathbf{s}_p(\mathbf{z})\|_2^2 \|\nabla h(\mathbf{x}-\mathbf{z})\|_2^2 && \text{(19a)} \\
&\qquad + \big(\nabla^\top \nabla h(\mathbf{x}-\mathbf{z})\big)^2\Big] \;, && \text{(19b)}
\end{aligned}$$

where the first inequality holds since $(a+b+c+d)^2 \le 4(a^2+b^2+c^2+d^2)$ for any $a,b,c,d \in \mathbb{R}$, and the second inequality is due to Cauchy-Schwarz inequality. The assumption $h \in \mathcal{C}_b^2$ immediately implies that the last term in (19) is uniformly bounded over $\mathbf{x}, \mathbf{z}$. For the first term, taking expectation over P and supremum over $\mathbf{z}$ gives

$$\begin{aligned}
&\sup_{\mathbf{z}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{X}\sim P}\big[\|\mathbf{s}_p(\mathbf{X})\|_2^2 \|\mathbf{s}_p(\mathbf{z})\|_2^2 h(\mathbf{X}-\mathbf{z})^2\big] \\
&\le \sup_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{s}_p(\mathbf{z})\|_2^2 \big(\mathbb{E}_{\mathbf{X}\sim P}\big[h(\mathbf{X}-\mathbf{z})^4\big]\big)^{1/2} \big(\mathbb{E}_{\mathbf{X}\sim P}\big[\|\mathbf{s}_p(\mathbf{X})\|_2^4\big]\big)^{1/2} \\
&= C_P^{1/2} \sup_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{s}_p(\mathbf{z})\|_2^2 \big(\mathbb{E}_{\mathbf{X}\sim P}\big[h(\mathbf{X}-\mathbf{z})^4\big]\big)^{1/2} \;,
\end{aligned}$$

where the first step holds by Hölder's inequality, and $C_P \coloneqq \mathbb{E}_{\mathbf{X}\sim P}\big[\|\mathbf{s}_p(\mathbf{X})\|_2^4\big]$. The RHS of the last line is finite by assumption. Similarly, the second term in (19) can also be bounded by Hölder's inequality as

$$\sup_{\mathbf{z}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{X}\sim P}\big[\|\mathbf{s}_p(\mathbf{x})\|_2^2 \|\nabla h(\mathbf{x}-\mathbf{z})\|_2^2\big] \;\le\; C_P^{1/2} \sup_{\mathbf{z}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{X}\sim P}\big(\big[\|\nabla h(\mathbf{x}-\mathbf{z})\|_2^4\big]\big)^{1/2} \;,$$

which is finite since $h \in \mathcal{C}_b^2$. For the third term in (19),

$$\sup_{\mathbf{z}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{X}\sim P}\big[\|\mathbf{s}_p(\mathbf{z})\|_2^2 \|\nabla h(\mathbf{x}-\mathbf{z})\|_2^2\big] \;=\; \sup_{\mathbf{z}\in\mathbb{R}^d} \|\mathbf{s}_p(\mathbf{z})\|_2^2 \mathbb{E}_{\mathbf{X}\sim P}\big[\|\nabla h(\mathbf{x}-\mathbf{z})\|_2^2\big] \;,$$

which is again finite by assumption. Substituting these bounds into (19), we have shown that $\sigma_\infty^2 = \sup_{\mathbf{z}\in\mathbb{R}^d} \mathbb{E}_{\mathbf{X}\sim P}[u_p(\mathbf{X},\mathbf{z})^2] < \infty$, which verifies condition (i).

To show (ii), we note that a translation-invariant kernel corresponds to choosing a constant weighting function $w(\mathbf{x}) \equiv 1$ in the tilted kernel of Lemma 2, so that we can leverage the derivations in the proof of Lemma 2. Setting $w(\mathbf{x}) \equiv 1$ in (28) to simplify $u_p(\mathbf{x},\mathbf{x})$ yields

$$\begin{aligned} \xi_P^2 \;\coloneqq\; \mathbb{E}_{\mathbf{X}\sim P}\big[u_p(\mathbf{X},\mathbf{X})^2\big] \;&=\; \mathbb{E}_{\mathbf{X}\sim P}\Big[\big(\|\mathbf{s}_p(\mathbf{X})\|_2^2 h(0) - \nabla^\top\nabla h(0)\big)^2\Big] \\ &\leq\; 2h(0)^2\mathbb{E}_{\mathbf{X}\sim P}\big[\|\mathbf{s}_p(\mathbf{X})\|_2^4\big] + 2\big|\nabla^\top\nabla h(0)\big|^2 \;<\; \infty\;, \end{aligned}$$

where the first inequality holds since $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b \in \mathbb{R}$. This proves condition (ii).

Having showed the moment conditions, we are now ready to prove the main theorem. Pick any $s \in [0,1)$ and any integer $m' \in \big[\frac{1}{2}n^{1-s}, n\big)$. Define $\epsilon_{m,n} \coloneqq m'/n$ and $m = n - m'$. Let $C, C_{P,1}, C_{P,2}$ be the constants defined in Lemma 8, and define the events

$$\mathcal{A}_1 \coloneqq \big\{D^2(\mathbb{X}_n) \;\geq\; \epsilon_{m,n}^2 u_p(\mathbf{z},\mathbf{z}) - t_1(m')\big\}\;, \tag{20a}$$

$$\mathcal{A}_2 \coloneqq \left\{q_{B,1-\alpha}^2(\mathbb{X}_n) \leq \frac{C\log(1/\alpha)}{n\sqrt{\delta/2}}\epsilon_{m,n}u_p(\mathbf{z},\mathbf{z}) + t_2(m')\right\}\;, \tag{20b}$$

where

$$t_1(m') \;\coloneqq\; \frac{2\sigma_\infty\epsilon_{m,n}}{\sqrt{\delta m/2}}\;, \qquad\qquad t_2(m') \;\coloneqq\; \frac{2C_{P,1}}{n\delta} + \frac{C_{P,2}\log(1/\alpha)}{n\sqrt{\delta/2}}\;,$$

and the dependence on m' in these quantities is through $m = n - m'$. It follows from Lemma 7 and Lemma 8 applied to $\delta/2$ that the event $\mathcal{A} \coloneqq \mathcal{A}_1 \cap \mathcal{A}_2$ occurs with probability $\Pr(\mathcal{A}) \geq 1 - \delta/2 - \delta/2 = 1 - \delta$. These events allow us to bound $D^2(\mathbb{X}_{n,\mathbf{z}})$ and $q_{B,1-\alpha}^2(\mathbb{X}_{n,\mathbf{z}})$, which then allows us to show that $D^2(\mathbb{X}_{n,\mathbf{z}}) > q_{B,1-\alpha}^2(\mathbb{X}_{n,\mathbf{z}})$ for sufficiently large $\mathbf{z}$. On $\mathcal{A}$, we have

$$\begin{aligned} D^2(\mathbb{X}_n) - q_{B,1-\alpha}^2(\mathbb{X}_{n,\mathbf{z}}) \;&\geq\; \epsilon_{m,n}^2 u_p(\mathbf{z},\mathbf{z}) - t_1(m') - \frac{C\log(1/\alpha)}{n\sqrt{\delta/2}}\epsilon_{m,n}u_p(\mathbf{z},\mathbf{z}) - t_2(m') \\ &=\; \epsilon_{m,n}u_p(\mathbf{z},\mathbf{z})\left(\epsilon_{m,n} - \frac{C'}{n\sqrt{\delta}}\right) - T(m')\;, \end{aligned}$$

where in the last line we have defined $C' \coloneqq \sqrt{2}C\log(1/\alpha)$ and $T(m') \coloneqq t_1(m') + t_2(m')$. This implies

$$\begin{aligned} &\Pr_{\mathbb{X}_m^*\sim P,\mathbf{W}}\big(D^2(\mathbb{X}_{n,\mathbf{z}}) > \hat{q}_{1-\alpha}^B(\mathbb{X}_{n,\mathbf{z}})\big) \\ &\geq\; \Pr_{\mathbb{X}_m^*\sim P,\mathbf{W}}\big(D^2(\mathbb{X}_{n,\mathbf{z}}) > \hat{q}_{1-\alpha}^B(\mathbb{X}_{n,\mathbf{z}}) \mid \mathcal{A}\big)\Pr_{\mathbb{X}_m^*\sim P,\mathbf{W}}(\mathcal{A}) \\ &\geq\; (1-\delta)\Pr_{\mathbb{X}_m^*\sim P,\mathbf{W}}\big(D^2(\mathbb{X}_{n,\mathbf{z}}) > \hat{q}_{1-\alpha}^B(\mathbb{X}_{n,\mathbf{z}}) \mid \mathcal{A}\big) \\ &\geq\; (1-\delta)\mathbb{1}\left\{\epsilon_{m,n}u_p(\mathbf{z},\mathbf{z})\left(\epsilon_{m,n} - \frac{C'}{n\sqrt{\delta}}\right) - T(m') \;>\; 0\right\}\;, \end{aligned} \tag{21}$$

where the first step holds due to the law of total probability. We claim that the indicator function in (21) takes value 1 for all $m' \in \big[\frac{1}{2}n^{1-s}, n\big)$ when the following inequalities hold

$$n^{\frac{(1-s)}{2}} \;>\; 4C' \;, \qquad\qquad \|\mathbf{s}_p(\mathbf{z})\|_2^2 \;>\; \frac{1}{h(0)}\Big(4T(n-1)n^{2s} + \nabla_1^\top \nabla_2 h(0)\Big) \;. \tag{22}$$

Indeed, since $\delta > 0$ was arbitrary, we can set $\delta = n^{-(1-s)}$, so we have

$$\epsilon_{m,n} - \frac{C'}{n\sqrt{\delta}} \;=\; \epsilon_{m,n} - \frac{C'}{n^{\frac{1}{2}+\frac{s}{2}}} \;\overset{(a)}{>}\; \epsilon_{m,n} - \frac{C'}{4C'n^s} \;=\; \epsilon_{m,n} - \frac{1}{4n^s} \;\overset{(b)}{\geq}\; \epsilon_{m,n} - \frac{\epsilon_{m,n}}{2} \;=\; \frac{\epsilon_{m,n}}{2} \;, \tag{23}$$

where (a) holds since the first condition in (22) implies $n^{\frac{1}{2}+\frac{s}{2}} = n^{s+\frac{(1-s)}{2}} \geq 4C'n^s$, and (b) holds since the assumption $m' \geq \frac{1}{2}n^{1-s}$ implies $n^s \leq n/(2m') = \epsilon_{m,n}/2$. Using this and the second inequality in (22),

$$\begin{aligned}
\epsilon_{m,n} u_p(\mathbf{z},\mathbf{z})\bigg(\epsilon_{m,n} - \frac{C'}{n\sqrt{\delta}}\bigg) \;&=\; \epsilon_{m,n}\big(\|\mathbf{s}_p(\mathbf{z})\|_2^2 h(0) - \nabla_1^\top \nabla_2 h(0)\big)\bigg(\epsilon_{m,n} - \frac{C'}{n\sqrt{\delta}}\bigg) \\
&>\; \epsilon_{m,n} \times 4T(n-1)n^{2s} \times \frac{\epsilon_{m,n}}{2} \\
&\geq\; \epsilon_{m,n} \times \frac{2T(n-1)}{\epsilon_{m,n}^2} \times \frac{\epsilon_{m,n}}{2} \\
&=\; T(n-1) \\
&\geq\; T(m') \;,
\end{aligned}$$

where the first equality holds by setting $w(\mathbf{x}) \equiv 1$ in (28) to compute $u_p(\mathbf{z},\mathbf{z})$, the second line follows from the second inequality in (22) and (23), the third line holds since $n^{2s} \geq n^{2s}/(m')^2 = \epsilon_{m,n}^{-2}$, and the last line holds as direct computation shows $T(n-1) \geq T(m')$ since $m' < n$ by assumption. To summarize, we have shown that for any n and $\mathbf{z}$ satisfying (22), the following holds

$$\Pr_{\mathbb{X}_m^* \sim P, \mathbf{W}}\big(D^2(\mathbb{X}_{n,\mathbf{z}}) > q_{\infty,1-\alpha}^2(\mathbb{X}_{n,\mathbf{z}})\big) \;\geq\; 1 - \delta \;=\; 1 - n^{-(1-s)} \;.$$

It then remains to show that there exists $\mathbf{z}$ not dependent on m' that satisfies the second inequality in (22), which would imply the claimed result. Such $\mathbf{z}$ always exists, because, for the second inequality in (22), the RHS does not depend on $\mathbf{z}$ nor m' and is finite since $\epsilon_{m,n} = m'/n \in (n^{-s}/2, 1)$ by assumption, while the LHS explodes with $\mathbf{z}$ as we have assumed $\mathbf{z} \mapsto \|\mathbf{s}_p(\mathbf{z})\|_2$ is unbounded. This concludes the proof. ■

A.1.4 Proof of Theorem 1

Proof It suffices to show the claim for any sequence of the form $\epsilon_n = n^{-s}$, where $s \in [0, 1)$ is arbitrary. Let the random variable M' be defined as at the start of Section A.1. For any random sample $\mathbb{X}_n$, define the event $\mathcal{A}(\mathbb{X}_n) = \{D^2(\mathbb{X}_n) > q_{\infty,1-\alpha}^2(\mathbb{X}_n)\}$. For n sufficiently large so that $n^{(1-s)/2} > 2\sqrt{2}C\log(1/\alpha)$, we can apply Proposition 9 to conclude that, for all n, there exists $\mathbf{z}_n \in \mathbb{R}^d$ such that the following holds for all $m' \in \big[\frac{1}{2}n^{1-s}, n\big) = \big[\frac{1}{2}\epsilon_n n, n\big)$,

$$\begin{aligned}\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(\mathcal{A}(\mathbb{X}_n) \mid M' = m'\right) \;&=\; \Pr_{\mathbb{X}^*_{n-m'} \sim P, \mathbf{W}}\left(D^2(\mathbb{X}_{n,\mathbf{z}}) > q^2_{\infty,1-\alpha}(\mathbb{X}_{n,\mathbf{z}})\right) \\ &\geq\; 1 - n^{-(1-s)} \;, \end{aligned} \tag{24}$$

where we have used the shorthand notation $\mathbb{X}_{n,\mathbf{z}_n} = \mathbb{X}^*_{n-m'} \cup \{\mathbf{z}_n\}^{m'}$. We will show that the sequence of probability measures $Q := (1-\epsilon_n)P + \epsilon_n \delta_{\mathbf{z}_n}$ satisfies

$$\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(D^2(\mathbb{X}_n) > q^2_{\infty,1-\alpha}(\mathbb{X}_n)\right) - \Pr_{\mathbb{X}^*_n \sim P, \mathbf{W}}\left(D^2(\mathbb{X}^*_n) > q^2_{\infty,1-\alpha}(\mathbb{X}^*_n)\right) \;\to\; 1 - \alpha \;, \tag{25}$$

which will imply the claimed result of our theorem.

Using a concentration inequality for Binomial distributions (e.g., Chung and Lu, 2002, Lemma 2.1) applied to the case of Binomial(n, ϵ_n), the event $\mathcal{B} := \{\epsilon_n n/2 \leq M' \leq 3\epsilon_n n/2\}$ holds with high probability

$$\begin{aligned}\Pr_{M'}(\mathcal{B}) \;&=\; 1 - \Pr_{M'}\left(\left|M' - \epsilon_n n\right| < \frac{\epsilon_n n}{2}\right) \\ &\geq\; 1 - \exp\left(-\frac{(\epsilon_n n)^2}{8\epsilon_n n}\right) - \exp\left(-\frac{(\epsilon_n n)^2}{2(\epsilon_n n + \epsilon_n n/6)}\right) \\ &=\; 1 - \exp\left(-\frac{\epsilon_n n}{8}\right) - \exp\left(-\frac{3\epsilon_n n}{7}\right) \\ &=:\; 1 - f(n) \;, \end{aligned} \tag{26}$$

where in the last line we have defined $f(n) = \exp(-\epsilon_n n/8) - \exp(-3\epsilon_n n/7)$. Define the index set I_n to be the set of integers in the interval $[\epsilon_n n/2, 3\epsilon_n n/3]$. Then

$$\begin{aligned} &\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(D^2(\mathbb{X}_n) > q^2_{\infty,1-\alpha}(\mathbb{X}_n)\right) \\ &\geq\; \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(\mathcal{A}(\mathbb{X}_n) \cap \mathcal{B}\right) \\ &\overset{(a)}{=}\; \sum_{m' \in I_n} \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(\mathcal{A}(\mathbb{X}_n) \mid M' = m'\right) \Pr(M' = m') \\ &\geq\; \left(\min_{m' \in I_n} \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(\mathcal{A}(\mathbb{X}_n) \mid M' = m'\right)\right) \sum_{m' = m'_0}^{n} \Pr(M' = m') \\ &\overset{(b)}{=}\; \min_{m' \in I_n} \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\left(\mathcal{A}(\mathbb{X}_n) \mid M' = m'\right) \Pr(\mathcal{B}) \\ &\overset{(c)}{\geq}\; \left(1 - n^{-(1-s)}\right)\left(1 - f(n)\right) \;\overset{(d)}{=}\; 1 - \mathcal{O}\left(n^{-(1-s)}\right) \;, \end{aligned} \tag{27}$$

where the first inequality follows from the law of total probability, (a) and (b) hold by the partition $\mathcal{B} = \cup_{m' \in I_n}\{M' = m'\}$ and again the law of total probability, (c) follows from (24) and (26), and (d) holds since the assumptions $\epsilon_n = n^{-s}$ and $s \in [0, 1)$ imply $f(n) = o\big(n^{-(1-s)}\big)$. On the other hand, the assumed moment conditions imply $\mathbb{E}_{\mathbf{X}^* \sim P}[u_p(\mathbf{X}^*, \cdot)] = 0$ as argued in the paragraph before (11) and $\mathbb{E}_{\mathbf{X}^*_1, \mathbf{X}^*_2 \sim P}[u_p(\mathbf{X}^*_1, \mathbf{X}^*_2)^2] \leq \mathbb{E}_{\mathbf{X}^* \sim P}[u_p(\mathbf{X}^*, \mathbf{X}^*)^2] < \infty$ as argued in Section A.1.3, so we can apply Arcones and Gine (1992, Theorem 3.5) with $r = 2$ to yield the asymptotic validity of the KSD test when $Q = P$, namely

$$\Pr_{\mathbb{X}^*_n \sim P, \mathbf{W}}\left(D^2(\mathbb{X}^*_n) > q^2_{\infty,1-\alpha}(\mathbb{X}^*_n)\right) \;\to\; \alpha \;.$$

Combining this with (27) shows the desired convergence (25). ■

A.1.5 Implications for KSDs on Non-Euclidean Spaces

The core argument underpinning the proofs in Theorem 1 is that the function $\mathbf{z} \mapsto u_p(\mathbf{z}, \mathbf{z})$ is unbounded on the support of the model, which in our case is $\mathbb{R}^d$. The proof per se does not rely on the support being an Euclidean space. Consequently, this argument can potentially be extended to KSDs constructed for non-Euclidean data, such as Riemannian data (Xu and Matsuda, 2021), sequence data with varying dimensionality (Baum et al., 2023), functional data (Wynne et al., 2022), or censored time-to-event data (Fernandez et al., 2020).

