Compress Then Test: Powerful Kernel Testing in Near-linear Time

Carles Domingo-Enrich
Courant Institute of Mathematical Sciences
NYU
cd2754@nyu.edu

Raaz Dwivedi
Harvard University
MIT
raaz@mit.edu

Lester Mackey
Microsoft Research New England
lmackey@microsoft.com

Abstract

Kernel two-sample testing provides a powerful framework for distinguishing any pair of distributions based on n sample points. However, existing kernel tests either run in n^2 time or sacrifice undue power to improve runtime. To address these shortcomings, we introduce Compress Then Test (CTT), a new framework for high-powered kernel testing based on sample compression. CTT cheaply approximates an expensive test by compressing each n point sample into a small but provably high-fidelity coreset. For standard kernels and subexponential distributions, CTT inherits the statistical behavior of a quadratic-time test—recovering the same optimal detection boundary—while running in near-linear time. We couple these advances with cheaper permutation testing, justified by new power analyses; improved time-vs.-quality guarantees for low-rank approximation; and a fast aggregation procedure for identifying especially discriminating kernels. In our experiments with real and simulated data, CTT and its extensions provide 20–200x speed-ups over state-of-the-art approximate MMD tests with no loss of power.

1 Introduction

Kernel two-sample tests based on the maximum mean discrepancy (MMD, Gretton et al., 2012a) can distinguish any pair of distributions given only a sufficiently large sample from each. However, standard MMD tests have prohibitive running times that scale quadratically in the sample size n. Gretton et al. (2012a); Zaremba et al. (2013); Yamada et al. (2019); Schrab et al. (2022) introduced faster approximate MMD tests based on subsampling, but each suffers from a fundamental time-quality trade-off barrier: for any pair of distributions, quadratic time is required to match the discrimination power of a standard MMD test (see Prop. 2). Our first contribution is a new subsampling approach called Compress Then Test (CTT) that accelerates testing by first compressing each sample. In Sec. 3, we prove that this approach pierces the aforementioned barrier, matching the quality of quadratic-time tests in near-linear time for subexponential distributions. Along the way, we develop refined analyses of permutation tests, establishing their discriminating power even when permutations are restricted to preserve group structure and relatively few (e.g., 39) permutations are employed. In our experiments with both real and synthetic data, the CTT time-quality trade-off curves dominate those of state-of-the-art subsampling approaches, providing $200\times$ speed-ups.

Zhao and Meng (2015) introduced an alternative, low-rank approach to fast approximate MMD testing that replaces the target kernel with a $\Theta(nr)$ time approximation based on r random Fourier features (RFFs, Rahimi and Recht, 2008). This method often performs well in practice, but the guarantees of Rahimi and Recht (2008); Zhao and Meng (2015); Sutherland and Schneider (2015); Sriperumbudur and Szabó (2015) require $\Omega(n^2)$ random features and hence $\Omega(n^3)$ time to match the power of a standard MMD test. By compressing before performing low-rank approximation, our second contribution, Low-Rank CTT (LR-CTT), allows a user to harness any effective low-rank approximation without sacrificing the improved time-quality guarantees of CTT. In our experiments, this hybrid test offers the best performance of all, outpacing both the CTT and RFF tests.

Finally, in the spirit of Schrab et al. (2021), we develop Aggregated CTT (ACTT) tests that improve power by rapidly identifying the most discriminating kernel in a collection of candidates. In our experiments, ACTT offers 100-200$\times$ speed-ups over the state-of-the-art efficient aggregated tests of Schrab et al. (2022).

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

2 Kernel Two-sample Testing

As a standing assumption, suppose that we observe $\mathbb{X}_m \triangleq (X_i)_{i=1}^m$ and $\mathbb{Y}_n \triangleq (Y_j)_{j=1}^n$, two independent sequences of datapoints drawn i.i.d. from unknown probability measures $\mathbb{P}$ and $\mathbb{Q}$ respectively. In two-sample testing, our goal is to decide whether the null hypothesis $\mathcal{H}_0 : \mathbb{P} = \mathbb{Q}$ or the alternative hypothesis $\mathcal{H}_1 : \mathbb{P} \neq \mathbb{Q}$ is correct. A test Δ is a binary function of $\mathbb{X}_m$ and $\mathbb{Y}_n$ such that the null hypothesis $\mathcal{H}_0$ is rejected if and only if $\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1$. The *size* or Type I error of the test is the probability that the null hypothesis is rejected when it is true, i.e., the probability $\Pr[\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1]$ when $\mathbb{P} = \mathbb{Q}$. A test is said to have *level* $\alpha \in (0, 1)$ if its Type I error is bounded by α for all probability distributions, i.e., $\sup_{\mathbb{P}=\mathbb{Q}} \Pr[\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1] \leq \alpha$. The Type II error of a test for a specific choice of $\mathbb{P} \neq \mathbb{Q}$ is the probability that the null hypothesis is accepted, i.e., $\Pr[\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0]$. For a given level α, our aim is to build a test with Type II error as small as possible for alternatives $\mathbb{Q}$ that are not too similar to $\mathbb{P}$. If $\Pr[\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 0] = \beta$, we say that the test has *power* $1 - \beta$ against the alternative $\mathbb{Q}$.

Kernel two-sample tests are popular because they can distinguish any pair of distributions given sufficiently large samples and a characteristic kernel $\mathbf{k}$ (Gretton et al., 2012a). A *characteristic kernel* is any positive-definite function $\mathbf{k}(x, y)$ (Steinwart and Christmann, 2008, Def. 4.15) satisfying $\mathbb{E}_{X\sim\mathbb{P}}\mathbf{k}(X, x) \neq \mathbb{E}_{Y\sim\mathbb{Q}}\mathbf{k}(Y, x)$ for some x whenever $\mathbb{P} \neq \mathbb{Q}$. Common examples include Gaussian, Matérn, B-spline, inverse multiquadric (IMQ), sech, and Wendland's compactly supported kernels on $\mathbb{R}^d$ (Dwivedi and Mackey, 2021). Kernel two-sample tests take the form $\Delta(\mathbb{X}_m,\mathbb{Y}_n) = \mathbf{1}[T(\mathbb{X}_m,\mathbb{Y}_n) > t_\alpha]$ where the test statistic $T(\mathbb{X}_m,\mathbb{Y}_n)$ is an estimate of the squared *maximum mean discrepancy* (MMD) between $\mathbb{P}$ and $\mathbb{Q}$,

$$\mathrm{MMD}^2_{\mathbf{k}}(\mathbb{P}, \mathbb{Q}) \triangleq \mathbb{E}_{X,X' \overset{\text{i.i.d.}}{\sim} \mathbb{P}}\mathbf{k}(X, X') + \mathbb{E}_{Y,Y' \overset{\text{i.i.d.}}{\sim} \mathbb{Q}}\mathbf{k}(Y, Y') - 2\mathbb{E}_{X\sim\mathbb{P}\perp\!\!\!\perp Y\sim\mathbb{Q}}\mathbf{k}(X, Y),$$

and t_α is a threshold chosen to ensure that the test has either finite-sample or asymptotic level α.

Quadratic-time or complete MMD tests The standard MMD test statistics defined in Gretton et al. (2012a) each require $\Theta(m^2 + n^2)$ kernel evaluations and hence computation that grows quadratically in the sample sizes. For example, Gretton et al. (2012a, Sec. 4.1) defines the squared sample MMD test statistic,

$$\mathrm{MMD}^2_{\mathbf{k}}(\mathbb{X}_m, \mathbb{Y}_n) \triangleq \tfrac{1}{m^2} \textstyle\sum_{1\leq i,i'\leq m} \mathbf{k}(X_i, X_{i'}) + \tfrac{1}{n^2} \textstyle\sum_{1\leq j,j'\leq n} \mathbf{k}(Y_j, Y_{j'}) - \tfrac{2}{mn} \textstyle\sum_{i=1}^m\sum_{j=1}^n \mathbf{k}(X_i, Y_j).$$

Gretton et al. (2012a, Lem. 6) also presents two unbiased estimators of $\mathrm{MMD}^2_{\mathbf{k}}(\mathbb{P}, \mathbb{Q})$ as test statistics:

$$\mathrm{MMD}^2_{u}(\mathbb{X}_m, \mathbb{Y}_n) \triangleq \tfrac{\sum_{1\leq i\neq i'\leq m} \mathbf{k}(X_i,X_{i'})}{m(m-1)} + \tfrac{\sum_{1\leq j\neq j'\leq n} \mathbf{k}(Y_j,Y_{j'})}{n(n-1)} - \tfrac{2\sum_{i=1}^m\sum_{j=1}^n \mathbf{k}(X_i,Y_j)}{mn}, \quad (1)$$

$$\mathrm{MMD}^2_{\text{up}}(\mathbb{X}_n, \mathbb{Y}_n) \triangleq \tfrac{\sum_{1\leq i\neq j\leq n} \mathbf{h}(X_i,X_j,Y_i,Y_j)}{n(n-1)}, \quad (2)$$

where $\mathbf{h}(x, x', y, y') = \mathbf{k}(x, x') + \mathbf{k}(y, y') - \mathbf{k}(x, y') - \mathbf{k}(x', y)$. The estimator (2) differs from the estimator (1) as it omits the diagonal cross-terms and is defined only when $m = n$.

Block MMD tests To improve computational cost through subsampling, Zaremba et al. (2013) introduced block MMD tests, or *B-tests* for short, that average $\frac{n}{B}$ independent instances of the quadratic estimator (2), each with sample size B, i.e.,

$$\mathrm{MMD}^2_B(\mathbb{X}_n, \mathbb{Y}_n) \triangleq \tfrac{B}{n} \textstyle\sum_{i=1}^{\frac{n}{B}} \eta_i(\mathbb{X}_n, \mathbb{Y}_n) \text{ with} \quad (3)$$
$$\eta_i(\mathbb{X}_n, \mathbb{Y}_n) \triangleq \tfrac{1}{B(B-1)} \textstyle\sum_{j,k=(i-1)B+1, j\neq k}^{iB} \mathbf{h}(X_j, X_k, Y_j, Y_k).$$

Consequently, the statistic computation takes time $\Theta(nB)$. Moreover, when $\frac{n}{B} \to \infty$, $\sqrt{\frac{n}{B}}\,\mathrm{MMD}^2_B$ has a Gaussian limit under the null that can be estimated to set t_α. Previously, Gretton et al. (2012a, Sec. 6) studied a particular instantiation of this test with $B = 2$.

Incomplete MMD tests Yamada et al. (2019) introduced an alternative $\Theta(\ell)$ time subsampling approximation based on *incomplete* MMD test statistics, $\mathrm{MMD}^2_{\text{inc}}(\mathbb{X}_n, \mathbb{Y}_n) \triangleq \frac{1}{\ell} \sum_{(i,j)\in\mathcal{D}} \mathbf{h}(X_i, X_j, Y_i, Y_j)$, with $\mathcal{D}$ a collection of ℓ ordered index pairs. Yamada et al. (2019) sampled pairs uniformly with replacement and set t_α using the Gaussian limit of $\sqrt{\ell}\,\mathrm{MMD}^2_{\text{inc}}$ as $\ell \to \infty$. Schrab et al. (2022) instead used deterministically pre-selected index pairs and a wild bootstrap setting of t_α described below.

Low-rank RFF tests Zhao and Meng (2015) proposed a complementary speed-up for MMD testing based on a low-rank MMD approximation of the form

$$\mathrm{MMD}^2_{\Phi_r}(\mathbb{X}_m, \mathbb{Y}_n) \triangleq \left\| \tfrac{1}{m} \textstyle\sum_{i=1}^m \Phi_r(x_i) - \tfrac{1}{n} \textstyle\sum_{i=1}^n \Phi_r(y_i) \right\|_2^2 \quad (4)$$

where Φ_r maps each sample point to an r-dimensional feature vector. Specifically, Zhao and Meng chose r RFFs to unbiasedly estimate $\mathrm{MMD}^2_{\mathbf{k}}$ in $\Theta((m + n)r)$ time.

Permutation tests For any of the aforementioned test statistics, one can alternatively set t_α using the following permutation approach to obtain a test with non-asymptotic level α (Romano and Wolf, 2005; Fromont et al., 2012). Let $\mathbb{U}$ be the concatenation of $\mathbb{X}_m$ and $\mathbb{Y}_n$. For each permutation σ of the indices $\{1, \ldots, m + n\}$, define the permuted samples $\mathbb{X}^\sigma_m = (U_{\sigma(i)})_{i=1}^m$, $\mathbb{Y}^\sigma_n = (U_{\sigma(m+j)})_{j=1}^n$ and the permuted statistic as $T^\sigma \triangleq T(\mathbb{X}^\sigma_m, \mathbb{Y}^\sigma_n)$. Sample $\mathcal{B}$ uniformly random permutations $(\sigma_b)_{b=1}^{\mathcal{B}}$ to obtain the values $T_b \triangleq T^{\sigma_b}$ and sort them in increasing order $(T_{(b)})_{b=1}^{\mathcal{B}}$. Finally, set $t_\alpha = T_{(\lceil(1-\alpha)(\mathcal{B}+1)\rceil)}$.

Wild bootstrap tests Similarly, when $m = n$, the following wild bootstrap approach employed by Fromont et al. (2012) yields a non-asymptotic level α by exchangeability and Romano and Wolf (2005, Lem. 1). For each vector $\epsilon \in \{\pm 1\}^n$, define $T^{\epsilon} \triangleq T(\mathbb{X}_n^{\epsilon}, \mathbb{Y}_n^{\epsilon})$ where $\mathbb{X}_n^{\epsilon_b}, \mathbb{Y}_n^{\epsilon_b}$ are constructed from $\mathbb{X}_n$ and $\mathbb{Y}_n$ by swapping X_i and Y_i if $\epsilon_i = -1$. Sample $\mathcal{B}$ i.i.d. vectors $(\epsilon_b)_{b=1}^{\mathcal{B}}$ uniformly from $\{\pm 1\}^n$, compute the values $T_b \triangleq T^{\epsilon_b}$, and finally set t_α as in the permutation approach.

3 Compress Then Test

This section introduces Compress Then Test (CTT), a new framework for testing with sample compression. CTT relies on a new test statistic, CORESETMMD, that we describe and analyze in Sec. 3.1. Sec. 3.2 then provides an analysis of the complete CTT procedure detailed in Alg. 1.

Algorithm 1: Compress Then Test, Δ_{CTT}

Input: Samples $(\mathbb{X}_m, \mathbb{Y}_n)$, # coresets s, compression level $\mathfrak{g}$, kernels $(\mathbf{k}, \mathbf{k}')$, failure prob. δ, # replicates $\mathcal{B}$, level α

Partition $\mathbb{X}_m$ into $s_m = \frac{sm}{m+n}$ equal-sized bins $(\mathbb{X}_m^{(i)})_{i=1}^{s_m}$
Partition $\mathbb{Y}_n$ into $s_n = \frac{sn}{m+n}$ equal-sized bins $(\mathbb{Y}_n^{(i)})_{i=1}^{s_n}$

// Identify coreset of size $2^{\mathfrak{g}}\sqrt{\frac{m+n}{s}}$ for each bin
for $i = 1, \dots, s_m$ **do**
  $\hat{\mathbb{X}}_m^{(i)} \leftarrow$ KT-COMPRESS$(\mathbb{X}_m^{(i)}, \mathfrak{g}, \mathbf{k}, \mathbf{k}', \delta)$
end
for $i = 1, \dots, s_n$ **do**
  $\hat{\mathbb{Y}}_n^{(i)} \leftarrow$ KT-COMPRESS$(\mathbb{Y}_n^{(i)}, \mathfrak{g}, \mathbf{k}, \mathbf{k}', \delta)$
end

// Compute CORESETMMD test statistic
$M_{\mathcal{B}+1} \leftarrow \mathrm{MMD}_{\mathbf{k}}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)$ for (5)
$\hat{\mathbb{X}}_m \coloneqq$ CONCAT$((\hat{\mathbb{X}}_m^{(i)})_{i=1}^{s_m})$ and $\hat{\mathbb{Y}}_n \coloneqq$ CONCAT$((\hat{\mathbb{Y}}_n^{(i)})_{i=1}^{s_n})$

// Simulate null by randomly permuting the s coresets $\mathcal{B}$ times
for $b = 1, \dots, \mathcal{B}$ **do**
  $(\hat{\mathbb{X}}_m^b, \hat{\mathbb{Y}}_n^b) \leftarrow$ PERMUTECORESETS$(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n, s)$
  $M_b \leftarrow \mathrm{MMD}_{\mathbf{k}}(\hat{\mathbb{X}}_m^b, \hat{\mathbb{Y}}_n^b)$
end

// Threshold test statistic
$R \leftarrow$ position of $M_{\mathcal{B}+1}$ in an increasing ordering of $(M_b)_{b=1}^{\mathcal{B}+1}$ with ties broken uniformly at random
if $R > b_\alpha \coloneqq \lceil (1-\alpha)(\mathcal{B}+1) \rceil$ **then return** 1 // reject null
else if $R < b_\alpha$ **then return** 0 // accept null
else return 1 with prob. $p_\alpha = b_\alpha - (1-\alpha)(\mathcal{B}+1)$ or else 0

3.1 MMD compression with CORESETMMD

At the heart of our testing strategy lies CORESETMMD (5), a new, inexpensive estimate for $\mathrm{MMD}_{\mathbf{k}}(\mathbb{P}, \mathbb{Q})$ that builds atop the KT-COMPRESS algorithm, a strategy introduced by Shetty et al. (2022, Ex. 4) to compress a given point sequence (see App. A for background on KT-COMPRESS). Given a coreset count s, a target compression level $\mathfrak{g}$, and an auxiliary kernel function $\mathbf{k}'$ used by KT-COMPRESS, CORESETMMD partitions each input sample into bins of size $\frac{m+n}{s}$, compresses each bin into a smaller coreset of points using KT-COMPRESS, concatenates the coresets to form the compressed approximations $\hat{\mathbb{X}}_m$ and $\hat{\mathbb{Y}}_n$ of size $2^{\mathfrak{g}} m\sqrt{\frac{s}{m+n}}$ and $2^{\mathfrak{g}} n\sqrt{\frac{s}{m+n}}$ respectively, and finally computes the MMD estimate $\mathrm{MMD}_{\mathbf{k}}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)$.

As we show in App. B, this strategy offers the following strong approximation error guarantees, expressed in terms of the KT-COMPRESS *error inflation factor* $\mathbf{R}_{\mathbf{k},\mathbf{k}'}/2^{\mathfrak{g}}$.

Lemma 1 (Quality of CORESETMMD)**.** *The* CORESETMMD *estimate* (5) *satisfies*[1]

$$|\,\mathrm{MMD}_{\mathbf{k}}(\mathbb{X}_m, \mathbb{Y}_n) - \mathrm{MMD}_{\mathbf{k}}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)| \le \frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{X}_m, \frac{m}{s_m}, \delta, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + \frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Y}_n, \frac{n}{s_n}, \delta, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}}, \tag{6}$$

with probability at least $1-\delta$ *conditional on* $(\mathbb{X}_m, \mathbb{Y}_n)$*, and*

$$|\,\mathrm{MMD}_{\mathbf{k}}(\mathbb{P}, \mathbb{Q}) - \mathrm{MMD}_{\mathbf{k}}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)| \le \frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, \frac{m}{s_m}, \delta, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + \frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, \frac{n}{s_n}, \delta, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} + c_\delta\Big(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{\|\mathbf{k}\|_\infty}{n}}\Big), \tag{7}$$

with probability at least $1-3\delta$ *for* $c_\delta \triangleq 2 + \sqrt{2\log(\frac{2}{\delta})}$.

Remark 1 (Beyond i.i.d. data)**.** *Our proof shows that the guarantee* (6) *holds more generally for any point sequences* $(\mathbb{X}_m, \mathbb{Y}_n)$ *generated independently of the randomness in* CORESETMMD.

Remark 2 (Beyond KT-COMPRESS)**.** CORESETMMD *and* CTT *are compatible with any compression scheme. In particular, when an alternative compression algorithm is used in place of* KT-COMPRESS *in Alg. 1, the conclusions of Lem. 1 and Thm. 1 can be straightforwardly generalized to accommodate the quality guarantees of that alternative.*

The first guarantee of Lem. 1 bounds the compression error introduced by substituting the compressed points $(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)$ for $(\mathbb{X}_m, \mathbb{Y}_n)$, while the second accounts also for the $\Theta(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}})$ random fluctuations of the quadratic-time statistic $\mathrm{MMD}_{\mathbf{k}}(\mathbb{X}_m, \mathbb{Y}_n)$ around the population estimand $\mathrm{MMD}_{\mathbf{k}}(\mathbb{P}, \mathbb{Q})$ (Gretton et al., 2012a). In either case, we find that CORESETMMD offers an order $O(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}})$ approximation—the same order as the quadratic-time $\mathrm{MMD}_{\mathbf{k}}(\mathbb{X}_m, \mathbb{Y}_n)$ estimate—up to the inflation factor $(1 + \mathbf{R}_{\mathbf{k},\mathbf{k}'}/2^{\mathfrak{g}})$.

The value $\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, \mathfrak{g})$ depends on the choice of the auxiliary kernel $\mathbf{k}'$ and the tail decay of $\mathbb{P}$ (see App. B.1 for details).[2] Two standard choices for $\mathbf{k}'$ are the target kernel $\mathbf{k}$ itself (Dwivedi and Mackey, 2022) or a *square-root kernel* $\mathbf{k}_{\text{rt}}$ satisfying $\mathbf{k}(x, y) = \int \mathbf{k}_{\text{rt}}(x, z)\mathbf{k}_{\text{rt}}(y, z)dz$. As detailed in Dwivedi and Mackey (2021), convenient square-root (or square-root dominating) kernels are available for a

[1] Unless otherwise specified, all of our results refer to an arbitrary setting of an algorithm's input arguments.

[2] The related value $\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{X}_m, m, \delta, \mathfrak{g})$ is $\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\cdot, m, \delta, \mathfrak{g})$ applied to the empirical distribution over $\mathbb{X}_m$.

Tails of $\mathbb{P}$	Choice of $\mathbf{k}'$	$\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, \mathfrak{g})$
Compact	Compact $\mathbf{k}_{\text{rt}}$	$(\log \frac{m}{\delta})^2$
Subexponential	Analytic $\mathbf{k}$	$(\log \frac{m}{\delta})^{\frac{3d+5}{2}}$
Subexponential	Subexponential $\mathbf{k}_{\text{rt}}$	$c_{m,\delta}(\log \frac{m}{\delta})^{\frac{d+5}{2}}$
ρ-Heavy-tailed	ρ-Heavy-tailed $\mathbf{k}_{\text{rt}}$	$(\frac{m}{\delta})^{\frac{d}{2\rho}}(\log \frac{m}{\delta})^{\frac{5}{2}}$

Table 1: **Error inflation due to compression.** We report the scaling of $\mathbf{R}_{\mathbf{k},\mathbf{k}'}$ in Lem. 1 up to constants depending on d under various assumptions on $\mathbf{k}'$ and the tail decay of $\mathbb{P}$ (see App. B.5 for the proof). Here $c_{m,\delta} \triangleq \sqrt{\log\log \frac{m}{\delta}}$.

Test name	MMD separation	Runtime
CTT (ours, Thm. 1)	$\frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P},m,\delta,\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + m^{-\frac{1}{2}}$	$4^{\mathfrak{g}} m \log m$
Complete MMD (Gretton et al., 2012a)	$m^{-\frac{1}{2}}$	m^2
Block MMD (Zaremba et al., 2013)	$(Bm)^{-\frac{1}{4}}$	Bm
Incomp. MMD (Yamada et al., 2019)	$\ell^{-\frac{1}{4}}$	ℓ

Table 2: **Detectable** $\mathrm{MMD}(\mathbb{P}, \mathbb{Q})$ **separation vs. runtime** for complete and approximate MMD tests. For subexponential $(\mathbb{P}, \mathbb{Q})$, CTT can detect $m^{-\frac{1}{2}}$ MMD separation in near-linear time, while the complete, block, and incomplete tests require quadratic time. See Sec. 3.2 for more details.

variety of popular kernels including Gaussian, Matérn, B-spline, inverse multiquadric (IMQ), sech, and Wendland's compactly supported $\mathbf{k}$.

Tab. 1 summarizes how $\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, \mathfrak{g})$ varies with $\mathbf{k}'$ and $\mathbb{P}$. For example, when $\mathbb{P}$, $\mathbb{Q}$, and $\mathbf{k}' = \mathbf{k}_{\text{rt}}$ are compactly supported, $\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, \mathfrak{g}) = O((\log \frac{m}{\delta})^2)$ and hence the compression error (6) of Lem. 1 becomes

$$O(\frac{(\log \frac{m}{\delta})^2}{2^{\mathfrak{g}}\sqrt{m}} + \frac{(\log \frac{n}{\delta})^2}{2^{\mathfrak{g}}\sqrt{n}}) = o(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}}),$$

when $\mathfrak{g} = \log_2(\omega(\log^2(\frac{m \vee n}{\delta})))$. More generally, the CORESETMMD compression error is asymptotically negligible relative to the usual error of $\mathrm{MMD}_{\mathbf{k}}(\mathbb{X}_m, \mathbb{Y}_n)$ whenever $\mathfrak{g} = \log_2(\omega(\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m \vee n, \delta, \mathfrak{g})))$. For example, if $\mathbb{P}$, $\mathbb{Q}$, and $\mathbf{k}' = \mathbf{k}_{\text{rt}}$ have have subexponential tails then, for some constant $c > 0$, the choice $\mathfrak{g} \geq c \log_2 \log(m \vee n)$ yields $o(\frac{1}{\sqrt{m}} + \frac{1}{\sqrt{n}})$ compression error. By Tab. 1, the same result holds when $\mathbf{k}' = \mathbf{k}$ is analytic. Together, these results cover all of the aforementioned popular kernels.

We next turn our attention to the running time of CORESETMMD. By Shetty et al. (2022, Ex. 4), the runtime of each KT-COMPRESS$(\mathbb{X}_m^{(i)}, \mathfrak{g}, \mathbf{k}, \mathbf{k}', \delta)$ call is dominated by $O(4^{\mathfrak{g}} \frac{m+n}{s}(\log_4(\frac{m+n}{s}) - \mathfrak{g}))$ kernel evaluations. Since $\mathrm{MMD}_{\mathbf{k}}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)$ can be computed using $O(4^{\mathfrak{g}} s(m+n))$ kernel evaluations once $\hat{\mathbb{X}}_m$ and $\hat{\mathbb{Y}}_n$ are available, the total runtime of CORESETMMD is

$$O(4^{\mathfrak{g}}(m+n)(s + \log_4(\frac{m+n}{s}) - \mathfrak{g})). \tag{8}$$

Notably, this runtime is $O((m+n)\log_4^{c+1}(m+n))$, *near-linear* in $m+n$, whenever $s = O(\log_4(m+n))$ and $\mathfrak{g} \leq c \log_4 \log(m+n)$, as in the subexponential and compact-support settings previously considered.

3.2 Compress Then Test

We are now prepared to discuss our complete CTT procedure defined in Alg. 1. CTT begins by computing the CORESETMMD test statistic described in Sec. 3.1 but then reuses the coresets to carry out a special form of the permutation test. Rather than permuting all $m+n$ points as in the standard permutation tests of Sec. 2, CTT keeps each coreset intact and only permutes the order of the s coresets when setting the test threshold.

The advantages of coreset reuse are threefold. First, the compression step can be carried out just once, irrespective of the number of permutations employed. Second, the same kernel evaluations used to compute the initial test statistic (5) can be reused to compute every permuted CORESETMMD. Indeed, when forming the initial test statistic, it suffices to store the s^2 sufficient statistics

$$a_{ij} = \frac{1}{|\hat{\mathbb{Z}}^{(i)}||\hat{\mathbb{Z}}^{(j)}|} \sum_{z \in \hat{\mathbb{Z}}^{(i)}, z' \in \hat{\mathbb{Z}}^{(j)}} \mathbf{k}(z, z') \quad \text{for}$$
$$(\hat{\mathbb{Z}}^{(1)}, \dots, \hat{\mathbb{Z}}^{(s)}) \triangleq (\hat{\mathbb{X}}_m^{(1)}, \dots, \hat{\mathbb{X}}_m^{(s_m)}, \hat{\mathbb{Y}}_n^{(1)}, \dots, \hat{\mathbb{Y}}_n^{(s_n)})$$

since each permuted CORESETMMD can be written as

$$\mathrm{MMD}^2_{\mathbf{k}}(\hat{\mathbb{X}}_m^b, \hat{\mathbb{Y}}_n^b) = \sum_{i,j=1}^{s} \frac{1 - 2|\mathbb{I}[i \leq s_m] - \mathbb{I}[j \leq s_m]|}{s^2} a_{\sigma(i)\sigma(j)}$$

for some permutation σ over the coreset indices $\{1, \dots, s\}$. Hence, the total running time of CTT is simply the running time (8) of a single CORESETMMD call plus $O(s^2\mathcal{B})$ arithmetic operations.

Finally, by keeping each coreset intact, CTT ensures that every coreset permutation $(\hat{\mathbb{X}}_m^b, \hat{\mathbb{Y}}_n^b)$ accurately approximates an analogous full-sample permutation that keeps each of the $(\mathbb{X}_m^{(i)})_{i=1}^{s_m}$ and $(\mathbb{Y}_n^{(i)})_{i=1}^{s_n}$ bins intact and only permutes the order of the s bins. One of the main contributions of this work is showing that such restricted permutation procedures provide high power even when s is set to a small value. However, before turning to power, we next show that CTT has a size exactly equal to the nominal level α for all sample sizes and all data distributions $\mathbb{P}$.

Proposition 1 (Finite-sample exactness of CTT)**.** *For any distribution $\mathbb{P}$, Compress Then Test (Alg. 1) has size (Type I error) exactly equal to the nominal level α, i.e.,*

$$\Pr[\Delta_{\mathrm{CTT}}(\mathbb{X}_m, \mathbb{Y}_n) = 1] = \alpha \quad \textit{whenever} \quad \mathbb{P} = \mathbb{Q}.$$

Remark 3 (Exchangeability)**.** *Our proof in App. C does not require datapoint independence and rather holds under the weaker condition that the point sequence $(X_1, \dots, X_m, Y_1, \dots, Y_n)$ is* exchangeable *under the null, i.e., the null distribution of this point sequence is invariant under permutation.*

Our proof of Prop. 1, based on exchangeability, parallels the size-no-larger-than-level proofs of Schrab et al. (2021); Albert et al. (2022) but includes a more detailed treatment of the case $R = b_\alpha$ to ensure the exactness of the Type I error, as in Hoeffding (1952).

We now provide a complementary upper bound on the Type II error of CTT (or equivalently, a lower bound on its power) under suitable assumptions on the MMD separation between $\mathbb{P}$ and $\mathbb{Q}$.

Theorem 1 (Power of CTT)**.** *Suppose Compress Then Test (Alg. 1) is run with $m \leq n$, level α, replication count $\mathcal{B} \geq \frac{1}{\alpha} - 1$, coreset count $s_m \geq \frac{32}{9}\log(\frac{2e}{\gamma})$ for $\gamma \triangleq \frac{\alpha}{4e}(\frac{\tilde{\beta}}{4})^{\frac{1}{\lfloor \alpha(\mathcal{B}+1) \rfloor}}$ and $\tilde{\beta} \triangleq \frac{\beta}{1+\beta/2}$. Then* CTT *has power*

$$\Pr[\Delta_{\mathrm{CTT}}(\mathbb{X}_m, \mathbb{Y}_n) = 1] \geq 1 - \beta$$

whenever $c' \,\mathrm{MMD}_{\mathbf{k}}(\mathbb{P}, \mathbb{Q}) / \sqrt{\log(1/\gamma)}$ is greater than

$$2c_{\tilde{\beta}/20s}\sqrt{\frac{\|\mathbf{k}\|_\infty}{m}} + \frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, \frac{m}{s_m}, \frac{\tilde{\beta}}{20s_m}, \mathfrak{g}) + \mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, \frac{m}{s_m}, \frac{\tilde{\beta}}{20s_n}, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}}$$

for c' a universal constant and c_δ as defined in Lem. 1.

Remark 4 (Valid parameter values)**.** *The* CTT *compression level $\mathfrak{g}$ is an integer in $\{0, \dots, \log_4(\frac{m+n}{s})\}$—a larger value provides more power but increases runtime. The failure probability δ and level α take arbitrary values in $(0, 1)$, while the coreset count $s \leq m + n$ and replicate count $\mathcal{B}$ are positive integers.*

The proof of Thm. 1 in App. D contains several novel arguments that may be of independent interest. First, using novel techniques based on order statistics, we show that $\mathcal{B} \geq \frac{1}{\alpha} - 1$ permutations suffice to obtain a powerful permutation test. Our arguments can be straightforwardly adapted to strengthen the analogous results for the complete (Schrab et al., 2021, Thm. 5) and incomplete (Schrab et al., 2022, Thm. 5.2) permutation tests. Compared with the $\mathcal{B} \geq \frac{3}{\alpha^2}(\log(\frac{8}{\beta}) + \alpha(1-\alpha))$ requirement of Schrab et al. (2021, 2022), our requirement eliminates all dependence on the target power $1 - \beta$ and improves the α dependence by a quadratic factor. Put in practical terms, by Thm. 1, $\mathcal{B} \geq 19$ permutations suffice for powerful permutation testing at level $\alpha = 0.05$ while $\mathcal{B} \geq 2613$ were previously required to guarantee power greater than the level. Second, we show that to obtain a powerful permutation test, one need not permute all $m + n$ datapoints; rather, it suffices to permute s bins where the number s can be chosen *independently* of the sample sizes.

In the end, Thm. 1 implies that CTT with a small number of coresets and permutations can detect distributional discrepancies of order $\frac{1}{\sqrt{m}}$—the same detection threshold enjoyed by the quadratic-time MMD tests (Gretton et al., 2012b, Thm. 13)—-up to the inflation factor $(1 + \mathbf{R}_{\mathbf{k},\mathbf{k}'}/2^{\mathfrak{g}})$. Since $\mathbf{R}_{\mathbf{k},\mathbf{k}'}/2^{\mathfrak{g}} = o(1)$ whenever $\mathfrak{g} = \log_2(\omega(\mathbf{R}_{\mathbf{k},\mathbf{k}'}))$ and the runtime of CTT is dominated by a single CoresetMMD computation, by setting $\mathbf{k}'$, $\mathfrak{g}$, and s as discussed in Sec. 3.1, CTT can recover the quadratic-time detection threshold in *near-linear* $O((m+n)\log_4^{c+1}(m+n))$ time for subexponential $(\mathbb{P}, \mathbb{Q})$ and subquadratic time for heavy-tailed $(\mathbb{P}, \mathbb{Q})$ with $\rho > 2d$ moments.

Our next result shows that such runtime improvements *cannot* be achieved by the state-of-the-art block and incomplete MMD tests of Sec. 2, as each requires quadratic time (i.e., $B = \Omega(m)$ or $\ell = \Omega(m^2)$) to match the order $\frac{1}{\sqrt{m}}$ detection threshold of a complete MMD test.

Proposition 2 (Power upper bounds for complete, block, and incomplete MMD tests)**.** *For any nominal level $\alpha \in (0, 1)$ and target Type II error $\beta \in (0, 1)$, there exists a constant $c_{\alpha,\beta}$ such that the following power upper bounds hold for all sample sizes m.*

(a) Asymptotic complete test: $\Pr[\Delta_{up}(\mathbb{X}_m, \mathbb{Y}_m) = 1] < 1 - \beta$ if $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) \leq \frac{c_{\alpha,\beta}}{\sqrt{m}}$.

(b) Asymptotic block test: $\Pr[\Delta_B(\mathbb{X}_m, \mathbb{Y}_m) = 1] < 1 - \beta$ if $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) \leq \frac{c_{\alpha,\beta}}{(Bm)^{1/4}}$ and $B, \frac{m}{B} \to \infty$.

(c) Asymptotic incomplete test: $\Pr[\Delta_{inc}(\mathbb{X}_m, \mathbb{Y}_m) = 1] < 1 - \beta$ if $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) \leq \frac{c_{\alpha,\beta}}{\ell^{1/4}}$, and $\frac{\ell}{m} \to c > 0$.

The proof of Prop. 2 in App. E uses the asymptotic distribution of each statistic under the null and alternative hypotheses (as derived by Gretton et al., 2007, 2009; Zaremba et al., 2013; Yamada et al., 2019) to upper bound the power (and hence lower bound the Type II error) of each test. Moreover, the proof reveals that these detectable $\mathrm{MMD}(\mathbb{P}, \mathbb{Q})$ separation rates are tight. For example, there also exists a constant $c'_{\alpha,\beta} > c_{\alpha,\beta}$ such that $\Pr[\Delta_{up}(\mathbb{X}_m, \mathbb{Y}_m) = 1] > 1 - \beta$ whenever $\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) \geq c'_{\alpha,\beta}/\sqrt{m}$. Tab. 2 summarizes the trade-off between detectable MMD separation and runtime for the complete and approximate MMD tests and highlights the improved trade-off offered by CTT.

In particular, the time-power trade-off of CTT improves significantly under the favorable settings of Tab. 1 (e.g., for compact $\mathbb{P}$ or subexponential $\mathbb{P}$ and analytic k in lower dimensions). While the improvements need not be as large for heavier-tailed distributions, less smooth kernels, and higher dimensions, even the worst-case trade-offs of CTT are no worse than prior methods' as $\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, \mathfrak{g}) = O\Big(2^{\mathfrak{g}/2} m^{1/4} \sqrt{\log(\frac{1}{\delta})}\Big)$ by Dwivedi and Mackey (2021, Rem. 2). That is, for arbitrary distributions, dimensions, and kernels, a user can comfortably use CTT as a drop-in replacement for the block and incomplete tests, as we should expect no worse power-time trade-off curves. That said, there is some overhead associated with compression, so a user may find the block and incomplete tests to be faster for small sample sizes.

4 CTT Extensions

In this section, we develop two extensions of CTT: first, a fast and powerful way to exploit an accurate low-rank kernel approximation and, second, a fast and powerful aggregation procedure for identifying a particularly discriminating kernel from amongst a collection of candidates.

