ADAPTIVE TWO-SAMPLE TESTING VIA DEPENDENCE-AWARE WITNESS AGGREGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Two-sample testing is a fundamental problem in statistics and machine learning. Existing tests are typically powerful only against particular classes of alternatives. Since the underlying discrepancy is unknown in practice, relying on a single test can lead to substantial power loss. We introduce an adaptive two-sample testing framework that combines complementary procedures through a unified witness-function perspective. Our key observation is that many seemingly heterogeneous tests can be represented by witness functions whose expectation differences reveal discrepancies between two distributions. This representation enables their signals to be aggregated within a common framework. To handle the dependence among witness statistics evaluated on the same data, we develop a dependence-aware aggregation procedure. We establish asymptotic Type-I error control and show that the proposed test is consistent whenever at least one constituent witness carries sufficiently strong signal. Extensive experiments across diverse distributional shifts demonstrate that our method consistently delivers competitive or superior power across regimes while maintaining valid Type-I error control.

1 INTRODUCTION

Two sample distribution testing asks whether two independent samples are drawn from the same distribution, which is a fundamental problem in statistics and machine learning. It underlies a broad range of applications, including dataset shift detection (Lipton et al., 2018; Rabanser et al., 2019), model evaluation (Lloyd & Ghahramani, 2015), large language models (LLMs) (Zhang et al., 2024), world models (Yu et al., 2026), and cancer diagnosis (Borgwardt et al., 2006). The problem becomes particularly challenging in modern high-dimensional settings, where the distributional discrepancy may take many different forms. For example, the difference may be weak but spread across many coordinates, or concentrated on a small number of features or directions, or manifested through more complex changes in scale or dependence structure.

A broad range of algorithms have been developed for detecting such differences, including maximum mean discrepancy (MMD) (Gretton et al., 2012), energy distance (ED) (Székely & Rizzo, 2004), kernel Fisher discriminant analysis (KFDA) (Kübler et al., 2022), classifier-based tests (C2ST) (Lopez-Paz & Oquab, 2017; Kim et al., 2021), rank-based tests (RT) (Hu & Lei, 2024), and projection-based methods (Wang et al., 2021; Hu & Lin, 2025; 2026). While these methods have achieved good results, they are sensitive to different structures of distributional discrepancy. For example, statistics that accumulate discrepancies across many coordinates can be powerful when the signal is broadly distributed, whereas projection- or max-type procedures can be more effective when the difference is sparse or concentrated along a small number of informative directions. More generally, different procedures emphasize different features of the distributions and therefore attain high power under different alternatives.

This creates an important difficulty. The structure of the discrepancy is typically unknown in practice, and hence it is unclear which procedure should be used. A test that is well suited to a sparse alternative may perform poorly when the discrepancy is dense, while a test that is effective for diffuse signals may lose substantial power when the difference is concentrated. Figure 1 illustrates this issue: under low-rank covariance shifts, the relative performance of individual tests changes substantially as the rank varies (left), while under sparse-to-dense variance shifts, different tests are

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

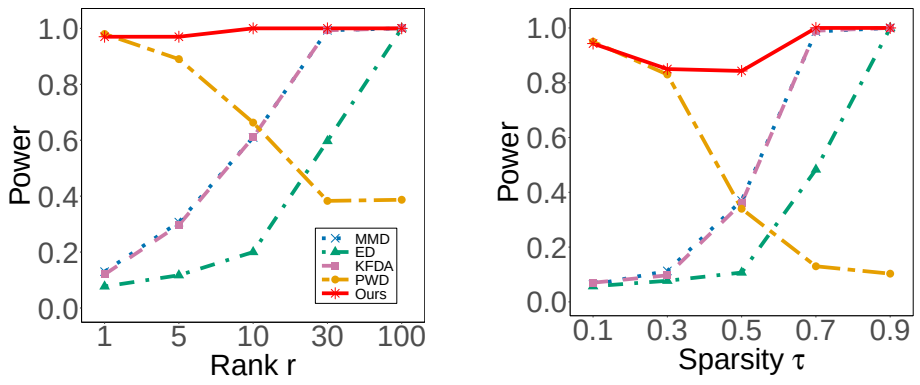

Figure 1: Motivation for adaptive two-sample testing. Empirical power under low-rank covariance shifts (left) and sparse-to-dense variance shifts (right). The relative performance of individual tests varies substantially with the underlying discrepancy structure, whereas the proposed procedure (Max-Cov) maintains strong power across a broad range of regimes.

favored at different sparsity levels (right). Thus, relying on a single procedure can lead to substantial power loss when the underlying discrepancy structure is unknown.

A natural solution is therefore to combine multiple complementary tests rather than committing to a single one. Ideally, such a procedure would automatically exploit whichever test is informative for the unknown alternative and thereby maintain strong power across a broad range of distributional discrepancies. This idea is simple in principle but difficult to realize in practice, because existing two-sample procedures are highly heterogeneous in their constructions; see Section 2 for representative examples. Consequently, it is not straightforward to combine them within a unified framework.

Our key observation is that many of these seemingly different procedures can nevertheless be understood through a common object: a *witness function*. Despite being derived from different discrepancy measures and optimization principles, many two-sample methods ultimately construct a function f for which the two distributions differ in expectation, $\mathbb{E}_P\{f(X)\} - \mathbb{E}_Q\{f(Y)\}$. The function f therefore acts as a witness to the discrepancy between P and Q. Integral probability metrics (IPMs) (Müller, 1997) provide a canonical example of this viewpoint, but the idea extends beyond IPMs and encompasses witnesses arising from distance-based, discriminative, classification-based, projection-based, and rank-based constructions. Table 1 summarizes representative methods and their associated witness functions, with the precise connections developed in Section 3.1.

This witness perspective provides a common language for combining otherwise heterogeneous two-sample procedures. Rather than attempting to aggregate their original test statistics directly, we first learn a collection of complementary witness functions and then evaluate their empirical mean differences on held-out data. In this way, different two-sample procedures are transformed into a common form: each produces a scalar signal measuring how strongly its learned witness separates the two distributions. The original problem of choosing among heterogeneous tests is thus converted into the problem of jointly assessing multiple witness-based mean differences.

However, this formulation reveals another issue. Because all learned witnesses are evaluated on the same testing sample, their corresponding statistics can be strongly dependent. Generic p-value combination methods (Fisher, 1925; Stouffer, 1949; Vovk & Wang, 2020; Liu & Xie, 2020) may fail to capture this dependence adequately, potentially leading to invalid Type-I error control or substantial power loss. We therefore explicitly account for the dependence among witness statistics and develop a dependence-aware aggregation procedure for adaptive two-sample testing.

Our framework follows a simple learn-then-test principle: we learn multiple complementary witness functions and combine their evidence in a dependence-aware manner. The resulting test does not need to know in advance what the underlying discrepancy is, and can exploit the signal whenever at least one constituent witness is sufficiently informative.

Table 1: Representative two-sample procedures and their witness functions.

Method	Discrepancy / optimization principle	Witness function
MMD	$\sup_{f\in\mathcal{H}_k,\|f\|_{\mathcal{H}_k}\le 1}\{\mathbb{E}f(X)-\mathbb{E}f(Y)\}$	$f_{\mathrm{MMD}}(z)=\frac{\mu_P(z)-\mu_Q(z)}{\|\mu_P-\mu_Q\|_{\mathcal{H}_k}}$
PWD	$\sup_{U\in\mathcal{V}_{d,k}}\sup_{g\in Lip_1}\{\mathbb{E}g(U^\top X)-\mathbb{E}g(U^\top Y)\}$	$f_{\mathrm{PWD}}(z)=g^\star(U^{\star\top}z)$
ED	$2\mathbb{E}\rho(X,Y)-\mathbb{E}\rho(X,X')-\mathbb{E}\rho(Y,Y')$	$f_{\mathrm{ED}}(z)=\mathbb{E}\rho(z,Y)-\mathbb{E}\rho(z,X)$
KFDA	$\sup_{f\in\mathcal{H}_k,\ f\neq 0}\frac{\langle\mu_P-\mu_Q,f\rangle_{\mathcal{H}_k}}{\langle f,(\Sigma_k+\lambda I)f\rangle_{\mathcal{H}_k}^{1/2}}$	$f_{\mathrm{KFDA}}(z)=\left[(\Sigma_k+\lambda I)^{-1}(\mu_P-\mu_Q)\right](z)$
C2ST	Classification accuracy	$f_{\mathrm{C2ST}}(z)=\{\mathbb{P}(g(X)>0)-\mathbb{P}(g(Y)>0)\}\mathbb{1}(g(z)>0)$
RT	Density-ratio ranking, $g=p/q$	$f_{\mathrm{RT}}(z)=\mathbb{P}(g(X)\le g(z))-\mathbb{P}(g(z)\le g(Y))$

The proposed framework is supported by both theoretical and empirical guarantees. We establish asymptotic Type-I error control for the dependence-aware test and derive power guarantees showing that the procedure is consistent whenever at least one constituent witness has sufficiently strong signal relative to its estimation and sampling errors. Thus, the method inherits the strengths of complementary constituent tests while adapting to the unknown form of the alternative. Empirically, as illustrated in Figure 1, the proposed procedure maintains strong power across settings in which the performance of individual tests varies substantially with rank and sparsity.

Our main contributions are summarized as follows.

- **A unified witness perspective.** We reveal a common witness-based representation underlying a broad class of heterogeneous two-sample procedures, providing a unified foundation for adaptive aggregation.
- **A dependence-aware adaptive test.** Building on this representation, we construct an adaptive test that aggregates complementary witness statistics while explicitly accounting for their dependence.
- **Rigorous theoretical guarantees.** We establish asymptotic Type-I error control and show that the proposed test is consistent whenever at least one constituent witness carries sufficiently strong signal.
- **Robust empirical performance.** Extensive experiments demonstrate that the proposed method maintains strong power across diverse distributional shifts, while individual tests can be highly sensitive to the structure of the alternative.

2 PRELIMINARIES

Let $\mathcal{X}$ be a separable metric space, typically a subset of $\mathbb{R}^d$. We observe two independent random samples in $\mathcal{X}$, $X_1,\dots,X_{n_x}\overset{i.i.d}{\sim}P$ and $Y_1,\dots,Y_{n_y}\overset{i.i.d}{\sim}Q$. Write $n=n_x+n_y$. The goal is to test the hypothesis

$$H_0: P=Q \qquad \text{versus} \qquad H_1: P\neq Q. \tag{1}$$

Existing two-sample procedures quantify the discrepancy between P and Q using different distance measures or discriminative criteria. We briefly review several popular methods.

MMD. The MMD (Gretton et al., 2012) is defined as

$$\sup_{f\in\mathcal{H}_k,\|f\|_{\mathcal{H}_k}\le 1}\{\mathbb{E}f(X)-\mathbb{E}f(Y)\},$$

where $\mathcal{H}_k$ is the reproducing kernel Hilbert space (RKHS) associated with a positive definite kernel $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$.

PWD. The PWD (Hu & Lin, 2026) is defined as

$$\sup_{U\in\mathcal{V}_{d,k}}\sup_{g\in Lip_1}\{\mathbb{E}g(U^\top X)-\mathbb{E}g(U^\top Y)\},$$

where $\mathcal{V}_{d,k} = \{V \in \mathbb{R}^{d\times k} : V^\top V = I_k\}$ and $Lip_1 = \{g : \mathbb{R}^k \to \mathbb{R} \mid g \text{ is 1-Lipschitz}, g(0) = 0\}$. For high-dimensional data, additional structural constraints, such as sparsity of the projection matrix U, can be imposed to improve estimation.

ED. Energy distance (ED) measures the discrepancy between two distributions through pairwise distances between and within the two populations. The generalized energy distance (Sejdinovic et al., 2013) is defined as

$$2\mathbb{E}\rho(X, Y) - \mathbb{E}\rho(X, X') - \mathbb{E}\rho(Y, Y'), \tag{2}$$

where $\rho : \mathcal{X} \times \mathcal{X} \to [0, +\infty)$ is a semimetric of negative type.

KFDA. The KFDA (Kübler et al., 2022) seeks a function in an RKHS by maximizing the mean separation between P and Q relative to the variability,

$$\sup_{f\in\mathcal{H}_k,\, f\neq 0} \frac{\langle \mu_P - \mu_Q, f\rangle_{\mathcal{H}_k}}{\langle f, (\Sigma_k + \lambda I) f\rangle_{\mathcal{H}_k}^{1/2}},$$

where $\lambda > 0$ and $\Sigma_k = \Sigma_P/r + \Sigma_Q/(1-r)$, $r = n_x/(n_x + n_y)$. Here, Σ_P and Σ_Q denote the covariance operators of $k(X, \cdot)$ and $k(Y, \cdot)$, respectively, and $\mu_P(z) = \mathbb{E}k(X, z)$, $\mu_Q(z) = \mathbb{E}k(Y, z)$.