A.2 Proof of Lemma 2

Proof Direct expansion of the Stein kernel in (1) with a tilted kernel k gives

$$\begin{aligned} u_p(\mathbf{x}, \mathbf{x}') \;=\;& \langle w(\mathbf{x})\mathbf{s}_p(\mathbf{x}),\ w(\mathbf{x}')\mathbf{s}_p(\mathbf{x}')\rangle h(\mathbf{x}, \mathbf{x}') \\ &+ \langle w(\mathbf{x}')\mathbf{s}_p(\mathbf{x}'),\ \nabla w(\mathbf{x})\rangle h(\mathbf{x} - \mathbf{x}') + \langle w(\mathbf{x}')\mathbf{s}_p(\mathbf{x}'), \nabla h(\mathbf{x} - \mathbf{x}')\rangle w(\mathbf{x}) \\ &+ \langle w(\mathbf{x})\mathbf{s}_p(\mathbf{x}),\ \nabla w(\mathbf{x}')\rangle h(\mathbf{x} - \mathbf{x}') - \langle w(\mathbf{x})\mathbf{s}_p(\mathbf{x}), \nabla h(\mathbf{x} - \mathbf{x}')\rangle w(\mathbf{x}') \\ &+ \langle \nabla w(\mathbf{x}),\ \nabla w(\mathbf{x}')\rangle h(\mathbf{x} - \mathbf{x}') + \langle w(\mathbf{x})\nabla h(\mathbf{x} - \mathbf{x}'),\ \nabla w(\mathbf{x}')\rangle \\ &- \langle \nabla w(\mathbf{x}),\ w(\mathbf{x}')\nabla h(\mathbf{x} - \mathbf{x}')\rangle - w(\mathbf{x})w(\mathbf{x}')\nabla^\top \nabla h(\mathbf{x} - \mathbf{x}') \\ =\;& \langle \mathbf{s}_{p,w}(\mathbf{x}),\ \mathbf{s}_{p,w}(\mathbf{x}')\rangle h(\mathbf{x} - \mathbf{x}') \\ &+ \langle \mathbf{s}_{p,w}(\mathbf{x}'),\ \nabla w(\mathbf{x})\rangle h(\mathbf{x} - \mathbf{x}') + \langle \mathbf{s}_{p,w}(\mathbf{x}'),\ \nabla h(\mathbf{x} - \mathbf{x}')\rangle w(\mathbf{x}) \\ &+ \langle \mathbf{s}_{p,w}(\mathbf{x}),\ \nabla w(\mathbf{x}')\rangle h(\mathbf{x} - \mathbf{x}') - \langle \mathbf{s}_{p,w}(\mathbf{x}),\ \nabla h(\mathbf{x} - \mathbf{x}')\rangle w(\mathbf{x}') \\ &+ \langle \nabla w(\mathbf{x}),\ \nabla w(\mathbf{x}')\rangle h(\mathbf{x} - \mathbf{x}') + \langle \nabla h(\mathbf{x} - \mathbf{x}'),\ \nabla w(\mathbf{x}')\rangle w(\mathbf{x}) \\ &- \langle \nabla w(\mathbf{x}),\ \nabla h(\mathbf{x} - \mathbf{x}')\rangle w(\mathbf{x}') - w(\mathbf{x})w(\mathbf{x}')\nabla^\top \nabla h(\mathbf{x} - \mathbf{x}')\ , \end{aligned}$$

where $\mathbf{s}_{p,w}(\cdot) \coloneqq w(\cdot)\mathbf{s}_p(\cdot)$ and $\nabla h(\mathbf{x} - \mathbf{x}')$ denotes ∇h evaluated at $\mathbf{x} - \mathbf{x}'$. By setting $\mathbf{x} = \mathbf{x}'$ and eliminating identical terms,

$$\begin{aligned} & u_p(\mathbf{x}, \mathbf{x}) \\ =\;& \|\mathbf{s}_{p,w}(\mathbf{x})\|_2^2 h(0) + 2\langle \mathbf{s}_{p,w}(\mathbf{x}),\ \nabla w(\mathbf{x})\rangle h(0) + \|\nabla w(\mathbf{x})\|_2^2 h(0) - w(\mathbf{x})^2 \nabla^\top \nabla h(0) \qquad (28) \\ \leq\;& \|\mathbf{s}_{p,w}(\mathbf{x})\|_2^2 h(0) + 2\|\mathbf{s}_{p,w}(\mathbf{x})\|_2 \|\nabla w(\mathbf{x})\|_2 h(0) + \|\nabla w(\mathbf{x})\|_2^2 h(0) + w(\mathbf{x})^2 |\nabla^\top \nabla h(0)|\ , \qquad (29) \end{aligned}$$

where the last line follows from Cauchy-Schwarz inequality. The RHS of (29) is bounded over $\mathbf{x}$ assuming the stated conditions on h, w and the supremum of $\|s_p(x)w(x)\|_2 = \|s_{p,w}(x)\|_2$. To prove the bound on $u_p(\mathbf{x}, \mathbf{x}')$, we denote by $\mathcal{H}_u$ the RKHS associated with u_p, which is a reproducing kernel (see, for example, Barp et al., 2024, Theorem 1). We use the reproducing property of the Stein reproducing kernel u_p and the Cauchy-Schwarz inequality for the corresponding RKHS norm $\|\cdot\|_{\mathcal{H}_u}$ to yield

$$u_p(\mathbf{x}, \mathbf{x}') \;=\; \langle u_p(\cdot, \mathbf{x}), u_p(\cdot, \mathbf{x}')\rangle_{\mathcal{H}_u} \leq \|u_p(\cdot, \mathbf{x})\|_{\mathcal{H}_u} \|u_p(\cdot, \mathbf{x}')\|_{\mathcal{H}_u} \;=\; u_p(\mathbf{x}, \mathbf{x})^{\frac{1}{2}} u_p(\mathbf{x}', \mathbf{x}')^{\frac{1}{2}}\ . \qquad (30)$$

Hence, $\sup_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^d} u_p(\mathbf{x},\mathbf{x}') \leq \infty$. In particular, $u_p(\mathbf{x},\mathbf{x})^{1/2}$ is well-defined, since $u_p(\mathbf{x},\mathbf{x}) \geq 0$, which is because the Stein kernel u_p is positive definite. ■

A.3 Proof of Theorem 3 and Related Preliminary Results

Overview of proof. We will show a general result in Proposition 14, which states that the robust-KSD test that rejects $H_0^{\mathrm{C}} : \mathcal{B}^{\mathrm{KSD}}(P;\theta)$ if $\Delta_\theta(\mathbb{X}_n) \coloneqq \max(0, D(\mathbb{X}_n) - \theta) > q_{B,1-\alpha}(\mathbb{X}_n)$ is qualitatively robust for any $\theta \geq 0$. This immediately shows Theorem 3 by setting $\theta = 0$. To show Proposition 14, we follow a similar approach in the proof of Theorem 1, where we will first show that the result holds conditionally on the number of contaminated data, and use a high-probability argument to complete the proof.

The rest of this section is organized as follows:

- Section A.3.1 shows Lemma 11, which states that the bootstrap quantiles $q_{B,1-\alpha}(\mathbb{X}_n)$ and $q_{B,1-\alpha}(\mathbb{X}_n^*)$ computed using any two samples $\mathbb{X}_n$ and $\mathbb{X}_n^*$ that differ by at most m' elements are close to each other with high probability.
- Section A.3.2 shows Lemma 12, which states that the difference in the exceedance probabilities of $\Delta_\theta(\mathbb{X}_n)$ and of $\Delta_\theta(\mathbb{X}_n^*)$ is small.
- Section A.3.3 shows Lemma 13, which states that the rejection probability of a robust-KSD test using $\mathbb{X}_n$ is close to that of a test using $\mathbb{X}_n^*$, assuming $\mathbb{X}_n$ and $\mathbb{X}_n^*$ differ by no more than $o(n^{1/2})$ elements. Its proof relies on Lemma 11 and Lemma 12.
- Section A.3.4 uses Lemma 13 to show Proposition 14, which states that the robust-KSD test is qualitatively robust for any $\theta \geq 0$. This result immediately implies Theorem 3, the proof of which is also contained in this subsection.

Lemma 11. *Assume k is a tilted kernel satisfying the conditions in Lemma 2. Then there exists absolute constants $C_1, C_2 > 0$ such that, for any $\delta > 0$ and any (deterministic) realizations $\mathbb{X}_n = \{\mathbf{x}_i\}_{i=1}^n$ and $\mathbb{X}_n^* = \{\mathbf{x}_i^*\}_{i=1}^n$ that differ by at most m' elements, we have*

$$\Pr_{\mathbf{W}}\left(\left|D^2_{\mathbf{W}}(\mathbb{X}_n) - D^2_{\mathbf{W}}(\mathbb{X}_n^*)\right| \ \leq\ \frac{\tau_\infty \epsilon_{m,n}^{\frac{1}{2}}}{n}\left(C_1 + C_2\log\left(\frac{8}{\delta}\right)\right)\right) \ \geq\ 1-\delta\ ,$$

where $D^2_{\mathbf{W}}(\mathbb{X}_n)$, defined in (3), is the bootstrap sample computed using $\mathbb{X}_n$, and $\epsilon_{m,n} = m'/n$. Moreover, the above inequality implies

$$\Pr_{\mathbf{W}}\left(|q_{\infty,1-\alpha}(\mathbb{X}_n) - q_{\infty,1-\alpha}(\mathbb{X}_n^*)| \ \leq\ \frac{\tau_\infty^{\frac{1}{2}} \epsilon_{m,n}^{\frac{1}{4}}}{n^{\frac{1}{2}}}\sqrt{C_1 + C_2\log\left(\frac{8}{\delta}\right)}\right) \ \geq\ 1-\delta\ .$$

Lemma 12. *Assume $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$ and that k is a tilted kernel satisfying the conditions in Lemma 2. For any $\delta, \gamma > 0$ and any integer $m' \in [0, n]$, it holds that*

$$\begin{aligned}&\sup_{\mathbf{z}_1,\ldots,\mathbf{z}_{m'}\in\mathbb{R}^d}\left|\Pr_{\mathbb{X}_m^*\sim P}\left(D\left(\mathbb{X}_{n-m'}^* \cup \{\mathbf{z}_i\}_{i=1}^{m'}\right) \ >\ \gamma\right) - \Pr_{\mathbb{X}_n^*\sim P}\left(D(\mathbb{X}_n^*) \ >\ \gamma\right)\right| \\ &\leq\ \Pr_{\mathbb{X}_n^*\sim P}\left(|D(\mathbb{X}_n^*) - \gamma| \ \leq\ t_{m,n,\delta}\right) + \delta\ ,\end{aligned}$$

where $t_{m,n,\delta} := \big(4\sqrt{2}\tau_\infty \epsilon_{m,n}/\sqrt{\delta n} + 2\tau_\infty \epsilon_{m,n}^2\big)^{1/2}$.

Lemma 13. *Assume* $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$ *and that* k *is a tilted kernel satisfying the conditions in Lemma 2. For any integers* $\theta \geq 0$*,* $\delta > 0$*, and any sequence* $f_n = o(n^{1/2})$*, there exists* n_0 *such that for any* $n \geq n_0$*, we have*

$$\max_{m' \leq f_n} \omega(m') \ < \ \delta \ ,$$

where the max *is over all non-negative integers* $m' \leq f_n$*, and*

$$\omega(m') \ := \ \big| \mathrm{Pr}_{\mathbb{X}_m^* \sim P, \mathbf{W}} \left(\Delta_\theta(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > q_{\infty,1-\alpha}(\mathbb{X}_m^* \cup \mathbb{Z}_{m'})\right) \\ - \mathrm{Pr}_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\Delta_\theta(\mathbb{X}_n^*) > q_{\infty,1-\alpha}(\mathbb{X}_n^*)\right)\big| \ .$$

Proposition 14. *Assume* $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$ *and that* k *is a tilted kernel satisfying the conditions in Lemma 2. If* $\theta \geq 0$ *and* $\epsilon_n = o(n^{-1/2})$*, then as* $n \to \infty$*,*

$$\sup_{Q \in \mathcal{P}(P;\epsilon_n)} \big| \mathrm{Pr}_{\mathbb{X}_n \sim Q, \mathbf{W}} \left(\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n)\right) - \mathrm{Pr}_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\Delta_\theta(\mathbb{X}_n^*) > q_{\infty,1-\alpha}(\mathbb{X}_n^*)\right)\big| \ \to \ 0 \ ,$$

where $\mathcal{P}(P;\epsilon_n)$ *is the Huber's contamination model defined in* (6)*.*

A.3.1 Proof of Lemma 11

Proof Without loss of generality, we assume $\mathbb{X}_n$ and $\mathbb{X}_n^*$ are ordered so that $\mathbf{x}_i = \mathbf{x}_i^*$ for $i = 1, \ldots, m$. We also define the $n \times n$ matrices $U := (u_{ij})_{1\leq i,j\leq n}$ and $U' := (u'_{ij})_{1\leq i,j\leq n}$, where $u_{ij} := u_p(\mathbf{x}_i, \mathbf{x}_j)$ and $u'_{ij} := u_p(\mathbf{x}_i^*, \mathbf{x}_j^*)$. Since u_p is a reproducing kernel, U and U' are Gram matrices, thus symmetric and positive semi-definite. Furthermore, we define $W_i^0 := W_i - 1$ for any $i = 1, \ldots, n$, and define the vectors $\mathbf{V}_1 := (W_1^0, \ldots, W_m^0, 0, \ldots, 0)^\top$ and $\mathbf{V}_2 := (0, \ldots, 0, W_{m+1}^0, \ldots, W_n^0)^\top$, so that $\mathbf{W}^0 = \mathbf{W} - 1 = \mathbf{V}_1 + \mathbf{V}_2$. It follows that the bootstrap sample $D^2_{\mathbf{W}}(\mathbb{X}_n)$ defined in (3) can be decomposed as

$$\begin{aligned} D^2_{\mathbf{W}}(\mathbb{X}_n) \ &= \ \frac{1}{n^2} \sum_{1\leq i,j\leq n} W_i^0 W_j^0 u_{ij} \ = \ \frac{1}{n^2} (\mathbf{W}^0)^\top U \mathbf{W}^0 \\ &= \ \frac{1}{n^2} \left(\mathbf{V}_1^\top U \mathbf{V}_1 + \mathbf{V}_2^\top U \mathbf{V}_2 + 2\mathbf{V}_1^\top U \mathbf{V}_2\right) \ . \end{aligned}$$

Similarly, we can decompose $D^2_{\mathbf{W}}(\mathbb{X}_n^*)$ as

$$D^2_{\mathbf{W}}(\mathbb{X}_n^*) \ = \ \frac{1}{n^2}\left(\mathbf{V}_1^\top U' \mathbf{V}_1 + \mathbf{V}_2^\top U' \mathbf{V}_2 + 2\mathbf{V}_1^\top U' \mathbf{V}_2\right) .$$

Since $\mathbf{x}_i = \mathbf{x}_i^*$ for all $i \leq m$ by construction, we have $u_{ij} = u'_{ij}$ for $1 \leq i,j \leq m$, so $\mathbf{V}_1^\top U \mathbf{V}_1 = \sum_{1\leq i,j\leq m} W_i^0 W_j^0 u_{ij} = \sum_{1\leq i,j\leq m} W_i^0 W_j^0 u'_{ij} = \mathbf{V}_1^\top U' \mathbf{V}_1$, and we can bound the following difference as

$$\begin{aligned} & n^2 \big| D^2_{\mathbf{W}}(\mathbb{X}_n) - D^2_{\mathbf{W}}(\mathbb{X}_n^*) \big| \\ & = \ \big| \mathbf{V}_2^\top U \mathbf{V}_2 + 2\mathbf{V}_1^\top U \mathbf{V}_2 - \mathbf{V}_2^\top U' \mathbf{V}_2 - 2\mathbf{V}_1^\top U' \mathbf{V}_2 \big| \\ & \leq \ |\mathbf{V}_2^\top U \mathbf{V}_2| + 2|\mathbf{V}_1^\top U \mathbf{V}_2| + |\mathbf{V}_2^\top U' \mathbf{V}_2| + 2|\mathbf{V}_1^\top U' \mathbf{V}_2| \\ & \leq \ |\mathbf{V}_2^\top U \mathbf{V}_2| + 2|\mathbf{V}_1^\top U \mathbf{V}_1|^{\frac{1}{2}} |\mathbf{V}_2^\top U \mathbf{V}_2|^{\frac{1}{2}} + |\mathbf{V}_2^\top U' \mathbf{V}_2| + 2|\mathbf{V}_1^\top U' \mathbf{V}_1|^{\frac{1}{2}} |\mathbf{V}_2^\top U' \mathbf{V}_2|^{\frac{1}{2}} \ , \qquad (31) \end{aligned}$$

where the last line follows from Cauchy-Schwarz inequality applied to the U-weighted inner product $(\mathbf{x},\mathbf{x}') \mapsto \mathbf{x}^\top U \mathbf{x}'$ and the U'-weighted inner product $(\mathbf{x},\mathbf{x}') \mapsto \mathbf{x}^\top U' \mathbf{x}'$, which are well-defined since U, U' are positive semi-definite. The proof proceeds by bounding each term separately.

We discuss how to bound $|\mathbf{V}_1 U \mathbf{V}_1|$ and $|\mathbf{V}_2 U \mathbf{V}_2|$, and the argument for U' is identical. We first define the $n \times n$ matrix with $(A_1)_{ij} = u_{ij}$ for $i, j \leq m$ and $(A_1)_{ij} = 0$, so that we can write $\mathbf{V}_1^\top U \mathbf{V}_1 = \mathbf{W}^\top A_1 \mathbf{W}$. We can apply the Hanson-Wright lemma (Lemma 10) to conclude that there exists positive constants C such that, for any $\delta > 0$ and any almost-sure sequence $\mathbb{X}_n$, the following event occurs with probability at least $1 - \delta/4$,

$$\begin{aligned}
|\mathbf{V}_1^\top A \mathbf{V}_1| \;=\; |\mathbf{W}^\top A_1 \mathbf{W}| \;&\leq\; \left| -\frac{1}{n} \sum_{1 \leq i,j \leq m} u_{ij} + \sum_{1 \leq i \leq m} u_{ii} \right| + C \log\left(\frac{8}{\delta}\right) \left(\sum_{1 \leq i,j \leq m} u_{ij}^2 \right)^{\frac{1}{2}} \\
&\leq\; \frac{1}{n} \sum_{1 \leq i,j \leq m} |u_{ij}| + \sum_{1 \leq i \leq m} |u_{ii}| + C \log\left(\frac{8}{\delta}\right) \left(\sum_{1 \leq i,j \leq m} u_{ij}^2 \right)^{\frac{1}{2}} \\
&\leq\; \frac{m^2}{n} \tau_\infty + m\tau_\infty + C m \tau_\infty \log\left(\frac{8}{\delta}\right) \\
&\leq\; 2m\tau_\infty + C m \tau_\infty \log\left(\frac{8}{\delta}\right) \;, \qquad (32)
\end{aligned}$$

where the second last inequality holds since $|u_{ij}| \leq \sup_{\mathbf{x},\mathbf{x}' \in \mathbb{R}^d} |u_p(\mathbf{x},\mathbf{x}')| \leq \tau_\infty$, and the last line holds since $m \leq n$. Similarly, defining the matrix A_2 with $(A_2)_{ij} = 0$ for $i, j \leq m$ and $(A_2)_{ij} = u_{ij}$ otherwise, we can write $\mathbf{V}_2^\top U \mathbf{V}_2 = \mathbf{W}^\top A_2 \mathbf{W}$, and the same argument as before shows that the following holds with probability at least $1 - \delta/4$,

$$\begin{aligned}
|\mathbf{v}_2^\top A \mathbf{v}_2| \;=\; |\mathbf{W}^\top A_2 \mathbf{W}| \;&\leq\; \left| -\frac{1}{n} \sum_{m < i,j \leq n} u_{ij} + \sum_{m < i \leq n} u_{ii} \right| + C \log\left(\frac{8}{\delta}\right) \left(\sum_{m < i,j \leq n} u_{ij}^2 \right)^{\frac{1}{2}} \\
&\leq\; \frac{(n-m)^2 \tau_\infty}{n} + (n-m)\tau_\infty + C(n-m)\tau_\infty \log\left(\frac{8}{\delta}\right) \\
&\leq\; 2(n-m)\tau_\infty + C(n-m)\tau_\infty \log\left(\frac{8}{\delta}\right) \;, \qquad (33)
\end{aligned}$$

Combining (32) and (33), we conclude that the following holds with probability at least $1 - \delta/2$,

$$\begin{aligned}
&|\mathbf{V}_2^\top U \mathbf{V}_2| + 2|\mathbf{V}_1^\top U \mathbf{V}_1|^{\frac{1}{2}} |\mathbf{V}_2^\top U \mathbf{V}_2|^{\frac{1}{2}} \\
&\leq\; \left(2(n-m)\tau_\infty + C(n-m)\tau_\infty \log\left(\frac{8}{\delta}\right) \right) \\
&\qquad + 2 \left(2m\tau_\infty + C m \tau_\infty \log\left(\frac{8}{\delta}\right) \right)^{\frac{1}{2}} \left(2(n-m)\tau_\infty + C(n-m)\tau_\infty \log\left(\frac{8}{\delta}\right) \right)^{\frac{1}{2}} \\
&=\; \left((n-m) + 2m^{\frac{1}{2}} (n-m)^{\frac{1}{2}} \right) \times \tau_\infty \left(2 + C \log\left(\frac{8}{\delta}\right) \right) \;.
\end{aligned}$$

The same argument shows that the last two terms of (31) can be bounded by the same quantity on an event with probability at least $1-\delta/2$. To conclude, we have shown that with probability at least $1-\delta$,

$$\begin{aligned}
\left|D^2_{\mathbf{W}}(\mathbb{X}_n) - D^2_{\mathbf{W}}(\mathbb{X}^*_n)\right| \;&\leq\; \frac{2}{n^2} \times \left((n-m) + 2(n-m)^{\frac{1}{2}} m^{\frac{1}{2}}\right) \times \tau_\infty \left(2 + C\log\left(\frac{8}{\delta}\right)\right) \\
&=\; \frac{2}{n} \times \frac{(n-m)^{\frac{1}{2}}}{n^{\frac{1}{2}}} \left(\frac{(n-m)^{\frac{1}{2}}}{n^{\frac{1}{2}}} + \frac{2m^{\frac{1}{2}}}{n^{\frac{1}{2}}}\right) \times \tau_\infty \left(2 + C\log\left(\frac{8}{\delta}\right)\right) \\
&\leq\; \frac{6\tau_\infty \epsilon^{\frac{1}{2}}_{m,n}}{n} \left(2 + C\log\left(\frac{8}{\delta}\right)\right) \;=\; \frac{\tau_\infty \epsilon^{\frac{1}{2}}_{m,n}}{n} \left(12 + 6C\log\left(\frac{8}{\delta}\right)\right) \;,
\end{aligned}$$

where in the second last line we have defined $\epsilon_{m,n} = (n-m)/n$ and used $n-m \leq n$ and $m \leq n$. Defining $C_1 = 12$ and $C_2 = 6C$ shows the first claim. The second claim can be shown by first noting that the above inequality implies that their $(1-\alpha)$-quantiles must satisfy

$$\left|q^2_{\infty,1-\alpha}(\mathbb{X}_n) - q^2_{\infty,1-\alpha}(\mathbb{X}^*_n)\right| \;\leq\; \frac{\tau_\infty \epsilon^{\frac{1}{2}}_{m,n}}{n} \left(C_1 + C_2\log\left(\frac{8}{\delta}\right)\right) \;=:\; \rho_{m,n,\delta} \;.$$

We will argue separately for the two cases $(q_{\infty,1-\alpha}(\mathbb{X}_n) + q_{\infty,1-\alpha}(\mathbb{X}^*_n)) \geq \rho^{1/2}_{m,n,\delta}$ and $(q_{\infty,1-\alpha}(\mathbb{X}_n) + q_{\infty,1-\alpha}(\mathbb{X}^*_n)) < \rho^{1/2}_{m,n,\delta}$. In the former case, the above inequality implies

$$\begin{aligned}
\left|q_{\infty,1-\alpha}(\mathbb{X}_n) - q_{\infty,1-\alpha}(\mathbb{X}^*_n)\right| \;&=\; \frac{\left|q_{\infty,1-\alpha}(\mathbb{X}_n) - q_{\infty,1-\alpha}(\mathbb{X}^*_n)\right| \times \left|q_{\infty,1-\alpha}(\mathbb{X}_n) + q_{\infty,1-\alpha}(\mathbb{X}^*_n)\right|}{q_{\infty,1-\alpha}(\mathbb{X}_n) + q_{\infty,1-\alpha}(\mathbb{X}^*_n)} \\
&=\; \frac{\left|q^2_{\infty,1-\alpha}(\mathbb{X}_n) - q^2_{\infty,1-\alpha}(\mathbb{X}^*_n)\right|}{q_{\infty,1-\alpha}(\mathbb{X}_n) + q_{\infty,1-\alpha}(\mathbb{X}^*_n)} \\
&\leq\; \frac{\rho_{m,n,\delta}}{\rho^{\frac{1}{2}}_{m,n,\delta}} \;=\; \rho^{\frac{1}{2}}_{m,n,\delta} \;.
\end{aligned}$$