4.1 Low-Rank CTT

Our first extension, called Low-Rank CTT (Alg. 2), allows the user to exploit an accurate low-rank kernel approximation without sacrificing the provable time-power trade-off improvements of CTT. Specifically, we consider $\Theta(nr)$-time low-rank MMD_{Φ_r} approximations of the form (4) with Φ_r selected so that the approximation error

$$\epsilon^2_{\Phi_r}(\mathbb{X}_m, \mathbb{Y}_n) = \sup_{x,y\in\mathbb{X}_m\cup\mathbb{Y}_n} |\mathbf{k}(x, y) - \Phi_r(x)^\top \Phi_r(y)|$$

is small. For example, Sriperumbudur and Szabó (2015, Thm. 1) show that $\epsilon_{\Phi_r}(\mathbb{X}_m, \mathbb{Y}_n) = O(r^{-1/4})$ and hence that $|\,\mathrm{MMD}^2_{\mathbf{k}}(\mathbb{X}_m, \mathbb{Y}_n) - \mathrm{MMD}^2_{\Phi_r}(\mathbb{X}_m, \mathbb{Y}_n)| \overset{(60)}{\leq} 4\epsilon^2_{\Phi_r}(\mathbb{X}_m, \mathbb{Y}_n) = O(r^{-1/2})$ with high probability when Φ_r consists of r random Fourier features and $(\mathbb{P}, \mathbb{Q})$ are compactly supported. However, since computing $\mathrm{MMD}_{\Phi_r}(\mathbb{X}_m, \mathbb{Y}_n)$ requires $\Theta((m + n)r)$ feature evaluations, this analysis requires $\Omega(m^3)$ time to match the order $\frac{1}{\sqrt{m}}$ detection threshold of a complete MMD test. Our following result, proved in App. F, shows that appropriate compression prior to low-rank approximation yields comparable power guarantees in just $O(4^{\mathfrak{g}}(m + n) \log r)$ time.

Theorem 2 (LR-CTT exactness and power)**.** *Low-Rank* CTT *(Alg. 2) has size exactly equal to the level α for all $\mathbb{P}$. If the replication count $\mathcal{B} \geq \frac{1}{\alpha} - 1$, the permutation bin count $s \geq \frac{m+n}{m} \frac{32}{9} \log(\frac{2e}{\gamma})$ for $(\gamma, \tilde{\beta})$ as in Thm. 1, and $m \leq n$, then* LR-CTT *has power*

$$\Pr[\Delta_{\text{LR-CTT}}(\mathbb{X}_m, \mathbb{Y}_n) = 1] \geq 1-\beta$$

when, for a universal constant c' and c_δ defined in Lem. 1,

$$c' \,\mathrm{MMD}_{\mathbf{k}}(\mathbb{P}, \mathbb{Q})/\sqrt{\log(1/\gamma)} \geq 2c_{\tilde{\beta}/20s_r} \sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}} + \frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, \frac{m}{s_{m,r}}, \frac{\tilde{\beta}}{20 s_{m,r}}, \mathfrak{g}) + \mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, \frac{m}{s_{m,r}}, \frac{\tilde{\beta}}{20 s_{n,r}}, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + \epsilon_{\Phi_r}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n).$$

Algorithm 2: Low-Rank CTT, $\Delta_{\text{LR-CTT}}$

Input: Samples $(\mathbb{X}_m, \mathbb{Y}_n)$, coreset size factor a, compression level $\mathfrak{g}$, kernels $(\mathbf{k}, \mathbf{k}')$, feature map Φ_r, failure prob. δ, # permutation bins s, # replicates $\mathcal{B}$, level α

Partition $\mathbb{X}_m$ into $s_{m,r} = \frac{4^{\mathfrak{g}} a^2 m}{r^2}$ equal-sized bins $(\mathbb{X}_m^{(i)})_{i=1}^{s_{m,r}}$
Partition $\mathbb{Y}_n$ into $s_{n,r} = \frac{4^{\mathfrak{g}} a^2 n}{r^2}$ equal-sized bins $(\mathbb{Y}_n^{(i)})_{i=1}^{s_{n,r}}$
// Identify coreset of size $\frac{r}{a}$ for each bin
for $i = 1, \dots, s_{m,r}$ **do**
  $\hat{\mathbb{X}}_m^{(i)} \leftarrow$ KT-COMPRESS$(\mathbb{X}_m^{(i)}, \mathfrak{g}, \mathbf{k}, \mathbf{k}', \delta)$
end
for $i = 1, \dots, s_{n,r}$ **do**
  $\hat{\mathbb{Y}}_n^{(i)} \leftarrow$ KT-COMPRESS$(\mathbb{Y}_n^{(i)}, \mathfrak{g}, \mathbf{k}, \mathbf{k}', \delta)$
end
// Compute LR-CORESETMMD test statistic

$$M_{\mathcal{B}+1} \leftarrow \mathrm{MMD}_{\Phi_r}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n) \quad \text{for} \tag{9}$$

$\hat{\mathbb{X}}_m := \mathrm{CONCAT}((\hat{\mathbb{X}}_m^{(i)})_{i=1}^{s_{m,r}})$ and $\hat{\mathbb{Y}}_n := \mathrm{CONCAT}((\hat{\mathbb{Y}}_n^{(i)})_{i=1}^{s_{n,r}})$
// Simulate null by randomly permuting s coresets $\mathcal{B}$ times
for $b = 1, \dots, \mathcal{B}$ **do**
  $(\hat{\mathbb{X}}_m^b, \hat{\mathbb{Y}}_n^b) \leftarrow$ PERMUTECORESETS$(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n, s)$
  $M_b \leftarrow \mathrm{MMD}_{\Phi_r}(\hat{\mathbb{X}}_m^b, \hat{\mathbb{Y}}_n^b)$
end
// Threshold test statistic
$R \leftarrow$ position of $M_{\mathcal{B}+1}$ in an increasing ordering of $(M_b)_{b=1}^{\mathcal{B}+1}$ with ties broken uniformly at random
if $R > b_\alpha := \lceil(1-\alpha)(\mathcal{B}+1)\rceil$ **then return** 1 // reject null
else if $R < b_\alpha$ **then return** 0 // accept null
else return 1 with prob. $p_\alpha = b_\alpha - (1 - \alpha)(\mathcal{B} + 1)$ or else 0

Specifically, to form a low-rank CORESETMMD test statistic (9), Low-Rank CTT (LR-CTT, Alg. 2) divides each sample into $s_{m,r}$ or $s_{n,r}$ equal-sized bins, forms a coreset for each bin using KT-COMPRESS with kernels $(\mathbf{k}, \mathbf{k}')$, and computes the low-rank approximation MMD_{Φ_r} using only the concatenated coreset points $\hat{\mathbb{X}}_m$ and $\hat{\mathbb{Y}}_n$. Then, just as in Alg. 1, LR-CTT selects an appropriate test statistic threshold albeit now manually partitioning $(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)$ into s coreset bins and permuting those bins. All told, the LR-CTT runtime is dominated by $O(4^{\mathfrak{g}}(m+n)(\log_4(\frac{2r}{a}) - \mathfrak{g}))$ kernel evaluations, $O(4^{\mathfrak{g}}(m + n)a)$ feature evaluations, and $O(s^2\mathcal{B})$ arithmetic operations. Importantly, when $a = O(\log r)$, the *logarithmic* dependence on the rank r means that, by Thm. 2, LR-CTT can recover the order $\frac{1}{\sqrt{m}}$ detection threshold of a complete MMD test in *near-linear* time for subexponential $(\mathbb{P}, \mathbb{Q})$ and subquadratic time for heavy-tailed $(\mathbb{P}, \mathbb{Q})$ with $\rho > 2d$ moments, even when the approximation error ϵ_{Φ_r} has slow (e.g., order $r^{-1/4}$) decay.

4.2 Aggregated CTT

Each of the tests considered so far assumes that a suitable kernel $\mathbf{k}$ has been pre-selected by the user. However, because the discriminating power of a kernel varies with the pair of distributions under consideration, it can be challenging to identify a single suitable kernel a priori. As a result, a variety of strategies have been introduced for automatically selecting discriminating kernels for MMD tests (see, e.g., Gretton et al., 2012b; Sutherland et al., 2017; Liu et al., 2020; Kübler et al., 2020). We highlight in particular the *aggregated MMD tests* of Schrab et al. (2021) which combine complete MMD tests with varying kernels into a single test with power comparable to the best individual test. Since these complete aggregated tests run in quadratic time, Schrab et al. (2022) recently introduced incomplete aggregated tests that trade off computation time and power exactly as in the single-kernel setting (see Tab. 2).

In Alg. 6 of App. G, we extend our Compress Then Test framework to form a more efficient aggregated test that we call Aggregated CTT (ACTT). Like past aggregated tests, Alg. 6 takes as input *any* indexed collection of kernels $(\mathbf{k}_\lambda)_{\lambda\in\Lambda}$ and accommodates nonnegative importance weights $(w_\lambda)_{\lambda\in\Lambda}$ with $\sum_{\lambda\in\Lambda} w_\lambda \leq 1$ reflecting prior beliefs about the suitability of each kernel. Like Alg. 1, ACTT then proceeds to partition $\mathbb{X}_m$ and $\mathbb{Y}_n$ into bins and to form a coreset for each bin using a parallel collection of auxiliary KT-COMPRESS kernels $(\mathbf{k}'_\lambda)_{\lambda\in\Lambda}$ scaled so that $\sup_z|\mathbf{k}'_\lambda(z,z)| = 1$. However, instead of forming a separate coreset for each candidate kernel, as one might if one were running a CTT test separately for each $\mathbf{k}_\lambda$, ACTT saves additional computation by forming a single coreset per bin using the combination kernels $\mathbf{k} = \sum_{\lambda\in\Lambda}\mathbf{k}_\lambda$ and $\mathbf{k}' = \sum_{\lambda\in\Lambda}\mathbf{k}'_\lambda$. These shared coresets are used to compute a CORESETMMD test statistic M_λ for each $\mathbf{k}_\lambda$, $\mathcal{B}_1$ permuted CORESETMMD statistics to estimate the null distribution for each $\mathbf{k}_\lambda$, and $\mathcal{B}_2$ permuted CORESETMMD statistics to estimate the size of the aggregated test. Finally, exactly as in Schrab et al. (2021, Alg. 1), ACTT selects a suitable rejection threshold for each test statistic M_λ and rejects the null whenever at least one M_λ exceeds its threshold. The total cost of ACTT is at most $|\Lambda|$ times that of single-kernel CTT (with $\mathcal{B} = \mathcal{B}_1$) plus the cost of $O(|\Lambda|(\mathcal{B}_1\log\mathcal{B}_1 + \mathcal{B}_2\mathcal{B}_3))$ arithmetic operations due to sorting and selecting thresholds.

Thm. 3, proved in App. G, shows that ACTT is *valid*, i.e., it has Type I error $\leq \alpha$ for all sample sizes and generating distributions, and that its power is comparable to that of the best $\mathbf{k}_\lambda$-CTT test run with compression level $\mathfrak{g} - \log_2|\Lambda|$. Moreover, by Thm. 1, each $\mathbf{k}_\lambda$-CTT test has power comparable to a complete $\mathbf{k}_\lambda$ test when $\mathfrak{g} = \log_2(|\Lambda|\omega(\mathbf{R}_{\mathbf{k}_\lambda,\mathbf{k}'_\lambda}))$. Therefore, by setting $\mathbf{k}'$, $\mathfrak{g}$, and s as discussed in Sec. 3.1, ACTT with $|\Lambda| = O(1)$ can recover the detection threshold of the best quadratic-time $\mathbf{k}_\lambda$ test in *near-linear* $O((m+n)\log_4^{c+1}(m+n))$ time for subexponential $(\mathbb{P},\mathbb{Q})$ and subquadratic time for heavy-tailed $(\mathbb{P},\mathbb{Q})$ with $\rho > 2d$ moments.

Theorem 3 (ACTT validity and power)**.** *For any distribution $\mathbb{P}$, ACTT (Alg. 6) has non-asymptotic level α, i.e.,*

$$\Pr[\Delta_{\text{ACTT}}(\mathbb{X}_m,\mathbb{Y}_n)=1] \leq \alpha \text{ whenever } \mathbb{P}=\mathbb{Q}. \quad (10)$$

Morever, with $m \leq n$, $\alpha \in (0,\frac{1}{e})$, and replicate counts $\mathcal{B}_1 \geq (\max_{\lambda\in\Lambda} w_\lambda^{-2})\frac{12}{\alpha^2}(\log(\frac{8}{\beta})+\alpha(1-\alpha))$, $\mathcal{B}_2 \geq \frac{8}{\alpha^2}\log(\frac{2}{\beta})$, and $\mathcal{B}_3 \geq \log_2(\frac{4}{\alpha}\min_{\lambda\in\Lambda} w_\lambda^{-1})$, ACTT *has power*

$$\Pr[\Delta_{\text{ACTT}}(\mathbb{X}_m,\mathbb{Y}_n)=1] \geq 1-\beta \quad (11)$$

whenever there exists a $\lambda \in \Lambda$ satisfying

$$\mathrm{MMD}_{\mathbf{k}_\lambda}(\mathbb{P},\mathbb{Q}) \geq c'\sqrt{\log(\tfrac{1}{\gamma_\lambda})}\,\varepsilon_{\text{AGG}}(\tfrac{\beta/(10s)}{4+\beta}) \quad (12)$$

and $s_m \geq \frac{32}{9}\log(\frac{2e}{\gamma_\lambda})$, where $\gamma_\lambda \triangleq \frac{\alpha w_\lambda}{8e}(\frac{\beta}{8+2\beta})^{\frac{1}{\lfloor \alpha w_\lambda(\mathcal{B}_1+1)/2\rfloor}}$, c' is a universal constant, and

$$\varepsilon_{\text{AGG}}(\delta) \triangleq 2c_\delta\sqrt{\tfrac{\|\mathbf{k}_\lambda\|_\infty}{m}} + c_\Lambda \max_{\lambda\in\Lambda}\frac{\mathbf{R}_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathbb{P},\frac{m}{s_m},\delta,\mathfrak{g})+\mathbf{R}_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathbb{Q},\frac{m}{s_m},\delta,\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}}$$

for c_δ as in Lem. 1 and $c_\Lambda \triangleq 2\sqrt{|\Lambda|(1+\log(|\Lambda|))}$.

5 Experiments

We now present seven experiments that illustrate the improved power-runtime trade-offs of CTT, LR-CTT, and ACTT over state-of-the-art approximate MMD tests. In all experiments, we use a Gaussian $\mathbf{k}' = \mathbf{k}$, $\alpha = 0.05$, $m = n = 4^9$, $s = 32$, and $\delta = \frac{1}{2}$. We report average rejection rates over 400 independent replications of each experiment with 95% Wilson (1927) confidence intervals. See App. H for additional details and github.com/microsoft/goodpoints for open-source Python code recreating all experiments.

CTT experiments We evaluate CTT in two settings with the kernel bandwidth set using the popular median heuristic (Chaudhuri et al., 2017). In the GAUSSIAN setting, $\mathbb{P}$ and $\mathbb{Q}$ are 10-dimensional Gaussians with identity covariances; the means have Euclidean distance 0.012 under the alternative and 0 under the null. The EMNIST setting is similar to the one considered by Kübler et al. (2020); Schrab et al. (2021), where $\mathbb{P}$ and $\mathbb{Q}$ denote distributions on downsampled 7×7 images of the EMNIST dataset (Cohen et al., 2017)—an extension of the MNIST dataset (LeCun et al., 2010) that also includes letters. Under the alternative hypothesis, $\mathbb{P}$ denotes a 2-mixture of uniform distributions based on parity of digits and letters with weight 0.49 (resp. 0.51) for even (resp. odd) parity, while $\mathbb{Q}$ puts equal weight 0.5 on both parities. Under the null hypothesis, we consider $\mathbb{P} = \mathbb{Q}$ = equally weighted mixture. We plot the test power results versus runtime in Fig. 1 with GAUSSIAN setting on top and EMNIST setting on the bottom.

Fig. 1 (left) shows that in both settings, the CTT time-power trade-off curve uniformly dominates those of the state-of-the-art subsampling approximations of Sec. 2: the wild bootstrap block (W-Block) and incomplete (W-Incomp.) tests and the asymptotic block (A-Block I and II) incomplete (A-Incomp.) tests. In particular, the CTT test

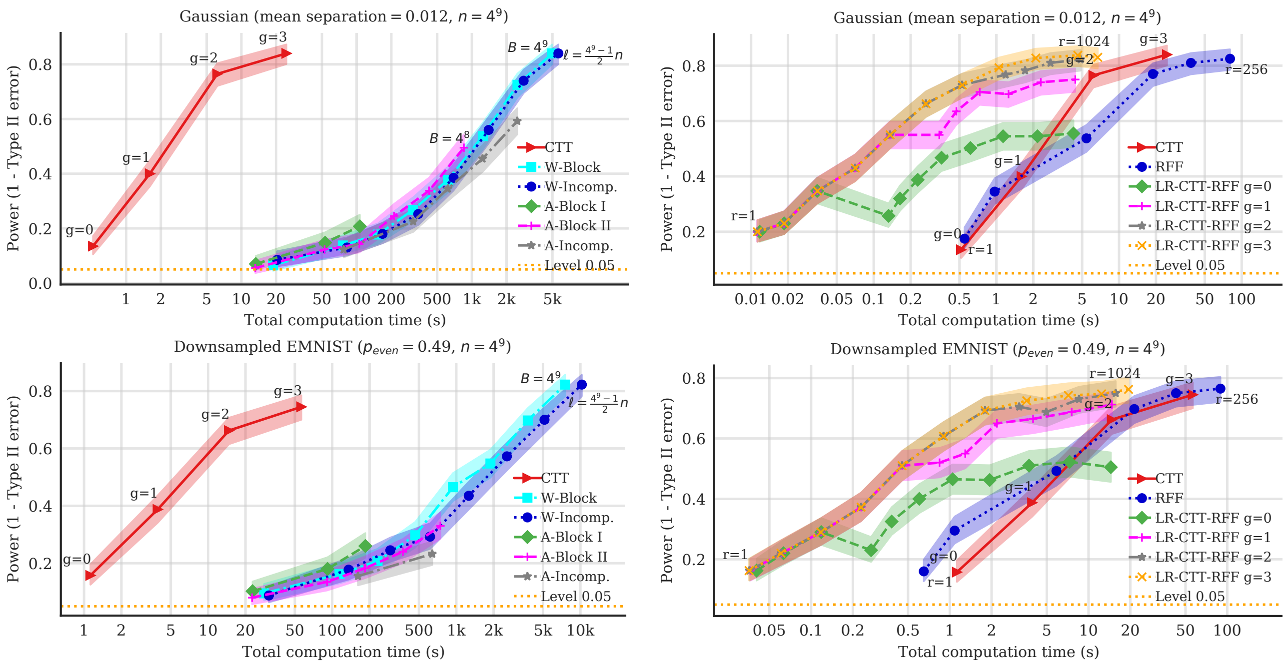

Figure 1: **Time-power trade-off curves** in the GAUSSIAN and EMNIST experimental settings comparing (*left*) CTT to five state-of-the-art approximate MMD tests based on subsampling and (*right*) LR-CTT to the state-of-the-art low-rank MMD test based on random Fourier features (RFF).

with $\mathfrak{g}=3$ achieves the same power as the wild bootstrap quadratic-time tests (W-block with $B=4^9$ and W-Incomp. with $\ell=\frac{(4^9-1)n}{2}$) while providing a $200\times$ speed-up. While CTT and the wild bootstrap tests are guaranteed to have Type I error controlled by α (Fig. 3 in App. H), the asymptotic tests violate their level constraint for large B or ℓ as the asymptotic approximation is poor for such settings. As a result, Fig. 1 displays only those points that respect the level constraint in the power plots. For consistency, we used $\mathcal{B}=39$ replicates for all the non-asymptotic tests.

LR-CTT experiments In the same settings, Fig. 1 (right) compares CTT, the state-of-the-art low-rank RFF test of Sec. 2, and LR-CTT with RFF Φ_r and $a = r/(4^{\mathfrak{g}}2^{\mathbb{I}[r>4^{\mathfrak{g}+1}]})$. We use $\mathcal{B}=39$ permutations to set the threshold for each test. We find that CTT and RFF produce comparable trade-off curves despite their distinct and complementary approximation strategies and that the combined LR-CTT test with $\mathfrak{g}\geq 2$ consistently yields the best performance, with 5–20$\times$ speed-ups over CTT or RFF alone.

ACTT experiments We compare our ACTT procedure in two different settings with the aggregated wild bootstrap incomplete test (W-Incomp.) of Schrab et al. (2022). For the BLOBS experiment of Gretton et al. (2012b, Fig. 1), and Sutherland et al. (2017, Fig. 2), $\mathbb{P}$ and $\mathbb{Q}$ are two-dimensional 3×3 grids of Gaussian mixture components with a grid spacing of 10. Each mixture component has identity covariance in $\mathbb{P}$, while for $\mathbb{Q}$ the ratio of eigenvalues for their covariance matrix is ϵ with diagonal entries set to 1; the null hypothesis corresponds to $\epsilon=1$. We consider the bandwidth set $\Lambda = \{2^i\lambda_0\}_{i=0}^{-4}$ for $\mathbf{k}_\lambda(x,y) = e^{-\|x-y\|_2^2/(2\lambda^2)}$, where λ_0 denotes the median heuristic bandwidth, and uniform weights $w_\lambda = 1/|\Lambda|$. We plot the results in Fig. 2 and observe that ACTT provides a uniform gain in the power-runtime curve over the aggregated WB incomplete test—a $100\times$ to $200\times$-speed up.

We perform the same comparison with the same configurations on the HIGGS experiment, a variation of the setting considered by Liu et al. (2020), which took the data from Baldi et al. (2014). While the original dataset has samples with 27 covariates belonging to two different classes (0 and 1), Liu et al. (2020) considers only four covariates of those, and we only use the first two of the four. We consider two settings for the alternative distribution: one in which $\mathbb{P}$ is sampled from the class 0 and $\mathbb{Q}$ is sampled from the class 1 (HIGGS: Fig. 2, *middle*) and a more challenging one in which $\mathbb{P}$ is sampled from the class 0 and $\mathbb{Q}$ is sampled from each class with equal probability (HIGGS MIXTURE: Fig. 2, *bottom*). We observe a $100\times$ to $200\times$-speed up over the aggregated WB incomplete test.

6 Connections and Conclusions

This paper introduced CTT, a new framework for kernel testing with compression; LR-CTT, a test that combines low-rank approximation and compression for added scalability; and ACTT, a fast and powerful procedure for aggregating kernel tests. While we have shown that CTT, LR-CTT, and ACTT offer better power-time trade-offs than

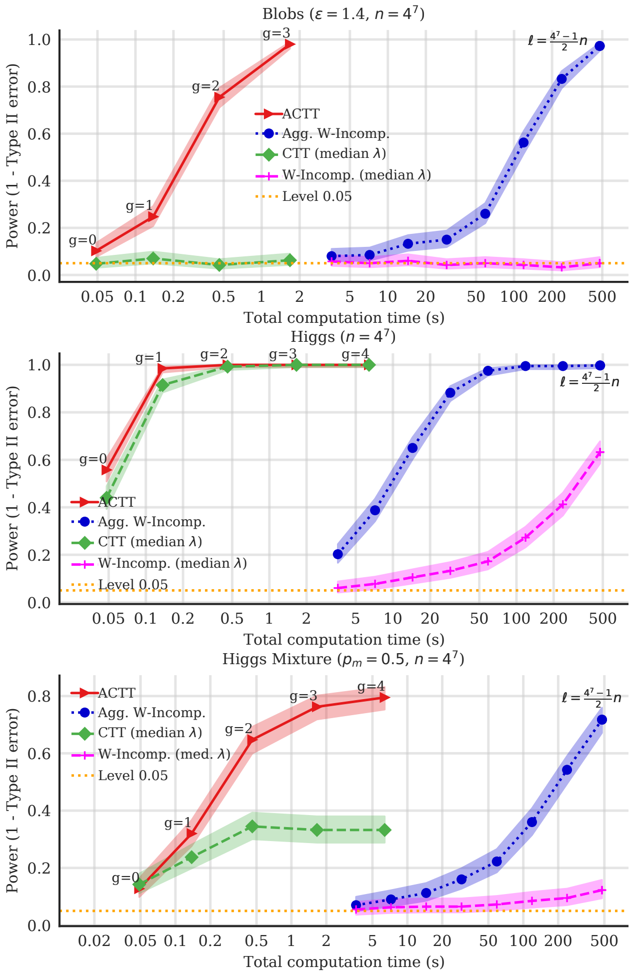

Figure 2: **Time-power trade-off curves** for ACTT and the state-of-the-art incomplete MMD aggregation test in the BLOBS and HIGGS experimental settings.

state-of-the-art approximate MMD tests, we highlight that there are other approaches to fast non-parametric testing based on alternative test statistics (see, e.g., Chwialkowski et al., 2015; Jitkrittum et al., 2016; Kirchler et al., 2020; Shekhar et al., 2022). A natural follow-up question is whether compression techniques can also improve the power-time trade-offs of those tests. A second opportunity for future work is to extend the CTT framework to other inferential tasks like independence and goodness-of-fit testing or kernel regression.

References

NIST Digital Library of Mathematical Functions. http://dlmf.nist.gov/, Release 1.1.6 of 2022-06-30. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds. (Cited on page 23.)

Mélisande Albert, Béatrice Laurent, Amandine Marrel, and Anouar Meynaoui. Adaptive test of independence based on HSIC measures. *The Annals of Statistics*, 50(2):858–879, 2022. (Cited on page 5.)

Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5(1), 2014. (Cited on page 8.)

Arin Chaudhuri, Deovrat Kakde, Carol Sadek, Laura Gonzalez, and Seunghyun Kong. The mean and median criteria for kernel bandwidth selection for support vector data description. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, nov 2017. (Cited on pages 7 and 42.)

Kacper P Chwialkowski, Aaditya Ramdas, Dino Sejdinovic, and Arthur Gretton. Fast two-sample testing with analytic representations of probability measures. *Advances in Neural Information Processing Systems*, 28, 2015. (Cited on page 9.)

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926, 2017. doi: 10.1109/IJCNN.2017.7966217. (Cited on page 7.)

Raaz Dwivedi and Lester Mackey. Kernel thinning. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 1753–1753. PMLR, 15–19 Aug 2021. (Cited on pages 2, 3, 6, 13, 18, 19, 20, and 40.)

Raaz Dwivedi and Lester Mackey. Generalized kernel thinning. In *International Conference on Learning Representations*, 2022. (Cited on pages 3, 13, 18, 20, 40, 41, and 42.)

Magalie Fromont, Béatrice Laurent, Matthieu Lerasle, and Patricia Reynaud-Bouret. Kernels based tests with non-asymptotic bootstrap approaches for two-sample problems. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 23.1–23.23. PMLR, 2012. (Cited on pages 2 and 3.)

James Gentle. *Computational Statistics*. 01 2009. ISBN 978-0-387-98143-7. doi: 10.1007/978-0-387-98144-4. (Cited on page 22.)

Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2007. (Cited on pages 5, 32, and 33.)

Arthur Gretton, Kenji Fukumizu, Zaïd Harchaoui, and Bharath K. Sriperumbudur. A fast, consistent kernel two-sample test. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009. (Cited on pages 5 and 32.)

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012a. (Cited on pages 1, 2, 3, 4, and 17.)

Arthur Gretton, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, Kenji Fukumizu, and Bharath K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012b. (Cited on pages 5, 7, and 8.)

Wassily Hoeffding. The Large-Sample Power of Tests Based on Permutations of Observations. *The Annals of Mathematical Statistics*, 23(2):169 – 192, 1952. doi: 10.1214/aoms/1177729436. (Cited on page 5.)

Wittawat Jitkrittum, Zoltán Szabó, Kacper Chwialkowski, and Arthur Gretton. Interpretable distribution features with maximum testing power. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, page 181–189. Curran Associates Inc., 2016. (Cited on page 9.)

Matthias Kirchler, Shahryar Khorasani, Marius Kloft, and Christoph Lippert. Two-sample testing using deep learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1387–1398. PMLR, 2020. (Cited on page 9.)

Jonas M. Kübler, Wittawat Jitkrittum, Bernhard Schölkopf, and Krikamol Muandet. Learning kernel tests without data splitting. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2020. (Cited on page 7.)

Erik Learned-Miller and Joseph DeStefano. A probabilistic upper bound on differential entropy. *IEEE Transactions on Information Theory*, 54(11):5223–5230, 2008. doi: 10.1109/TIT.2008.929937. (Cited on page 22.)

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010. (Cited on page 7.)

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang 0001, Arthur Gretton, and Dougal J. Sutherland. Learning deep kernels for non-parametric two-sample tests. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6316–6326. PMLR, 2020. (Cited on pages 7 and 8.)

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008. (Cited on page 1.)

Joseph P Romano and Michael Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100 (469):94–108, 2005. (Cited on pages 2 and 3.)

Mark Rudelson and Roman Vershynin. Hanson-Wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1 – 9, 2013. (Cited on page 29.)

Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. MMD aggregated two-sample test, 2021. (Cited on pages 1, 5, 7, 21, 22, 39, and 42.)

Antonin Schrab, Ilmun Kim, Benjamin Guedj, and Arthur Gretton. Efficient aggregated kernel tests using incomplete u-statistics, 2022. (Cited on pages 1, 2, 5, 7, 8, and 42.)

Robert Serfling. *Approximation Theorems of Mathematical Statistics*, volume 162. John Wiley & Sons, 2009. (Cited on page 33.)

Shubhanshu Shekhar, Ilmun Kim, and Aaditya Ramdas. A permutation-free kernel two-sample test. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. (Cited on page 9.)

Abhishek Shetty, Raaz Dwivedi, and Lester Mackey. Distribution compression in near-linear time. In *International Conference on Learning Representations*, 2022. (Cited on pages 3, 4, 13, 14, 15, 16, 17, 18, 39, 40, 41, and 42.)

Bharath Sriperumbudur and Zoltán Szabó. Optimal rates for random fourier features. *Advances in neural information processing systems*, 28, 2015. (Cited on pages 1 and 6.)

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008. (Cited on page 2.)

Danica J Sutherland and Jeff Schneider. On the error of random fourier features. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 862–871, 2015. (Cited on page 1.)

Danica J. Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alex Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*, 2017. (Cited on pages 7 and 8.)

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. (Cited on page 20.)

Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. (Cited on page 7.)

Makoto Yamada, Denny Wu, Yao-Hung Hubert Tsai, Hirofumi Ohta, Ruslan Salakhutdinov, Ichiro Takeuchi, and Kenji Fukumizu. Post selection inference with incomplete maximum mean discrepancy estimator. In *International Conference on Learning Representations*, 2019. (Cited on pages 1, 2, 4, 5, 34, and 42.)

Wojciech Zaremba, Arthur Gretton, and Matthew Blaschko. B-test: A non-parametric, low variance kernel two-sample test. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. (Cited on pages 1, 2, 4, 5, and 42.)

Ji Zhao and Deyu Meng. FastMMD: Ensemble of circular discrepancy for efficient two-sample test. *Neural Computation*, 27(6):1345–1372, 2015. (Cited on pages 1 and 2.)

Appendix

A Background on KT-COMPRESS

This section details the KT-COMPRESS algorithm of Shetty et al. (2022, Ex. 4). In a nutshell, KT-COMPRESS (Alg. 3) takes as input a point sequence of size n, a compression level $\mathfrak{g}$, two kernel functions $(\mathbf{k}, \mathbf{k}')$, and a failure probability δ. It then combines the COMPRESS algorithm of Shetty et al. (2022, Alg. 1) with the generalized kernel thinning (KT) algorithm of Dwivedi and Mackey (2021, 2022, Alg. 1) to output a coreset of $2^{\mathfrak{g}}\sqrt{n}$ input points that together closely approximate the input in terms of $\mathrm{MMD}_{\mathbf{k}}$. KT-COMPRESS proceeds by calling the recursive procedure COMPRESS, which uses KT with kernels $(\mathbf{k}, \mathbf{k}')$ as an intermediate halving algorithm. The KT algorithm itself consists of two subroutines: (1) KT-SPLIT (Alg. 5a), which splits a given input point sequence into two equal halves with small approximation error in the $\mathbf{k}'$ reproducing kernel Hilbert space and (2) KT-SWAP (Alg. 5b), which selects the best approximation amongst the KT-SPLIT coresets and a baseline coreset (that simply selects every other point in the sequence) and then iteratively refines the selected coreset by swapping out each element in turn for the non-coreset point that most improves $\mathrm{MMD}_{\mathbf{k}}$ error. As in Shetty et al. (2022, Rem. 3), we symmetrize the output of KT by returning either the KT coreset or its complement with equal probability.

For this work, we develop a slight modification of the original KT-COMPRESS algorithm, which we use in all experiments and in the released implementation. When the compression level $\mathfrak{g} = 0$ and the number of input points passed to COMPRESS is $n = 4$, instead of running the usual COMPRESS algorithm, we run OPTHALVE4 (Alg. 4) which identifies the coreset of size two that optimally approximates the input point sequence in terms of $\mathrm{MMD}_{\mathbf{k}}$ and then returns either that coreset or its complement with equal probability.

Algorithm 3: KT-COMPRESS – Identify coreset of size $2^{\mathfrak{g}}\sqrt{n}$

Input: point sequence $\mathcal{S}_{\text{in}}$ of size n, compression level $\mathfrak{g}$, kernels $(\mathbf{k}, \mathbf{k}')$, failure probability δ

return COMPRESS$(\mathcal{S}_{\text{in}}, \mathfrak{g}, \mathbf{k}, \mathbf{k}', \frac{\delta}{n4^{\mathfrak{g}+1}(\log_4 n - \mathfrak{g})})$

function COMPRESS $(\mathcal{S}, \mathfrak{g}, \mathbf{k}, \mathbf{k}', \delta)$ *:*

  if $|\mathcal{S}| = 4^{\mathfrak{g}}$ **then return** $\mathcal{S}$

  Partition $\mathcal{S}$ into four arbitrary subsequences $\{\mathcal{S}_i\}_{i=1}^4$ each of size $n/4$

  for $i = 1, 2, 3, 4$ **do**

    $\widetilde{\mathcal{S}}_i \leftarrow$ COMPRESS$(\mathcal{S}_i, \mathfrak{g}, \mathbf{k}, \mathbf{k}', \delta)$ // run COMPRESS recursively to return coresets of size $2^{\mathfrak{g}} \cdot \sqrt{\frac{|\mathcal{S}|}{4}}$

  end

  $\widetilde{\mathcal{S}} \leftarrow$ CONCATENATE$(\widetilde{\mathcal{S}}_1, \widetilde{\mathcal{S}}_2, \widetilde{\mathcal{S}}_3, \widetilde{\mathcal{S}}_4)$ // combine the coresets to obtain a coreset of size $2 \cdot 2^{\mathfrak{g}} \cdot \sqrt{|\mathcal{S}|}$

return KT$(\widetilde{\mathcal{S}}, \mathbf{k}, \mathbf{k}', |\widetilde{\mathcal{S}}|^2\delta)$ // halve the coreset to size $2^{\mathfrak{g}}\sqrt{|\mathcal{S}|}$ via symmetrized kernel thinning

function KT $(\mathcal{S}, \mathbf{k}, \mathbf{k}', \delta)$ *:*

  // Identify kernel thinning coreset containing $\lfloor |\mathcal{S}|/2 \rfloor$ input points

  $\mathcal{S}_{\text{KT}} \leftarrow$ KT-SWAP$(\mathbf{k}$, KT-SPLIT$(\mathbf{k}', \mathcal{S}, \delta))$

return $\mathcal{S}_{\text{KT}}$ with probability $\frac{1}{2}$ and the complementary coreset $\mathcal{S} \setminus \mathcal{S}_{\text{KT}}$ otherwise

Algorithm 4: OPTHALVE4 – Optimal four-point halving

Input: kernel $\mathbf{k}$, point sequence $\mathcal{S}_{\text{in}} = (x_i)_{i=1}^4$

`K12_plus_K43` $\leftarrow \mathbf{k}(x_1, x_2) + \mathbf{k}(x_4, x_3)$; `K41_plus_K23` $\leftarrow \mathbf{k}(x_4, x_1) + \mathbf{k}(x_2, x_3)$; `K42_plus_K13` $\leftarrow \mathbf{k}(x_4, x_2) + \mathbf{k}(x_1, x_3)$

if `K12_plus_K43` $<$ `K41_plus_K23` **then**

  if `K12_plus_K43` $<$ `K42_plus_K13` **then**

    return (x_3, x_4) with probability $\frac{1}{2}$ and (x_1, x_2) otherwise

  end

  return (x_2, x_4) with probability $\frac{1}{2}$ and (x_1, x_3) otherwise

end

if `K41_plus_K23` $<$ `K42_plus_K13` **then**

  return (x_1, x_4) with probability $\frac{1}{2}$ and (x_2, x_3) otherwise

end

return (x_2, x_4) with probability $\frac{1}{2}$ and (x_1, x_3) otherwise

Algorithm 5a: KT-SPLIT – Divide points into candidate coresets of size $\lfloor n/2 \rfloor$

Input: kernel $\mathbf{k}'$, point sequence $\mathcal{S}_{\text{in}} = (x_i)_{i=1}^n$, failure probability δ

$\mathcal{S}^{(1)}, \mathcal{S}^{(2)} \leftarrow \{\}$ // Initialize empty coresets: $\mathcal{S}^{(1)}, \mathcal{S}^{(2)}$ have size i after round i
$\sigma \leftarrow 0$ // Initialize swapping parameter
for $i = 1, 2, \ldots, \lfloor n/2 \rfloor$ **do**
  // Consider two points at a time
  $(x, x') \leftarrow (x_{2i-1}, x_{2i})$
  // Compute swapping threshold $\mathfrak{a}$
  $\mathfrak{a}, \sigma \leftarrow$`get_swap_params`$(\sigma, \mathfrak{b}, \frac{\delta}{n})$ with $\mathfrak{b}^2 = \mathbf{k}'(x,x) + \mathbf{k}'(x',x') - 2\mathbf{k}'(x,x')$
  // Assign one point to each coreset after probabilistic swapping
  $\theta \leftarrow \sum_{j=1}^{2i-2}(\mathbf{k}'(x_j, x) - \mathbf{k}'(x_j, x')) - 2\sum_{z\in\mathcal{S}^{(1)}}(\mathbf{k}'(z, x) - \mathbf{k}'(z, x'))$
  $(x, x') \leftarrow (x', x)$ *with probability* $\min(1, \frac{1}{2}(1 - \frac{\theta}{\mathfrak{a}})_+)$
  $\mathcal{S}^{(1)}$`.append`(x); $\mathcal{S}^{(2)}$`.append`(x')
end
return $(\mathcal{S}^{(1)}, \mathcal{S}^{(2)})$, candidate coresets of size $\lfloor n/2 \rfloor$

function `get_swap_params`$(\sigma, \mathfrak{b}, \delta)$:
  $\mathfrak{a} \leftarrow \max(\mathfrak{b}\sigma\sqrt{2\log(2/\delta)}, \mathfrak{b}^2)$
  $\sigma^2 \leftarrow \sigma^2 + \mathfrak{b}^2(1 + (\mathfrak{b}^2 - 2\mathfrak{a})\sigma^2/\mathfrak{a}^2)_+$
return $(\mathfrak{a}, \sigma)$

Algorithm 5b: KT-SWAP – Identify and refine the best candidate coreset

Input: kernel $\mathbf{k}$, point sequence $\mathcal{S}_{\text{in}} = (x_i)_{i=1}^n$, candidate coresets $(\mathcal{S}^{(1)}, \mathcal{S}^{(2)})$

$\mathcal{S}^{(0)} \leftarrow$`baseline_coreset`$(\mathcal{S}_{\text{in}},$ `size`$= \lfloor n/2 \rfloor)$ // Compare to baseline (e.g., standard thinning)
$\mathcal{S}_{\text{KT}} \leftarrow \mathcal{S}^{(\ell^\star)}$ for $\ell^\star \leftarrow \arg\min_{\ell\in\{0,1,2\}} \text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}^{(\ell)})$ // Select best coreset
// Swap out each point in $\mathcal{S}_{\text{KT}}$ for best alternative in $\mathcal{S}_{\text{in}}$ while ensuring no point is repeated in $\mathcal{S}_{\text{KT}}$
for $i = 1, \ldots, \lfloor n/2 \rfloor$ **do**
  $\mathcal{S}_{\text{KT}}[i] \leftarrow \arg\min_{z\in\{\mathcal{S}_{\text{KT}}[i]\}\cup(\mathcal{S}_{\text{in}}\setminus\mathcal{S}_{\text{KT}})} \text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}_{\text{KT}}$ with $\mathcal{S}_{\text{KT}}[i] = z)$
end
return $\mathcal{S}_{\text{KT}}$, refined coreset of size $\lfloor n/2 \rfloor$

B Proof of Lem. 1: Quality of CoresetMMD

We first provide a discussion on the error inflation factor and then prove the two claims in Lem. 1.

