Supplement to "Kernel-based tests for Likelihood-Free Hypothesis Testing"

Patrik Róbert Gerber[*]
Department of Mathematics, MIT
Cambridge, MA 02139
prgerber@mit.edu

Tianze Jiang[*]
Department of Mathematics, MIT
Cambridge, MA 02139
tjiang@mit.edu

Yury Polyanskiy[*]
Department of EECS, MIT
Cambridge, MA 02139
yp@mit.edu

Rui Sun[*]
Department of Mathematics, MIT
Cambridge, MA 02139
eruisun@mit.edu

Contents

[*]Equal contribution.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

A Notation

We use $A \gtrsim B, A \lesssim B, A \asymp B$ to denote $A = \Omega(B), B = \Omega(A)$ and $A = \Theta(B)$ respectively, where the hidden constants depend on untracked parameters multiplicatively.[2]

We write $\mathrm{TV}, \mathrm{KL}, \chi^2$ for total-variation, KL-divergence and χ^2-divergence, respectively. We write $D(P_{Y|X}\|Q_{Y|X}|P_X) = \mathbb{E}_{X\sim P_X} D(P_{Y|X}\|Q_{Y|X})$ as the *conditional divergence* for any probability measures P, Q on two variables X, Y and divergence $D \in \{\mathrm{TV}, \mathrm{KL}, \chi^2\}$.

We write ℓ^p for the usual ℓ^p sequence space and L^p for the usual L^p space with respect to the Lebesgue measure. Both the ℓ^p norm and the L^p norm are written as $\|\cdot\|_p$ if no ambiguity arises.

For real numbers $a, b \in \mathbb{R}$ we also write $\max\{a, b\}$ as $a \vee b$ and $\min\{a, b\}$ as $a \wedge b$.

We use $\vec{1}_d$ to denote an d-dimensional all 1's vector.

For an integer $k \in \mathbb{Z}^+$, we write $[k]$ as a short notation for the set $\{1, 2, \ldots, k\}$.

In the proofs of Theorem 3.2 and Theorem 3.3, we use $\overset{!}{=}$ for an equality that we are trying to prove.

B Applications of Theorem 3.2

Usually, minimax rates of testing are proven under separation assumptions using more traditional measures of distance such as L^p, where $p \in [1, \infty]$. In this section we show one example of how Theorem 3.2 can be used to recover known results, and also obtain some novel results under L^2-separation and L^1-separation.

B.1 Bounded Discrete Distributions Under L^2/L^1-Separation

Sample Complexity Upper Bounds Let $\mathcal{P}_{\mathsf{Db}}(k, C)$ be the set of all discrete distributions P supported on $[k] = \{1, 2, \ldots, k\}$ satisfying $\max_{1\le i\le k} p(i) \le C/k$, where p is the probability mass function of P (here $\sum_{i=1}^k p(k) = 1$). For distributions P_X, P_Y, P_Z we shall write p_X, p_Y, p_Z as their probability mass functions, respectively.

Let us apply Theorem 3.2 with underlying space $\mathcal{X} = [k]$ and measure $\mu = \frac{1}{k}\sum_{i=1}^k \delta_i$. Take the kernel $K(x, y) = \mathbb{1}\{x = y\} = \sum_{i=1}^k \mathbb{1}\{x = y = i\}$, and note that for any two distributions P_X, P_Y we have

$$\mathrm{MMD}^2(P_X, P_Y) = \mathbb{E}\left[K(X, X') + K(Y, Y') - 2K(X, Y)\right] = \sum_i |p_X(i) - p_Y(i)|^2$$

where $(X, X', Y, Y') \sim P_X^{\otimes 2} \otimes P_Y^{\otimes 2}$. So the corresponding MMD is the ℓ^2-distance on probability mass functions. Note also that $K = \sum_{i=1}^k \frac{1}{k}\left(\sqrt{k}\mathbb{1}\{x = i\}\right)\left(\sqrt{k}\mathbb{1}\{y = i\}\right)$, where $\left\{\sqrt{k}\mathbb{1}\{x = i\}\right\}_{i=1}^k$ forms an orthonormal basis of $L^2(\mu)$. So K has only one nonzero eigenvalue, namely

$$\lambda_1 = \lambda_2 = \ldots = \lambda_k = 1/k,$$

of multiplicity k. Suppose that we observe samples X, Y, Z of size n, n, m from $P_X, P_Y, P_Z \in \mathcal{P}_{\mathsf{Db}}(k, C)$, where $\mathrm{MMD}(P_X, P_Y) = \sqrt{\sum_i |p_X(i) - p_Y(i)|^2} \ge \epsilon$. Plugging into Theorem 3.2 shows that:

Proposition B.1. *For any two $P_X, P_Y \in \mathcal{P}_{\mathsf{Db}}(k, C)$, if the ℓ^2-distance between p_X, p_Y is at least ϵ, then testing (mLFHT) is possible at total error α using n simulation samples and m real data samples provided that*

$$\begin{aligned} \min\{m, n\} &\gtrsim \frac{C\|\lambda\|_\infty \log(1/\alpha)(1+R)^2}{\delta^2\epsilon^2} \asymp \frac{\log(1/\alpha)(1+R)^2}{k\epsilon^2\delta^2}, \\ \min\{n, \sqrt{mn}\} &\gtrsim \frac{C\|\lambda\|_2 \sqrt{\log(1/\alpha)}}{\epsilon^2\delta} \asymp \frac{\sqrt{\log(1/\alpha)}}{\sqrt{k}\epsilon^2\delta}. \end{aligned} \tag{1}$$

[2] For example, the first equation in (1) means that there exists a constant c independent of $\alpha, k, \epsilon, \delta, R$, such that $\min\{m, n\} \ge c\frac{\log(1/\alpha)(1+R)^2}{k\epsilon^2\delta^2}$.

where R is defined as in the assumption (iii) of Section 3.2.

We can convert the above results to measure separation with respect to total variation (recall $\mathrm{TV}(p,q) = \frac{1}{2}\sum_i |p(i)-q(i)| = \frac{1}{2}\|p-q\|_1$) using the AM-QM inequality $\|p_X - p_Y\|_1 \leq \sqrt{k}\|p_X - p_Y\|_2$. Then, taking $R \asymp \alpha \asymp \delta = \Theta(1)$ recovers the minimax optimal results of [3; 7; 8], for LFHT over the class $\mathcal{P}_{\mathsf{Db}}$. Note that analogous results for two-sample testing follow from the above using the reduction presented in Section 3.5.

Sample Complexity Lower Bounds Recall the definition of $J^\star_\epsilon$ and note that $\|\lambda\|^2_{2,J} = \frac{\min(J-1,k)}{k^2}$ for all $J \geq 2$. By Corollary 3.6 we see that $J^\star_\epsilon \gtrsim k$ as soon as $\epsilon \lesssim 1/k$. Thus, for $\epsilon \lesssim 1/k$ the necessity of

$$m \gtrsim \frac{\log(1/\alpha)}{k\epsilon^2\delta^2},\ n \gtrsim \frac{\sqrt{\log(1/\alpha)}}{\sqrt{k}\epsilon^2} \text{ and } m + \sqrt{mn} \gtrsim \frac{\sqrt{\log(1/\alpha)}}{\sqrt{k}\epsilon^2\delta} \tag{2}$$

follows by Theorem 3.3. Here it is crucial to note that when $\delta = \Theta(1)$, we have

$$m + \sqrt{mn} \gtrsim \frac{\sqrt{\log(1/\alpha)}}{\sqrt{k}\epsilon^2} \text{ and } n \gtrsim \frac{\sqrt{\log(1/\alpha)}}{\sqrt{k}\epsilon^2} \iff \sqrt{mn} \gtrsim \frac{\sqrt{\log(1/\alpha)}}{\sqrt{k}\epsilon^2} \text{ and } n \gtrsim \frac{\sqrt{\log(1/\alpha)}}{\sqrt{k}\epsilon^2}$$

and hence the upper bound (1) meets with the lower bound (2) provided $R \asymp \delta = \Theta(1)$. Once again, setting $R \asymp \delta \asymp \alpha = \Theta(1)$ we the optimal lower bounds recovering the results of [3] (in the regime $\epsilon \lesssim 1/k$). In short we can also recover the following result for LFHT.

Proposition B.2 ([3, Theorem 1, adapted])**.** *On the class $P_{\mathsf{Db}}(k,C)$, using n simulation samples and m real data samples, if*

$$n \gtrsim \frac{1}{\sqrt{k}\epsilon^2}, \quad m \gtrsim \frac{1}{k\epsilon^2}, \quad \sqrt{mn} \gtrsim \frac{1}{\sqrt{k}\epsilon^2}, \tag{3}$$

then for any two distributions $P_X, P_Y \in \mathcal{P}_{\mathsf{Db}}(k,C)$ with $\|p_X - p_Y\|_2 \geq \epsilon$, testing (LFHT) is possible with a total error of 1%. Conversely, to ensure the existence of a procedure that can test (LFHT) with a total error of 1% for any $P_X, P_Y \in \mathcal{P}_{\mathsf{Db}}(k,C)$ with $\|p_X - p_Y\|_2 \geq \epsilon$, the number of observations (n,m) must satisfy

$$n \gtrsim \frac{1}{\sqrt{k}\epsilon^2}, \quad m \gtrsim \frac{1}{k\epsilon^2}, \quad \sqrt{mn} \gtrsim \frac{1}{\sqrt{k}\epsilon^2}. \tag{4}$$

The implied constants in (3) and (4) do not depend on k and ϵ, but may differ.

B.2 β-Hölder Smooth Densities on $[0,1]^d$ Under L^2/L^1-Separation

Sample Complexity Upper Bounds Let $\mathcal{P}_{\mathsf{H}}(\beta,d,C)$ be the set of all distributions on $[0,1]^d$ with β-Hölder smooth Lebesgue-density p satisfying $\|p\|_{\mathcal{C}^\beta} \leq C$ for some constant $C > 1$, where

$$\|p\|_{\mathcal{C}^\beta} \triangleq \max_{0\leq|\alpha|\leq\lceil\beta-1\rceil} \|f^{(\alpha)}\|_\infty + \sup_{x\neq y\in[0,1]^d,|\alpha|=\lceil\beta-1\rceil} \frac{|f^{(\alpha)}(x) - f^{(\alpha)}(y)|}{\|x-y\|_2^{\beta-\lceil\beta-1\rceil}},$$

where $\lceil\beta-1\rceil$ is the largest integer strictly smaller than β and $|\alpha| = \sum_i \alpha_i$ is the norm of a multi-index $\alpha \in \mathbb{N}^d$. Abusing notation, we also use $\mathcal{P}_{\mathsf{H}}(\beta,d,C)$ to denote the set of all corresponding density functions.

We take $K(x,y) = \sum_j \mathbb{1}\{x,y \in B_j\}$, where $\{B_j\}_{j\in[\kappa]^d}$ is the j'th cell of the regular grid of size κ^d on $[0,1]^d$, i.e., $B_j = [(j-\vec{1}_d)/\kappa, j/\kappa]$ for $j \in [\kappa]^d$. Clearly there are κ^d nonzero eigenvalues, each equal to 1. The following approximation result is due to Ingster [5], see also [1, Lemma 7.2].

Lemma B.3. *Let $f,g \in \mathcal{P}_{\mathsf{H}}(\beta,d,C)$ with $\|f-g\|_2 \geq \epsilon$. Then, there exist constants c,c' independent of ϵ such that for any $\kappa \geq c\epsilon^{-1/\beta}$,*

$$\mathrm{MMD}(f,g) \geq c'\|f-g\|_2.$$

Now, suppose that we have samples X,Y,Z of size n,n,m from $P_X,P_Y,P_Z \in \mathcal{P}_{\mathsf{H}}(\beta,d,C)$ with densities p_X,p_Y,p_Z such that $\|p_X - p_Y\|_2 \geq \epsilon$. Then, Theorem 3.2 combined with Lemma B.3 and the choice $\kappa \asymp \epsilon^{-1/\beta}$ shows that

Proposition B.4. *Testing* (mLFHT) *on* $\mathcal{P}_{\mathsf{H}}(\beta, d, C)$ *at total error* α *using* n *simulation and* m *real data samples is possible provided*

$$\min\{m, n\} \gtrsim \frac{C\|\lambda\|_\infty \log(1/\alpha)(1+R)^2}{\delta^2\epsilon^2} \asymp \frac{\log(1/\alpha)(1+R)^2}{\delta^2\epsilon^2},$$
$$\min\{n, \sqrt{nm}\} \gtrsim \frac{C\|\lambda\|_2 \sqrt{\log(1/\alpha)}}{\epsilon^2\delta} \asymp \frac{\sqrt{\log(1/\alpha)}}{\epsilon^{(2\beta+d/2)/\beta}\delta},$$

where ϵ *is an* L^2*-distance lower bound between* P_X, P_Y *and* R *is defined as in the assumption* (iii) *of Section 3.2.*

Setting $R \asymp \alpha \asymp \delta = \Theta(1)$ recovers the optimal results of [3] for the class $\mathcal{P}_{\mathsf{H}}$. Once again, identical results under L^1 separation follow from Jensen's inequality $\|\cdot\|_{L^1([0,1]^d)} \leq \|\cdot\|_{L^2([0,1]^d)}$. Note that analogous results for two-sample testing follow from the above using the reduction presented in Section 3.5.