When instead $(q_{\infty,1-\alpha}(\mathbb{X}_n) + q_{\infty,1-\alpha}(\mathbb{X}^*_n)) < \rho^{1/2}_{m,n,\delta}$, we have

$$\begin{aligned}
\left|q_{\infty,1-\alpha}(\mathbb{X}_n) - q_{\infty,1-\alpha}(\mathbb{X}^*_n)\right| \;&\leq\; \max(q_{\infty,1-\alpha}(\mathbb{X}_n),\; q_{\infty,1-\alpha}(\mathbb{X}^*_n)) \\
&\leq\; q_{\infty,1-\alpha}(\mathbb{X}_n) + q_{\infty,1-\alpha}(\mathbb{X}^*_n) \\
&\leq\; \rho^{\frac{1}{2}}_{m,n,\delta} \;,
\end{aligned}$$

which combined with the previous inequality shows the second claim. ■

A.3.2 Proof of Lemma 12

Proof Pick $\{\mathbf{z}_i\}_{i=1}^{m'} \subset \mathbb{R}^d$, and define $\mathbb{Z}_{m'} := \{\mathbf{z}_i\}_{i=1}^{m'}$. Decomposing the test statistic $D^2(\mathbb{X}^*_m \cup \mathbb{Z}_{m'})$ using a similar approach as in (10), we have

$$D^2(\mathbb{X}^*_m \cup \mathbb{Z}_{m'}) \;=\; \frac{m^2}{n^2} D^2(\mathbb{X}^*_m) + \frac{2}{n^2} S_m + \frac{1}{n^2} \sum_{m<i,j\leq n} u_p(\mathbf{z}_i, \mathbf{z}_j) \;, \tag{34}$$

where $S_m \coloneqq \sum_{i=1}^{m} T_i$ and $T_i \coloneqq \sum_{j=m+1}^{n} u_p(\mathbf{X}_i^*, \mathbf{z}_j)$. Similarly, the test statistic computed using the pure sample $\mathbb{X}_n^*$ can be written as

$$D^2(\mathbb{X}_n^*) \;=\; \frac{m^2}{n^2} D^2(\mathbb{X}_m^*) + \frac{2}{n^2} S_m^* + \frac{1}{n^2} \sum_{m<i,j\leq n} u_p(\mathbf{X}_i^*, \mathbf{X}_j^*) \;,$$

where $S_m^* \coloneqq \sum_{i=1}^{m} T_i^*$ and $T_i^* \coloneqq \sum_{j=m+1}^{n} u_p(\mathbf{X}_i^*, \mathbf{X}_j^*)$. Using a triangle inequality to bound their difference,

$$\begin{aligned} n^2 \big| D^2(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) - D^2(\mathbb{X}_n^*) \big| \;&=\; \Bigg| 2S_m + \sum_{m<i,j\leq n} u_p(\mathbf{z}_i, \mathbf{z}_j) - 2S_m^* - \sum_{m<i,j\leq n} u_p(\mathbf{X}_i, \mathbf{X}_j) \Bigg| \\ &\leq\; 2|S_m| + 2|S_m^*| + \sum_{m<i,j\leq n} |u_p(\mathbf{z}_i, \mathbf{z}_j)| + |u_p(\mathbf{X}_i, \mathbf{X}_j)| \\ &\leq\; 2|S_m| + 2|S_m^*| + 2(n-m)^2 \tau_\infty \;, \end{aligned} \tag{35}$$

where the last line holds since $\sup_{\mathbf{x},\mathbf{x}' \in \mathbb{R}^d} u_p(\mathbf{x}, \mathbf{x}') \leq \tau_\infty$ under the assumed kernel conditions due to Lemma 2. We then bound $|S_m|$ and $|S_m^*|$ with high probability. When $\mathbb{E}_{\mathbf{X}^* \sim P}[\|\mathbf{s}_p(\mathbf{X}^*)\|_2] < \infty$, we have $\mathbb{E}_{\mathbf{X}^* \sim P}[u_p(\mathbf{X}^*, \cdot)] = 0$ as argued in the paragraph before (11), so it is straightforward to see that $\mathbb{E}_{\mathbf{X}_i^* \sim P}[T_i] = 0$ for all i, and thus S_m is a sum of zero-mean i.i.d. random variables T_i. Therefore,

$$\mathrm{Var}(S_m^2) \;=\; m \mathbb{E}_{\mathbf{X}_1^* \sim P}\big[T_1^2\big] \;=\; m \sum_{m<j,l\leq n} \mathbb{E}_{\mathbf{X}_1^* \sim P}[u_p(\mathbf{X}_1^*, \mathbf{z}_j) u_p(\mathbf{X}_1^*, \mathbf{z}_l)] \;\leq\; m(n-m)^2 \tau_\infty^2 \;,$$

where the last step holds again since $\sup_{\mathbf{x},\mathbf{x}' \in \mathbb{R}^d} u_p(\mathbf{x}, \mathbf{x}') \leq \tau_\infty$. For any $\delta > 0$, Chebyshev's inequality implies that the event $\mathcal{A}_1 \coloneqq \{|S_m| \leq \sqrt{m}(n-m)\tau_\infty / \sqrt{\delta/2}\}$ occurs with probability at least $1 - \delta/2$. In other words, $|S_m|$ can be upper bounded on the high-probability event $\mathcal{A}_1$. A similar argument applied to S_m^* shows that the event $\mathcal{A}_2 \coloneqq \big\{|S_m^*| \leq \sqrt{m}(n-m)\tau_\infty / \sqrt{\delta/2}\big\}$ also occurs with probability at least $1 - \delta/2$. On $\mathcal{A} \coloneqq \mathcal{A}_1 \cap \mathcal{A}_2$, which occurs with probability at least $1 - \delta$, we have by (35) that

$$\begin{aligned} \big| D^2(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) - D^2(\mathbb{X}_n^*) \big| \;&\leq\; \frac{4\sqrt{m}(n-m)\tau_\infty}{n^2 \sqrt{\delta/2}} + \frac{2(n-m)^2 \tau_\infty}{n^2} \\ &\leq\; \frac{4\sqrt{2}(n-m)\tau_\infty}{n^{3/2}\sqrt{\delta}} + \frac{2(n-m)^2 \tau_\infty}{n^2} \\ &=\; 4\sqrt{2}\epsilon_{m,n} \frac{\tau_\infty}{\sqrt{\delta n}} + 2\tau_\infty \epsilon_{m,n}^2 \;\eqqcolon\; t_{m,n,\delta}^2 \;. \end{aligned} \tag{36}$$

where the second step holds since $\sqrt{m} \leq \sqrt{n}$, and in the last line we have substituted $\epsilon_{m,n} = (n-m)/n$. We claim that this implies that the following holds on $\mathcal{A}$

$$\big| D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) - D(\mathbb{X}_n^*) \big| \;\leq\; t_{m,n,\delta} \;. \tag{37}$$

To see this, we first assume $D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + D(\mathbb{X}_n^*) \geq t_{m,n,\delta}$. In this case, (36) implies

$$\begin{aligned}
\big|D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) - D(\mathbb{X}_n^*)\big| \;&=\; \frac{\big|D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) - D(\mathbb{X}_n^*)\big| \times \big|D(\mathbb{X}_n) + D(\mathbb{X}_n^*)\big|}{D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + D(\mathbb{X}_n^*)} \\
&=\; \frac{D^2(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) - D^2(\mathbb{X}_n^*)}{D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + D(\mathbb{X}_n^*)} \\
&\leq\; t_{m,n,\delta} \; .
\end{aligned}$$

On the other hand, when $D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + D(\mathbb{X}_n^*) < t_{m,n,\delta}$, we have

$$\big|D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) - D(\mathbb{X}_n^*)\big| \;\leq\; \max\big(D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}),\; D(\mathbb{X}_n^*)\big) \;\leq\; D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + D(\mathbb{X}_n^*) \;<\; t_{m,n,\delta} \; ,$$

where the first inequality holds since $D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}), D(\mathbb{X}_n^*)$ are non-negative. Combining these two cases shows (37). It then follows that, for any $\gamma > 0$,

$$\begin{aligned}
&\mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > \gamma\big) \\
&=\; \mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(\{D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > \gamma\} \cap \mathcal{A}\big) + \mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(\{D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > \gamma\} \cap \mathcal{A}^{\mathsf{c}}\big) \\
&\leq\; \mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(D(\mathbb{X}_n^*) + t_{m,n,\delta} > \gamma\big) + \delta \\
&=\; \mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(D(\mathbb{X}_n^*) > \gamma - t_{m,n,\delta}\big) + \delta \; ,
\end{aligned}$$

where the first step follows from the law of total probability, and the second line holds since $\mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(\{D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > \gamma\} \cap \mathcal{A}^{\mathsf{c}}\big) \leq \mathrm{Pr}(\mathcal{A}^{\mathsf{c}}) \leq \delta$. This implies one side of the desired inequality

$$\begin{aligned}
&\mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > \gamma\big) - \mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(D(\mathbb{X}_n^*) > \gamma\big) \\
&\leq\; \mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(D(\mathbb{X}_n^*) > \gamma - t_{m,n,\delta}\big) - \mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(D(\mathbb{X}_n^*) > \gamma\big) + \delta \\
&=\; \mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(\gamma - t_{m,n,\delta} \;\leq\; D(\mathbb{X}_n^*) \;\leq\; \gamma\big) + \delta \\
&\leq\; \mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(\big|D(\mathbb{X}_n^*) - \gamma\big| \;\leq\; t_{m,n,\delta}\big) + \delta \; ,
\end{aligned}$$

where in the last line we have used

$$\begin{aligned}
\mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(\gamma - t_{m,n,\delta} \;\leq\; D(\mathbb{X}_n^*) \;\leq\; \gamma\big) \;&\leq\; \mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(\gamma - t_{m,n,\delta} \;\leq\; D(\mathbb{X}_n^*) \;\leq\; \gamma + t_{m,n,\delta}\big) \\
&=\; \mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(\big|D(\mathbb{X}_n^*) - \gamma\big| \;\leq\; t_{m,n,\delta}\big) \; .
\end{aligned}$$

A similar argument shows the other direction

$$\begin{aligned}
&\mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > \gamma\big) - \mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(D(\mathbb{X}_n^*) > \gamma\big) \\
&\geq\; \mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(\{D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > \gamma\} \cap \mathcal{A}\big) - \mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(D(\mathbb{X}_n^*) > \gamma\big) \\
&\geq\; \mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(D(\mathbb{X}_n^*) > \gamma + t_{m,n,\delta}\big) - \mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(D(\mathbb{X}_n^*) > \gamma\big) \\
&=\; -\,\mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(\gamma \leq D(\mathbb{X}_n^*) \leq \gamma + t_{m,n,\delta}\big) \\
&\geq\; -\,\mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(\gamma - t_{m,n,\delta} \leq D(\mathbb{X}_n^*) \leq \gamma + t_{m,n,\delta}\big) \\
&=\; -\,\mathrm{Pr}_{\mathbb{X}_n^* \sim P}\big(\big|D(\mathbb{X}_n^*) - \gamma\big| \;\leq\; t_{m,n,\delta}\big) \; ,
\end{aligned}$$

where the first step holds by the law of total probability, and the second step follows from (37). ∎

A.3.3 Proof of Lemma 13

Proof Fix any $\theta, \gamma \geq 0$. Let m', n be any positive integers with $m' \leq f_n$, where $f_n = o(n^{1/2})$. Define $m = n - m'$. Pick $Z_{m'} = \{\mathbf{z}_i\}_{i=1}^{m'} \subset \mathbb{R}^d$, and denote the LHS of the inequality by

$$\begin{aligned} T(\mathbf{z}_1, \ldots, \mathbf{z}'_m) \ \coloneqq \ & \big| \Pr_{\mathbb{X}_m^* \sim P, \mathbf{W}} \big(\Delta_\theta(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > q_{\infty, 1-\alpha}(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) \big) \\ & - \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \big(\Delta_\theta(\mathbb{X}_n^*) > q_{\infty, 1-\alpha}(\mathbb{X}_n^*) \big) \big| \ . \end{aligned}$$

The LHS of the inequality can be bounded as

$$\begin{aligned} T(\mathbf{z}_1, \ldots, \mathbf{z}'_m) \ = \ & \big| \Pr_{\mathbb{X}_m^* \sim P, \mathbf{W}} \big(D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > q_{\infty, 1-\alpha}(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + \theta \big) \\ & \quad - \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \big(D(\mathbb{X}_n^*) > q_{\infty, 1-\alpha}(\mathbb{X}_n^*) + \theta \big) \big| \\ \leq \ & \Big| \Pr_{\mathbb{X}_m^* \sim P, \mathbf{W}} \big(D(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) > q_{\infty, 1-\alpha}(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + \theta \big) \\ & \quad - \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \big(D(\mathbb{X}_n^*) > q_{\infty, 1-\alpha}(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + \theta \big) \Big| \\ & + \Big| \Pr_{\mathbb{X}_m^* \sim P, \mathbf{W}} \big(D(\mathbb{X}_n^*) > q_{\infty, 1-\alpha}(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + \theta \big) \\ & \quad - \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \big(D(\mathbb{X}_n^*) > q_{\infty, 1-\alpha}(\mathbb{X}_n^*) + \theta \big) \Big| \\ =: \ & T_1 + T_2 \ , \end{aligned} \tag{38}$$

where the first equality holds since, for any $\gamma \geq 0$, it can be checked that the inequality $\Delta_\theta = \max(0, D(\mathbb{X}_n) - \theta) > \gamma$ holds if and only if $D(\mathbb{X}_n) < \gamma + \theta$. We will now bound the terms T_1 and T_2, respectively. Denote for brevity $\gamma_{\mathbf{z}} \coloneqq q_{\infty, 1-\alpha}(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) + \theta$ and $\gamma_* \coloneqq q_{\infty, 1-\alpha}(\mathbb{X}_n^*) + \theta$. Fix any $\delta > 0$. Applying Lemma 12 with $\gamma = \gamma_{\mathbf{z}}$ yields

$$\begin{aligned} T_1 \ & \leq \ \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \big(\big| D(\mathbb{X}_n^*) - \gamma_{\mathbf{z}} \big| \ \leq \ t_{m,n,\delta} \big) + \delta \\ & \leq \ \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \big(\big| D(\mathbb{X}_n^*) - \gamma_* \big| \ \leq \ t_{m,n,\delta} + |\gamma_{\mathbf{z}} - \gamma_*| \big) + \delta \ , \end{aligned} \tag{39}$$

where $t_{m,n,\delta} \coloneqq \big(4\sqrt{2}\tau_\infty \epsilon_{m,n} / \sqrt{\delta n} + 2\tau_\infty \epsilon_{m,n}^2 \big)^{1/2}$ is defined in Lemma 12, and (39) follows from a triangle inequality. Moreover, by Lemma 11, there exists an event, say $\mathcal{A}$, with probability at least $1 - \delta$ such that, on $\mathcal{A}$,

$$|\gamma_{\mathbf{z}} - \gamma_*| \ = \ \big| q_{\infty, 1-\alpha}(\mathbb{X}_m^* \cup \mathbb{Z}_{m'}) - q_{\infty, 1-\alpha}(\mathbb{X}_n^*) \big| \ \leq \ \frac{\tau_\infty^{\frac{1}{2}} \epsilon_{m,n}^{\frac{1}{4}}}{n^{\frac{1}{2}}} \sqrt{C_1 + C_2 \log\left(\frac{8}{\delta}\right)} \ =: \ \rho_{m,n,\delta} \ .$$

Combining this with (39) implies that T_1 can be bounded as

$$\begin{aligned} T_1 \ & \leq \ \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \big(\big\{ \big| D(\mathbb{X}_n^*) - \gamma_* \big| \ \leq \ t_{m,n,\delta} + |\gamma_{\mathbf{z}} - \gamma_*| \big\} \cap \mathcal{A} \big) + \Pr(\mathcal{A}^{\mathrm{c}}) + \delta \\ & \leq \ \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \big(\big| D(\mathbb{X}_n^*) - \gamma_* \big| \ \leq \ t_{m,n,\delta} + \rho_{m,n,\delta} \big) + 2\delta \ , \end{aligned} \tag{40}$$

where the second line holds since $\Pr(\mathcal{A}^{\mathrm{c}}) \leq \delta$. To bound T_2, we first note that

$$\begin{aligned} T_2 \ & = \ \big| \Pr_{\mathbf{X}_n^* \sim P} \big(D(\mathbb{X}_n^*) > \gamma_{\mathbf{z}} \big) - \Pr_{\mathbf{X}_n^* \sim P} \big(D(\mathbb{X}_n^*) > \gamma_* \big) \big| \\ & = \ \Pr_{\mathbf{X}_n^* \sim P} \big(\min(\gamma_{\mathbf{z}}, \ \gamma_*) \ < \ D(\mathbb{X}_n^*) \ \leq \ \max(\gamma_{\mathbf{z}}, \ \gamma_*) \big) \\ & \leq \ \Pr_{\mathbf{X}_n^* \sim P} \big(\big\{ \min(\gamma_{\mathbf{z}}, \ \gamma_*) \ < \ D(\mathbb{X}_n^*) \ \leq \ \max(\gamma_{\mathbf{z}}, \ \gamma_*) \big\} \cap \mathcal{A} \big) + \Pr(\mathcal{A}^{\mathrm{c}}) \\ & \leq \ \Pr_{\mathbf{X}_n^* \sim P} \big(\big| D(\mathbb{X}_n^*) - \gamma_* \big| \ \leq \ \rho_{m,n,\delta} \big) + \delta \ , \end{aligned} \tag{41}$$

where the second line holds since $|\Pr(Y > a) - \Pr(Y > b)| = \Pr(\min(a, b) < Y \leq \max(a, b))$ for any constants a, b and random variable Y, and the last line holds since $\Pr(\mathcal{A}^{\mathsf{c}}) \leq \delta$ and since on $\mathcal{A}$ we have

$$\begin{aligned} \gamma_* - \rho_{m,n,\delta} \;\leq\; \gamma_* - \big|\gamma_{\mathbf{z}} - \gamma_*\big| \;\leq\; \min(\gamma_{\mathbf{z}},\, \gamma_*) \;\leq\; \max(\gamma_{\mathbf{z}},\, \gamma_*) \;&\leq\; \gamma_* + \big|\gamma_{\mathbf{z}} - \gamma_*\big| \\ &\leq\; \gamma_* + \rho_{m,n,\delta}\,. \end{aligned}$$

Substituting the bounds (40) and (41) into (38) gives

$$\begin{aligned} T(\mathbf{z}_1, \ldots, \mathbf{z}'_m) \;&\leq\; \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}}\big(\big|D(\mathbb{X}_n^*) - \gamma_*\big| \;\leq\; t_{m,n,\delta} + \rho_{m,n,\delta}\big) \\ &\qquad + \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}}\big(\big|D(\mathbb{X}_n^*) - \gamma_*\big| \;\leq\; \rho_{m,n,\delta}\big) + 3\delta \\ &\leq\; 2\Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}}\big(\big|D(\mathbb{X}_n^*) - \gamma_*\big| \;\leq\; t_{m,n,\delta} + \rho_{m,n,\delta}\big) + 3\delta\,, \end{aligned} \tag{42}$$

where the last inequality holds since $t_{m,n,\delta} \geq 0$. Since $m' \leq f_n$ and $f_n = o(n^{1/2})$ by assumption, we have $\epsilon_{m,n} = m'/n \leq f_n/n = o(n^{-1/2})$. This implies

$$\begin{aligned} t_{m,n,\delta} \;=\; \Big(4\sqrt{2}\tau_\infty \delta^{-\frac{1}{2}} \epsilon_{m,n} n^{-\frac{1}{2}} + 2\tau_\infty \epsilon_{m,n}^2\Big)^{\frac{1}{2}} \;&\leq\; \Big(4\sqrt{2}\tau_\infty \delta\Big)^{\frac{1}{2}} \epsilon_{m,n}^{\frac{1}{2}} n^{-\frac{1}{4}} + \sqrt{2}\tau_\infty^{\frac{1}{2}} \epsilon_{m,n} \\ &\in\; o\Big(n^{-\frac{1}{2}}\Big)\,, \end{aligned}$$

which holds since τ_∞, δ are constants, and

$$\begin{aligned} \rho_{m,n,\delta} \;=\; \frac{\tau_\infty^{\frac{1}{2}} \epsilon_{m,n}^{\frac{1}{4}}}{n^{\frac{1}{2}}} \sqrt{C_1 + C_2 \log\big(8n^{2(1-2s)}\big)} \;&=\; \frac{\tau_\infty^{\frac{1}{2}} \epsilon_{m,n}^{\frac{1}{4}}}{n^{\frac{1}{2}}} \sqrt{C'_1 + C'_2 \log(n)} \\ &\in\; o\Big(n^{-\frac{1}{2}}\Big)\,, \end{aligned} \tag{43}$$

where we have defined $C'_1 \coloneqq C_1 + C_2 \log 8$ and $C'_2 \coloneqq 2C_2(1 - 2s)$, and the last step holds since $\epsilon_{m,n}^{1/4}\sqrt{\log(n)} = o\big(n^{-1/8}\sqrt{\log(n)}\big) = o(1)$.

Define $\eta(m') \coloneqq t_{m,n,\delta} + \rho_{m,n,\delta}$, where the dependence on m' is through $m = n - m'$. Then the above derivations show that $\eta(m') = o(n^{-1/2})$ for all $m' \leq f_n$. In particular, $\eta(f_n) = o(n^{-1/2})$. Moreover, substituting these into (42) gives

$$\begin{aligned} T(\mathbf{z}_1, \ldots, \mathbf{z}'_m) \;&\leq\; 2\Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}}\big(\big|D(\mathbb{X}_n^*) - \gamma_*\big| \;\leq\; \eta(m')\big) + 3\delta \\ &\leq\; 2\Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}}\big(\big|D(\mathbb{X}_n^*) - \gamma_*\big| \;\leq\; \eta\,(f_n)\big) + 3\delta\,, \end{aligned}$$

where the last line holds since $m' \leq f_n$ by assumption and $m' \mapsto \eta(m')$ is monotone increasing by direct computation. Taking supremum over $\mathbb{Z}_{m'} \subset \mathbb{R}^d$ and $m' \leq f_n$ gives

$$\max_{m' \leq f_n} \sup_{\mathbf{z}_1, \ldots, \mathbf{z}_{m'} \in \mathbb{R}^d} T(\mathbf{z}_1, \ldots, \mathbf{z}'_m) \;\leq 2\Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}}\big(\big|D(\mathbb{X}_n^*) - \gamma_*\big| \;\leq\; \eta\,(f_n)\big) + 3\delta$$

It remains to show that the first term on the RHS can be bounded by δ, from which the claimed result in Lemma 13 would follow by redefining δ. To proceed, we write

$$\begin{aligned} &\Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}}\big(\,|\Delta_\theta(\mathbb{X}_n^*) - q_{\infty,1-\alpha}(\mathbb{X}_n^*)| \;\leq\; \eta\,(f_n)\,\big) \\ &=\; \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}}\big(\Delta_\theta(\mathbb{X}_n^*) \;\leq\; q_{\infty,1-\alpha}(\mathbb{X}_n^*) + \eta\,(f_n)\,\big) \\ &\quad - \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}}\big(\Delta_\theta(\mathbb{X}_n^*) \;\leq\; q_{\infty,1-\alpha}(\mathbb{X}_n) - \eta\,(f_n)\,\big) \\ &=\; \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}}\big(\sqrt{n}D(\mathbb{X}_n^*) \;\leq\; \sqrt{n}\zeta_{n,\theta} + \sqrt{n}\eta\,(f_n)\,\big) && \text{(44a)} \\ &\quad - \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}}\big(\sqrt{n}D(\mathbb{X}_n^*) \;\leq\; \sqrt{n}\zeta_{n,\theta} - \sqrt{n}\eta\,(f_n)\,\big) && \text{(44b)} \end{aligned}$$

where the first equality holds again because $|\Pr(Y > a) - \Pr(Y > b)| = \Pr(\min(a, b) < Y \leq \max(a, b))$ for any constants a, b and random variable Y, the second equality holds by noting that $\Delta_\theta(\mathbb{X}_n^*) = \max(0, D(\mathbb{X}_n^*) - \theta) \leq t$ for some $t > 0$ if and only if $D(\mathbb{X}_n^*) - \theta \leq t$, and (44) holds by multipling $\sqrt{n}$ on both sides of the inequalities within the probabilities and defining $\zeta_{n,\theta} \coloneqq q_{\infty,1-\alpha}(\mathbb{X}_n^*) + \theta$.