B.1 On the KT-Compress error inflation factor

Given a point sequence $\mathcal{S}_{\text{in}}$, a positive integer n, and a scalar $\delta \in (0,1)$, the inflation factor $\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathcal{S}_{\text{in}}, n, \delta, \mathfrak{g})$ denotes the smallest scalar of the form

$$\begin{aligned}\mathbf{R}^2_{\mathbf{k},\mathbf{k}'}(\mathcal{S}_{\text{in}}, n, \delta, \mathfrak{g}) =& 256(\log_4 n - \mathfrak{g} - 1)(C_{\mathbf{k},\mathbf{k}'}(\mathcal{S}_{\text{in}}) + \mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathcal{S}_{\text{in}}, \delta, 2^{\mathfrak{g}+1}\sqrt{n})\sqrt{\log(\tfrac{3n(\log_4 n - \mathfrak{g} - 1)}{\delta})})^2 \\ &\cdot (\sqrt{\log(n+1)} + \sqrt{\log(2/\delta)})^2,\end{aligned} \tag{13}$$

where $C_{\mathbf{k},\mathbf{k}'}(\mathcal{S}_{\text{in}})$ and $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathcal{S}_{\text{in}}, \delta, 2^{\mathfrak{g}+1}\sqrt{n})$ are any scalars satisfying the property that, on an event of probability at least $1 - \frac{\delta}{2}$, every KT-Compress (Alg. 3) call to KT with an input of size ℓ (that is a subset of $\mathcal{S}_{\text{in}}$) is $\mathbf{k}$-sub-Gaussian (see Shetty et al. (2022, Def. 2)) with parameters

$$a_{\ell,n} = \frac{2C_{\mathbf{k},\mathbf{k}'}(\mathcal{S}_{\text{in}})}{\ell} \quad \text{and} \quad v_{\ell,n} = \frac{2\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathcal{S}_{\text{in}},\delta,\ell)}{\ell}\sqrt{\log(\tfrac{12n4^{\mathfrak{g}}(\log_4 n - \mathfrak{g})}{\ell\delta})}. \tag{14}$$

See App. B.4.1 for the valid values of $C_{\mathbf{k},\mathbf{k}'}$, and $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}$ derived in prior work for standard choices of $\mathbf{k}'$. (Notably, $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}$ is non-decreasing in its last argument.)

The factor $\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, n, \delta, \mathfrak{g})$ is the population analogue of $\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Y}_n, n, \delta, \mathfrak{g})$ and is defined as the smallest scalar of the form

$$\mathbf{R}^2_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n,\delta,\mathfrak{g}) \triangleq 256(\log_4 n-\mathfrak{g}-1)(C_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n,\delta)+\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n,\delta,2^{\mathfrak{g}+1}\sqrt{n})\sqrt{\log(\tfrac{3n(\log_4 n-\mathfrak{g}-1)}{\delta})})^2 \cdot (\sqrt{\log(n+1)}+\sqrt{\log(2/\delta)})^2,$$

where $C_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n,\delta)$ and $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n,\delta,2^{\mathfrak{g}+1}\sqrt{n})$ satisfy

$$\mathbb{P}\big[C_{\mathbf{k},\mathbf{k}'}(\mathbb{Y}_n) \le C_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n,\delta) \text{ and } \mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{Y}_n,\delta,2^{\mathfrak{g}+1}\sqrt{n}) \le \mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n,\delta,2^{\mathfrak{g}+1}\sqrt{n})\big] \ge 1-\delta/2 \tag{15}$$

and $C_{\mathbf{k},\mathbf{k}'}(\mathbb{Y}_n)$ and $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{Y}_n,\delta,2^{\mathfrak{g}+1}\sqrt{n})$ satisfy the sub-Gaussian property (14) defined above when $\mathcal{S}_{\mathrm{in}} = \mathbb{Y}_n$. See App. B.4.2 for upper bounds on the quantities $C_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n,\delta)$ and $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n,\delta,2^{\mathfrak{g}+1}\sqrt{n})$ for standard choices of $\mathbf{k}'$ and App. B.5 for how that translates to a scaling for $\mathbf{R}_{\mathbf{k},\mathbf{k}'}$ for the settings in Tab. 1.

B.2 Proof of claim (6)

We follow the notation of Shetty et al. (2022, Apps. A and C) and note that KT-COMPRESS (Alg. 3) is the COMPRESS algorithm of Shetty et al. (2022) with, in the notation of Shetty et al. (2022, Example 4), HALVE $=$ KT$(\frac{\ell^2}{n4^{\mathfrak{g}+1}(\beta_n+1)}\delta)$ and $\beta_n \triangleq \log_4 n-\mathfrak{g}-1$ for an input of size ℓ.

We associate with each algorithm ALG and each input point sequence $\mathcal{S}_{\mathrm{in}}$ of size n and output $\mathcal{S}_{\mathrm{out}}$ of size n_{out} the measure difference

$$\phi_{\mathrm{ALG}}(\mathcal{S}_{\mathrm{in}}) = \tfrac{1}{n}\textstyle\sum_{x\in\mathcal{S}_{\mathrm{in}}}\delta_x - \tfrac{1}{n_{out}}\sum_{x\in\mathcal{S}_{\mathrm{out}}}\delta_x,$$

and the unnormalized measure difference

$$\psi_{\mathrm{ALG}}(\mathcal{S}_{\mathrm{in}}) = n\cdot\phi_{\mathrm{ALG}}(\mathcal{S}_{\mathrm{in}}) = \textstyle\sum_{x\in\mathcal{S}_{\mathrm{in}}}\delta_x - \tfrac{n}{n_{out}}\sum_{x\in\mathcal{S}_{\mathrm{out}}}\delta_x.$$

Shetty et al. (2022, Eqn. 18) show that for their COMPRESS algorithm, the following holds:

$$\psi_{\mathrm{C}}(\mathcal{S}_{\mathrm{in}}) = \sqrt{n}2^{-\mathfrak{g}-1}\textstyle\sum_{i=0}^{\beta_n}\sum_{j=1}^{4^i}2^{-i}\psi_H(\mathcal{S}^{\mathrm{in}}_{i,j}), \tag{16}$$

where $\beta_n \triangleq \log_4 n-\mathfrak{g}-1$, $(\mathcal{S}^{\mathrm{in}}_{i,j})_{j\in[4^i]}$ are the 4^i coresets of size $n_i = 2^{\mathfrak{g}+1-i}\sqrt{n}$ resulting from i recursive calls to the COMPRESS algorithm, and ψ_H is the unnormalized measure difference for HALVE. Substituting $\mathcal{S}_{\mathrm{in}} \leftarrow \mathbb{X}^{(r)}_m$, we get that for $r=1,\dots,s$, $\psi_{\mathrm{C}}(\mathbb{X}^{(r)}_m) = \sqrt{m/s_m}\cdot 2^{-\mathfrak{g}-1}\sum_{i=0}^{\beta_{m/s_m}}\sum_{j=1}^{4^i}2^{-i}\psi_H(\mathbb{X}^{(r)}_{m,i,j})$, where $\mathbb{X}^{(r)}_{m,i,j}$ is defined analogously to $\mathcal{S}^{\mathrm{in}}_{i,j}$. If we let C+ be the algorithm that maps $\mathbb{X}_m$ to $\hat{\mathbb{X}}_m$ (and $\mathbb{Y}_n$ to $\hat{\mathbb{Y}}_n$), we obtain

$$\psi_{\mathrm{C+}}(\mathbb{X}^{(r)}_m) = \sqrt{m/s_m}\cdot 2^{-\mathfrak{g}-1}\textstyle\sum_{r=1}^{s_m}\sum_{i=0}^{\beta_{m/s_m}}\sum_{j=1}^{4^i}2^{-i}\psi_H(\mathbb{X}^{(r)}_{m,i,j}).$$

Similarly,

$$\psi_{\mathrm{C+}}(\mathbb{Y}^{(r)}_n) = \sqrt{n/s_n}\cdot 2^{-\mathfrak{g}-1}\textstyle\sum_{r=1}^{s_n}\sum_{i=0}^{\beta_{n/s_n}}\sum_{j=1}^{4^i}2^{-i}\psi_H(\mathbb{Y}^{(r)}_{n,i,j}).$$

Following App. C.1 from Shetty et al. (2022), if one numbers the elements of $\mathcal{S}_{\mathrm{in}}$ as $(x_1,\dots,x_n)$, and defines the $n\times n$ kernel matrix $\mathbf{K} \triangleq (\mathbf{k}(x_i,x_j))_{i,j=1}^n$, one obtains

$$u_{k,j} \triangleq \mathbf{K}^{\frac{1}{2}}\textstyle\sum_{i=1}^n e_i\Big(\mathbf{1}(x_i\in\mathcal{S}^{\mathrm{in}}_{k,j}) - 2\cdot\mathbf{1}(x_i\in\mathcal{S}^{\mathrm{out}}_{k,j})\Big), \text{ and } u_{\mathrm{C}} \triangleq \sum_{k=0}^{\log_4 n-\mathfrak{g}-1}\sum_{j=1}^{4^k} w_{k,n}u_{k,j},$$

where $w_{k,n} \triangleq \frac{\sqrt{n}}{2^{\mathfrak{g}+1+k}}$. Then, we have

$$n^2\cdot\mathrm{MMD}^2_{\mathbf{k}}(\mathcal{S}_{\mathrm{in}},\mathcal{S}_{\mathrm{C}}) = \|u_{\mathrm{C}}\|_2^2, \quad \text{and}$$
$$\mathbb{E}[u_{k,j}|(u_{k',j'}: j'\in[4^{k'}], k'>k)] = 0 \quad \text{for} \quad k=0,\dots,\log_4 n-\mathfrak{g}-2,$$

and $u_{k,j}$ for $j\in[4^k]$ are conditionally independent given $(u_{k',j'}: j'\in[4^{k'}], k'>k)$. This follows easily from (16). For any $u\in\mathbb{R}^n$ for arbitrary n, we also define

$$\mathbf{M}_u \triangleq \begin{pmatrix} 0 & u^\top \\ u & \mathbf{0}_{n\times n}\end{pmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}.$$

For any $u \in \mathbb{R}^n$, the matrix $\mathbf{M}_u$ satisfies

$$\|\mathbf{M}_u\|_{\text{op}} = \|u\|_2 = \lambda_{\max}(\mathbf{M}_u), \quad \text{and} \quad \mathbf{M}_u^q \preceq \|u\|_2^q \mathbf{I}_{n+1} \text{ for all } q \in \mathbb{N}.$$

Defining the shorthand $\mathbf{M}_{k,n} \triangleq \mathbf{M}_{w_{k,n} u_{k,j}}$, we find that

$$n \, \mathrm{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}_{\text{C}}) = \|u_{\text{C}}\|_2 = \lambda_{\max}(\mathbf{M}_{u_{\text{C}}}) = \lambda_{\max}(\textstyle\sum_{k=0}^{\log_4 n - \mathfrak{g} - 1} \sum_{j=1}^{4^k} \mathbf{M}_{k,j}),$$

Let $\mathbb{U} = (U_i)_{i=1}^{m+n}$ be the sequence obtained as the concatenation of $\mathbb{X}_m$ and $\mathbb{Y}_n$. Define the matrix $\mathbf{K}_{\mathbb{X}_m, \mathbb{Y}_n} \triangleq (\mathbf{k}(U_i, U_j))_{i,j=1}^{m+n}$. Substituting $\mathcal{S}_{\text{in}} \leftarrow \mathbb{X}_m^{(r)}$, we can write

$$u_{k,j,\mathbb{X}_m}^{(r)} \triangleq \mathbf{K}_{\mathbb{X}_m, \mathbb{Y}_n}^{\frac{1}{2}} \textstyle\sum_{i=1}^n e_i \Big(\mathbf{1}(x_i \in \mathbb{X}_{m,k,j}^{(r)}) - 2 \cdot \mathbf{1}(x_i \in \mathbb{X}_{m,k,j}^{(r),\text{out}}) \Big),$$
$$\text{and } u_{\text{C},\mathbb{X}_m}^{(r)} \triangleq \textstyle\sum_{k=0}^{\log_4 (m/s_m) - \mathfrak{g} - 1} \sum_{j=1}^{4^k} w_{k,m/s_m} u_{k,j,\mathbb{X}_m}^{(r)}, \quad \mathbf{M}_{k,j,\mathbb{X}_m}^{(r)} \triangleq \mathbf{M}_{w_{k,m/s} u_{k,j,\mathbb{X}_m}^{(r)}}$$

Hence, we can define

$$u_{\text{C}+,\mathbb{X}_m} \triangleq \textstyle\sum_{r=1}^{s_m} \sum_{k=0}^{\log_4 (m/s_m) - \mathfrak{g} - 1} \sum_{j=1}^{4^k} w_{k,m/s_m} u_{k,j,\mathbb{X}_m}^{(r)}.$$

Analogously,

$$u_{\text{C}+,\mathbb{Y}_n} \triangleq \textstyle\sum_{r=1}^{s_n} \sum_{k=0}^{\log_4 (n/s_n) - \mathfrak{g} - 1} \sum_{j=1}^{4^k} w_{k,n/s_n} u_{k,j,\mathbb{Y}_n}^{(r)}.$$

Also, note that

$$\begin{aligned} |\, \mathrm{MMD}_{\mathbf{k}}(\mathbb{X}_m, \mathbb{Y}_n) - \mathrm{MMD}_{\mathbf{k}}\Big(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n\Big)| &= |\|(\mathbb{P} - \mathbb{Q})\mathbf{k}\|_{\mathbf{k}} - \|(\mathbb{P}_m - \mathbb{Q}_n)\mathbf{k}\|_{\mathbf{k}}| \le \|(\mathbb{P} - \mathbb{Q} - (\mathbb{P}_m - \mathbb{Q}_n))\mathbf{k}\|_{\mathbf{k}} \\ &= \|(\mathbb{P} - \mathbb{P}_m)\mathbf{k} - (\mathbb{Q} - \mathbb{Q}_n)\mathbf{k}\|_{\mathbf{k}} \end{aligned}$$

which implies that

$$\begin{aligned} &|\, \mathrm{MMD}_{\mathbf{k}}(\mathbb{X}_m, \mathbb{Y}_n) - \mathrm{MMD}_{\mathbf{k}}\Big(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n\Big)| \le \|u_{\text{C}+,\mathbb{X}_m}/m - u_{\text{C}+,\mathbb{Y}_n}/n\|_2 \\ &= \Big\| \textstyle\sum_{r=1}^{s_m} \big(\frac{1}{m} \sum_{k=0}^{\log_4 (m/s_m) - \mathfrak{g} - 1} \sum_{j=1}^{4^k} w_{k,m/s_m} u_{k,j,\mathbb{X}_m}^{(r)} - \frac{1}{n} \sum_{k=0}^{\log_4 (n/s_n) - \mathfrak{g} - 1} \sum_{j=1}^{4^k} w_{k,n/s_n} u_{k,j,\mathbb{Y}_n}^{(r)} \big) \Big\|_2 \\ &= \lambda_{\max}\big(\textstyle\sum_{r=1}^{s_m} \big(\frac{1}{m} \sum_{k=0}^{\log_4 (m/s_m) - \mathfrak{g} - 1} \sum_{j=1}^{4^k} \mathbf{M}_{k,j,\mathbb{X}_m}^{(r)} - \frac{1}{n} \sum_{k=0}^{\log_4 (n/s_n) - \mathfrak{g} - 1} \sum_{j=1}^{4^k} \mathbf{M}_{k,j,\mathbb{Y}_n}^{(r)} \big)\big). \end{aligned}$$

Now we apply the sub-Gaussian matrix Freedman inequality (Shetty et al., 2022, Lem. 4). The zero-mean condition on the matrices holds following the argument in Shetty et al. (2022, Sec. C.3.1), while for the moment bounds we use the approach in their Sec. C.3.2. Namely, we use that for any $q \in 2\mathbb{N}$,

$$(\mathbf{M}_{k,j,\mathbb{X}_m}^{(r)})^q = \mathbf{M}_{w_{k,m/4} u_{k,j,\mathbb{X}_m}^{(r)}}^q \preceq \|w_{k,m/4} u_{k,j,\mathbb{X}_m}^{(r)}\|_2^q \mathbf{I}_{n+1} = w_{k,m/4}^q \|u_{k,j,\mathbb{X}_m}^{(r)}\|_2^q \mathbf{I}_{m+n+1},$$

and similarly, $(\mathbf{M}_{k,j,\mathbb{X}_m}^{(r)})^q \preceq w_{k,m/4}^q \|u_{k,j,\mathbb{X}_m}^{(r)}\|_2^q \mathbf{I}_{m+n+1}$. Shetty et al. (2022, Lem. 5) prove that for any non-negative random variable Z,

$$\mathbb{P}[Z > a + v\sqrt{t}] \le e^{-t} \text{ for all } t \ge 0 \implies \mathbb{E}[Z^q] \le (2a + 2v)^q (\tfrac{q}{2})! \text{ for all } q \in 2\mathbb{N}.$$

In their case, their $\mathbf{k}$-sub Gaussian assumption on HALVE implies that

$$\mathbb{P}[\|u_{k,j}\|_2 \ge \ell'_k (a_{\ell'_k} + v_{\ell'_k} \sqrt{t}) \mid (u_{k',j'} : j' \in [4^{k'}], k' > k)] \le e^{-t} \text{ for all } t \ge 0,$$

for suitable scalar sequences $\{a_\ell, v_\ell\}$ (also see (14)), where $\ell'_k \triangleq \sqrt{n} 2^{\mathfrak{g}+1-k}$, which yields the moment bound $\mathbb{E}[\|u_{k,j}\|_2^q \mid (u_{k',j'} : j' \in [4^{k'}], k' > k)] \le (\frac{q}{2})! (2\ell'_k (a_{\ell'_k} + v_{\ell'_k}))^q$. Under the same assumption on HALVE, we obtain analogously that

$$\mathbb{E}[\|u_{k,j,\mathbb{X}_m}^{(r)}\|_2^q \mid (u_{k',j',\mathbb{X}_m}^{(r)} : j' \in [4^{k'}], k' > k)] \le (\tfrac{q}{2})! (2\ell'_{k,m/s_m} (a_{\ell'_{k,m/s_m}} + v_{\ell'_{k,m/s_m}}))^q,$$
$$\mathbb{E}[\|u_{k,j,\mathbb{Y}_n}^{(r)}\|_2^q \mid (u_{k',j',\mathbb{Y}_n}^{(r)} : j' \in [4^{k'}], k' > k)] \le (\tfrac{q}{2})! (2\ell'_{k,n/s_n} (a'_{\ell'_{k,n/s_n}} + v'_{\ell'_{k,n/s_n}}))^q,$$

where $\ell'_{k,m/s_m} \triangleq \sqrt{m/s_m} \cdot 2^{\mathfrak{g}+1-k}$, $\ell'_{k,n/s_n} \triangleq \sqrt{n/s_n} \cdot 2^{\mathfrak{g}+1-k}$.

Now let $\{a_{\ell,n}, v_{\ell,n}\}$ denote the scalar sequences from (14) so that COMPRESS with HALVE $= \mathrm{KT}(\frac{\ell^2}{n4^{\mathfrak{g}+1}(\beta_n+1)}\delta)$ for input of size ℓ, every HALVE call invoked by COMPRESS is $\mathbf{k}$-sub-Gaussian with parameters $a_{\ell,n}, v_{\ell,n}$ on an event of probability at least $1 - \frac{\delta}{2}$. We define

$$\sigma^2 \triangleq \textstyle\sum_{r=1}^{s_m}\big(\sum_{k=0}^{\log_4(m/s_m)-\mathfrak{g}-1} \sum_{j=1}^{4^k}(\frac{2}{m} w_{k,m/s_m} \ell'_{k,m/s_m} (a_{\ell'_{k,m/s_m},m/s_m} + v_{\ell'_{k,m/s_m},m/s_m}))^2 \\ + \sum_{k=0}^{\log_4(n/s_n)-\mathfrak{g}-1} \sum_{j=1}^{4^k}(\frac{2}{m} w_{k,n/s_n} \ell'_{k,n/s_n} (a'_{\ell'_{k,n/s_n},n/s_n} + v'_{\ell'_{k,n/s_n},n/s_n}))^2\big),$$

which when combined with the expressions for $a_{\ell,n}$ and $v_{\ell,n}$ from (14), yields that

$$\begin{aligned}\sigma^2 &= \textstyle\sum_{k=0}^{\log_4(m/s_m)-\mathfrak{g}-1}\Bigg(\frac{4}{2^{\mathfrak{g}}\sqrt{m}}\bigg(C_{\mathbf{k},\mathbf{k}'}(\mathbb{X}_m) + \mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{X}_m, \delta, 2^{\mathfrak{g}+1-k}\sqrt{\frac{m}{s_m}})\sqrt{\log(\frac{6\cdot 4^{\mathfrak{g}}\sqrt{m/s_m}(\beta_{m/s_m}+1)}{2^{\mathfrak{g}-k}\delta})}\bigg)\Bigg)^2 && (17)\\ &\quad + \textstyle\sum_{k=0}^{\log_4(n/s_n)-\mathfrak{g}-1}\Bigg(\frac{4}{2^{\mathfrak{g}}\sqrt{n}}\bigg(C_{\mathbf{k},\mathbf{k}'}(\mathbb{Y}_n) + \mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{Y}_n, \delta, 2^{\mathfrak{g}+1-k}\sqrt{\frac{n}{s_n}})\sqrt{\log(\frac{6\cdot 4^{\mathfrak{g}}\sqrt{n/s_n}(\beta_{n/s_n}+1)}{2^{\mathfrak{g}-k}\delta})}\bigg)\Bigg)^2 \\ &\le \tfrac{16(\log_4(m/s_m)-\mathfrak{g}-1)}{4^{\mathfrak{g}}m}\bigg(C_{\mathbf{k},\mathbf{k}'}(\mathbb{X}_m) + \mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{X}_m, \delta, 2^{\mathfrak{g}+1}\sqrt{\frac{m}{s_m}})\sqrt{\log(\frac{3m(\log_4(m/s_m)-\mathfrak{g}-1)}{s_m\delta})}\bigg)^2 \\ &\quad + \tfrac{16(\log_4(n/s_n)-\mathfrak{g}-1)}{4^{\mathfrak{g}}n}\bigg(C_{\mathbf{k},\mathbf{k}'}(\mathbb{Y}_n) + \mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{Y}_n, \delta, 2^{\mathfrak{g}+1}\sqrt{\frac{n}{s_n}})\sqrt{\log(\frac{3n(\log_4(n/s_n)-\mathfrak{g}-1)}{s_n\delta})}\bigg)^2, && (18)\end{aligned}$$

where in the last inequality we also use the fact that $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}$ is non-decreasing in its last argument. Now, by the sub-Gaussian matrix Freedman inequality as stated in (Shetty et al., 2022, Lem. 4), we obtain that

$$\begin{aligned}&\Pr(|\,\mathrm{MMD}_{\mathbf{k}}(\mathbb{X}_m, \mathbb{Y}_n) - \mathrm{MMD}_{\mathbf{k}}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)| > \sigma\sqrt{8(\log(m+n+1)+t)}) && (19)\\ &\le \Pr(\lambda_{\max}(\textstyle\sum_{r=1}^{4}(\frac{1}{m}\sum_{k=0}^{\beta_{m/4}}\sum_{j=1}^{4^k}\mathbf{M}^{(r)}_{k,j,\mathbb{X}_m} - \frac{1}{n}\sum_{k=0}^{\beta_{n/4}}\sum_{j=1}^{4^k}\mathbf{M}^{(r)}_{k,j,\mathbb{Y}_n})) > \sigma\sqrt{8(\log(m+n+1)+t)}) \\ &\le \tfrac{\delta}{2} + e^{-t}, \qquad \text{for all} \quad t \ge 0.\end{aligned}$$

The term $\frac{\delta}{2}$ does not come from the sub-Gaussian matrix Freedman inequality but rather from the conditioning on the event for which (14) holds. Equation (19) in turn implies that $\Pr(|\,\mathrm{MMD}_{\mathbf{k}}(\mathbb{X}_m, \mathbb{Y}_n) - \mathrm{MMD}_{\mathbf{k}}\Big(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n\Big)| > \sigma(\sqrt{8\log(m+n+1)} + \sqrt{8t})) \le \frac{\delta}{2} + e^{-t}$. Equivalently, for any $\delta > 0$, with probability at least $1 - \delta$,

$$|\,\mathrm{MMD}_{\mathbf{k}}(\mathbb{X}_m, \mathbb{Y}_n) - \mathrm{MMD}_{\mathbf{k}}\Big(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n\Big)| \le \sqrt{8}\sigma(\sqrt{\log(m+n+1)} + \sqrt{\log(2/\delta)}). \tag{20}$$

Putting together the bound (20) with the upper bound (18) and the definition (13) of $\mathbf{R}_{\mathbf{k},\mathbf{k}'}$ immediately yields the claimed bound (6) and we are done.

B.3 Proof of claim (7)

Note that

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) - \mathrm{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n) = \mathrm{MMD}(\mathbb{P}, \mathbb{Q}) - \mathrm{MMD}(\mathbb{X}_m, \mathbb{Y}_n) + \mathrm{MMD}(\mathbb{X}_m, \mathbb{Y}_n) - \mathrm{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n).$$

The second term in the display above can be bounded via (6) and the first term via the following result from Gretton et al. (2012a):

Lemma 2 (Adapted from Theorem 7, Gretton et al. (2012a))**.** *Assume that* $\|\mathbf{k}\|_\infty < +\infty$. *Then,*

$$\Pr\bigg[|\mathrm{MMD}(\mathbb{X}_m, \mathbb{Y}_n) - \mathrm{MMD}(\mathbb{P}, \mathbb{Q})| > 2\bigg(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{\|\mathbf{k}\|_\infty}{n}}\bigg) + \epsilon\bigg] \le 2\exp\bigg(\tfrac{-\epsilon^2 mn}{2\|\mathbf{k}\|_\infty(m+n)}\bigg).$$

Using Lem. 2 with $\delta = 2\exp\bigg(\frac{-\epsilon^2 mn}{2\|\mathbf{k}\|_\infty(m+n)}\bigg)$, which is equivalent to $\epsilon = \sqrt{\frac{2\|\mathbf{k}\|_\infty(m+n)}{mn}\log\big(\frac{2}{\delta}\big)}$, we obtain that

$$\Pr\bigg[|\mathrm{MMD}(\mathbb{X}_m, \mathbb{Y}_n) - \mathrm{MMD}(\mathbb{P}, \mathbb{Q})| > 2\big(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{\|\mathbf{k}\|_\infty}{n}}\big) + \sqrt{\tfrac{2\|\mathbf{k}\|_\infty(m+n)}{mn}\log\big(\tfrac{2}{\delta}\big)}\bigg] \le \delta,$$

where the bound (18) on σ depends on $(\mathbb{X}_m, \mathbb{Y}_n)$. Next, the bound (6) states that with probability at least $1-\delta$,

$$|\operatorname{MMD}(\mathbb{X}_m, \mathbb{Y}_n) - \operatorname{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)| \leq \frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{X}_m, m, \delta, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + \frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Y}_n, n, \delta \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}},$$

and the definition of $\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, \mathfrak{g})$ (see the discussion around (15) in App. B.1) implies that

$$\mathbb{P}\big[\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{X}_m, m, \delta, \mathfrak{g}) \leq \mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, \mathfrak{g}) \quad \text{and} \quad \mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Y}_n, n, \delta, \mathfrak{g}) \leq \mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, n, \delta, \mathfrak{g})\big] \geq 1-\delta.$$

Putting the pieces together yields the bound (7) with probability at least $1-3\delta$ as claimed.

B.4 Bounds on $C_{\mathbf{k},\mathbf{k}'}$ and $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}$

First, we discuss bounds on the sample-based quantities $C_{\mathbf{k},\mathbf{k}'}(\mathbb{X}_m)$ and $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{X}_m, \delta, 2^{\mathfrak{g}+1}\sqrt{n})$ defined in (14) followed by bounds on its population analog $C_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta)$ and $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, 2^{\mathfrak{g}+1}\sqrt{n})$ defined in (15).

B.4.1 Bounds on sample-level quantities ($C_{\mathbf{k},\mathbf{k}'}(\mathbb{X}_m)$ and $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{X}_m, \delta, 2^{\mathfrak{g}+1}\sqrt{n})$)

We discuss the default choices $\mathbf{k}' = \mathbf{k}_{\mathrm{rt}}$ and $\mathbf{k}' = \mathbf{k}$ and the more general case one-by-one.

Case I: $\mathbf{k}' = \mathbf{k}_{\mathrm{rt}}$ For this case, we follow the discussion from Dwivedi and Mackey (2021, Sec. 3.1). Let $L_{\mathbf{k}_{\mathrm{rt}}}$ denote the Lipschitz constant of $\mathbf{k}_{\mathrm{rt}}$ and define $\tau_{\mathbf{k}_{\mathrm{rt}}}(R) \triangleq (\sup_x \int_{\|y\|_2 \geq R} \mathbf{k}_{\mathrm{rt}}^2(x, x-y)dy)^{\frac{1}{2}}$,

$$\mathfrak{S}_{\mathbf{k}_{\mathrm{rt}},\ell} \triangleq \min\{r : \sup_{\substack{x,y:\\ \|x-y\|_2 \geq r}} |\mathbf{k}_{\mathrm{rt}}(x,y)| \leq \tfrac{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}{\ell}\}, \quad \mathfrak{S}^\dagger_{\mathbf{k}_{\mathrm{rt}},\ell} \triangleq \min\{r : \tau_{\mathbf{k}_{\mathrm{rt}}}(r) \leq \tfrac{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}{\ell}\},$$
$$\mathfrak{S}_{\mathbb{X}_m} \triangleq \max_{x \in \mathbb{X}_m} \|x\|_2, \quad \text{and} \quad \mathfrak{S}_{\mathbb{X}_m,\mathbf{k}_{\mathrm{rt}},\ell} \triangleq \min\big(\mathfrak{S}_{\mathbb{X}_m}, \ell^{1+\frac{1}{d}} \mathfrak{S}_{\mathbf{k}_{\mathrm{rt}},\ell} + \ell^{\frac{1}{d}} \tfrac{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}{L_{\mathbf{k}_{\mathrm{rt}}}}\big),$$

and the kernel thinning inflation factor

$$\mathfrak{N}_{\mathbf{k}_{\mathrm{rt}}}(\ell,d,\delta,R) \triangleq 37\sqrt{\log\big(\tfrac{3\ell}{\delta}\big)}\left[\sqrt{\log\big(\tfrac{8}{\delta}\big)} + 5\sqrt{d\log\Big(2 + 2\tfrac{L_{\mathbf{k}_{\mathrm{rt}}}}{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}\big(\mathfrak{S}_{\mathbf{k}_{\mathrm{rt}},\ell} + R\big)\Big)}\right].$$

Then using Dwivedi and Mackey (2021, Thm. 1) in Shetty et al. (2022, Example 4), we find that

$$C_{\mathbf{k},\mathbf{k}_{\mathrm{rt}}}(\mathbb{X}_m) = 2\|\mathbf{k}_{\mathrm{rt}}\|_{\infty,\mathrm{in}} \quad \text{and} \quad \mathfrak{M}_{\mathbf{k},\mathbf{k}_{\mathrm{rt}}}(\mathbb{X}_m, \delta, \ell) = \|\mathbf{k}_{\mathrm{rt}}\|_{\infty,\mathrm{in}}(\max(\mathfrak{S}_{\mathbb{X}_m}, \mathfrak{S}^\dagger_{\mathbf{k}_{\mathrm{rt}},\ell/2}))^{\frac{d}{2}} \cdot \mathfrak{N}_{\mathbf{k}_{\mathrm{rt}}}(\ell,d,\delta,\mathfrak{S}_{\mathbb{X}_m,\mathbf{k}_{\mathrm{rt}},\ell}),$$

and $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{Y}_n, \delta, \ell)$ defined analogously by replacing $\mathbb{X}_m$ by $\mathbb{Y}_n$ and m by n. Here $\|\mathbf{k}'\|_{\infty,\mathrm{in}} = \sup_{x \in \mathbb{X}_m} \mathbf{k}'(x,x)$. We note that the bounds in Dwivedi and Mackey (2021); Shetty et al. (2022) are stated with $\|\mathbf{k}'\|_\infty \triangleq \sup_x \mathbf{k}'(x,x)$ instead of $\|\mathbf{k}'\|_{\infty,\mathrm{in}}$ (and note that $\|\mathbf{k}'\|_{\infty,\mathrm{in}} \leq \|\mathbf{k}'\|_\infty$). However, as noted in Dwivedi and Mackey (2022, App. B), all the results of Dwivedi and Mackey (2021) (which is what Shetty et al. (2022) build on) go through with $\|\mathbf{k}'\|_\infty$ replaced by $\|\mathbf{k}'\|_{\infty,\mathrm{in}}$ thereby yielding the result stated above.

Case II: $\mathbf{k}' = \mathbf{k}$ For this case, we follow the discussion in (Dwivedi and Mackey, 2022, Sec 2.2). In particular, for a set $\mathcal{A} \subset \mathbb{R}^d$ and scalar $\varepsilon > 0$, define the $\mathbf{k}$ *covering number* $\mathcal{N}_{\mathbf{k}}(\mathcal{A}, \varepsilon)$ with $\mathcal{M}_{\mathbf{k}}(\mathcal{A}, \varepsilon) \triangleq \log \mathcal{N}_{\mathbf{k}}(\mathcal{A}, \varepsilon)$ as the minimum cardinality of a set $\mathcal{C} \subset \mathbb{B}_{\mathbf{k}} \triangleq \{f : \|f\|_{\mathbf{k}} \leq 1\}$ satisfying

$$\mathbb{B}_{\mathbf{k}} \subseteq \bigcup_{h \in \mathcal{C}} \{g \in \mathbb{B}_{\mathbf{k}} : \sup_{x \in \mathcal{A}} |h(x) - g(x)| \leq \varepsilon\}.$$

Then choosing $\varepsilon = \frac{\sqrt{\|\mathbf{k}\|_{\infty,\mathrm{in}}}}{\ell/2}$ in the notation of Dwivedi and Mackey (2022, Thm. 2) and combining that result with Shetty et al. (2022, Example 4), we conclude that we can use the following bounds

$$C_{\mathbf{k},\mathbf{k}}(\mathbb{X}_m) = 2\sqrt{\|\mathbf{k}\|_{\infty,\mathrm{in}}} \quad \text{and} \quad \mathfrak{M}_{\mathbf{k},\mathbf{k}}(\mathbb{X}_m, \delta, \ell) = \sqrt{\tfrac{8\|\mathbf{k}\|_{\infty,\mathrm{in}}}{3} \log\big(\tfrac{12\log\ell}{\delta}\big)\big[\log\big(\tfrac{8}{\delta}\big) + \mathcal{M}_{\mathbf{k}}(\mathcal{A}_{\mathbb{X}_m}, (\ell/2)^{-1})\big]}, \qquad (21)$$

where $\mathcal{A}_{\mathbb{X}_m} = \{x : \|x\|_2 \leq \mathfrak{S}_{\mathbb{X}_m}\}$. We can define $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{Y}_n, \delta, \ell)$ analogously by replacing $\mathbb{X}_m$ by $\mathbb{Y}_n$ and m by n.

Case III: General $\mathbf{k}'$ When $\mathbf{k}'$ is neither of the two default choices ($\mathbf{k}$ or $\mathbf{k}_{\mathrm{rt}}$) like in ACTT, then the expressions for $\mathbf{R}_{\mathbf{k},\mathbf{k}'}$ can be derived using Dwivedi and Mackey (2022, Thm. 2-4) and Dwivedi and Mackey (2021, Thm. 1-2). For instance, when the RKHS of $\mathbf{k}$ is contained in the RKHS of $\mathbf{k}'$, we can apply the sub-Gaussian tail bounds for a single f (Dwivedi and Mackey (2021, Thm. 1)) and then apply a union bound with a covering argument for the ball $\{\|f\|_{\mathbf{k}} \leq 1\}$ (Dwivedi and Mackey (2022, Thm. 2)), in which case $\mathbf{R}_{\mathbf{k},\mathbf{k}'}$ also scales with $\sup_{\|f\|_{\mathbf{k}} \leq 1} \|f\|_{\mathbf{k}'}$. See Rem. 6 for an example of this case.