C2ST. The classifier-based test (C2ST) (Lopez-Paz & Oquab, 2017) reduces the two-sample problem to binary classification by training a classifier to distinguish observations from P and Q, with out-of-sample classification performance used as the test statistic.

RT. Hu & Lei (2024) proposed a conformal rank-based test procedure that compares the relative ordering of observations under the two distributions. In the population setting, the optimal ranking score is given by the density ratio $g(z) = p(z)/q(z)$, where p and q denote the densities of P and Q, respectively. The test is then constructed from the ranks induced by g, with larger values indicating stronger evidence that an observation is more likely to arise from one distribution than the other.

The procedures above are derived from different principles and can exhibit substantially different power under different alternatives. In the next section, we show that they nevertheless share a common representation that motivates our adaptive testing framework.

3 METHODOLOGY

Although a wide range of two-sample tests have been developed, individual methods are typically tailored to particular forms of distributional discrepancy and may lose power under other alternatives. Since the underlying discrepancy is unknown in practice, selecting an appropriate test can be difficult. We first show that many representative procedures admit a common witness-based representation. Building on this perspective, we then develop an adaptive two-sample test that aggregates complementary witnesses to achieve robust power across diverse alternatives.

3.1 A UNIFIED WITNESS PERSPECTIVE

A central insight underlying our framework is that many widely used two-sample procedures admit a common representation through a witness function f, with the distributional discrepancy reflected by its difference in expectations under P and Q. For an informative witness, the population mean discrepancy $\Delta_f(P, Q) = \mathbb{E}_P f(X) - \mathbb{E}_Q f(Y)$ satisfies $\Delta_f(P, Q) = 0$ if $P = Q$ and $\Delta_f(P, Q) > 0$ whenever the witness successfully distinguishes P from Q. The original testing problem (1) can therefore be represented as

$$H_0' : \Delta_f(P, Q) = 0, \qquad \text{versus} \qquad H_1' : \Delta_f(P, Q) > 0. \tag{3}$$

This provides a common representation for otherwise heterogeneous two-sample procedures. We next show that the representative methods reviewed in Section 2 all admit such a witness-based formulation.

MMD. For MMD, the population witness function achieving the supremum is

$$f_{\mathrm{MMD}}(z) = \frac{\mu_P(z) - \mu_Q(z)}{\|\mu_P - \mu_Q\|_{\mathcal{H}_k}}.$$

Set $f_{\mathrm{MMD}} = 0$ when $\mu_P = \mu_Q$. The corresponding mean discrepancy is

$$\mathbb{E}f_{\mathrm{MMD}}(X) - \mathbb{E}f_{\mathrm{MMD}}(Y) = \|\mu_P - \mu_Q\|_{\mathcal{H}_k} = \mathrm{MMD}(P, Q),$$

which is zero under $P = Q$ and positive whenever the kernel distinguishes P and Q .

PWD. Let $(U^\star, g^\star)$ denote the solution of the PWD optimization problem. The corresponding witness is

$$f_{\mathrm{PWD}}(z) = g^\star(U^{\star\top} z).$$

Its mean discrepancy satisfies

$$\mathbb{E}f_{\mathrm{PWD}}(X) - \mathbb{E}f_{\mathrm{PWD}}(Y) = \mathrm{PWD}(P, Q).$$

Thus, the PWD discrepancy is also directly represented as a difference in expectations of its witness.

ED. For generalized ED, define

$$f_{\mathrm{ED}}(z) = \mathbb{E}\rho(z, Y) - \mathbb{E}\rho(z, X).$$

Then

$$\mathbb{E}f_{\mathrm{ED}}(X) - \mathbb{E}f_{\mathrm{ED}}(Y) = 2\mathbb{E}\rho(X, Y) - \mathbb{E}\rho(X, X') - \mathbb{E}\rho(Y, Y') = \mathrm{ED}_\rho(P, Q).$$

Hence, the generalized ED can likewise be expressed as a mean discrepancy associated with f_{ED}.

KFDA. For KFDA, the optimal witness function is proportional to

$$f_{\mathrm{KFDA}}(z) = \left[(\Sigma_k + \lambda I)^{-1}(\mu_P - \mu_Q)\right](z).$$

The corresponding mean discrepancy is

$$\mathbb{E}f_{\mathrm{KFDA}}(X) - \mathbb{E}f_{\mathrm{KFDA}}(Y) = \langle \mu_P - \mu_Q, (\Sigma_k + \lambda I)^{-1}(\mu_P - \mu_Q)\rangle_{\mathcal{H}_k}.$$

The quantity is zero under $P = Q$ and positive for $P \neq Q$, provided that k is characteristic.

C2ST. For the sign-based C2ST, Liu et al. (2020) showed that the statistic can be represented as an MMD with kernel $k_{\mathrm{C2ST}}(x, y) = \mathbb{1}(g(x) > 0)\mathbb{1}(g(y) > 0)/4$, where $g : \mathcal{X} \to \mathbb{R}$ denotes the classification score. A corresponding witness can therefore be taken as

$$f_{\mathrm{C2ST}}(z) = \big\{\mathbb{P}\big(g(X) > 0\big) - \mathbb{P}\big(g(Y) > 0\big)\big\}\mathbb{1}\big(g(z) > 0\big),$$

where an arbitrary positive scaling factor is immaterial. Let $acc = \{\mathbb{P}(g(X) > 0) + \mathbb{P}(g(Y) \leq 0)\}/2$. Its mean discrepancy is

$$\mathbb{E}f_{\mathrm{C2ST}}(X) - \mathbb{E}f_{\mathrm{C2ST}}(Y) = \{\mathbb{P}(g(X) > 0) - \mathbb{P}(g(Y) > 0)\}^2 = 4(acc - 1/2)^2,$$

which is 0 if $P = Q$ and becomes positive when g successfully separates the two distributions.

RT. Recall that $g(z) = p(z)/q(z)$ denotes the population density ratio. A corresponding rank-based witness is

$$f_{\mathrm{RT}}(z) = \mathbb{P}\big(g(X) \leq g(z)\big) - \mathbb{P}\big(g(z) \leq g(Y)\big).$$

Under H_0, $g = 1$ almost surely and $f_{\mathrm{RT}} = 0$ almost surely, so its mean discrepancy is zero. Under alternatives for which $g(X)$ and $g(Y)$ have continuous distributions, its mean discrepancy is

$$\mathbb{E}f_{\mathrm{RT}}(X) - \mathbb{E}f_{\mathrm{RT}}(Y) = 1 - 2\mathbb{P}(g(X) \leq g(Y)).$$

Under these alternatives, the density ratio assigns stochastically larger values to observations from P, with $\mathbb{P}(g(X) \leq g(Y)) < 1/2$, yielding a positive discrepancy.

These examples demonstrate that a broad class of two-sample procedures can be expressed through witness functions whose mean differences capture distributional discrepancies. Importantly, this perspective is more general than the classical IPM framework. The witness may arise from an IPM optimization (MMD and PWD), a distance-based construction (ED), a discriminative criterion (KFDA), classification (C2ST), or ranking (RT). This common witness-based representation provides a unified foundation for aggregating otherwise heterogeneous procedures within a single testing framework.

3.2 AN ADAPTIVE TWO-SAMPLE TEST

Building on the common witness representation, we propose an adaptive two-sample test that aggregates complementary witnesses learned from different procedures. Since the population witnesses are unknown, they must be estimated from data. To separate witness learning from subsequent statistical inference, we adopt a sample-splitting strategy. One subsample is used to estimate the witnesses, while an independent subsample is used to construct the final test statistic.

Specifically, we split each of the two samples into disjoint fitting and testing subsamples, with sizes $(\tilde{n}_x, \tilde{n}_y)$ and $(\bar{n}_x, \bar{n}_y)$, respectively, where $\bar{n}_x = n_x - \tilde{n}_x$ and $\bar{n}_y = n_y - \tilde{n}_y$. Let $\mathcal{D}_{\text{fit}}$ denote the fitting subsample, and suppose that m witness functions are learned from $\mathcal{D}_{\text{fit}}$. We collect them as $\hat{f} = (\hat{f}_1, \ldots, \hat{f}_m)^\top$. For simplicity, we treat m as fixed in our theoretical analysis. The theory can be extended to allow m to grow with n with minor modifications.

Let $X_1^{\text{test}}, \ldots, X_{\bar{n}_x}^{\text{test}}$ and $Y_1^{\text{test}}, \ldots, Y_{\bar{n}_y}^{\text{test}}$ denote the testing observations. We define the vector of empirical witness mean differences by

$$\hat{S}_n = \frac{1}{\bar{n}_x} \sum_{i=1}^{\bar{n}_x} \hat{f}(X_i^{\text{test}}) - \frac{1}{\bar{n}_y} \sum_{i=1}^{\bar{n}_y} \hat{f}(Y_i^{\text{test}}).$$

Under the null hypothesis, each component of $\hat{S}_n$ is centered around zero, while under the alternative at least one component can exhibit a nonzero mean when the corresponding witness captures the distributional discrepancy.

To account for the potentially strong dependence among the witness statistics, we estimate their covariance structure using the testing sample. Let $\hat{C}_X$ and $\hat{C}_Y$ be the sample covariance matrices of $\hat{f}(X_i^{\text{test}})$ and $\hat{f}(Y_i^{\text{test}})$, respectively, each centered at its own testing-sample mean and normalized by its sample size, $\bar{n}_x$ or $\bar{n}_y$. Define $\hat{\Sigma} = (\bar{n}_y \hat{C}_X + \bar{n}_x \hat{C}_Y)/(\bar{n}_x + \bar{n}_y)$. We then whiten the vector of witness mean differences and define the proposed test statistic as

$$T_n = \left(\frac{\bar{n}_x \bar{n}_y}{\bar{n}_x + \bar{n}_y} \right)^{1/2} \left\| \hat{\Sigma}^{-1/2} \hat{S}_n \right\|_\infty .$$

When $\hat{\Sigma}$ is singular, we define $\hat{\Sigma}^{-1/2} = (\hat{\Sigma}^+)^{1/2}$, where $\hat{\Sigma}^+$ is the Moore–Penrose pseudoinverse. The whitening step accounts for the covariance among the witness statistics and yields asymptotically independent standard normal coordinates under the null, allowing evidence from heterogeneous procedures to be aggregated on a common standardized scale. For a significance level $\alpha \in (0, 1)$, we reject H_0 if $T_n > q_{1-\alpha}$, where $q_{1-\alpha}$ is the $1 - \alpha$ quantile of $\max_{j=1,\ldots,m} |Z_j|$ with $Z \sim N(0, I_m)$, as established in Theorem 1. We refer to this covariance-adjusted maximum statistic T_n as Max-Cov.

Remark 1. *Although* (3) *suggests a one-sided test, the whitening transformation $\hat{\Sigma}^{-1/2} \hat{S}_n$ mixes the coordinates and may change their signs. Thus, we adopt a two-sided statistic, which is invariant to such sign changes and performs well in practice.*

Remark 2. *A Hotelling-type quadratic statistic, referred to as* Hotelling*, can also be considered,*

$$T_n' = \frac{\bar{n}_x \bar{n}_y}{\bar{n}_x + \bar{n}_y} \hat{S}_n^\top \hat{\Sigma}^{-1} \hat{S}_n.$$

For $\alpha \in (0, 1)$, we reject H_0 if $T_n' > q_{1-\alpha}'$, where $q_{1-\alpha}'$ is the $1 - \alpha$ quantile of $\chi^2(m)$.

Alternatively, we can standardize each witness statistic separately,

$$T_n'' = \left(\frac{\bar{n}_x \bar{n}_y}{\bar{n}_x + \bar{n}_y} \right)^{1/2} \max_{j=1,\ldots,m} \frac{\hat{S}_{n,j}}{\hat{\Sigma}_{jj}^{1/2}},$$

which we refer to as Max-Var. *$\hat{\Sigma}_{jj}$ is the jth diagonal element of $\hat{\Sigma}$. Since entrywise standardization preserves the sign of each mean discrepancy, a one-sided statistic is natural in this case. However, the limiting null distribution of T_n'' depends on the unknown correlation structure among the witness statistics. A permutation procedure or Gaussian multiplier bootstrap can be used to approximate its null distribution and determine the critical value or p-value.*

4 THEORETICAL RESULTS

We write $\mathbb{E}_\star$ and $\mathrm{Var}_\star$ for expectation and variance conditional on $\mathcal{D}_{\mathrm{fit}}$, and define

$$\Sigma = \frac{\bar{n}_y \mathrm{Var}_\star\{\hat{f}(X)\} + \bar{n}_x \mathrm{Var}_\star\{\hat{f}(Y)\}}{\bar{n}_x + \bar{n}_y}.$$

We begin by stating the assumptions required for our theoretical analysis. Assumption 1 requires the sample-size ratio n_x/n_y to be bounded away from zero and infinity. Assumption 2 imposes mild regularity conditions on the second and fourth moments of the learned witnesses.