Under the assumed conditions on k and assuming $\mathbb{E}_{\mathbf{X}^* \sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$, Gorham and Mackey (2017, Proposition 1) shows that $\mathbb{E}_{\mathbf{X}^* \sim P}[u_p(\mathbf{X}^*, \cdot)] = 0$, so the V-statistic $D^2(\mathbb{X}_n^*)$ is *degenerate* (see, e.g., Serfling, 2009, Chapter 6), and classic results on the asymptotics of degenerate V-statistics (Serfling, 2009, Theorem 6.4.1 A) shows that $nD^2(\mathbb{X}_n^*)$ converges weakly to a non-negative distribution. Since the square-root function $x \mapsto \sqrt{x}$ is continuous on $[0, \infty)$ and continuous functions preserve weak limits by the Continuous Mapping Theorem (Van der Vaart, 2000, Theorem 2.3), the squared-root statistic, $\sqrt{n}D(\mathbb{X}_n^*)$, also converges weakly to a non-negative distribution. In particular, the scaled quantile $\sqrt{n}q_{\infty,1-\alpha}(\mathbb{X}_n^*)$, and thus also $\sqrt{n}\zeta_{n,\theta}$, converge to a positive number as $n \to \infty$. Also since $\eta\,(f_n) = o(n^{-1/2})$ as argued below (43), we have $\sqrt{n}\eta\,(f_n) \to 0$. In summary, we have shown that

$$\begin{aligned}
&\lim_{n\to\infty} \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(|\Delta_\theta(\mathbb{X}_n^*) - q_{\infty,1-\alpha}(\mathbb{X}_n^*)| \ \leq \ \eta\,(f_n) \right) \\
&= \lim_{n\to\infty} \Big(\Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\sqrt{n}D(\mathbb{X}_n^*) \ \leq \ \sqrt{n}\zeta_{n,\theta} + \sqrt{n}\eta\,(f_n) \right) \\
&\qquad\qquad - \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\sqrt{n}D(\mathbb{X}_n^*) \ \leq \ \sqrt{n}\zeta_{n,\theta} - \sqrt{n}\eta\,(f_n) \right) \Big) \\
&= \lim_{n\to\infty} \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\sqrt{n}D(\mathbb{X}_n^*) \ \leq \ \sqrt{n}\zeta_{n,\theta} \right) - \Pr_{\mathbb{X}_n^* \sim P, \mathbf{W}} \left(\sqrt{n}D(\mathbb{X}_n^*) \ \leq \ \sqrt{n}\zeta_{n,\theta} \right) \\
&= 0\,,
\end{aligned}$$

which completes the proof. ■

A.3.4 Proofs for Proposition 14 and Theorem 3

Proof of Proposition 14 For each n and $R \in \mathcal{P}(\mathbb{R}^d)$, define $Q = (1 - \epsilon_n)P + \epsilon_n R$. Let $\mathbb{X}_n$ and $\mathbb{X}_n^*$ be random samples drawn from Q and P, respectively. As argued in Section A.1, each random variable in $\mathbb{X}_n$ can be written as $\mathbf{X}_i \overset{d}{=} (1 - \xi_i)\mathbf{X}_i^* + \xi_i \mathbf{Z}_i$, where $\mathbf{Z}_i \sim R$ and $\xi_i \sim \text{Bernoulli}(n, \epsilon_n)$ are independent, and $\overset{d}{=}$ denotes equality in distribution. Define the random variable $M' = \sum_{i=1}^n \xi_i$, then $M' \sim \text{Binomial}(n, \epsilon_n)$.

Pick $\delta > 0$. Given any sequence $\epsilon_n = o(n^{-1/2})$, there must exists $f_n = o(n^{1/2})$ such that $\epsilon_n \leq f_n/n$ and $f_n \to \infty$ as $n \to \infty$. Take such f and define the event $\mathcal{B} \coloneqq \{M' - \epsilon_n n \leq f_n\}$. Intuitively, $\mathcal{B}$ is the event where the number of outliers M' does not deviate from its mean $\epsilon_n n$ by more than f_n. Our proof proceeds by first showing that $\mathcal{B}$ occurs with high probability, then proving that the claimed result holds on this event.

To show $\mathcal{B}$ occurs with high probability, we use a concentration inequality for Binomial distributions; see, e.g., Chung and Lu (2002, Lemma 2.1, Eq. 2.2). Applying this result to $\text{Binomial}(n, \epsilon_n)$, we have

$$\Pr_{M'}(\mathcal{B}^{\mathsf{c}}) \ = \ \Pr_{M'}\left(M' - \epsilon_n n > f_n\right) \ \leq \ \exp\left(- \frac{f_n^2}{2(\epsilon_n n + f_n/3)} \right) \ =: \ t_n\,. \qquad (45)$$

Define the event $\mathcal{A}(\mathbb{X}_n) = \{\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n)\}$ and similarly for $\mathcal{A}(\mathbb{X}_n^*)$. Using the above inequality to decompose $\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(\mathcal{A}(\mathbb{X}_n))$ yields

$$\begin{aligned}
&\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(\mathcal{A}(\mathbb{X}_n)) \\
\leq\ & \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(\mathcal{A}(\mathbb{X}_n) \cap \mathcal{B}) + \Pr(\mathcal{B}^{\mathsf{c}}) \\
=\ & \sum_{m' \leq \epsilon_n n + f_n} \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(\mathcal{A}(\mathbb{X}_n) \mid M' = m') \Pr(M' = m') + \Pr(\mathcal{B}^{\mathsf{c}}) \\
=\ & \sum_{m' \leq \epsilon_n n + f_n} \Pr_{\mathbb{X}_{n-m'}^* \sim P,\, \mathbb{Z}_{m'} \sim R, \mathbf{W}}(\mathcal{A}(\mathbb{X}_{n-m'}^* \cup \mathbb{Z}_{m'})) \Pr(M' = m') + \Pr(\mathcal{B}^{\mathsf{c}}) \\
\leq\ & \max_{m' \leq \epsilon_n n + f_n} \Pr_{\mathbb{X}_{n-m'}^* \sim P,\, \mathbb{Z}_{m'} \sim R, \mathbf{W}}(\mathcal{A}(\mathbb{X}_{n-m'}^* \cup \mathbb{Z}_{m'})) \sum_{m' \leq \epsilon_n n + f_n} \Pr(M' = m') + \Pr(\mathcal{B}^{\mathsf{c}}) \\
=\ & \max_{m' \leq \epsilon_n n + f_n} \Pr_{\mathbb{X}_{n-m'}^* \sim P,\, \mathbb{Z}_{m'} \sim R, \mathbf{W}}(\mathcal{A}(\mathbb{X}_{n-m'}^* \cup \mathbb{Z}_{m'})) \Pr(\mathcal{B}) + \Pr(\mathcal{B}^{\mathsf{c}}) \\
\leq\ & \max_{m' \leq \epsilon_n n + f_n} \Pr_{\mathbb{X}_{n-m'}^* \sim P,\, \mathbb{Z}_{m'} \sim R, \mathbf{W}}(\mathcal{A}(\mathbb{X}_{n-m'}^* \cup \mathbb{Z}_{m'})) + t_n ,
\end{aligned}$$

where in the last line we have used (45) and that $\Pr(\mathcal{B}) \leq 1$. This implies

$$\begin{aligned}
&\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(\mathcal{A}(\mathbb{X}_n)) - \Pr_{\mathbb{X}_n^* \sim P}(\mathcal{A}(\mathbb{X}_n^*)) \\
\leq\ & \max_{m' \leq \epsilon_n n + f_n} \big(\Pr_{\mathbb{X}_{n-m'}^* \sim P,\, \mathbb{Z}_{m'} \sim R, \mathbf{W}}(\mathcal{A}(\mathbb{X}_{n-m'}^* \cup \mathbb{Z}_{m'})) - \Pr_{\mathbb{X}_n^* \sim P}(\mathcal{A}(\mathbb{X}_n^*)) \big) + t_n \\
\leq\ & \max_{m' \leq \epsilon_n n + f_n} \sup_{\mathbf{z}_1, \ldots, \mathbf{z}_{m'} \in \mathbb{R}^d} \big| \Pr_{\mathbb{X}_m^* \sim P}(\mathcal{A}(\mathbb{X}_m^* \cup \{\mathbf{z}_i\}_{i=1}^{m'})) - \Pr_{\mathbb{X}_n^* \sim P}(\mathcal{A}(\mathbb{X}_n^*)) \big| + t_n \\
=:\ & \max_{m' \leq \epsilon_n n + f_n} \omega(m') + t_n ,
\end{aligned}$$

where in the second inequality we have taken supremum over all possible values of $\mathbb{Z}_{m'}$, and in the last line we have defined

$$\omega(m') := \sup_{\mathbf{z}_1, \ldots, \mathbf{z}_{m'} \in \mathbb{R}^d} \left| \Pr_{\mathbb{X}_{n-m'}^* \sim P} \left(\mathcal{A}\big(\mathbb{X}_{n-m'}^* \cup \{\mathbf{z}_i\}_{i=1}^{m'}\big) \right) - \Pr_{\mathbb{X}_n^* \sim P} \left(\mathcal{A}(\mathbb{X}_n^*) \right) \right| .$$

Similarly, we have the lower bound

$$\begin{aligned}
&\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(\mathcal{A}(\mathbb{X}_n)) \\
\geq\ & \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(\mathcal{A}(\mathbb{X}_n) \cap \mathcal{B}) \\
=\ & \sum_{m' \leq \epsilon_n n + f_n} \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(\mathcal{A}(\mathbb{X}_n) \mid M' = m') \Pr(M' = m') \\
=\ & \sum_{m' \leq \epsilon_n n + f_n} \Pr_{\mathbb{X}_{n-m'}^* \sim P,\, \mathbb{Z}_{m'} \sim R, \mathbf{W}}(\mathcal{A}(\mathbb{X}_{n-m'}^* \cup \mathbb{Z}_{m'})) \Pr(M' = m') \\
\geq\ & \min_{m' \leq \epsilon_n n + f_n} \Pr_{\mathbb{X}_{n-m'}^* \sim P,\, \mathbb{Z}_{m'} \sim R, \mathbf{W}}(\mathcal{A}(\mathbb{X}_{n-m'}^* \cup \mathbb{Z}_{m'})) \sum_{m' \leq \epsilon_n n + f_n} \Pr(M' = m') \\
=\ & \min_{m' \leq \epsilon_n n + f_n} \Pr_{\mathbb{X}_{n-m'}^* \sim P,\, \mathbb{Z}_{m'} \sim R, \mathbf{W}}(\mathcal{A}(\mathbb{X}_{n-m'}^* \cup \mathbb{Z}_{m'})) \Pr(\mathcal{B}) ,
\end{aligned}$$

where in the second last line we have taken infimum over the values of $\mathbb{Z}_{m'}$. Taking infimum over $\mathbf{z}_1, \ldots, \mathbf{z}_{m'}$, the last line is lower bounded by

$$\begin{aligned}
&\min_{m' \leq \epsilon_n n + f_n} \Pr_{\mathbb{X}^*_{n-m'} \sim P,\, \mathbb{Z}_{m'} \sim R, \mathbf{W}}(\mathcal{A}(\mathbb{X}^*_{n-m'} \cup \mathbb{Z}_{m'})) \Pr(\mathcal{B}) \\
\geq\ &\min_{m' \leq \epsilon_n n + f_n} \inf_{\mathbf{z}_1, \ldots, \mathbf{z}_{m'}} \Pr_{\mathbb{X}^*_{n-m'} \sim P}(\mathcal{A}(\mathbb{X}^*_{n-m'} \cup \{\mathbf{z}_i\}_{i=1}^{m'})) \Pr(\mathcal{B}) \\
\geq\ &\min_{m' \leq \epsilon_n n + f_n} \inf_{\mathbf{z}_1, \ldots, \mathbf{z}_{m'}} \Pr_{\mathbb{X}^*_{n-m'} \sim P}(\mathcal{A}(\mathbb{X}^*_{n-m'} \cup \{\mathbf{z}_i\}_{i=1}^{m'})) \times (1 - t_n) \\
=\ &\min_{m' \leq \epsilon_n n + f_n} \inf_{\mathbf{z}_1, \ldots, \mathbf{z}_{m'}} \Big(\Pr_{\mathbb{X}^*_{n-m'} \sim P}(\mathcal{A}(\mathbb{X}^*_{n-m'} \cup \{\mathbf{z}_i\}_{i=1}^{m'})) \\
&\qquad\qquad - t_n \Pr_{\mathbb{X}^*_{n-m'} \sim P}(\mathcal{A}(\mathbb{X}^*_{n-m'} \cup \{\mathbf{z}_i\}_{i=1}^{m'})) \Big) \\
\geq\ &\min_{m' \leq \epsilon_n n + f_n} \inf_{\mathbf{z}_1, \ldots, \mathbf{z}_{m'}} \Pr_{\mathbb{X}^*_{n-m'} \sim P}(\mathcal{A}(\mathbb{X}^*_{n-m'} \cup \{\mathbf{z}_i\}_{i=1}^{m'})) - t_n \ , \qquad (46)
\end{aligned}$$

where the second inequality holds since (45) implies $\Pr(\mathcal{B}) \geq 1 - t_n$, and the last line holds as a probability is always smaller than or equal to 1. This implies

$$\begin{aligned}
&\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(\mathcal{A}(\mathbb{X}_n)) - \Pr_{\mathbb{X}^*_n \sim P}(\mathcal{A}(\mathbb{X}^*_n)) \\
\geq\ &\min_{m' \leq \epsilon_n n + f_n} \inf_{\mathbf{z}_1, \ldots, \mathbf{z}_{m'}} \left(\Pr_{\mathbb{X}^*_{n-m'} \sim P} \left(\mathcal{A}\big(\mathbb{X}^*_{n-m'} \cup \{\mathbf{z}_i\}_{i=1}^{m'}\big) \right) - \Pr_{\mathbb{X}^*_n \sim P} \left(\mathcal{A}(\mathbb{X}^*_n) \right) \right) - t_n \\
\geq\ &- \max_{m' \leq \epsilon_n n + f_n} \sup_{\mathbf{z}_1, \ldots, \mathbf{z}_{m'} \in \mathbb{R}^d} \left| \Pr_{\mathbb{X}^*_{n-m'} \sim P} \left(\mathcal{A}\big(\mathbb{X}^*_{n-m'} \cup \{\mathbf{z}_i\}_{i=1}^{m'}\big) \right) - \Pr_{\mathbb{X}^*_n \sim P} \left(\mathcal{A}(\mathbb{X}^*_n) \right) \right| - t_n \\
=\ &- \max_{m' \leq \epsilon_n n + f_n} \omega(m') - t_n \ .
\end{aligned}$$

We have therefore shown that

$$\left| \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(\mathcal{A}(\mathbb{X}_n)) - \Pr_{\mathbb{X}^*_n \sim P}(\mathcal{A}(\mathbb{X}^*_n)) \right| \ \leq\ \max_{m' \leq \epsilon_n n + f_n} \omega(m') + t_n \ . \qquad (47)$$

It remains to bound the two terms on the RHS of the above inequality. The second term can be bounded by noting that, since $\epsilon_n \leq f_n / n$ and $f_n = o(n^{1/2})$, it is clear that t_n, defined in (45), converges to 0 as $n \to \infty$, so $t_n \leq \delta/2$ for sufficiently large n. To bound the first term, since $\epsilon_n n + f_n \leq 2 f_n = o(n^{1/2})$ and the assumptions in Lemma 13 hold, we can apply Lemma 13 to conclude that there exists n_0 such that for any $n \geq n_0$, we have

$$\max_{m' \leq \epsilon_n n + f_n} \omega(m') \ <\ \frac{\delta}{2} \ .$$

Using these arguments to bound (47), we have shown that, for sufficiently large n,

$$\left| \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}(\mathcal{A}(\mathbb{X}_n)) - \Pr_{\mathbb{X}^*_n \sim P}(\mathcal{A}(\mathbb{X}^*_n)) \right| \ \leq\ \frac{\delta}{2} + \frac{\delta}{2} \ =\ \delta \ .$$

Taking the supremum over $Q \in \mathcal{P}(P; \epsilon_n)$ then implies, for sufficiently large n,

$$\sup_{Q \in \mathcal{P}(P; \epsilon_n)} \left| \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}} \left(\Delta_\theta(\mathbb{X}_n) > q_{\infty, 1-\alpha}(\mathbb{X}_n) \right) - \Pr_{\mathbb{X}^*_n \sim P, \mathbf{W}} \left(\Delta_\theta(\mathbb{X}^*_n) > q_{\infty, 1-\alpha}(\mathbb{X}^*_n) \right) \right| \ \leq\ \delta \ .$$

Since $\delta > 0$ was arbitrary, this shows the claim. ■

Proof of Theorem 3 This result immediately follows from Proposition 14 by setting $\theta = 0$, in which case

$$\begin{aligned}
&\sup_{Q \in \mathcal{P}(P;\epsilon_n)} \left| \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}} \left(D^2(\mathbb{X}_n) > q^2_{\infty,1-\alpha}(\mathbb{X}_n) \right) - \Pr_{\mathbb{X}^*_n \sim P, \mathbf{W}} \left(D^2(\mathbb{X}^*_n) > q^2_{\infty,1-\alpha}(\mathbb{X}^*_n) \right) \right| \\
&= \sup_{Q \in \mathcal{P}(P;\epsilon_n)} \left| \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}} \left(D(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n) \right) - \Pr_{\mathbb{X}^*_n \sim P, \mathbf{W}} \left(D(\mathbb{X}^*_n) > q_{\infty,1-\alpha}(\mathbb{X}^*_n) \right) \right| \\
&= \sup_{Q \in \mathcal{P}(P;\epsilon_n)} \left| \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}} \left(\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n) \right) - \Pr_{\mathbb{X}^*_n \sim P, \mathbf{W}} \left(\Delta_\theta(\mathbb{X}^*_n) > q_{\infty,1-\alpha}(\mathbb{X}^*_n) \right) \right| \\
&\to 0 ,
\end{aligned}$$

where the second line holds by taking the square-root on both sides of the inequalities, and the last equality holds since $D(\mathbb{X}_n) = \Delta_\theta(\mathbb{X}_n)$ when $\theta = 0$. ■

A.4 P-Targeted Kernel Stein Discrepancy

This section states preliminary results which will be needed to prove Theorem 4. The main tool we use is the following generalized definition of KSD, which was originally proposed in Shi and Mackey (2024, Definition 4).

Definition 3 (P-KSD). *Given $P \in \mathcal{P}(\mathbb{R}^d)$ with a Lebesgue density $p \in \mathcal{C}^1$ and a reproducing kernel $k \in \mathcal{C}_b^{(1,1)}$, the* P-targeted (Langevin) kernel Stein discrepancy *(P-KSD) between two probability measures $Q, R \in \mathcal{P}(\mathbb{R}^d)$ is defined as*

$$\mathbb{S}_P(Q, R) \coloneqq \sup_{f \in \mathcal{B}} \left| \mathbb{E}_{\mathbf{X} \sim Q}[(\mathcal{A}_P f)(\mathbf{X})] - \mathbb{E}_{\mathbf{Y} \sim R}[(\mathcal{A}_P f)(\mathbf{Y})] \right| , \tag{48}$$

where $\mathcal{B} \coloneqq \left\{ h = (h_1, \ldots, h_d) : \ h_j \in \mathcal{H}_k \text{ and } \sum_{j=1}^d \|h_j\|^2_{\mathcal{H}_k} \leq 1 \right\}$ is the unit ball in the d-times Cartesian product $\mathcal{H}_k^d$ of the RKHS $\mathcal{H}_k$ associated with k.

When any input probability measure, say Q, is an empirical measure based on a sample $\mathbb{X}_n = \{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$, we will abuse the notation by writing $\mathbb{S}_P(Q, R) = \mathbb{S}_P(\mathbb{X}_n, R)$ to emphasize the dependence on $\mathbb{X}_n$.

It can be shown that P-KSD is a *Maximum Mean Discrepancy* (MMD, Fortet and Mourier, 1953; Müller, 1997) with the Stein reproducing kernel u_p (Barp et al., 2024, Theorem 1). When either of the two arguments coincides with P, the P-KSD reduces to the standard KSD, as we will show in the next section. The main benefit of working with P-KSD rather than KSD is that P-KSD satisfies both symmetry and a triangle inequality (Shi and Mackey, 2024, Lemma 1), i.e., $\mathbb{S}_P(Q, R) = \mathbb{S}_P(R, Q)$ and $\mathbb{S}_P(Q, R) \leq \mathbb{S}_P(Q, R') + \mathbb{S}_P(R', R)$, for any $Q, R, R' \in \mathcal{P}(\mathbb{R}^d)$ for which they are defined. These properties allow us to show several lemmas, which will be crucial in proving the validity of our robust-KSD tests in later sections.

- Lemma 15 shows that the standard KSD can be written as a P-KSD.
- Lemma 16 shows that $\mathbb{S}_P(Q, R)$ is equivalent to an MMD, and that it admits a closed form involving expectations over Q and R.
- Lemma 17 shows that the P-KSD-projection of a probability measure onto a (standard) KSD-ball centered at P admits a closed-form expression.

- Lemma 18 restates Tolstikhin et al. (2017, Proposition A.1), which is a McDiarmid-type inequality for MMD. It also applies to P-KSD, as an P-KSD is also an MMD. We will use it in Section D to derive a robust-KSD test that is well-calibrated for finite samples. Other deviation bounds instead of the McDiarmid bound could also be used to construct a similar test;see Remark 10 for a discussion.