Sample Complexity Lower Bounds The kernel defined in the previous paragraph is not suitable for constructing lower bounds over the class $\mathcal{P}_{\mathsf{H}}$ because its eigenfunctions do not necessarily lie in $\mathcal{P}_{\mathsf{H}}$. It would be possible to consider a different kernel that is more adapted to this problem/class but we do not pursue this here.

B.3 $(\beta, 2)$-Sobolev Smooth Densities on $\mathbb{R}^d$ Under L^2-Separation

Sample Complexity Upper Bounds Let $\mathcal{P}_{\mathsf{S}}(\beta, d, C)$ be the class of distributions that are supported on $\mathbb{R}^d$ and whose Lebesgue density p satisfies $\|p\|_{\beta,2} \leq C$, where

$$\|p\|_{\beta,2} \triangleq \left\|(1 + \|\cdot\|)^\beta \mathcal{F}[p]\right\|_2 \tag{5}$$

and $\mathcal{F}$ denotes the Fourier transform. Again, abusing notation, we write $\mathcal{P}_{\mathsf{S}}(\beta, d, C)$ both as the set of distributions and the set of density functions.

We take the Gaussian kernel $G_\sigma(x, y) = \sigma^{-d}\exp(-\|x - y\|_2^2/\sigma^2)$ on $\mathcal{X} = \mathbb{R}^d$ with base measure $\mathrm{d}\mu(x) = \exp(-x^2)\mathrm{d}x$. In [9] the authors showed that the two-sample test that thresholds the Gaussian MMD with appropriately chosen variance σ^2 achieves the minimax optimal sample complexity over $\mathcal{P}_{\mathsf{S}}$, when separation is measured by L^2. A key ingredient in their proof is the following inequality.

Lemma B.5 ([9, Lemma 5])**.** *Let* $f, g \in \mathcal{P}_{\mathsf{S}}(\beta, d, C)$ *with* $\|f - g\|_2 \geq \epsilon$*. Then, there exist constants* c, c' *independent of* ϵ *such that for any* $\sigma \leq c\,\epsilon^{1/\beta}$*, we have*

$$\mathrm{MMD}(f, g) \geq c'\|f - g\|_2.$$

Now, suppose that we have samples X, Y, Z of sizes n, n, m from $P_X, P_Y, P_Z \in \mathcal{P}_{\mathsf{S}}(\beta, d, C)$ for some constant C with densities p_X, p_Y, p_Y satisfying $\|p_X - p_Y\|_2 \geq \epsilon$.

Note that the heat-semigroup is an L^2-contraction ($\|\lambda\|_\infty \leq 1$) and that

$$\|\lambda\|_2^2 = \int G_\sigma(x, y)^2 \mathrm{d}\mu(x)\mathrm{d}\mu(y) \asymp \sigma^{-d}$$

up to constants depending on the dimension. Theorem 3.2 combined with Lemma B.5 and a choice $\sigma \asymp \epsilon^{1/\beta}$ yields the following result.

Proposition B.6. *Testing* (mLFHT) *over the class* $\mathcal{P}_{\mathsf{S}}$ *with total error* α *is possible provided*

$$\min\{m, n\} \gtrsim \frac{C\|\lambda\|_\infty \log(1/\alpha)(1+R)^2}{\delta^2\epsilon^2} \asymp \frac{\log(1/\alpha)(1+R)^2}{\delta^2\epsilon^2}$$
$$\min\{n, \sqrt{nm}\} \gtrsim \frac{C\|\lambda\|_2 \sqrt{\log(1/\alpha)}}{\epsilon^2\delta} \asymp \frac{\sqrt{\log(1/\alpha)}}{\epsilon^{(2\beta+d/2)/\beta}\delta},$$

where ϵ *is the lower bound on the* L^2*-distance between* P_X, P_Y *and* R *is defined as in the assumption* (iii) *of Section 3.2.*

Taking $R \asymp \delta \asymp \alpha = \Theta(1)$ above, we obtain new results for LFHT and using the reduction from two-sample testing given in Section 3.5 we partly recover [9, Theorem 5]. Only partly, because the above requires bounded density with respect to our base measure $\mathrm{d}\mu(x) = \exp(-x^2)\mathrm{d}x$.

Sample Complexity Lower Bounds Note that our lower bound Theorem 3.3 doesn't apply because the top eigenfunction of the Gaussian kernel is not constant. Once again, a more careful choice of the base measure (or kernel) might lead to a more suitable argument for the lower bound. We leave such pursuit as open.

C Black-box Boosting of Success Probability

In this section we briefly describe how upper bounds on the minimax sample complexity in the constant error probability regime ($\alpha = \Theta(1)$) can be used to obtain the dependence $\log(1/\alpha)$ in the small error probability regime ($\alpha = o(1)$). We will argue abstractly in a way that applies to the setting of Theorem 3.2.

Suppose that from some distributions $P_1, P_2, \ldots, P_k$ we take samples $X^1, X^2, \ldots, X^k$ of size $n_1, n_2, \ldots, n_k$ respectively and are able to decide between two hypotheses H_0 and H_1 (fixed but arbitrary) with total error probability at most $1/3$. Call this test as $\Psi(X^1, \ldots, X^k) \in \{0, 1\}$, so that

$$\mathbb{P}(\Psi(X^1, \ldots, X^k) = 0 | H_0) \geq 2/3 \qquad \text{and} \qquad \mathbb{P}(\Psi(X^1, \ldots, X^k) = 1 | H_1) \geq 2/3.$$

Now, to each an error of $o(1)$, instead, we take $18n_1 \log(2/\alpha), \ldots, 18n_k \log(2/\alpha)$ observations from P_1 through P_k, and split each sample into $18\log(2/\alpha)$ equal sized batches $\{X^{i,j}\}_{i\in[k], j\in[18\log(2/\alpha)]}$. Here $18\log(2/\alpha)$ is assumed to be an integer without loss of generality. The split samples form $18\log(2/\alpha)$ independent binary random variables

$$A_j \triangleq \Psi(X^{1,j}, \ldots, X^{k,j})$$

for $j = 1, 2, \ldots, 18\log(2/\alpha)$. We claim that the majority voting test

$$\Psi_\alpha(\{X^{i,j}\}_{i,j}) = \begin{cases} 1 & \text{if } \bar{A} \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

tests H_0 against H_1 with total probability of error at most α, where

$$\bar{A} \triangleq \frac{1}{18\log(2/\alpha)} \sum_{j=1}^{18\log(2/\alpha)} A_j.$$

Indeed, by Hoeffding's inequality, we have

$$\mathbb{P}\left(\bar{A} \geq 1/2 \middle| H_0\right) \leq \alpha/2$$
$$\mathbb{P}\left(\bar{A} \leq 1/2 \middle| H_1\right) \leq \alpha/2.$$

Therefore, in the remainder of our upper bound proofs, we only focus on achieving a constant probability of error ($\alpha = \Theta(1)$) as the logarithmic dependence follows by the above.

Remark C.1. As mentioned in the discussion succeeding Corollary 3.6, we do conjecture the *tight* dependence in the upper bound to be $\sqrt{\log(\alpha^{-1})}$ instead of $\log(\alpha^{-1})$ shown by this method.

D Proof of Theorem 3.2

D.1 Notation and Technical Tools

We use the expansion

$$K(x, y) = \sum_\ell \lambda_\ell e_\ell(x) e_\ell(y)$$

extensively, where $\lambda \triangleq (\lambda_1, \lambda_2, \ldots)$ are K's eigenvalues (regarded as an integral operator on $L^2(\mu)$) in non-increasing order and $e_1, e_2, \ldots$ are the corresponding eigenfunctions forming an orthonormal

basis for $L^2(\mu)$, and convergence is to be understood in $L^2(\mu)$. We use the notation $\langle \cdot \rangle \triangleq \int \cdot \, \mathrm{d}\mu$. For all $u \in L^2(\mu)$ we define

$$u_\ell \triangleq \langle u e_\ell \rangle, \quad u_{\ell\ell'} \triangleq \langle u e_\ell e_{\ell'} \rangle, \quad \ell = 1, 2, \ldots$$

and consequently $u = \sum_\ell u_\ell e_\ell$. We also define

$$K[u](\cdot) \triangleq \int K(t, \cdot) u(t) \mu(\mathrm{d}t) = \sum_\ell \lambda_\ell u_\ell e_\ell(\cdot),$$

where the second equality follows from the orthonormality of $\{e_\ell\}_{\ell=1}^\infty$. Note that the RKHS embedding satisfies $\theta_u \triangleq \int K(x, \cdot) u(x) \mathrm{d}\mu(x) = K[u]$. Now, for P_X we write

$$x_\ell \triangleq (p_X)_\ell = \langle p_X e_\ell \rangle, \quad x_{\ell\ell'} \triangleq (p_X)_{\ell\ell'} = \langle p_X e_\ell e_{\ell'} \rangle, \quad \ell, \ell' = 1, 2, \ldots$$

where p_X is the μ-density of P_X. The similar notations also apply to P_Y, P_Z. The following identities will be very useful in our proofs.

Lemma D.1. *For each identity below, let $f, g, h \in L^2(\mu)$ be such that the quantity is well defined. Then,*

$$\|\theta_f\|_{\mathcal{H}_K}^2 = \sum_\ell \lambda_\ell f_\ell^2 \tag{6}$$

$$\mathrm{MMD}^2(f, g) = \sum_\ell \lambda_\ell (f_\ell - g_\ell)^2 \tag{7}$$

$$\|K[f]\|_2^2 = \sum_\ell \lambda_\ell^2 f_\ell^2 \tag{8}$$

$$\sum_\ell \lambda_\ell f_\ell g_\ell = \langle f K[g] \rangle = \langle K[f] g \rangle \tag{9}$$

$$\sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} h_{\ell\ell'} f_\ell g_{\ell'} = \langle h K[f] K[g] \rangle \tag{10}$$

$$\sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} g_{\ell\ell'} f_{\ell\ell'} = \sum_\ell \lambda_\ell \langle f e_\ell K[g e_\ell] \rangle. \tag{11}$$

Suppose that f, g are probability densities with respect to μ that are bounded by C. Then

$$0 \le \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} g_{\ell\ell'} f_{\ell\ell'} \le C^2 \|\lambda\|_2^2. \tag{12}$$

Proof. We prove each claim, starting with (6). Clearly

$$\begin{aligned}
\|\theta_f\|_{\mathcal{H}_K}^2 &= \|K[f]\|_{\mathcal{H}_K}^2 \\
&= \left\| \int K(x, \cdot) f(x) \mathrm{d}\mu(x) \right\|_{\mathcal{H}_K}^2 \\
&= \iint \langle K(x, \cdot), K(y, \cdot) \rangle_{\mathcal{H}_K} f(x) f(y) \mathrm{d}\mu(x) \mathrm{d}\mu(y) \\
&= \iint K(x, y) f(x) f(y) \mathrm{d}\mu(x) \mathrm{d}\mu(y) \\
&= \sum_\ell \lambda_\ell f_\ell^2
\end{aligned}$$

as required. The second claim (7) follows immediately from (6) by definition. For (8) by orthogonality we have

$$\begin{aligned}
\|K[f]\|_2^2 &= \| \sum_\ell \lambda_\ell f_\ell e_\ell \|_2^2 \\
&= \sum_\ell \lambda_\ell^2 f_\ell^2.
\end{aligned}$$

For (9) by the definition of $K[\cdot]$ we have

$$\begin{aligned}\sum_\ell \lambda_\ell f_\ell g_\ell &= \left\langle \left(\sum_\ell \lambda_\ell f_\ell e_\ell \right) g \right\rangle \\ &= \langle K[f] g \rangle.\end{aligned}$$

For (10) we can write

$$\begin{aligned}\sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} h_{\ell\ell'} f_\ell g_{\ell'} &= \sum_\ell \lambda_\ell f_\ell \left\langle \left(\sum_{\ell'} \lambda_{\ell'} g_{\ell'} e_{\ell'} \right) h e_\ell \right\rangle \\ &= \sum_\ell \lambda_\ell f_\ell \langle K[g] h e_\ell \rangle \\ &= \langle K[g] h K[f] \rangle.\end{aligned}$$

Finally, for (11) we have

$$\begin{aligned}\sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} f_{\ell\ell'} g_{\ell\ell'} &= \sum_\ell \lambda_\ell \left\langle \left(\sum_{\ell'} \lambda_{\ell'} g_{\ell\ell'} e_{\ell'} \right) f e_\ell \right\rangle \\ &= \sum_\ell \lambda_\ell \langle K[g e_\ell] f e_\ell \rangle.\end{aligned}$$

Suppose now that f, g are probability densities with respect to μ that are bounded by $C > 0$. Let X, Y be independent random variables following the densities f, g. Then

$$\begin{aligned}\sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} f_{\ell\ell'} g_{\ell\ell'} &= \mathbb{E}\left[\left(\sum_\ell \lambda_\ell e_\ell(X) e_\ell(Y) \right)^2 \right] \\ &\leq C^2 \int_{\mathcal{X}} \int_{\mathcal{X}} \left(\sum_\ell \lambda_\ell e_\ell(x) e_\ell(y) \right)^2 \mathrm{d}\mu(x) \mathrm{d}\mu(y) \\ &= C^2 \|\lambda\|_2^2\end{aligned}$$

as claimed, where we used that the e_ℓ are orthonormal. □

D.2 Mean and Variance Computation

We take $\pi = \delta/2$. Our statistic reads

$$\begin{aligned}-T(X,Y,Z) + \gamma(X,Y,\pi) &= \langle \theta_{\widehat{P}_Z} - (\bar{\pi}\theta_{\widehat{P}_X} + \pi\theta_{\widehat{P}_Y}), \theta_{\widehat{P}_X} - \theta_{\widehat{P}_Y} \rangle_{u,\mathcal{H}_K} \\ &= \frac{1}{nm} \underbrace{\sum_{ij} k(X_i, Z_j)}_{\mathsf{I}} - \frac{1}{nm} \underbrace{\sum_{ij} k(Y_i, Z_j)}_{\mathsf{II}} - \frac{2\bar{\pi}}{n(n-1)} \underbrace{\sum_{i<i'} k(X_i, X_{i'})}_{\mathsf{III}} \\ &\quad + \frac{2\pi}{n(n-1)} \underbrace{\sum_{i<i'} k(Y_i, Y_{i'})}_{\mathsf{IV}} + \frac{\bar{\pi} - \pi}{n^2} \underbrace{\sum_{ij} k(X_i, Y_j)}_{\mathsf{V}}.\end{aligned}$$