Assumption 1. *For $0 < c_1 \le c_2 < 1$, $n_x/(n_x+n_y), \tilde{n}_x/n_x, \tilde{n}_y/n_y \in [c_1, c_2]$.*

Assumption 2. *Let $v_j = \Sigma_{jj} > 0$ and $V = \mathrm{diag}(v_1, \ldots, v_m)$. There exist $\kappa > 0$ and a deterministic sequence $L_n > 0$ with $L_n/n \to 0$ such that, almost surely with respect to $\mathcal{D}_{\mathrm{fit}}$, $V^{-1/2}\Sigma V^{-1/2} \succeq \kappa I_m$ and $\mathbb{E}_\star\big|\hat{f}_j(X) - \mathbb{E}_\star \hat{f}_j(X)\big|^4 / v_j^2 \le L_n$, $\mathbb{E}_\star\big|\hat{f}_j(Y) - \mathbb{E}_\star \hat{f}_j(Y)\big|^4 / v_j^2 \le L_n$.*

Theorem 1 (Null distribution)**.** *Suppose that Assumptions 1 and 2 hold. Under H_0, we have $T_n \xrightarrow{d} \max_{1\le j\le m} |Z_j|$, where $Z = (Z_1, \ldots, Z_m)^\top \sim N(0, I_m)$. Moreover,*

$$\sup_{t\in\mathbb{R}} \left| \mathbb{P}(T_n \le t) - \mathbb{P}\left(\max_{1\le j\le m} |Z_j| \le t \right) \right| \le \min\left\{ 1,\ c\sqrt{\frac{L_n}{n}} \right\},$$

where $c > 0$ depends only on m, c_1, c_2, κ.

Theorem 1 establishes the asymptotic null distribution of T_n, yielding an explicit critical value and avoiding resampling for calibration. Notably, the result is agnostic to the specific witness-learning algorithms and remains valid in high-dimensional settings under the stated conditions. The Berry–Esseen type bound further quantifies the approximation error, establishing asymptotic control of the size at the nominal significance level α, $|\mathbb{P}(T_n > q_{1-\alpha}) - \alpha| \to 0$.

While the test size is insensitive to the specific witness learning algorithm, the choice and quality of the learned witnesses directly affect power. Because our adaptive testing procedure is formulated as a general framework, we first establish a generic power guarantee in Theorem 2 without specifying the particular witnesses. We then specialize this result to the constituent witnesses used in Section 5, deriving more explicit sufficient conditions for consistency in Appendix D.

For each fitted witness $\hat{f}_j$, let f_j denote the corresponding population witness. We have

$$\hat{S}_{n,j} - \Delta_{f_j}(P,Q) = \big\{\hat{S}_{n,j} - \Delta_{\hat{f}_j}(P,Q)\big\} + \big\{\Delta_{\hat{f}_j}(P,Q) - \Delta_{f_j}(P,Q)\big\},$$

where $\Delta_{\hat{f}_j}(P,Q) = \mathbb{E}_\star \hat{S}_{n,j}$. The two terms represent the testing-sample fluctuation around the conditional mean and the population signal error caused by fitting, respectively.

Theorem 2 (Power guarantee)**.** *Suppose Assumptions 1 and 2 hold. For deterministic sequences $a_{n,j}, b_{n,j} \ge 0$,*

$$\mathbb{P}\Big(|\hat{S}_{n,j} - \Delta_{\hat{f}_j}(P,Q)| > a_{n,j}\Big) \to 0 \quad \textit{and} \quad \mathbb{P}\Big(|\Delta_{\hat{f}_j}(P,Q)| < \Delta_{f_j}(P,Q) - b_{n,j}\Big) \to 0. \tag{4}$$

If there exists a coordinate $j \in \{1, \ldots, m\}$ such that

$$\Delta_{f_j}(P,Q) > a_{n,j} + b_{n,j} \quad \textit{and} \quad \frac{v_j}{n\{\Delta_{f_j}(P,Q) - a_{n,j} - b_{n,j}\}^2} \xrightarrow{P} 0, \tag{5}$$

then for every fixed $\alpha \in (0,1)$, $\mathbb{P}(T_n > q_{1-\alpha}) \to 1$.

Theorem 2 establishes a general power guarantee for our adaptive testing procedure. Consistency follows as long as there exists at least one witness whose population signal is sufficiently large relative to the corresponding estimation errors: the testing-sample estimation error $a_{n,j}$ and the loss $b_{n,j}$ incurred by learning the witness from the fitting data. Because the theorem does not specialize to any particular witness construction, both the error terms and the signal condition are stated in generic form, without imposing specific convergence rates.

The bound involving $\Delta_{\hat{f}_j}(P,Q)$ is one-sided. It only requires the fitted witness to retain the population signal up to an error $b_{n,j}$, and therefore allows $|\Delta_{\hat{f}_j}(P,Q)|$ to exceed $\Delta_{f_j}(P,Q)$. Moreover, $a_{n,j}$ and $b_{n,j}$ need not be negligible relative to the population signal. It is sufficient that the residual signal $\Delta_{f_j}(P,Q) - a_{n,j} - b_{n,j}$ remains positive and dominates the stochastic scale $\sqrt{v_j/n}$ as specified in (5).

Table 2: Empirical Type-I error rates of different methods under the null models.

Model	Single method				Naive aggregation				Proposed		
	MMD	ED	KFDA	PWD	Bonf	Fisher	Cauchy	Stou	Max-Cov	Max-Var	Hotelling
Gaussian	0.027	0.023	0.023	0.063	0.027	0.083	0.040	0.120	0.067	0.047	0.070
Gaussian mixture	0.050	0.060	0.047	0.073	0.030	0.133	0.060	0.163	0.067	0.057	0.060

Table 3: Empirical power of different methods under the alternative models.

Model	Setting	Single method				Naive aggregation		Proposed		
		MMD	ED	KFDA	PWD	Bonf	Cauchy	Max-Cov	Max-Var	Hotelling
A	$r = 1$	0.130	0.077	0.120	0.980	0.850	0.830	0.970	0.637	0.987
	$r = 5$	0.307	0.117	0.297	0.890	0.757	0.793	0.970	0.433	0.997
	$r = 10$	0.610	0.200	0.610	0.663	0.657	0.737	1.000	0.380	1.000
	$r = 30$	0.993	0.597	0.993	0.383	1.000	1.000	1.000	0.987	1.000
	$r = 100$	1.000	1.000	1.000	0.387	1.000	1.000	1.000	1.000	1.000
B	$\tau = 0.1$	0.063	0.057	0.070	0.950	0.870	0.870	0.943	0.953	0.957
	$\tau = 0.3$	0.113	0.077	0.097	0.830	0.693	0.707	0.850	0.880	0.907
	$\tau = 0.5$	0.370	0.107	0.360	0.340	0.293	0.390	0.843	0.483	0.923
	$\tau = 0.7$	0.987	0.483	0.990	0.130	0.960	0.970	1.000	0.943	1.000
	$\tau = 0.9$	1.000	1.000	1.000	0.103	1.000	1.000	1.000	1.000	1.000
C	$\tau = 0.1$	0.677	0.670	0.740	1.000	0.997	1.000	1.000	1.000	1.000
	$\tau = 0.3$	0.640	0.647	0.617	0.763	0.673	0.793	0.710	0.927	0.820
	$\tau = 0.5$	0.647	0.640	0.563	0.413	0.513	0.643	0.373	0.697	0.467
	$\tau = 0.7$	0.587	0.583	0.523	0.250	0.447	0.570	0.217	0.557	0.353
	$\tau = 0.9$	0.603	0.613	0.543	0.257	0.420	0.583	0.260	0.493	0.310
D	$\beta = 0.2$	0.123	0.080	0.363	0.700	0.563	0.583	0.670	0.510	0.803
	$\beta = 0.5$	0.313	0.187	0.953	0.997	0.997	1.000	0.997	0.970	1.000
E	$\beta = 0.2$	0.050	0.047	0.070	0.833	0.673	0.683	0.893	0.277	0.937
	$\beta = 0.5$	0.080	0.057	0.157	1.000	0.990	0.980	1.000	0.463	1.000

5 Numerical Experiments

5.1 Simulation studies

We evaluate the three proposed aggregation statistics: Max-Cov (T_n), Hotelling (T_n'), and Max-Var (T_n''). We compare the proposed procedure with the individual witness-based MMD, ED, KFDA, and PWD tests, as well as naive p-value aggregation of these four methods using Bonferroni (Bonf), Fisher, Cauchy and Stouffer (Stou) combination rules (Fisher, 1925; Stouffer, 1949; Liu & Xie, 2020). The proposed procedure aggregates the witnesses learned by the four constituent methods. Detailed implementation and hyperparameter selection procedures are provided in Appendix B. We set $n_x = n_y = 250$, $d = 500$ and $\alpha = 0.05$, and conduct 300 independent Monte Carlo repetitions.

Null models. We evaluate Type-I error control under the Gaussian distribution $N(0_d, \Sigma)$, where $\Sigma = (r_{ij})_{i,j=1}^d$ with $r_{ij} = 0.5^{|i-j|}$, and the Gaussian mixture $0.5N(-1_d, I_d) + 0.5N(1_d, I_d)$. As shown in Table 2, the four individual tests and all variants of the proposed procedure maintain Type-I error close to the nominal level across the two null settings. In contrast, the Fisher and Stouffer combination rules exhibit inflated Type-I error, likely because they do not adequately account for the dependence among the constituent p-values. The Bonferroni and Cauchy combination rules maintain satisfactory Type-I error control and are therefore retained for the subsequent power comparisons.

Alternative models. To evaluate power across different forms of distributional shift, we consider five alternatives, covering low-rank covariance shift, sparse-to-dense variance and mean shifts, and two non-Gaussian settings. For Models A–C, we vary the rank or sparsity of the shift while fixing its overall signal magnitude, measured by the Frobenius norm for covariance differences and the Euclidean norm for mean differences. This allows us to study how test performance changes with the structure of the alternative rather than with signal strength. Models D and E consider non-Gaussian alternatives. Model D changes the marginal distributions while preserving their means and variances, whereas Model E changes the joint distribution while preserving all univariate marginals.

In Models A–C, r and τ control the rank and sparsity of the discrepancy, respectively, while β controls the extent of the distributional change in Models D and E. Full model specifications are provided in Appendix B.1.

Table 4: Empirical power on MNIST.

n_x	σ	Single method				Naive aggregation		Proposed		
		MMD	ED	KFDA	PWD	Bonf	Cauchy	Max-Cov	Max-Var	Hotelling
350	80	0.167	0.177	0.347	0.100	0.197	0.270	0.377	0.143	0.437
	100	0.263	0.263	0.440	0.187	0.280	0.373	0.510	0.220	0.620
	120	0.407	0.413	0.493	0.317	0.383	0.500	0.600	0.307	0.660
500	80	0.253	0.273	0.590	0.170	0.433	0.490	0.603	0.360	0.680
	100	0.440	0.453	0.700	0.357	0.560	0.613	0.727	0.483	0.773
	120	0.637	0.643	0.767	0.530	0.667	0.750	0.780	0.607	0.867
800	80	0.463	0.473	0.893	0.253	0.773	0.827	0.863	0.720	0.907
	100	0.743	0.743	0.953	0.600	0.907	0.937	0.927	0.853	0.983
	120	0.873	0.877	0.977	0.883	0.973	0.983	0.947	0.933	0.990

Table 3 reveals pronounced regime dependence among the individual tests. In Models A and B, PWD is particularly powerful when the covariance discrepancy is concentrated in a small number of directions or coordinates, whereas MMD and KFDA become more competitive as the signal is spread more broadly. Consequently, the best individual method changes substantially with the structure of the alternative. In contrast, the proposed aggregation procedures remain much more stable across these regimes. In particular, Max-Cov and Hotelling achieve high power throughout Models A and B and substantially outperform the individual tests in several intermediate settings, such as Model A with $r = 10$ and Model B with $\tau = 0.5$, where the constituent tests attain only moderate power. Similar behavior is observed in the non-Gaussian Models D and E. Overall, Max-Cov performs consistently well across the considered alternatives, while Hotelling generally achieves the strongest power and Max-Var is particularly competitive for the mean-shift alternatives in Model C. The Bonferroni and Cauchy combinations improve robustness relative to individual tests, but are generally less powerful than the proposed dependence-aware aggregation. Additional comparison results with other methods are provided in Appendix C.1.