The MMD has already been studied extensively in the context of both Bayesian and frequentist robust parameter estimation; see, e.g., Briol et al. (2019); Chérief-Abdellatif and Alquier (2020, 2022); Dellaporta et al. (2022); Dellaporta and Damoulas (2023); Alquier and Gerber (2024). A key assumption to ensure robustness in all of these papers is that the kernel is assumed to be bounded, and this is exactly what we are able to achieve with the KSD and P-KSD through Lemma 2.

A.4.1 Properties of P-KSD

Lemma 15 (KSD as P-KSD)**.** *Assume $k \in \mathcal{C}_b^{(1,1)}$ and $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$. For any $Q \in \mathcal{P}(\mathbb{R}^d)$ such that $D(Q,P) < \infty$, we have $\mathbb{S}_P(Q,P) = \mathbb{S}_P(P,Q) = D(Q,P)$.*

Proof of Lemma 15 This follows directly from the symmetry of P-KSD (Shi and Mackey, 2024, Lemma 1) and that the assumed conditions imply $\mathbb{E}_{\mathbf{X}\sim P}[(\mathcal{A}_P f)(\mathbf{X})] = 0$ for all $f \in \mathcal{B}$ (Gorham and Mackey, 2017, Proposition 1), so one of the two expectations in (48) vanishes when either of the two arguments coincides with P. ■

Barp et al. (2024, Theorem 1) shows that a KSD is equivalent to an MMD with the Stein kernel u_p. Since P-KSD is a KSD by Lemma 15, we can conclude that the P-KSD is also an MMD with kernel u_p, and that P-KSD has a double-expectation form similar to MMD.

Lemma 16 (P-KSD closed-form)**.** *Let $Q, R \in \mathcal{P}(\mathbb{R}^d)$ and $k \in \mathcal{C}_b^{(1,1)}$. Denote by $\mathcal{H}_u$ the RKHS associated with the Stein kernel $\mathcal{H}_k$. If $\mathbb{E}_{\mathbf{X}\sim Q}[u_p(\mathbf{X},\mathbf{X})] < \infty$ and $\mathbb{E}_{\mathbf{Y}\sim R}[u_p(\mathbf{Y},\mathbf{Y})] < \infty$, then the following two identities hold*

$$\begin{aligned}\mathbb{S}_P^2(Q,R) &= \left\|\mathbb{E}_{\mathbf{X}\sim Q}[u_p(\cdot,\mathbf{X})] - \mathbb{E}_{\mathbf{Y}\sim R}[u_p(\cdot,\mathbf{Y})]\right\|_{\mathcal{H}_u}^2 \\ &= \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[u_p(\mathbf{X},\mathbf{X}')] + \mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim R}[u_p(\mathbf{Y},\mathbf{Y}')] - 2\mathbb{E}_{\mathbf{X}\sim Q,\mathbf{Y}\sim R}[u_p(\mathbf{X},\mathbf{Y})] .\end{aligned}$$

Proof of Lemma 16 The assumed conditions ensure $\mathbb{S}_P^2(Q,R)$ is well-defined. Since P-KSD is an MMD with reproducing kernel u_p, we can apply Gretton et al. (2012, Lemma 4) to conclude the first equality, and Gretton et al. (2012, Lemma 6) to yield the second. ■

Lemma 17 (P-KSD projection)**.** *Let $R \in \mathcal{P}(\mathbb{R}^d)$ and assume $\mathbb{E}_{\mathbf{Y}\sim R}[u_p(\mathbf{Y},\mathbf{Y})] < \infty$. Further assume the conditions in Lemma 15 hold. For any $\theta > 0$,*

$$\inf_{Q\in\mathcal{B}^{\mathrm{KSD}}(P;\theta)} \mathbb{S}_P(R,Q) \;=\; \max\big(0,\ \mathbb{S}_P(R,P) - \theta\big) \;=\; \max\big(0,\ D(R,P) - \theta\big) .$$

Proof of Lemma 17 The P-KSD $\mathbb{S}_P(R,P)$ is well-defined under the assumed conditions, and hence so is $\mathbb{S}_P\big(\lambda R + (1-\lambda)P, P\big)$ for any $\lambda \in [0,1]$. Also since P-KSD is an MMD with (potentially unbounded) reproducing kernel u_p, the claim then follows by Sun and Zou (2023, Proposition 3) and noting that their proof extends directly to potentially unbounded kernels. ■

Lemma 18 (P-KSD deviation bound)**.** *Assume k is a tilted kernel satisfying the conditions in Lemma 2 and $P \in \mathcal{P}(\mathbb{R}^d)$ has a density $p \in \mathcal{C}^1$. For any random sample $\mathbb{X}_n$ drawn from $Q \in \mathcal{P}(\mathbb{R}^d)$ and any $\alpha > 0$,*

$$\mathrm{Pr}_{\mathbb{X}_n \sim Q}\left(\mathbb{S}_P(\mathbb{X}_n, Q) > \sqrt{\frac{\tau_\infty}{n}} + \sqrt{\frac{-2\tau_\infty \log \alpha}{n}}\right) \;\leq\; \alpha \; .$$

Proof of Lemma 18 Let Q_n denote the empirical distribution based on $\mathbb{X}_n$. By the first equality in Lemma 16, we can write $\mathbb{S}_P(\mathbb{X}_n, Q) = \|\mathbb{E}_{\mathbf{X}\sim Q_n}[u_p(\cdot, \mathbf{X})] - \mathbb{E}_{\mathbf{X}\sim Q}[u_p(\cdot, \mathbf{X})]\|_{\mathcal{H}_u}$. Under the assumptions in Lemma 2, the function $u_p(\cdot, \mathbf{x}) : \mathbb{R}^d \to \mathcal{H}_u$ is continuous for all $\mathbf{x} \in \mathbb{R}^d$. Moreover, $\|u_p(\cdot, \mathbf{x})\|_{\mathcal{H}_u} = u_p(\mathbf{x}, \mathbf{x})$ by the reproducing property of the Stein kernel u_p (see also the argument before (30)), and $\sup_{\mathbf{x}\in\mathbb{R}^d} u_p(\mathbf{x}, \mathbf{x}) < \infty$ again by Lemma 2. We can hence apply the McDiarmid-type inequality for MMD in Tolstikhin et al. (2017, Proposition A.1) to conclude the claimed result. ■

A.5 Validity of the Bootstrap Approach

We provide a proof for the validity of the bootstrap approach in Theorem 4. We first discuss the intuition, and then present two lemmas in Section A.5.1. The proof of Theorem 4 will be presented in Section A.5.2.

The bootstrap procedure used in robust-KSD is the same as the one in the standard KSD test, and it is not immediately obvious that this gives a valid decision threshold. For the standard KSD test, the V-statistic $D^2(\mathbb{X}_n)$ is *degenerate* under the point null $H_0 : Q = P$, i.e., $\mathbb{E}_{\mathbf{X}\sim Q}[u_p(\cdot, P)] = 0$ when $Q = P$ (Liu et al., 2016, Theorem 4.1), and classic bootstrapping methods for degenerate V-statistics (Arcones and Gine, 1992; Huskova and Janssen, 1993) show that the bootstrap quantile $q_{\infty,1-\alpha}$ is a valid decision threshold. However, the same argument cannot be applied directly to our robust test, because our null H_0^{C} is a composite one that contains not only P but also other distributions $Q \neq P$, in which case $D^2(\mathbb{X}_n)$ is no longer degenerate as shown by Liu et al. (2016, Theorem 4.1). Nevertheless, our proof shows that $q_{\infty,1-\alpha}$ is still a correct decision threshold.

Intuition of the proof. Let $\mathbb{X}_n = \{\mathbf{X}_i\}_{i=1}^n$ be a random sample from some $Q \in \mathcal{P}(\mathbb{R}^d)$. We will first show that the rejection probability of the robust-KSD test can be bounded as

$$\mathrm{Pr}_{\mathbb{X}_n\sim Q}\left(\Delta_\theta(\mathbb{X}_n) \;>\; q_{\infty,1-\alpha}(\mathbb{X}_n)\right) \;\leq\; \mathrm{Pr}_{\mathbb{X}_n\sim Q}\left(\mathbb{S}_P(\mathbb{X}_n, Q) \;>\; q_{\infty,1-\alpha}(\mathbb{X}_n)\right) \,,$$

where $\mathbb{S}_P(\mathbb{X}_n, Q)$ is the P-KSD introduced in Section A.4. We will then prove that the RHS of the above inequality converges to the prescribed test level α as $n \to \infty$ by showing that $q_{\infty,1-\alpha}(\mathbb{X}_n)$ is a valid bootstrap approximation for the $(1-\alpha)$-quantile of the distribution of $\mathbb{S}_P(\mathbb{X}_n, Q)$.

To see this, we first use the closed-form expression for P-KSD (Lemma 16) to write

$$\begin{aligned}\mathbb{S}_P^2(\mathbb{X}_n, Q) &= \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q_n}[u_p(\mathbf{X},\mathbf{X}')] - 2\mathbb{E}_{\mathbf{X}\sim Q_n,\mathbf{Y}\sim Q}[u_p(\mathbf{X},\mathbf{Y})] + \mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim Q}[u_p(\mathbf{Y},\mathbf{Y}')] \\ &= \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q_n}[u_p(\mathbf{X},\mathbf{X}';Q)] \\ &= \frac{1}{n^2}\sum_{1\le i,j\le n} u_p(\mathbf{X}_i,\mathbf{X}_j;Q)\ , \end{aligned} \tag{49}$$

where we have defined

$$u_p(\mathbf{x},\mathbf{x}';Q) \ := \ u_p(\mathbf{x},\mathbf{x}') - \mathbb{E}_{\mathbf{Y}\sim Q}[u_p(\mathbf{x},\mathbf{Y})] - \mathbb{E}_{\mathbf{Y}'\sim Q}[u_p(\mathbf{Y}',\mathbf{x}')] + \mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim Q}[u_p(\mathbf{Y},\mathbf{Y}')]\ .$$

In other words, $\mathbb{S}_P^2(\mathbb{X}_n, Q)$ is a V-statistic with symmetric function $u_p(\cdot,\cdot;Q)$.

Our key observation is that $\mathbb{S}_P^2(\mathbb{X}_n, Q)$ is the second-order term in the *Hoeffding's decomposition* (Arcones and Gine, 1992, Eq. 3.2) of the V-statistic estimate $D^2(\mathbb{X}_n, P)$ for the standard KSD defined in (2). This suggests that (3) is an appropriate bootstrap sample. Crucially, this argument does *not* require $Q = P$, unlike in the proof of the standard KSD test. We formalize this argument in the next section.

A naive alternative approach to bootstrap $\mathbb{S}_P^2(\mathbb{X}_n, Q)$ is to note that it is degenerate for any Q and use standard bootstrapping techniques for degenerate V-statistics (e.g., Arcones and Gine, 1992, Theorem 3.5). However, this approach is not applicable here since, to compute the bootstrap sample, it requires the function $u_p(\mathbf{x},\mathbf{x}';Q)$ to be computable at any $\mathbf{x},\mathbf{x}'$, which is not the case since $u_p(\mathbf{x},\mathbf{x}';Q)$ involves intractable expectations over Q (recall that u_p has mean zero under P, but not necessarily under Q).

A.5.1 Bootstrapping the P-KSD Estimate

Our proof for Theorem 4 relies on Arcones and Gine (1992, Lemma 3.4), which states that the asymptotic distribution of the second-order term in the Hoeffding's decomposition of a V-statistic can be approximated by a bootstrap distribution. We restate this result for the case of $\mathbb{S}_P^2(\mathbb{X}_n, Q)$ in Lemma 19.

Lemma 19 (Arcones and Gine (1992), Lemma 3.4)**.** *Let $\mathbb{X}_\infty = \{\mathbf{X}_i\}_{i=1}^\infty$ be a random sample where $\mathbf{X}_i \sim Q$ are independent, and for any n define $\mathbb{X}_n := \{\mathbf{X}_i\}_{i=1}^n$. Let Q_n be the empirical measure based on $\mathbb{X}_n$ and define $\mathbb{X}_n^* = \{\mathbf{X}_i^*\}_{i=1}^n$, where $\mathbf{X}_i^* \sim Q_n$ are independent conditionally on $\mathbb{X}_n$. Assume $\mathbb{E}_{\mathbf{X}\sim Q}[u_p(\mathbf{X},\mathbf{X})^2] < \infty$. Then, as $n \to \infty$, the following holds almost surely,*

$$\sup_{t\in\mathbb{R}} \left| \Pr_{\mathbf{X}_n^*\sim Q_n}\left(n\cdot \mathbb{S}_P^2(\mathbb{X}_n^*,\ \mathbb{X}_n)\ \le\ t \mid \mathbb{X}_\infty\right) - \Pr_{\mathbb{X}_n\sim Q}\left(n\cdot \mathbb{S}_P^2(\mathbb{X}_n,\ Q)\ \le\ t\right)\right| \ \to\ 0\ .$$

Using this result, we can prove that a valid bootstrap approximation for the distribution of $\mathbb{S}_P(\mathbb{X}_n, Q)$ can be obtained by using the bootstrap sample $D_{\mathbf{W}}(\mathbb{X}_n)$ defined in (3). This is summarized in the next lemma.

Lemma 20. *Assume $\mathbb{E}_{\mathbf{X}\sim Q}[u_p(\mathbf{X},\mathbf{X})^2] < \infty$ and let $\mathbf{W} := (W_1,\ldots,W_n)$ be a random vector distributed as Multinomial$(n; 1/n,\ldots,1/n)$. Under the notation in Lemma 19, the following holds as $n \to \infty$,*

$$\sup_{t\in\mathbb{R}} \left| \Pr_{\mathbf{W}}\left(\sqrt{n}\cdot D_{\mathbf{W}}(\mathbb{X}_n)\ \le\ t \mid \mathbb{X}_\infty\right) - \Pr_{\mathbb{X}_n\sim Q}\left(\sqrt{n}\cdot \mathbb{S}_P(\mathbb{X}_n, Q)\ \le\ t\right)\right| \ \to\ 0\ . \tag{50}$$

Proof of Lemma 20 As argued in (49), the squared P-KSD $\mathbb{S}_P^2(\mathbb{X}_n, Q)$ can be written as a V-statistic with symmetric function $u_p(\cdot,\cdot;Q)$. Moreover, direct computation shows that $\mathbb{E}_{\mathbf{X}\sim Q}[u_p(\cdot,\mathbf{X};Q)] \equiv 0$, so the symmetric function $u_p(\cdot,\cdot;Q)$ is *Q-degenerate of order 1* (Arcones and Gine, 1992, pp. 5). On the other hand, it is well-known that the weighted bootstrap form can be equivalently written as an Efron's resampled bootstrap statistic (Janssen, 1994; Dehling and Mikosch, 1994; Janssen, 1997), i.e.,

$$\begin{aligned}
D^2_{\mathbf{W}}(\mathbb{X}_n) &= \frac{1}{n^2}\sum_{1\le i,j\le n}(W_i-1)(W_j-1)u_p(\mathbf{X}_i,\mathbf{X}_j)\\
&= \frac{1}{n^2}\sum_{1\le i,j\le n} W_iW_ju_p(\mathbf{X}_i,\mathbf{X}_j) - W_iu_p(\mathbf{X}_i,\mathbf{X}_j) - W_ju_p(\mathbf{X}_i,\mathbf{X}_j) + u_p(\mathbf{X}_i,\mathbf{X}_j)\\
&\overset{d}{=} \frac{1}{n^2}\sum_{1\le i,j\le n} u_p(\mathbf{X}_i^*,\mathbf{X}_j^*) - u_p(\mathbf{X}_i^*,\mathbf{X}_j) - u_p(\mathbf{X}_i,\mathbf{X}_j^*) + u_p(\mathbf{X}_i,\mathbf{X}_j)\\
&= \frac{1}{n^2}\sum_{1\le i,j\le n} u_p(\mathbf{X}_i^*,\mathbf{X}_j^*;Q_n)\\
&= \mathbb{S}_P^2(\mathbb{X}_n^*,\mathbb{X}_n)\ ,
\end{aligned}$$

where $\mathbf{X}_i^*$ and Q_n are defined in Lemma 19, the notation $\overset{d}{=}$ denotes equality in distribution, the second last line follows by the definition of $u_p(\cdot,\cdot;Q_n)$, and the last line follows from Lemma 16. Under the assumed moment condition, we can apply Lemma 19 together with the above derivation to conclude that a version of (50) with the *squared* statistics holds, i.e.,

$$\sup_{t\in\mathbb{R}}\left|\Pr_{\mathbf{W}}\left(n\cdot D^2_{\mathbf{W}}(\mathbb{X}_n)\ \le\ t \mid \mathbb{X}_\infty\right) - \Pr_{\mathbb{X}_n\sim Q}\left(n\cdot \mathbb{S}_P^2(\mathbb{X}_n,Q)\ \le\ t\right)\right|\ \to\ 0\ .$$

The claim then follows by noting that the mapping $u \mapsto \sqrt{u}$ is everywhere continuous on $[0,\infty)$ and that weak convergence is preserved by continuous function by the Continuous Mapping Theorem (Van der Vaart, 2000, Theorem 2.3). ■

A.5.2 Proof of Theorem 4

Proof of Theorem 4 We first show that $\mathbb{E}_{\mathbf{X}\sim Q}[u_p(\mathbf{X},\mathbf{X})^2] < \infty$, so in particular $\mathbb{E}_{\mathbf{X}\sim Q}[u_p(\mathbf{X},\mathbf{X})] < \infty$ and we can apply Lemma 17. Defining $\mathbf{s}_{p,w}(x) = w(x)\mathbf{s}_p(x)$, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}\sim Q}[u_p(\mathbf{X},\mathbf{X})^2]\ &\le\ \mathbb{E}_{\mathbf{X}\sim Q}\Bigg[\Big(\|\mathbf{s}_{p,w}(\mathbf{x})\|_2^2h(0) + 2\|\mathbf{s}_{p,w}(\mathbf{x})\|_2\|\nabla w(\mathbf{x})\|_2h(0) + \|\nabla w(\mathbf{x})\|_2^2h(0)\\
&\qquad + w(\mathbf{x})^2\big|\nabla^\top\nabla h(0)\big|\Big)^2\Bigg]\\
&\le\ 4\Big(\mathbb{E}_{\mathbf{X}\sim Q}\big[\|\mathbf{s}_{p,w}(\mathbf{x})\|_2^4\big]h^2(0) + 2\mathbb{E}_{\mathbf{X}\sim Q}\big[\|\mathbf{s}_{p,w}(\mathbf{x})\|_2^2\|\nabla w(\mathbf{x})\|_2^2\big]h^2(0)\\
&\qquad + \mathbb{E}_{\mathbf{X}\sim Q}\big[\|\nabla w(\mathbf{x})\|_2^4\big]h^2(0) + \mathbb{E}_{\mathbf{X}\sim Q}\big[w(\mathbf{x})^4\big]\big|\nabla^\top\nabla h(0)\big|^2\Big)\ ,
\end{aligned}$$

where the second line follows from (29), and the last line holds since $(a+b+c+d)^2 \leq 4(a^2+b^2+c^2+d^2)$ for any $a,b,c,d \in \mathbb{R}$. The RHS of the last inequality is finite under the assumed conditions on k.

We are now ready to prove the first result of our theorem. For any $Q \in \mathcal{B}^{\mathrm{KSD}}(P;\theta) \cap \mathcal{P}(\mathbb{R}^d; w)$, we have

$$\begin{aligned}
&\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\big(\Delta_\theta(\mathbb{X}_n) \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n)\big) \\
&= \ \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\big(\max(0, D(\mathbb{X}_n) - \theta) \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n)\big) \\
&= \ \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\big(\inf_{Q' \in \mathcal{B}^{\mathrm{KSD}}(P;\theta)} \mathbb{S}_P(\mathbb{X}_n, Q') \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n)\big) \\
&\leq \ \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\big(\mathbb{S}_P(\mathbb{X}_n, Q) \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n)\big) \ , \qquad (51)
\end{aligned}$$

where the first equality holds by Lemma 17 and noting that $D(\mathbb{X}_n)$ is equivalent to the KSD between the empirical measure based on $\mathbb{X}_n$ and P, and the last line holds since $Q \in \mathcal{B}^{\mathrm{KSD}}(P;\theta)$. To show the first claim of our theorem, it suffices to show that the RHS of (51) converges to α as $n \to \infty$. Defining the following bootstrapping error

$$\delta_n \ \coloneqq \ \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\big(\mathbb{S}_P(\mathbb{X}_n, Q) \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n)\big) - \Pr_{\mathbf{W}}\big(D_{\mathbf{W}}(\mathbb{X}_n) \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n) \mid \mathbb{X}_\infty\big) \ ,$$

we can write the RHS of (51) as

$$\begin{aligned}
\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\big(\mathbb{S}_P(\mathbb{X}_n, Q) \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n)\big) \ &= \ \Pr_{\mathbf{W}}\big(D_{\mathbf{W}}(\mathbb{X}_n) \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n) \mid \mathbb{X}_\infty\big) + \delta_n \\
&= \ \alpha + \delta_n \ , \qquad (52)
\end{aligned}$$

where the last step holds since $q_{\infty,1-\alpha}(\mathbb{X}_n)$ is the $(1-\alpha)$-quantile of the conditional distribution of $D_{\mathbf{W}}(\mathbb{X}_n)$ given $\mathbb{X}_\infty$. Moreover, since we have shown that $\mathbb{E}_{\mathbf{X} \sim Q}[u_p(\mathbf{X}, \mathbf{X})^2] < \infty$, we can apply Lemma 20 to conclude that $\delta_n \to 0$ as $n \to \infty$. Hence, the probability on the LHS of the above equation converges to α. Combining with (51), we have proved that $\lim_{n\to\infty} \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\big(\Delta_\theta(\mathbb{X}_n) \ > \ q_{\infty,1-\alpha}(\mathbb{X}_n)\big) \leq \alpha$. Taking supremum over all $Q \in \mathcal{B}^{\mathrm{KSD}}(P;\theta) \cap \mathcal{P}(\mathbb{R}^d; w)$ shows the first claim of our theorem.