Recall that $\nu = \arg\min_{\nu' \in \mathbb{R}} \mathrm{MMD}(P_Z, \bar{\nu}' P_X + \nu' P_Y)$. Let us write $z = \bar{\nu}x + \nu y + r$ for $1 - \bar{\nu} = \nu$, where the residual term is denoted as $r \in L^2(\mu)$. Let $\theta_r = \int r(t) K(t, \cdot) \mu(\mathrm{d}t)$ be the mean embedding of r. Under both hypotheses we assume that $\|\theta_r\|_{\mathcal{H}_K} \leq R \cdot \mathrm{MMD}(P_X, P_Y)$, moreover $\langle \theta_r, \theta_{P_Y} - \theta_{P_X} \rangle_{\mathcal{H}_K} = 0$ by the definition of ν. We look at each of the $5 + \binom{5}{2} = 15$ terms of the

variance separately.

$$\begin{aligned}
\mathrm{var}(\mathsf{I}) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Big\{ n(n-1)m(z_{\ell\ell'} - z_\ell z_{\ell'}) x_\ell x_{\ell'} + nm(m-1)(x_{\ell\ell'} - x_\ell x_{\ell'}) z_\ell z_{\ell'} \\
&\qquad + nm(x_{\ell\ell'} z_{\ell\ell'} - x_\ell x_{\ell'} z_\ell z_{\ell'}) \Big\} \\
\mathrm{var}(\mathsf{II}) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Big\{ n(n-1)m(z_{\ell\ell'} - z_\ell z_{\ell'}) y_\ell y_{\ell'} + nm(m-1)(y_{\ell\ell'} - y_\ell y_{\ell'}) z_\ell z_{\ell'} \\
&\qquad + nm(y_{\ell\ell'} z_{\ell\ell'} - y_\ell y_{\ell'} z_\ell z_{\ell'}) \Big\} \\
\mathrm{var}(\mathsf{III}) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Big\{ \binom{n}{2} (x_{\ell\ell'}^2 - x_\ell^2 x_{\ell'}^2) + (\binom{n}{2}^2 - \binom{n}{2} - \binom{4}{2}\binom{n}{4})(x_{\ell\ell'} - x_\ell x_{\ell'}) x_\ell x_{\ell'} \Big\} \\
\mathrm{var}(\mathsf{IV}) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Big\{ \binom{n}{2} (y_{\ell\ell'}^2 - y_\ell^2 y_{\ell'}^2) + (\binom{n}{2}^2 - \binom{n}{2} - \binom{4}{2}\binom{n}{4})(y_{\ell\ell'} - y_\ell y_{\ell'}) y_\ell y_{\ell'} \Big\} \\
\mathrm{var}(\mathsf{V}) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Big\{ n^2(n-1)(y_{\ell\ell'} - y_\ell y_{\ell'}) x_\ell x_{\ell'} + n^2(n-1)(x_{\ell\ell'} - x_\ell x_{\ell'}) y_\ell y_{\ell'} \\
&\qquad + n^2(x_{\ell\ell'} y_{\ell\ell'} - x_\ell x_{\ell'} y_\ell y_{\ell'}) \Big\}
\end{aligned}$$

For the cross terms we obtain

$$\begin{aligned}
\mathrm{cov}(\mathsf{I}, \mathsf{II}) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} n^2 m(z_{\ell\ell'} - z_\ell z_{\ell'}) x_\ell y_{\ell'} \\
\mathrm{cov}(\mathsf{I}, \mathsf{III}) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} n(n-1)m(x_{\ell\ell'} - x_\ell x_{\ell'}) z_\ell x_{\ell'} \\
\mathrm{cov}(\mathsf{I}, \mathsf{IV}) &= 0 \\
\mathrm{cov}(\mathsf{I}, \mathsf{V}) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} n^2 m(x_{\ell\ell'} - x_\ell x_{\ell'}) z_\ell y_{\ell'} \\
\mathrm{cov}(\mathsf{II}, \mathsf{III}) &= 0 \\
\mathrm{cov}(\mathsf{II}, \mathsf{IV}) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} n(n-1)m(y_{\ell\ell'} - y_\ell y_{\ell'}) z_\ell y_{\ell'} \\
\mathrm{cov}(\mathsf{II}, \mathsf{V}) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} n^2 m(y_{\ell\ell'} - y_\ell y_{\ell'}) z_\ell x_{\ell'} \\
\mathrm{cov}(\mathsf{III}, \mathsf{IV}) &= 0 \\
\mathrm{cov}(\mathsf{III}, \mathsf{V}) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} n^2 (n-1)(x_{\ell\ell'} - x_\ell x_{\ell'}) x_\ell y_{\ell'} \\
\mathrm{cov}(\mathsf{IV}, \mathsf{V}) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} n^2 (n-1)(y_{\ell\ell'} - y_\ell y_{\ell'}) y_\ell x_{\ell'}.
\end{aligned}$$

Note that $\binom{n}{2}^2 - \binom{n}{2} - \binom{n}{2}\binom{n}{4} = n(n-1)^2 - n(n-1)$. Collecting terms, and simplifying, we get the coefficient of the $\frac{1}{n}$ term:

$$\begin{aligned}
\mathrm{Coef}\left(\frac{1}{n}\right) = \sum_{\ell,\ell'} \lambda_\ell \lambda_{\ell'} \Bigg(& \underbrace{(x_{\ell\ell'} - x_\ell x_{\ell'}) z_\ell z_{\ell'}}_{\mathrm{var(I)}} + \underbrace{(y_{\ell\ell'} - y_\ell y_{\ell'}) z_\ell z_{\ell'}}_{\mathrm{var(II)}} + \underbrace{4\bar{\pi}^2 (x_{\ell\ell'} - x_\ell x_{\ell'}) x_\ell x_{\ell'}}_{\mathrm{var(III)}} \\
& + \underbrace{4\pi^2 (y_{\ell\ell'} - y_\ell y_{\ell'}) y_\ell y_{\ell'}}_{\mathrm{var(IV)}} + \underbrace{(\bar{\pi} - \pi)^2 (y_{\ell\ell'} - y_\ell y_{\ell'}) x_\ell x_{\ell'} + (\bar{\pi} - \pi)^2 (x_{\ell\ell'} - x_\ell x_{\ell'}) y_\ell y_{\ell'}}_{\mathrm{var(V)}} \\
& - \underbrace{4\bar{\pi}(x_{\ell\ell'} - x_\ell x_{\ell'}) z_\ell x_{\ell'}}_{\mathrm{cov(I,III)}} + \underbrace{2(\bar{\pi} - \pi)(x_{\ell\ell'} - x_\ell x_{\ell'}) z_\ell y_{\ell'}}_{\mathrm{cov(I,V)}} \\
& - \underbrace{4\pi(y_{\ell\ell'} - y_\ell y_{\ell'}) z_\ell y_{\ell'}}_{\mathrm{cov(II,IV)}} - \underbrace{2(\bar{\pi} - \pi)(y_{\ell\ell'} - y_\ell y_{\ell'}) z_\ell x_{\ell'}}_{\mathrm{cov(II,V)}} \\
& - \underbrace{4\bar{\pi}(\bar{\pi} - \pi)(x_{\ell\ell'} - x_\ell x_{\ell'}) x_\ell y_{\ell'}}_{\mathrm{cov(III,V)}} + \underbrace{4\pi(\bar{\pi} - \pi)(y_{\ell\ell'} - y_\ell y_{\ell'}) y_\ell x_{\ell'}}_{\mathrm{cov(IV,V)}} \Bigg).
\end{aligned}$$

After expanding z_ℓ as $z_\ell = \bar{\nu} x_\ell + \nu y_\ell + r_\ell$, we split the calculation into multiple parts to simplify it. First, we focus on terms that are multiplied by $(x_{\ell\ell'} - x_\ell x_{\ell'})$ and do not contain r_ℓ or $r_{\ell'}$. Using Lemma D.1 extensively and the fact that $\bar{\pi} = 1 - \pi, \bar{\nu} = 1 - \nu$, we find that the sum of these terms equals

$$\begin{aligned}
& \bar{\nu}^2 \langle xK[x]^2 \rangle + \nu^2 \langle xK[y]^2 \rangle + 2\bar{\nu}\nu \langle xK[x]K[y] \rangle - \bar{\nu}^2 \langle xK[x] \rangle^2 - \nu^2 \langle xK[y] \rangle^2 - 2\bar{\nu}\nu \langle xK[x] \rangle \langle xK[y] \rangle \\
& + 4\bar{\pi}^2 \langle xK[x]^2 \rangle - 4\bar{\pi}^2 \langle xK[x] \rangle^2 + (\bar{\pi} - \pi)^2 \langle xK[y]^2 \rangle - (\bar{\pi} - \pi)^2 \langle xK[y] \rangle^2 \\
& - 4\bar{\pi}\bar{\nu} \langle xK[x]^2 \rangle - 4\bar{\pi}\nu \langle xK[x]K[y] \rangle + 4\bar{\pi}\bar{\nu} \langle xK[x] \rangle^2 + 4\bar{\pi}\nu \langle xK[x] \rangle \langle xK[y] \rangle \\
& + 2(\bar{\pi} - \pi)\bar{\nu} \langle xK[x]K[y] \rangle + 2(\bar{\pi} - \pi)\nu \langle xK[y]^2 \rangle - 2(\bar{\pi} - \pi)\bar{\nu} \langle xK[x] \rangle \langle xK[y] \rangle - 2(\bar{\pi} - \pi)\nu \langle xK[y] \rangle^2 \\
& - 4\bar{\pi}(\bar{\pi} - \pi) \langle xK[x]K[y] \rangle + 4\bar{\pi}(\bar{\pi} - \pi) \langle xK[x] \rangle \langle xK[y] \rangle \\
= & (\bar{\nu} - 2\bar{\pi})^2 \Big(\langle xK[x - y]^2 \rangle - \langle xK[x - y] \rangle^2 \Big) \\
\leq & C \left\| \lambda \right\|_\infty \mathrm{MMD}^2(P_X, P_Y).
\end{aligned}$$

Similarly, the terms involving $(y_{\ell\ell'} - y_\ell y_{\ell'})$ but not r_ℓ or $r_{\ell'}$ sum up to the quantity

$$(\nu - 2\pi)^2 \Big(\langle yK[x - y]^2 \rangle - \langle yK[x - y] \rangle^2 \Big) \leq C \|\lambda\|_\infty \mathrm{MMD}^2(P_X, P_Y).$$

Next, collecting the terms involving both $(x_{\ell\ell'} - x_\ell x_{\ell'})$ and r_ℓ or $r_{\ell'}$ we get

$$\begin{aligned}
& 2\bar{\nu} \langle xK[r]K[x] \rangle + 2\nu \langle xK[r]K[y] \rangle + \langle xK[r]^2 \rangle - 2\bar{\nu} \langle xK[x] \rangle \langle xK[r] \rangle - 2\nu \langle xK[y] \rangle \langle xK[r] \rangle - \langle xK[r] \rangle^2 \\
& - 4\bar{\pi} \langle xK[x]K[r] \rangle + 4\bar{\pi} \langle xK[x] \rangle \langle xK[r] \rangle \\
& + 2(\bar{\pi} - \pi) \langle xK[y]K[r] \rangle - 2(\bar{\pi} - \pi) \langle xK[y] \rangle \langle xK[r] \rangle \\
= & \, 2(\bar{\nu} - 2\bar{\pi}) \Big(\langle xK[r]K[x - y] \rangle - \langle xK[r] \rangle \langle xK[x - y] \rangle \Big) + \langle xK[r]^2 \rangle - \langle xK[r] \rangle^2 \\
\lesssim & C \left\| \lambda \right\|_\infty (R + R^2) \, \mathrm{MMD}^2(P_X, P_Y).
\end{aligned}$$

Finally, collecting the terms involving both $(y_{\ell\ell'} - y_\ell y_{\ell'})$ and r_ℓ or $r_{\ell'}$ we get

$$\begin{aligned}
& 2(\nu - 2\pi) \Big(\langle yK[r]K[y - x] \rangle - \langle yK[r] \rangle \langle yK[y - x] \rangle \Big) + \langle yK[r]^2 \rangle - \langle yK[r] \rangle^2 \\
\lesssim & C \|\lambda\|_\infty (R + R^2) \, \mathrm{MMD}^2(P_X, P_Y).
\end{aligned}$$