B.4.2 Bounds on population-level quantities ($C_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, \mathfrak{g})$ and $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, 2^{\mathfrak{g}+1}\sqrt{n})$)

Define

$$C'_{\mathbf{k},\mathbf{k}'} = \begin{cases} 2\|\mathbf{k}_{\mathrm{rt}}\|_\infty & \text{when} \quad \mathbf{k}' = \mathbf{k}_{\mathrm{rt}} \\ 2\sqrt{\|\mathbf{k}\|_\infty} & \text{when} \quad \mathbf{k}' = \mathbf{k} \end{cases}.$$

Then for the choices of $C_{\mathbf{k},\mathbf{k}'}(\cdot)$ in App. B.4.1, we have $\max\{C_{\mathbf{k},\mathbf{k}'}(\mathbb{X}_m), C_{\mathbf{k},\mathbf{k}'}(\mathbb{Y}_n)\} \leq C'_{\mathbf{k},\mathbf{k}'}$ almost surely, where $\|\mathbf{k}\|_\infty \triangleq \sup_{x\in\mathcal{X}} \mathbf{k}(x,x)$. Thus if we set $C_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta) = C_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, n, \delta) = C'_{\mathbf{k},\mathbf{k}'}$, to satisfy (15), it remains to determine $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, 2^{\mathfrak{g}+1}\sqrt{n})$ such that $\mathbb{P}\big[\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{X}_m, \delta, 2^{\mathfrak{g}+1}\sqrt{n}) \leq \mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, 2^{\mathfrak{g}+1}\sqrt{n})\big] \geq 1 - \delta/2$ for the choices of $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}$ in App. B.4.1.

We now derive a suitable expression for these population-level quantities. Following Dwivedi and Mackey (2021), we define $\tau_{\mathbb{P}}(R) \triangleq \mathbb{P}(\mathbb{B}^c(0,R))$ and $\tau_{\mathbb{Q}}(R) \triangleq \mathbb{Q}(\mathbb{B}^c(0,R))$ where $\mathbb{B}^c(x,R) = \{y : \|x-y\|_2 \geq R\}$. The following result (proven using results on order statistics that we later develop in App. D.3) shows that we can upper-bound $\mathfrak{S}_{\mathbb{X}_m}$ and $\mathfrak{S}_{\mathbb{Y}_n}$ with high probability by a quantities that depend on $\mathbb{P}$ and m, and $\mathbb{Q}$ and n, respectively.

Lemma 3. *Define $\mathfrak{S}_{\mathbb{P},m,\delta} \triangleq \tau_{\mathbb{P}}^{-1}(\delta/m)$. With probability at least $1-\delta$, we have that $\mathfrak{S}_{\mathbb{X}_m} \leq \mathfrak{S}_{\mathbb{P},m,\delta}$. Similarly, with probability at least $1-\delta$, we have that $\mathfrak{S}_{\mathbb{Y}_n} \leq \mathfrak{S}_{\mathbb{Q},n,\delta} \triangleq \tau_{\mathbb{Q}}^{-1}(\delta/n)$.*

Proof. The random variable $\mathfrak{S}_{\mathbb{X}_m} \triangleq \max_{x\in\mathbb{X}_m} \|x\|_2$ is the m-th order statistic for m samples of $\mathbb{P}$ (Def. 1). Since the function $\tau_{\mathbb{P}}(R) \triangleq \mathbb{P}(\mathbb{B}^c(0,R)) = 1 - \mathbb{P}(\mathbb{B}(0,R))$ is one minus the cumulative function of the random variable $\|x\|_2, x \sim \mathbb{P}$, we obtain that $\tau_{\mathbb{P}}(\mathfrak{S}_{\mathbb{X}_m})$ is the first order statistic for m samples of the uniform distribution over $[0,1]$. Applying Lem. 5(iv) on $1 - \tau_{\mathbb{P}}(\mathfrak{S}_{\mathbb{X}_m})$, we obtain that

$$\Pr(\tau_{\mathbb{P}}(\mathfrak{S}_{\mathbb{X}_m}) < x) = \Pr(1 - \tau_{\mathbb{P}}(\mathfrak{S}_{\mathbb{X}_m}) > 1 - x) = \tbinom{m}{m-1} x^{m+1-m} = mx.$$

Hence, with probability at least $1-\delta$, $\mathfrak{S}_{\mathbb{X}_m} < \tau_{\mathbb{P}}^{-1}(\delta/m)$, and similarly $\mathfrak{S}_{\mathbb{Y}_n} < \tau_{\mathbb{P}}^{-1}(\delta/n)$. □

Now we can set

$$\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, \ell) = \|\mathbf{k}_{\mathrm{rt}}\|_\infty \max(\mathfrak{S}_{\mathbb{P},m,\delta/2}, \mathfrak{S}^\dagger_{\mathbf{k}_{\mathrm{rt}},\ell/2})^{\frac{d}{2}}\, \mathfrak{N}_{\mathbf{k}_{\mathrm{rt}}}(\ell, d, \delta, \mathfrak{S}_{\mathbb{P},m,\delta/2,\mathbf{k}_{\mathrm{rt}},\ell}) \text{ when } \mathbf{k}' = \mathbf{k}_{\mathrm{rt}} \quad \text{and} \tag{22}$$

$$\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, \ell) = \sqrt{\tfrac{8\|\mathbf{k}\|_\infty}{3} \log(\tfrac{12\log m}{\delta})\big[\log(\tfrac{8}{\delta}) + \mathcal{M}_{\mathbf{k}}(\mathcal{A}_{\mathbb{P},m,\delta}, (\ell/2)^{-1})\big]} \text{ when } \mathbf{k}' = \mathbf{k}, \tag{23}$$

where $\mathfrak{S}_{\mathbb{P},m,\delta/2,\mathbf{k}_{\mathrm{rt}},\ell} \triangleq \min\big(\mathfrak{S}_{\mathbb{P},m,\delta/2}, \ell^{1+\frac{1}{d}}\mathfrak{S}_{\mathbf{k}_{\mathrm{rt}},\ell} + \ell^{\frac{1}{d}}\frac{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}{L_{\mathbf{k}_{\mathrm{rt}}}}\big)$ and $\mathcal{A}_{\mathbb{P},m,\delta} = \big\{x : \|x\|_2 \leq \mathfrak{S}_{\mathbb{P},m,\delta/2}\big\}$. By Lem. 3, we have that with probability at least $1-\delta/2$, $\mathfrak{S}_{\mathbb{X}_m} \leq \mathfrak{S}_{\mathbb{P},m,\delta/2}$, and by construction (see App. B.4.1), $\mathbb{P}(\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{X}_m, \delta, \ell) \leq \mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, \ell)) \geq 1 - \delta/2$ as needed above.

B.5 Proof of Tab. 1

We begin by showing the bounds on $\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, \mathfrak{g})$ for the cases in which $\mathbf{k}' = \mathbf{k}_{\mathrm{rt}}$, and we include the case in which $\mathbf{k}'$ and $\mathbb{P}$ are sub-Gaussian for completeness. For $\mathfrak{g} \leq \log m$, equation (13) implies that the quantity $\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, \mathfrak{g})$ is of order $c_d\sqrt{\|\mathbf{k}'\|_\infty \log(m+n)\log(m)\log(\frac{m}{\delta})} \cdot \mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, 2^{\mathfrak{g}+1}\sqrt{m})$. Hence, we seek to upper-bound $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, 2^{\mathfrak{g}+1}\sqrt{m})$. In the table we replace the factor $\sqrt{\log(m+n)\log(m)}$ by $\log(\frac{m}{\delta})$ for simplcity.

First, note that upper bounds on $\tau_{\mathbb{P}}(x) \triangleq \mathbb{P}(\mathbb{B}^c(0,x))$, which are the usual notion of tail bounds on distributions, are equivalent to upper bounds on $\mathfrak{S}_{\mathbb{P},m,\delta} \triangleq \tau_{\mathbb{P}}^{-1}(\delta/m)$. Namely,

- **r-Compact:** $\tau_{\mathbb{P}}(x) = 0,\ \forall x > r \Leftrightarrow \tau_{\mathbb{P}}^{-1}(\delta/m) \leq r,\ \forall \delta, m,$
- **σ-Sub-Gaussian:** $\tau_{\mathbb{P}}(x) \leq 2e^{-\frac{x^2}{2\sigma^2}} \Leftrightarrow \tau_{\mathbb{P}}^{-1}(\delta/m) \leq \sqrt{2\sigma^2\log(\frac{2m}{\delta})},\ \forall \delta, m,$
- **σ, λ-subexponential:** $\tau_{\mathbb{P}}(x) \leq 2\max\{e^{-\frac{x^2}{2\sigma^2}}, e^{-\frac{x}{2\lambda}}\} \Leftrightarrow \tau_{\mathbb{P}}^{-1}(\delta/m) \leq \max\{\sqrt{2\sigma^2\log(\frac{2m}{\delta})}, 2\lambda\log(\frac{2m}{\delta})\},\ \forall \delta, m,$
- **ρ-Heavy-Tailed:** $\tau_{\mathbb{P}}(x) \leq c_d r^{-\rho} \Leftrightarrow \tau_{\mathbb{P}}^{-1}(\delta/m) \leq (\frac{c_d m}{\delta})^{1/\rho}, \forall \delta, m.$

Second, define $\widetilde{\mathfrak{S}}_{\mathbf{k}_{\mathrm{rt}},m} = \max\{\mathfrak{S}_{\mathbf{k}_{\mathrm{rt}},m}, \mathfrak{S}^{\dagger}_{\mathbf{k}_{\mathrm{rt}},m}\}$. Following Dwivedi and Mackey (2021), we formulate bounds on the decay of $\mathbf{k}'$ in terms of bounds on $\mathfrak{S}_{\mathbf{k}_{\mathrm{rt}}}$.

In Tab. 1, we consider four different growth conditions for the input point radii $\mathfrak{S}_{\mathbb{X}_m}$ arising from four forms of the target distribution and kernel tail decay (assuming same decay for both $\mathbb{P}$ and $\mathbf{k}'$): (1) **Compact**: $\mathfrak{S}_{\mathbb{P},m,\delta} \precsim_d r, \widetilde{\mathfrak{S}}_{\mathbf{k}_{\mathrm{rt}},m} \precsim_d 1$, (2) **Sub-Gaussian**: $\mathfrak{S}_{\mathbb{P},m,\delta} \precsim_d \sigma\sqrt{\log(m/\delta)}, \widetilde{\mathfrak{S}}_{\mathbf{k}_{\mathrm{rt}},m} \precsim_d \sqrt{\log m}$, (3) **subexponential**: $\mathfrak{S}_{\mathbb{P},m,\delta} \precsim_d \lambda \log(m/\delta), \widetilde{\mathfrak{S}}_{\mathbf{k}_{\mathrm{rt}},m} \precsim_d \log m$, and (4) **Heavy-Tailed**: $\mathfrak{S}_{\mathbb{P},m,\delta} \precsim_d (m/\delta)^{1/\rho}, \widetilde{\mathfrak{S}}_{\mathbf{k}_{\mathrm{rt}},m} \precsim_d m^{1/\rho}$). Here, the notation $\precsim_d$ means that factors depending on d and δ are hidden. The first condition holds when $\mathbb{P}$ is supported on a compact set like the unit cube $[0,1]^d$.

To get the bounds in the table, we observe that

$$\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, 2^{\mathfrak{g}+1}\sqrt{m}) = O_d\big(\max(\mathfrak{S}_{\mathbb{P},m,\delta/6}, \mathfrak{S}^{\dagger}_{\mathbf{k}_{\mathrm{rt}},2^{\mathfrak{g}}\sqrt{m}})^{\frac{d}{2}}\sqrt{\log(\tfrac{m}{\delta})\cdot\log(\max(\mathfrak{S}_{\mathbb{P},m,\delta/6}, \mathfrak{S}_{\mathbf{k}_{\mathrm{rt}},2^{\mathfrak{g}}\sqrt{m}}))}\big),$$

where O_d hides constants that depend on d. We now plug in the bounds on $\mathfrak{S}_{\mathbb{P},m,\delta}, \widetilde{\mathfrak{S}}_{\mathbf{k}_{\mathrm{rt}},m}$ for each of the four cases to obtain the following scaling for $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, 2^{\mathfrak{g}+1}\sqrt{m})$ (and simplifying expressions by using $\mathfrak{g} \le \frac{1}{2}\log_2 m$):

- r**-Compact:** $O_d(r^{\frac{d}{2}}\sqrt{\log(\frac{m}{\delta})\cdot\log r})$
- σ**-Sub-Gaussian:** $O_d(\sigma^{\frac{d}{2}}\log(\frac{m}{\delta})^{\frac{d+2}{4}}\sqrt{\log(\log(\frac{m}{\delta}))})$
- σ, λ**-subexponential:** $O_d(\lambda^{\frac{d}{2}}\log(\frac{m}{\delta})^{\frac{d+2}{2}}\sqrt{\log(\log(\frac{m}{\delta}))})$
- ρ**-Heavy-Tailed:** $O_d((\frac{m}{\delta})^{\frac{d}{2\rho}}\log(\frac{m}{\delta}))$

To show the bound for $\mathbf{k}' = \mathbf{k}$ with an analytic $\mathbf{k}$ and $\mathbb{P}$ with subexponential tails, we follow the pointers in App. B.4.1. Putting together (23) and Dwivedi and Mackey (2022, Thm. 2), we find that for this case $\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, \mathfrak{g})$ is of order $c_d\sqrt{\|\mathbf{k}\|_\infty \log(m+n)\log(m)\log(\frac{m}{\delta})}\cdot\mathfrak{M}_{\mathbf{k},\mathbf{k}}(\mathbb{P}, m, \delta, 2^{\mathfrak{g}+1}\sqrt{m})$, and doing algebra with (21), we conclude that $\mathfrak{M}_{\mathbb{P},m,\mathbf{k}}(\delta)$ scales linearly with the square-root of the log-covering number $\mathcal{M}_{\mathbf{k}}$. Dwivedi and Mackey (2022, Prop. 2(a)) states that the kernel covering number $\mathcal{M}_{\mathbf{k}}$ admits the scaling $(\log(1/\varepsilon))^{d+1}$ times the Euclidean covering number in $\mathbb{R}^d$ that admits a scaling of r^d for a Euclidean ball of radius r (see Wainwright (2019, Lem. 5.7)). Consequently, using the LOGGROWTH $\mathcal{M}_{\mathbf{k}}$ and subexponential $\mathbb{P}$ column with $\omega = d+1$ in Dwivedi and Mackey (2022, Tab. 2) shows that for this case $\mathfrak{M}_{\mathbf{k},\mathbf{k}}(\mathbb{P}, m, \delta, 2^{\mathfrak{g}+1}\sqrt{m}) = O(\log\frac{m}{\delta})^{\frac{3d+2}{2}}$, which in turn implies the corresponding scaling in Tab. 1 for $\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m, \delta, \mathfrak{g})$, where once again we have used the fact that $\mathfrak{g} \le \frac{1}{2}\log_2 m$ to simplify expressions.

C Proof of Prop. 1: Finite-sample exactness of CTT

We will prove the result under the weaker assumption that the point sequence $(X_1, \dots, X_m, Y_1, \dots, Y_n)$ is exchangeable. Under this assumption the statistics $(M_b)_{b=1}^{\mathcal{B}+1}$ are also exchangeable. Since R represents the position of $M_{\mathcal{B}+1}$ in a sorted ordering of $(M_b)_{b=1}^{\mathcal{B}+1}$ with ties broken uniformly at random and all positions in $\{1, \dots, \mathcal{B}+1\}$ are equally likely under exchangeability,

$$\begin{aligned}
&\Pr[R = b_\alpha] = 1/(\mathcal{B}+1), \\
&\Pr[R > b_\alpha] = (\mathcal{B}+1-b_\alpha)/(\mathcal{B}+1), \quad \text{and} \\
&\Pr[R < b_\alpha] = (b_\alpha - 1)/(\mathcal{B}+1).
\end{aligned}$$

Therefore, the CTT probability of rejection is

$$\begin{aligned}
\Pr[\Delta(\mathbb{X}_m, \mathbb{Y}_n) = 1] &= \Pr[R > b_\alpha] + \Pr[R = b_\alpha]p_\alpha \\
&= (\mathcal{B}+1-b_\alpha)/(\mathcal{B}+1) + (b_\alpha - (1-\alpha)(\mathcal{B}+1))/(\mathcal{B}+1) = \alpha.
\end{aligned}$$

D Proof of Thm. 1: Power of CTT

We first state a detailed version of Thm. 1.

D.1 Thm. 4: Power of CTT, detailed

In this section we will prove the following theorem, which is the detailed statement of the result in Thm. 1.

Theorem 4 (Power of CTT, detailed)**.** *Suppose Compress Then Test (Alg. 1) is run with level α, replication count $\mathcal{B} \geq \frac{1}{\alpha} - 1$, and coreset count $s_m \geq (32/9)\log(\frac{2e}{\gamma})$ for $\gamma \triangleq \frac{\alpha}{4e}(\frac{\beta}{4+2\beta})^{\frac{1}{\lfloor \alpha(\mathcal{B}+1) \rfloor}}$. Let $\tilde{\beta} = \beta/(1+\beta/2)$. Then* CTT *has power*

$$\Pr[\Delta_{\text{CTT}}(\mathbb{X}_m, \mathbb{Y}_n) = 1] \geq 1-\beta$$

whenever

$$\mathrm{MMD}_{\mathbf{k}}(\mathbb{P}, \mathbb{Q}) \geq 32\Big(\frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m/s_m, \tilde{\beta}/6), \mathfrak{g}}{2^{\mathfrak{g}}\sqrt{m}} + \frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, n/s_n, \tilde{\beta}/6, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} + c_{\tilde{\beta}/6}\big(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{\|\mathbf{k}\|_\infty}{n}}\big) + 2\big(\sqrt{\tfrac{9}{32}}+1\big)\big(2+c'\sqrt{\log(\gamma)}\big) \times$$
$$\big(\frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m/s_m, \tilde{\beta}/(20s_m), \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + \sqrt{\tfrac{s_n}{s_m}}\frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, n/s_n, \tilde{\beta}/(20s_n), \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} + c_{\tilde{\beta}/(20s)}\big(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{s_n\|\mathbf{k}\|_\infty}{s_m n}}\big)\big)\Big).$$

Remark 5. *Thm. 1 follows from this result as*

$$\sqrt{s_n/s_m}\cdot \mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, n/s_n, \tilde{\beta}/(20s_n), \mathfrak{g})/(2^{\mathfrak{g}}\sqrt{n}) = \mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, n/s_n, \tilde{\beta}/(20s_n), \mathfrak{g})/(2^{\mathfrak{g}}\sqrt{m})$$

and since $20s > 6$ and $m \leq n$.

We introduce some notation that we use throughout the proof. First, we let $(M_{(b)})_{b=1}^{\mathcal{B}}$ be the increasing ordering of the permuted MMD values. Recall that $b_\alpha = \lceil (1-\alpha)(\mathcal{B}+1) \rceil$, and that R is the position of $M_{\mathcal{B}+1}$ after sorting $(M_b)_{b=1}^{\mathcal{B}+1}$ increasingly with ties broken at random.

We note $R \leq b_\alpha$ is a necessary condition to accept the null hypothesis, and we show that it implies that $M_{\mathcal{B}+1} \triangleq \mathrm{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n) \leq M_{(b_\alpha)}$. To prove this, assume the contrapositive: if $M_{\mathcal{B}+1} > M_{(b_\alpha)}$, then forcibly the position R of $M_{\mathcal{B}+1}$ within an increasing ordering of $(M_b)_{b=1}^{\mathcal{B}}$ is greater than b_α. Hence,

$$\Pr[\Delta_{\text{CTT}}(\mathbb{X}_m, \mathbb{Y}_n) = 0] \leq \Pr[R \leq b_\alpha] \leq \Pr[\mathrm{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n) \leq M_{(b_\alpha)}].$$

Hence, to prove Thm. 1 (or Thm. 4) it suffices to show that $\Pr[\mathrm{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n) \leq M_{(b_\alpha)}] \leq \beta$.

D.2 Recasting the power lower bound into a high-probability threshold upper bound

We start with the following result that follows a structure similar to Schrab et al. (2021, Lem. 2).

Lemma 4 (Upper bound on acceptance probability from upper bound on threshold)**.** *Let $1 \geq \beta > 0$ arbitrary, and define $\tilde{\beta} = \frac{\beta}{1+\frac{\beta}{2}}$. Define the function*

$$Z(m,n,\beta) \triangleq \frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m/s_m, \tilde{\beta}/6, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + \frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, n/s_n, \tilde{\beta}/6, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} + c_{\tilde{\beta}/6}\big(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{\|\mathbf{k}\|_\infty}{n}}\big), \tag{24}$$

which is equal to the upper bound in (7) when we make the choice $\delta = \tilde{\beta}/6$. If $\Pr\big[\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) \geq Z(m,n,\beta) + M_{(b_\alpha)}\big] \geq \frac{1}{1+\frac{\beta}{2}}$ then $\Pr[\mathrm{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n) \leq M_{(b_\alpha)}] \leq \beta$.

Proof. Define the events $\mathfrak{A} := \{\mathrm{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n) \leq M_{(b_\alpha)}\}$, and $\mathfrak{B} := \{\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) \geq Z(m,n,\beta) + M_{(b_\alpha)}\}$.

By assumption, we have $\Pr[\mathfrak{B}] \geq 1 - \frac{\tilde{\beta}}{2}$, and we want to show $\Pr[\mathfrak{A}] \leq \beta$. Note that

$$\begin{aligned}
\Pr[\mathfrak{A}|\mathfrak{B}] &= \Pr\Big[\mathrm{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n) \leq M_{(b_\alpha)} \,|\, \mathfrak{B}\Big] \\
&\leq \Pr\Big[\mathrm{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n) \leq \mathrm{MMD}(\mathbb{P}, \mathbb{Q}) - Z(m,n,\beta) \,|\, \mathfrak{B}\Big] \\
&\leq \tfrac{1}{\Pr[\mathfrak{B}]}\Pr\Big[\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) - \mathrm{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n) \geq Z(m,n,\beta)\Big] \\
&\leq \tfrac{1}{1-\frac{\tilde{\beta}}{2}}\Pr\Big[\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) - \mathrm{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n) \geq Z(m,n,\beta)\Big].
\end{aligned} \tag{25}$$

Equation (7) in Lem. 1 shows that with probability at least $1 - \frac{\tilde{\beta}}{2}$,

$$|\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) - \mathrm{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)| \leq Z(m, n, \beta).$$

Thus, the right-hand side of (25) is upper-bounded by $\frac{\tilde{\beta}}{2} \cdot \frac{1}{1-\frac{\tilde{\beta}}{2}} = \frac{\beta}{2}$, where we used that

$$\tilde{\beta} = \frac{\beta}{1+\frac{\beta}{2}} \Leftrightarrow \frac{\tilde{\beta}}{2} = \frac{\frac{\beta}{2}}{1+\frac{\beta}{2}} \Leftrightarrow \frac{\beta}{2} = \frac{\frac{\tilde{\beta}}{2}}{1-\frac{\tilde{\beta}}{2}}.$$

We conclude the proof:

$$\begin{aligned}\Pr(\mathfrak{A}) &= \Pr(\mathfrak{A}|\mathfrak{B})\Pr(\mathfrak{B}) \\ &\quad + \Pr(\mathfrak{A}|\mathfrak{B}^c)\Pr(\mathfrak{B}^c) \leq \tfrac{\beta}{2} \cdot 1 + 1 \cdot \tfrac{\beta}{2} = \beta.\end{aligned}$$

□

D.3 High-probability bound on the threshold

Given Lem. 4, the remainder of the proof of Thm. 1 is devoted to checking that $\Pr\big[\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) \geq Z(m, n, \beta) + M_{(b_\alpha)}\big] \geq 1/(1 + \frac{\beta}{2})$ holds, which involves getting a high-probability upper-bound on $M_{(b_\alpha)}$. Schrab et al. (2021) use an approach based on the Dvoretzky-Kiefer-Wolfowitz theorem, which forces them to use a number of permutations $\mathcal{B}$ larger than a threshold which is larger than the values used in practice. We employ more precise techniques based on order statistics (pioneered in this setting by Learned-Miller and DeStefano (2008)) which give tight results for any $\mathcal{B}$ as long as $\mathcal{B} \geq \alpha^{-1} - 1$. We focus on the case of permutations instead of wild bootstrap, but the arguments could be adapted for the wild boostrap case, which is in fact simpler.

Definition 1 (k-th order statistic)**.** *Given n i.i.d. variables $(Y_k)_{k=1}^n$, and define the variables $(Y_{(k)})_{k=1}^n$ as the result of sorting $(Y_k)_{k=1}^n$ in increasing order. For any $1 \leq k \leq n$, the variable $Y_{(k)}$ is known as the k-th order statistic.*

It is well known that the k-th order statistic for n samples of the uniform distribution on $[0, 1]$ is distributed according to the beta distribution $\mathrm{Beta}(k, n + 1 - k)$ (Gentle, 2009, p.63). The CDF of the distribution $\mathrm{Beta}(k, n + 1 - k)$ is equal to the regularized incomplete beta function $I_x(k, n + 1 - k)$, which is defined below.

Given positive $a, b \in \mathbb{R}$ and $x \in [0, 1]$, the regularized incomplete beta function is defined as $I_x(a, b) = \frac{B(x;a,b)}{B(a,b)}$, where $B(x; a, b) = \int_0^x t^{a-1}(1-t)^{b-1}\, dt$ is the incomplete beta function and $B(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}\, dt$ is the beta function.

Lemma 5 (Properties of the regularized incomplete beta function)**.** *The following statements regarding the regularized incomplete beta function and order statistics hold:*

(i) *For any integers $m \leq n$ and $x \in [0, 1)$, we have that*

$$I_x(m, n + 1 - m) = \textstyle\sum_{j=m}^n \binom{n}{j} x^j (1 - x)^{n-j}. \tag{26}$$

(ii) *For any $m \leq n$ and $x \in [0, 1)$, we have that $\frac{\partial^k}{\partial x^k} I_x(m, n + 1 - m)|_{x=0} = 0$ for any $0 \leq k < m$, and that*

$$\frac{\partial^m}{\partial x^m} I_x(m, n + 1 - m)|_{x=0} = \binom{n}{m} m! = \frac{n!}{(n-m)!}, \tag{27}$$

$$\frac{\partial^{m+1}}{\partial x^{m+1}} I_x(m, n + 1 - m)|_{x=0} = -\mathbf{1}_{m<n} \frac{n!m}{(n-m-1)!} \tag{28}$$

(iii) *For any $x \in [0, 1)$, there exists $z \in [0, x)$ such that $I_x(m, n + 1 - m) = \binom{n}{m} x^m - m\binom{n}{m+1} z^{m+1}$.*

(iv) *Let $Y_{(m)}$ be the m-th order statistic (Def. 1) for n samples of the uniform distribution on $[0, 1]$. For any $x \in [0, 1]$, we have that*

$$\binom{n}{m-1} x^{n+1-m} - (n + 1 - m)\binom{n}{m-2} x^{n+2-m} \mathbf{1}_{m>1} \leq \Pr[Y_{(m)} > 1 - x] \leq \binom{n}{m-1} x^{n+1-m}.$$

Proof. We prove each part separately.

(i) This part follows directly from NIS, Eq. 8.17.5.

(ii) The statement $\frac{\partial^k}{\partial x^k} I_x(m, n+1-m)|_{x=0} = 0$ for any $0 \leq k < m$ holds because by (26), $I_x(m, n+1-m)$ can be expressed as a polynomial in x where all the terms are of power at least m.

To obtain $\frac{\partial^m}{\partial x^m} I_x(m, n+1-m)|_{x=0}$, we multiply by $m!$ the coefficient of $I_x(m, n+1-m)$ for the term of degree m, which is the term of degree m of the polynomial $\binom{n}{m} x^m (1-x)^{n-m}$.

To obtain $\frac{\partial^{m+1}}{\partial x^{m+1}} I_x(m, n+1-m)|_{x=0}$, we multiply by $(m+1)!$ the coefficient of $I_x(m, n+1-m)$ for the term of degree $m+1$, which is the term of degree $m+1$ of the polynomial $\binom{n}{m} x^m (1-x)^{n-m}$ plus the term of degree $m+1$ of the polynomial $\binom{n}{m+1} x^{m+1} (1-x)^{n-m-1}$ (if the latter term exists). Thus, when $m+1 \leq n$, $(m+1)!$ times the coefficient of $I_x(m, n+1-m)$ for the term of degree $m+1$ reads:

$$\begin{aligned}
&- \tbinom{n}{m}(n-m)(m+1)! + \tbinom{n}{m+1}(m+1)! \\
&= -\tfrac{n!}{m!(n-m)!}(n-m)(m+1)! + \tfrac{n!}{(m+1)!(n-m-1)!}(m+1)! \\
&= -\tfrac{n!(m+1)}{(n-m-1)!} + \tfrac{n!}{(n-m-1)!} \\
&= -\tfrac{n!m}{(n-m-1)!}.
\end{aligned}$$

When $m = n$, we obtain 0 instead.

(iii) By the residual form of Taylor's theorem, we have that for any $y \in [0, 1)$,

$$I_y(m, n+1-m) = \tfrac{1}{m!} \tfrac{\partial^m}{\partial x^m} I_x(m, n+1-m)|_{x=0} y^m + \tfrac{1}{(m+1)!} \tfrac{\partial^{m+1}}{\partial x^{m+1}} I_x(m, n+1-m)|_{x=0} z^{m+1},$$

where $z \in [0, y]$. Substituting the expressions from (27) and (28) into this equation, we obtain that

$$I_y(m, n+1-m) = \tfrac{n!}{(n-m)!m!} y^m - \tfrac{n!m}{(m+1)!(n-m-1)!} z^{m+1} \mathbf{1}_{m<n} = \tbinom{n}{m} y^m - m\tbinom{n}{m+1} z^{m+1} \mathbf{1}_{m<n}.$$

(iv) By NIS, Eq. 8.17.4, for any a, b non-negative and $x \in [0, 1]$, we have that $I_x(a, b) = 1 - I_{1-x}(a, b)$. The variable $Y_{(m)}$ is distributed according to $\mathrm{Beta}(m, n+1-m)$, which means that

$$\begin{aligned}
\Pr[Y_{(m)} > 1-x] &= 1 - \Pr[Y_{(m)} \leq 1-x] = 1 - I_{1-x}(m, n+1-m) = I_x(n+1-m, m) \\
&= I_x(n+1-m, n+1-(n+1-m)) \\
&= \tbinom{n}{n+1-m} x^{n+1-m} - (n+1-m)\tbinom{n}{n+2-m} z^{n+2-m} \mathbf{1}_{n+1-m<n} \\
&= \tbinom{n}{m-1} x^{n+1-m} - (n+1-m)\tbinom{n}{m-2} z^{n+2-m} \mathbf{1}_{m>1}
\end{aligned}$$

In the second-to-last equality we plugged the result from Lem. 5(iii), replacing m by $n+1-m$. Since $z \in [0, x]$, the result follows.

□

Let F be the CDF of the random variable $M_\sigma \triangleq \mathrm{MMD}(\hat{\mathbb{X}}^\sigma_m, \hat{\mathbb{Y}}^\sigma_n)$,i.e. $F(x) = P(M_\sigma \leq x)$. We define the random map $\mathfrak{F}$ as

$$\mathfrak{F}(x) = \begin{cases} F(x) & \text{if } F \text{ continuous at } x \\ \mathrm{Unif}(\lim_{y \to x^-} F(y), F(x)) & \text{otherwise} \end{cases}$$

Note that by definition, for all x we have that $\Pr(M_\sigma < x) \leq \mathfrak{F}(x) \leq \Pr(M_\sigma \leq x)$. Also, by construction $\mathfrak{F}(M_\sigma)$ is distributed uniformly over $[0, 1]$.

Lemma 6 (High probability bound on the threshold from quantile of the CDF F)**.** *For an arbitrary $\alpha' \in (0, 1)$, we define the random variable*

$$q_{1-\alpha'}(\mathbb{X}_m, \mathbb{Y}_n) \triangleq \inf \left\{ x \in \mathbb{R} : 1 - \alpha' \leq F(x) \right\}.$$

Given $\mathbb{X}_m, \mathbb{Y}_n$, we have that with probability at least $1 - \frac{\delta}{2}$,

$$M_{(b_\alpha)} \leq q_{1-\alpha^*}(\mathbb{X}_m, \mathbb{Y}_n),$$

where $\alpha^* = \left(\frac{\delta}{2\binom{\mathcal{B}}{\lfloor \alpha(\mathcal{B}+1) \rfloor}} \right)^{1/\lfloor \alpha(\mathcal{B}+1) \rfloor}$.

Proof. Note that $M_{(b_\alpha)}$ is the b_α-th order statistic for the $\mathcal{B}$ samples $(M_b)_{b=1}^{\mathcal{B}}$. Since the random map $\mathfrak{F}$ is increasing, this implies that $\mathfrak{F}(M_{(b_\alpha)})$ is the b_α-th order statistic for the $\mathcal{B}$ samples $(\mathfrak{F}(M_b))_{b=1}^{\mathcal{B}}$. As stated above, $(\mathfrak{F}(M_b))_{b=1}^{\mathcal{B}}$ are uniform i.i.d. variables over $[0,1]$, which means that $\mathfrak{F}(M_{(b_\alpha)})$ is the b_α-th order statistic for $\mathcal{B}$ samples of the uniform distribution over $[0,1]$. Applying Lem. 5(iv) with $n=\mathcal{B}$, $m=b_\alpha=\lceil(1-\alpha)(\mathcal{B}+1)\rceil$, we obtain that for any $x\in[0,1)$,

$$\Pr[\mathfrak{F}(M_{(b_\alpha)})>1-x] \leq \binom{\mathcal{B}}{b_\alpha-1}x^{\mathcal{B}+1-b_\alpha} = \binom{\mathcal{B}}{\mathcal{B}+1-b_\alpha}x^{\mathcal{B}+1-b_\alpha} = \binom{\mathcal{B}}{\lfloor\alpha(\mathcal{B}+1)\rfloor}x^{\lfloor\alpha(\mathcal{B}+1)\rfloor}.$$

Since

$$\binom{\mathcal{B}}{\lfloor\alpha(\mathcal{B}+1)\rfloor}x^{\lfloor\alpha(\mathcal{B}+1)\rfloor}=\delta/2 \quad \Leftrightarrow \quad x=\Big(\frac{\delta}{2\binom{\mathcal{B}}{\lfloor\alpha(\mathcal{B}+1)\rfloor}}\Big)^{1/\lfloor\alpha(\mathcal{B}+1)\rfloor},$$

we obtain that with probability at least $1-\frac{\delta}{2}$,

$$\mathfrak{F}(M_{(b_\alpha)}) < 1-\Big(\frac{\delta}{2\binom{\mathcal{B}}{\lfloor\alpha(\mathcal{B}+1)\rfloor}}\Big)^{1/\lfloor\alpha(\mathcal{B}+1)\rfloor}.$$

For any $\epsilon>0$, we have that given $\mathbb{X}_m,\mathbb{Y}_n$, $F(x-\epsilon)=\Pr(M_\sigma\leq x-\epsilon)\leq\Pr(M_\sigma<x)\leq\mathfrak{F}(x)$. Hence, with probability at least $1-\frac{\delta}{2}$, $F(M_{(b_\alpha)}-\epsilon)\leq 1-\Big(\frac{\delta}{2\binom{\mathcal{B}}{\lfloor\alpha(\mathcal{B}+1)\rfloor}}\Big)^{1/\lfloor\alpha(\mathcal{B}+1)\rfloor}$.