Now assume $Q \notin \mathcal{B}^{\mathrm{KSD}}(P;\theta)$ and $Q \in \mathcal{P}(\mathbb{R}^d; w)$, so in particular $\theta - D(Q,P) < 0$. We have

$$\begin{aligned}
&\Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\big(\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}(\mathbb{X}_n)\big) \\
&= \ \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\big(D(\mathbb{X}_n) - \theta > q_{\infty,1-\alpha}(\mathbb{X}_n)\big) \qquad (53) \\
&= \ \Pr_{\mathbb{X}_n \sim Q, \mathbf{W}}\Big(\sqrt{n}\big(D^2(\mathbb{X}_n) - D^2(Q,P)\big) > \sqrt{n}(q_{\infty,1-\alpha}(\mathbb{X}_n) + \theta)^2 - \sqrt{n} D^2(Q,P)\Big) \ , \qquad (54)
\end{aligned}$$

where (53) holds since it can be checked that $\Delta_\theta(\mathbb{X}_n) > \gamma$ if and only if $D(\mathbb{X}_n) - \theta > \gamma$ for any $\gamma \geq 0$. The argument in Liu et al. (2016, Theorem 4.1) shows that that $\sqrt{n}(D^2(\mathbb{X}_n) - D^2(Q,P))$ converges weakly to a Gaussian limit assuming $\mathbb{E}_{\mathbf{X} \sim Q}[u_p(\mathbf{X}, \mathbf{X}')^2] < \infty$, which holds since (30) and Jensen's inequality imply

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}, \mathbf{X}' \sim Q}[u_p(\mathbf{X}, \mathbf{X}')^2] \ \leq \ \mathbb{E}_{\mathbf{X}, \mathbf{X}' \sim Q}[u_p(\mathbf{X}, \mathbf{X}) u_p(\mathbf{X}', \mathbf{X}')] \ &= \ \big(\mathbb{E}_{\mathbf{X} \sim Q}[u_p(\mathbf{X}, \mathbf{X})]\big)^2 \\
&\leq \ \mathbb{E}_{\mathbf{X} \sim Q}[u_p(\mathbf{X}, \mathbf{X})^2] \ ,
\end{aligned}$$

which is finite as shown before. On the other hand, the weak convergence of $\sqrt{n}(D^2(\mathbb{X}_n) - D^2(Q,P))$ also implies $q_{\infty,1-\alpha}(\mathbb{X}_n) \to 0$ as $n \to \infty$, so when $D(Q,P) > \theta$,

$$\lim_{n\to\infty} (q_{\infty,1-\alpha}(\mathbb{X}_n) + \theta)^2 - D^2(Q,P) \;=\; \theta^2 - D^2(Q,P) \;<\; 0 \;.$$

Combing these arguments shows that in the probability in (54), the LHS converges to a non-degenerate distribution, while the RHS tends to $-\infty$. This implies (54) converges to 1, thus proving the second claim. ■

Remark 9. *If the bootstrapped quantile $q_{\infty,1-\alpha}(\mathbb{X}_n)$ is replaced by the quantile $q^*_{1-\alpha}(\mathbb{X}_n)$ of the distribution of the P-KSD $\mathbb{S}_P(\mathbb{X}_n, Q)$, then the bootstrap approximation error in* (52) *vanishes, i.e., $\delta_n = 0$ for all n. In that case, we have the following stronger, uniform level control*

$$\limsup_{n\to\infty} \sup_{Q\in\mathcal{B}^{\mathrm{KSD}}(P;\theta)\cap\mathcal{P}(\mathbb{R}^d;w)} \mathrm{Pr}_{\mathbb{X}_n\sim Q,\mathbf{W}}\left(\Delta_\theta(\mathbb{X}_n) \;>\; q^*_{1-\alpha}(\mathbb{X}_n)\right) \;\leq\; \alpha \;.$$

In fact, the inequality holds for any finite n, since the RHS of (52) *equals to α for all n.*

A.6 Connections Between KSD Balls and Contamination Models

This section provides proofs for the results in Section 4.2.

- Section A.6.1 states and proves Lemma 21, an intermediary result which provides a decomposition of the KSD under Huber's contamination models.
- Section A.6.2 proves Proposition 5 using Lemma 21.
- Section A.6.3 proves Proposition 6.
- Section A.6.4 provides an example of using Proposition 6 to bound the KSD between t- and Gaussian distributions.

A.6.1 Statement and Proof of Lemma 21

Lemma 21. *Let $\epsilon \in [0,1]$ and $Q = (1-\epsilon)P + \epsilon R$, for any $R \in \mathcal{P}(\mathbb{R}^d)$ such that $\mathbb{E}_{\mathbf{Y}\sim R}[u_p(\mathbf{Y},\mathbf{Y})^{1/2}] < \infty$. Assume $k \in \mathcal{C}_b^{(1,1)}$ and $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$. Then $D(Q,P) = \epsilon D(R,P)$.*

Proof of Lemma 21 The assumption $\mathbb{E}_{\mathbf{Y}\sim R}[u_p(\mathbf{Y},\mathbf{Y})^{1/2}] < \infty$ implies that $D(R,P)$ is well-defined by Gorham and Mackey (2017, Proposition 2). Moreover, when $k \in \mathcal{C}_b^{(1,1)}$ and $\mathbb{E}_{\mathbf{X}\sim P}[\|\mathbf{s}_p(\mathbf{X})\|_2] < \infty$, Gorham and Mackey (2015, Proposition 1) shows that $\mathbb{E}_{\mathbf{X}\sim P}[u_p(\mathbf{X},\cdot)] = 0$. Using this and the linearity of expectation,

$$\begin{aligned} D^2(Q,P) \;&=\; (1-\epsilon)^2\mathbb{E}_{\mathbf{X},\mathbf{X}'\sim P}[u_p(\mathbf{X},\mathbf{X}')] + 2(1-\epsilon)\epsilon\mathbb{E}_{\mathbf{X}\sim P,\mathbf{Y}\sim R}[u_p(\mathbf{X},\mathbf{Y})] \\ &\quad + \epsilon^2\mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim R}[u_p(\mathbf{Y},\mathbf{Y}')] \\ &=\; (1-\epsilon)^2 D^2(P,P) + \epsilon^2 D^2(R,P) \\ &=\; \epsilon^2 D^2(R,P) \;. \end{aligned}$$

Taking square-root of both sides gives the desired result. ■

A.6.2 Proof of Proposition 5

Proof The proof follows a similar approach as Chérief-Abdellatif and Alquier (2022, Lemma 3.3). Under the assumed kernel conditions, Lemma 2 shows that $\sup_{\mathbf{x},\mathbf{x}' \in \mathbb{R}^d} u_p(\mathbf{x},\mathbf{x}') \leq \tau_\infty < \infty$, which implies $D^2(Q,P) = \mathbb{E}_{\mathbf{X},\mathbf{X}' \sim Q}[u_p(\mathbf{X},\mathbf{X}')] \leq \tau_\infty < \infty$ for any probability measure $Q \in \mathcal{P}(\mathbb{R}^d)$. In particular, $D(Q,P)$ is well-defined for any $Q \in \mathcal{P}(\mathbb{R}^d)$ and the conditions in Lemma 21 are met. For any $Q = (1-\epsilon)P + \epsilon R$ with $Q \in \mathcal{P}(P;\epsilon_0)$, applying Lemma 21 shows that

$$D(Q,P) \; = \; \epsilon D(R,P) \; \leq \; \epsilon_0 D(R,P) \; \leq \; \epsilon_0 \tau_\infty^{\frac{1}{2}} \; . \tag{55}$$

To show that this bound is tight, we can lower bound the KSD as

$$\begin{aligned} \sup_{Q \in \mathcal{P}(P;\epsilon_0)} D(Q,P) \; \geq \; \sup_{\mathbf{z} \in \mathbb{R}^d, \epsilon \in [0,\epsilon_0]} D\big((1-\epsilon)P + \epsilon \delta_{\mathbf{z}}, P\big) \; &= \; \sup_{\mathbf{z} \in \mathbb{R}^d, \epsilon \in [0,\epsilon_0]} \epsilon D(\delta_{\mathbf{z}}, P) \\ &= \; \epsilon_0 \sup_{\mathbf{z} \in \mathbb{R}^d} u_p(\mathbf{z},\mathbf{z})^{\frac{1}{2}} \\ &= \; \epsilon_0 \tau_\infty^{\frac{1}{2}} \; , \end{aligned}$$

where the first inequality holds since $R = \delta_{\mathbf{z}}$ is only one possible type of perturbation, the first equality follows from Lemma 21, the second equality holds because the supremum over ϵ is reached at ϵ_0, and the last step holds as $\sup_{\mathbf{z} \in \mathbb{R}^d} u_p(\mathbf{z},\mathbf{z}) = \tau_\infty$ by definition. Combining with the upper bound (55), we have shown that $\sup_{Q \in \mathcal{P}(P;\epsilon_0)} D(Q,P) = \epsilon_0 \tau_\infty^{1/2}$.

A.6.3 Proof of Proposition 6

Under the assumed conditions on k, Lemma 2 shows that the Stein kernel is bounded by $\tau_\infty = \sup_{\mathbf{x} \in \mathbb{R}^d} u_p(\mathbf{x},\mathbf{x}) < \infty$, so in particular $D(Q,P)$ is well-defined for all $Q \in \mathcal{P}(\mathbb{R}^d)$. Under the assumed integrability condition, we have $\mathbb{E}_{\mathbf{X} \sim P}[u_p(\mathbf{X},\cdot)] = 0$ as argued in the paragraph before (11). We can therefore rewrite the squared KSD as

$$\begin{aligned} D^2(Q,P) \; &= \; \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} u_p(\mathbf{x},\mathbf{x}') q(\mathbf{x}) q(\mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \\ &= \; \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} u_p(\mathbf{x},\mathbf{x}') (q(\mathbf{x}) - p(\mathbf{x}))(q(\mathbf{x}') - p(\mathbf{x}')) \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \; , \end{aligned}$$

Using the assumed bound $|q(\mathbf{x}) - p(\mathbf{x})| \leq \delta(\mathbf{x})$ and Lemma 2, we can bound the RHS of the above line by

$$\begin{aligned} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} u_p(\mathbf{x},\mathbf{x}') \delta(\mathbf{x}) \delta(\mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \; &\leq \; \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} |u_p(\mathbf{x},\mathbf{x}')| \delta(\mathbf{x}) \delta(\mathbf{x}') \, \mathrm{d}\mathbf{x} \, \mathrm{d}\mathbf{x}' \\ &\leq \; \tau_\infty \bigg(\int_{\mathbb{R}^d} \delta(\mathbf{x}) \, \mathrm{d}\mathbf{x} \bigg)^2 \; = \; \tau_\infty \delta_0^2 \; . \end{aligned}$$

Taking the square-root of both sides implies $D(Q,P) \leq \tau_\infty^{1/2} \delta_0$. ■

A.6.4 KSD Balls and Fat Tails

The next proposition serves as an example for how Proposition 6 can be applied to design a robust KSD test when the model is Gaussian but data are drawn from t-distributions that are moment-matched to the model.

Proposition 22. *Let $P = \mathcal{N}(0,1)$ and let $Q_\nu = t_\nu \sqrt{(\nu-2)/\nu}$, where t_ν is the t-distribution with degree-of-freedom (dof) $\nu > 2$. Denote by p and q_ν their probability density functions, and by F_∞ and F_ν their cumulative distribution functions, respectively. Then $p(x)$ and $q_\nu(x)$ have exactly two intersections $a_1 < a_2$ on $(0,\infty)$. If furthermore k satisfies the conditions in Lemma 2, then* $\mathrm{KSD}(Q_\nu, P) \leq 4\tau_\infty^{1/2}(F_\nu(a_1) - F_\infty(a_1) + F_\infty(a_2) - F_\nu(a_2))$.

This result suggests that, to ensure quantitative robustness against the scaled t-distribution with ν degrees-of-freedom, we can choose the uncertainty radius in the robust-KSD test to be $\theta = 4\tau_\infty^{1/2}(F_\nu(a_1) - F_\infty(a_1) + F_\infty(a_2) - F_\nu(a_2))$. The intersection points a_1, a_2 do not have a closed form; instead, we approximate them by numerical solvers, which is trivial for this one-dimensional problem, and the numerical error is negligible as evidenced empirically in Figure 4.

Proof of Proposition 22 For $\nu > 2$, the change-of-variable formula shows that the scaled t-distribution Q_ν has density function $q(x) = Z_\nu q_\nu^*(x)$, where $Z_\nu := \Gamma(\frac{\nu+1}{2}) / \big(\sqrt{\pi(\nu-2)}\Gamma(\frac{\nu}{2})\big)$, $q_\nu^*(x) := (1 + \frac{x^2}{\nu-2})^{-(\nu+1)/2}$, and $\Gamma(\cdot)$ is the Gamma function (see, e.g., Forbes et al., 2011, Chapter 8.1). To show $q_\nu(x)$ intersects with $p(x)$ at exactly two points on $(0,\infty)$, we first show that $q_\nu(x)$ is *strong super-Gaussian* (Palmer et al., 2010), i.e., $x \mapsto \log q_\nu(\sqrt{x})$ is convex on $[0,\infty)$. This is immediate by writing

$$\log q_\nu(\sqrt{x}) \; = \; \log Z_\nu - \frac{\nu+1}{2} \log\left(1 + \frac{x}{\nu-2}\right) \; ,$$

and noting that $u \mapsto \log(1+u)$ is concave on $[0,\infty)$. Palmer et al. (2010, Theorem 2) shows that any symmetric and strongly super-Gaussian density belongs to the class of *density cross inequalities* (DC_+), which are symmetric densities that cross a Gaussian density of equal variance exactly four times on $\mathbb{R}$, and take higher values than that normal density at $x = 0$ and as $x \to \infty$. Since in this case both Q_ν and P have unit variance, we conclude that q_ν and p intersect at exactly four points on $\mathbb{R}$. By symmetry of these densities about $x = 0$, this shows that q_ν and p must intersect at exactly *two* points on $(0,\infty)$.

Call these two intersections a_1, a_2 and assume without loss of generality $0 < a_1 < a_2$. It then follows that $q_\nu(x) \geq p(x)$ on $[0, a_1) \cup [a_2, \infty)$ and $q_\nu(x) \leq p(x)$ on $[a_1, a_2)$. We therefore have

$$\begin{aligned}
\delta_0 \; &:= \; \int_{\mathbb{R}} |p(x) - q_\nu(x)| \, \mathrm{d}x \\
&= \; 2\int_0^{a_1} (q_\nu(x) - p(x)) \, \mathrm{d}x + 2\int_{a_1}^{a_2} (p(x) - q_\nu(x)) \, \mathrm{d}x + 2\int_{a_2}^{\infty} (q_\nu(x) - p(x)) \, \mathrm{d}x \\
&= \; 2(F_\nu(a_1) - F_\infty(a_1)) + 2\big(F_\infty(a_2) - F_\infty(a_1) - (F_\nu(a_2) - F_\nu(a_1))\big) \\
&\quad + 2\big((1 - F_\nu(a_2)) - (1 - F_\infty(a_2))\big) \\
&= \; 2(F_\nu(a_1) - F_\infty(a_1) + F_\infty(a_2) - F_\infty(a_1) - F_\nu(a_2) + F_\nu(a_1) + F_\infty(a_2) - F_\nu(a_2)) \\
&= \; 4(F_\nu(a_1) - F_\infty(a_1) + F_\infty(a_2) - F_\nu(a_2)) \; ,
\end{aligned}$$

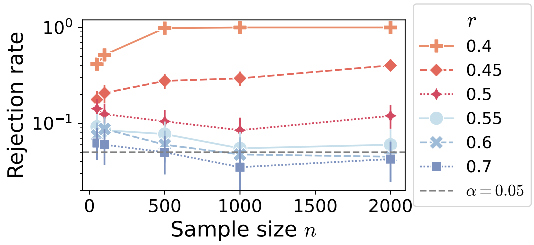


Figure 8: Rejection probability (in log scale) against sample size under the univariate contaminated Gaussian model with outlier $z = 10$. The contamination ratio scales as $\epsilon_n = n^{-r}$ for different values of r.

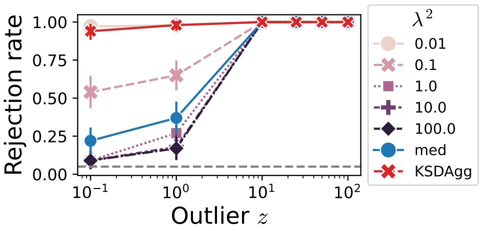


Figure 9: Rejection probability for an IMQ kernel with bandwidth λ. "med" is the median heuristic. "KSDAgg" is the test of Schrab et al. (2022). The dashed line shows $\alpha = 0.05$. All tests all reject H_0 for large z.

where in the second line we have again used the symmetry of normal and t densities. Applying Proposition 6 completes the proof. ■

Appendix B. Complementary Experiments

This section includes additional experimental results. Section B.1 discusses the rate requirement in Theorem 3 and demonstrates that the KSD test is no longer qualitatively robust when this condition is not met. Section B.2 reviews the robust MMD tests included in Section 5.1 and presents implementation details. Section B.4 includes an ablation study on the choice of kernel bandwidths in the robust-KSD test. Section B.6 studies its scalability with dimension. Section B.8 compares robust-KSD using two different bootstrap methods: the weighted bootstrap used throughout this work, and a wild bootstrap that is more prevalent in the kernel testing literature.

B.1 Decay Rate of Contamination Ratio

We show empirically that the rate requirement in the qualitative robustness result Theorem 3 is not an artifact of the proof, i.e., the tilted-KSD test is no longer qualitatively robust to Huber's contamination models with contamination ratio ϵ_n if $\epsilon_n = n^{-r}$ for any $r \leq 1/2$, where n is the sample size. We run the standard KSD test with a tilted kernel under the contaminated Gaussian model as described in Section 5.1 with dimension $d = 1$ and outlier $z = 10$. The contamination ratio is chosen to be $\epsilon_n = n^{-r}$ for different choices of r, and the probability of rejection as n grows is shown in Figure 8. The results are averaged over 400 repetitions instead of 100 repetitions as in Section 5.1 to reduce numerical uncertainty. When $r > 0.5$, the rejection probability converges to the prescribed test level α, which is the limit of the rejection probability *without* contamination. This aligns with Theorem 3. However, this rejection probability no longer converges to α when $r \leq 0.5$, thus showing that the tilted-KSD test is no longer qualitatively robust.

B.2 Implementation Details for the Robust MMD Tests

This section provides implementation details for the two robust MMD-based tests included in Section 5.1. Both tests use the Maximum Mean Discrepancy (MMD; Gretton et al., 2012, Definition 2) between the data-generating distribution Q and the model P as a measure of their disparity. Given a reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, the squared MMD between Q and P are defined as

$$D^2_{\mathrm{MMD}}(Q, P) \ = \ \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim Q}[k(\mathbf{X}, \mathbf{X}')] + \mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim P}[k(\mathbf{Y}, \mathbf{Y}')] - 2\mathbb{E}_{\mathbf{X}\sim Q,\mathbf{Y}\sim P}[k(\mathbf{X}, \mathbf{Y})] \ .$$

Given independent samples $\mathbb{X}_n = \{\mathbf{X}_i\}_{i=1}^n$ from Q and $\mathbb{Y}_m = \{\mathbf{Y}_j\}_{j=1}^m$ from P, the MMD can be estimated directly by $D^2_{\mathrm{MMD}}(Q_n, P_m)$, where Q_n and P_m are the empirical measures formed by $\mathbb{X}_n$ and $\mathbb{Y}_m$, respectively (Gretton et al., 2012). The computational cost of such estimate scales with sample size n as $\mathcal{O}((n + m)^2 + mC_{\mathrm{sim}})$, where C_{sim} is the cost of simulating one datum from P, compared with $\mathcal{O}(n^2 + nC_{\mathrm{score}})$ for KSD, where C_{score} denotes the cost of a single score evaluation. In our experiment in Section 5.1, P is a Gaussian model, so both simulation and score evaluation are trivial and $C_{\mathrm{sim}}, C_{\mathrm{score}}$ are both small. Hence, we set $m = n$ so that the costs of MMD and of KSD are comparable. Moreover, we will use the squared-exponential kernel $k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|_2^2/(2\gamma^2))$ and set γ^2 using the median heuristic. Squared-exponential kernels are popular for MMDs and are also used in the original papers that proposed the two robust MMD tests (Sun and Zou, 2023; Schrab and Kim, 2024).

B.2.1 The MMD-Dev Test

The robust MMD test of Sun and Zou (2023, Eq. 38), which we refer to as *MMD-Dev*, targets the hypotheses

$$H_0^{\mathrm{MMD}} : \ Q \in \mathcal{B}^{\mathrm{MMD}}(P; \theta_{\mathrm{MMD}}) \ , \qquad H_1^{\mathrm{MMD}} : \ Q \notin \mathcal{B}^{\mathrm{MMD}}(P; \theta_{\mathrm{MMD}}) \ , \tag{56}$$

where $\theta_{\mathrm{MMD}} \geq 0$ and $\mathcal{B}^{\mathrm{MMD}}(P; \theta_{\mathrm{MMD}}) = \{Q \in \mathcal{P} : D_{\mathrm{MMD}}(Q, P) \leq \theta_{\mathrm{MMD}}\}$ is the MMD ball centered at P and with radius θ_{MMD}. Given a test level $\alpha \in (0, 1)$, MMD-Dev rejects H_0^{MMD} if $D_{\mathrm{MMD}}(Q_n, P_n) > \theta_{\mathrm{MMD}} + \gamma_n$, where $\gamma_n = \sqrt{2K/n}(1 + \sqrt{-\log \alpha})$. Sun and Zou (2023, Theorem 4) shows that this test controls the Type-I error for any finite n, and is asymptotically optimal against certain alternatives. The decision threshold γ_n is derived using a McDiarmid-type deviation inequality for MMDs due to Gretton et al. (2012, Theorem 8).

In our experiments, we chose $\theta_{\mathrm{MMD}} = \epsilon_0\sqrt{2}$. This ensures that $\mathcal{B}^{\mathrm{MMD}}(P; \theta_{\mathrm{MMD}})$ contains Huber's contamination models with contamination ratios up to ϵ_0, which we show in the next result. Notably, although we can compare the MMD-Dev test with our robust-KSD test on Huber's contamination models, they are in general *not* directly comparable, since they target different null hypotheses.

Lemma 23 (MMD balls and Huber's model). *Let k be a reproducing kernel with $0 < k(\mathbf{x}, \mathbf{x}') \leq K < \infty$, for all $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$. For any $\epsilon_0 \in [0, 1]$, if $\theta_{\mathrm{MMD}} = \epsilon_0\sqrt{2K}$, then $\mathcal{P}(P; \epsilon_0) \subset \mathcal{B}^{\mathrm{MMD}}(P; \theta_{\mathrm{MMD}})$, where $\mathcal{P}(P; \epsilon_0)$ is the Huber's contamination model defined in* (6).