Similarly we get

$$\begin{aligned}
\mathrm{Coef}\left(\frac{1}{m}\right) = \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Bigg(& \underbrace{(z_{\ell\ell'} - z_\ell z_{\ell'}) x_\ell x_{\ell'}}_{\mathrm{var(I)}} + \underbrace{(z_{\ell\ell'} - z_\ell z_{\ell'}) y_\ell y_{\ell'}}_{\mathrm{var(I)}} + \underbrace{2(z_{\ell\ell'} - z_\ell z_{\ell'}) x_\ell y_{\ell'}}_{\mathrm{cov(I,II)}} \Bigg) \\
= & \langle zK[x - y]^2 \rangle - \langle zK[x - y] \rangle^2 \\
\lesssim & C \|\lambda\|_\infty \, \mathrm{MMD}^2(P_X, P_Y).
\end{aligned}$$

The remaining coefficients don't rely on subtle cancellations, and simple bounds yield

$$\begin{aligned}
\mathrm{Coef}\left(\frac{1}{n(n-1)}\right) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Bigg(\underbrace{4\bar{\pi}^2 \left(\frac{1}{2}(x_{\ell\ell'}^2 - x_\ell^2 x_{\ell'}^2) - (x_{\ell\ell'} - x_\ell x_{\ell'}) x_\ell x_{\ell'} \right)}_{\mathrm{var(III)}} \\
&\qquad + \underbrace{4\pi^2 \left(\frac{1}{2}(y_{\ell\ell'}^2 - y_\ell^2 y_{\ell'}^2) - (y_{\ell\ell'} - y_\ell y_{\ell'}) y_\ell y_{\ell'} \right)}_{\mathrm{var(IV)}} \Bigg) \\
&\lesssim C^2 \|\lambda\|_2^2 \\
\mathrm{Coef}\left(\frac{1}{nm}\right) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Bigg(\underbrace{-(z_{\ell\ell'} - z_\ell z_{\ell'}) x_\ell x_{\ell'} - (x_{\ell\ell'} - x_\ell x_{\ell'}) z_\ell z_{\ell'} + (x_{\ell\ell'} z_{\ell\ell'} - x_\ell x_{\ell'} z_\ell z_{\ell'})}_{\mathrm{var(I)}} \\
&\qquad - \underbrace{(z_{\ell\ell'} - z_\ell z_{\ell'}) y_\ell y_{\ell'} - (y_{\ell\ell'} - y_\ell y_{\ell'}) z_\ell z_{\ell'} + (y_{\ell\ell'} z_{\ell\ell'} - y_\ell y_{\ell'} z_\ell z_{\ell'})}_{\mathrm{var(I)}} \Bigg) \\
&\lesssim C^2 \|\lambda\|_2^2 \\
\mathrm{Coef}\left(\frac{1}{n^2}\right) &= \sum_{\ell\ell'} \lambda_\ell \lambda_{\ell'} \Bigg(\underbrace{(\bar{\pi} - \pi)\left(-(y_{\ell\ell'} - y_\ell y_{\ell'}) x_\ell x_{\ell'} - (x_{\ell\ell'} - x_\ell x_{\ell'}) y_\ell y_{\ell'} + (x_{\ell\ell'} y_{\ell\ell'} - x_\ell x_{\ell'} y_\ell y_{\ell'})\right)}_{\mathrm{var(V)}} \Bigg) \\
&\lesssim C^2 \|\lambda\|_2^2.
\end{aligned}$$

Summarizing, we've found that

$$\begin{aligned}
\mathrm{var}(T(X,Y,Z) - \gamma(X,Y,\pi)) \lesssim \left(\frac{1}{n} + \frac{1}{m}\right) C\|\lambda\|_\infty (1+R^2)\,\mathrm{MMD}^2(P_X, P_Y) \\
+ \left(\frac{1}{n^2} + \frac{1}{nm}\right) C^2 \|\lambda\|_2^2.
\end{aligned} \tag{13}$$

Using that $\langle \theta_r, \theta_{P_Y} - \theta_{P_X} \rangle_{\mathcal{H}_K} = 0$, we compute the expectation to be

$$\mathbb{E}\left[-T(X,Y,Z) + \gamma(X,Y,\pi)\right] = (\pi - \nu)\,\mathrm{MMD}^2(P_X, P_Y).$$

Taking $\pi \triangleq \delta/2$ and applying Chebyshev's inequality shows that there exists a universal constant $c > 0$, such that the testing problem is possible at constant error probability (say $\alpha = 5\%$), provided that the sample sizes m, n satisfy the following inequalities:

$$\begin{aligned}
\min\{m, n\} &\geq c \frac{C\|\lambda\|_\infty (1+R^2)}{\delta^2 \epsilon^2} \\
\min\{n, \sqrt{nm}\} &\geq c \frac{C\|\lambda\|_2}{\delta \epsilon^2}.
\end{aligned}$$

By repeated sample splitting and majority voting (see Appendix C), we can boost the success probability of this test to the desired level $1 - \alpha$ by incurring a multiplicative $\Theta(\log(1/\alpha))$ factor on the sample sizes n, m, which yields the desired result.

E Proof of Theorem 3.3

E.1 Information theoretic tools

Our lower bounds rely on the method of two fuzzy hypotheses [14]. Given a measurable space $\mathcal{S}$, let $\mathcal{M}(\mathcal{S})$ denote the set of all probability measures on $\mathcal{S}$. We call subsets $H \subseteq \mathcal{M}(\mathcal{S})$ hypotheses. The following is the main technical result that our proofs rely on.

Lemma E.1. *Take hypotheses $H_0, H_1 \subseteq \mathcal{M}(\mathcal{S})$ and $P_0, P_1 \in \mathcal{M}(\mathcal{S})$ random with $\mathbb{P}(P_i \in H_i) = 1$. Then*

$$\inf_{\psi} \max_{i=0,1} \sup_{P \in H_i} P(\psi \neq i) \geq \frac{1}{2}\left(1 - \mathrm{TV}(\mathbb{E}P_0, \mathbb{E}P_1)\right),$$

where the infimum is over all tests $\psi : \mathcal{X} \to \{0, 1\}$.

Proof. For any ψ

$$\begin{aligned}\max_{i=0,1} \sup_{\mathbb{P}_i \in H_i} \mathbb{P}_i(\psi \neq i) &\geq \frac{1}{2} \sup_{\mathbb{P}_i \in H_i} \left(\mathbb{P}_0(\psi = 1) + \mathbb{P}_1(\psi = 0)\right) \\ &\geq \frac{1}{2}\,\mathbb{E}\left[P_0(\psi = 1) + P_1(\psi = 0)\right].\end{aligned}$$

Optimizing over ψ we get that the RHS above is equal to $\frac{1}{2}(1 - \mathrm{TV}(\mathbb{E}\,P_0, \mathbb{E}\,P_1))$ as required. □

Therefore, to prove a lower bound on the minimax sample complexity of testing with total error probability α, we just need to construct two random measures $P_i \in H_i$ such that $1 - \mathrm{TV}(\mathbb{E}\,P_0, \mathbb{E}\,P_1) = \Omega(\alpha)$. In our proofs we also use the following standard results on f-divergences.

Lemma E.2 ([12, Section 7])**.** *For any probability distributions P, Q the inequalities*

$$1 - \mathrm{TV}(P, Q) \geq \frac{1}{2}\exp(-\mathrm{KL}(P\|Q)) \geq \frac{1}{2}\frac{1}{1 + \chi^2(P\|Q)}$$

hold.

Lemma E.3 (Chain rule for χ^2-divergence)**.** *Let $P_{X,Y}, Q_{X,Y}$ be probability measures such that the marginals on X are equal ($P_X = Q_X$). Then*

$$\chi^2(P_{X,Y}\|Q_{X,Y}) = \chi^2(P_{Y|X}\|Q_{Y|X}|P_X).$$

Proof. Let $P_{X,Y}, Q_{X,Y}$ have densities p, q with respect to some μ. Then, by some abuse of notation, we have

$$\begin{aligned}\chi^2(P_{X,Y}\|Q_{X,Y}) &= -1 + \int \frac{p(x,y)^2}{q(x,y)}\mathrm{d}\mu(x,y) \\ &= -1 + \int \frac{p(y|x)^2 p(x)}{q(y|x)}\mathrm{d}\mu(x,y) \\ &= \int p(x) \int \left(\frac{p(y|x)^2}{q(y|x)} - 1\right)\mathrm{d}\mu(y,x) \\ &= \chi^2(P_{Y|X}\|Q_{Y|X}|P_X).\end{aligned}$$

□

E.2 Constructing hard instances

Recall that in the statement of Theorem 3.3, we assume that $\mu(\mathcal{X}) = 1$, $\sup_{x \in \mathcal{X}} K(x,x) \leq 1$ and $\int K(x,y)\mu(\mathrm{d}x) \equiv \lambda_1$. Let $f_0 \equiv 1$ and for each $\eta \in \{\pm 1\}^{\mathbb{N}}$ define

$$f_\eta = 1 + \epsilon \underbrace{\sum_{j \geq 2} \rho_j \eta_j e_j}_{\triangleq g_\eta} \tag{14}$$

where $\{\rho_j\}_{j \geq 2}$ is chosen as $\rho_j = \mathbb{1}\{2 \leq j \leq J\}\sqrt{\lambda_j}/\|\lambda\|_{2,J}$, where we define $\|\lambda\|_{2,J} = \sqrt{\sum_{2 \leq j \leq J} \lambda_j^2}$ for some $J \geq 2$. Notice that $\int f_\eta(x)\mu(\mathrm{d}x) = \mu(\mathcal{X}) = 1$ due to orthogonality of the eigenfunctions. Assume from here on that J is chosen so that for all η we have $f_\eta(x) \geq 1/2$ for all $x \in \mathcal{X}$. This makes f_η into a valid probability density with respect to the base measure μ. Before continuing, we prove the following Lemma, which gives a lower bound on the maximal J for which $f_\eta \geq 1/2$ for all η.

Lemma E.4. *$J \le J_\epsilon^\star$ holds provided $2\epsilon\sqrt{J-1} \le \|\lambda\|_{2J}$.*

Proof of Lemma E.4. Notice that

$$\|e_j\|_\infty = \sup_{x\in\mathcal{X}} \langle K(x,\cdot), e_j\rangle_{\mathcal{H}} \le \sup_{x\in\mathcal{X}} \|K(x,\cdot)\|_{\mathcal{H}}\|e_j\|_{\mathcal{H}} \le \frac{1}{\sqrt{\lambda_j}}, \tag{15}$$

where we use $\|K(x,\cdot)\|_{\mathcal{H}} = \sqrt{K(x,x)}$. We have

$$\begin{aligned}\|g_\eta\|_\infty &= \epsilon\|\sum_{j\ge 2}\rho_j\eta_j e_j\|_\infty = \epsilon \sup_{x\in\mathcal{X}}\langle K(x,\cdot), \sum_{j\ge 2}\rho_j\eta_j e_j\rangle_{\mathcal{H}} \\ &\le \epsilon\|\sum_{j\ge 2}\rho_j\eta_j e_j\|_{\mathcal{H}} = \epsilon\sqrt{\sum_{j\ge 2}\rho_j^2/\lambda_j} = \frac{\epsilon\sqrt{J-1}}{\|\lambda\|_{2,J}},\end{aligned}$$

and the result follows. □

Note that Lemma E.4 immediately gives us a proof of Corollary 3.6.

Proof of Corollary 3.6. Suppose that J is such that $\sum_{j=2}^J \lambda_j^2 \ge c^2\|\lambda\|_2^2$. Then, by Lemma E.4, if $\epsilon \le \|\lambda\|_{2J}/(2\sqrt{J-1})$ then $J \le J_\epsilon^\star$. By assumption, this is implied by the inequality $\epsilon \le c\|\lambda\|_2/(2\sqrt{J-1})$, and the result follows. □

Continuing with our proof, note that by construction we have

$$\mathrm{MMD}^2(f_0, f_\eta) = \sum_{j\ge 2}\lambda_j\rho_j^2 = \epsilon^2, \quad \forall \eta \in \{\pm 1\}^{\mathbb{N}}. \tag{16}$$

E.2.1 Lower Bound on m

Again, we apply Lemma E.1 with the new (deterministic) construction

$$P_0 = f_0^{\otimes n} \otimes f_{\mathbb{1}}^{\otimes n} \otimes (1+\delta\epsilon g_{\mathbb{1}})^{\otimes m} \qquad P_1 = f_0^{\otimes n} \otimes f_{\mathbb{1}}^{\otimes n} \otimes f_0^{\otimes m}, \tag{17}$$

where we write $f_{\mathbb{1}} = f_{(1,1,\dots)}$ and similarly for $g_{\mathbb{1}}$. By the data-processing inequality for χ^2-divergence (also by Lemma E.3), we may drop the first $2n$ coordinates and obtain

$$\begin{aligned}\chi^2(\mathbb{E}\,P_0, \mathbb{E}\,P_1) &= \chi^2((1+\delta\epsilon g_{\mathbb{1}})^{\otimes m}\|f_0^{\otimes m}) \\ &= (1+\delta^2\epsilon^2)^m - 1 \\ &\le \exp(\delta^2\epsilon^2 m) - 1.\end{aligned}$$

By Lemma E.2 we

$$1 - \mathrm{TV}(\mathbb{E}\,P_0, \mathbb{E}\,P_1) \gtrsim \frac{1}{\chi^2(\mathbb{E}\,P_0, \mathbb{E}\,P_1) - 1} \ge \exp(-\delta^2\epsilon^2 m) \overset{!}{=} \Omega(\alpha).$$

The lower bound $m \gtrsim \log(1/\alpha)/(\delta\epsilon)^2$ now follows readily.