5.2 Real data applications

We complement our simulation study with an experiment on MNIST (LeCun et al., 1998) to evaluate the tests on real image data. Distributional differences are introduced by adding Gaussian noise to a subset of images. Let P denote the distribution of the original images, normalized to $[0, 1]$ and represented as 784-dimensional vectors. The perturbed distribution is then modeled as

$$Q_{\pi,\sigma} = (1-\pi)P + \pi P_\sigma, \qquad P_\sigma = \mathcal{L}\Big(\mathrm{clip}_{[0,1]}(Z+\varepsilon)\Big),$$

where $Z \sim P$, $\varepsilon \sim \mathcal{N}(0, (\sigma/255)^2 I_{784})$ is independent of Z, and $\mathcal{L}$ denotes the distribution of a random vector. The parameters π and σ control the fraction of perturbed images and the noise magnitude, respectively. We consider $n_x = n_y \in \{350, 500, 800\}$, $\pi \in \{0, 0.1, 0.15\}$ and $\sigma \in \{80, 100, 120\}$, where $\pi = 0$ corresponds to the null.

Table 4 reports empirical power for $\pi = 0.1$ over 300 repetitions at significance level $\alpha = 0.05$. Results for remaining settings are provided in Appendix C.2. The proposed aggregation procedures generally outperform the individual tests and naive aggregation. Hotelling achieves the highest empirical power in all reported settings, while Max-Cov also performs strongly and often outperforms the best constituent test. Bonferroni and Cauchy improve robustness over the individual tests but are typically less powerful than the proposed dependence-aware aggregation.

6 Conclusion

We propose an adaptive two-sample testing framework that combines complementary tests through a unified witness-function perspective while accounting for dependence among the resulting statistics. The method enjoys asymptotic Type-I error control and consistent power whenever at least one constituent witness is sufficiently informative, and shows robust performance across diverse discrepancy structures. This is particularly useful in real applications where the form of distributional difference is typically unknown in advance. Moreover, the framework is open and extensible, allowing future advances in two-sample testing to be naturally incorporated as additional witnesses.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, improving the clarity and presentation of the manuscript, and checking notation consistency. All AI-assisted revisions were reviewed by the authors. The authors take full responsibility for the final content of the paper.

ETHICS STATEMENT

This work develops a general statistical methodology for two-sample testing and does not involve human subjects, private or sensitive data, or potentially harmful applications. The empirical studies use standard benchmark and publicly available datasets. We are not aware of any specific ethical concerns arising from this work.

REPRODUCIBILITY STATEMENT

We provide detailed implementation procedures and experimental settings in the Appendix, together with complete proofs for the theoretical results. We will release the source code to facilitate reproducibility.

REFERENCES

Vidmantas Bentkus. A Lyapunov-type bound in $\mathbb{R}^d$. *Theory of Probability and Its Applications*, 49: 311–323, 2005.

Felix Biggs, Antonin Schrab, and Arthur Gretton. Mmd-fuse: Learning and combining kernels for two-sample testing without data splitting. In *Advances in Neural Information Processing Systems*, 2023.

Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

Louis H. Y. Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Stein's method*. Springer Science & Business Media, 2010.

Ronald A Fisher. *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd, 1925.

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

Xiaoyu Hu and Jing Lei. A two-sample conditional distribution test using conformal prediction and weighted rank sum. *Journal of the American Statistical Association*, 119(546):1136–1154, 2024.

Xiaoyu Hu and Zhenhua Lin. Two-sample distribution tests in high dimensions via max-sliced wasserstein distance and bootstrapping. *Biometrika*, 112(2), 2025.

Xiaoyu Hu and Zhenhua Lin. Neural wasserstein two-sample tests. *arXiv:2601.21732*, 2026.

Ilmun Kim, Aaditya Ramdas, Aarti Singh, and Larry Wasserman. Classification accuracy as a proxy for two-sample testing. *The Annals of Statistics*, 49(1):411–434, 2021.

Jonas M Kübler, Wittawat Jitkrittum, Bernhard Schölkopf, and Krikamol Muandet. A witness two-sample test. In *International Conference on Artificial Intelligence and Statistics*, pp. 1403–1419, 2022.

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

Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*, pp. 3122–3130, 2018.

Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning*, pp. 6316–6326, 2020.

Yaowu Liu and Jun Xie. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529): 393–402, 2020.

James R Lloyd and Zoubin Ghahramani. Statistical model criticism using kernel two sample tests. In *Advances in Neural Information Processing Systems*, 2015.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations*, 2017.

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

Stephan Rabanser, Stephan Günnemann, and Zachary Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems*, 2019.

Haskell P. Rosenthal. On the subspaces of L^p $(p > 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics*, 8:273–303, 1970.

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

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013.

Shubhanshu Shekhar, Ilmun Kim, and Aaditya Ramdas. A permutation-free kernel two-sample test. In *Advances in Neural Information Processing Systems*, 2022.

Samuel A Stouffer. *Adjustment during Army Life*. Princeton University Press, 1949.

Gábor J. Székely and M. L. Rizzo. Testing for equal distributions in high dimension. *InterStat*, 2004.

Vladimir Vovk and Ruodu Wang. Combining p-values via averaging. *Biometrika*, 107(4):791–808, 2020.

Jie Wang, Rui Gao, and Yao Xie. Two-sample test using projected wasserstein distance. In *IEEE International Symposium on Information Theory (ISIT)*, pp. 3320–3325, 2021.

Zhoushun Yu, Xiaoyu Hu, and Xiangyu Xu. QQWorld: Quantile-quantile matching for world model regularization. *arXiv:2607.28415*, 2026.

Shuhai Zhang, Yiliao Song, Jiahao Yang, Yuanqing Li, Bo Han, and Mingkui Tan. Detecting machine-generated texts by multi-population aware optimization for maximum mean discrepancy. In *International Conference on Learning Representations*, 2024.

A Additional Related Work

Learn-then-test two-sample procedures. Although several methods follow a learn-then-test paradigm, where a function is learned on a training split and then used to construct a test statistic on held-out data (Lopez-Paz & Oquab, 2017; Kübler et al., 2022; Shekhar et al., 2022), they typically learn a single function within a prescribed model or kernel family. Their power can therefore remain sensitive to the form of the alternative, as also observed in Section 5. More importantly, they do not address how to combine heterogeneous tests built on different discrepancy measures and learning principles. Our framework instead identifies a common witness-based representation and builds on it to develop a dependence-aware aggregation procedure for combining complementary tests.

Aggregated kernel tests. MMDAgg (Schrab et al., 2023) and MMD-FUSE (Biggs et al., 2023) improve robustness by aggregating MMD statistics over multiple kernels, and thus adapt only within the MMD family. In contrast, our method aggregates heterogeneous tests built from different discrepancy measures and learning principles, thereby achieving adaptation beyond a single discrepancy family. Numerical comparisons with MMDAgg and MMD-FUSE are provided in Appendix C.1.

B Implementation Details

B.1 Model setup

Let 0_d and 1_d be d-dimensional vectors with elements all equal to 0 and 1, respectively. Let I_d denote the $d \times d$ identity matrix. The following models are used to evaluate power in Section 5.1.

- Model A: let $X \sim N(0_d, I_d)$ and $Y \sim N(0_d, I_d + 8r^{-0.5} U_r U_r^\top)$, where $U_r \in \mathbb{R}^{d\times r}$ has orthonormal columns satisfying $U_r^\top U_r = I_r$.
- Model B: let $X \sim N(0_d, \Sigma)$ and $Y \sim N(0_d, \Sigma')$, where $\Sigma_{ij} = 0.5^{|i-j|}$ and Σ' is defined by replacing the diagonal entries of Σ with $\Sigma'_{jj} = 1 + 5s^{-0.5}$ for $j = 1, \ldots, s$ and $\Sigma'_{jj} = 1$ for $j = s+1, \ldots, d$. Set $s = \lfloor d^\tau \rfloor$, where $\lfloor \cdot \rfloor$ denotes rounding down to the nearest integer.
- Model C: let $X \sim N(0_d, \Sigma)$ and $Y \sim N(\mu, \Sigma)$, where Σ is as in Model B and $\mu_j = s^{-0.5}$ for $j = 1, \ldots, s$ and 0 otherwise. Set $s = \lfloor d^\tau \rfloor$.
- Model D: let $X \sim 0.5N(-1_d, I_d) + 0.5N(1_d, I_d)$ and $Y = (Y_1^\top, Y_2^\top)^\top$, where $Y_1 \sim N(0_s, 2R)$ and $Y_2 \sim 0.5N(-1_{d-s}, I_{d-s}) + 0.5N(1_{d-s}, I_{d-s})$ are independent, with $R_{ij} = 0.5^{|i-j|}$. Set $s = [5\beta]$, where $[\cdot]$ denotes rounding to the nearest integer.
- Model E: let $X \sim 0.5N(-1_d, I_d) + 0.5N(1_d, I_d)$ and $Y \sim 0.5N(-1_d, \Sigma) + 0.5N(1_d, \Sigma)$, where $\Sigma = \begin{pmatrix} R & 0 \\ 0 & I_{d-s} \end{pmatrix}$, $R = (r_{ij})_{i,j=1}^s$ and $r_{ij} = 0.9^{|i-j|}$. Set $s = [100\beta]$.

B.2 Witness construction

We include four methods for witness construction and aggregation in the proposed procedure: maximum mean discrepancy (MMD), energy distance (ED), kernel Fisher discriminant analysis (KFDA), and projected Wasserstein distance (PWD). Each constituent method is also evaluated individually to assess the benefit of the proposed aggregation. We further compare against naive p-value aggregation of the four individual tests. Specifically, the implementations of the witness-based MMD, KFDA and PWD procedures follow Kübler et al. (2022) and Hu & Lin (2026). We provide the implementation details below for completeness.

Each sample is split randomly into fitting and testing subsamples. Denote their sizes by $(\tilde{n}_x, \tilde{n}_y)$ and $\bar{n}_x, \bar{n}_y$, respectively, and write

$$\mathcal{X}_{\text{fit}} = \{X_i^{\text{fit}}\}_{i=1}^{\tilde{n}_x}, \qquad \mathcal{Y}_{\text{fit}} = \{Y_j^{\text{fit}}\}_{j=1}^{\tilde{n}_y},$$
$$\mathcal{X}_{\text{test}} = \{X_i^{\text{test}}\}_{i=1}^{\bar{n}_x}, \quad \mathcal{Y}_{\text{test}} = \{Y_j^{\text{test}}\}_{j=1}^{\bar{n}_y}.$$

The sample splitting ratio is $\eta_{\text{split}} = 0.5$. Throughout, $z \in \mathbb{R}^d$ denotes an observation at which a witness is evaluated, and $\|\cdot\|_2$ denotes Euclidean norm.

MMD witness. MMD compares similarity to the two fitting groups. For candidate c, let $\sigma = c\, s(\mathcal{X}_{\text{fit}}, \mathcal{Y}_{\text{fit}})$. The Gaussian kernel k_σ measures similarity between points $u, v \in \mathbb{R}^d$:

$$k_\sigma(u, v) = \exp\left(-\frac{\|u - v\|_2^2}{2\sigma^2}\right).$$

The difference in average similarities gives g_{MMD}, with normalization factor q_{MMD}:

$$g_{\mathrm{MMD}}(z) = \frac{1}{\tilde{n}_x}\sum_{i=1}^{\tilde{n}_x} k_\sigma(z, X_i^{\mathrm{fit}}) - \frac{1}{\tilde{n}_y}\sum_{j=1}^{\tilde{n}_y} k_\sigma(z, Y_j^{\mathrm{fit}}),$$

$$q_{\mathrm{MMD}} = \frac{1}{\tilde{n}_x^2}\sum_{i,i'=1}^{\tilde{n}_x} k_\sigma(X_i^{\mathrm{fit}}, X_{i'}^{\mathrm{fit}}) + \frac{1}{\tilde{n}_y^2}\sum_{j,j'=1}^{\tilde{n}_y} k_\sigma(Y_j^{\mathrm{fit}}, Y_{j'}^{\mathrm{fit}}) - \frac{2}{\tilde{n}_x\tilde{n}_y}\sum_{i=1}^{\tilde{n}_x}\sum_{j=1}^{\tilde{n}_y} k_\sigma(X_i^{\mathrm{fit}}, Y_j^{\mathrm{fit}}).$$

Dividing by $\sqrt{q_{\mathrm{MMD}}}$ gives the witness $\hat{f}_{\mathrm{MMD}}$:

$$\hat{f}_{\mathrm{MMD}}(z) = \frac{g_{\mathrm{MMD}}(z)}{\sqrt{q_{\mathrm{MMD}}}}.$$

The within-group sums include diagonal terms.