Proof of Lemma 23 For any $Q \in \mathcal{P}(\mathbb{R}^d)$, we define its kernel mean embedding (Berlinet and Thomas-Agnan, 2004, Chapter 4) as $\mu_Q(\cdot) \coloneqq \mathbb{E}_{\mathbf{X}\sim Q}[k(\mathbf{X}, \cdot)]$, which is well-defined since k is bounded. By Sriperumbudur et al. (2010, Theorem 1), the MMD between any $Q, P \in \mathcal{P}(\mathbb{R}^d)$ can be equivalently written as $D_{\mathrm{MMD}}(Q, P) = \|\mu_Q - \mu_P\|_{\mathcal{H}_k}$, where $\mathcal{H}_k$ is the RKHS associated with k. Pick $\epsilon_0 \in [0, 1]$ and $R \in \mathcal{P}$. For any $\epsilon \in [0, \epsilon_0]$,

$$\begin{aligned} D_{\mathrm{MMD}}((1-\epsilon)P + \epsilon R, P) \ = \ \|\mu_{(1-\epsilon)P+\epsilon R} - \mu_P\|_{\mathcal{H}_k} &= \|(1-\epsilon)\mu_P + \epsilon\mu_R - \mu_P\|_{\mathcal{H}_k} \\ &= \epsilon\|\mu_R - \mu_P\|_{\mathcal{H}_k} \\ &= \epsilon D_{\mathrm{MMD}}(R, P) \ , \end{aligned}$$

where the second equality holds due to the linearity of the expectation operator. Moreover,

$$\begin{aligned} D^2_{\mathrm{MMD}}(R, P) &= \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim R}[k(\mathbf{X}, \mathbf{X}')] + \mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim P}[k(\mathbf{Y}, \mathbf{Y}')] - 2\mathbb{E}_{\mathbf{X}\sim R,\mathbf{Y}\sim P}[k(\mathbf{X}, \mathbf{Y})] \\ &\leq \mathbb{E}_{\mathbf{X},\mathbf{X}'\sim R}[k(\mathbf{X}, \mathbf{X}')] + \mathbb{E}_{\mathbf{Y},\mathbf{Y}'\sim P}[k(\mathbf{Y}, \mathbf{Y}')] \\ &\leq 2K \ , \end{aligned}$$

where the last line holds since $\sup_{\mathbf{x},\mathbf{x}'\in\mathbb{R}^d} k(\mathbf{x}, \mathbf{x}') \leq K$ by assumption. This shows that $D_{\mathrm{MMD}}((1-\epsilon)P + \epsilon R, P) \leq \epsilon\sqrt{2K} \leq \epsilon_0\sqrt{2K}$, so the claim holds. ■

B.2.2 The dcMMD Test

The dcMMD test (Schrab and Kim, 2024) targets different hypotheses. Given $\epsilon_0 \in [0, 1]$, the null assumes that the observed sample $\mathbb{X}_n$ is generated by firstly drawing i.i.d. random variables from some probability measure, then corrupting them by replacing at most a proportion of ϵ_0 of the sample by arbitrary values. The hypotheses can be formalized as follows:

$$\begin{aligned} H_0^{\mathrm{dcMMD}} &: \ \text{At least } (1-\epsilon_0) \times n \text{ random variables in } \mathbb{X}_n \text{ are i.i.d. from } P. \\ H_1^{\mathrm{dcMMD}} &: \ \text{Otherwise} \ . \end{aligned}$$

The dcMMD test rejects H_0^{dcMMD} if $D_{\mathrm{MMD}}(Q_n, P_n) > q_\alpha^{\mathrm{MMD}} + 2\epsilon_0\sqrt{2K}$, where q_α^{MMD} is the empirical quantile of B permutation samples $\{T^b_{\mathrm{MMD}}\}_{b=1}^B$, and each T^b_{MMD} is computed by *(i)* randomly permuting $\mathbb{X}_n \cup \mathbb{Y}_n$, *(ii)* partitioning the permuted set into two subsets $\mathbb{X}_n^b$ and $\mathbb{Y}_n^b$ of size n, and *(iii)* computing $T^b_{\mathrm{MMD}} \coloneqq D_{\mathrm{MMD}}(Q_n^b, P_n^b)$, where Q_n^b and P_n^b are the empirical measures based on $\mathbb{X}_n^b$ and $\mathbb{Y}_n^b$, respectively. We use $B = 500$, which is the default setup in Schrab and Kim (2024).

The dcMMD test gives stronger calibration guarantee than Huber's contamination model because it controls contamination proportion no larger than ϵ_0 for *any realization* of the sample $\mathbb{X}_n$, rather than in expectation as required by Huber's model. However, it is not necessarily stronger than KSD-balls or MMD-balls, since it does not account for the case where *all* samples are under mild perturbation.

B.3 Approximating the Supremum of the Stein Kernel

As described in Section 4.2, in our experiments we approximated the intractable supremum $\tau_\infty = \sup_{\mathbf{x}\in\mathbb{R}^d} u_p(\mathbf{x}, \mathbf{x})$ by $\max_{i=1,\ldots,n} u_p(\mathbf{x}_i, \mathbf{x}_i)$. We now study how this approximation

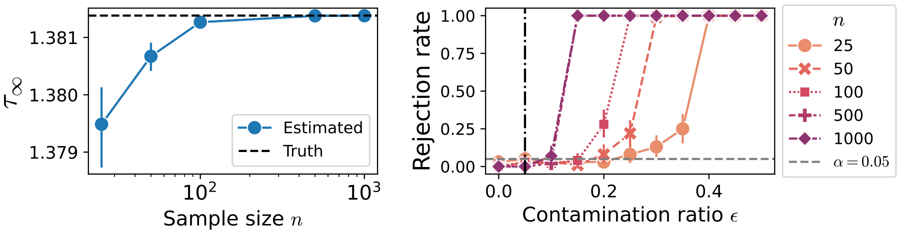


Figure 10: *Left.* Estimated τ_∞ using the trick described in Section 4.2 compared with the ground-truth; the estimate becomes more accurate with larger samples. *Right.* Rejection probability of robust-KSD under an outlier-contaminated Gaussian model; the Type-I error remains calibrated even for small sample sizes.

affects the performance of the robust-KSD test. We run the same contaminated Gaussian experiment in Section 5.1 with $d = 1$ and outlier location $z = 10$. For this model, $\mathbf{x} \mapsto u_p(\mathbf{x}, \mathbf{x})$ is a simple, univariate function, so we can compute the ground-truth τ_∞ accurately by numerical optimization. Clearly, the ground truth τ_∞ is equivalent to taking $n = \infty$. This is evidenced by the left plot in Figure 10, which shows that the finite-sample approximation $\max_{i=1,\ldots,n} u_p(\mathbf{x}_i, \mathbf{x}_i)$ becomes more accurate as n increases. The right plot of Figure 10 shows the rejection probability of robust-KSD with θ set to control at most $\epsilon_0 = 0.05$ proportion of contamination. Remarkably, even when this approximation under-estimates τ_∞ (which is the case for small n), robust-KSD still remains well-calibrated. This is not surprising: the supremum τ_∞ represents the maximal contribution of a single datum $\mathbf{x}$ taking arbitrary values in the *entire* sample space $\mathbb{R}^d$, regardless of whether it is present in the *observed* data set. However, in practice, it suffices to control the contribution of any single datum in the observed data set.

B.4 Ablation Study for Kernel Bandwidth in the Standard KSD Test

We evaluate how the choice of the kernel bandwidth affects the performance of the *standard* KSD test. We run the same contaminated Gaussian experiment as in the previous subsection. The standard KSD test uses an IMQ kernel with bandwidth $\lambda^2 \in \Lambda \cup \{\lambda^2_{\text{med}}\}$, where $\Lambda = \{0.01, 0.1, 1, 10, 100\}$ and λ^2_{med} is the bandwidth chosen by median heuristic. We also include KSDAgg, which also uses an IMQ kernel for a fair comparison. As shown in Figure 9, all standard IMQ-KSD tests reject the point null with high probability for large outlier values, *regardless* of the bandwidth value. This suggests that no fixed bandwidth can ensure robustness, which is consistent with Theorem 1, and is because the Stein kernel is unbounded regardless of the choice of λ, thus the non-robustness issue persists. The KSDAgg test is even more sensitive to contamination than the IMQ-KSD tests. As noted in Section 5.1, this is because KSDAgg is designed to optimally combine multiple bandwidths to boost the test power against all alternatives, including contaminated models.

B.5 Gaussian Mean-Shift Experiment

We run a Gaussian mean-shift experiment to demonstrate the performance of our robust test against model deviations other than contamination. We use a Gaussian model $\mathcal{N}(0, I_d)$ in

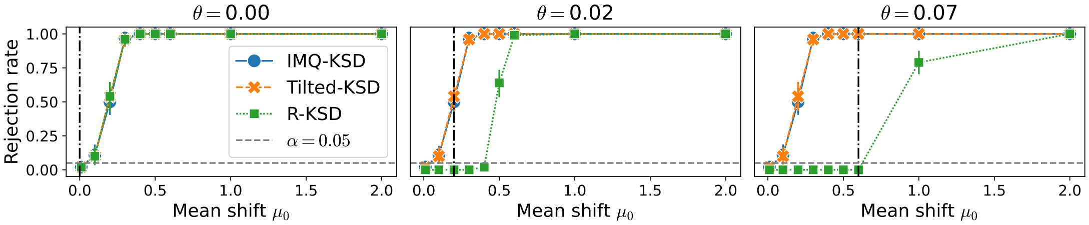


Figure 11: Probability of rejection against the mean-shift μ_0 under a Gaussian model in $d = 50$ dimensions. *Black dash-dot line.* Uncertainty radius θ, which is set to be the KSD value $D(Q_{\mu_0}, P)$ corresponding to different values of μ_0.

dimension $d = 50$, and draw data from $Q_{\mu_0} = \mathcal{N}(\mu_0 e_1, I_d)$, where $e_1 = (1, 0, \ldots, 0)^\top \in \mathbb{R}^d$ and $\mu_0 \in \mathbb{R}$ is a mean-shift. The uncertainty radius θ is chosen to be the Tilted-KSD value corresponding to $\mu_0 = 0, 0.2$ and 0.6, respectively. The purpose of this experiment is to demonstrate that the robust-KSD test is well-calibrated when $D(Q_{\mu_0}, P) \leq \theta$ and consistent when $D(Q_{\mu_0}, P) > \theta$. As shown in Figure 11, the standard tests reject with probability higher than robust-KSD. This is again expected since the standard tests are not robust to contamination. For mean-shift values not greater than the black vertical line, $Q_{\mu_0} \in \mathcal{B}^{\mathrm{KSD}}(P; \theta)$, so we are under the null hypothesis and robust-KSD rejects no more frequently than the level, showing that it is well-calibrated. For larger μ_0, robust-KSD rejects with probability approaching one, thus showing its power. Moreover, when $\mu_0 = 0$ so that $\theta = 0$, robust-KSD becomes identical to the standard Tilted-KSD test, showing that robust-KSD is indeed a generalization of the standard test.

B.6 Scalability with Dimension

We run the Gaussian mean-shift experiment in Section B.5 in different dimensions; other experimental setups remain unchanged. The uncertainty radius for R-KSD is chosen to be the (non-squared) KSD value corresponding to mean-shift $\mu = 0.3$. Results are reported in Figure 12. As dimension increases, both the two standard tests and R-KSD have declining power in correctly rejecting for large values of μ. This shows that both the standard and the robust KSD tests suffer from the *curse-of-dimensionality*, a known issue for kernel-based tests (Huang et al., 2023; Reddi et al., 2015; Ramdas et al., 2015). Unsurprisingly, this issue is more prominent for the R-KSD test. This is because the KSD-ball $\mathcal{B}^{\mathrm{KSD}}(P; \theta)$ could potentially include more distributions in higher dimensions.

B.7 Different Contamination Distributions

In most of our experiments in Section 5, the data-generating distribution takes the form $Q = (1 - \epsilon_0)P + \epsilon_0 R$ with the contamination distribution $R = \delta_{\mathbf{z}}$ being a Dirac delta taking a single value at $\mathbf{z}$. We now investigate how the choice of R affects the empirical performance of R-KSD. Importantly, our results in Section 4 make no assumption on the form of R.

We set $P = \mathcal{N}(0, 1)$ and generate data from $Q = (1 - \epsilon)P + \epsilon\mathcal{N}(10, \sigma^2)$ with varying $\sigma > 0$. Following Proposition 5, we choose $\theta = \epsilon_0 \tau_\infty^{1/2}$, and fix the contamination tolerance to be

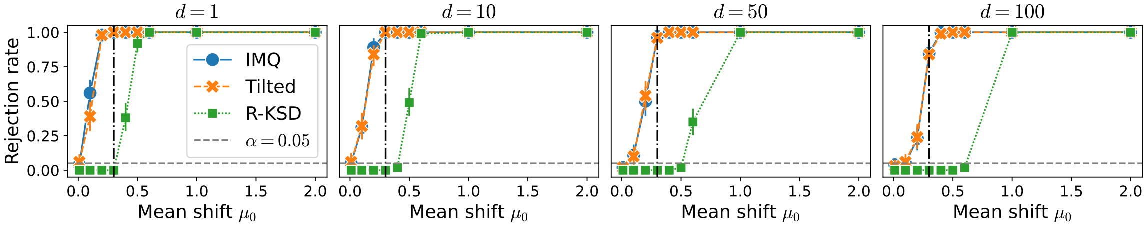


Figure 12: Probability of rejection against the mean shift under a Gaussian model in various dimensions. *Grey dotted line.* Test level $\alpha = 0.05$. *Black dash-dot line.* Uncertainty radius θ, which is set to be the KSD value corresponding to $\mu = 0.3$.

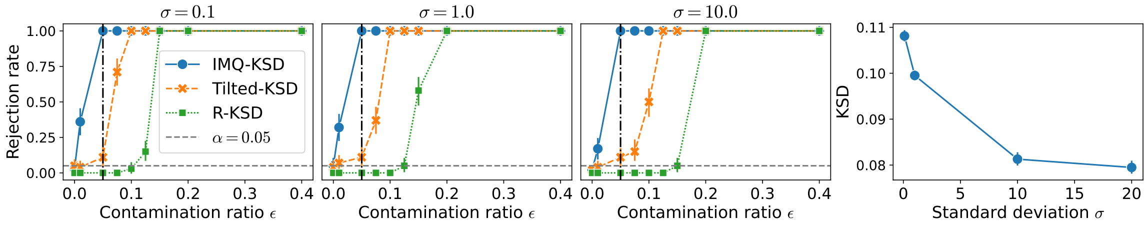


Figure 13: Mixture-of-Gaussian experiment. *Left.* Probability of rejection under different standard deviations for the contamination component. *Right.* KSD estimates. *Grey dotted line.* Test level $\alpha = 0.05$. *Black dash-dot line.* Uncertainty radius $\theta = \epsilon_0 \tau_\infty^{1/2}$, with τ_∞ estimated following Section 4.2.

$\epsilon_0 = 0.05$. Since Proposition 5 holds for all R, we expect R-KSD to remain (asymptotically) valid in this setting.

Results are shown in Figure 13. For larger σ, all tests, including our R-KSD, saw a exhibit reduced test power. This is because, when σ is large, the noise component R becomes more dispersed, making it harder for KSD to detect discrepancies. This can be confirmed by the rightmost plot, which shows that the KSD between Q and P decreases with σ. Crucially, our R-KSD test is able to control the Type-I error regardless of the value of σ.

B.8 Weighted Bootstrap and Wild Bootstrap

We compare the robust-KSD test using the weighted bootstrap described in Section 2.2 (equivalent to the Efron's bootstrap) against the wild bootstrap due to Leucht and Neumann (2013). For the standard KSD test, weighted bootstrap was used in Liu et al. (2016), while the wild bootstrap is more popular in the literature (Chwialkowski et al., 2016; Schrab et al., 2022; Liu et al., 2023). Compared with weighted bootstrap, the wild approach is more flexible as it can work for dependent samples (Chwialkowski et al., 2014), and theoretical guarantees on the power of the standard test were only proved with the wild bootstrap (Schrab et al., 2022). In this work, we have used the weighted approach since we can then leverage existing theoretical results to show the validity of the test; see Section A.5.1. This is for simplicity rather than necessity.

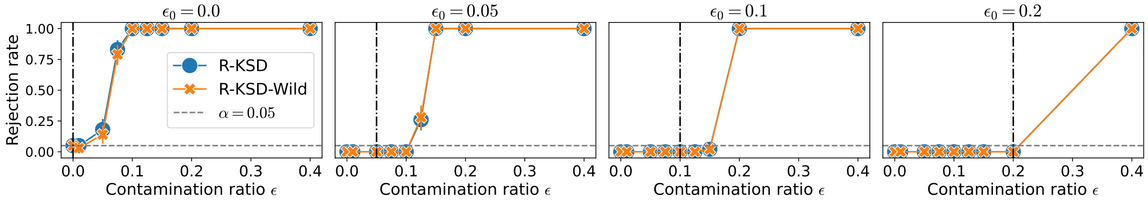


Figure 14: Probability of rejection under an outlier-contaminated Gaussian model using the weighted bootstrap and the wild bootstrap. *Black dash-dot vertical line.* Uncertainty radius $\theta = \epsilon_0 \tau_\infty^{1/2}$, with τ_∞ estimated following Section 4.2.

We numerically compare these two bootstrap methods using the contaminated Gaussian model in Section 5.1 and show that the use of weighted bootstrap does not negatively impact the test. Figure 14 shows the rejection probability of the robust-KSD test using the weighted bootstrap (R-KSD) and wild bootstrap (R-KSD-Wild). The uncertainty radius is chosen to be $\theta = \epsilon_0 \tau_\infty^{1/2}$ for different values of ϵ_0, following Section 5.1. In particular, $\epsilon_0 = 0$ corresponds to setting $\theta = 0$, in which case robust-KSD reduces to the standard KSD. As evident from the plot, these two bootstrap methods produce almost identical results regardless of the value of θ.

Appendix C. Extension to U-statistics

In this work, we have focused primarily on using the V-statistic (2) for estimating the KSD. An alternative estimator commonly used in the literature (Liu et al., 2016; Jitkrittum et al., 2017; Schrab et al., 2022; Liu et al., 2023) is the following U-statistic $\frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} u_p(\mathbf{X}_i, \mathbf{X}_j)$. In this section, we discuss briefly how to extend our key results to this U-statistic estimate.

- **Extending the qualitative (non-)robustness results:** The main ingredients for proving Theorem 1 are *(i)* a deviation bound for the V-statistic $D^2(\mathbb{X}_n)$ as shown in Lemma 7, and *(ii)* a concentration bound for the bootstrapped quantile $q^2_{\infty,1-\alpha}(\mathbb{X}_n)$ as shown in Lemma 8. A U-statistic counterpart of both results can be shown by following the same proof technique and replacing $D^2(\mathbb{X}_n)$ with the U-statistic. For example, with the U-statistic, a decomposition similar to (10) can be derived by summing over only the non-diagonal terms $i \neq j$ and replacing normalizing factors of the form $1/n^2$ to $1/(n(n-1))$. Similarly, the proof of Theorem 3 can be extended to U-statistics by following the same steps in Section A.3.

- **Extending the robust-KSD test:** To adapt the robust-KSD test to the U-statistic, we first note that rejecting the null if $\Delta_\theta(\mathbb{X}_n) > q_{\infty,1-\alpha}$ is equivalent to rejecting the null if $D^2(\mathbb{X}_n) > (q_{\infty,1-\alpha} + \theta)^2$, as argued in (53) and the paragraph thereafter. A natural approach is therefore to replace the V-statistic $D^2(\mathbb{X}_n)$ by the U-statistic, and the bootstrap quantile $q_{\infty,1-\alpha}$ by the square-root of the bootstrap quantile formed with the U-statistic. However, since the bootstrap samples (3) based on U-statistics can take *negative* values, the bootstrap quantile can also take negative values, rendering its square-root $q_{\infty,1-\alpha}$ undefined. One solution is to never reject the null when this happens. This has little impact when $D^2(Q, P)$ is large, since the bootstrap samples are then likely to take positive values. However, when $D^2(Q, P) \approx 0$, the bootstrap samples are more prone to taking negative values, thus making the test conservative.

Appendix D. A Non-Asymptotically Valid Robust KSD Test

The robust-KSD test introduced in Section 4 is only well-calibrated when $n \to \infty$. In this appendix, we now derive a robust test that is well-calibrated with finite samples. An immediate consequence is that a stronger, uniform Type-I error control can be achieved.

The test rejects the null H_0^{C} in (8) if

$$\Delta_\theta(\mathbb{X}_n) \ > \ \gamma_n \ ,$$

where $\gamma_n = \sqrt{\tau_\infty / n} + \sqrt{-2\tau_\infty(\log \alpha)/n}$, the constant τ_∞ is defined in Lemma 2, and $\Delta_\theta(\mathbb{X}_n)$ is defined in (9). We call this test *robust-KSD-Dev* (R-KSD-Dev).

The decision threshold γ_n is based on the concentration bound Lemma 18, which is a deviation bound of the McDiarmid's type (McDiarmid et al., 1989). This test can be viewed as a counterpart of the robust MMD test of Sun and Zou (2023), which is also constructed using McDiarmid's inequality. The next result, proved at the end of this section, shows its finite-sample validity under H_0^{C} as well as its consistency under H_1^{C}.

Theorem 24. *Let $\mathbb{X}_\infty = \{\mathbf{X}_i\}_{i=1}^\infty$ be a sequence of independent random variables following Q. Suppose k is a tilted kernel satisfying the conditions in Lemma 2, and let $\theta \geq 0$. Then*

1. *Under H_0^{C}, for any n, it holds that $\sup_{Q \in \mathcal{B}^{\mathrm{KSD}}(P;\theta)} \Pr_{\mathbb{X}_n \sim Q}\big(\Delta_\theta(\mathbb{X}_n) > \gamma_n\big) \ \leq \ \alpha$.*
2. *Under H_1^{C}, it holds that $\Pr_{\mathbb{X}_n \sim Q}\big(\Delta_\theta(\mathbb{X}_n) > \gamma_n\big) \to 1$, as $n \to \infty$.*

Remark 10 (Alternative deviation bounds). *Alternative deviation bounds to Lemma 18 can also be used to construct similar tests. More precisely, Theorem 24 holds for any threshold γ_n that satisfies*

$$\Pr_{\mathbb{X}_n \sim Q}(\mathbb{S}_P(\mathbb{X}_n, Q) > \gamma_n) \ \leq \ \alpha \ ,$$

where $\mathbb{S}_P(\mathbb{X}_n, Q) = \big\| \mathbb{E}_{\mathbf{X} \sim Q_n}[u_p(\cdot, \mathbf{X})] - \mathbb{E}_{\mathbf{X} \sim Q}[u_p(\cdot, \mathbf{Y})] \big\|_{\mathcal{H}_u}$ is a Hilbert-space norm, as shown in Lemma 16. Thus, any deviation bound for Hilbert-space norms may be applied. Examples include another McDiarmid's bound of Gretton et al. (2012, Theorem 7), the empirical Bernstein bounds of Wolfer and Alquier (2022, Theorem A.1) and Martinez-Taboada and Ramdas (2024, Corollary 1), as well as the Hilbert-space valued Hoeffding bound of Pinelis (1994, Theorem 3.5), particularly its i.i.d. variant (Chatalic et al., 2022, Lemma E.1).