E.2.2 Lower Bound on n

Once again, we apply Lemma E.1 to the new construction

$$P_0 = f_0^{\otimes n} \otimes f_\eta^{\otimes n} \otimes f_0^{\otimes m}, \qquad P_1 = f_\eta^{\otimes n} \otimes f_0^{\otimes n} \otimes f_0^{\otimes m}, \tag{18}$$

where we put a uniform prior on $\eta \in \{\pm 1\}^{\mathbb{N}}$ as before. Using the subadditivity of total variation under products, we compute

$$\begin{aligned}\mathrm{TV}(\mathbb{E}\,P_0, \mathbb{E}\,P_1) &= \mathrm{TV}(f_0^{\otimes n} \otimes \mathbb{E}\,f_\eta^{\otimes n}, \mathbb{E}[f_\eta^{\otimes n}] \otimes f_0^{\otimes n}) \\ &\le 2\,\mathrm{TV}(\mathbb{E}\,f_\eta^{\otimes n}, f_0^{\otimes n}).\end{aligned}$$

Just as in Appendix E.2.3 we upper bound by the χ^2-divergence to get

$$\begin{aligned}
\chi^2(\mathbb{E}\, f_\eta^{\otimes n} \| f_0^{\otimes n}) &= -1 + \mathbb{E}_{\eta\eta'} \int \prod_{i=1}^{n} (f_\eta(x_i) f_{\eta'}(x_i)) \mu(\mathrm{d}x_1) \dots \mu(\mathrm{d}x_n) \\
&\le -1 + \mathbb{E} \exp(n\epsilon^2 \sum_{j\ge 2} \rho_j^2 \eta_j \eta_j') \\
&= -1 + \prod_{j\ge 2} \cosh(n\epsilon^2 \rho_j^2) \\
&\le -1 + \exp(n^2\epsilon^4 \sum_{j\ge 2} \rho_j^4) \\
&= -1 + \exp(n^2\epsilon^4 / \|\lambda\|_{2,J}^2).
\end{aligned}$$

Again, by Lemma E.2 we obtain

$$1 - \mathrm{TV}(\mathbb{E}\, P_0, \mathbb{E}\, P_1) \gtrsim \frac{1}{\chi^2(\mathbb{E}\, P_0 \| \,\mathbb{E}\, P_1) - 1} \ge \exp(-n^2\epsilon^4/\|\lambda\|_{2,J}^2) \overset{!}{=} \Omega(\alpha).$$

The lower bound $n \gtrsim \sqrt{\log(1/\alpha)} \|\lambda\|_{2,J}/\epsilon^2$ now follows readily.

E.2.3 Lower Bound on $m \cdot n$

We take a uniform prior on η and consider the random measures

$$P_0 = f_0^{\otimes n} \otimes f_\eta^{\otimes n} \otimes ((1-\delta) f_0 + \delta f_\eta)^{\otimes m} \qquad \text{and} \qquad P_1 = f_0^{\otimes n} \otimes f_\eta^{\otimes n} \otimes f_0^{\otimes m}. \tag{19}$$

Our goal is to apply Lemma E.1 to P_0, P_1. Notice that $(1-\delta) f_0 + \delta f_\eta = 1 + \delta\epsilon g_\eta$. Let us write X, Y, Z for the marginals first n, second n and last m coordinates of P_0 and P_1. By the data processing inequality and the chain rule Lemma E.3 we have

$$\begin{aligned}
\chi^2(\mathbb{E}\, P_0 \| \,\mathbb{E}\, P_1) &= \chi^2((\mathbb{E}\, P_0)_{Y,Z} \| (\mathbb{E}\, P_1)_{Y,Z}) \\
&= \chi^2((\mathbb{E}\, P_0)_{Z|Y} \| (\mathbb{E}\, P_1)_{Z|Y} | (\mathbb{E}\, P_0)_Y) \\
&= \mathbb{E}\, \chi^2 \left(\mathbb{E}\left[(1+\delta\epsilon g_\eta)^{\otimes m} \middle| Y \right] \| f_0^{\otimes m} \right) =: (\dagger).
\end{aligned}$$

Notice that the expectation inside the χ^2-divergence is with respect to η given the variables Y, or in other words, over the posterior of η with uniform prior given n observations from the density $1 + \epsilon g_\eta = f_\eta$. The outer expectation is over Y. Given Y, let η and η' be i.i.d. from said posterior. We get the bound

$$\begin{aligned}
(\dagger) + 1 &\le \mathbb{E} \int \prod_{i=1}^{m} (1 + \delta\epsilon g_\eta(x_i))(1 + \delta\epsilon g_{\eta'}(x_i)) \mu(\mathrm{d}x_i) \\
&= \mathbb{E}(1 + \delta^2\epsilon^2 \sum_{j\ge 2} \rho_j^2 \eta_j \eta_j')^m \\
&\le \mathbb{E} \exp(\delta^2\epsilon^2 m \sum_{j\ge 2} \rho_j^2 \eta_j \eta_j').
\end{aligned}$$

Define the collections of variables $\eta_{-j} = \{\eta_j\}_{j\ge 2} \setminus \{\eta_j\}$ and η'_{-j} similarly. We shall prove the following claim:

$$\mathbb{E}\left[\exp(\delta^2\epsilon^2 m \rho_j^2 \eta_j \eta_j') \middle| \, \eta_{-j} \eta'_{-j} \right] \le \exp(c\delta^2\epsilon^4(\delta^2 m^2 + mn)\rho_j^4) \tag{20}$$

for some universal constant $c > 0$. Assuming that (20) holds, by induction we can show that

$$\begin{aligned}
(\dagger) + 1 &\le \exp(c\delta^2(\delta^2 m^2 + mn)\epsilon^4 \sum_{j\ge 2} \rho_j^4) \\
&= \exp(c\delta^2(\delta^2 m^2 + mn)\epsilon^4 / \|\lambda\|_{2,J}^2).
\end{aligned}$$

Thus, if $mn + \delta^2 m^2 = o\left(\|\lambda\|_{2,J}^2/(\delta^2\epsilon^4)\right)$ then testing is impossible.

We now prove (20). Since the variable $\eta'_j\eta'_j$ is either 1 or -1, we have

$$\mathbb{E}\left[\exp(\delta^2\epsilon^2 m\rho_j^2\eta_j\eta'_j)\middle|\,\eta_{-j}\eta'_{-j}\right] = (e^{\delta^2\epsilon^2 m\rho_j^2} - e^{-\delta^2\epsilon^2 m\rho_j^2})\cdot\mathbb{P}(\eta_j\eta'_j = 1|\eta_{-j}\eta'_{-j}) + e^{-\delta^2\epsilon^2 m\rho_j^2}.$$

Let us write $\eta_{\pm 1,j}$ for the vector of signs equal to η but whose j'th coordinate is ± 1 respectively. Looking at the probability above, and using the independence of η, η' given Y, we have

$$\begin{aligned}\mathbb{P}(\eta_j\eta'_j = 1|Y,\eta_{-j},\eta'_{-j}) &= \mathbb{P}(\eta_j = 1|Y,\eta_{-j})^2 + \mathbb{P}(\eta_j = -1|Y,\eta_{-j})^2\\ &= \frac{1}{4}\frac{(f_{\eta_{1j}}^{\otimes n}(Y))^2 + (f_{\eta_{-1j}}^{\otimes n}(Y))^2}{\left(\frac{1}{2}f_{\eta_{1j}}^{\otimes n}(Y) + \frac{1}{2}f_{\eta_{-1j}}^{\otimes n}(Y)\right)^2}.\end{aligned}$$

Taking the expectation $\mathbb{E}[\,\cdot\,|\eta_{-j},\eta'_{-j}]$ and using the HM-AM inequality $(\frac{1}{2}(x+y))^{-1} \le \frac{1}{2}(\frac{1}{x}+\frac{1}{y})$ valid for all $x, y > 0$ gives

$$\begin{aligned}\mathbb{P}(\eta_j\eta'_j = 1|\eta_{-j},\eta'_{-j}) &= \frac{1}{4}\int\frac{(\prod_{i=1}^n f_{\eta_{1j}}(x_i))^2 + (\prod_{i=1}^n f_{\eta_{-1j}}(x_i))^2}{\frac{1}{2}\prod_{i=1}^n f_{\eta_{1j}}(x_i) + \frac{1}{2}\prod_{i=1}^n f_{\eta_{-1j}}(x_i)}\mu(\mathrm{d}x_1)\dots\mu(\mathrm{d}x_n)\\ &\le \frac{1}{4} + \frac{1}{8}\int\left(\frac{(\prod_{i=1}^n f_{\eta_{1j}}(x_i))^2}{\prod_{i=1}^n f_{\eta_{-1j}}(x_i)} + \frac{(\prod_{i=1}^n f_{\eta_{-1j}}(x_i))^2}{\prod_{i=1}^n f_{\eta_{1j}}(x_i)}\right)\mu(\mathrm{d}x_1)\dots\mu(\mathrm{d}x_n) = (\star).\end{aligned}$$

Note that $f_{\eta_{1j}} = f_{\eta_{-1j}} + 2\epsilon\rho_j e_j$. Using the lower bound $f_{\eta_{\pm 1j}}(x) \ge \frac{1}{2}$ for all $x \in \mathcal{X}$ and the inequality $1 + x \le \exp(x)$, we get

$$\begin{aligned}(\star) &\le \frac{1}{4} + \frac{1}{8}\left[\left(1 + \int\frac{4\epsilon^2\rho_j^2 e_j^2(x)}{f_{\eta_{-1j}}(x)}\mu(\mathrm{d}x)\right)^n + \left(1 + \int\frac{4\epsilon^2\rho_j^2 e_j^2(x)}{f_{\eta_{1j}}(x)}\mu(\mathrm{d}x)\right)^n\right]\\ &\le \frac{1}{4}(1 + e^{8\epsilon^2 n\rho_j^2}).\end{aligned}$$

Recall that $(\star)$ is a probability so $(\star) \le 1$, and we obtain

$$(\star) \le \frac{1}{4}(1 + e^{8\epsilon^2 n\rho_j^2 \wedge \ln 3}).$$

Putting it together and applying Lemma E.5 we get

$$\begin{aligned}\text{LHS of (20)} &\le (e^{\delta^2\epsilon^2 m\rho_j^2} - e^{-\delta^2\epsilon^2 m\rho_j^2})\frac{1}{4}(1 + e^{8\epsilon^2 n\rho_j^2\wedge\ln 3}) + e^{-\delta^2\epsilon^2 m\rho_j^2}\\ &\le e^{c\delta^2\epsilon^4\rho_j^4(\delta^2 m^2 + mn)}\end{aligned}$$

for universal $c = 16 > 0$. Thus, by Lemma E.2 we obtain

$$1 - \mathrm{TV}(\mathbb{E}\,P_0, \mathbb{E}\,P_1) \gtrsim \frac{1}{\chi^2(\mathbb{E}\,P_0, \mathbb{E}\,P_1) + 1} \ge \exp(-c\delta^2\epsilon^4(\delta^2 m^2 + mn)/\|\lambda\|_{2,J}^2) \overset{!}{=} \Omega(\alpha).$$

The necessity of

$$mn + \delta^2 m^2 \gtrsim \frac{\log(1/\alpha)\|\lambda\|_{2,J}^2}{\delta^2\epsilon^4}$$

follows immediately. In fact, this was even stronger than stated in Theorem 3.3 with $mn + m^2 \gtrsim \log(1/\alpha)\|\lambda\|_{2,J}^2/(\delta^2\epsilon^4)$.[3]

Lemma E.5. *For $a, b \ge 0$, the following inequality holds:*

$$\frac{1}{4}(e^a - e^{-a})(1 + e^{b\wedge\ln 3}) + e^{-a} \le e^{2(ab+a^2)}.$$

Proof. If $b \ge \ln 3$ or $a \ge 1$ we have:

$$\text{LHS} \le \frac{1}{4}(e^a - e^{-a})(1 + e^{\ln 3}) + e^{-a} = e^a \le e^{\frac{b}{\ln 3}a + a^2}.$$

[3] We have $mn + m^2 \le (\sqrt{mn} + m)^2 \le 2(mn + m^2)$, so $\sqrt{mn} + m \asymp \sqrt{mn + m^2}$.