Hence, if we define $\alpha^*=\Big(\frac{\delta}{2\binom{\mathcal{B}}{\lfloor\alpha(\mathcal{B}+1)\rfloor}}\Big)^{1/\lfloor\alpha(\mathcal{B}+1)\rfloor}$, we obtain that

$$M_{(b_\alpha)}-\epsilon \leq \inf\big\{x\in\mathbb{R}:1-\alpha^*\leq F(x)\big\} \triangleq q_{1-\alpha^*}(\mathbb{X}_m,\mathbb{Y}_n)$$

Since $\epsilon>0$ is arbitrary, we conclude that $M_{(b_\alpha)}\leq q_{1-\alpha^*}(\mathbb{X}_m,\mathbb{Y}_n)$ with probability at least $1-\frac{\delta}{2}$. $\square$

Recall that $(\hat{\mathbb{X}}_m^{(i)})_{i=1}^{s_m}$, $(\hat{\mathbb{Y}}_n^{(i)})_{i=1}^{s_n}$ are the outputs of KT-Compress on inputs $(\mathbb{X}_m^{(i)})_{i=1}^{s_m}$, $(\mathbb{Y}_n^{(i)})_{i=1}^{s_n}$. For $i=1,\dots,s_m$, $j=1,\dots,s_n$, denote

$$\begin{gathered}\hat{\mathbb{P}}_m=\frac{1}{|\hat{\mathbb{X}}_m|}\sum_{x\in\hat{\mathbb{X}}_m}\delta_x,\quad \hat{\mathbb{Q}}_n=\frac{1}{|\hat{\mathbb{Y}}_n|}\sum_{y\in\hat{\mathbb{Y}}_n}\delta_y,\\ \hat{\mathbb{P}}_m^{(i)}=\frac{1}{|\hat{\mathbb{X}}_m^{(i)}|}\sum_{x\in\hat{\mathbb{X}}_m^{(i)}}\delta_x,\quad \hat{\mathbb{Q}}_n^{(j)}=\frac{1}{|\hat{\mathbb{Y}}_n^{(j)}|}\sum_{y\in\hat{\mathbb{Y}}_n^{(j)}}\delta_y,\\ \hat{\mathbb{S}}_{m+n}^{(i)}=\hat{\mathbb{P}}_m^{(i)},\quad \hat{\mathbb{S}}_{m+n}^{(s_m+j)}=\hat{\mathbb{Q}}_n^{(j)}\end{gathered}$$

We can write

$$\begin{aligned}\mathrm{MMD}^2(\hat{\mathbb{X}}_m,\hat{\mathbb{Y}}_n)&=\langle(\hat{\mathbb{P}}_m-\hat{\mathbb{Q}}_n)\mathbf{k},(\hat{\mathbb{P}}_m-\hat{\mathbb{Q}}_n)\mathbf{k}\rangle_{\mathbf{k}}\\ &=\frac{1}{s_m^2s_n^2}\sum_{i=1}^{s_m}\sum_{i'=1}^{s_m}\sum_{j=1}^{s_n}\sum_{j'=1}^{s_n}\langle(\hat{\mathbb{P}}_m^i-\hat{\mathbb{Q}}_n^j)\mathbf{k},(\hat{\mathbb{P}}_m^{i'}-\hat{\mathbb{Q}}_n^{j'})\mathbf{k}\rangle_{\mathbf{k}}\\ &=\frac{1}{s_m^2s_n^2}\sum_{i=1}^{s_m}\sum_{i'=1}^{s_m}\sum_{j=1}^{s_n}\sum_{j'=1}^{s_n}\langle(\hat{\mathbb{S}}_{m+n}^i-\hat{\mathbb{S}}_{m+n}^{s_m+j})\mathbf{k},(\hat{\mathbb{S}}_{m+n}^{i'}-\hat{\mathbb{S}}_{m+n}^{s_m+j'})\mathbf{k}\rangle_{\mathbf{k}}\\ &=\frac{1}{s_m^2s_n^2}\big(\sum_{i\neq i'\in\{1,\dots,s_m\}}\sum_{j\neq j'\in\{1,\dots,s_n\}}\langle(\hat{\mathbb{S}}_{m+n}^i-\hat{\mathbb{S}}_{m+n}^{s_m+j})\mathbf{k},(\hat{\mathbb{S}}_{m+n}^{i'}-\hat{\mathbb{S}}_{m+n}^{s_m+j'})\mathbf{k}\rangle_{\mathbf{k}} \qquad (29)\\ &+\sum_{i=1}^{s_m}\sum_{j\neq j'\in\{1,\dots,s_n\}}\langle(\hat{\mathbb{S}}_{m+n}^i-\hat{\mathbb{S}}_{m+n}^{s_m+j})\mathbf{k},(\hat{\mathbb{S}}_{m+n}^{i}-\hat{\mathbb{S}}_{m+n}^{s_m+j'})\mathbf{k}\rangle_{\mathbf{k}}\\ &+\sum_{j=1}^{s_n}\sum_{i\neq i'\in\{1,\dots,s_m\}}\langle(\hat{\mathbb{S}}_{m+n}^i-\hat{\mathbb{S}}_{m+n}^{s_m+j})\mathbf{k},(\hat{\mathbb{S}}_{m+n}^{i'}-\hat{\mathbb{S}}_{m+n}^{s_m+j})\mathbf{k}\rangle_{\mathbf{k}}\\ &+\sum_{i=1}^{s_m}\sum_{j=1}^{s_n}\langle(\hat{\mathbb{S}}_{m+n}^i-\hat{\mathbb{S}}_{m+n}^{s_m+j})\mathbf{k},(\hat{\mathbb{S}}_{m+n}^{i}-\hat{\mathbb{S}}_{m+n}^{s_m+j})\mathbf{k}\rangle_{\mathbf{k}}\big)\end{aligned}$$

By assuming $m\leq n$, let $L:=\{l_1,...,l_m\}$ be an m-tuple uniformly drawn without replacement from $\{1,\dots,n\}$. Then, we can write (29) as

$$\begin{aligned}\mathrm{MMD}^2(\hat{\mathbb{X}}_m,\hat{\mathbb{Y}}_n)&=\frac{s_n-1}{s_ns_m^2}\mathbb{E}_L[\sum_{i\neq i'\in\{1,\dots,s_m\}}\langle(\hat{\mathbb{S}}_{m+n}^i-\hat{\mathbb{S}}_{m+n}^{s_m+l_i})\mathbf{k},(\hat{\mathbb{S}}_{m+n}^{i'}-\hat{\mathbb{S}}_{m+n}^{s_m+l_{i'}})\mathbf{k}\rangle_{\mathbf{k}}] \qquad (30)\\ &+\frac{1}{s_m^2s_n^2}\big(\sum_{i=1}^{s_m}\sum_{j\neq j'\in\{1,\dots,s_n\}}\langle(\hat{\mathbb{S}}_{m+n}^i-\hat{\mathbb{S}}_{m+n}^{s_m+j})\mathbf{k},(\hat{\mathbb{S}}_{m+n}^{i}-\hat{\mathbb{S}}_{m+n}^{s_m+j'})\mathbf{k}\rangle_{\mathbf{k}}\\ &+\sum_{j=1}^{s_n}\sum_{i\neq i'\in\{1,\dots,s_m\}}\langle(\hat{\mathbb{S}}_{m+n}^i-\hat{\mathbb{S}}_{m+n}^{s_m+j})\mathbf{k},(\hat{\mathbb{S}}_{m+n}^{i'}-\hat{\mathbb{S}}_{m+n}^{s_m+j})\mathbf{k}\rangle_{\mathbf{k}}\\ &+\sum_{i=1}^{s_m}\sum_{j=1}^{s_n}\langle(\hat{\mathbb{S}}_{m+n}^i-\hat{\mathbb{S}}_{m+n}^{s_m+j})\mathbf{k},(\hat{\mathbb{S}}_{m+n}^{i}-\hat{\mathbb{S}}_{m+n}^{s_m+j})\mathbf{k}\rangle_{\mathbf{k}}\big).\end{aligned}$$

This holds because for any $i \neq i' \in \{1, \dots, s_m\}$,

$$\mathbb{E}_L[\langle(\hat{\mathbb{P}}_m^{(i)} - \hat{\mathbb{Q}}_n^{(s_m+l_i)})\mathbf{k}, (\hat{\mathbb{P}}_m^{(i')} - \hat{\mathbb{Q}}_n^{(s_m+l_{i'})})\mathbf{k}\rangle_{\mathbf{k}}] = \frac{1}{s_n(s_n-1)} \sum_{j\neq j' \in\{1,\dots,s_n\}} \langle(\hat{\mathbb{S}}_{m+n}^{(i)} - \hat{\mathbb{S}}_{m+n}^{(s_m+j)})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{(i')} - \hat{\mathbb{S}}_{m+n}^{(s_m+j')})\mathbf{k}\rangle_{\mathbf{k}}$$

Recall also that $\mathbb{U}_{m+n} = (U_i)_{i=1}^{m+n}$, with $U_i = X_i$ for $i = 1, \dots, m$ and $U_{m+j} = Y_j$ for $j = 1, \dots, n$. Equivalently, we can write that $\mathbb{U}_{m+n} = (\mathbb{U}_{m+n}^{(i)})_{i=1}^{s}$, with $\mathbb{U}_{m+n}^{(i)} = \mathbb{X}_m^{(i)}$ and $\mathbb{U}_{m+n}^{(s_m+j)} = \mathbb{Y}_n^{(j)}$ for $i = 1, \dots, s_m$, $j = 1, \dots, s_n$. Analogously, we define $\hat{\mathbb{U}}_{m+n} = (\hat{\mathbb{U}}_{m+n}^{(i)})_{i=1}^{s}$, with $\hat{\mathbb{U}}_{m+n}^{(i)} = \hat{\mathbb{X}}_m^{(i)}$ and $\hat{\mathbb{U}}_{m+n}^{(s_m+j)} = \hat{\mathbb{Y}}_n^{(j)}$ for $i = 1, \dots, s_m$, $j = 1, \dots, s_n$.

Given a permutation $\sigma : \{1, \dots, s\} \to \{1, \dots, s\}$, we write $\mathbb{U}_{m+n}^{\sigma} = (\mathbb{U}_{m+n}^{(\sigma(i))})_{i=1}^{s}$, $\hat{\mathbb{U}}_{m+n}^{\sigma} = (\hat{\mathbb{U}}_{m+n}^{(\sigma(i))})_{i=1}^{s}$, and $\mathbb{X}_m^{\sigma} = (\mathbb{U}_{m+n}^{(\sigma(i))})_{i=1}^{s_m}$, $\hat{\mathbb{X}}_m^{\sigma} = (\hat{\mathbb{U}}_{m+n}^{(\sigma(i))})_{i=1}^{s_m}$, $\mathbb{Y}_n^{\sigma} = (\mathbb{U}_{m+n}^{(\sigma(i))})_{i=s_m+1}^{s}$, $\hat{\mathbb{Y}}_n^{\sigma} = (\hat{\mathbb{U}}_{m+n}^{(\sigma(i))})_{i=s_m+1}^{s}$. Analogously to (30), we can write

$$\begin{aligned}
\mathrm{MMD}^2(\hat{\mathbb{X}}_m^{\sigma}, \hat{\mathbb{Y}}_n^{\sigma}) &= \tfrac{1}{s_m^2 s_n^2} \sum_{i=1}^{s_m} \sum_{i'=1}^{s_m} \sum_{j=1}^{s_n} \sum_{j'=1}^{s_n} \langle(\hat{\mathbb{S}}_{m+n}^{(\sigma(i))} - \hat{\mathbb{S}}_{m+n}^{(\sigma(s_m+j))})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{(\sigma(i'))} - \hat{\mathbb{S}}_{m+n}^{(\sigma(s_m+j'))})\mathbf{k}\rangle_{\mathbf{k}} \\
&= \tfrac{(s_m-1)(s_n-1)}{s_m s_n} \mathbb{E}_L[\mathcal{M}^{\sigma,L}] \\
&+ \tfrac{1}{s_m^2 s_n^2} \big(\sum_{i=1}^{s_m} \sum_{j\neq j'\in\{1,\dots,s_n\}} \langle(\hat{\mathbb{S}}_{m+n}^{i} - \hat{\mathbb{S}}_{m+n}^{\sigma(s_m+j)})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{i} - \hat{\mathbb{S}}_{m+n}^{\sigma(s_m+j')})\mathbf{k}\rangle_{\mathbf{k}} \\
&+ \sum_{j=1}^{s_n} \sum_{i\neq i'\in\{1,\dots,s_m\}} \langle(\hat{\mathbb{S}}_{m+n}^{\sigma(i)} - \hat{\mathbb{S}}_{m+n}^{\sigma(s_m+j)})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\sigma(i')} - \hat{\mathbb{S}}_{m+n}^{\sigma(s_m+j)})\mathbf{k}\rangle_{\mathbf{k}} \\
&+ \sum_{i=1}^{s_m} \sum_{j=1}^{s_n} \langle(\hat{\mathbb{S}}_{m+n}^{\sigma(i)} - \hat{\mathbb{S}}_{m+n}^{\sigma(s_m+j)})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\sigma(i)} - \hat{\mathbb{S}}_{m+n}^{\sigma(s_m+j)})\mathbf{k}\rangle_{\mathbf{k}} \big),
\end{aligned} \tag{31}$$

where we use the short-hand

$$\mathcal{M}^{\sigma,L} \triangleq \tfrac{1}{s_m(s_m-1)} \sum_{i\neq i'\in\{1,\dots,s_m\}} \langle(\hat{\mathbb{S}}_{m+n}^{(\sigma(i))} - \hat{\mathbb{S}}_{m+n}^{(\sigma(s_m+l_i))})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{(\sigma(i'))} - \hat{\mathbb{S}}_{m+n}^{(\sigma(s_m+l_{i'}))})\mathbf{k}\rangle_{\mathbf{k}}. \tag{32}$$

The following proposition, whose proof is deferred to App. D.3.1, provides a tail upper-bound on the random variable $\mathrm{MMD}^2(\hat{\mathbb{X}}_m^{\sigma}, \hat{\mathbb{Y}}_n^{\sigma})$.

Proposition 3 (Tail bound on $\mathrm{MMD}^2(\hat{\mathbb{X}}_m^{\sigma}, \hat{\mathbb{Y}}_n^{\sigma})$ conditioned on $\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n$)**.** *Let σ be a uniformly random permutation over $\{1, \dots, s\}$. Let $\delta, \delta'' \in (0, 1)$, and $\delta \in (0, e^{-1})$. There is an event $\mathcal{A}$ of probability at least $1 - \delta$ concerning the draw of $\hat{\mathbb{X}}_m$, $\hat{\mathbb{Y}}_n$, such that conditioned on $\mathcal{A}$, with probability at least $1 - \delta' - \delta''$ on the draw of σ,*

$$\mathrm{MMD}^2(\hat{\mathbb{X}}_m^{\sigma}, \hat{\mathbb{Y}}_n^{\sigma}) \leq \tfrac{2}{s_m}(\log(\tfrac{2}{\delta''}) + 1)\,\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) + \tfrac{c' \log(1/\delta')+2}{s_m}\big(\,\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) W(m, n, \delta/(5s)) + W(m, n, \delta/(5s))^2\big)$$

where c' is a universal constant and

$$W(m, n, \delta) = 2\big(\tfrac{\sqrt{s_m}\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P},m/s_m,\delta,\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + \tfrac{\sqrt{s_n}\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n/s_n,\delta,\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} + c_\delta\big(\sqrt{\tfrac{s_m\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{s_n\|\mathbf{k}\|_\infty}{n}}\big)\big). \tag{33}$$

Here, the KT-COMPRESS *error inflation factors $\mathbf{R}_{\mathbf{k},\mathbf{k}'}$ and the factor c_δ are defined as in Lem. 1.*

Combining Lem. 6 with the tail bound from Prop. 3 yields the following high-probability upper bound on the threshold $M_{(b_\alpha)}$.

Corollary 1 (High probability bound on the threshold)**.** *Assume that $b_\alpha \triangleq \lceil(\mathcal{B}+1)(1-\alpha)\rceil \leq \mathcal{B}$, and that* KT-COMPRESS *calls are run with value $\delta^*/(5s_m)$ and $\delta^*/(5s_n)$ respectively. Then, with probability at least $1 - \frac{\delta}{2}$,*

$$\begin{aligned}
M_{(b_\alpha)} &\leq \hat{Z}(m, n, \alpha, \delta) \\
&\triangleq \sqrt{\tfrac{2}{s_m}(\log(\tfrac{2}{\delta^*}) + 1)}\,\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) + \sqrt{\tfrac{1}{s_m}(2 + c' \log(1/\delta^*))}\big(\sqrt{\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) W(m, n, \delta/(20s))} \\
&+ W(m, n, \delta/(20s))\big).
\end{aligned} \tag{34}$$

where

$$\delta^* \triangleq (\tfrac{\delta}{2})^{1/k_\alpha} \tfrac{\alpha}{4e}, \quad k_\alpha \triangleq \lfloor \alpha(\mathcal{B}+1) \rfloor,$$

Proof. By Lem. 6, with probability at least $1 - \frac{\delta}{4}$, we have that

$$M_{(b_\alpha)} \leq q_{1-\alpha^*}(\mathbb{X}_m, \mathbb{Y}_n),$$

where $\alpha^* = \big(\frac{\delta}{4\binom{\mathcal{B}}{k_\alpha}}\big)^{1/k_\alpha}$. Conditioned on the event $\mathcal{A}$ with $\delta \leftarrow \delta/4$ (i.e. $\Pr(\mathcal{A}) \geq 1-\delta/4$), and setting $\delta' = \delta'' = \alpha^*/2$ we have that with probability at least $1-\alpha^*$ on the choice of σ,

$$\begin{aligned}\mathrm{MMD}(\hat{\mathbb{X}}_m^\sigma, \hat{\mathbb{Y}}_n^\sigma) &\leq \big(\tfrac{2}{s_m}(\log(\tfrac{2\cdot 2}{\alpha^*})+1)\,\mathrm{MMD}^2(\mathbb{P},\mathbb{Q}) \\ &\quad + \tfrac{1}{s_m}\big(\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(20s)) + W(m,n,\delta/(20s))^2\big)(2+c'\log(2/\alpha^*))\big)^{1/2} \\ &\leq \sqrt{\tfrac{2}{s_m}(\log(\tfrac{2\cdot 2}{\alpha^*})+1)}\,\mathrm{MMD}(\mathbb{P},\mathbb{Q}) \\ &\quad + \sqrt{\tfrac{1}{s_m}(2+c'\log(2/\alpha^*))}\big(\sqrt{\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(20s))} + W(m,n,\delta/(20s))\big).\end{aligned}$$

An application of Prop. 3 yields, conditioned on the event $\mathcal{A}$,

$$\begin{aligned}q_{1-\alpha^*}(\mathbb{X}_m,\mathbb{Y}_n) &\leq \sqrt{\tfrac{2}{s_m}(\log(\tfrac{2\cdot 2}{\alpha^*})+1)}\,\mathrm{MMD}(\mathbb{P},\mathbb{Q}) \\ &\quad + \sqrt{\tfrac{1}{s_m}(2+c'\log(2/\alpha^*))}\big(\sqrt{\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(20s))} + W(m,n,\delta/(20s))\big).\end{aligned} \tag{35}$$

Since the probability of $\mathcal{A}$ is at least $1-\delta/4$, we obtain that with probability at least $1-\delta/2$,

$$\begin{aligned}M_{(b_\alpha)} &\leq \sqrt{\tfrac{2}{s_m}(\log(\tfrac{2\cdot 2}{\alpha^*})+1)}\,\mathrm{MMD}(\mathbb{P},\mathbb{Q}) \\ &\quad + \sqrt{\tfrac{1}{s_m}(2+c'\log(2/\alpha^*))}\big(\sqrt{\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(20s))} + W(m,n,\delta/(20s))\big).\end{aligned}$$

Using the fact $(\frac{n}{k})^k \leq \binom{n}{k} \leq (\frac{en}{k})^k$, we obtain that $\log(\frac{\mathcal{B}}{k_\alpha}) \leq \log(\binom{\mathcal{B}}{k_\alpha}^{\frac{1}{k_\alpha}}) \leq \log(\frac{\mathcal{B}}{k_\alpha})+1$. Furthermore, $\log(\frac{\mathcal{B}}{k_\alpha}) \leq \log(\frac{2\mathcal{B}}{\alpha(\mathcal{B}+1)}) \leq \log(\frac{2}{\alpha})$ since $k_\alpha \geq 1$. Consequently, $\log(\binom{\mathcal{B}}{k_\alpha}^{\frac{1}{k_\alpha}}) \leq \log\frac{2e}{\alpha}$, so that the separation rate in Thm. 1 is independent of $\mathcal{B}$ (up to the condition $\mathcal{B}+1 \geq \alpha^{-1}$). Equivalently, $\binom{\mathcal{B}}{k_\alpha}^{\frac{1}{k_\alpha}} \leq \frac{2e}{\alpha}$, and $\alpha^* \geq (\frac{\delta}{4})^{1/k_\alpha}\frac{\alpha}{2e}$, and $\delta^* \geq (\frac{\delta}{4})^{1/k_\alpha}\frac{\alpha}{4e}$. Plugging this into (35) concludes the proof. □

D.3.1 Proof of Prop. 3: Tail bound on $\mathrm{MMD}^2(\hat{\mathbb{X}}_m^\sigma, \hat{\mathbb{Y}}_n^\sigma)$ conditioned on $\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n$

We will first show the following lemma, which gives a high-probability upper bound on the expectation $\mathbb{E}_L[\mathcal{M}^{\sigma,L}]$, where $\mathcal{M}^{\sigma,L}$ is defined in equation (32), and on the rest of the terms that appear in the right-hand side of (31).

Lemma 7 (Bounding the right-hand side of (31))**.** *Let σ be a uniformly random permutation over $\{1,\ldots,s\}$, and $L := \{l_1,...,l_m\}$ a uniformly random m-tuple of elements from $\{1,\ldots,n\}$ without replacement. Let $\delta,\delta'' \in (0,1)$, and $\delta' \in (0,e^{-1})$. Conditioned on the event $\mathcal{A}$ defined in Lem. 9, which has probability at least $1-\delta$, we have that with probability at least $1-\delta'-\delta''$ on the choice of σ,*

$$\begin{aligned}\mathbb{E}_L[\mathcal{M}^{\sigma,L}] &\leq \tfrac{2}{s_m-1}\log(\tfrac{2}{\delta''})\,\mathrm{MMD}^2(\mathbb{P},\mathbb{Q}) \\ &\quad + \tfrac{c'\log(1/\delta')}{\sqrt{s_m(s_m-1)}}\big(\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(5s)) + W(m,n,\delta/(5s))^2\big)\end{aligned}$$

and

$$\begin{aligned}&\tfrac{1}{s_m^2 s_n^2}\big(\textstyle\sum_{i=1}^{s_m}\sum_{j\neq j'\in\{1,\ldots,s_n\}}\langle(\hat{\mathbb{S}}^i_{m+n} - \hat{\mathbb{S}}^{\sigma(s_m+j)}_{m+n})\mathbf{k}, (\hat{\mathbb{S}}^i_{m+n} - \hat{\mathbb{S}}^{\sigma(s_m+j')}_{m+n})\mathbf{k}\rangle_{\mathbf{k}} \\ &+ \textstyle\sum_{j=1}^{s_n}\sum_{i\neq i'\in\{1,\ldots,s_m\}}\langle(\hat{\mathbb{S}}^{\sigma(i)}_{m+n} - \hat{\mathbb{S}}^{\sigma(s_m+j)}_{m+n})\mathbf{k}, (\hat{\mathbb{S}}^{\sigma(i')}_{m+n} - \hat{\mathbb{S}}^{\sigma(s_m+j)}_{m+n})\mathbf{k}\rangle_{\mathbf{k}} \\ &+ \textstyle\sum_{i=1}^{s_m}\sum_{j=1}^{s_n}\langle(\hat{\mathbb{S}}^{\sigma(i)}_{m+n} - \hat{\mathbb{S}}^{\sigma(s_m+j)}_{m+n})\mathbf{k}, (\hat{\mathbb{S}}^{\sigma(i)}_{m+n} - \hat{\mathbb{S}}^{\sigma(s_m+j)}_{m+n})\mathbf{k}\rangle_{\mathbf{k}}\big) \\ &\leq \tfrac{s-1}{s_m s_n}(\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(5s)) + W(m,n,\delta/(5s))^2 + \mathrm{MMD}^2(\mathbb{P},\mathbb{Q}))\end{aligned} \tag{36}$$

simultaneously, where c' is a universal constant and $W(m,n,\delta)$ is as defined in (33).

Proof. Given a vector $\epsilon = (\epsilon_i)_{i=1}^s \in \{\pm 1\}^s$, an m-tuple $L := \{l_1,...,l_m\}$ of elements from $\{1,\ldots,n\}$ without replacement, and a permutation σ on $\{1,\ldots,s\}$, define the permutation σ^ϵ on $\{1,\ldots,s\}$ as

$$\begin{cases}\sigma^{\epsilon,L}(i) = \epsilon_i\sigma(i) + (1-\epsilon_i)\sigma(s_m+l_i) & \text{for } i\in\{1,\ldots,s_m\}, \\ \sigma^{\epsilon,L}(s_m+l_i) = (1-\epsilon_i)\sigma(i) + \epsilon_i\sigma(s_m+l_i) & \text{for } i\in\{1,\ldots,s_m\}, \\ \sigma^{\epsilon,L}(j) = \sigma(j) & \text{otherwise.}\end{cases}$$

Using these objects, we have that

$$\begin{aligned}\mathcal{M}^{\sigma^{\epsilon,L},L} &\triangleq \frac{1}{s_m(s_m-1)} \sum_{i\neq i' \in \{1,\dots,s_m\}} \langle (\hat{\mathbb{S}}_{m+n}^{\sigma_{\epsilon,L}(i)} - \hat{\mathbb{S}}_{m+n}^{\sigma_{\epsilon,L}(s_m+l_i)})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\sigma_{\epsilon,L}(i')} - \hat{\mathbb{S}}_{m+n}^{\sigma_{\epsilon,L}(s_m+l_{i'})})\mathbf{k}\rangle_{\mathbf{k}} \\ &= \frac{1}{s_m(s_m-1)} \sum_{i\neq i' \in \{1,\dots,s_m\}} \epsilon_i \epsilon_{i'} \langle (\hat{\mathbb{S}}_{m+n}^{\sigma(i)} - \hat{\mathbb{S}}_{m+n}^{\sigma(s_m+l_i)})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\sigma(i')} - \hat{\mathbb{S}}_{m+n}^{\sigma(s_m+l_{i'})})\mathbf{k}\rangle_{\mathbf{k}} \end{aligned} \tag{37}$$

Note that given a fixed m-tuple L, if σ is distributed uniformly over the permutations of $\{1,\dots,s\}$, and ϵ contains i.i.d. Rademacher variables, then $\sigma^{\epsilon,L}$ is distributed uniformly over the permutations of $\{1,\dots,s\}$ as well. Hence, $\mathcal{M}^{\sigma,L}$ has the same distribution as $\mathcal{M}^{\sigma^{\epsilon,L},L}$, conditioned on $\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n$ and L (and consequently conditioned on $\mathbb{X}_m$ and $\mathbb{Y}_n$). Given σ, L, define the function $\rho_{\sigma,L} : \{1,\dots,s_m\} \to \{-1,0,1\}$ as

$$\rho_{\sigma,L}(i) = \begin{cases} 1 & \text{if } \sigma(i) \in \{1,\dots,s_m\} \text{ and } \sigma(s_m+l_i) \in \{s_m+1,\dots,s\} \\ -1 & \text{if } \sigma(i) \in \{s_m+1,\dots,s\} \text{ and } \sigma(s_m+l_i) \in \{1,\dots,s_m\} \\ 0 & \text{if } \sigma(i),\sigma(s_m+l_i) \in \{1,\dots,s_m\} \text{ or } \sigma(i),\sigma(s_m+l_i) \in \{s_m+1,\dots,s\} \end{cases}$$

We can rewrite (37) as

$$\begin{aligned}\mathcal{M}^{\sigma^{\epsilon,L},L} = \frac{1}{s_m(s_m-1)} \Big(&\sum_{i\neq i' \in \rho_{\sigma,L}^{-1}(\{-1,1\})} \tilde{\epsilon}_i \tilde{\epsilon}_{i'} \langle (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m})\mathbf{k}, (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m})\mathbf{k}\rangle_{\mathbf{k}} \\ &+ \sum_{i\neq i' \in \rho_{\sigma,L}^{-1}(\{0\})} \tilde{\epsilon}_i \tilde{\epsilon}_{i'} \langle (\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i)} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_i)})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i')} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_{i'})})\mathbf{k}\rangle_{\mathbf{k}} \\ &+ 2\sum_{i \in \rho_{\sigma,L}^{-1}(\{-1,1\}), i' \in \rho_{\sigma,L}^{-1}(\{0\})} \tilde{\epsilon}_i \tilde{\epsilon}_{i'} \langle (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i')} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_{i'})})\mathbf{k}\rangle_{\mathbf{k}} \Big)\end{aligned}$$

where we introduced $\tilde{\epsilon} = (\tilde{\epsilon}_i)_{i=1}^s$ and the permutation $\tilde{\sigma}$, defined as:

$$\tilde{\epsilon}_i = \begin{cases} -\epsilon_i & \text{if } \rho_\sigma(i) = -1, \\ \epsilon_i & \text{otherwise} \end{cases}$$

$$\tilde{\sigma}(i) = \begin{cases} \sigma(s_m+l_i) & \text{if } i \in \{1,\dots,s\} \text{ and } \rho_\sigma(i) = -1 \\ \sigma(j) & \text{if } i = s_m + l_j \text{ and } \rho_\sigma(j) = -1 \\ \sigma(i) & \text{otherwise.} \end{cases}$$

Note that conditioned on σ, $\tilde{\epsilon}$ is still a vector of i.i.d. Rademacher variables. Now, we will apply Lem. 8 on $\tilde{\epsilon} = (\tilde{\epsilon}_i)_{i=1}^s \in \mathbb{R}^s$ and the matrix $A = (A_{i,i'})_{i,i'=1}^s \in \mathbb{R}^{s\times s}$ defined as

$$A_{i,i'} = \begin{cases} 0 & \text{if } i = i' \\ \frac{1}{s_m(s_m-1)} (\langle (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m})\mathbf{k}, (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m})\mathbf{k}\rangle_{\mathbf{k}} - \mathrm{MMD}^2(\mathbb{P},\mathbb{Q})) & \text{if } i \neq i' \in \rho_\sigma^{-1}(\{\pm 1\}), \\ \frac{1}{s_m(s_m-1)} \langle (\hat{\mathbb{P}}_m^{(\tilde{\sigma}(i))} - \hat{\mathbb{Q}}_n^{(\tilde{\sigma}(s_m+i)-s_m)})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{(\tilde{\sigma}(i'))} - \hat{\mathbb{S}}_{m+n}^{(\tilde{\sigma}(s_m+l_{i'}))})\mathbf{k}\rangle_{\mathbf{k}} & \text{if } i \in \rho_\sigma^{-1}(\{\pm 1\}), i' \in \rho_\sigma^{-1}(\{0\}), \\ \frac{1}{s_m(s_m-1)} \langle (\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i)} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_i)})\mathbf{k}, (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m})\mathbf{k}\rangle_{\mathbf{k}} & \text{if } i \in \rho_\sigma^{-1}(\{0\}), i' \in \rho_\sigma^{-1}(\{\pm 1\}), \\ \frac{1}{s_m(s_m-1)} \langle (\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i)} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_i)})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(l_{i'})} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_{i'})})\mathbf{k}\rangle_{\mathbf{k}} & \text{if } i \neq i' \in \rho_\sigma^{-1}(\{0\}). \end{cases} \tag{38}$$

We develop the expressions that appear in Lem. 8. First, note that $\mathbb{E}[\tilde{\epsilon}^\top A \tilde{\epsilon}] = \mathrm{Tr}[A] = 0$. Also,

$$\begin{aligned}\tilde{\epsilon}^\top A \tilde{\epsilon} &= \frac{1}{s_m(s_m-1)} \Big(\sum_{i\neq i' \in \rho_\sigma^{-1}(\{-1,1\})} \tilde{\epsilon}_i \tilde{\epsilon}_{i'} (\langle (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m})\mathbf{k}, (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m})\mathbf{k}\rangle_{\mathbf{k}} - \mathrm{MMD}^2(\mathbb{P},\mathbb{Q})) \\ &+ 2\sum_{i \in \rho_\sigma^{-1}(\{-1,1\}), i' \in \rho_\sigma^{-1}(\{0\})} \tilde{\epsilon}_i \tilde{\epsilon}_{i'} \langle (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+i)-s_m})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i')} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_{i'})})\mathbf{k}\rangle_{\mathbf{k}} \\ &+ \sum_{i\neq i' \in \rho_\sigma^{-1}(\{0\})} \tilde{\epsilon}_i \tilde{\epsilon}_{i'} \langle (\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i)} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_i)})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i')} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s+i')})\mathbf{k}\rangle_{\mathbf{k}} \Big) \\ &= \mathcal{M}^{\sigma^{\epsilon,L},L} - \Big(\frac{1}{s_m(s_m-1)} \sum_{i\neq i' \in \rho_\sigma^{-1}(\{-1,1\})} \tilde{\epsilon}_i \tilde{\epsilon}_{i'}\Big) \mathrm{MMD}^2(\mathbb{P},\mathbb{Q}) = \mathcal{M}^{\sigma^{\epsilon,L},L} - c_{\tilde{\epsilon}} \mathrm{MMD}^2(\mathbb{P},\mathbb{Q}),\end{aligned} \tag{39}$$

where we defined $c_{\tilde{\epsilon}} \triangleq \frac{1}{s_m(s_m-1)} \sum_{i\neq i' \in \rho_\sigma^{-1}(\{-1,1\})} \tilde{\epsilon}_i \tilde{\epsilon}_{i'}$, and

$$\|A\|_{\mathrm{op}}^2 \leq \|A\|_{\mathrm{F}}^2$$

$$= \frac{1}{s_m^2(s_m-1)^2} \Big(\sum_{i\neq i' \in \rho_\sigma^{-1}(\{-1,1\})} (\langle (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s})\mathbf{k}, (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m})\mathbf{k}\rangle_{\mathbf{k}} - \mathrm{MMD}^2(\mathbb{P},\mathbb{Q}))^2 \tag{40}$$

$$+ 2\sum_{i \in \rho_\sigma^{-1}(\{-1,1\}), i' \in \rho_\sigma^{-1}(\{0\})} \langle (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i')} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_{i'})})\mathbf{k}\rangle_{\mathbf{k}}^2 \tag{41}$$

$$+ \sum_{i\neq i' \in \rho_\sigma^{-1}(\{0\})} \langle (\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i)} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_i)})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i')} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_{i'})})\mathbf{k}\rangle_{\mathbf{k}}^2 \Big) \tag{42}$$

Using that $\mathrm{MMD}^2(\mathbb{P},\mathbb{Q}) = \langle(\mathbb{P}-\mathbb{Q})\mathbf{k}, (\mathbb{P}-\mathbb{Q})\mathbf{k}\rangle_{\mathbf{k}}$, we upper-bound each of the terms in the right-hand side of (40):

$$\begin{aligned}
&|\langle(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m})\mathbf{k}, (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+i')-s_m})\mathbf{k}\rangle_{\mathbf{k}} - \langle(\mathbb{P}-\mathbb{Q})\mathbf{k}, (\mathbb{P}-\mathbb{Q})\mathbf{k}\rangle_{\mathbf{k}}| \\
&\leq |\langle(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m})\mathbf{k}, (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m})\mathbf{k} - (\mathbb{P}-\mathbb{Q})\mathbf{k}\rangle_{\mathbf{k}}| \\
&+ |\langle(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m})\mathbf{k} - (\mathbb{P}-\mathbb{Q})\mathbf{k}, (\mathbb{P}-\mathbb{Q})\mathbf{k}\rangle_{\mathbf{k}}| \\
&\leq \|(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m})\mathbf{k}\|_{\mathbf{k}} \cdot \|(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')} - \mathbb{P})\mathbf{k} - (\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m} - \mathbb{Q})\mathbf{k}\|_{\mathbf{k}} \\
&+ \|(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \mathbb{P})\mathbf{k} - (\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+i)-s_m} - \mathbb{Q})\mathbf{k}\|_{\mathbf{k}} \cdot \|(\mathbb{P}-\mathbb{Q})\mathbf{k}\|_{\mathbf{k}} \qquad (43) \\
&\leq \|(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \mathbb{P})\mathbf{k} - (\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+i)-s_m} - \mathbb{Q})\mathbf{k}\|_{\mathbf{k}} \cdot \|(\mathbb{P}-\mathbb{Q})\mathbf{k}\|_{\mathbf{k}} + \|(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')} - \mathbb{P})\mathbf{k} - (\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+i')-s_m} - \mathbb{Q})\mathbf{k}\|_{\mathbf{k}} \cdot \|(\mathbb{P}-\mathbb{Q})\mathbf{k}\|_{\mathbf{k}} \\
&+ \|(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \mathbb{P})\mathbf{k} - (\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m} - \mathbb{Q})\mathbf{k}\|_{\mathbf{k}} \cdot \|(\hat{\mathbb{P}}_m^{\sigma(i')} - \mathbb{P})\mathbf{k} - (\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m} - \mathbb{Q})\mathbf{k}\|_{\mathbf{k}} \\
&= (\mathrm{MMD}(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)}, \mathbb{P}) + \mathrm{MMD}(\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m}, \mathbb{Q}) + \mathrm{MMD}(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')}, \mathbb{P}) + \mathrm{MMD}(\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m}, \mathbb{Q})) \cdot \mathrm{MMD}(\mathbb{P},\mathbb{Q}) \\
&+ (\mathrm{MMD}(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)}, \mathbb{P}) + \mathrm{MMD}(\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m}, \mathbb{Q}))(\mathrm{MMD}(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')}, \mathbb{P}) + \mathrm{MMD}(\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m}, \mathbb{Q})).
\end{aligned}$$

An analogous but simpler approach yields bounds for the terms in (41) and (42):

$$\begin{aligned}
&|\langle(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i')} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_{i'})})\mathbf{k}\rangle_{\mathbf{k}}| \leq \|(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m})\mathbf{k}\|_{\mathbf{k}} \cdot \|(\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i')} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_{i'})})\mathbf{k}\|_{\mathbf{k}} \\
&= \mathrm{MMD}(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)}, \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m}) \cdot \mathrm{MMD}(\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i')}, \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_{i'})}) \\
&\leq (\mathrm{MMD}(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)}, \mathbb{P}) + \mathrm{MMD}(\mathbb{P},\mathbb{Q}) + \mathrm{MMD}(\mathbb{Q}, \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m})) \cdot (\mathrm{MMD}(\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i')}, \mathbb{S}) + \mathrm{MMD}(\mathbb{S}, \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_{i'})})), \qquad (44) \\
&|\langle(\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i)} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_i)})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i')} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_{i'})})\mathbf{k}\rangle_{\mathbf{k}}| \leq \mathrm{MMD}(\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i)}, \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_i)}) \cdot \mathrm{MMD}(\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i')}, \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_{i'})}) \\
&\leq (\mathrm{MMD}(\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i)}, \mathbb{S}) + \mathrm{MMD}(\mathbb{S}, \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_i)}))(\mathrm{MMD}(\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i')}, \mathbb{S}) + \mathrm{MMD}(\mathbb{S}, \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_{i'})}))
\end{aligned}$$

where $\mathbb{S}$ stands for $\mathbb{P}$ or $\mathbb{Q}$ as needed. Applying Lem. 9, we obtain that if KT-COMPRESS calls are run with value $\delta/(5s_m)$ and $\delta/(5s_n)$ respectively, conditioned on the event $\mathcal{A}$ we have that simultaneously for any $i, i' \in \{1, \dots, s_m\}$,

$$\begin{aligned}
&|\langle(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m})\mathbf{k}, (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m})\mathbf{k}\rangle_{\mathbf{k}} - \langle(\mathbb{P}-\mathbb{Q})\mathbf{k}, (\mathbb{P}-\mathbb{Q})\mathbf{k}\rangle_{\mathbf{k}}|, \\
&\leq \mathrm{MMD}(\mathbb{P},\mathbb{Q}) W(m,n,\delta/(5s)) + W(m,n,\delta/(5s))^2, \\
&|\langle(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i')} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_{i'})})\mathbf{k}\rangle_{\mathbf{k}}| \leq W(m,n,\delta/(5s))\,\mathrm{MMD}(\mathbb{P},\mathbb{Q}) + \tfrac{1}{2}W(m,n,\delta/(5s))^2 \qquad (45) \\
&|\,\mathrm{MMD}^2(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)}, \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m}) - \mathrm{MMD}^2(\mathbb{P},\mathbb{Q})| \leq \mathrm{MMD}(\mathbb{P},\mathbb{Q}) W(m,n,\delta/(5s)) + W(m,n,\delta/(5s))^2, \\
&|\langle(\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i)} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_i)})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i')} - \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_{i'})})\mathbf{k}\rangle_{\mathbf{k}}| \leq W(m,n,\delta/(5s))^2 \\
&\mathrm{MMD}^2(\hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(i)}, \hat{\mathbb{S}}_{m+n}^{\tilde{\sigma}(s_m+l_i)}) \leq W(m,n,\delta/(5s))^2,
\end{aligned}$$

where $W(m,n,\delta/(5s))$ is defined as in (33), and where we used that $\delta/(5s) \leq \delta/(5s_m)$ and $\delta/(5s) \leq \delta/(5s_n)$ since $s_m, s_n \leq s$. We conclude that conditioned on the event $\mathcal{A}$,

$$\begin{aligned}
\|A\|_{\mathrm{op}}^2 \leq \|A\|_{\mathrm{F}}^2 &\leq \tfrac{1}{s_m^2(s_m-1)^2}\big(|\rho_\sigma^{-1}(\{-1,1\})|(|\rho_\sigma^{-1}(\{-1,1\})|-1)(\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(5s)) + W(m,n,\delta/(5s))^2)^2 \qquad (46) \\
&+ |\rho_\sigma^{-1}(\{-1,1\})||\rho_\sigma^{-1}(\{0\})|(\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(5s)) + \tfrac{1}{2}W(m,n,\delta/(5s))^2)^2 \\
&+ |\rho_\sigma^{-1}(\{0\})|(|\rho_\sigma^{-1}(\{0\})|-1)W(m,n,\delta/(5s))^4\big) \\
&\leq \eta(m,n,\delta)^2 \triangleq \tfrac{1}{s_m(s_m-1)}(\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(5s)) + W(m,n,\delta/(5s))^2)^2
\end{aligned}$$