ED witness. ED follows the same construction using a kernel derived from distances. For candidate exponent β, denote the powered distance by d_β and its associated kernel by k_β:

$$d_\beta(u, v) = \|u - v\|_2^\beta,$$
$$k_\beta(u, v) = \frac{1}{2}\big(\|u\|_2^\beta + \|v\|_2^\beta - d_\beta(u, v)\big).$$

Let g_{ED} be the difference between the two fitting-sample averages of $k_\beta(z, \cdot)$:

$$g_{\mathrm{ED}}(z) = \frac{1}{2\tilde{n}_x}\sum_{i=1}^{\tilde{n}_x} \|X_i^{\mathrm{fit}}\|_2^\beta - \frac{1}{2\tilde{n}_y}\sum_{j=1}^{\tilde{n}_y} \|Y_j^{\mathrm{fit}}\|_2^\beta + \frac{1}{2\tilde{n}_y}\sum_{j=1}^{\tilde{n}_y} d_\beta(z, Y_j^{\mathrm{fit}}) - \frac{1}{2\tilde{n}_x}\sum_{i=1}^{\tilde{n}_x} d_\beta(z, X_i^{\mathrm{fit}}).$$

The normalization factor q_{ED} and witness $\hat{f}_{\mathrm{ED}}$ are

$$q_{\mathrm{ED}} = \frac{1}{\tilde{n}_x\tilde{n}_y}\sum_{i=1}^{\tilde{n}_x}\sum_{j=1}^{\tilde{n}_y} d_\beta(X_i^{\mathrm{fit}}, Y_j^{\mathrm{fit}}) - \frac{1}{2\tilde{n}_x^2}\sum_{i,i'=1}^{\tilde{n}_x} d_\beta(X_i^{\mathrm{fit}}, X_{i'}^{\mathrm{fit}}) - \frac{1}{2\tilde{n}_y^2}\sum_{j,j'=1}^{\tilde{n}_y} d_\beta(Y_j^{\mathrm{fit}}, Y_{j'}^{\mathrm{fit}}),$$

$$\hat{f}_{\mathrm{ED}}(z) = \frac{g_{\mathrm{ED}}(z)}{\sqrt{q_{\mathrm{ED}}}}.$$

The normalizer q_{ED} is half the empirical energy distance with within-group divisors $\tilde{n}_x^2$ and $\tilde{n}_y^2$.

KFDA witness. The original Fisher criterion seeks a function whose squared difference of group means is large relative to its within-group variation. KFDA applies this criterion in a kernel space. We use a regularized version and compute the resulting witness through a linear system.

For candidate (c, λ), set $\sigma = c\,s(\mathcal{X}_{\mathrm{fit}}, \mathcal{Y}_{\mathrm{fit}})$. Write the pooled fitting sample as $Z = (X_1^{\mathrm{fit}}, \ldots, X_{\tilde{n}_x}^{\mathrm{fit}}, Y_1^{\mathrm{fit}}, \ldots, Y_{\tilde{n}_y}^{\mathrm{fit}})$ and its Gaussian kernel matrix as K, where $K_{ij} = k_\sigma(Z_i, Z_j)$. Let $\pi_x = n_x/(n_x + n_y)$ denote the full-sample proportion of observations from P, corresponding to r in the theoretical analysis. Let I_k denote the $k \times k$ identity matrix and $\mathbf{1}_k$ the vector of k ones. Define the group-mean weights δ, centering matrix C_k, and groupwise centering-and-scaling matrix

H by

$$\delta = \begin{pmatrix} \tilde{n}_x^{-1}\mathbf{1}_{\tilde{n}_x} \\ -\tilde{n}_y^{-1}\mathbf{1}_{\tilde{n}_y} \end{pmatrix},$$
$$C_k = I_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^\top,$$
$$H = \begin{pmatrix} \dfrac{C_{\tilde{n}_x}}{2\pi_x\tilde{n}_x} & 0 \\ 0 & \dfrac{C_{\tilde{n}_y}}{2(1-\pi_x)\tilde{n}_y} \end{pmatrix}.$$

Let $\gamma \in \mathbb{R}^{\tilde{n}_x+\tilde{n}_y}$ be the coefficients of a witness expanded over kernels centered at the fitting observations. Its regularized Fisher ratio, denoted by $J_\lambda(\gamma)$, is

$$J_\lambda(\gamma) = \frac{(\gamma^\top K\delta)^2}{\gamma^\top KHK\gamma + \lambda\gamma^\top K\gamma}.$$

The numerator is the squared difference of witness means. The denominator combines the two within-group variances with weights $1/(2\pi_x)$ and $1/\{2(1-\pi_x)\}$, respectively, and adds a penalty on the witness's squared RKHS norm.

A maximizing direction can be obtained from the normal equation

$$(KHK + \lambda K)\gamma = K\delta.$$

In our implementation, we obtain the coefficients by solving

$$(HK + \lambda I_{\tilde{n}_x+\tilde{n}_y})\gamma = \delta.$$

Multiplying this system by K recovers the preceding equation, so it yields the same witness without inverting K. The fitted witness $\hat{f}_{\mathrm{KFDA}}$ is

$$\hat{f}_{\mathrm{KFDA}}(z) = \sum_{i=1}^{\tilde{n}_x+\tilde{n}_y} \gamma_i k_\sigma(z, Z_i).$$

The empirical covariance operator induced by H equals $\hat{C}_{r,F}/2$, where $r = \pi_x$ and $\hat{C}_{r,F}$ is the fitting-sample covariance operator used in the power analysis. Here λ denotes the implementation regularization parameter. The corresponding regularization parameter in the theoretical analysis is 2λ, and the implementation witness is twice the theoretical witness evaluated at that parameter value. When $n_x = n_y$, the two diagonal blocks of H reduce to $C_{\tilde{n}_x}/\tilde{n}_x$ and $C_{\tilde{n}_y}/\tilde{n}_y$, respectively. Bandwidth and regularization are selected on the fitting subsample by the cross-validation procedure below.

PWD witnesses. PWD learns a low-dimensional projection before fitting a scalar witness. Let $\mathcal{G} = \mathcal{K} \times \mathcal{R}$ be its parameter grid. For each pair (k, ρ), a matrix $U \in \mathbb{R}^{d\times k}$ with orthonormal columns defines the projection $z \mapsto U^\top z$, with $U^\top U = I_k$.

A coupling matrix $\Pi \in \mathbb{R}^{\tilde{n}_x\times\tilde{n}_y}$ assigns mass from one fitting group to the other: its entry Π_{ij} is the mass matched from X_i^{fit} to Y_j^{fit}. Equal mass per observation gives the feasible set $\mathcal{U}_{\tilde{n}_x,\tilde{n}_y}$:

$$\mathcal{U}_{\tilde{n}_x,\tilde{n}_y} = \left\{ \Pi \geq 0 : \begin{matrix} \Pi\mathbf{1}_{\tilde{n}_y} = \tilde{n}_x^{-1}\mathbf{1}_{\tilde{n}_x}, \\ \Pi^\top\mathbf{1}_{\tilde{n}_x} = \tilde{n}_y^{-1}\mathbf{1}_{\tilde{n}_y} \end{matrix} \right\}.$$

For a fixed projection, the transport distance $W(U)$ is the minimum total matching cost:

$$W(U) = \min_{\Pi\in\mathcal{U}_{\tilde{n}_x,\tilde{n}_y}} \sum_{i=1}^{\tilde{n}_x}\sum_{j=1}^{\tilde{n}_y} \Pi_{ij}\,\|U^\top(X_i^{\mathrm{fit}} - Y_j^{\mathrm{fit}})\|_2.$$

To find a projection that separates the groups while encouraging sparse weights, we penalize the sum of absolute entries, denoted by $\|U\|_1$, and solve

$$\max_{U^\top U = I_k} \quad W(U) - \rho\|U\|_1.$$

On the fitted projection $\widehat{U}_{k,\rho}$, a scalar network $f_\omega : \mathbb{R}^k \to \mathbb{R}$, with parameters ω, learns the witness by minimizing the loss $L(\omega)$:

$$L(\omega) = \frac{1}{\tilde{n}_x} \sum_{i=1}^{\tilde{n}_x} f_\omega(\widehat{U}_{k,\rho}^\top X_i^{\text{fit}}) - \frac{1}{\tilde{n}_y} \sum_{j=1}^{\tilde{n}_y} f_\omega(\widehat{U}_{k,\rho}^\top Y_j^{\text{fit}}).$$

With $\widehat{\omega}$ denoting the fitted network parameters, the PWD witness $\hat{f}_{k,\rho}$ is

$$\hat{f}_{k,\rho}(z) = -f_{\widehat{\omega}}(\widehat{U}_{k,\rho}^\top z).$$

The sign aligns the witness with a positive P-minus-Q mean difference. Each network has three ReLU hidden layers of width 50, 150, or 250 for $k = 1, 5$, or 10, respectively, followed by a scalar linear output. All linear layers use spectral normalization. With the projection fixed, we train for 1,000 full-batch Adam steps at learning rate 10^{-3} and moving-average coefficients $(0.5, 0.999)$. A separate fit for each admissible parameter pair yields nine witnesses on the full grid, all retained for joint testing.

Candidate parameters. MMD and KFDA use a Gaussian kernel whose bandwidth σ sets the scale of similarity. For fitting samples A and B, let $s(A, B)$ be the median positive pairwise Euclidean distance in the pooled sample. A candidate multiplier c determines the bandwidth through

$$\sigma = c\, s(A, B).$$

The median is recomputed for each cross-validation fit and for the final fit on the entire fitting subsample. The selected multiplier therefore defines a bandwidth relative to the distance scale of each fitting sample.

Alongside c, we select the ED distance exponent β, KFDA regularization strength λ, PWD projection dimension k, and sparsity penalty ρ. Denote the candidate sets for $(c, \beta, \lambda, k, \rho)$ by $(\mathcal{C}, \mathcal{E}, \mathcal{L}, \mathcal{K}, \mathcal{R})$, respectively:

$$\mathcal{C} = \{0.5, 1, 2\}, \qquad \mathcal{E} = \{0.5, 1, 1.5\},$$
$$\mathcal{L} = \{10^{-4}, 10^{-3}, 10^{-2}\},$$
$$\mathcal{K} = \{1, 5, 10\}, \qquad \mathcal{R} = \{0.01, 0.1, 1\}.$$

Each method uses these candidates as follows, with CV denoting cross-validation.

Method	Candidate parameters	Treatment
MMD	$c \in \mathcal{C}$	Select one by five-fold CV
ED	$\beta \in \mathcal{E}$	Select one by five-fold CV
KFDA	$(c, \lambda) \in \mathcal{C} \times \mathcal{L}$	Select one by five-fold CV
PWD	$(k, \rho) \in \mathcal{K} \times \mathcal{R}$	Retain all admissible pairs

The grids are fixed, with PWD projection dimensions restricted to $k \leq d$.

Remark B.1. *Following Hu & Lin (2026), we retain the witnesses obtained from all candidate hyperparameter settings for PWD, since cross-validation would require repeatedly solving the PWD optimization problem. For the other constituent methods, hyperparameter tuning is substantially cheaper, so we use cross-validation to select a single witness.*

Selecting and refitting witnesses. Parameter selection favors witnesses whose validation mean difference is large relative to its variability. For MMD, ED, and KFDA, we divide each fitting group randomly into five folds of nearly equal size. Let $I_{X,\ell}$ and $I_{Y,\ell}$ index fold ℓ, for $\ell = 1, \ldots, 5$. Each witness is fitted on four folds and scored on the fifth, using a common partition across a method's candidates.