If $b < \ln 3$ and $a < 1$, we have

$$e^b \le 1 + \frac{2}{\ln 3} b \le 1 + 2b, \quad \frac{e^a + e^{-a}}{2} \le e^{a^2}, \quad \frac{e^a - e^{-a}}{2} \le \frac{e - e^{-1}}{2} a \le 2a,$$

and then

$$\begin{aligned}
\frac{1}{4}(e^a - e^{-a})(1 + e^b) + e^{-a} &= \frac{1}{2}(e^a + e^{-a}) + \frac{e^b - 1}{4}(e^a - e^{-a}) \\
&\le e^{a^2} + 2ab \\
&\le e^{a^2}(1 + 2ab) \\
&\le e^{a^2 + 2ab}
\end{aligned}$$

The result follows from $\ln 3 > 1$. □

F Proofs From Section 4

F.1 Heuristic Justification of the Objective (7)

As usual, let X, Y, Z denotes samples of sizes n, n, m from P_X, P_Y, P_Z respectively. Let us give a heuristic justification for using the training objective defined in (7) for the purpose of obtaining a kernel for LFHT/mLFHT. Note that originally it was proposed as a training objective for kernels to be used in two sample testing. Recall that our test for LFHT can be written as

$$\Psi_{1/2}(X, Y, Z) = \mathbb{1}\Big\{T_{\mathsf{LF}} \ge 0\Big\}$$

where

$$T_{\mathsf{LF}} = \mathrm{MMD}_u^2(\widehat{P}_Z, \widehat{P}_Y; K) - \mathrm{MMD}_u^2(\widehat{P}_Z, \widehat{P}_X; K),$$

Heuristically, to maximize the power of (mLFHT), we would like to maximize the following population quantity

$$J_{\mathsf{LF}} \triangleq \frac{\mathbb{E}_0[T_{\mathsf{LF}}] - \mathbb{E}_1[T_{\mathsf{LF}}]}{\sqrt{\mathrm{var}_0(T_{\mathsf{LF}})}}$$

where

$$\begin{aligned}
\mathbb{E}_0[T_{\mathsf{LF}}] &= \mathbb{E}_{X,Y,Z}[T_{\mathsf{LF}} | P_Z = P_X] = +\mathrm{MMD}^2(P_X, P_Y; K), \\
\mathbb{E}_1[T_{\mathsf{LF}}] &= \mathbb{E}_{X,Y,Z}[T_{\mathsf{LF}} | P_Z = P_Y] = -\mathrm{MMD}^2(P_X, P_Y; K).
\end{aligned}$$

Let $T_{\mathsf{TS}} = \mathrm{MMD}_u(\hat{P}_X, \hat{P}_Y)$ be the usual statistic that is thresholded for two-sample testing. Then, a computation analogous to that in Section D.2 show (cf. (13)) that

$$\begin{aligned}
\mathrm{var}_0(T_{\mathsf{LF}}) &\approx \frac{A(K)}{n} + \frac{A(K)}{m} + \frac{B(K)}{n^2} + \frac{B(K)}{mn}, \\
\mathrm{var}_0(T_{\mathsf{TS}}) &\approx \frac{A(K)}{n} + \frac{B(K)}{n^2}
\end{aligned}$$

for some $A(K)$ and $B(K)$. Therefore, we have approximately

$$J_{\mathsf{LF}} \approx \frac{2\,\mathrm{MMD}^2(P_X, P_Y; K)}{\sqrt{1 + \frac{n}{m}}\sqrt{\mathrm{var}_0(T_{\mathsf{TS}})}} \approx 2\sqrt{\frac{m}{m+n}}\,\widehat{J}(X, Y; K)$$

which only differs from our optimization objective defined in (7) by a constant factor.

Second, notice that $\frac{\mathrm{MMD}(P_X, P_Y; K)}{\sqrt{\mathrm{var}(T_{\mathsf{TS}})}}$ depends only on $P_X - P_Y$ and that $((1-\delta)P_X + \delta P_Y) - P_X \propto P_Y - P_X$, therefore it is sensible to use (7) as our training objective for is also sensible for (mLFHT), and we don't even need to observe the sample Z.

F.2 Proof of Proposition 4.1

Proof. In this proof we regard $\mathcal{D} \triangleq (X^{\mathsf{tr}}, X^{\mathsf{ev}}, Y^{\mathsf{tr}}, Y^{\mathsf{ev}})$ and the parameters of the kernel ω as fixed. Recall that we are looking at the problem mLFHT with a misspecification parameter $R = 0$ (see Theorem 3.2). Given a test set $\{z_i\}_{i\in[m]}$, our test statistic is $T(\{z_i\}_{i\in[m]}) = \frac{1}{m}\sum_{i=1}^{m} f(z_i)$ where

$$f(z_i) = \frac{1}{n_{\mathsf{ev}}} \sum_{j=1}^{n_{\mathsf{ev}}} \Big(K_\omega(z_i, Y_j^{\mathsf{ev}}) - K_\omega(z_i, X_j^{\mathsf{ev}}) \Big).$$

In Phase 3 of Algorithm 1, we observe the value $\widehat{T} = T(Z) = \frac{1}{m}\sum_{i=1}^{m} f(Z_i)$ and reject the null hypothesis for large values of $\widehat{T}$. Thus, the p-value is defined as

$$p = p(Z, \mathcal{D}) \triangleq \mathbb{P}_{\widetilde{Z}\sim P_X^{\otimes m}}(T(\widetilde{Z}) > \widehat{T}).$$

Phase 2 of our Algorithm 1 produces random variables $T_1, \ldots, T_k$ that all have the distribution of $T(\{\widetilde{Z}_i\}_{i\in[m]})$, so that $\mathbb{1}\{T_r \geq \widehat{T}\}$ $(r = 1, \ldots, k)$ are unbiased estimates of the p-value. However, the T_i are not independent, because they sample from the finite collection of calibration samples X^{cal}. However, as $n_{\mathsf{cal}} \to \infty$ the covariances between T_{r_1}, T_{r_2} for $r_1 \neq r_2$ tend to zero, and we obtain a consistent estimate of p. □

F.3 Proof of Proposition 4.2

Proof. The test statistic $T(X, Y, Z)$ in (3) is given by

$$T(X, Y, Z) = \frac{1}{m}\sum_{i=1}^{m} f_K(Z_i)$$

where

$$f_K(z) = \theta_{\widehat{P}_Y}(z) - \theta_{\widehat{P}_X}(z).$$

This simplifies to (consider $K(x, y) = f(x)f(y)$)

$$f_K(z) = \left(\frac{1}{n}\sum_{j=1}^{n} f(Y_j) - \frac{1}{n}\sum_{j=1}^{n} f(X_j) \right) f(z) = C(X, Y) f(z).$$

where $C(X, Y)$ does not depend on z. Therefore, for any witness function f, we obtain the desired additive test. □

F.4 Additive Test Statistics

In this section we prove accordingly that the test statistics of all of **MMD-M/G/O**, **SCHE**, **LBI**, **UME**, **RFM** are of the form $T_f(Z) = \frac{1}{m}\sum_{i=1}^{m} f(Z_i)$ (where f might depends on X, Y). The test is to compare $T_f(Z)$ with some threshold $\gamma(X, Y)$.

Note that in the setting of Algorithm 1, the X and Y here correspond to X^{ev} and Y^{ev}.

MMD-M/G/O As described in (3) we have

$$T_f(Z) = \frac{1}{m}\sum_{i=1}^{m} \left(\frac{1}{n}\sum_{j=1}^{n} \left(K(Z_i, Y_j) - K(Z_i, X_j) \right) \right).$$

SCHE As described in Section 4.2 we have

$$T_f(Z) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{\phi(Z_i) > t\}.$$

LBI As described in Section 4.2 we have

$$T_f(Z) = \frac{1}{m}\sum_{i=1}^{m} \log\left(\frac{\phi(Z_i)}{1 - \phi(Z_i)} \right).$$

UME As described in [6], the UME statistic evaluates the squared witness function at J_q test locations $W = \{w_k\}_{k=1}^{J_q} \subset \mathcal{X}$. Formally for any two distributions P, Q we define

$$U^2(P,Q) = \|\theta_Q - \theta_P\|^2_{L^2(W)} = \frac{1}{J_q}\sum_{k=1}^{J_q}(\theta_Q(w_k) - \theta_P(w_k))^2.$$

However, we note a crucial difference that their result only considers the case of $n = m$, and their proposed estimator for $U^2(P_Z, P_X)$ can not be naturally extended to the case of $n \neq m$. Here we generalize it to $m \neq n$ where we (conveniently) use a biased estimate of their distance. Given samples X, Y, Z and a set of witness locations W, the test statistic is a (biased yet) consistent estimator of $U^2(P_Z, P_Y) - U^2(P_Z, P_X)$. Let $\psi_W(z) = \frac{1}{\sqrt{J_q}}(K(z, w_1), \ldots, K(z, w_{J_q})) \in \mathbb{R}^{|W|}$ be the "feature function," then:

$$\begin{aligned}\widehat{U}^2(Z,X) &= \left\| \frac{1}{m}\sum_{i=1}^{m}\psi_W(Z_i) - \frac{1}{n}\sum_{j=1}^{n}\psi_W(X_i)\right\|_2^2 \\ &= \left\| \frac{1}{m}\sum_{i=1}^{m}\psi_W(Z_i)\right\|_2^2 + \left\| \frac{1}{n}\sum_{j=1}^{n}\psi_W(X_i)\right\|_2^2 - \frac{2}{mn}\sum_{1\le i\le m, 1\le j\le n}\langle \psi_W(Z_i), \psi_W(X_j)\rangle\end{aligned}$$

Here $\langle\cdot,\cdot\rangle$ denotes the usual inner product. Therefore, the difference between distances is

$$\widehat{U}^2(Z,Y) - \widehat{U}^2(Z,X) = \frac{1}{m}\sum_{i=1}^{m}\left\langle \psi_W(Z_i), \frac{2}{n}\sum_{j=1}^{n}(\psi_W(X_j) - \psi_W(Y_j))\right\rangle + F(X,Y)$$

where F is sum function based only on X, Y. This is clearly an additive statistic for Z.

RFM Algorithm 1 in [13] describes a method for learning a kernel from data given a binary classification task. For convenience lets concatenate the data to $X^{\text{RFM}} = (X, Y) \in \mathbb{R}^{2n\times d}$ and labels $y^{\text{RFM}} = (\vec{0}_n, \vec{1}_n) \in \mathbb{R}^{1\times 2n}$. Given a learned kernel K, we write the Gram matrix as $(K(X^{\text{RFM}}, X^{\text{RFM}}))_{i,j} = K(X_i^{\text{RFM}}, X_j^{\text{RFM}})$ $(1 \le i, j \le 2n)$. Let $K(X^{\text{RFM}}, z)$ be a column vector with components $K(X_i^{\text{RFM}}, z)$ $(1 \le i \le 2n)$. The classifier is then defined as

$$f^{\text{RFM}}(z) = y^{\text{RFM}} \cdot K(X^{\text{RFM}}, X^{\text{RFM}})^{-1} \cdot K(X^{\text{RFM}}, z). \tag{21}$$

Though in [13] the kernel learned from RFM is used to construct a classifier as in Equation (21), since RFM is a feature learning method, we also apply the RFM kernel to our MMD test, namely

$$f^{\text{RFM to MMD}}(z) = \frac{1}{n}\sum_{j=1}^{n}\left(K(z, Y_j) - K(z, X_j)\right).$$

G Application: Diffusion Models vs CIFAR

We defer a more fine-grained detail to our code submission, which includes executable programs (with PyTorch) once the data-generating script from DDPM has been run (see README in the ./codes/CIFAR folder).

G.1 Dataset Details

We use the CIFAR-10 dataset available online at https://www.cs.toronto.edu/~kriz/cifar.html, which contains 50000 colored images of size 32×32 with 10 classes. For the diffusion generated images, we use the SOTA Hugging Face model (DDPM) that can be found at https://huggingface.co/google/ddpm-CIFAR-10-32. We generated 10000 artificial images for our experiments. The code can be found at our code supplements.

For dataset balancing, we randomly shuffled the CIFAR-10 dataset and used 10000 images as data in our code. Most of our experiments are conducted with the null P_X as CIFAR images, and the alternate as $P_Y = \frac{2}{3}\cdot\text{CIFAR} + \frac{1}{3}\cdot\text{DDPM}$. To this end, we matched 20000 images from CIFAR to

Figure 4: Data visualization for CIFAR-10 (left) vs DDPM diffusion generated images (right)

belong to the alternate hypothesis, and the remaining 30000 images to stay in the null hypothesis. For the alternate dataset, we simply sample without replacement from the $20000 + 10000$ mixture. This sampled distribution is *almost* the same as mixing (so long as the sample bank is large enough compared to the acquired data, so that each item in the alternate has close to $1/3$ probability of being in DDPM, which is indeed the case).

G.2 Experiment Setup and Benchmarks

We use a standard deep Conv-net [11], which has been employed for SOTA GAN discriminator tasks in similar settings. It has four convolutional layers and one fully connected layer outputting the feature space of size $(300, 1)$. For SCHE and LBI, we simply added a linear layer of $(300, 2)$ after applying ReLU to the 300-dimensional layer and used the cross-entropy loss to train the network. Note that this is equivalent to first fixing the feature space and then performing logistic regression to the feature space. For kernels, we add extra trainable parameters after the 300-d feature output.

For the MMD-based tests, we simply train the kernel on the neural net and evaluate our objective. For UME, we used a slightly generalized version of the original statistic in [6] which allows for comparison on randomly selected witness locations in the null hypothesis with $m \neq n$ (see Appendix F.4). The kernel is trained using our heuristic (see (7) and Appendix F.1), with MMD replaced by UME. The formula for UME variance can be found in [6]. For RFM, we use Algorithm 1 in [13] to learn a kernel on (stochastic batched) samples, and then use our MMD test on the trained kernel.

We use 80 training epochs for most of our code from the CNN architecture (for classifiers, this is well after interpolating the training data and roughly when validation loss stops decreasing), and a batch size of 32 which has a slight empirical benefit compared to larger batch sizes. The learning rates are tuned separately in MMD methods for optimality, whereas for classifiers they follow the discriminator's original setting from [11]. In Phase 2 of Algorithm 1, we choose $k = 1000$ for the desired precision while not compromising runtime. For each task, we run 10 independent models and report their performances as the mean and standard deviation of those 10 runs as estimates. We refer to a full set of hyper-parameters in our code implementation.

Our code is implemented in Python 3.7 (PyTorch 1.1) and was ran on an NVIDIA RTX 3080 GPU equipped with a standard torch library and dataset extensions. Our code setup for feature extraction is similar to that of [10]. For benchmark implementations, our code follows from the original code templated provided by the cited papers.