Applying Lem. 10, we get that with probability at least $1-\delta''$,

$$|\tfrac{1}{s_m}\textstyle\sum_{i\in\rho_\sigma^{-1}(\{-1,1\})}\tilde{\epsilon}_i| \leq |\tfrac{1}{s_m}\sum_{i=1}^{s_m}\epsilon_i| \leq \sqrt{\tfrac{2}{s_m}\log(\tfrac{2}{\delta''})},$$

and this implies that with probability at least $1-\delta''$,

$$\begin{aligned}
c_{\tilde{\epsilon}} &\triangleq \tfrac{1}{s_m(s_m-1)}\textstyle\sum_{i\neq i'\in\rho_\sigma^{-1}(\{-1,1\})}\tilde{\epsilon}_i\tilde{\epsilon}_{i'} = \tfrac{1}{s_m(s_m-1)}\big((\sum_{i\in\rho_\sigma^{-1}(\{-1,1\})}\tilde{\epsilon}_i)^2 - \sum_{i\in\rho_\sigma^{-1}(\{-1,1\})}\tilde{\epsilon}_i^2\big) \qquad (47) \\
&\leq \tfrac{1}{s_m(s_m-1)}(2s_m\log(\tfrac{2}{\delta''}) - |\rho_\sigma^{-1}(\{-1,1\})|) \leq \tfrac{2}{s_m-1}\log(\tfrac{2}{\delta''}).
\end{aligned}$$

Conditioned on the event $\mathcal{A}$ defined in Lem. 9, we obtain that for any $x \geq 0$,

$$\begin{aligned}
\Pr_\sigma(\mathbb{E}_L[\mathcal{M}^{\sigma^{\epsilon,L},L}] \geq x) &\overset{\text{(i)}}{=} \Pr_{\sigma,\tilde{\epsilon}}(\mathbb{E}_L[\tilde{\epsilon}^\top A \tilde{\epsilon} - c_{\tilde{\epsilon}} \operatorname{MMD}^2(\mathbb{P},\mathbb{Q})] \geq x) \\
&\overset{\text{(ii)}}{\leq} \delta'' + e^{-\lambda^* x} \mathbb{E}_{\sigma,\tilde{\epsilon}}[\exp(\lambda^* \mathbb{E}_L[\tilde{\epsilon}^\top A \tilde{\epsilon} - \tfrac{2}{s_m - 1} \log(\tfrac{2}{\delta''}) \operatorname{MMD}^2(\mathbb{P},\mathbb{Q})])] \\
&\overset{\text{(iii)}}{\leq} \delta'' + e^{-\lambda^* x} \mathbb{E}_{\sigma,\tilde{\epsilon},L}[\exp(\lambda^* (\tilde{\epsilon}^\top A \tilde{\epsilon} - \tfrac{2}{s_m - 1} \log(\tfrac{2}{\delta''}) \operatorname{MMD}^2(\mathbb{P},\mathbb{Q})))] \\
&= \delta'' + \mathbb{E}_{\sigma,L}[e^{-\lambda^*(x + \frac{2}{s_m-1} \log(\frac{2}{\delta''}) \operatorname{MMD}^2(\mathbb{P},\mathbb{Q}))} \mathbb{E}_{\tilde{\epsilon}}[\exp(\lambda^* \tilde{\epsilon}^\top A \tilde{\epsilon})]] \\
&\overset{\text{(iv)}}{\leq} \delta'' + \mathbb{E}_{\sigma,L}\big[\exp(-\tfrac{\lambda^*}{K^2}(x + \tfrac{2}{s_m-1} \log(\tfrac{2}{\delta''}) \operatorname{MMD}^2(\mathbb{P},\mathbb{Q})) + c''(\lambda^*)^2 \|A\|_{\mathrm{F}}^2)\big] \\
&\overset{\text{(v)}}{\leq} \delta'' + \exp(-\tfrac{\lambda^*}{K^2}(x + \tfrac{2}{s_m-1} \log(\tfrac{2}{\delta''}) \operatorname{MMD}^2(\mathbb{P},\mathbb{Q})) + c''(\lambda^*)^2 \eta(m,n,\delta)^2) \\
&\overset{\text{(vi)}}{\leq} \delta'' + \exp\big(-c \min\{\tfrac{(x + \frac{2}{s_m-1} \log(\frac{2}{\delta''}) \operatorname{MMD}^2(\mathbb{P},\mathbb{Q}))^2}{K^4 \eta(m,n,\delta)^2}, \tfrac{x + \frac{2}{s_m-1} \log(\frac{2}{\delta''}) \operatorname{MMD}^2(\mathbb{P},\mathbb{Q})}{K^2 \eta(m,n,\delta)}\}\big)
\end{aligned}$$

Here, (i) holds by (39), (ii) holds by (47) and the application of a Chernoff bound; the value of λ^* is to be set at a later point. Inequality (iii) holds by the convexity of the exponential function, and (iv) follows from Lem. 8 (48). (v) holds because conditioned on $\mathcal{A}$, $\|A\|_{\mathrm{F}}^2 \leq \eta(m,n,\delta)^2$ by equation (46), and (vi) follows from Lem. 8 (49). If we set

$$\delta' = \exp\big(-c \min\{\tfrac{(x + \frac{2}{s_m-1} \log(\frac{2}{\delta''}) \operatorname{MMD}^2(\mathbb{P},\mathbb{Q}))^2}{K^4 \eta(m,n,\delta)^2}, \tfrac{x + \frac{2}{s_m-1} \log(\frac{2}{\delta''}) \operatorname{MMD}^2(\mathbb{P},\mathbb{Q})}{K^2 \eta(m,n,\delta)}\}\big) \in (0, e^{-1})$$

we have that

$$x = \tfrac{c_0 (\operatorname{MMD}(\mathbb{P},\mathbb{Q}) W(m,n,\delta/(5s)) + W(m,n,\delta/(5s))^2) \log(1/\delta')}{\sqrt{s_m(s_m-1)}} + \tfrac{2}{s_m-1} \log(\tfrac{2}{\delta''}) \operatorname{MMD}^2(\mathbb{P},\mathbb{Q}),$$

where we defined $c_0 = K^2/c$ and we used that $\delta \in (0, e^{-1})$. We conclude that conditioned on $\mathcal{A}$, with probability at least $1 - \delta' - \delta''$,

$$\mathbb{E}_L[\mathcal{M}^{\sigma^{\epsilon,L},L}] \leq \tfrac{c_0 (\operatorname{MMD}(\mathbb{P},\mathbb{Q}) W(m,n,\delta/(5s)) + W(m,n,\delta/(5s))^2) \log(1/\delta')}{\sqrt{s_m(s_m-1)}} + \tfrac{2}{s_m-1} \log(\tfrac{2}{\delta''}) \operatorname{MMD}^2(\mathbb{P},\mathbb{Q}).$$

To show (36), we use the same arguments of (43), (44) and (45). □

Plugging the results of Lem. 7 into the right-hand side of (31) shows that conditioned on the event $\mathcal{A}$, with probability at least $1 - \delta' - \delta''$,

$$\begin{aligned}
\operatorname{MMD}^2(\hat{\mathbb{X}}_m^\sigma, \hat{\mathbb{Y}}_n^\sigma) \leq \tfrac{(s_m-1)(s_n-1)}{s_m s_n} \big(&\tfrac{2}{s_m-1} \log(\tfrac{2}{\delta''}) \operatorname{MMD}^2(\mathbb{P},\mathbb{Q}) \\
&+ \tfrac{c' \log(1/\delta')}{\sqrt{s_m(s_m-1)}} \big(\operatorname{MMD}(\mathbb{P},\mathbb{Q}) W(m,n,\delta/(5s)) + W(m,n,\delta/(5s))^2 \big) \big) \\
+ \tfrac{s-1}{s_m s_n} (\operatorname{MMD}(\mathbb{P},\mathbb{Q}) &W(m,n,\delta/(5s)) + W(m,n,\delta/(5s))^2 + \operatorname{MMD}^2(\mathbb{P},\mathbb{Q})),
\end{aligned}$$

which concludes the proof of Prop. 3 upon simplification.

Lemma 8 (Hanson-Wright inequality, Rudelson and Vershynin (2013), Thm. 1.1, adapted)**.** *Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with sub-Gaussian independent components X_i which satisfy $\mathbb{E} X_i = 0$, and $\mathbb{E}[X^\top A X] = 0$, and $\|X_i\|_{\psi_2} = \inf\{K' | \mathbb{E}[\exp(X_i^2/(K')^2)] < 2\} \leq K$. Let A be an $n \times n$ matrix. Then, there exists $c, c', c'' > 0$ and such that for every $t \geq 0$, $\lambda \leq c'/\|A\|_{op}$,*

$$\Pr(X^\top A X > t) \leq e^{-\lambda t} \mathbb{E}[e^{\lambda X^\top A X}] \leq \exp(-\lambda t/K^2 + c'' \lambda^2 \|A\|_F^2). \tag{48}$$

Optimizing this bound over $\lambda \leq c_0/\|A\|_{op}$, one obtains

$$\Pr(X^\top A X > t) \leq \exp(-\min\{\tfrac{t^2}{K^4 \|A\|_F^2}, \tfrac{t}{K^2 \|A\|_{op}}\}). \tag{49}$$

Lemma 9 (Simultaneous bound on the MMD errors of $\hat{\mathbb{P}}_m^i$ and $\hat{\mathbb{Q}}_n^j$)**.** *Suppose that* KT-COMPRESS *is run with value $\delta/(5s_m)$ and $\delta/(5s_n)$, i.e. $\hat{\mathbb{X}}_m^{(i)} \leftarrow$ KT-COMPRESS$(\mathbb{X}_m^{(i)}, \mathfrak{g}, \mathbf{k}, \mathbf{k}', \delta/(5s_m))$ and $\hat{\mathbb{Y}}_n^{(i)} \leftarrow$ KT-COMPRESS$(\mathbb{Y}_n^{(i)}, \mathfrak{g}, \mathbf{k}, \mathbf{k}', \delta/(5s_n))$. Then, with probability at least $1-\delta$, we have that simultaneously for all $i \in \{1, \dots, s_m\}$, $j \in \{1, \dots, s_n\}$,*

$$\begin{aligned}\mathrm{MMD}(\hat{\mathbb{P}}_m^i, \mathbb{P}) &\leq \tfrac{\sqrt{s_m}\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m/s_m, \delta/(5s_m), \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + c_{\delta/(5s_m)}\sqrt{\tfrac{s_m\|\mathbf{k}\|_\infty}{m}}, \\ \mathrm{MMD}(\hat{\mathbb{Q}}_n^j, \mathbb{Q}) &\leq \tfrac{\sqrt{s_n}\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, n/s_n, \delta/(5s_n), \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} + c_{\delta/(5s_n)}\sqrt{\tfrac{s_n\|\mathbf{k}\|_\infty}{n}}.\end{aligned}$$

We define $\mathcal{A}$ as the event that these s conditions take place simultaneously, and observe that $\Pr(\mathcal{A}) \geq 1-\delta$.

Proof. Using the notation in App. B.2, we can write that

$$\begin{aligned}\mathrm{MMD}(\hat{\mathbb{P}}_m^i, \mathbb{P}_m^i) &\leq \big\|u_{\mathrm{C}+,\mathbb{X}_m^i} \cdot s_m/m\big\|_2 = \Big\|\tfrac{s_m}{m}\textstyle\sum_{k=0}^{\log_4(m/s)-\mathfrak{g}-1}\sum_{j=1}^{4^k} w_{k,m/s_m} u_{k,j,\mathbb{X}_m}^{(i)}\Big\|_2 \\ &= \lambda_{\max}\big(\tfrac{s_m}{m}\textstyle\sum_{k=0}^{\log_4(m/s_m)-\mathfrak{g}-1}\sum_{j=1}^{4^k}\mathbf{M}_{k,j,\mathbb{X}_m}^{(i)}\big).\end{aligned}$$

We define

$$\begin{aligned}\sigma_i^2 &\triangleq \textstyle\sum_{k=0}^{\log_4(m/s_m)-\mathfrak{g}-1}\sum_{j=1}^{4^k}\big(\tfrac{2s_m}{m} w_{k,m/s_m}\ell'_{k,m/s_m}\big(a_{\ell'_{k,m/s_m},m/s_m}(\tfrac{\delta}{5s_m}) + v_{\ell'_{k,m/s_m},m/s_m}(\tfrac{\delta}{5s_m})\big)\big)^2 \\ &= s_m\textstyle\sum_{k=0}^{\log_4(m/s_m)-\mathfrak{g}-1}\big(\tfrac{1}{2^{\mathfrak{g}}\sqrt{m}}\big(C_a\sqrt{\|\mathbf{k}'\|_\infty} + C_v\sqrt{\|\mathbf{k}'\|_\infty\log\big(\tfrac{6\cdot 4^{\mathfrak{g}}\sqrt{m/s_m}(\beta_{m/s_m}+1)}{2^{\mathfrak{g}-k}\delta/(5s_m)}\big)}\mathfrak{M}_{\mathbb{X}_m^i,\mathbf{k}'}\big)\big)^2 \\ &\leq \tfrac{s_m(\log_4(m/s_m)-\mathfrak{g}-1)\|\mathbf{k}'\|_\infty}{4^{\mathfrak{g}}m}\big(C_a + C_v\sqrt{\log\big(\tfrac{15m(\log_4(m/s_m)-\mathfrak{g}-1)}{\delta}\big)}\mathfrak{M}_{\mathbb{X}_m^i,\mathbf{k}'}\big)^2\end{aligned}$$

When comparing this equation with (17), we have replaced δ by $\delta/(5s_m)$. We obtain that

$$\Pr(\mathrm{MMD}(\hat{\mathbb{P}}_m^i, \mathbb{P}_m^i) > \sigma_i\sqrt{8(\log(m+1)+t)}) \leq \tfrac{\delta}{10s_m} + e^{-t}, \qquad \text{for all} \quad t \geq 0.$$

Equivalently, with probability at least $1-\delta/(5s_m)$,

$$\mathrm{MMD}(\hat{\mathbb{P}}_m^i, \mathbb{P}_m^i) \leq \sqrt{8}\sigma_i(\sqrt{\log(m+1)} + \sqrt{\log(2s_m/\delta)}).$$

An application of Lem. 2 with $\mathbb{Q} = \mathbb{P}$ and $n = +\infty$ yields

$$\Pr\Big[\mathrm{MMD}(\mathbb{P}_m^i, \mathbb{P}) > (2 + \sqrt{2\log(\tfrac{10s}{\delta})})\sqrt{\tfrac{s\|\mathbf{k}\|_\infty}{m}}\Big] \leq \tfrac{\delta}{5s}.$$

Hence, with probability at least $1-2\delta/(5s_m)$,

$$\begin{aligned}\mathrm{MMD}(\hat{\mathbb{P}}_m^i, \mathbb{P}) &\leq \mathrm{MMD}(\hat{\mathbb{P}}_m^i, \mathbb{P}_m^i) + \mathrm{MMD}(\mathbb{P}_m^i, \mathbb{P}) \\ &\leq \sqrt{8}\sigma_i(\sqrt{\log(m+1)} + \sqrt{\log(2s_m/\delta)}) + (2 + \sqrt{2\log(\tfrac{2s_m}{\delta})})\sqrt{\tfrac{s_m\|\mathbf{k}\|_\infty}{m}}.\end{aligned}$$

Defining $\mathfrak{M}_{\mathbb{P},m,\mathbf{k}'}(\delta)$ as in (22) and (23), we obtain that with probability at least $1-\delta/(10s_m)$, $\mathfrak{M}_{\mathbb{X}_m^i,\mathbf{k}'} \leq \mathfrak{M}_{\mathbb{P},m/s_m,\mathbf{k}'}(\delta/(5s_m))$. Hence, with probability at least $1-5\delta/(10s_m) = 1-\delta/(2s_m)$,

$$\mathrm{MMD}(\hat{\mathbb{P}}_m^i, \mathbb{P}) \leq \tfrac{\sqrt{s_m}\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m/s_m, \delta/(5s_m), \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + c_{\delta/(5s_m)}\sqrt{s_m\tfrac{\|\mathbf{k}\|_\infty}{m}},$$

Using a union bound, we obtain the result. □

Lemma 10 (Chernoff bound for sums of Rademacher variables)**.** *Let $\epsilon = (\epsilon_i)_{i=1}^k$ be i.i.d. Rademacher variables. We have that for any $x > 0$,*

$$\Pr\big(|\tfrac{1}{k}\big(\textstyle\sum_{i=1}^k \epsilon_i\big)| > x\big) \leq e^{-D((1+x)/2||1/2)k} + e^{-D((1-x)/2||1/2)k} \leq 2\exp(-\tfrac{x^2k}{2}),$$

where $D(x||y) = x\log(\frac{x}{y}) + (1-x)\log(\frac{1-x}{1-y})$.

Proof. The first inequality holds by the Chernoff-Hoeffding theorem. The second inequality holds because $D((1+x)/2||1/2) = D((1-x)/2||1/2)$, and because for $p \geq 1/2$, we have that $D(p+x||p) \geq \frac{x^2}{2p(1-p)}$. □

D.4 Concluding the proof of Thm. 4

The following result is the basis for Thm. 4.

Lemma 11 (Putting everything together)**.** *Let $\beta \in (0,1)$ be arbitrary, and define $\tilde{\beta} = \frac{\beta}{1+\frac{\beta}{2}}$. Let $\alpha \in (0,1)$ and suppose that $k_\alpha \triangleq \lfloor \alpha(\mathcal{B}+1) \rfloor \geq 1$. Assume that* CTT *is run with $\mathcal{B} \geq \frac{1}{\alpha} - 1$ and $\delta = \min\{\frac{\tilde{\beta}}{6}, (\frac{\tilde{\beta}}{2})^{1/k_\alpha} \frac{\alpha}{30es}\}$. Define the function*

$$\tilde{Z}(m,n,\alpha,\beta) \triangleq Z(m,n,\beta) + \hat{Z}(m,n,\alpha,\tilde{\beta}),$$

where the functions Z and $\hat{Z}$ are defined in Lem. 4 (equation (24)) and Cor. 1 (equation (34)), respectively. If $\mathrm{MMD}(\mathbb{P},\mathbb{Q}) \geq \tilde{Z}(m,n,\alpha,\beta)$, then

$$\Pr[\mathrm{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n) \leq M_{(b_\alpha)}] \leq \beta. \tag{50}$$

Proof. Using Lem. 4, it suffices to see that with probability at least $1 - \frac{\tilde{\beta}}{2}$,

$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) \geq Z(m,n,\beta) + M_{(b_\alpha)}.$$

Cor. 1 implies that with probability at least $1 - \frac{\tilde{\beta}}{2}$,

$$M_{(b_\alpha)} \leq \hat{Z}(m,n,\alpha,\tilde{\beta}).$$

Hence, with probability at least $1 - \frac{\tilde{\beta}}{2}$, $\tilde{Z}(m,n,\alpha,\beta) \geq Z(m,n,\beta) + M_{(b_\alpha)}$. Using the assumption that $\mathrm{MMD}(\mathbb{P},\mathbb{Q}) \geq \tilde{Z}(m,n,\alpha,\beta)$, we obtain that (50) holds. $\square$

Proof of Thm. 4 To go from the statement of Lem. 11 to the one of Thm. 4, we write the function $\tilde{Z}(m,n,\alpha,\beta)$ in terms of its arguments, as follows:

$$\begin{aligned}
\tilde{Z}(m,n,\alpha,\beta) &= \tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P},m/s_m,\tilde{\beta}/6,\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + \tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n/s_n,\tilde{\beta}/6,\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} + c_{\tilde{\beta}/6}\big(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{\|\mathbf{k}\|_\infty}{n}}\big) \\
&\quad + \sqrt{\tfrac{2}{s_m}\log(\tfrac{8e^2}{\alpha}(\tfrac{4}{\tilde{\beta}})^{1/k_\alpha})}\,\mathrm{MMD}(\mathbb{P},\mathbb{Q}) \\
&\quad + \sqrt{\tfrac{1}{s_m}(2 + c'\log(\tfrac{4e}{\alpha}(\tfrac{4}{\tilde{\beta}})^{1/k_\alpha}))}\big(\sqrt{\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\tilde{\beta}/(20s))} + W(m,n,\tilde{\beta}/(20s))\big).
\end{aligned}$$

If we define

$$\begin{aligned}
&a = 1 - \sqrt{\tfrac{2}{s_m}\log(\tfrac{8e^2}{\alpha}(\tfrac{4}{\tilde{\beta}})^{1/k_\alpha})}, \\
&b = \sqrt{\tfrac{1}{s_m}(2 + c'\log(\tfrac{4e}{\alpha}(\tfrac{2}{\tilde{\beta}})^{1/k_\alpha}))W(m,n,\tilde{\beta}/(20s))}, \\
&c = \tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P},m/s_m,\tilde{\beta}/6,\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + \tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n/s_n,\tilde{\beta}/6,\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} + c_{\tilde{\beta}/6}\big(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{\|\mathbf{k}\|_\infty}{n}}\big) + \sqrt{\tfrac{2+c'\log(\frac{4e}{\alpha}(\frac{4}{\tilde{\beta}})^{1/k_\alpha})}{s_m}}W(m,n,\tilde{\beta}/(20s)), \\
&x = \sqrt{\mathrm{MMD}(\mathbb{P},\mathbb{Q})},
\end{aligned} \tag{51}$$

and we assume that $a > 0$, we can rewrite the condition $\mathrm{MMD}(\mathbb{P},\mathbb{Q}) \geq \tilde{Z}(m,n,\alpha,\beta)$ as $ax^2 - bx - c \geq 0$, which together with the positivity constraint on x is equivalent to $x \geq \frac{b+\sqrt{b^2+4ac}}{2a}$. A sufficient condition for this is $\sqrt{\mathrm{MMD}(\mathbb{P},\mathbb{Q})} \geq \frac{b}{a} + \sqrt{\frac{c}{a}}$, and yet another sufficient condition is $\mathrm{MMD}(\mathbb{P},\mathbb{Q}) \geq 2(\frac{b^2}{a^2} + \frac{c}{a})$. Since $0 < a < 1$ by assumption, the right-hand side of this equation is upper-bounded by

$$\begin{aligned}
\tfrac{2}{a^2}(b^2 + c) &= \tfrac{2}{(1-\sqrt{\frac{2}{s_m}\log(\frac{8e^2}{\alpha}(\frac{4}{\tilde{\beta}})^{1/k_\alpha})})^2}\Bigg(\tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P},m/s_m,\tilde{\beta}/6,\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + \tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n/s_n,\tilde{\beta}/6,\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} + c_{\tilde{\beta}/6}\big(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{\|\mathbf{k}\|_\infty}{n}}\big) \\
&\quad + 2\big(\tfrac{1}{s_m}(2 + c'\log(\tfrac{4e}{\alpha}(\tfrac{4}{\tilde{\beta}})^{1/k_\alpha})) + \sqrt{\tfrac{1}{s_m}(2 + c'\log(\tfrac{4e}{\alpha}(\tfrac{4}{\tilde{\beta}})^{1/k_\alpha}))}\big) \times \\
&\quad \big(\tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P},m/s_m,\tilde{\beta}/(20s_m),\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m/s_m}} + \tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n/s_n,\tilde{\beta}/(20s_n),\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n/s_n}} + c_{\tilde{\beta}/(20s)}\big(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m/s_m}} + \sqrt{\tfrac{\|\mathbf{k}\|_\infty}{n/s_n}}\big)\big)\Bigg).
\end{aligned}$$

We have that

$$\sqrt{\tfrac{2}{s_m}\log(\tfrac{8e^2}{\alpha}(\tfrac{4}{\tilde{\beta}})^{1/k_\alpha})} \leq \tfrac{3}{4} \Leftrightarrow s_m \geq \tfrac{32}{9}\log(\tfrac{8e^2}{\alpha}(\tfrac{4}{\tilde{\beta}})^{1/k_\alpha}) \tag{52}$$

Under (52), we have that

$$\frac{2}{(1-\sqrt{\frac{2}{s_m}\log(\frac{8e^2}{\alpha}(\frac{4}{\tilde{\beta}})^{1/k_\alpha})})^2} \leq \frac{2}{(1-\frac{3}{4})^2} = 32,$$

$$\tfrac{1}{\sqrt{s_m}}(2 + c'\log(\tfrac{4e}{\alpha}(\tfrac{4}{\tilde{\beta}})^{1/k_\alpha})) \leq \frac{2+c'\log(\frac{4e}{\alpha}(\frac{4}{\tilde{\beta}})^{1/k_\alpha})}{\sqrt{\frac{9}{32}\log(\frac{8e^2}{\alpha}(\frac{4}{\tilde{\beta}})^{1/k_\alpha})}} \leq \sqrt{\tfrac{32}{9}}(2 + c'\sqrt{\log(\tfrac{4e}{\alpha}(\tfrac{4}{\tilde{\beta}})^{1/k_\alpha})})$$

and consequently,

$$\begin{aligned}\tfrac{2}{a^2}(b^2+c) \leq 32\Bigg(&\frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P},m/s_m,\tilde{\beta}/6)}{2^{\mathfrak{g}}\sqrt{m}} + \frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n/s_n,\tilde{\beta}/6,\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} + c_{\tilde{\beta}/6}\big(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}}+\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{n}}\big)\\ &+ 2\big(\sqrt{\tfrac{9}{32}}+1\big)\big(2+c'\sqrt{\log(\tfrac{4e}{\alpha}(\tfrac{4}{\tilde{\beta}})^{1/k_\alpha})}\big)\bigg(\frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P},m/s_m,\tilde{\beta}/(20s),\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}}\\ &+ \sqrt{\tfrac{s_n}{s_m}}\frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n/s_n,\tilde{\beta}/(20s),\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} + c_{\tilde{\beta}/(20s)}\big(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{s_n\|\mathbf{k}\|_\infty}{s_m n}}\big)\bigg)\Bigg)\end{aligned}$$

The final result follows.

E Proof of Prop. 2: Power upper bounds for complete, block, and incomplete MMD tests

We prove the three parts of Prop. 2 one by one.

E.1 Proof of Prop. 2(a)

According to Gretton et al. (2007, Thm. 8) (see also Gretton et al. (2009, Eq. 2) for the exact formulation), the complete unbiased test with statistic $\mathrm{MMD}^2_{up}(\mathbb{X}_m, \mathbb{Y}_m)$ has the following asymptotic distribution under the null hypothesis:

$$m\,\mathrm{MMD}^2_{up}(\mathbb{X}_m, \mathbb{Y}_m) \to \textstyle\sum_{l=1}^\infty \lambda_l(z_l^2 - 2),$$

where $z_l \sim N(0, 2)$ i.i.d., $\to$ denotes convergence in distribution and λ_i are the solutions to the eigenvalue equation

$$\textstyle\int \tilde{k}(x_i, x_j)\psi_l(x_i)dP(x_i) = \lambda_l\psi_l(x_j),$$

where $\tilde{k}(x_i, x_j) = k(x_i, x_j) - \mathbb{E}_x k(x_i, x) - \mathbb{E}_x k(x, x_j) + \mathbb{E}_{x,x'}k(x, x')$. Hence, the variance of $\mathrm{MMD}^2_{up}(\mathbb{X}_m, \mathbb{Y}_m)$ is (asymptotically)

$$\mathrm{Var}(\mathrm{MMD}^2_{up}(\mathbb{X}_m, \mathbb{Y}_m)) = \tfrac{1}{m^2}\textstyle\sum_{l=1}^\infty \lambda_l^2 \mathrm{Var}(z_l^2 - 2) = \tfrac{4}{m^2}\textstyle\sum_{l=1}^\infty \lambda_l^2,$$

where the last equality holds because z_l^2 is distributed like a chi-squared distribution of one degree of freedom scaled by $\sqrt{2}$, which has variance 4. Since the asymptotic threshold $t_{1-\alpha}$ of level α for $\mathrm{MMD}^2_{up}(\mathbb{X}_m, \mathbb{Y}_m)$ is of the order of the standard deviation of $\mathrm{MMD}^2_{up}(\mathbb{X}_m, \mathbb{Y}_m)$, we can write

$$t_{1-\alpha} = \tfrac{K_{1-\alpha}}{m}, \tag{53}$$

where the constant $K_{1-\alpha}$ is of the order of $2\sqrt{\sum_{l=1}^\infty \lambda_l^2}$ and depends on α.

Under the alternative, $\mathrm{MMD}^2_{up}(\mathbb{X}_m, \mathbb{Y}_m)$ converges in distribution to a Gaussian according to

$$m^{1/2}(\mathrm{MMD}^2_{up}(\mathbb{X}_m, \mathbb{Y}_m) - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q})) \to N(0, \sigma_{up}^2), \tag{54}$$

where $\sigma_{up}^2 = 4(\mathbb{E}_{x,y}(\mathbb{E}_{x',y'}\mathbf{h}(x,x',y,y'))^2 - (\mathbb{E}_{x,y,x',y'}\mathbf{h}(x,x',y,y'))^2)$ (Gretton et al., 2007, Sec. 6), (Serfling, 2009, Sec. 5.5). Let $z = (x,y)$, $z' = (x',y')$ and $\mathbf{h}(z,z') = \mathbf{h}(x,x',y,y')$. If $\langle \cdot, \cdot \rangle_{\mathbf{k}}$ denotes the RKHS inner product, note that

$$\begin{aligned}
&|\mathbb{E}_{z'}[\mathbf{h}(z,z')]| \\
&= |\mathbb{E}_{x',y'}[\mathbf{k}(x,x') + \mathbf{k}(y,y') - \mathbf{k}(x,y') - \mathbf{k}(x',y)]| \\
&= \left| \int \mathbf{k}(x,x')\, d(\mathbb{P}-\mathbb{Q})(x') - \int \mathbf{k}(y,x')\, d(\mathbb{P}-\mathbb{Q})(x') \right| = \left| \int (\mathbf{k}(x,x') - \mathbf{k}(y,x'))\, d(\mathbb{P}-\mathbb{Q})(x') \right| \\
&= \left| \int \mathbf{k}(x'',x')\, d(\mathbb{P}-\mathbb{Q})(x')\, d(\delta_x - \delta_y)(x'') \right| = \left| \left\langle \int \mathbf{k}(\cdot,x')\, d(\mathbb{P}-\mathbb{Q})(x'), \int \mathbf{k}(\cdot,x')\, d(\delta_x - \delta_y)(x') \right\rangle_{\mathbf{k}} \right| \\
&\leq \left\| \int \mathbf{k}(\cdot,x')\, d(\mathbb{P}-\mathbb{Q})(x') \right\|_{\mathbf{k}} \left\| \int \mathbf{k}(\cdot,x')\, d(\delta_x - \delta_y)(x') \right\|_{\mathbf{k}} \\
&= \mathrm{MMD}(\mathbb{P},\mathbb{Q})(\mathbf{k}(x,x) + \mathbf{k}(y,y) - 2\mathbf{k}(x,y)).
\end{aligned} \tag{55}$$

Hence, using equation (55) we obtain an upper bound on σ_{up}^2:

$$\begin{aligned}
\sigma_{up}^2 &:= 4(\mathbb{E}_z[(\mathbb{E}_{z'}[\mathbf{h}(z,z')])^2] - (\mathbb{E}_{z,z'}[\mathbf{h}(z,z')])^2) \leq 4\mathbb{E}_z[(\mathbb{E}_{z'}[\mathbf{h}(z,z')])^2] \\
&\leq 4\mathrm{MMD}^2(\mathbb{P},\mathbb{Q})\mathbb{E}_{x,y}[(\mathbf{k}(x,x) + \mathbf{k}(y,y) - 2\mathbf{k}(x,y))^2] = 4\mathrm{MMD}^2(\mathbb{P},\mathbb{Q})\mathbb{E}_z[\mathbf{h}(z,z)^2].
\end{aligned} \tag{56}$$

Asymptotically, we obtain that under the alternative distribution

$$\begin{aligned}
&\Pr(\mathrm{MMD}_{up}^2(\mathbb{X}_m,\mathbb{Y}_m) < t_{1-\alpha}) \\
&= \Pr(\tfrac{\sqrt{m}}{\sigma_{up}}(\mathrm{MMD}_{up}^2(\mathbb{X}_m,\mathbb{Y}_m) - \mathrm{MMD}^2(\mathbb{P},\mathbb{Q})) < \tfrac{\sqrt{m}}{\sigma_{up}}(t_{1-\alpha} - \mathrm{MMD}^2(\mathbb{P},\mathbb{Q}))) \\
&= \Phi(\tfrac{\sqrt{m}}{\sigma_{up}}(t_{1-\alpha} - \mathrm{MMD}^2(\mathbb{P},\mathbb{Q}))),
\end{aligned}$$

where Φ denotes the CDF of a standard Gaussian. Hence, the condition that the Type II error be upper-bounded by $\beta \in (0, 1/2)$ translates to

$$\begin{aligned}
&\tfrac{\sqrt{m}}{\sigma_{up}}(t_{1-\alpha} - \mathrm{MMD}^2(\mathbb{P},\mathbb{Q})) \leq \Phi^{-1}(\beta) \\
&\qquad \Leftrightarrow \mathrm{MMD}^2(\mathbb{P},\mathbb{Q}) \geq t_{1-\alpha} - \tfrac{\sigma_{up}}{\sqrt{m}}\Phi^{-1}(\beta) = t_{1-\alpha} + \tfrac{\sigma_{up}}{\sqrt{m}}\Phi^{-1}(1-\beta)
\end{aligned} \tag{57}$$

Replacing $t_{1-\alpha}$ by its expression in (53) and using the upper bound (56), we get that a sufficient condition for (57) is

$$\mathrm{MMD}^2(\mathbb{P},\mathbb{Q}) - 2\sqrt{\tfrac{\mathbb{E}_z[\mathbf{h}(z,z)^2]}{m}}\Phi^{-1}(1-\beta)\,\mathrm{MMD}(\mathbb{P},\mathbb{Q}) - \tfrac{K_{1-\alpha}}{m} \geq 0.$$

Solving the corresponding second-degree equation, this is equivalent to

$$\mathrm{MMD}(\mathbb{P},\mathbb{Q}) \geq \tfrac{\sqrt{\mathbb{E}_z[\mathbf{h}(z,z)^2]}\Phi^{-1}(1-\beta) + \sqrt{\mathbb{E}_z[\mathbf{h}(z,z)^2]\Phi^{-1}(1-\beta)^2 + K_{1-\alpha}}}{\sqrt{m}} = O(1/\sqrt{m})$$

A necessary condition for (57) to hold is $\mathrm{MMD}(\mathbb{P},\mathbb{Q}) \geq \sqrt{\tfrac{K_{1-\alpha}}{m}} = \Omega(1/\sqrt{m})$, which concludes the proof of this part.