A candidate θ is c for MMD, β for ED, or (c, λ) for KFDA. Write $h_{\theta,-\ell}$ for the witness fitted without fold ℓ. On the validation fold, let $r_{X,\ell} = |I_{X,\ell}|$ and $r_{Y,\ell} = |I_{Y,\ell}|$ be the group sizes, $\bar{f}_{X,\ell}, \bar{f}_{Y,\ell}$ the

witness means, and $v_{X,\ell}, v_{Y,\ell}$ their empirical variances:

$$v_{X,\ell} = \frac{1}{r_{X,\ell}} \sum_{i \in I_{X,\ell}} \{h_{\theta,-\ell}(X_i^{\text{fit}}) - \bar{f}_{X,\ell}\}^2,$$
$$v_{Y,\ell} = \frac{1}{r_{Y,\ell}} \sum_{j \in I_{Y,\ell}} \{h_{\theta,-\ell}(Y_j^{\text{fit}}) - \bar{f}_{Y,\ell}\}^2.$$

The fold score $S_\ell(\theta)$ divides the difference of validation means by its estimated standard error. Its five-fold average $\mathrm{CV}(\theta)$ determines the selected candidate $\widehat{\theta}$:

$$S_\ell(\theta) = \frac{\bar{f}_{X,\ell} - \bar{f}_{Y,\ell}}{\sqrt{\max\{v_{X,\ell}/r_{X,\ell} + v_{Y,\ell}/r_{Y,\ell},\, 10^{-12}\}}},$$
$$\mathrm{CV}(\theta) = \frac{1}{5}\sum_{\ell=1}^{5} S_\ell(\theta),$$
$$\widehat{\theta} \in \arg\max_{\theta} \mathrm{CV}(\theta).$$

Refitting the selected candidates on the entire fitting subsample gives one witness each for MMD, ED, and KFDA. Together with the PWD family, these form the collection $\mathcal{H}$:

$$\mathcal{H} = \{\hat{f}_{\mathrm{MMD}}, \hat{f}_{\mathrm{ED}}, \hat{f}_{\mathrm{KFDA}}\} \cup \{\hat{f}_{k,\rho} : (k,\rho) \in \mathcal{G}\}.$$

The collection $\mathcal{H}$ is the output of witness fitting and remains fixed for all subsequent tests.

B.3 Individual constituent methods

On the held-out sample, each fitted witness contributes a difference of group means. MMD, ED, and KFDA each test a single difference; PWD tests its family of differences jointly. We use significance level $\alpha = 0.05$ throughout and reject when the p-value is below α.

Scalar calibration for MMD, ED, and KFDA. For method $j \in \{\mathrm{MMD}, \mathrm{ED}, \mathrm{KFDA}\}$, denote the testing-sample witness means by $\bar{f}_{X,j}, \bar{f}_{Y,j}$ and their unbiased sample variances by $\hat{v}_{X,j}, \hat{v}_{Y,j}$. The standardized mean difference $\widehat{Z}_j$ gives a one-sided p-value p_j through the standard normal cumulative distribution function Φ:

$$\widehat{Z}_j = \frac{\bar{f}_{X,j} - \bar{f}_{Y,j}}{\sqrt{\max\{\hat{v}_{X,j}/\bar{n}_x + \hat{v}_{Y,j}/\bar{n}_y,\, 10^{-12}\}}},$$
$$p_j = 1 - \Phi(\widehat{Z}_j).$$

The individual PWD test. For PWD, let $\hat{f}_{\mathrm{PWD}}$ collect the $\hat{f}_{k,\rho}$ values, ordered by ρ and then k. Its dimension $|\mathcal{G}|$ is the number of parameter pairs. The corresponding p-value is computed following Hu & Lin (2026).

B.4 Naive Aggregation across methods

Write (p_1, p_2, p_3, p_4) for the MMD, ED, KFDA, and PWD p-values. The Bonferroni rule (Bonf) adjusts their minimum, while the equally weighted Cauchy rule combines their tangent transforms into a statistic C:

$$p_{\mathrm{Bonf}} = \min\left\{1,\, 4 \min_{1 \le j \le 4} p_j\right\},$$
$$C = \frac{1}{4}\sum_{j=1}^{4} \tan\{\pi(1/2 - p_j)\},$$
$$p_{\mathrm{Cauchy}} = \frac{1}{2} - \frac{1}{\pi}\arctan(C).$$

For completeness, Fisher combines log p-values into a statistic F, and Stouffer combines normal scores into a statistic S. Let χ^2_8 denote a chi-squared random variable with eight degrees of freedom and Φ^{-1} the standard normal quantile function. Their combinations are

$$F = -2\sum_{j=1}^{4} \log p_j,$$
$$p_{\text{Fisher}} = \Pr\{\chi^2_8 \geq F\},$$
$$S = \frac{1}{2}\sum_{j=1}^{4} \Phi^{-1}(1-p_j),$$
$$p_{\text{Stouffer}} = 1 - \Phi(S).$$

All four rules act directly on the individual p-values. Fisher and Stouffer use their standard independence-based reference distributions.

B.5 Proposed dependence-aware aggregation

Define their combined vector $\hat{f}$ by

$$\hat{f}(z) = \begin{pmatrix} \hat{f}_{\text{PWD}}(z) \\ \hat{f}_{\text{MMD}}(z) \\ \hat{f}_{\text{ED}}(z) \\ \hat{f}_{\text{KFDA}}(z) \end{pmatrix} \in \mathbb{R}^{|\mathcal{G}|+3}.$$

The PWD block contributes its entire family, followed by one witness from each remaining method. The combined dimension m is

$$m = |\mathcal{G}| + 3 = 12$$

for the full grid.

Let A_i, B_j collect their values on the testing observations, with group averages $\bar{A}, \bar{B}$:

$$A_i = \hat{f}(X_i^{\text{test}}), \qquad B_j = \hat{f}(Y_j^{\text{test}}),$$
$$\bar{A} = \frac{1}{\bar{n}_x}\sum_{i=1}^{\bar{n}_x} A_i, \qquad \bar{B} = \frac{1}{\bar{n}_y}\sum_{j=1}^{\bar{n}_y} B_j.$$

The scaled mean-difference vector w is

$$w = \sqrt{\frac{\bar{n}_x \bar{n}_y}{\bar{n}_x + \bar{n}_y}}(\bar{A} - \bar{B}), \qquad n_{\text{eff}} = \frac{\bar{n}_x \bar{n}_y}{\bar{n}_x + \bar{n}_y}.$$

Within-group covariance estimates $\hat{C}_X, \hat{C}_Y$ give the covariance estimate $\hat{\Sigma}$ for w:

$$\hat{C}_X = \frac{1}{\bar{n}_x}\sum_{i=1}^{\bar{n}_x}(A_i - \bar{A})(A_i - \bar{A})^\top,$$
$$\hat{C}_Y = \frac{1}{\bar{n}_y}\sum_{j=1}^{\bar{n}_y}(B_j - \bar{B})(B_j - \bar{B})^\top, \tag{6}$$
$$\hat{\Sigma} = \frac{\bar{n}_y}{\bar{n}_x + \bar{n}_y}\hat{C}_X + \frac{\bar{n}_x}{\bar{n}_x + \bar{n}_y}\hat{C}_Y.$$

Let $\hat{\Sigma}^+$ denote the Moore–Penrose pseudoinverse of $\hat{\Sigma}$, obtained by inverting its nonzero eigenvalues. We use its symmetric square root R for covariance adjustment and the diagonal standard deviations s_r for componentwise scaling:

$$R = (\hat{\Sigma}^+)^{1/2},$$
$$s_r = \sqrt{\max\{\hat{\Sigma}_{rr}, 10^{-12}\}}, \qquad r = 1, \ldots, q.$$

Table 5: Empirical Type-I error rates of more baselines and our methods.

Model	$\mathrm{MMD}_{\mathrm{full}}$	$\mathrm{ED}_{\mathrm{full}}$	MMDAgg	MMD-FUSE	Max-Cov	Max-Var	Hotelling
Gaussian	0.023	0.033	0.020	0.037	0.067	0.047	0.070
Gaussian mixture	0.060	0.050	0.043	0.050	0.067	0.057	0.060

Here $\hat{\Sigma}_{rr}$ is the rth diagonal entry. For a vector u, we write $\|u\|_\infty = \max_r |u_r|$.

We define the Max-Cov, Max-Var, and Hotelling statistics, denoted by T_n, T_n'', and T_n', as

$$T_n = \|Rw\|_\infty, \qquad \text{(Max-Cov)},$$
$$T_n'' = \max_{1\le r\le m} \frac{w_r}{s_r}, \quad \text{(Max-Var)},$$
$$T_n' = w^\top \hat{\Sigma}^+ w, \qquad \text{(Hotelling)}.$$

Remark B.2. *In implementation, we use the Moore–Penrose pseudoinverse to improve numerical stability. This does not affect the asymptotic theory, since $\mathbb{P}(\hat{\Sigma} \succ 0) \to 1$, and the pseudoinverse coincides with the ordinary inverse whenever $\hat{\Sigma}$ is positive definite.*

C Additional Numerical Results

C.1 Comparisons with other methods

We additionally compare our method with full-sample MMD (Gretton et al., 2012) and ED (Székely & Rizzo, 2004), as well as MMDAgg (Schrab et al., 2023) and MMD-FUSE (Biggs et al., 2023). The full-sample MMD ($\mathrm{MMD}_{\mathrm{full}}$) uses a Gaussian kernel with bandwidth selected by the median heuristic , while the full-sample ED ($\mathrm{ED}_{\mathrm{full}}$) uses the classical energy-distance statistic. MMDAgg and MMD-FUSE are implemented using the default procedures recommended in their original papers. By aggregating multiple kernels within the MMD framework, MMDAgg and MMD-FUSE provide a complementary comparison with our cross-method witness aggregation.

Table 5 shows that all methods maintain satisfactory Type-I error control under both null models. As shown in Table 6, the $\mathrm{MMD}_{\mathrm{full}}$, $\mathrm{ED}_{\mathrm{full}}$, MMDAgg and MMD-FUSE can achieve high power in some settings but remain sensitive to the structure of the alternative. In contrast, Max-Cov and Hotelling exhibit more stable performance across models and regimes, with particularly large gains in Models A, D, and E and more robust power across sparsity levels in Model B. These results further support the benefit of aggregating heterogeneous witnesses for robust performance across diverse alternatives.

C.2 More results on real data

We report additional MNIST results for $\pi = 0$ in Table 7 and for $\pi = 0.15$ in Table 8. All considered methods maintain satisfactory Type-I error control, and the power results exhibit patterns similar to those observed for $\pi = 0.1$.

D Power Guarantee for Specific Constituent Witnesses

The constituent witnesses considered in numerical experiments are derived from PWD, ED, MMD, and KFDA. We first introduce the notation used in this section.

For PWD, we consider the ℓ_1-constrained projection Wasserstein distance and estimate the corresponding witness using the same algorithm as in Hu & Lin (2026). Let $\mathcal{N}(D, \mathbf{p}, s, B)$ denote the class of ReLU networks with D hidden layers, width vector $\mathbf{p}$, at most s nonzero weights and biases, all weights and biases bounded in absolute value by one, and outputs uniformly bounded in absolute value by B. Let $\mathcal{C}^{(1)}$ be a nonempty finite set of candidate pairs (k, τ), where $1 \le k \le d$ and $k \le \tau \le k\sqrt{d}$. Denote by $PW^{(1)}_{k,\tau}(P, Q)$ the PWD with the additional constraint $\sum_{j,l} |U_{jl}| \le \tau$.

Table 6: Empirical power of more baselines and our methods.

Model	Setting	$\mathrm{MMD}_{\mathrm{full}}$	$\mathrm{ED}_{\mathrm{full}}$	MMDAgg	MMD-FUSE	Max-Cov	Max-Var	Hotelling
A	$r = 1$	0.050	0.063	0.317	0.167	0.970	0.637	0.987
	$r = 5$	0.087	0.093	0.890	0.547	0.970	0.433	0.997
	$r = 10$	0.173	0.167	0.993	0.877	1.000	0.380	1.000
	$r = 30$	0.563	0.563	1.000	1.000	1.000	0.987	1.000
	$r = 100$	1.000	1.000	0.990	1.000	1.000	1.000	1.000
B	$\tau = 0.1$	0.040	0.033	0.413	0.293	0.943	0.953	0.957
	$\tau = 0.3$	0.050	0.053	0.897	0.863	0.850	0.880	0.907
	$\tau = 0.5$	0.103	0.113	0.997	1.000	0.843	0.483	0.923
	$\tau = 0.7$	0.380	0.397	0.997	1.000	1.000	0.943	1.000
	$\tau = 0.9$	0.997	0.997	1.000	1.000	1.000	1.000	1.000
C	$\tau = 0.1$	0.900	0.903	0.857	0.903	1.000	1.000	1.000
	$\tau = 0.3$	0.867	0.873	0.843	0.873	0.710	0.927	0.820
	$\tau = 0.5$	0.840	0.827	0.820	0.843	0.373	0.697	0.467
	$\tau = 0.7$	0.830	0.833	0.797	0.827	0.217	0.557	0.353
	$\tau = 0.9$	0.850	0.850	0.800	0.840	0.260	0.493	0.310
D	$\beta = 0.2$	0.060	0.057	0.067	0.057	0.670	0.510	0.803
	$\beta = 0.5$	0.063	0.063	0.453	0.143	0.997	0.970	1.000
E	$\beta = 0.2$	0.033	0.033	0.053	0.033	0.893	0.277	0.937
	$\beta = 0.5$	0.033	0.037	0.127	0.047	1.000	0.463	1.000

Table 7: Empirical Type-I error rates on MNIST for $\pi = 0$.

n_x	σ	Single method				Naive aggregation		Proposed		
		MMD	ED	KFDA	PWD	Bonf	Cauchy	Max-Cov	Max-Var	Hotelling
350	0	0.057	0.040	0.060	0.053	0.040	0.060	0.053	0.060	0.070
500	0	0.070	0.063	0.053	0.063	0.043	0.057	0.057	0.057	0.040
800	0	0.043	0.043	0.047	0.050	0.033	0.047	0.077	0.050	0.060

For ED, we specialize to $\rho_p(x, y) = \|x - y\|_2^p$ for a fixed $p \in (0, 2)$, and denote the resulting discrepancy by $ED_p(P, Q)$.