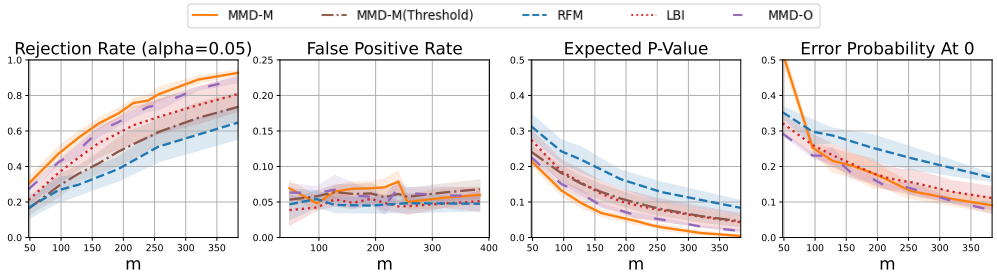

Figure 5: Relevant plots following the setting in Figure 2 (in the main text) of fixing $n_{\text{tr}} = 1920$ and varying sample size m in the x-axis for the comparison with missing benchmarks. Errorbars are projected showing standard deviation across 10 runs. We replaced part (d) in Figure 2 (in the main text) to a sanity check in our FPR when thresholded at $\alpha = 0.05$.

G.3 Sample Allocation

We make a comment on why (4) is *different* from just thresholding $\widehat{\text{MMD}}^2(Z, Y^{\text{tr}}) - \widehat{\text{MMD}}^2(Z, X^{\text{tr}})$ at 0, which was what we did in part (c) of Figure 2 (and hence the difference along the curve of MMD-M vs Figure 1). Our theory assumes that the samples are i.i.d. conditioned on the kernel being chosen already. However, in the experiments, the kernel is dependent on the training data. Therefore, to evaluate the MMD estimate (between experimentations), one needs extra data that does not intersect with training.

In fact, it can be experimentally shown by comparing Figure 1 and Figure 2(c) that doing so (while reducing the sample complexity on n_{ev}) hurts performance. Indeed, we found out that when $X^{\text{ev}}, Y^{\text{ev}}$ are non-intersecting with training, performance is (almost) always better at a cost of hurting the overall sample complexity of n.

G.4 Remarks on Results

Figure 5 lists all of our benchmarks in the setting of Figure 2 (in the main text) on missing benchmarks, where the last figure is replaced by the false positive rate at thresholding at $\alpha = 0.05$ to verify our results. As mentioned in the main text, our MMD-M method consistently outperforms other benchmarks on both the expected p-value (of alternate) and rejection rate at $\alpha = 0.05$, while all of our tests observe a empirical false positive rate close to $\alpha = 0.05\%$ (Part (b)), showing the consistency of methods.

H Application: Higgs-Boson Detection

H.1 Dataset Details

We use the Higgs dataset available online at `http://archive.ics.uci.edu/ml/datasets/HIGGS`, produced using Monte Carlo simulations [2]. The dataset is nearly balanced, containing $5,829,122$ signal instances and $5,170,877$ background instances. Each instance is a 28-dimensional vector, consisting of 28 features. The first 21 features are kinematic properties measured by the detectors in the accelerator, such as momentum and energy. The last 7 properties are *invariant masses*, derived from the first 21 features.

H.2 Experiment Setup and Training Models

The modified Algorithm 1 is shown in Algorithm 2 and Algorithm 3. Compared with Algorithm 2, we implement the thresholding trick (Section 4.3) in Algorithm 3.

H.2.1 Configuration and Model Architecture

We implement all methods in Python 3.9 and PyTorch 1.13 and run them on an NVIDIA Quadro RTX 8000 GPU.

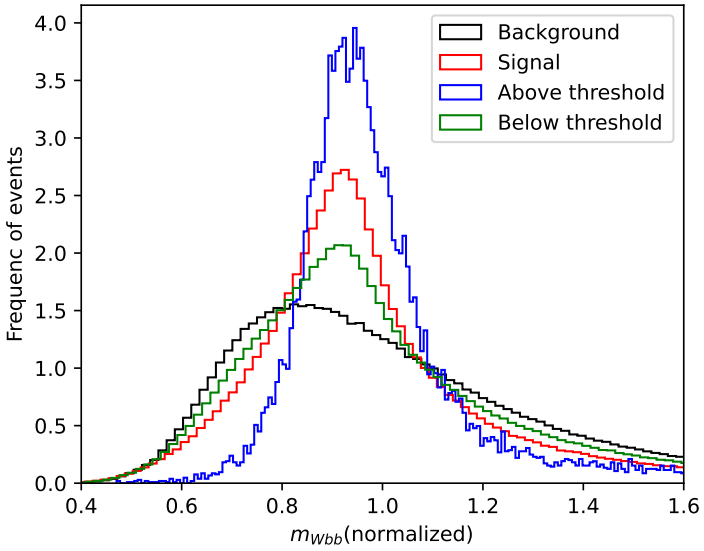

Figure 6: This figure visualizes the distribution of the 26th feature, the invariant mass m_{Wbb}. The red and black lines are the histograms of the original dataset. We employ MMD-M as a classifier, trained and evaluated using $n_{\text{tr}} = 1.3 \times 10^6$ and $n_{\text{ev}} = n_{\text{opt}} = 2 \times 10^4$ through Algorithm 3. The blue(green) line represents all instances z's whose "witness scores" $f(z; X^{\text{ev}}, Y^{\text{ev}})$'s are larger(smaller) than t_{opt}.

For all classifier-based methods in this study (SCHE and LBI), we adopt the same architecture as previously proposed in [2]. The classifiers are six-layer neural networks with 300 hidden units in each layer, all employing the tanh activation function. For SCHE, the output layer is a single sigmoid unit and we utilize the binary cross-entropy loss for training. For LBI, the output layer is a linear unit and we utilize the binary cross entropy loss combined with a logit function (which is more numerically stable than simply using a sigmoid layer followed by a cross entropy loss).

For all MMD-based methods (MMD-M, MMD-G, MMD-O, and UME), the networks φ and φ' are both six-layer neural networks with 300 ReLU units in each layer. The feature space, which is the output of the neural network φ, is set to be 100-dimensional. Here UME has the same kernel architecture as MMD-M, and the number of test locations is set to be $J_q = 4096$. For RFM, we adopt the same architecture as in [13], where the kernel is $K_M(x, y) = \exp(-\gamma(x-y)^T M(x-y))$ with a constant γ and a learnable positive semi-definite matrix M. We set $\gamma \equiv 1$.

The neural networks are initialized using the default setting in PyTorch, and the bandwidths σ, σ' are initialized using the *median heuristic* [4]. The parameter τ is initially set to 0.5. For UME, the witness locations W are initially randomly sampled from the training set. For RFM, the initial M equals the median bandwidth times an identity matrix.

H.2.2 Training

The size of our training set, denoted as n_{tr}, varies from 1.0×10^2 to 1.6×10^6. For a given n_{tr}, we select the first n_{tr} datapoints from each class of the Higgs dataset to form X^{tr} and Y^{tr}, i.e., $|X^{\text{tr}}| = |Y^{\text{tr}}| = n_{\text{tr}}$. Subsequently, we randomly select $n_{\text{validation}} = \min(\sqrt{10 n_{\text{tr}}}, 0.1 n_{\text{tr}})$ points from each of $X_{\text{tr}}, Y_{\text{tr}}$ to constitute the validation set, while the remainder of $X_{\text{tr}}, Y_{\text{tr}}$ are used for running gradient descent. The optimizer is set to be a minibatch SGD, with a batch size of 1024, a learning rate of 0.001, and a momentum of 0.99. Training is halted once the validation loss stops to decrease for 10 epochs, then we choose the checkpoint (saved for each epoch) with the smallest validation loss thus far as our trained model. Beyond the general setting above, in RFM a batch size of 1024 doesn't work well and instead we use a batch size of $20,000$.

H.3 Evaluating the Performance

H.3.1 Evaluating the p-Value with the Methodology of Algorithm 1

We call the "witness score" of an instance $z \in \mathcal{X}$ as

$$f(z; X^{\mathsf{ev}}, Y^{\mathsf{ev}}) = \frac{1}{n_{\mathsf{cal}}} \sum_{i=1}^{n_{\mathsf{cal}}} \left(k(z, Y_i^{\mathsf{ev}}) - k(z, X_i^{\mathsf{ev}}) \right). \tag{22}$$

For a vector of instances $Z = (Z_1, \ldots, Z_m)$, we write

$$f(Z; X^{\mathsf{ev}}, Y^{\mathsf{ev}}) = (f(Z_1; X^{\mathsf{ev}}, Y^{\mathsf{ev}}), \ldots, f(Z_m; X^{\mathsf{ev}}, Y^{\mathsf{ev}})).$$

The testing procedure is summarized in Phases 2, 3 and 4 in Algorithm 2 and Algorithm 3. In the Higgs experiment, we utilize the Gaussian approximation method to determine the p-values when the witness function f is not thresholded, which allows us to reach very small p-values and errors under limited computational resource. In cases where the score function f is thresholded by a value t, using the Binomial distribution as in Algorithm 3 is more precise and also fast enough.

Given a trained kernel K trained on X^{tr} and Y^{tr}, we set $X^{\mathsf{ev}} = X^{\mathsf{tr}}$ and $Y^{\mathsf{ev}} = Y^{\mathsf{tr}}$, and accordingly $n_{\mathsf{ev}} = n_{\mathsf{tr}}$. This results in a more efficient use of data (since we reuse $X^{\mathsf{tr}}, Y^{\mathsf{tr}}$ also as $X^{\mathsf{ev}}, Y^{\mathsf{ev}}$). Then, out of the untouched portion of the data, we randomly choose $n_{\mathsf{cal}} = 20,000$ datapoints from both classes to populate X^{cal} and Y^{cal}, i.e., $|X^{\mathsf{cal}}| = |Y^{\mathsf{cal}}| = n_{\mathsf{cal}} = 20,000$. In addition to the general setting above, for RFM, we need to solve a $2n_{\mathsf{ev}}$-dimensional linear equation during inference, which arises from the inverse matrix in Equation (21) (solving $K(X^{\mathrm{RFM}}, X^{\mathrm{RFM}})\boldsymbol{u} = (y^{\mathrm{RFM}})^T$ for $\boldsymbol{u} \in \mathbb{R}^{2n_{\mathsf{ev}}}$). So we set $n_{\mathsf{ev}} = \min(n_{\mathsf{tr}}, 10,000)$ that $X_{\mathsf{ev}}, Y_{\mathsf{ev}}$ are randomly sampled from the training set.

In order to compare different benchmarks, we evaluate the expected significance of discovery on a mixture of 1000 backgrounds and 100 signals. For each benchmark and each n_{tr}, we train 10 independent models. Then for each trained model we proceed through the Phases 2, 3 (and 4) in Algorithm 2 and Algorithm 3 by 10 times for 10 different $(X^{\mathsf{ev}}, X^{\mathsf{cal}}, X^{\mathsf{opt}}, Y^{\mathsf{ev}}, Y^{\mathsf{cal}}, Y^{\mathsf{opt}})$. The mean and standard deviation from these 100 runs are reported in Figure 7.

We also display in Figure 8 the trade-off b (m, n_{ev}) and (m, n_{tr}) to reach certain levels of significance of discovery in MMD-M. From the bottom left plot, we see that the (averaged) significance is not sensitive to n_{ev} when $\lg n_{\mathsf{ev}}$ is large. So taking $n_{\mathsf{ev}} = 20,000$ is sufficient.

H.3.2 Evaluating the Error of the Test (4)

We set the parameters to be $\delta = 0.1$ and $\pi = \frac{1}{2}\delta$ in our experiments. As explained Appendix G.3, here we no longer take $X^{\mathsf{ev}} = X^{\mathsf{tr}}$. Empirically, taking $X^{\mathsf{ev}} = X^{\mathsf{tr}}$ yields a very bad threshold $\gamma(X^{\mathsf{ev}}, Y^{\mathsf{ev}}, \pi)$.[4] Instead, X^{ev} is sampled from untouched datapoints other than X^{tr}, and the same applies for Y. We still take $n_{\mathsf{ev}} = n_{\mathsf{tr}}$ here, resulting in a total size of $n_{\mathsf{ev}} + n_{\mathsf{tr}} = 2n_{\mathsf{tr}}$. Specifically, when $n_{\mathsf{ev}} \geq 10,000$, computing a $n_{\mathsf{ev}} \times n_{\mathsf{ev}}$ Gram matrix becomes computationally expensive, so we adopt Monte Carlo method to compute $\gamma(X^{\mathsf{ev}}, Y^{\mathsf{ev}}, \pi)$, in which we subsample $10,000$ points from X^{ev} and Y^{ev} to calculate γ and repeat this process 100 times.

Again, we utilize the Gaussian approximation. Recall that the test is to compare $T = \frac{1}{m}\sum_{i=1}^{m} f(Z_i)$ with γ. The type 1 and type 2 error are estimated as $\mathrm{CDF}_{\mathcal{N}(0,1)}\left(-\frac{\gamma(X^{\mathsf{ev}}, Y^{\mathsf{ev}}, \pi) - \mathbb{E}[f|H_0]}{\sqrt{\mathrm{var}(f|H_0)/m}}\right)$ and $\mathrm{CDF}_{\mathcal{N}(0,1)}\left(-\frac{\mathbb{E}[f|H_1] - \gamma(X^{\mathsf{ev}}, Y^{\mathsf{ev}}, \pi)}{\sqrt{\mathrm{var}(f|H_1)/m}}\right)$ for the witness function f, which can be estimated efficiently using the calibration samples $X^{\mathsf{cal}}, Y^{\mathsf{cal}}$.