E.2 Proof of Prop. 2(b)

By the definition of $\mathrm{MMD}_B^2(\mathbb{X}_m,\mathbb{Y}_m)$ in (3), it is the average of m/B independent instances $\eta_i(\mathbb{X}_m,\mathbb{Y}_m)$ of the quadratic estimator (2), each with sample size B. Hence, in the regime $m/B \to \infty$, we have

$$\tfrac{m}{B}(\mathrm{MMD}_B^2(\mathbb{X}_m,\mathbb{Y}_m) - \mathrm{MMD}^2(\mathbb{P},\mathbb{Q})) \to N(0, \mathrm{Var}(\eta_i(\mathbb{X}_m,\mathbb{Y}_m))).$$

Using the argument of App. E.1, under the null hypothesis, we obtain that asymptotically $\mathrm{Var}(\eta_i(\mathbb{X}_m,\mathbb{Y}_m)) = \frac{4}{B^2}\sum_{l=1}^{\infty}\lambda_l^2$. Hence, under the null hypothesis, $\sqrt{\frac{Bm}{4\sum_{l=1}^{\infty}\lambda_l^2}}\mathrm{MMD}_B^2(\mathbb{X}_m,\mathbb{Y}_m) \to N(0,1)$. We derive the expression for the threshold $t_{1-\alpha}$ corresponding to the level α:

$$\begin{aligned}
&\Pr(\mathrm{MMD}_B^2(\mathbb{X}_m,\mathbb{Y}_m) < t_{1-\alpha}) = \Pr(\sqrt{\tfrac{Bn}{4\sum_{l=1}^{\infty}\lambda_l^2}}\mathrm{MMD}_B^2(\mathbb{X}_m,\mathbb{Y}_m) < \sqrt{\tfrac{Bn}{4\sum_{l=1}^{\infty}\lambda_l^2}}t_{1-\alpha}) \\
&= \Phi(\sqrt{\tfrac{Bm}{4\sum_{l=1}^{\infty}\lambda_l^2}}t_{1-\alpha}) = 1 - \alpha.
\end{aligned}$$

This implies that

$$t_{1-\alpha} = \sqrt{\tfrac{4\sum_{l=1}^{\infty}\lambda_l^2}{Bm}}\Phi^{-1}(1-\alpha). \tag{58}$$

Reusing (54), we have that asymptotically, under the alternative hypothesis, $\mathrm{Var}(\eta_i(\mathbb{X}_m, \mathbb{Y}_m)) = \sigma_{up}^2/B$. Hence, under the alternative hypothesis,

$$\tfrac{\sqrt{m}}{\sigma_{up}}(\mathrm{MMD}_B^2(\mathbb{X}_m, \mathbb{Y}_m) - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q})) \to N(0, 1)).$$

We conclude that asymptotically,

$$\begin{aligned}
&\Pr(\mathrm{MMD}_B^2(\mathbb{X}_m, \mathbb{Y}_m) < t_{1-\alpha}) \\
&= \Pr(\tfrac{\sqrt{m}}{\sigma_{up}}(\mathrm{MMD}_B^2(\mathbb{X}_m, \mathbb{Y}_m) - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q})) < \tfrac{\sqrt{m}}{\sigma_{up}}(t_{1-\alpha} - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}))) \\
&= \Phi(\tfrac{\sqrt{m}}{\sigma_{up}}(t_{1-\alpha} - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}))).
\end{aligned}$$

Hence, the condition that the Type II error be upper-bounded by $\beta \in (0, 1/2)$ translates to

$$\tfrac{\sqrt{m}}{\sigma_{up}}(t_{1-\alpha} - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q})) \le \Phi^{-1}(\beta) \Leftrightarrow \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) \ge t_{1-\alpha} + \tfrac{\sigma_{up}}{\sqrt{m}}\Phi^{-1}(1-\beta) \tag{59}$$

Replacing $t_{1-\alpha}$ by its expression in (58) and using the upper bound (56), we get that a necessary condition for (59) to hold is

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) \ge \sqrt{t_{1-\alpha}} = (\tfrac{4\sum_{l=1}^{\infty}\lambda_l^2}{Bm})^{1/4}\sqrt{\Phi^{-1}(1-\alpha)} = \Omega(1/(Bm)^{1/4}).$$

Also, a sufficient condition for (59) is

$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) - 2\sqrt{\tfrac{\mathbb{E}_z[\mathbf{h}(z,z)^2]}{m}}\Phi^{-1}(1-\beta)\,\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) - 2\sqrt{\tfrac{\sum_{l=1}^{\infty}\lambda_l^2}{Bm}}\Phi^{-1}(1-\alpha) \ge 0.$$

Solving the corresponding second-degree equation, this is equivalent to

$$\begin{aligned}
\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) &\ge \sqrt{\tfrac{\mathbb{E}_z[\mathbf{h}(z,z)^2]}{m}}\Phi^{-1}(1-\beta) + \sqrt{\tfrac{\mathbb{E}_z[\mathbf{h}(z,z)^2]}{m}\Phi^{-1}(1-\beta)^2 + 2\sqrt{\tfrac{\sum_{l=1}^{\infty}\lambda_l^2}{Bm}}\Phi^{-1}(1-\alpha)} \\
&= O(1/(Bm)^{1/4}).
\end{aligned}$$

E.3 Proof of Prop. 2(c)

Yamada et al. (2019, Cor. 3) show that when the pairs in the design $\mathcal{D}$ are chosen i.i.d. (with replacement), and $\lim_{m,|\mathcal{D}|\to\infty} m^{-2}|\mathcal{D}| = 0$, $0 < \gamma = \lim_{m,|\mathcal{D}|\to\infty} m^{-1}|\mathcal{D}| < \infty$, the incomplete MMD statistic $\mathrm{MMD}_{\text{inc}}^2(\mathbb{X}_m, \mathbb{Y}_m)$ is asymptotically distributed according to

$$\begin{cases} |\mathcal{D}|^{1/2}\,\mathrm{MMD}_{\text{inc}}^2(\mathbb{X}_m, \mathbb{Y}_m) \to N(0, \sigma^2) & \text{if } \mathbb{P} = \mathbb{Q} \\ |\mathcal{D}|^{1/2}(\mathrm{MMD}_{\text{inc}}^2(\mathbb{X}_m, \mathbb{Y}_m) - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q})) \to N(0, \sigma^2 + \gamma\sigma_{up}^2), & \text{if } \mathbb{P} \ne \mathbb{Q}. \end{cases}$$

where $\sigma^2 = \mathbb{E}_{z,z'}(\mathbf{h}(z, z') - \mathbb{E}_{z,z'}\mathbf{h}(z, z'))^2$ and σ_{up}^2 is defined in (56).

We derive the expression for the threshold $t_{1-\alpha}$ corresponding to the level α using the asymptotic distribution under the null hypothesis:

$$\begin{aligned}
&\Pr(\mathrm{MMD}_{\text{inc}}^2(\mathbb{X}_m, \mathbb{Y}_m) < t_{1-\alpha}) = \Pr(\tfrac{|\mathcal{D}|^{1/2}}{\sigma}\,\mathrm{MMD}_{\text{inc}}^2(\mathbb{X}_m, \mathbb{Y}_m) < \tfrac{|\mathcal{D}|^{1/2}}{\sigma}t_{1-\alpha}) \\
&= \Phi(\tfrac{|\mathcal{D}|^{1/2}}{\sigma}t_{1-\alpha}) = 1 - \alpha \Leftrightarrow t_{1-\alpha} = \tfrac{\sigma}{|\mathcal{D}|^{1/2}}\Phi^{-1}(1-\alpha).
\end{aligned}$$

And then, under the alternative hypothesis and asymptotically,

$$\begin{aligned}
&\Pr(\mathrm{MMD}_{\text{inc}}^2(\mathbb{X}_m, \mathbb{Y}_m) < t_{1-\alpha}) \\
&= \Pr(\tfrac{|\mathcal{D}|^{1/2}}{\sqrt{\sigma^2+\gamma^2\sigma_{up}^2}}(\mathbb{X}_m, \mathbb{Y}_m) - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q})) < \tfrac{|\mathcal{D}|^{1/2}}{\sqrt{\sigma^2+\gamma^2\sigma_{up}^2}}(t_{1-\alpha} - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}))) \\
&= \Phi(\tfrac{|\mathcal{D}|^{1/2}}{\sqrt{\sigma^2+\gamma^2\sigma_{up}^2}}(t_{1-\alpha} - \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}))).
\end{aligned}$$

Proceeding as in App. E.2, we obtain that a necessary condition for the Type II error to be upper-bounded by $\beta \in (0, 1/2)$ is that

$$\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) \geq \sqrt{t_{1-\alpha}} = \sqrt{\tfrac{\sigma}{|\mathcal{D}|^{1/2}} \Phi^{-1}(1-\alpha)} = \Omega(1/|\mathcal{D}|^{1/4}),$$

and that a sufficient condition is

$$\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) \geq t_{1-\alpha} + \tfrac{\sigma + 2\gamma \mathrm{MMD}(\mathbb{P},\mathbb{Q}) \sqrt{\mathbb{E}_z[\mathbf{h}(z,z)^2]}}{|\mathcal{D}|^{1/2}} \Phi^{-1}(1-\beta)$$

In order to derive this, we used that $\sqrt{\sigma^2 + 4\gamma^2 \mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) \mathbb{E}_z[\mathbf{h}(z,z)^2]} \leq \sigma + 2\gamma \mathrm{MMD}(\mathbb{P}, \mathbb{Q}) \sqrt{\mathbb{E}_z[\mathbf{h}(z,z)^2]}$. Solving the corresponding second-degree equation, this is equivalent to

$$\begin{aligned}\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) &\geq \tfrac{\gamma \sqrt{\mathbb{E}_z[\mathbf{h}(z,z)^2]}}{|\mathcal{D}|^{1/2}} \Phi^{-1}(1-\beta) + \sqrt{\tfrac{\gamma^2 \mathbb{E}_z[\mathbf{h}(z,z)^2]}{|\mathcal{D}|} \Phi^{-1}(1-\beta)^2 + t_{1-\alpha} + \tfrac{\sigma \Phi^{-1}(1-\beta)}{|\mathcal{D}|^{1/2}}} \\ &= O(1/|\mathcal{D}|^{1/4}).\end{aligned}$$

F Proof of Thm. 2: LR-CTT exactness and power

The proof of Thm. 2 follows the structure of the proof of Thm. 1. We first introduce a detailed statement of the result of Thm. 2.

Theorem 5. *Low-Rank* CTT *(Alg. 2) has size exactly equal to the level α for all $\mathbb{P}$. Suppose Low-Rank* CTT *(Alg. 2) is run with level α, replication count $\mathcal{B} \geq \frac{1}{\alpha} - 1$, and coreset count $s_m \geq (32/9) \log(\frac{2e}{\gamma})$ for $\gamma \triangleq \frac{\alpha}{4e} (\frac{\beta}{4+2\beta})^{\frac{1}{\lfloor \alpha(\mathcal{B}+1) \rfloor}}$. Let $\tilde{\beta} = \beta/(1+\beta/2)$. Then* LR-CTT *has power*

$$\Pr[\Delta_{\mathrm{CTT}}(\mathbb{X}_m, \mathbb{Y}_n) = 1] \geq 1 - \beta$$

whenever

$$\begin{aligned}\mathrm{MMD}_{\mathbf{k}}(\mathbb{P}, \mathbb{Q}) \geq 32 \Big(&\tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m/s_{m,r}, \tilde{\beta}/6, \mathfrak{g})}{2^{\mathfrak{g}} \sqrt{m}} + \tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, n/s_{n,r}, \tilde{\beta}/6, \mathfrak{g})}{2^{\mathfrak{g}} \sqrt{n}} + c_{\tilde{\beta}/6} \big(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{\|\mathbf{k}\|_\infty}{n}} \big) + 2\epsilon_{\Phi_r}(\mathbb{X}_m, \mathbb{Y}_n) \\ &+ 2^{1+1/4} \big(\sqrt{\tfrac{9}{32}} + 1 \big) \big(2 + c' \sqrt{\log(\gamma)} \big) \times \Big(\epsilon_{\Phi_r}(\mathbb{X}_m, \mathbb{Y}_n) + \tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m/s_{m,r}, \tilde{\beta}/(20 s_{m,r}), \mathfrak{g})}{2^{\mathfrak{g}} \sqrt{m}} \\ &+ \sqrt{\tfrac{s_n}{s_m}} \tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, n/s_{n,r}, \tilde{\beta}/(20 s_{n,r}), \mathfrak{g})}{2^{\mathfrak{g}} \sqrt{n}} + c_{\tilde{\beta}/(20 s_r)} \big(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{s_n \|\mathbf{k}\|_\infty}{s_m n}} \big) \Big) \Big).\end{aligned}$$

It is important to remark that Alg. 2 involves two separate parameters: the number of compression bins $s_r := s_{m,r} + s_{n,r}$ and the number of permutation bins $s = s_m + s_n$. The former is always larger or equal than the latter and in particular s divides s_r, that is, the compressed outputs of s_r/s compression bins are grouped together into a single permutation bin.

We formulate a statement which is analogous to Lem. 4, but for the LR-CTT test statistic, and with a slightly different lower bound $Z(m, n, \beta)$.

Lemma 12 (Upper bound on acceptance probability from upper bound on threshold)**.** *Let $1 \geq \beta > 0$ arbitrary, and define $\tilde{\beta} = \frac{\beta}{1+\frac{\beta}{2}}$. Define the function*

$$Z(m, n, \beta) \triangleq \tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m/s_{m,r}, \tilde{\beta}/6, \mathfrak{g})}{2^{\mathfrak{g}} \sqrt{m}} + \tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, n/s_{n,r}, \tilde{\beta}/6, \mathfrak{g})}{2^{\mathfrak{g}} \sqrt{n}} + c_{\tilde{\beta}/6} \big(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{\|\mathbf{k}\|_\infty}{n}} \big) + 2\epsilon_{\Phi_r}(\mathbb{X}_m, \mathbb{Y}_n),$$

which is equal to the upper bound in (7) when we make the choice $\delta = \tilde{\beta}/6$. If $\Pr\big[\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) \geq Z(m, n, \beta) + M_{(b_\alpha)}\big] \geq \frac{1}{1+\frac{\beta}{2}}$ then $\Pr[\mathrm{MMD}_{\Phi_r}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n) \leq M_{(b_\alpha)}] \leq \beta$. Here, $\mathbf{R}_{\mathbf{k},\mathbf{k}'}$ are defined in App. B.1, and $\epsilon^2_{\Phi_r}(\mathbb{P}_m, \mathbb{Q}_n) = \sup_{x,x' \in \mathrm{supp}(\mathbb{P}_m) \cup \mathrm{supp}(\mathbb{Q}_m)} |\langle \Phi_r(x), \Phi_r(x') \rangle - \mathbf{k}(x, x')|$.

Proof. The proof structure is the same as for Lem. 4, but in this case we must use instead that with probability at least $1 - \frac{\tilde{\beta}}{2}$,

$$\begin{aligned}|\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) - \mathrm{MMD}_{\Phi_r}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)| &\leq |\mathrm{MMD}(\mathbb{P}, \mathbb{Q}) - \mathrm{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)| \\ &\quad + |\mathrm{MMD}_{\Phi_r}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n) - \mathrm{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)|\end{aligned}$$

The first term in the right-hand side is upper-bounded by

$$\frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m/s_{m,r}, \tilde{\beta}/6, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + \frac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, n/s_{n,r}, \tilde{\beta}/6, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} + c_{\tilde{\beta}/6}\big(\sqrt{\frac{\|\mathbf{k}\|_\infty}{m}} + \sqrt{\frac{\|\mathbf{k}\|_\infty}{n}}\big),$$

while the second term can be upper-bounded as follows, since MMD is nonnegative:

$$\begin{aligned}
&|\mathrm{MMD}_{\Phi_r}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n) - \mathrm{MMD}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)| \\
&\leq \sqrt{|\mathrm{MMD}^2_{\Phi_r}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n) - \mathrm{MMD}^2(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)|} \\
&\leq \sqrt{|\langle \Phi_r(\hat{X}_m) - \Phi_r(\hat{Y}_n), \Phi_r(\hat{X}_m) - \Phi_r(\hat{Y}_n)\rangle - \langle (\hat{\mathbb{P}}_m - \hat{\mathbb{Q}}_n)\mathbf{k}, (\hat{\mathbb{P}}_m - \hat{\mathbb{Q}}_n)\mathbf{k}\rangle_{\mathbf{k}}|} \\
&\leq \bigg(|\langle \Phi_r(\hat{X}_m), \Phi_r(\hat{X}_m)\rangle - \langle \hat{\mathbb{P}}_m\mathbf{k}, \hat{\mathbb{P}}_m\mathbf{k}\rangle_{\mathbf{k}}| + 2|\langle \Phi_r(\hat{X}_m), \Phi_r(\hat{Y}_n)\rangle - \langle \hat{\mathbb{P}}_m\mathbf{k}, \hat{\mathbb{Q}}_n\mathbf{k}\rangle_{\mathbf{k}}| \\
&\qquad + |\langle \Phi_r(\hat{Y}_n), \Phi_r(\hat{Y}_n)\rangle - \langle \hat{\mathbb{Q}}_n\mathbf{k}, \hat{\mathbb{Q}}_n\mathbf{k}\rangle_{\mathbf{k}}| \bigg)^{1/2} \leq 2 \sup_{x,x' \in \mathbb{X}_m \cup \mathbb{Y}_n} \sqrt{|\langle \Phi_r(x), \Phi_r(x')\rangle - \mathbf{k}(x,x')|} = 2\epsilon_{\Phi_r}(\mathbb{P}_m, \mathbb{Q}_n).
\end{aligned} \tag{60}$$

□

The following proposition, an analog of Prop. 3, establishes a high-probability upper-bound on the MMD obtained by permuting the data samples.

Lemma 13 (Tail bound on $\mathrm{MMD}^2_{\Phi_r}(\hat{\mathbb{X}}^\sigma_m, \hat{\mathbb{Y}}^\sigma_n)$ conditioned on $\hat{\mathbb{X}}_m$, $\hat{\mathbb{Y}}_n$, and Φ_r)**.** *Let σ be a uniformly random permutation over $\{1, \ldots, s\}$. Let $\delta, \delta'' \in (0,1)$, and $\delta \in (0, e^{-1})$. There is an event $\mathcal{A}$ of probability at least $1-\delta$ concerning the draw of $\hat{\mathbb{X}}_m$, $\hat{\mathbb{Y}}_n$, such that conditioned on $\mathcal{A}$, with probability at least $1 - \delta' - \delta''$ on the draw of σ,*

$$\begin{aligned}
\mathrm{MMD}^2_{\Phi_r}(\hat{\mathbb{X}}^\sigma_m, \hat{\mathbb{Y}}^\sigma_n) &\leq \tfrac{2}{s_m}(\log(\tfrac{2}{\delta''}) + 1)\,\mathrm{MMD}^2(\mathbb{P}, \mathbb{Q}) \\
&\quad + \tfrac{c'\log(1/\delta')+2}{s_m}\big(2\epsilon^4_{\Phi_r}(\mathbb{P}_m, \mathbb{Q}_n) + 2(\mathrm{MMD}(\mathbb{P}, \mathbb{Q})W(m,n,\delta/(5s)) + W(m,n,\delta/(5s))^2)^2\big)^{1/2}
\end{aligned}$$

where $s_r = s_m + s_n$, c' is a universal constant and

$$W(m,n,\delta) = 2\big(\tfrac{\sqrt{s_m}\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P}, m/s_{m,r}, \delta, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + \tfrac{\sqrt{s_n}\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q}, n/s_{n,r}, \delta, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} + c_\delta\big(\sqrt{\tfrac{s_m\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{s_n\|\mathbf{k}\|_\infty}{n}}\big)\big). \tag{61}$$

Here, the KT-COMPRESS *error inflation factors $\mathbf{R}_{\mathbf{k},\mathbf{k}'}$ and the factor c_δ are defined as in Lem. 1, and $\xi(\mathbb{P}_m, \mathbb{Q}_n) = \sup_{x,x' \in \mathrm{supp}(\mathbb{P}_m) \cup \mathrm{supp}(\mathbb{Q}_m)} |\langle \Phi_r(x), \Phi_r(x')\rangle - \mathbf{k}(x,x')|$.*

Proof. For an arbitrary distribution $\mathbb{P}$, we denote $\Phi_r(\mathbb{P}) = \mathbb{E}_{x\sim\mathbb{P}}\Phi_r(x) \in \mathbb{R}^r$. We can write

$$\begin{aligned}
\mathrm{MMD}^2_{\Phi_r}(\hat{\mathbb{X}}^\sigma_m, \hat{\mathbb{Y}}^\sigma_n) &= \langle \Phi_r(\hat{\mathbb{P}}^\sigma_m) - \Phi_r(\hat{\mathbb{Q}}^\sigma_n), \Phi_r(\hat{\mathbb{P}}^\sigma_m) - \Phi_r(\hat{\mathbb{Q}}^\sigma_n)\rangle \\
&= \tfrac{1}{s_m^2 s_n^2}\textstyle\sum_{i=1}^{s_m}\sum_{i'=1}^{s_m}\sum_{j=1}^{s_n}\sum_{j'=1}^{s_n}\langle \Phi_r(\hat{\mathbb{S}}^{(\sigma(i))}_{m+n}) - \Phi_r(\hat{\mathbb{S}}^{(\sigma(s_m+j))}_{m+n}), \Phi_r(\hat{\mathbb{S}}^{(\sigma(i'))}_{m+n}) - \Phi_r(\hat{\mathbb{S}}^{(\sigma(s_m+j'))}_{m+n})\rangle
\end{aligned}$$

Remark that this is the analog of (31) when one replaces $\langle(\hat{\mathbb{S}}^{(\sigma(i))}_{m+n} - \hat{\mathbb{S}}^{(\sigma(s_m+j))}_{m+n})\mathbf{k}, (\hat{\mathbb{S}}^{(\sigma(i'))}_{m+n} - \hat{\mathbb{S}}^{(\sigma(s_m+j'))}_{m+n})\mathbf{k}\rangle_{\mathbf{k}}$ by $\langle \Phi_r(\hat{\mathbb{S}}^{(\sigma(i))}_{m+n}) - \Phi_r(\hat{\mathbb{S}}^{(\sigma(s_m+j))}_{m+n}), \Phi_r(\hat{\mathbb{S}}^{(\sigma(i'))}_{m+n}) - \Phi_r(\hat{\mathbb{S}}^{(\sigma(s_m+j'))}_{m+n})\rangle$. Following the analogy and using the construction from Prop. 3, we will apply the Hanson-Wright inequality (Lem. 8) to the matrix A defined as

$$A_{i,i'} = \begin{cases}
0 & \text{if } i = i' \\
\frac{\langle \Phi_r(\hat{\mathbb{P}}^{\tilde{\sigma}(i)}_m) - \Phi_r(\hat{\mathbb{Q}}^{\tilde{\sigma}(s_m+l_i)-s_m}_n), \Phi_r(\hat{\mathbb{P}}^{\tilde{\sigma}(i')}_m) - \Phi_r(\hat{\mathbb{Q}}^{\tilde{\sigma}(s_m+l_{i'})-s_m}_n)\rangle - \mathrm{MMD}^2_{\mathbf{k}}(\mathbb{P},\mathbb{Q})}{s_m(s_m-1)} & \text{if } i \neq i' \in \rho_\sigma^{-1}(\{\pm 1\}), \\
\frac{1}{s_m(s_m-1)}\langle \Phi_r(\hat{\mathbb{P}}^{\tilde{\sigma}(i)}_m) - \Phi_r(\hat{\mathbb{Q}}^{\tilde{\sigma}(s_m+i)-s_m}_n), \Phi_r(\hat{\mathbb{S}}^{\tilde{\sigma}(i')}_{m+n}) - \Phi_r(\hat{\mathbb{S}}^{\tilde{\sigma}(s_m+l_{i'})}_{m+n})\rangle & \text{if } i \in \rho_\sigma^{-1}(\{\pm 1\}), i' \in \rho_\sigma^{-1}(\{0\}), \\
\frac{1}{s_m(s_m-1)}\langle \Phi_r(\hat{\mathbb{S}}^{\tilde{\sigma}(i)}_{m+n}) - \Phi_r(\hat{\mathbb{S}}^{\tilde{\sigma}(s_m+l_i)}_{m+n}), \Phi_r(\hat{\mathbb{P}}^{\tilde{\sigma}(i')}_m) - \Phi_r(\hat{\mathbb{Q}}^{\tilde{\sigma}(s_m+l_{i'})-s_m}_n)\rangle & \text{if } i \in \rho_\sigma^{-1}(\{0\}), i' \in \rho_\sigma^{-1}(\{\pm 1\}), \\
\frac{1}{s_m(s_m-1)}\langle \Phi_r(\hat{\mathbb{S}}^{\tilde{\sigma}(i)}_{m+n}) - \Phi_r(\hat{\mathbb{S}}^{\tilde{\sigma}(s_m+l_i)}_{m+n}), \Phi_r(\hat{\mathbb{S}}^{\tilde{\sigma}(l_{i'})}_{m+n}) - \Phi_r(\hat{\mathbb{S}}^{\tilde{\sigma}(s_m+l_{i'})}_{m+n})\rangle & \text{if } i \neq i' \in \rho_\sigma^{-1}(\{0\}).
\end{cases} \tag{62}$$

which is the analog of the matrix defined in (38). Note that

$$
\begin{aligned}
&|\langle \Phi_r(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)}) - \Phi_r(\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m}), \Phi_r(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')}) - \Phi_r(\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m}) \rangle - \mathrm{MMD}_{\mathbf{k}}^2(\mathbb{P},\mathbb{Q})| \\
&\leq |\langle \Phi_r(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)}) - \Phi_r(\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m}), \Phi_r(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')}) - \Phi_r(\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m}) \rangle \\
&\qquad - \langle (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m})\mathbf{k}, (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+i')-s_m})\mathbf{k} \rangle_{\mathbf{k}}| \\
&\qquad + |\langle (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m})\mathbf{k}, (\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')} - \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+i')-s_m})\mathbf{k} \rangle_{\mathbf{k}} - \langle (\mathbb{P}-\mathbb{Q})\mathbf{k}, (\mathbb{P}-\mathbb{Q})\mathbf{k} \rangle_{\mathbf{k}}|.
\end{aligned} \tag{63}
$$

The last term in the right-hand side is formally the same as the one upper-bounded in (43). However, there is a difference: in this case $\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)}$ is not the empirical distribution of the output of KT-COMPRESS on the points of $\mathbb{X}_m^{(\tilde{\sigma}(i))}$, but rather the distribution corresponding to the concatenation of the outputs of KT-COMPRESS on the s_r/s compression bins contained in $\mathbb{X}_m^{(\tilde{\sigma}(i))}$. As a result, an adaptation of the argument yields

$$
\begin{aligned}
&|\langle \Phi_r(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)}) - \Phi_r(\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m}), \Phi_r(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')}) - \Phi_r(\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m}) \rangle - \mathrm{MMD}_{\mathbf{k}}^2(\mathbb{P},\mathbb{Q})| \\
&\leq \mathrm{MMD}(\mathbb{P},\mathbb{Q}) W(m,n,\delta/(5s_r)) + W(m,n,\delta/(5s_r))^2,
\end{aligned}
$$

where the function W defined in (61) is slightly different from the one in Prop. 3 in that the arguments of the error inflation factors are $m/s_{m,r}$ and $n/s_{n,r}$ instead of m/s_m and n/s_n.

We can bound the first term of (63) by

$$
\begin{aligned}
&|\langle \Phi_r(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)}), \Phi_r(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')}) \rangle - \langle \hat{\mathbb{P}}_m^{\tilde{\sigma}(i)}\mathbf{k}, \hat{\mathbb{P}}_m^{\tilde{\sigma}(i')}\mathbf{k} \rangle_{\mathbf{k}}| + |\langle \Phi_r(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i)}), \Phi_r(\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m}) \rangle - \langle \hat{\mathbb{P}}_m^{\tilde{\sigma}(i)}\mathbf{k}, \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m}\mathbf{k} \rangle_{\mathbf{k}}| \\
&+ |\langle \Phi_r(\hat{\mathbb{P}}_m^{\tilde{\sigma}(i')}), \Phi_r(\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m}) \rangle - \langle \hat{\mathbb{P}}_m^{\tilde{\sigma}(i')}\mathbf{k}, \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m}\mathbf{k} \rangle_{\mathbf{k}}| \\
&+ |\langle \Phi_r(\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m}), \Phi_r(\hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m}) \rangle - \langle \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_i)-s_m}\mathbf{k}, \hat{\mathbb{Q}}_n^{\tilde{\sigma}(s_m+l_{i'})-s_m}\mathbf{k} \rangle_{\mathbf{k}}| \\
&\leq 4 \sup_{x,x' \in \mathrm{supp}(\mathbb{P}_m) \cup \mathrm{supp}(\mathbb{Q}_n)} |\langle \Phi_r(x), \Phi_r(x') \rangle - \mathbf{k}(x,x')|.
\end{aligned}
$$

The bound in the right-hand side follows from applying Lem. 14. The other cases in (62) admit similar upper-bounds which in this case rely on (44). In analogy with (46), we obtain that

$$
\begin{aligned}
\|A\|_{\mathrm{op}}^2 \leq \|A\|_{\mathrm{F}}^2 \leq \tfrac{1}{s_m^2(s_m-1)^2} \big(&|\rho_\sigma^{-1}(\{-1,1\})|(|\rho_\sigma^{-1}(\{-1,1\})|-1)\big(2\xi^2(\mathbb{P}_m,\mathbb{Q}_n) \\
&\qquad + 2(\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(5s_r)) + W(m,n,\delta/(5s_r))^2)^2\big) \\
&+ |\rho_\sigma^{-1}(\{-1,1\})||\rho_\sigma^{-1}(\{0\})|(2\xi^2(\mathbb{P}_m,\mathbb{Q}_n) \\
&\qquad + 2(\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(5s_r)) + \tfrac{1}{2}W(m,n,\delta/(5s_r))^2)^2) \\
&+ |\rho_\sigma^{-1}(\{0\})|(|\rho_\sigma^{-1}(\{0\})|-1)W(m,n,\delta/(5s_r))^4\big) \\
\leq \eta(m,n,\delta)^2 \triangleq \tfrac{1}{s_m(s_m-1)} \big(&2\epsilon_{\Phi_r}^4(\mathbb{P}_m,\mathbb{Q}_n) + 2(\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(5s_r)) + W(m,n,\delta/(5s_r))^2)^2\big)
\end{aligned}
$$

To prove this inequality, we used that $(a+b)^2 \leq 2a^2 + 2b^2$ for any $a, b \geq 0$.

Mirroring the proof of Lem. 7, we establish that conditioned on the event $\mathcal{A}$ defined in Lem. 9, which has probability at least $1-\delta$, we have that with probability at least $1-\delta'-\delta''$ on the choice of σ,

$$
\begin{aligned}
\mathbb{E}_L[\mathcal{M}^{\sigma,L}] &\leq \tfrac{2}{s_m-1}\log(\tfrac{2}{\delta''})\,\mathrm{MMD}^2(\mathbb{P},\mathbb{Q}) \\
&\quad + \tfrac{c'\log(1/\delta')}{\sqrt{s_m(s_m-1)}}(2\epsilon_{\Phi_r}^4(\mathbb{P}_m,\mathbb{Q}_n) + 2(\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(5s_r)) + W(m,n,\delta/(5s_r))^2)^2)^{1/2}
\end{aligned}
$$

and

$$
\begin{aligned}
&\tfrac{1}{s_m^2 s_n^2}\big(\textstyle\sum_{i=1}^{s_m} \sum_{j \neq j' \in \{1,\dots,s_n\}} \langle (\hat{\mathbb{S}}_{m+n}^{i} - \hat{\mathbb{S}}_{m+n}^{\sigma(s_m+j)})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{i} - \hat{\mathbb{S}}_{m+n}^{\sigma(s_m+j')})\mathbf{k} \rangle_{\mathbf{k}} \\
&+ \textstyle\sum_{j=1}^{s_n} \sum_{i \neq i' \in \{1,\dots,s_m\}} \langle (\hat{\mathbb{S}}_{m+n}^{\sigma(i)} - \hat{\mathbb{S}}_{m+n}^{\sigma(s_m+j)})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\sigma(i')} - \hat{\mathbb{S}}_{m+n}^{\sigma(s_m+j)})\mathbf{k} \rangle_{\mathbf{k}} \\
&+ \textstyle\sum_{i=1}^{s_m} \sum_{j=1}^{s_n} \langle (\hat{\mathbb{S}}_{m+n}^{\sigma(i)} - \hat{\mathbb{S}}_{m+n}^{\sigma(s_m+j)})\mathbf{k}, (\hat{\mathbb{S}}_{m+n}^{\sigma(i)} - \hat{\mathbb{S}}_{m+n}^{\sigma(s_m+j)})\mathbf{k} \rangle_{\mathbf{k}}\big) \\
&\leq \tfrac{s-1}{s_m s_n}((2\epsilon_{\Phi_r}^4(\mathbb{P}_m,\mathbb{Q}_n) + 2(\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(5s_r)) + W(m,n,\delta/(5s_r))^2)^2)^{1/2} + \mathrm{MMD}^2(\mathbb{P},\mathbb{Q}))
\end{aligned}
$$

simultaneously, where c' is a universal constant and $W(m,n,\delta)$ is as defined in (33).

Mirroring the final step of the proof of Prop. 3, we rely on these two equations to show that conditioned on the event $\mathcal{A}$, with probability at least $1 - \delta' - \delta''$,

$$\begin{aligned}
\mathrm{MMD}^2_{\Phi_r}(\hat{\mathbb{X}}^\sigma_m, \hat{\mathbb{Y}}^\sigma_n) \leq \tfrac{(s_m-1)(s_n-1)}{s_m s_n}\Big(&\tfrac{2}{s_m-1}\log(\tfrac{2}{\delta''})\,\mathrm{MMD}^2(\mathbb{P},\mathbb{Q}) \\
&+ \tfrac{c'\log(1/\delta')}{\sqrt{s_m(s_m-1)}}(2\epsilon^4_{\Phi_r}(\mathbb{P}_m,\mathbb{Q}_n) + 2(\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(5s_r)) + W(m,n,\delta/(5s_r))^2)^2)^{1/2}\Big) \\
&+ \tfrac{s-1}{s_m s_n}((2\epsilon^4_{\Phi_r}(\mathbb{P}_m,\mathbb{Q}_n) + 2(\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(5s_r)) + W(m,n,\delta/(5s_r))^2)^2)^{1/2} + \mathrm{MMD}^2(\mathbb{P},\mathbb{Q})),
\end{aligned}$$

which concludes the proof upon simplification.

□

Lemma 14. *Let $\mathbb{S}$ and $\mathbb{S}'$ be arbitrary distributions. We have that*

$$|\langle \Phi_r(\mathbb{S}), \Phi_r(\mathbb{S}')\rangle - \langle \mathbb{S}\mathbf{k}, \mathbb{S}'\mathbf{k}\rangle_{\mathbf{k}}| \leq \sup_{x\in\mathrm{supp}(\mathbb{S}), x'\in\mathrm{supp}(\mathbb{S}')} |\langle \Phi_r(x), \Phi_r(x')\rangle - \mathbf{k}(x,x')|. \tag{64}$$

Proof. The right-hand side of (64) is equal to

$$\begin{aligned}
&|\mathbb{E}_{x\sim\mathbb{S},x'\sim\mathbb{S}'}\langle \Phi_r(x), \Phi_r(x')\rangle - \mathbb{E}_{x\sim\mathbb{S},x'\sim\mathbb{S}'}\mathbf{k}(x,x')| \leq \mathbb{E}_{x\sim\mathbb{S},x'\sim\mathbb{S}'}|\langle \Phi_r(x), \Phi_r(x')\rangle - \mathbf{k}(x,x')| \\
&\leq \sup_{x\in\mathrm{supp}(\mathbb{S}), x'\in\mathrm{supp}(\mathbb{S}')} |\langle \Phi_r(x), \Phi_r(x')\rangle - \mathbf{k}(x,x')|.
\end{aligned}$$

□

We proceed to prove Thm. 5. Reproducing the argument of Cor. 1, if $b_\alpha \triangleq \lceil(\mathcal{B}+1)(1-\alpha)\rceil \leq \mathcal{B}$, and KT-COMPRESS calls are run with value $\delta^*/(5s_m)$ and $\delta^*/(5s_n)$ respectively, then with probability at least $1 - \frac{\delta}{2}$,

$$\begin{aligned}
M_{(b_\alpha)} &\leq \hat{Z}(m,n,\alpha,\delta) \\
&\triangleq \sqrt{\tfrac{2}{s_m}(\log(\tfrac{2}{\delta^*})+1)}\,\mathrm{MMD}(\mathbb{P},\mathbb{Q}) + \sqrt{\tfrac{1}{s_m}(2 + c'\log(1/\delta^*))}\cdot 2^{1/4}(\epsilon_{\Phi_r}(\mathbb{P}_m,\mathbb{Q}_n) \\
&\qquad + \sqrt{\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\delta/(20s_r))} + W(m,n,\delta/(20s_r))).
\end{aligned} \tag{65}$$

where $M_{(b_\alpha)}$ is the threshold of Alg. 2, and

$$\delta^* \triangleq (\tfrac{\delta}{2})^{1/k_\alpha}\tfrac{\alpha}{4e}, \quad k_\alpha \triangleq \lfloor \alpha(\mathcal{B}+1)\rfloor.$$

Then, we have that the statement in Lem. 11 holds with the redefined $\hat{Z}(m,n,\alpha,\tilde{\beta})$ given by (65). Consequently, the function $\tilde{Z}(m,n,\alpha,\beta)$ now reads

$$\begin{aligned}
\tilde{Z}(m,n,\alpha,\beta) = &\tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P},m/s_{m,r},\tilde{\beta}/6,\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + \tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n/s_{n,r},\tilde{\beta}/6,\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} + c_{\tilde{\beta}/6}\big(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{\|\mathbf{k}\|_\infty}{n}}\big) \\
&+ \sqrt{\tfrac{2}{s_m}\log(\tfrac{8e^2}{\alpha}(\tfrac{4}{\tilde{\beta}})^{1/k_\alpha})}\,\mathrm{MMD}(\mathbb{P},\mathbb{Q}) \\
&+ \sqrt{\tfrac{\sqrt{2}}{s_m}(2 + c'\log(\tfrac{4e}{\alpha}(\tfrac{4}{\tilde{\beta}})^{1/k_\alpha}))}(\epsilon_{\Phi_r}(\mathbb{P}_m,\mathbb{Q}_n) + \sqrt{\mathrm{MMD}(\mathbb{P},\mathbb{Q})W(m,n,\tilde{\beta}/(20s_r))} \\
&\qquad + W(m,n,\tilde{\beta}/(20s_r))).
\end{aligned}$$

We also reduce the remainder of the proof to a second-degree inequality analogous to (51), but in this case the coefficients read

$$\begin{aligned}
a &= 1 - \sqrt{\tfrac{2}{s_m}\log(\tfrac{8e^2}{\alpha}(\tfrac{4}{\tilde{\beta}})^{1/k_\alpha})}, \\
b &= \sqrt{\tfrac{\sqrt{2}}{s_m}(2 + c'\log(\tfrac{4e}{\alpha}(\tfrac{2}{\tilde{\beta}})^{1/k_\alpha}))W(m,n,\tilde{\beta}/(20s_r))}, \\
c &= \tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{P},m/s_{m,r},\tilde{\beta}/6,\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + \tfrac{\mathbf{R}_{\mathbf{k},\mathbf{k}'}(\mathbb{Q},n/s_{n,r},\tilde{\beta}/6,\mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} + c_{\tilde{\beta}/6}\big(\sqrt{\tfrac{\|\mathbf{k}\|_\infty}{m}} + \sqrt{\tfrac{\|\mathbf{k}\|_\infty}{n}}\big) \\
&\quad + \sqrt{\tfrac{\sqrt{2}}{s_m}(2 + c'\log(\tfrac{4e}{\alpha}(\tfrac{4}{\tilde{\beta}})^{1/k_\alpha}))}(\epsilon_{\Phi_r}(\mathbb{P}_m,\mathbb{Q}_n) + W(m,n,\tilde{\beta}/(20s_r))), \\
x &= \sqrt{\mathrm{MMD}(\mathbb{P},\mathbb{Q})},
\end{aligned}$$

Proceeding just like in the case of CTT, the proof is concluded (recall that the function W is slightly different in this case).