Denote the positive-definite kernels used in MMD and KFDA by k and K_F, respectively, and let $\mathcal{H}_k$ and $\mathcal{H}_{K_F}$ be their corresponding separable RKHSs. For KFDA, let $\phi_F(x) = K_F(x, \cdot)$, $r = n_x/n$, and define

$$C_{r,F} = \frac{\mathrm{Cov}_P\{\phi_F(X)\}}{r} + \frac{\mathrm{Cov}_Q\{\phi_F(Y)\}}{1 - r}.$$

Let $D_{\mathrm{KFDA},\lambda}(P, Q)$ denote the KFDA optimization value defined in Section 2. Finally, let $\Sigma^{(1)}$ denote the principal submatrix of Σ corresponding to the PWD witnesses.

Assumption D.1. *There exist constants $0 < c_1' \le c_2' < \infty$ such that*

$$c_1' I \preceq \Sigma^{(1)} \preceq c_2' I$$

almost surely with respect to the fitting data.

Assumption D.2. *There exists a constant $0 < \eta < \infty$ such that*

$$\sup_{\|u\|_2=1} \|u^\top X\|_{\psi_2} \le \eta, \qquad \sup_{\|u\|_2=1} \|u^\top Y\|_{\psi_2} \le \eta,$$

where $X \sim P$, $Y \sim Q$, and $\|W\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\exp(W^2/t^2) \le 2\right\}$.

Table 8: Empirical power on MNIST for $\pi = 0.15$.

n_x	σ	Single method				Naive aggregation		Proposed		
		MMD	ED	KFDA	PWD	Bonf	Cauchy	Max-Cov	Max-Var	Hotelling
350	80	0.423	0.430	0.740	0.493	0.597	0.667	0.780	0.493	0.847
	100	0.707	0.723	0.843	0.700	0.773	0.860	0.820	0.663	0.943
	120	0.880	0.877	0.907	0.777	0.907	0.940	0.830	0.843	0.960
500	80	0.670	0.673	0.963	0.643	0.883	0.927	0.933	0.823	0.973
	100	0.883	0.880	0.983	0.893	0.970	0.987	0.957	0.933	0.997
	120	0.977	0.977	0.993	0.933	0.990	0.993	0.963	0.990	1.000
800	80	0.900	0.903	1.000	0.940	1.000	1.000	0.990	1.000	1.000
	100	0.990	0.983	1.000	0.997	1.000	1.000	0.997	1.000	1.000
	120	1.000	1.000	1.000	0.997	1.000	1.000	1.000	1.000	1.000

Assumption D.3. *Assume*

$$\sup_{x\in\mathcal{X}} k(x,x) \le 1, \qquad \sup_{x\in\mathcal{X}} K_F(x,x) \le 1.$$

For deterministic $\lambda > 0$ and $N \ge 1$, possibly depending on n,

$$N_F(\lambda) = \operatorname{tr}\left\{C_{r,F}(C_{r,F} + \lambda I)^{-1}\right\} \le N, \qquad n\lambda^2 \to \infty.$$

Set $\theta_n = \eta\sqrt{2\log(dn)}$ and define

$$R_n^{(1)}(k,\tau,d) = \tau k^{1/2}\{\log(dn)\}^{3/2}\left\{n^{-\frac{1}{k+2}} + n^{-1/2}(dk)^{1/2} + n^{-\frac{2}{k+2}}\log n + n^{-1}dk\log n\right\},$$

$$R_{n,M} = \sqrt{\frac{\log\log n}{n}}, \qquad R_{n,E,p} = \eta^{p/2}d^{p/4}\sqrt{\frac{\log\log n}{n}}, \qquad R_{n,F} = \frac{N^{1/4}}{\sqrt{1+\lambda}}\sqrt{\frac{\log\log n}{n}}.$$

Here N is the effective-dimension bound in Assumption D.3. Let $q_{1-\alpha}$ be the $(1-\alpha)$-quantile of $\max_{1\le j\le m}|Z_j|$, where $Z \sim N(0, I_m)$.

Theorem D.1. *Suppose that Assumptions 1, 2, D.1–D.3 hold. Use the fitting-sample versions of the MMD, ED and KFDA witnesses defined in Section 3.1. For each $(k,\tau) \in \mathcal{C}^{(1)}$, consider the ReLU network class*

$$\mathcal{N}\left(D, (k, 2k, 6(k+1)N_{\mathrm{NN}}, \ldots, 6(k+1)N_{\mathrm{NN}}, 1), s, 4k^{1/2}\tau\theta_n\right),$$

where

$$D \asymp \log n, \qquad s \asymp N_{\mathrm{NN}}D, \qquad N_{\mathrm{NN}} \ge 2^k \vee (4k^{1/2}\tau\theta_n + 1)e^k,$$

and

$$N_{\mathrm{NN}} \asymp n^{\frac{k}{2+k}}(\log n)^{-\frac{3k}{2+k}}.$$

Construct T_n from these $m = |\mathcal{C}^{(1)}| + 3$ fitted witnesses. If there exists $(k_\star, \tau_\star) \in \mathcal{C}^{(1)}$ such that, for all sufficiently large n,

$$\max\left\{\frac{PW^{(1)}_{k_\star,\tau_\star}(P,Q)}{R_n^{(1)}(k_\star,\tau_\star,d)}, \frac{\mathrm{MMD}(P,Q)}{R_{n,M}}, \frac{\sqrt{ED_p(P,Q)/2}}{R_{n,E,p}}, \frac{D_{\mathrm{KFDA},\lambda}(P,Q)}{R_{n,F}}\right\} \ge c_a$$

for some sufficiently large constant $c_a > 0$, then

$$\mathbb{P}(T_n > q_{1-\alpha}) \to 1 \qquad \text{as } n \to \infty.$$

Theorem D.1 specializes the generic power guarantee in Theorem 2 to the four constituent tests used in our procedure. It provides explicit detection conditions for PWD, MMD, ED, and KFDA through the rates $R_n^{(1)}$, $R_{n,M}$, $R_{n,E,p}$, and $R_{n,F}$, respectively. Consistency is guaranteed as long as at least one constituent discrepancy exceeds its corresponding detection rate by a sufficiently large constant factor. Thus, the adaptive test inherits the strength of the most informative constituent witness and does not require all component tests to be powerful simultaneously.

Remark D.3 (KFDA implementation and theoretical scaling). *Let λ_{impl} be the KFDA regularization parameter in Appendix B, and set $\lambda = 2\lambda_{\mathrm{impl}}$ in Theorem D.1. The implementation covariance operator is $\hat{C}_{r,F}/2$. Consequently, its fitted KFDA witness is twice the KFDA witness analyzed in the proof of Theorem D.1.*

This scaling also preserves the stated power conclusion. If the KFDA coordinate is multiplied by two, its conditional mean difference is multiplied by two and its conditional variance by four. Thus the ratio $v_j/\Delta_{\hat{f}_j}(P,Q)^2$ in Lemma G.1 is unchanged. The full estimated covariance matrix undergoes a congruence transformation by an invertible diagonal matrix, so its positive definiteness is unchanged. The normalized moment and eigenvalue conditions in Assumption 2 are likewise unchanged. Hence the KFDA non-rejection bounds, and therefore the conclusion of Theorem D.1, also apply to the implementation witness, with $\lambda = 2\lambda_{\mathrm{impl}}$.

E Proof of Theorem 1

Notation. Throughout, $n = n_x + n_y$, and $\bar{n} = \bar{n}_x + \bar{n}_y$ is the size of the pooled testing subsample, so that $\bar{n} \le n$. The conditional expectation $\mathbb{E}_\star$ and variance $\mathrm{Var}_\star$ are as in Section 4, and $\mathbb{P}_\star$ denotes probability conditional on $\mathcal{D}_{\mathrm{fit}}$. As in Section 3.2, $\hat{f} = (\hat{f}_1, \dots, \hat{f}_m)^\top$ denotes the vector of learned witnesses, and Σ is the matrix defined in Section 4. Conditionally on $\mathcal{D}_{\mathrm{fit}}$, the witnesses $\hat{f}$ are fixed functions and the testing observations are independent of them. The sample covariance matrices $\hat{C}_X$ and $\hat{C}_Y$ are centered at their own testing-sample means and normalized by $\bar{n}_x$ and $\bar{n}_y$, respectively. Put

$$n_{\mathrm{eff}} = \frac{\bar{n}_x \bar{n}_y}{\bar{n}}, \qquad \text{so that} \qquad T_n = \sqrt{n_{\mathrm{eff}}}\, \big\| \hat{\Sigma}^{-1/2} \hat{S}_n \big\|_\infty .$$

The argument below uses the inverse square root only on an event on which $\hat{\Sigma} \succ 0$. For vectors, $\|\cdot\|_2$ denotes the Euclidean norm; for matrices, $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_{\mathrm{op}}$ denote the Frobenius and operator norms. All matrix square roots are symmetric. Constants denoted by C may change from line to line but do not depend on n.

Outline. Let $W = \sqrt{n_{\mathrm{eff}}}\, \Sigma^{-1/2} \hat{S}_n$ be the mean difference whitened by the true Σ, and let $T_n^0 = \|W\|_\infty$. The proof compares T_n with T_n^0, and T_n^0 with $G = \max_j |Z_j|$.

The second comparison concerns a normalized sum of independent vectors. It follows from a multivariate Berry–Esseen bound (Bentkus, 2005).

The first comparison is different in kind. T_n differs from T_n^0 only through the estimation error of $\hat{\Sigma}$, which is bounded by a random bandwidth D_n. The comparison therefore reduces to showing that T_n^0 rarely falls within distance D_n of any fixed level. This is handled by a randomized concentration inequality (Chen et al., 2010).

Lemmas E.1 and E.2 supply the matrix perturbation bound and the moment bounds, and Lemma E.3 combines them.

Proof of Theorem 1. Since $\mathbb{P}(T_n \le t) = \mathbb{E}\{\mathbb{P}_\star(T_n \le t)\}$ and a difference of probabilities is at most one in absolute value, the triangle inequality and (16)–(17) in Lemma E.3 give

$$\begin{aligned}
&\sup_{t \in \mathbb{R}} |\mathbb{P}(T_n \le t) - \mathbb{P}(G \le t)| \\
&\le \mathbb{E}\left[1 \wedge \left\{ \sup_t |\mathbb{P}_\star(T_n \le t) - \mathbb{P}_\star(T_n^0 \le t)| + \sup_t |\mathbb{P}_\star(T_n^0 \le t) - \mathbb{P}(G \le t)| \right\} \right] \\
&\le \mathbb{E}[1 \wedge \{C(x_n + x_n^2)\}],
\end{aligned}$$

where x_n is defined in Lemma E.3. If $x_n \le 1$, then $C(x_n + x_n^2) \le 2Cx_n$; if $x_n > 1$, then $1 \wedge \{C(x_n + x_n^2)\} \le 1 = 1 \wedge (C' x_n)$ for any $C' \ge 1$. Hence $\mathbb{E}[1 \wedge \{C(x_n + x_n^2)\}] \le \mathbb{E}[1 \wedge (C' x_n)]$ with $C' = \max(2C, 1)$.

By (18), almost surely, $x_n \le \kappa^{-1} m \sqrt{L_n/n}$. Taking $c = C' \kappa^{-1} m$ gives

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}(T_n \le t) - \mathbb{P}\left(\max_{1 \le j \le m} |Z_j| \le t \right) \right| \le \min\left\{ 1, c\sqrt{L_n/n} \right\}.$$

The constant c depends only on m, c_1, c_2, κ. Since $L_n/n \to 0$, the bound tends to zero, proving the asserted convergence in distribution. □

Remark E.4. *If the covariance condition in Assumption 2 of Hu & Lin (2026) and the exponential-moment condition in Theorem 1 of Hu & Lin (2026) are imposed on our full witness vector, they read*

$$c_1' I_m \preceq \Sigma \preceq c_2' I_m, \qquad \max_{1\le j\le m} \mathbb{E}_\star e^{\hat{f}_j(X)^2/M_n^2} \le 2, \qquad \max_{1\le j\le m} \mathbb{E}_\star e^{\hat{f}_j(Y)^2/M_n^2} \le 2$$

almost surely, for constants $0 < c_1' \le c_2' < \infty$ and deterministic $M_n > 0$. Together with Assumption 1, these imply the covariance and fourth-moment bounds in Assumption 2, with $\kappa = c_1'/c_2'$ and $L_n = CM_n^2 \log(e + M_n)$, where C depends only on c_1, c_2, c_1', c_2'.