We consider both the regimes of fixing kernels and varying kernels (training kernel based on n). The results are shown in the top plot in Figure 1 and the top plot in Figure 8. For each point on the plot, we train 30 independent models and test each model 10 times, and report the average of these 300 runs. In both plots, we observe the asymmetric m vs n trade-off.

[4] If the kernel $K(\cdot,\cdot) = K_{X^{\mathsf{tr}}, Y^{\mathsf{tr}}}(\cdot,\cdot)$ is independent of $X^{\mathsf{ev}}, Y^{\mathsf{ev}}$, then we have $\gamma(X^{\mathsf{ev}}, Y^{\mathsf{ev}}, \delta/2) \approx \frac{1}{2}\left(\mathbb{E}_{Z\sim P_X}[T(X^{\mathsf{ev}}, Y^{\mathsf{ev}}, Z)] + \mathbb{E}_{Z\sim \delta P_Y + (1-\delta)P_X}[T(X^{\mathsf{ev}}, Y^{\mathsf{ev}}, Z)]\right)$. However this is no longer true if $(X^{\mathsf{tr}}, Y^{\mathsf{tr}})$ and $(X^{\mathsf{ev}}, Y^{\mathsf{ev}})$ intersect.

Algorithm 2 Estimate the significance of discovery of an input Z_{test}, using the original statistic

Input: $(X^{\mathsf{tr}}, X^{\mathsf{ev}}, X^{\mathsf{cal}})$, $(Y^{\mathsf{tr}}, Y^{\mathsf{ev}}, Y^{\mathsf{cal}})$; parametrized kernel K_ω; input Z_{test}.
 # Phase 1: Kernel training on X^{tr} and Y^{tr}
$\omega \leftarrow \arg\max_\omega^{\text{optimizer}} \hat{J}(X^{\mathsf{tr}}, Y^{\mathsf{tr}}; K_w)$ *# maximize objective $\widehat{J}(X^{\mathsf{tr}}, Y^{\mathsf{tr}}; K_\omega)$ as in (7)*
 # Phase 2: Distributional calibration of test statistic
$\text{Scores}^{(0)} \leftarrow f(X^{\mathsf{cal}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}})$ *# Scores$^{(0)}$ has a length of n_{cal}*
$\text{Scores}^{(1)} \leftarrow f(Y^{\mathsf{cal}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}})$ *# Scores$^{(1)}$ has a length of n_{cal}*
$\theta_0 \leftarrow \text{mean}(\text{Scores}^{(0)})$ *# estimate $\mathbb{E}[f(Z)|Z \sim P_X]$*
$\theta_1 \leftarrow \text{mean}(\text{Scores}^{(1)})$ *# estimate $\mathbb{E}[f(Z)|Z \sim P_Y]$*
$\sigma_0 \leftarrow \text{std}(\text{Scores}^{(0)})$ *# estimate $\sqrt{\text{var}[f(Z)|Z \sim P_X]}$*
 # Phase 3: Inference with input Z_{test}
$m \leftarrow \text{length}(Z_{\text{test}})$
$T \leftarrow T_f(Z_{\text{test}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}}) = \text{mean}(f(Z_{\text{test}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}}))$ *# compute test statistic*
$Z_{\text{discovery}} \leftarrow \frac{T-\theta_0}{\sigma_0/\sqrt{m}}$
Output: Estimated significance: $Z_{\text{discovery}}$

Algorithm 3 Estimate the significance of discovery of an input Z_{test}, applying the thresholding trick

Input: $(X^{\mathsf{tr}}, X^{\mathsf{ev}}, X^{\mathsf{cal}}, X^{\mathsf{opt}})$, $(Y^{\mathsf{tr}}, Y^{\mathsf{ev}}, Y^{\mathsf{cal}}, Y^{\mathsf{opt}})$; parametrized kernel K_ω; input Z_{test}.
 # Phase 1: Kernel training on X^{tr} and Y^{tr}
$\omega \leftarrow \arg\max_\omega^{\text{optimizer}} \hat{J}(X^{\mathsf{tr}}, Y^{\mathsf{tr}}; K_w)$ *# maximize objective $\widehat{J}(X^{\mathsf{tr}}, Y^{\mathsf{tr}}; K_\omega)$ as in (7)*
 # Phase 2: Find the best threshold
$\text{Scores}^{(0)} \leftarrow f(X^{\mathsf{opt}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}})$
$\text{Scores}^{(1)} \leftarrow f(Y^{\mathsf{opt}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}})$ *# witness function as in (22)*
for $i = 1, 2, ..., 2n_{\mathsf{opt}}$ **do**
  $t = (\text{Scores}^{(0)} \cup \text{Scores}^{(1)})[i]$
  $\text{TP}, \text{TN} = \text{mean}(\text{Scores}^{(1)} > t), \text{mean}(\text{Scores}^{(0)} < t)$ *# true positive and true negative rate*
  $\text{power}_i = \frac{\text{TP}+\text{TN}-1}{\sqrt{\text{TN}(1-\text{TN})}}$ *# find t to maximize the (estimated) p-value*
end for
$t_{\mathsf{opt}} = (\text{Scores}^{(0)} \cup \text{Scores}^{(1)})[\arg\max_i \text{power}_i]$
 # Phase 3: Distributional calibration of test statistic (under null hypothesis)
$\text{Scores}^{(0)} \leftarrow (f(X^{\mathsf{cal}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}}) > t)$ *# Scores$^{(0)} \in \{0,1\}^{n_{\mathsf{ev}}}$*
$\text{Scores}^{(1)} \leftarrow (f(Y^{\mathsf{cal}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}}) > t)$ *# Scores$^{(1)} \in \{0,1\}^{n_{\mathsf{ev}}}$*
$\theta_0 \leftarrow \text{mean}(\text{Scores}^{(0)})$ *# estimate $\mathbb{E}[f_t(Z)|Z \sim P_X] \in [0,1]$*
$\theta_1 \leftarrow \text{mean}(\text{Scores}^{(1)})$ *# estimate $\mathbb{E}[f_t(Z)|Z \sim P_Y] \in [0,1]$*
 # Phase 4: Inference with input Z_{test}
$m \leftarrow \text{length}(Z_{\text{test}})$
$T \leftarrow T_f(Z_{\text{test}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}}) = \text{mean}(f(Z_{\text{test}}; X^{\mathsf{ev}}, Y^{\mathsf{ev}}) > t)$ *# compute test statistic*
$Z_{\text{discovery}} \leftarrow \text{CDF}^{-1}_{\mathcal{N}(0,1)}(\text{CDF}_{\text{Bin}(m,\theta_0)}(T))$
Output: Estimated significance: $Z_{\text{discovery}}$

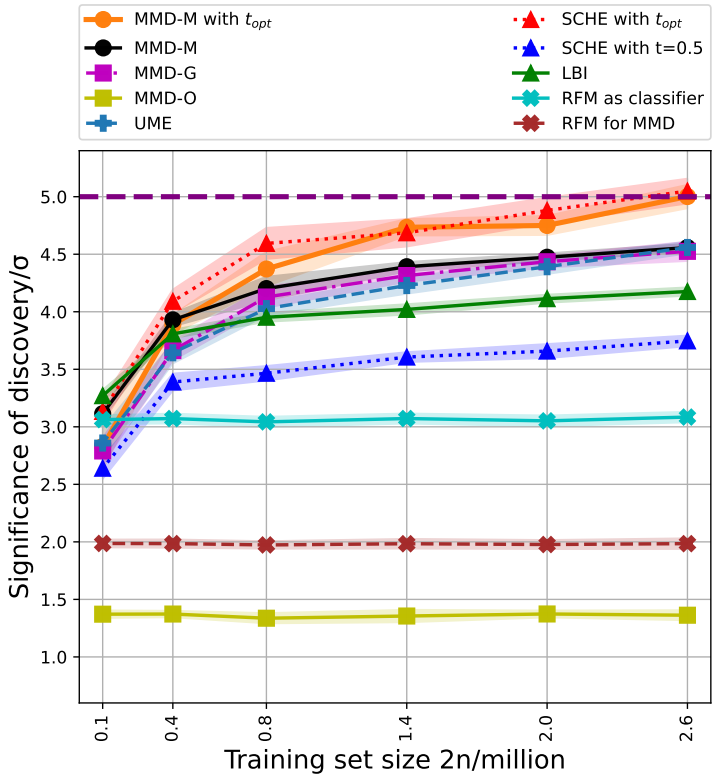

Figure 7: Complete image of Figure 1 in the main text. The mean and standard deviation are calculated based on 100 runs. See Appendix H for details.

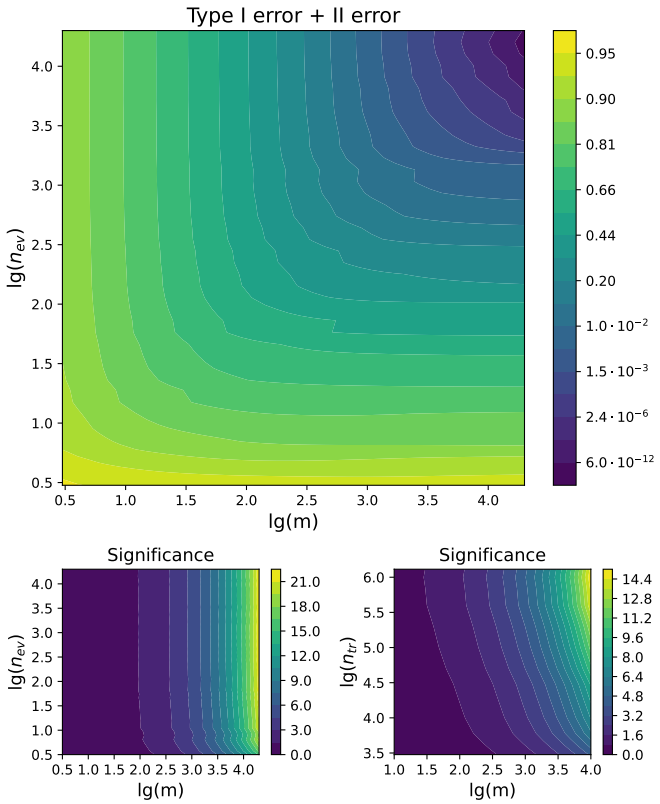

Figure 8: The top plot displays the (m, n_{ev}) trade-off to reach certain levels of total error using $n_{\text{tr}} = 1.3 \times 10^6$ in MMD-M. The bottom figures show the trade-off of (m, n_{ev}) and (m, n_{tr}) to reach certain level of significance of discovery in MMD-M. In the bottom left figure, we fix $n_{\text{tr}} = 1.3 \times 10^6$. In the bottom right figure, we fix $n_{\text{ev}} = 20,000$. See Appendix H for details.

References

[1] Arias-Castro, E., Pelletier, B., and Saligrama, V. Remember the curse of dimensionality: The case of goodness-of-fit testing in arbitrary dimension. *Journal of Nonparametric Statistics*, 30 (2):448–471, 2018.

[2] Baldi, P., Sadowski, P., and Whiteson, D. Searching for exotic particles in high-energy physics with deep learning. *Nature communications*, 5(1):1–9, 2014.

[3] Gerber, P. R. and Polyanskiy, Y. Likelihood-free hypothesis testing. *CoRR*, abs/2211.01126, 2022. doi: 10.48550/arXiv.2211.01126.

[4] Gretton, A., Sejdinovic, D., Strathmann, H., Balakrishnan, S., Pontil, M., Fukumizu, K., and Sriperumbudur, B. K. Optimal kernel choice for large-scale two-sample tests. *Advances in neural information processing systems*, 25, 2012.

[5] Ingster, Y. I. Minimax testing of nonparametric hypotheses on a distribution density in the l_p metrics. *Theory of Probability & Its Applications*, 31(2):333–337, 1987. doi: 10.1137/1131042.

[6] Jitkrittum, W., Kanagawa, H., Sangkloy, P., Hays, J., Schölkopf, B., and Gretton, A. Informative features for model comparison. *Advances in Neural Information Processing Systems*, 31, 2018.

[7] Kelly, B. G., Tularak, T., Wagner, A. B., and Viswanath, P. Universal hypothesis testing in the learning-limited regime. In *2010 IEEE International Symposium on Information Theory*, pages 1478–1482. IEEE, 2010.

[8] Kelly, B. G., Wagner, A. B., Tularak, T., and Viswanath, P. Classification of homogeneous data with large alphabets. *IEEE transactions on information theory*, 59(2):782–795, 2012.

[9] Li, T. and Yuan, M. On the optimality of gaussian kernel based nonparametric tests against smooth alternatives. *arXiv preprint arXiv:1909.03302*, 2019.

[10] Liu, F., Xu, W., Lu, J., Zhang, G., Gretton, A., and Sutherland, D. J. Learning deep kernels for non-parametric two-sample tests. In *International conference on machine learning*, pages 6316–6326. PMLR, 2020.

[11] Lopez-Paz, D. and Oquab, M. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

[12] Polyanskiy, Y. and Wu, Y. *Information Theory: From Coding to Learning*. Cambridge University Press, 2023+.

[13] Radhakrishnan, A., Beaglehole, D., Pandit, P., and Belkin, M. Feature learning in neural networks and kernel machines that recursively learn features. *arXiv preprint arXiv:2212.13881*, 2022.

[14] Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519.