Remark 11 (Justification for McDiarmid-type bound). *We opted for the McDiarmid-type bound in Lemma 18 because, in our setting, it is tighter than the bounds from Gretton et al. (2012), Wolfer and Alquier (2022) and Chatalic et al. (2022). In particular, while Wolfer and Alquier (2022) claim their empirical Bernstein bound outperforms McDiarmid's inequality, we find that this only holds for their bound tailored to translation-invariant kernels (Wolfer and Alquier, 2022, Theorem 3.1), but not for the non-translation-invariant version (Wolfer and Alquier, 2022, Theorem A.1), which our setting requires since we have assumed Lemma 2. Moreover, although the bound in Lemma 18 is worse than the empirical Bernstein bound of Martinez-Taboada and Ramdas (2024), we find that the difference is only marginal, and the former is considerably easier to implement.*

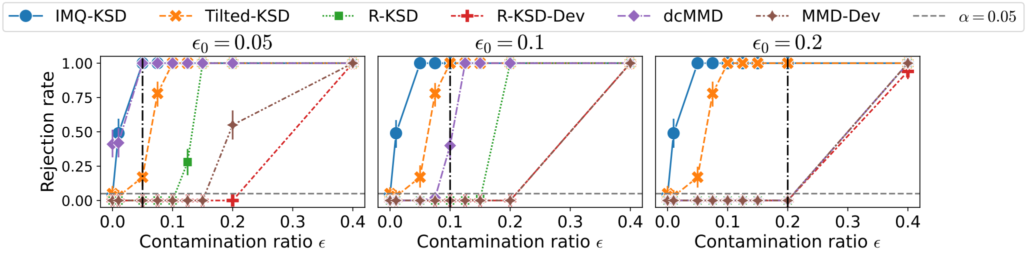


Figure 15: Rejection probability under an outlier-contaminated Gaussian model. *Black dash-dot vertical line.* Uncertainty radius $\theta = \epsilon_0 \tau_\infty^{1/2}$, with τ_∞ estimated following Section 4.2.

In Figure 15, we compare R-KSD-Dev with the bootstrap-based robust-KSD test, the robust MMD test and the standard tests under the outlier-contaminated Gaussian model in Section 5.1. We set the uncertainty radius to be $\theta = \epsilon_0 \tau_\infty^{1/2}$ for various values of ϵ_0, where τ_∞ is estimated following Section 4.2. As expected, R-KSD-Dev controls Type-I error against Huber's contamination with maximal proportion ϵ_0. However, when $\epsilon > \epsilon_0$, it has lower power compared with robust-KSD and the MMD-based tests. This is because R-KSD-Dev uses a deviation bound to construct the decision threshold, which is a uniform bound over all data-generating distributions and is therefore conservative. For the same reason, we expect tests constructed using the alternative deviation bounds discussed in Remark 10 to suffer from similar conservativeness.

Since the robust-KSD-Dev test has finite-sample guarantee on the Type-I error, it might be preferable to the bootstrap-based robust-KSD test when the sample size is small. On the other hand, the conservative nature of the robust-KSD-Dev test suggests it might only be effective for identifying the most aberrant behaviors.

Proof of Theorem 24 Proceeding as in the proof of Theorem 4, for any $Q \in \mathcal{B}^{\mathrm{KSD}}(P; \theta)$, we have the inequality

$$\Pr_{\mathbb{X}_n \sim Q}\left(\Delta_\theta(\mathbb{X}_n) \ > \ \gamma_n\right) \ \leq \ \Pr_{\mathbb{X}_n \sim Q}\left(\mathbb{S}_P(\mathbb{X}_n, Q) \ > \ \gamma_n\right) ,$$

By the deviation bound in Lemma 18, the RHS is bounded by α, thus proving the first claim.

We now fix $Q \notin \mathcal{B}^{\mathrm{KSD}}(P; \theta)$ so that we are under H_1^{C}. Since $\Delta_\theta(\mathbb{X}_n) = \max(D(\mathbb{X}_n) - \theta, 0) \geq D(\mathbb{X}_n) - \theta$, we have

$$\begin{aligned}\Pr_{\mathbb{X}_n \sim Q}\left(\Delta_\theta(\mathbb{X}_n) > \gamma_n\right) \ &\geq \ \Pr_{\mathbb{X}_n \sim Q}\left(D(\mathbb{X}_n) - \theta > \gamma_n\right) \\ &= \ \Pr_{\mathbb{X}_n \sim Q}\Big(\sqrt{n}\big(D(\mathbb{X}_n) - D(Q, P)\big) > \sqrt{n}\big(\gamma_n + \theta - D(Q, P)\big)\Big) \ .\end{aligned}$$

The same argument in Section A.5.2 shows that $\sqrt{n}\big(D(\mathbb{X}_n) - D(Q, P)\big)$ converges weakly to a Gaussian distribution. On the other hand, since $D(Q, P) > \theta$ under H_1^{C} and $\gamma_n \to 0$ as $n \to \infty$, the term $\sqrt{n}\big(\gamma_n + \theta - D(Q, P)\big) \to -\infty$. Therefore, the probability in the last line converges to one, thus proving the second claim. ■

References

Radoslaw Adamczak. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20(none):1 – 13, 2015.

Pierre Alquier and Mathieu Gerber. Universal robust regression via maximum mean discrepancy. *Biometrika*, 111(1):71–92, 2024.

Matias Altamirano, François-Xavier Briol, and Jeremias Knoblauch. Robust and scalable Bayesian online changepoint detection. In *International Conference on Machine Learning*, pages 642–663, 2023.

Matias Altamirano, François-Xavier Briol, and Jeremias Knoblauch. Robust and conjugate Gaussian process regression. In *International Conference on Machine Learning*, pages 1155–1185, 2024.

Alan Nawzad Amin, Eli N. Weinstein, and Debora Susan Marks. A kernelized Stein discrepancy for biological sequences. In *International Conference of Machine Learning*, pages 718–767, 2023.

Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E. Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gretton, Christophe Ley, Qiang Liu, Lester Mackey, Chris J. Oates, Gesine Reinert, and Yvik Swan. Stein's method meets computational statistics: A review of some recent developments. *Statistical Science*, 38(1): 120 – 139, 2023.

Miguel A. Arcones and Evarist Gine. On the bootstrap of U and V statistics. *The Annals of Statistics*, 20(2):655–674, 1992.

Krishnakumar Balasubramanian, Tong Li, and Ming Yuan. On the optimality of kernel-embedding based goodness-of-fit tests. *Journal of Machine Learning Research*, 22(1):1–45, 2021.

Alessandro Barp, François-Xavier Briol, Andrew Duncan, Mark Girolami, and Lester Mackey. Minimum Stein discrepancy estimators. *Advances in Neural Information Processing Systems*, 32:12964–12976, 2019.

Alessandro Barp, Carl-Johann Simon-Gabriel, Mark Girolami, and Lester Mackey. Targeted separation and convergence with kernel discrepancies. *Journal of Machine Learning Research*, 25(378):1–50, 2024.

Jerome Baum, Heishiro Kanagawa, and Arthur Gretton. A kernel Stein test of goodness of fit for sequential models. In *International Conference on Machine Learning*, pages 1936–1953. PMLR, 2023.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics.* Springer Science & Business Media, New York, 2004.

Kenneth L. Blackard, Theodore S. Rappaport, and Charles W. Bostian. Measurements and models of radio frequency impulsive noise for indoor wireless communications. *IEEE Journal on selected areas in communications*, 11(7):991–1001, 1993.

François-Xavier Briol, Alessandro Barp, Andrew B. Duncan, and Mark Girolami. Statistical inference for generative models with maximum mean discrepancy. *arXiv:1906.05944*, pages 1–57, 2019.

Stéphane Canu and Alex Smola. Kernel methods and the exponential family. *Neurocomputing*, 69(7-9):714–720, 2006.

Antoine Chatalic, Nicolas Schreuder, Lorenzo Rosasco, and Alessandro Rudi. Nyström kernel mean embeddings. In *International Conference on Machine Learning*, pages 3006–3024. PMLR, 2022.

Badr-Eddine Chérief-Abdellatif and Pierre Alquier. MMD-Bayes: Robust Bayesian estimation via maximum mean discrepancy. In *Advances in Approximate Bayesian Inference (AABI)*, pages 1–21, 2020.

Badr-Eddine Chérief-Abdellatif and Pierre Alquier. Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence. *Bernoulli*, 28(1):181–213, 2022.

Kyung Hyun Cho, Tapani Raiko, and Alexander Ilin. Gaussian-Bernoulli deep Boltzmann machine. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2013.

Fan Chung and Linyuan Lu. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6(2):125–145, 2002.

Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2606–2615, 2016.

Kacper P Chwialkowski, Dino Sejdinovic, and Arthur Gretton. A wild bootstrap for degenerate kernel tests. *Advances in Neural Information Processing Systems*, 27, 2014.

Anand G. Dabak and D.H. Johnson. Geometrically based robust detection. *Conference on Information Sciences and Systems, Johns Hopkins University, Baltimore*, pages 73–77, May 1994.

Ralph B D'Agostino. *Goodness-of-Fit Techniques.* Routledge, 1986.

Herold Dehling and Thomas Mikosch. Random quadratic forms and the bootstrap for U-statistics. *Journal of Multivariate Analysis*, 51(2):392–413, 1994.

Charita Dellaporta and Theodoros Damoulas. Robust Bayesian Inference for Berkson and Classical Measurement Error Models. *arXiv:2306.01468*, 2023.

Charita Dellaporta, Jeremias Knoblauch, Theodoros Damoulas, and François-Xavier Briol. Robust Bayesian inference for simulator-based models via the MMD posterior bootstrap. In *International Conference on Artificial Intelligence and Statistics*, pages 943–970, 2022.

Gerardo Duran-Martin, Matias Altamirano, Alexander Y. Shestopaloff, Jeremias Knoblauch, Matt Jones, François-Xavier Briol, and Kevin Murphy. Outlier-robust Kalman filtering through generalised Bayes. In *International Conference on Machine Learning*, pages 12138–12171, 2024.

Michael Fauß and Abdelhak M Zoubir. Old bands, new tracks—revisiting the band model for robust hypothesis testing. *IEEE Transactions on signal Processing*, 64(22):5875–5886, 2016.

Michael Fauß, Abdelhak M. Zoubir, and H. Vincent Poor. Minimax robust detection: Classic results and recent advances. *IEEE Transactions on Signal Processing*, 69:2252–2283, 2021.

Tamara Fernandez, Nicolas Rivera, Wenkai Xu, and Arthur Gretton. Kernelized Stein discrepancy tests of goodness-of-fit for time-to-event data. In *International Conference on Machine Learning*, pages 3112–3122. PMLR, 2020.

Asja Fischer and Christian Igel. Training restricted Boltzmann machines: An introduction. *Pattern Recognition*, 47(1):25–39, 2014.

Catherine Forbes, Merran Evans, Nicholas Hastings, and Brian Peacock. *Statistical Distributions*. John Wiley & Sons, 4 edition, 2011.

Robert Fortet and Edith Mourier. Convergence de la répartition empirique vers la répartition théorique. *Annales scientifiques de l'École Normale Supérieure*, 3e série, 70(3):267–285, 1953.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE transactions on neural networks and learning systems*, 25(5):845–869, 2013.

Kenji Fukumizu, Arthur Gretton, Gert Lanckriet, Bernhard Schölkopf, and Bharath K Sriperumbudur. Kernel choice and classifiability for RKHS embeddings of probability distributions. *Advances in Neural Information Processing Systems*, 22, 2009.

Rui Gao, Liyan Xie, Yao Xie, and Huan Xu. Robust hypothesis testing using Wasserstein uncertainty sets. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Ruize Gao, Feng Liu, Jingfeng Zhang, Bo Han, Tongliang Liu, Gang Niu, and Masashi Sugiyama. Maximum mean discrepancy test is aware of adversarial attacks. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3564–3575. PMLR, 18–24 Jul 2021.

Gene H. Golub and Charles F. van Loan. *Matrix Computations*. JHU Press, fourth edition, 2013. ISBN 1421407949 9781421407944.

Jackson Gorham and Lester Mackey. Measuring sample quality with Stein's method. *Advances in Neural Information Processing Systems*, 28, 2015.

Jackson Gorham and Lester Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR, 2017.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

Gökhan Gül and Abdelhak M Zoubir. Robust hypothesis testing with α-divergence. *IEEE Transactions on Signal Processing*, 64(18):4737–4750, 2016.

Robert Hafner. Construction of minimax-tests for bounded families of probability-densities. *Metrika*, 40(1):1–23, 1993.

Omar Hagrass, Bharath Sriperumbudur, and Bing Li. Spectral regularized kernel two-sample tests. *The Annals of Statistics*, 52(3):1076–1101, 2024a.

Omar Hagrass, Bharath K. Sriperumbudur, and Bing Li. Spectral regularized kernel goodness-of-fit tests. *Journal of Machine Learning Research*, 25(309):1–52, 2024b.

Omar Hagrass, Bharath Sriperumbudur, and Krishnakumar Balasubramanian. Minimax optimal goodness-of-fit testing with kernel Stein discrepancy. *arXiv preprint arXiv:2404.08278*, 2025.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

Robert V. Hogg, Elliot A. Tanis, and Dale L. Zimmerman. *Probability and Statistical Inference*, volume 993. Macmillan New York, 1977.

Kevin H. Huang, Xing Liu, Andrew Duncan, and Axel Gandy. A high-dimensional convergence theorem for U-statistics with applications to kernel-based testing. In *Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pages 3827–3918, 2023.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964.

Peter J. Huber. A robust version of the probability ratio test. *The Annals of Mathematical Statistics*, pages 1753–1758, 1965.

Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. John Wiley & Sons, 2011.

Jonathan Huggins and Lester Mackey. Random feature Stein discrepancies. *Advances in Neural Information Processing Systems*, 31, 2018.

Marie Huskova and Paul Janssen. Consistency of the generalized bootstrap for degenerate U-statistics. *The Annals of Statistics*, 21(4):1811–1823, 1993.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

Yu I Ingster. Minimax testing of nonparametric hypotheses on a distribution density in the L_p metrics. *Theory of Probability & Its Applications*, 31(2):333–337, 1987.

Yuri I. Ingster. Asymptotically minimax hypothesis testing for nonparametric alternatives. I, II, III. *Math. Methods Statist*, 2(2):85–114, 1993.

Paul Janssen. Weighted Bootstrapping of U-Statistics. *Journal of Statistical Planning and Inference*, 38(1):31–41, 1994. ISSN 0378-3758.

Paul Janssen. Bootstrapping U-statistics. *South African Statistical Journal*, 31(2):185–216, 1997.

Wittawat Jitkrittum, Wenkai Xu, Zoltán Szabó, Kenji Fukumizu, and Arthur Gretton. A linear-time kernel goodness-of-fit test. In *International Conference on Neural Information Processing Systems*, pages 261–270. Curran Associates Inc., 2017.

Heishiro Kanagawa, Arthur Gretton, and Lester Mackey. Controlling moments with kernel Stein discrepancies. *arXiv preprint arXiv:2211.05408*, 2022.

S. Kassam. Robust hypothesis testing for bounded classes of probability densities (corresp.). *IEEE Transactions on Information Theory*, 27(2):242–247, 1981.

Oscar Key, Arthur Gretton, François-Xavier Briol, and Tamara Fernandez. Composite goodness-of-fit tests with kernels. *Journal of Machine Learning Research*, 26(51):1–60, 2025.

A.N. Kolmogorov. Sulla determinazione empirica di una legge didistribuzione. *Giorn Dell'inst Ital Degli Att*, 4:89–91, 1933.

Diane Lambert. Qualitative robustness of tests. *Journal of the American Statistical Association*, 77(378):352–357, 1982.

Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

Erich Leo Lehmann and Joseph P Romano. *Testing Statistical Hypotheses*. Springer Cham, fourth edition, 2022.

Anne Leucht and Michael H. Neumann. Dependent wild bootstrap for degenerate U- and V-statistics. *Journal of Multivariate Analysis*, 117:257–280, 2013. ISSN 0047-259X.

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

Qiang Liu, Jason Lee, and Michael Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 276–284, 2016.

Xing Liu, Andrew B. Duncan, and Axel Gandy. Using perturbation to improve goodness-of-fit tests based on kernelized Stein discrepancy. In *International Conference on Machine Learning*, Proceedings of Machine Learning Research, pages 21527–21547, 2023.

Diego Martinez-Taboada and Aaditya Ramdas. Empirical Bernstein in smooth Banach spaces. *arXiv preprint arXiv:2409.06060*, 2024.

Takuo Matsubara, Jeremias Knoblauch, François-Xavier Briol, and Chris J Oates. Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):997–1022, 2022.

Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

Chris J Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79 (3):695–718, 2017.

Frédéric Ouimet. Explicit formulas for the joint third and fourth central moments of the multinomial distribution. *arXiv preprint arXiv:2006.09059*, 2020.

Jason A. Palmer, Ken Kreutz-Delgado, and Scott Makeig. Strong Sub- and Super-Gaussianity. In Vincent Vigneron, Vicente Zarzoso, Eric Moreau, Rémi Gribonval, and Emmanuel Vincent, editors, *Latent Variable Analysis and Signal Separation*, pages 303–310, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15995-4.

Iosif Pinelis. Optimum bounds for the distributions of martingales in Banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

Marc Postman, John Peter Huchra, and Margaret J Geller. Probes of large-scale structure in the corona borealis region. *Astronomical Journal (ISSN 0004-6256), vol. 92, Dec. 1986, p. 1238-1247.*, 92:1238–1247, 1986.

Yu V Prokhorov. Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214, 1956.

Yichen Qin and Carey E Priebe. Robust hypothesis testing via L_q-likelihood. *Statistica Sinica*, pages 1793–1813, 2017.

Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

Sashank Reddi, Aaditya Ramdas, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the high dimensional power of a linear-time two sample test under mean-shift alternatives. In *Artificial Intelligence and Statistics*, pages 772–780. PMLR, 2015.

Helmut Rieder. Qualitative robustness of rank tests. *The Annals of Statistics*, 10(1):205–211, 1982.

Giorgio Rizzoni and PS Min. Detection of sensor failures in automotive engines. *IEEE Transactions on Vehicular Technology*, 40(2):487–500, 1991.

Kathryn Roeder. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411):617–624, 1990.

Antonin Schrab and Ilmun Kim. Robust kernel hypothesis testing under data corruption. *arXiv preprint arXiv:2405.19912*, 2024.

Antonin Schrab, Benjamin Guedj, and Arthur Gretton. KSD aggregated goodness-of-fit test. *Advances in Neural Information Processing Systems*, 35:32624–32638, 2022.

Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. MMD aggregated two-sample test. *Journal of Machine Learning Research*, 24(194):1–81, 2023.

Robert J Serfling. *Approximation Theorems of Mathematical Statistics.* John Wiley & Sons, 2009.

Xiaofeng Shao. The dependent wild bootstrap. *Journal of the American Statistical Association*, 105(489):218–235, 2010.

Abhishek B Sharma, Leana Golubchik, and Ramesh Govindan. Sensor faults: Detection methods and prevalence in real-world datasets. *ACM Transactions on Sensor Networks (TOSN)*, 6(3):1–39, 2010.

Jiaxin Shi and Lester Mackey. A finite-particle convergence rate for stein variational gradient descent. *Advances in Neural Information Processing Systems*, 36, 2024.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Zhongchang Sun and Shaofeng Zou. Kernel robust hypothesis testing. *IEEE Transactions on Information Theory*, 2023.

Ilya Tolstikhin, Bharath K. Sriperumbudur, and Krikamol Muandet. Minimax estimation of kernel mean embeddings. *Journal of Machine Learning Research*, 18(86):1–47, 2017.

Aad W Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000.

Li K Wenliang and Heishiro Kanagawa. Blindness of score-based methods to isolated components and mixing proportions. *arXiv preprint arXiv:2008.10087*, 2020.

Geoffrey Wolfer and Pierre Alquier. Variance-aware estimation of kernel mean embedding. *arXiv preprint arXiv:2210.06672*, 2022.

George Wynne, Mikołaj Kasprzak, and Andrew B. Duncan. A spectral representation of kernel Stein discrepancy with application to goodness-of-fit tests for measures on infinite dimensional Hilbert spaces. *arXiv preprint arXiv:2206.04552*, 2022.

Wenkai Xu and Takeru Matsuda. A Stein goodness-of-fit test for directional distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 320–330. PMLR, 2020.

Wenkai Xu and Takeru Matsuda. Interpretable Stein goodness-of-fit tests on Riemannian manifold. In *International Conference on Machine Learning*, pages 11502–11513. PMLR, 2021.

Wenkai Xu and Gesine Reinert. A Stein goodness-of-test for exponential random graph models. In *International Conference on Artificial Intelligence and Statistics*, pages 415–423. PMLR, 2021.

Jiasen Yang, Qiang Liu, Vinayak Rao, and Jennifer Neville. Goodness-of-fit testing for discrete distributions via Stein discrepancy. In *International Conference on Machine Learning*, pages 5561–5570. PMLR, 2018.

Jiasen Yang, Vinayak Rao, and Jennifer Neville. A Stein–Papangelou goodness-of-fit test for point processes. In *International Conference on Artificial Intelligence and Statistics*, pages 226–235. PMLR, 2019.

Pengfei Yang and Biao Chen. Robust Kullback-Leibler divergence and universal hypothesis testing for continuous distributions. *IEEE Transactions on Information Theory*, 65(4): 2360–2373, 2018.

Liangwei Zhang, Jing Lin, Bin Liu, Zhicong Zhang, Xiaohui Yan, and Muheng Wei. A review on deep learning applications in prognostics and health management. *Ieee Access*, 7:162415–162438, 2019.

Mingtian Zhang, Oscar Key, Peter Hayes, David Barber, Brooks Paige, and François-Xavier Briol. Towards healing the blindness of score matching. *arXiv preprint arXiv:2209.07396*, 2022.