G Proof of Thm. 3: ACTT validity and power

Algorithm 6: Aggregated CTT, Δ_{ACTT}

Input: Samples $(\mathbb{X}_m, \mathbb{Y}_n)$, # coresets s, compression level $\mathfrak{g}$, kernels $(\mathbf{k}_\lambda, \mathbf{k}'_\lambda)_{\lambda\in\Lambda}$, importance weights $(w_\lambda)_{\lambda\in\Lambda}$, failure prob. δ, # replicates $(\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3)$, level α

Partition $\mathbb{X}_m$ into $s_m = \frac{sm}{m+n}$ equal-sized bins $(\mathbb{X}_m^{(i)})_{i=1}^{s_m}$
Partition $\mathbb{Y}_n$ into $s_n = \frac{sn}{m+n}$ equal-sized bins $(\mathbb{Y}_n^{(i)})_{i=1}^{s_n}$

// Identify coresets of size $2^{\mathfrak{g}}\sqrt{\frac{m+n}{s}}$ using sum of kernels
$\mathbf{k} \leftarrow \sum_{\lambda\in\Lambda} \mathbf{k}_\lambda$; $\quad \mathbf{k}' \leftarrow \sum_{\lambda\in\Lambda} \mathbf{k}'_\lambda$
for $i = 1, \ldots, s_m$ **do**
  $\hat{\mathbb{X}}_m^{(i)} \leftarrow$ KT-COMPRESS$(\mathbb{X}_m^{(i)}, \mathfrak{g}, \mathbf{k}, \mathbf{k}', \delta)$
end
for $i = 1, \ldots, s_n$ **do**
  $\hat{\mathbb{Y}}_n^{(i)} \leftarrow$ KT-COMPRESS$(\mathbb{Y}_n^{(i)}, \mathfrak{g}, \mathbf{k}, \mathbf{k}', \delta)$
end
// Compute CORESETMMD for each candidate parameter λ
$\hat{\mathbb{X}}_m \coloneqq$ CONCAT$((\hat{\mathbb{X}}_m^{(i)})_{i=1}^{s_m})$; $\quad \hat{\mathbb{Y}}_n \coloneqq$ CONCAT$((\hat{\mathbb{Y}}_n^{(i)})_{i=1}^{s_n})$
for $\lambda \in \Lambda$ **do** $M_\lambda \leftarrow \mathrm{MMD}_{\mathbf{k}_\lambda}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)$
// Simulate null for each λ by randomly permuting s coresets
for $\ell = 1, 2$ **and** $b = 1, \ldots, \mathcal{B}_\ell$ **do**
  $(\hat{\mathbb{X}}_m^{\ell,b}, \hat{\mathbb{Y}}_n^{\ell,b}) \leftarrow$ PERMUTECORESETS$(\mathbb{X}_m, \mathbb{Y}_n, s)$
  for $\lambda \in \Lambda$ **do** $M_{b,\lambda,\ell} \leftarrow \mathrm{MMD}_{\mathbf{k}_\lambda}(\hat{\mathbb{X}}_m^b, \hat{\mathbb{Y}}_n^b)$
end
for $\lambda \in \Lambda$ **do** Sort $(M_{b,\lambda,1})_{b=1}^{\mathcal{B}_1}$ increasingly into $(M_{(b),\lambda,1})_{b=1}^{\mathcal{B}_1}$
// Estimate largest rejection threshold for each M_λ statistic that ensures aggregated test size $\leq \alpha$
$u_{\min} \leftarrow 0$ **and** $u_{\max} \leftarrow \min_{\lambda\in\Lambda} w_\lambda^{-1}$
for $i = 1, \ldots, \mathcal{B}_3$ **do**
  $u \leftarrow \frac{u_{\min}+u_{\max}}{2}$; **for** $\lambda \in \Lambda$ **do** $b_{u,\lambda} \leftarrow \lceil(\mathcal{B}_1+1)(1-uw_\lambda)\rceil$
  $P_u \leftarrow \frac{1}{\mathcal{B}_2}\sum_{b=1}^{\mathcal{B}_2} \mathbf{1}[\max_{\lambda\in\Lambda}(M_{b,\lambda,2} - M_{(b_{u,\lambda}),\lambda,1}) > 0]$
  if $P_u \leq \alpha$ **then** $u_{\min} \leftarrow u$ **else** $u_{\max} \leftarrow u$
end
// Reject null if any test statistic M_λ exceeds its threshold
$\hat{u}_\alpha \leftarrow u_{\min}$; $\quad$ **for** $\lambda \in \Lambda$ **do** $b'_{\alpha,\lambda} \leftarrow \lceil(\mathcal{B}_1+1)(1-\hat{u}_\alpha w_\lambda)\rceil$
if $M_\lambda > M_{(b'_{\alpha,\lambda}),\lambda,1}$ **for some** $\lambda \in \Lambda$ **then return** 1 (reject null)
else return 0 (accept null)

The validity statement in (10) follows from exactly the same argument as Schrab et al. (2021, Prop. 8), replacing the estimate $\mathrm{MMD}_{up}(\mathbb{X}_m, \mathbb{Y}_n)$ with parameter λ by CORESETMMD$(\mathbb{X}_m, \mathbb{Y}_n)$ with parameter λ.

Let $\Delta_{\text{CTT},\lambda}$ denote the output of a modified CTT (Alg. 1) with level $\alpha w_\lambda/2$, $\mathcal{B} = \mathcal{B}_1$, $\mathbf{k} = \sum_{\lambda\in\Lambda} \mathbf{k}_\lambda$ and $\mathbf{k}' = \sum_{\lambda\in\Lambda} \mathbf{k}'_\lambda$ that uses $\mathbf{k}_\lambda$ (in place of $\mathbf{k}$) to compute CORESETMMD. Then using arguments from Schrab et al. (2021, Proof of Thm. 9, up to their equation 25), we find that

$$\Pr[\Delta_{\text{ACTT}}(\mathbb{X}_m, \mathbb{Y}_n) = 1] \geq \max_{\lambda\in\Lambda} \Pr[\Delta_{\text{CTT},\lambda}(\mathbb{X}_m,\mathbb{Y}_n)=1] - \tfrac{\beta}{2}. \tag{66}$$

We claim that for λ such that (12) holds, we have $\Pr[\Delta_{\text{CTT},\lambda}(\mathbb{X}_m, \mathbb{Y}_n) = 1] \geq 1 - \frac{\beta}{2}$, which when put together with (66) immediately implies the claimed power in (11). We now establish our power claim for this modified CTT test.

To do so, we claim that KT-COMPRESS with $\mathbf{k}$ and $\mathbf{k}' = \sum_{\lambda\in\Lambda} \mathbf{k}'_\lambda$—referred to as KT-COMPRESS-AGG—is $\mathbf{k}_\lambda$-sub-Gaussian (Shetty et al., 2022, Def. 3) with parameters $a''_{\ell,n}$ and $v''_{\ell,n}$ (the analog of $(a_{\ell,n}, v_{\ell,n})$ from (14) in our notation) simultaneously for all $\lambda \in \Lambda$, on an event of probability $1 - \delta/2$, where

$$a''_{\ell,n} = \tfrac{4}{\ell}\sqrt{\textstyle\sum_{\lambda\in\Lambda}\Big(C^2_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathcal{S}_{\text{in}}) + \mathfrak{M}^2_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathcal{S}_{\text{in}}, \delta, \ell) \cdot \log|\Lambda|\Big)}, \tag{67}$$

$$v''_{\ell,n} = \tfrac{2}{\ell}\sqrt{\log(\tfrac{12n4^{\mathfrak{g}}(\beta_n+1)}{\ell\delta}) \cdot \textstyle\sum_{\lambda\in\Lambda} \mathfrak{M}^2_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathcal{S}_{\text{in}}, \delta, \ell)},$$

and $C_{\mathbf{k},\mathbf{k}'}$ and $\mathfrak{M}_{\mathbf{k},\mathbf{k}'}$ were defined in (14). Deferring the proof of this claim to the end of this section, we proceed with the proof.

Using (67) and repeating the arguments from the proof of Lem. 1 (after (14)), we conclude the following analog of (6) for the CORESETMMD estimate with the output $\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n$ of KT-COMPRESS-AGG: With probability at least $1-\delta$, we have

$$\begin{aligned}
&|\operatorname{MMD}_{\mathbf{k}_\lambda}(\mathbb{X}_m, \mathbb{Y}_n) - \operatorname{MMD}_{\mathbf{k}_\lambda}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)|^2 \\
&\leq 1024(\sqrt{\log(m+n+1)} + \sqrt{\log(2/\delta)})^2 \\
&\quad \cdot |\Lambda| \Bigg[\tfrac{(\log_4(m/s_m) - \mathfrak{g} - 1)}{4^{\mathfrak{g}} m} \Big(C_{\Lambda,\mathbf{k}'}(\mathbb{X}_m) + \Big(\sqrt{\log|\Lambda|} + \sqrt{\log(\tfrac{3m(\log_4(m/s_m) - \mathfrak{g} - 1)}{s_m \delta})} \Big) \mathfrak{M}'_{\Lambda,\mathbf{k}'}(\mathbb{X}_m, \delta, 2^{\mathfrak{g}+1}\sqrt{m/s_m}) \Big)^2 \\
&\qquad + \tfrac{(\log_4(n/s_n) - \mathfrak{g} - 1)}{4^{\mathfrak{g}} n} \Big(C_{\Lambda,\mathbf{k}'}(\mathbb{Y}_n) + \Big(\sqrt{\log|\Lambda|} + \sqrt{\log(\tfrac{3n(\log_4(n/s_n) - \mathfrak{g} - 1)}{s_n \delta})} \Big) \mathfrak{M}'_{\Lambda,\mathbf{k}'}(\mathbb{Y}_n, \delta, 2^{\mathfrak{g}+1}\sqrt{n/s_n}) \Big)^2 \Bigg],
\end{aligned} \tag{68}$$

where $C_{\Lambda,\mathbf{k}'}(\mathcal{S}_{\mathrm{in}}) \triangleq \max_{\lambda\in\Lambda} C_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathcal{S}_{\mathrm{in}})$ and $\mathfrak{M}'_{\Lambda,\mathbf{k}'}(\mathcal{S}_{\mathrm{in}}, \delta, \ell) \triangleq \max_{\lambda\in\Lambda} \mathfrak{M}_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathcal{S}_{\mathrm{in}}, \delta, \ell)$. Putting (68) together with the definitions (13), we find that

$$|\operatorname{MMD}_{\mathbf{k}_\lambda}(\mathbb{X}_m, \mathbb{Y}_n) - \operatorname{MMD}_{\mathbf{k}_\lambda}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)| \leq \underbrace{2\sqrt{|\Lambda|(1+\log(|\Lambda|))}}_{=c_\Lambda} \cdot \max_{\lambda\in\Lambda} \Big(\tfrac{\mathbf{R}_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathbb{X}_m, \frac{m}{s_m}, \delta, \mathfrak{g})}{4^{\mathfrak{g}} m} + \tfrac{\mathbf{R}_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathbb{Y}_n, \frac{n}{s_n}, \delta, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} \Big).$$

with probability at least $1-\delta$. Propagating this result further in the proof of Lem. 1 implies the following analog of (7):

$$\begin{aligned}
|\operatorname{MMD}_{\mathbf{k}_\lambda}(\mathbb{P}, \mathbb{Q}) - \operatorname{MMD}_{\mathbf{k}_\lambda}(\hat{\mathbb{X}}_m, \hat{\mathbb{Y}}_n)| &\leq c_\delta \big(\sqrt{\tfrac{\|\mathbf{k}_\lambda\|_\infty}{m}} + \sqrt{\tfrac{\|\mathbf{k}_\lambda\|_\infty}{n}} \big) \\
&\quad + c_\Lambda \max_{\lambda\in\Lambda} \Big(\tfrac{\mathbf{R}_{\mathbf{k}'_\lambda}(\mathbb{P}, \frac{m}{s_m}, \delta, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{m}} + \tfrac{\mathbf{R}_{\mathbf{k}'_\lambda}(\mathbb{Q}, \frac{n}{s_n}, \delta, \mathfrak{g})}{2^{\mathfrak{g}}\sqrt{n}} \Big),
\end{aligned} \tag{69}$$

with probability at least $1-3\delta$ with $c_\delta \triangleq 2+\sqrt{2\log(\frac{2}{\delta})}$ as in Lem. 1.

We now apply Thm. 1 to characterize the power of the modified CTT (corresponding to $\Delta_{\mathrm{CTT},\lambda}$) described above. In particular, substituting $\alpha \leftarrow \frac{\alpha w_\lambda}{2}$, $\beta \leftarrow \frac{\beta}{2}$, $\gamma \leftarrow \gamma_\lambda$ in Thm. 1, noting $s_m \geq \frac{32}{9}\log(\frac{2e}{\gamma_\lambda})$, and using the definition of $\varepsilon_{\mathrm{AGG}}$ along with (69) in the proof of Thm. 1, we conclude that

$$\Pr[\Delta_{\mathrm{CTT},\lambda}(\mathbb{X}_m, \mathbb{Y}_n)] \geq 1 - \tfrac{\beta}{2} \quad \text{whenever} \quad \operatorname{MMD}_{\mathbf{k}_\lambda}(\mathbb{P}, \mathbb{Q}) \geq c' \sqrt{\log(\tfrac{1}{\gamma_\lambda})}\, \varepsilon_{\mathrm{AGG}}(\tfrac{\beta/(10s)}{4+\beta})$$

for some universal constant c', and yielding the desired claim when $m \leq n$. It remains to prove our earlier claim (67).

Proof of (67) Note that, for any probability measures $(\mathbb{P}', \mathbb{Q}')$,

$$\operatorname{MMD}_{\mathbf{k}_\lambda}(\mathbb{P}', \mathbb{Q}') \leq \operatorname{MMD}_{\mathbf{k}}(\mathbb{P}', \mathbb{Q}') \tag{70}$$

whenever the right-hand side is well-defined, since for any two kernels $\mathbf{k}_1, \mathbf{k}_2$ with well-defined $\operatorname{MMD}_{\mathbf{k}_1+\mathbf{k}_2}(\mathbb{P}', \mathbb{Q}')$, we have

$$\operatorname{MMD}^2_{\mathbf{k}_1+\mathbf{k}_2}(\mathbb{P}', \mathbb{Q}') = (\mathbb{P}' - \mathbb{Q}')(\mathbf{k}_1 + \mathbf{k}_2)(\mathbb{P}' - \mathbb{Q}') = \operatorname{MMD}^2_{\mathbf{k}_1}(\mathbb{P}', \mathbb{Q}') + \operatorname{MMD}^2_{\mathbf{k}_2}(\mathbb{P}', \mathbb{Q}'). \tag{71}$$

In the terminology of Shetty et al. (2022, Def. 3), we next establish that the halving algorithm KT(δ) (Shetty et al., 2022, Ex. 2) underlying KT-COMPRESS is $\mathbf{k}_\lambda$-sub-Gaussian when run with $\mathbf{k}$ and split kernel $\mathbf{k}'$. To proceed, we can suitably adapt the proof of Thm. 4 of Dwivedi and Mackey (2022) (which in turn is an adaptation of Dwivedi and Mackey (2021, Thm. 2-4)).

We begin by instantiating the notation of Dwivedi and Mackey (2022). Given an input coreset $\mathcal{S}_{\mathrm{in}}$, let $\mathcal{S}_{\mathrm{split},1}$ denote the first coreset output by the KT-SPLIT step and $\mathcal{S}_{\mathrm{out}}$ denote the output of size n_{out} after the KT-SWAP step. Then using (70) and the definition of KT-SWAP, we have

$$\operatorname{MMD}^2_{\mathbf{k}_\lambda}(\mathcal{S}_{\mathrm{in}}, \mathcal{S}_{\mathrm{out}}) \leq \operatorname{MMD}^2_{\mathbf{k}}(\mathcal{S}_{\mathrm{in}}, \mathcal{S}_{\mathrm{out}}) \overset{(i)}{\leq} \operatorname{MMD}^2_{\mathbf{k}}(\mathcal{S}_{\mathrm{in}}, \mathcal{S}_{\mathrm{split},1}) \overset{(71)}{=} \textstyle\sum_{\lambda\in\Lambda} \operatorname{MMD}^2_{\mathbf{k}_\lambda}(\mathcal{S}_{\mathrm{in}}, \mathcal{S}_{\mathrm{split},1}), \tag{72}$$

where the inequality (i) follows directly from the definition of KT-SWAP (Dwivedi and Mackey, 2022, Eqn. 27). Hence it remains to show that KT-SPLIT(δ) is $\mathbf{k}_\lambda$-sub-Gaussian for each λ.

To proceed, we modify the Dwivedi and Mackey (2022, Proof of Thm. 4). In particular, replacing $\mathbf{k}^\dagger$ (in their notation) with $\mathbf{k}'$, and $\|\mathbf{k}^\dagger\|_\infty$ with $\|\mathbf{k}'\|_{\infty,\text{in}}$ throughout their proof[3] we conclude, with analogy to Shetty et al. (2022, Ex. 2), that KT-SPLIT(δ) is $\mathbf{k}_\lambda$-sub-Gaussian with parameters $v_{\lambda,\ell}$ and $a_{\lambda,\ell}$ satisfying

$$a_{\lambda,\ell} = \frac{C_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathcal{S}_{\text{in}})}{n_{\text{out}}} \quad \text{and} \quad v_{\lambda,\ell} = \frac{\mathfrak{M}_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathcal{S}_{\text{in}},\delta,\ell)}{n_{\text{out}}}\sqrt{\log(\tfrac{6n_{\text{out}}\log_2(\ell/n_{\text{out}})}{\delta})}, \tag{73}$$

for input (a subset of $\mathcal{S}_{\text{in}}$) of size ℓ and output of size n_{out}, for $C_{\mathbf{k}_\lambda,\mathbf{k}'}$ and $\mathfrak{M}_{\mathbf{k}_\lambda,\mathbf{k}'}$ defined in (14) (also see Rem. 6).

Next, we use an auxiliary result proven at the end of this section.

Lemma 15 (Tail bounds for sum of non-centered sub-Gaussian random variables)**.** *Consider non-negative random variables $Z_1,\dots,Z_\ell$ such that for $i\in[\ell]$, we have $\mathbb{P}\big[Z_i \geq a_i + v_i\sqrt{t}\big] \leq e^{-t}$ for all $t\geq 0$, with some suitable scalars $\{a_i,v_i\}_{i=1}^\ell$. Then $\mathbb{P}\Big[\sqrt{\sum_{i=1}^\ell Z_i^2} \geq \widetilde{\alpha} + \widetilde{\beta}\sqrt{t}\Big] \leq e^{-t}$ for $t\geq 0$, where*

$$\widetilde{\alpha}^2 \triangleq 2\textstyle\sum_{i=1}^\ell (a_i^2 + v_i^2\log\ell) \leq 4\ell\log\ell\cdot\max_i\max\{a_i^2,v_i^2\} \quad \textit{and} \quad \widetilde{\beta}^2 \triangleq \sum_{i=1}^\ell v_i^2 \leq \ell\max_i v_i^2.$$

Putting Lem. 15 together with (72) and (73), we conclude that on an event of probability at least $1-\delta/2$ and simultaneously for all $\lambda\in\Lambda$, KT-COMPRESS-AGG (KT(δ) with aggregated kernels as above) is $\mathbf{k}_\lambda$-sub-Gaussian with parameters (a'_ℓ, v'_ℓ) given by

$$a'_\ell = \tfrac{1}{n_{\text{out}}}\sqrt{2\textstyle\sum_{\lambda\in\Lambda}\Big(C^2_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathcal{S}_{\text{in}}) + \mathfrak{M}^2_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathcal{S}_{\text{in}},\delta,\ell)\cdot\log|\Lambda|\Big)} \quad \text{and}$$
$$v'_\ell = \tfrac{1}{n_{\text{out}}}\sqrt{\log(\tfrac{6n_{\text{out}}\log_2(\ell/n_{\text{out}})}{\delta})}\cdot\sqrt{\textstyle\sum_{\lambda\in\Lambda}\mathfrak{M}^2_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathcal{S}_{\text{in}},\delta,\ell)}.$$

for input (a subset of $\mathcal{S}_{\text{in}}$) of size ℓ and output of size n_{out}. Now the arguments of Shetty et al. (2022, Ex. 4) imply that on an event of probability at least $1-\delta/2$, every HALVE call invoked by COMPRESS (for KT-COMPRESS-AGG) is $\mathbf{k}_\lambda$-sub-Gaussian with parameters $a''_{\ell,n}$ and $v''_{\ell,n}$ (the analog of $(a_{\ell,n}, v_{\ell,n})$ in our notation (14)), where

$$a''_{\ell,n} = \tfrac{2\sqrt{2}}{\ell}\sqrt{2\textstyle\sum_{\lambda\in\Lambda}\Big(C^2_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathcal{S}_{\text{in}}) + \mathfrak{M}^2_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathcal{S}_{\text{in}},\delta,\ell)\cdot\log|\Lambda|\Big)} \quad \text{and}$$
$$v''_{\ell,n} = \tfrac{2}{\ell}\sqrt{\log(\tfrac{12n4^{\mathfrak{g}}(\beta_n+1)}{\ell\delta})\cdot\textstyle\sum_{\lambda\in\Lambda}\mathfrak{M}^2_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathcal{S}_{\text{in}},\delta,\ell)},$$

as claimed in (67).

Proof of Lem. 15 Collect the scalars $\{a_i\}$ (resp. $\{v_i\}$) into vector $a\in\mathbb{R}^\ell$ (resp. $v\in\mathbb{R}^\ell$) such that the i-th coordinate of a (resp. v) is equal to a_i (resp. v_i). A direct union bound yields that with probability at least $1-\ell e^{-t}$, we have

$$\begin{aligned}\textstyle\sum_{i=1}^\ell Z_i^2 \leq \sum_{i=1}^\ell (a_i + v_i\sqrt{t})^2 = \sum_{i=1}^\ell a_i^2 + v_i^2 t + 2a_iv_i\sqrt{t} &= \|a\|_2^2 + \big\|v\sqrt{t}\big\|_2^2 + 2\langle a, v\sqrt{t}\rangle \\ &\overset{(i)}{\leq} \|a\|_2^2 + \big\|v\sqrt{t}\big\|_2^2 + 2\|a\|_2\big\|v\sqrt{t}\big\|_2 \\ &= (\|a\|_2 + \big\|v\sqrt{t}\big\|_2)^2,\end{aligned}$$

where step (i) follows from Cauchy-Schwarz's inequality. Substituting $t\leftarrow t+\log\ell$, we conclude that

$$\mathbb{P}\Big[\sqrt{\textstyle\sum_{i=1}^\ell Z_i^2} \geq \|a\|_2 + \|v\|_2\sqrt{\log\ell} + \sqrt{t}\|v\|_2\Big] \leq e^{-t}.$$

The lemma now follows once we note that $\widetilde{\beta} = \|v\|_2$ and

$$\widetilde{\alpha}^2 = 2(\|a\|_2^2 + \|v\|_2^2\log\ell) \geq (\|a\|_2 + \|v\|_2\sqrt{\log\ell})^2.$$

[3] The remark in Dwivedi and Mackey (2022, Footnote 5) implies that the arguments work both with $\|\mathbf{k}'\|_\infty$ and $\|\mathbf{k}'\|_{\infty,\text{in}}$.

Remark 6. If the aggregated kernel satisfies $\mathbf{k}' = \sum_{\lambda\in\Lambda} \mathbf{k}_\lambda$ with each $\mathbf{k}_\lambda$ normalized, i.e., $\|\mathbf{k}_\lambda\|_\infty = 1$. In this case, Dwivedi and Mackey (2022, Eq. (23)) shows that for any $\lambda \in \Lambda$ and any f in the RKHS of $\mathbf{k}_\lambda$, we have $\|f\|_{\mathbf{k}_\lambda} \leq \|f\|_{\mathbf{k}'}$. Then, repeating arguments as in Dwivedi and Mackey (2022, App. F, Proof of Thm. 4), we find that

$$C_{\mathbf{k}_\lambda,\mathbf{k}'}(\mathcal{S}_{\mathrm{in}}) = 2\sqrt{\|\mathbf{k}'\|_\infty} = 2\sqrt{|\Lambda|} \quad \text{and} \quad \mathfrak{M}_{\mathbf{k}_\lambda,\mathbf{k}'} = \sqrt{|\Lambda|} \cdot \mathfrak{M}_{\mathbf{k}_\lambda,\mathbf{k}_\lambda}$$

where $\mathfrak{M}_{\mathbf{k},\mathbf{k}}$ is defined in (21).

H Experiment details and supplementary results

Here we provide the details deferred from Sec. 5 along with supplementary results.

Optimal four-point halving As discussed in App. A, we modify the KT-COMPRESS algorithm of Shetty et al. (2022, Ex. 4) slightly so that whenever an input of size 4 is being compressed into an coreset of size 2, we return an optimal coreset of size 2 that minimizes $\mathrm{MMD}_{\mathbf{k}}$ between the input point set and the output. This optimal coreset is also symmetrized so the either the coreset or its complement is returned with equal probability. See Alg. 4.

Details on the code All computations related to kernel and MMD evaluations are written using identical Cython commands to ensure both consistent runtime comparisons across methods and faster runtimes overall. Our code can be easily extended to cover other MMD tests and can be used as a benchmark to assess power-time trade-off curves.

Additional details for CTT experiments on GAUSSIAN and EMNIST

- The bandwidth of the Gaussian kernel is selected according the median heuristic, which is a popular heuristic in kernel methods (Chaudhuri et al., 2017) that prescribes the usage of kernels of the form $k(x, y) = \exp(-\|x-y\|^2/(2\hat{\sigma}^2))$, where $\hat{\sigma}$ is the median of the pairwise distances between different points in the sequence $\mathbb{X}_m \cup \mathbb{Y}_n$. Unless otherwise specified, we used the median heuristic to select all bandwidths in our experiments. Since computing the median among all pairs is expensive, we selected 512 points from $\mathbb{X}_m$ and 512 points from $\mathbb{Y}_n$ uniformly at random and computed the median of all $\binom{1024}{2}$ pairwise distances among them.
- For wild bootstrap block and incomplete tests, we use the fast computation procedure proposed by Schrab et al. (2021, 2022), which computes the terms $\mathbf{h}(X_i, X_j, Y_i, Y_j)$ only once for each pair $i \neq j$. This is the main advantage of the wild bootstrap approach over the permutation approach. The wild bootstrap incomplete test is the same test studied by Schrab et al. (2022).
- Both in Asymp. Block I and II, the threshold is computed via the CLT using an estimate of the variance of the estimator. In Asymp. Block II, the estimate of the variance is obtained from the variance of the n/B block MMD estimates. Asymp. Block II was considered as a baseline by Yamada et al. (2019). In Asymp. Block I, the estimate of the variance is obtained by sampling a Rademacher vector length n and flipping the corresponding elements of $\mathbb{X}_m$, $\mathbb{Y}_n$ to obtain a new pair of sets of n/B blocks of size B, and computing the empirical variance of these n/B block MMD estimates. Since computations of $\mathbf{h}(X_i, X_j, Y_i, Y_j)$ are reused, Asymp. Block I is almost as fast as Asymp. Block II. Asymp. Block I was proposed first chronologically by Zaremba et al. (2013) in the paper that introduced block tests, although they used a permutation instead of a Rademacher variable, which made the method twice as slow.

Additional details for LR-CTT experiments on GAUSSIAN and EMNIST

- The bandwidth selection is as described above.

Additional details for ACTT experiments on BLOBS and HIGGS

- We use the permutation approach and take $\mathcal{B}_1 = 299$ permutations, $\mathcal{B}_2 = 200$ permutations, and $\mathcal{B}_3 = 20$ iterations.
- As suggested by Schrab et al. (2021), the ACTT experiments set Λ as multiples of the bandwidth given by the median heuristic. We computed the median heuristic bandwidth λ_0 as in the CTT experiments, and we set $\Lambda = \{2^{-i}\lambda_0 | i \in \{0, \ldots, 4\}\}$.
- The aggregated wild bootstrap incomplete test is the same test studied by Schrab et al. (2022).

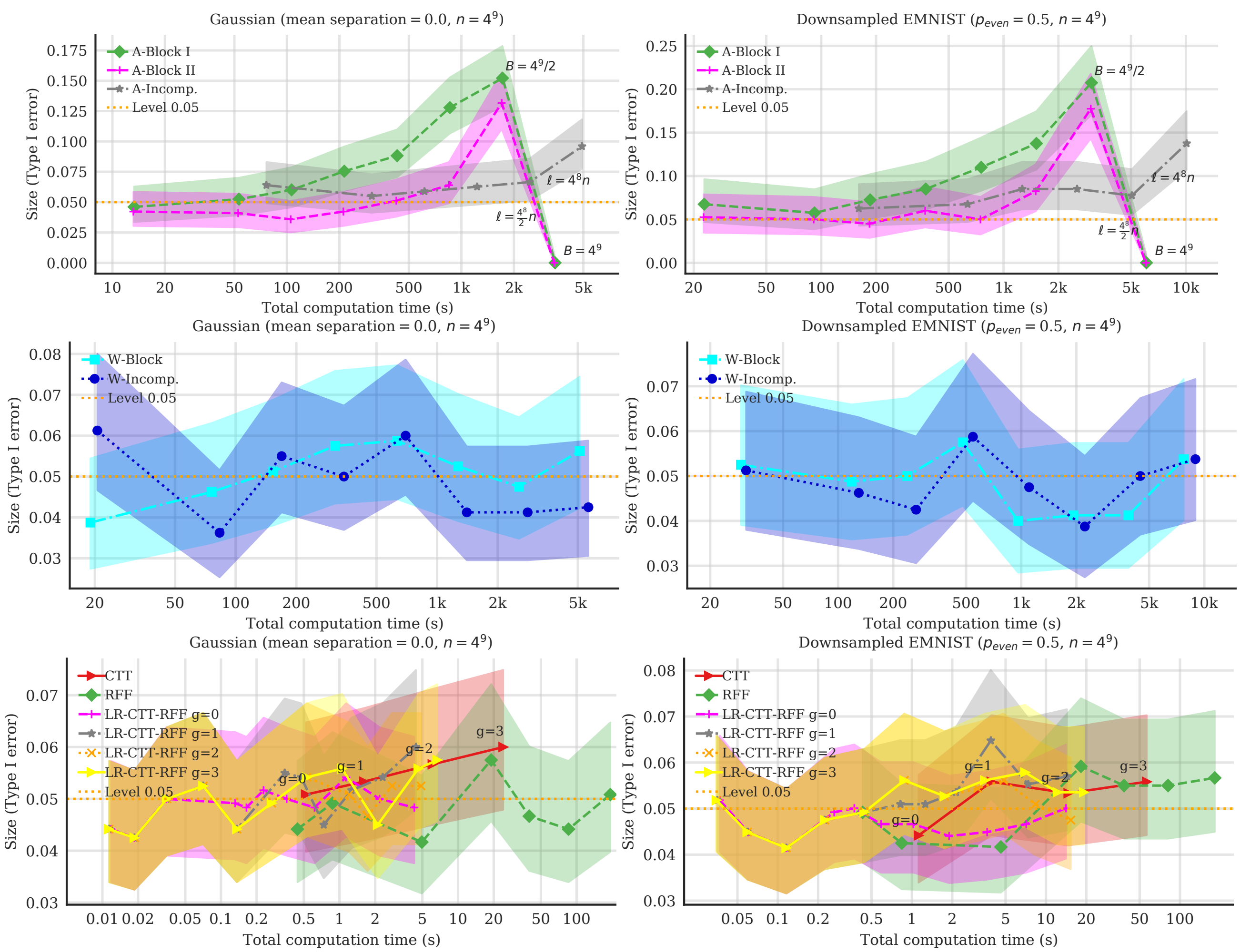

Figure 3: **Estimated test size** with 95% Wilson confidence intervals in the GAUSSIAN (*left*) and EMNIST (*right*) experimental settings of Fig. 1. **Top:** Asymptotic block and incomplete tests with 800 (*left*) and 400 (*right*) independent test repetitions. **Middle:** Non-asymptotic wild bootstrap block and incomplete with 800 independent test repetitions. **Bottom:** Non-asymptotic CTT, RFF, and LR-CTT with 1200 independent test repetitions.

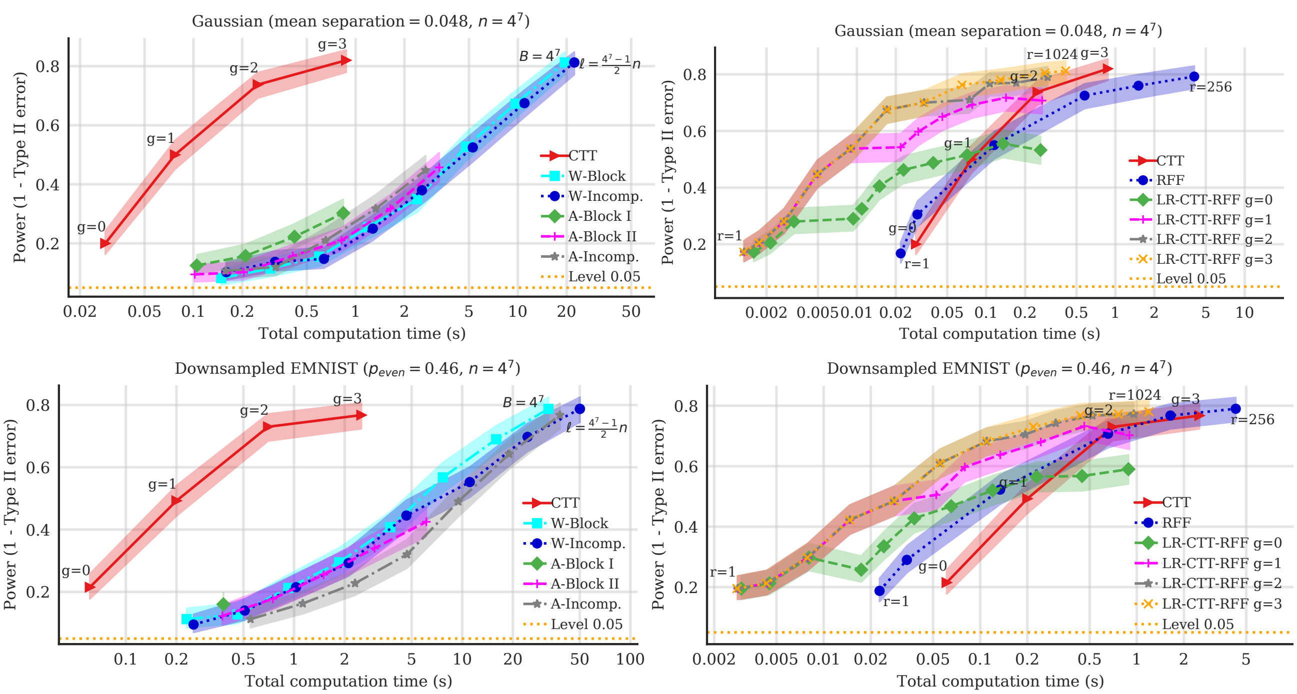

Figure 4: **Time-power trade-off curves** in the GAUSSIAN and EMNIST experimental settings comparing (*left*) CTT to five state-of-the-art approximate MMD tests based on subsampling and (*right*) LR-CTT to the state-of-the-art low-rank MMD test based on random Fourier features (RFF). These plots are like those in Fig. 1, but for a smaller sample size: $n = 4^7$ instead of $n = 4^9$.

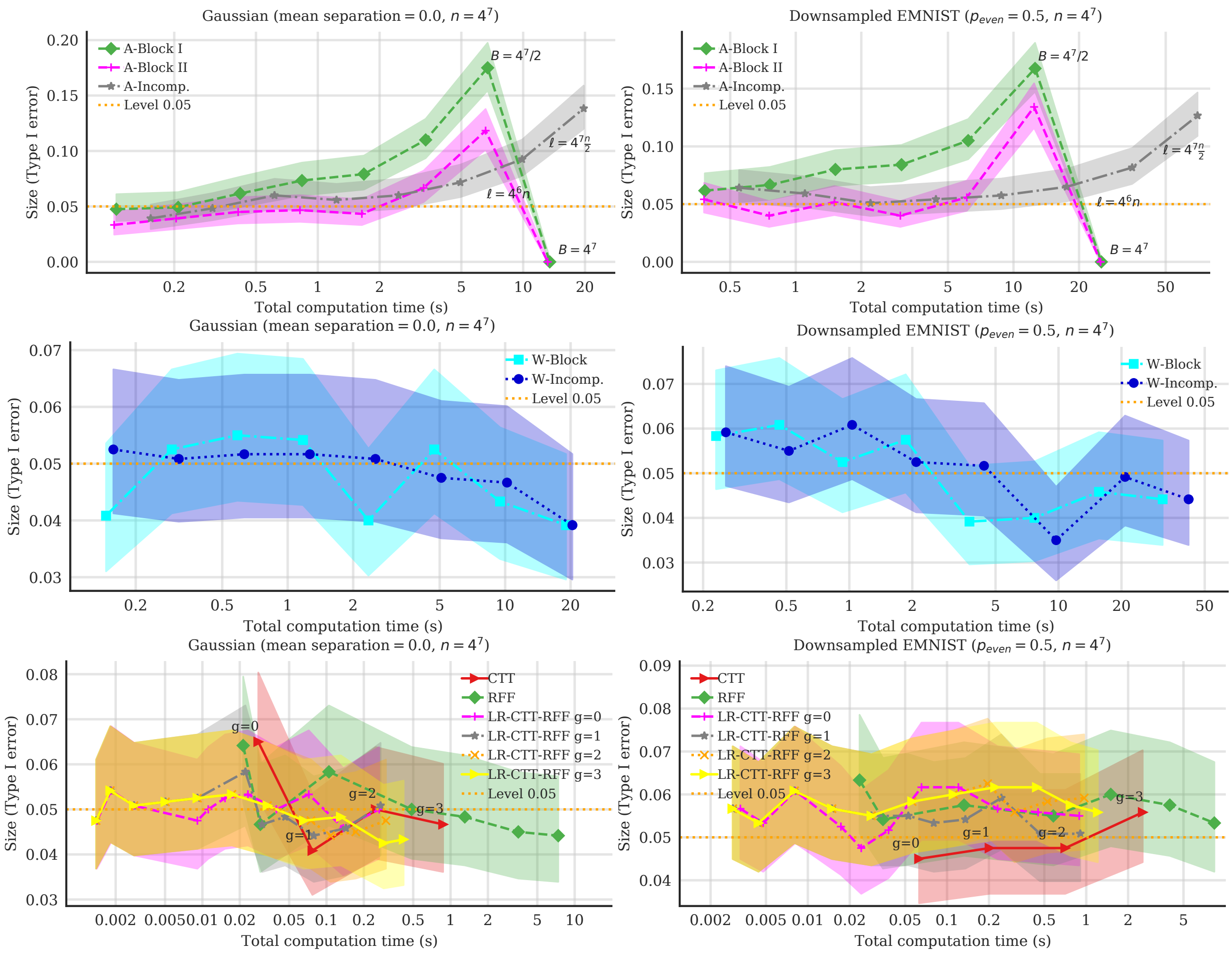

Figure 5: **Estimated test size** with 95% Wilson confidence intervals in the GAUSSIAN (*right*) and EMNIST (*left*) experimental settings of Fig. 4, i.e. with $n = 4^7$. **Top:** Asymptotic block and incomplete tests with 1200 independent test repetitions. **Middle:** Non-asymptotic wild bootstrap block and incomplete with 1200 independent test repetitions. **Bottom:** Non-asymptotic CTT, RFF, and LR-CTT with 1200 independent test repetitions.