Justification of Remark E.4. Throughout, condition on the fitting data. The covariance bounds imply $c_1' \le v_j \le c_2'$, and hence

$$V^{-1/2}\Sigma V^{-1/2} \succeq c_1' V^{-1} \succeq \frac{c_1'}{c_2'} I_m.$$

We prove the result only for X, as the argument for Y is analogous. By Assumption 1, both weights in the definition of Σ are bounded away from zero. Therefore,

$$\mathrm{Var}_\star\{\hat{f}_j(X)\} \le C,$$

where C depends only on c_1, c_2, c_2'. The sub-Gaussian tail condition, Jensen's inequality, and $(x - y)^2 \le 2x^2 + 2y^2$ give

$$\mathbb{E}_\star \exp\left\{ \frac{|\hat{f}_j(X) - \mathbb{E}_\star \hat{f}_j(X)|^2}{4M_n^2} \right\} \le 2.$$

Integrating the resulting tail bound, with a cutoff at $4M_n\sqrt{\log(e + M_n)}$, yields

$$\begin{aligned}
&\mathbb{E}_\star |\hat{f}_j(X) - \mathbb{E}_\star \hat{f}_j(X)|^4 \\
&\quad \le 16 M_n^2 \log(e + M_n)\, \mathrm{Var}_\star\{\hat{f}_j(X)\} \\
&\qquad + 8 \int_{4M_n\sqrt{\log(e+M_n)}}^{\infty} t^3 \exp\left(-\frac{t^2}{4M_n^2}\right) dt \\
&\quad \le CM_n^2 \log(e + M_n) + CM_n^4 \{1 + \log(e + M_n)\} e^{-4\log(e+M_n)} \\
&\quad \le CM_n^2 \log(e + M_n).
\end{aligned}$$

Dividing by $v_j^2 \ge (c_1')^2$ proves the required relative fourth-moment bound. □

E.1 Technical lemmas

Lemma E.1. *Let $\Sigma \succ 0$ and let $\hat{\Sigma}$ be symmetric. If $E = \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - I_m$ satisfies $\|E\|_{\mathrm{op}} \le 1/2$, then $\hat{\Sigma} \succ 0$ and*

$$\left\|\hat{\Sigma}^{-1/2}\Sigma^{1/2} - I_m\right\|_{\mathrm{F}} \le \sqrt{2}\, \|E\|_{\mathrm{F}}. \tag{7}$$

Proof. Let $B = \Sigma^{1/2}$. Since $\|E\|_{\mathrm{op}} \le 1/2$, we have $I_m + E \succeq I_m/2$, and hence $\hat{\Sigma} = B(I_m + E)B \succeq B^2/2 \succ 0$. Put $A = \hat{\Sigma}^{1/2}$ and $H = A^{-1}B - I_m$, so that the left-hand side of (7) equals $\|H\|_{\mathrm{F}}$. Since $B^2 - A^2 = \Sigma - \hat{\Sigma} = -BEB$, the matrix H solves the Sylvester equation

$$AH + HB = (B - A) + (A^{-1}B^2 - B) = A^{-1}(B^2 - A^2) = -A^{-1}BEB.$$

Choose orthogonal matrices O_1, O_2 such that $O_1^\top A O_1 = \mathrm{diag}(a_1, \ldots, a_m)$ and $O_2^\top B O_2 = \mathrm{diag}(b_1, \ldots, b_m)$, where all $a_j, b_k > 0$.

Multiply the Sylvester equation by $O_1^\top$ on the left and by O_2 on the right. On the left-hand side, $O_1^\top A = \mathrm{diag}(a_j) O_1^\top$ and $BO_2 = O_2\,\mathrm{diag}(b_k)$, so the (j, k) entry becomes $(a_j + b_k)(O_1^\top H O_2)_{jk}$.

On the right-hand side, $O_1^\top A^{-1}BEBO_2 = (O_1^\top A^{-1}BEO_2)\,\mathrm{diag}(b_k)$, so the (j,k) entry is $-b_k(O_1^\top A^{-1}BEO_2)_{jk}$. Hence

$$(O_1^\top HO_2)_{jk} = -\frac{b_k}{a_j+b_k}(O_1^\top A^{-1}BEO_2)_{jk}.$$

The Frobenius norm is invariant under orthogonal transformations, and $0 < b_k/(a_j+b_k) < 1$. Therefore

$$\begin{aligned}\|H\|_{\mathrm{F}}^2 = \|O_1^\top HO_2\|_{\mathrm{F}}^2 &= \sum_{j,k}\frac{b_k^2}{(a_j+b_k)^2}|(O_1^\top A^{-1}BEO_2)_{jk}|^2 \\ &\le \|O_1^\top A^{-1}BEO_2\|_{\mathrm{F}}^2 = \|A^{-1}BE\|_{\mathrm{F}}^2.\end{aligned}$$

Let $e_1,\dots,e_m$ denote the columns of E. Bounding each quadratic form by the operator norm gives

$$\|A^{-1}BE\|_{\mathrm{F}}^2 = \mathrm{tr}(E^\top BA^{-2}BE) = \sum_k e_k^\top(BA^{-2}B)e_k \le \|BA^{-2}B\|_{\mathrm{op}}\,\|E\|_{\mathrm{F}}^2.$$

Finally, $BA^{-2}B = B\hat\Sigma^{-1}B = B\,B^{-1}(I_m+E)^{-1}B^{-1}B = (I_m+E)^{-1}$, and $\|(I_m+E)^{-1}\|_{\mathrm{op}} \le (1-\|E\|_{\mathrm{op}})^{-1} \le 2$. Combining the last three displays gives (7). □

For the next lemma, let $q_1,\dots,q_{\bar n}$ be independent and identically distributed vectors in $\mathbb{R}^m$, with $\mathbb{E}q_i = 0$, $\mathbb{E}(q_iq_i^\top) = I_m$ and $\beta = \mathbb{E}\|q_1\|_2^4 < \infty$. Partition the pooled index set $\{1,\dots,\bar n\}$ into two disjoint subsets I_x, I_y of sizes $\bar n_x, \bar n_y$, where $\bar n \le n$ and $\min(\bar n_x,\bar n_y) \ge an$ for a fixed $a > 0$. Set

$$\omega_x = \frac{\bar n_y}{\bar n}, \qquad \omega_y = \frac{\bar n_x}{\bar n}, \qquad w_i = \begin{cases}\sqrt{\omega_x/\bar n_x}, & i\in I_x,\\ -\sqrt{\omega_y/\bar n_y}, & i\in I_y.\end{cases}$$

For $s\in\{x,y\}$, write $\bar q_s = \bar n_s^{-1}\sum_{i\in I_s} q_i$, and define

$$\begin{aligned}W &= \sum_i w_iq_i, \qquad x = \sqrt{\beta/n},\\ E_n &= \sum_i w_i^2(q_iq_i^\top - I_m) - \sum_{s\in\{x,y\}}\omega_s\bar q_s\bar q_s^\top.\end{aligned} \tag{8}$$

Let $W^{(i)}$ and $E_n^{(i)}$ be obtained by replacing q_i by zero in these expressions, with all weights and denominators unchanged. Put $D_n = 2\|E_n\|_{\mathrm{F}}\|W\|_2$ and $D_n^{(i)} = 2\|E_n^{(i)}\|_{\mathrm{F}}\|W^{(i)}\|_2$.

The weights encode the two-sample contrast as a single weighted sum. Since $\omega_x\bar n_x = \omega_y\bar n_y = n_{\mathrm{eff}}$, we have $\sum_{i\in I_x} w_i = \sqrt{n_{\mathrm{eff}}} = -\sum_{i\in I_y} w_i$, and $W = \sqrt{n_{\mathrm{eff}}}(\bar q_x - \bar q_y)$. Moreover, using $\omega_s \le 1$ and $\bar n_s \ge an$,

$$\sum_i w_i = 0, \qquad \sum_i w_i^2 = \omega_x+\omega_y = 1, \qquad \max_i w_i^2 \le \frac{1}{an}, \qquad \sum_i w_i^4 \le \max_i w_i^2\sum_i w_i^2 \le \frac{1}{an}. \tag{9}$$

The matrix E_n and the scalar D_n are auxiliary. As shown in the proof of Lemma E.3, E_n is the whitened error of the pooled covariance estimator: $I_m + E_n = \Sigma^{-1/2}\hat\Sigma\Sigma^{-1/2}$. The terms $\omega_s\bar q_s\bar q_s^\top$ appear because each group covariance is centered at its own sample mean. The quantity D_n will serve as the random bandwidth controlling $|T_n - T_n^0|$. Its leave-one-out version $D_n^{(i)}$ is a function of $(q_k)_{k\ne i}$ only, which is what the randomized concentration argument requires.

Lemma E.2. *In the setting above, the following bounds hold with a constant C depending only on m and a:*

$$\mathbb{E}\|W\|_2^4 \le C(1+x^2), \tag{10}$$

$$\mathbb{E}\|E_n\|_{\mathrm{F}}^2 + \max_i \mathbb{E}\|E_n^{(i)}\|_{\mathrm{F}}^2 \le Cx^2, \tag{11}$$

$$\sum_i |w_i|\mathbb{E}\{\|q_i\|_2\,|D_n - D_n^{(i)}|\} \le C(x+x^2), \tag{12}$$

where $x = \sqrt{\beta/n}$ and $\beta = \mathbb{E}\|q_1\|_2^4 < \infty$.

Proof. Vector Rosenthal inequality. Let $h_1, \dots, h_{\bar{n}}$ be real coefficients and $v = \sum_i h_i q_i$. For each coordinate j, the summands $h_i q_{ij}$ are independent with mean zero and $\mathbb{E}(h_i q_{ij})^2 = h_i^2$, because $\mathbb{E}(q_i q_i^\top) = I_m$. Rosenthal's inequality (Rosenthal, 1970, Theorem 3) with $p = 4$ gives

$$\mathbb{E}|v_j|^4 \le C\Big\{\sum_i h_i^4\, \mathbb{E}q_{1j}^4 + \Big(\sum_i h_i^2\Big)^2\Big\}.$$

Since $\|v\|_2^4 = (\sum_j v_j^2)^2 \le m \sum_j v_j^4$ and $\sum_j \mathbb{E}q_{1j}^4 \le \mathbb{E}(\sum_j q_{1j}^2)^2 = \beta$, summing over j yields

$$\mathbb{E}\Big\|\sum_i h_i q_i\Big\|_2^4 \le C_m\Big\{\beta \sum_i h_i^4 + \Big(\sum_i h_i^2\Big)^2\Big\}. \tag{13}$$

Taking the trace of $\mathbb{E}(q_1 q_1^\top) = I_m$ gives $\mathbb{E}\|q_1\|_2^2 = m$. Jensen's inequality then gives $\beta \ge (\mathbb{E}\|q_1\|_2^2)^2 = m^2 \ge 1$, and hence

$$n^{-2} \le n^{-1} \le x^2. \tag{14}$$

Fourth moments of W and $\bar{q}_s$. Taking $h_i = w_i$ in (13) and using (9) gives

$$\mathbb{E}\|W\|_2^4 \le C(1 + \beta/n) = C(1 + x^2),$$

which is (10). Taking $h_i = \bar{n}_s^{-1}$ for $i \in I_s$ and $h_i = 0$ otherwise gives $\sum_i h_i^4 = \bar{n}_s^{-3}$ and $(\sum_i h_i^2)^2 = \bar{n}_s^{-2}$, so

$$\mathbb{E}\|\bar{q}_s\|_2^4 \le C\big(\beta/\bar{n}_s^3 + \bar{n}_s^{-2}\big).$$

Bound on E_n. Put $A_n = \sum_i w_i^2 (q_i q_i^\top - I_m)$, so that $E_n = A_n - \omega_x \bar{q}_x \bar{q}_x^\top - \omega_y \bar{q}_y \bar{q}_y^\top$. The summands of A_n are independent with mean zero, so the cross terms vanish in $\mathbb{E}\|A_n\|_{\mathrm{F}}^2$. Moreover,

$$\|q_i q_i^\top - I_m\|_{\mathrm{F}}^2 = \mathrm{tr}\{(q_i q_i^\top - I_m)^2\} = \|q_i\|_2^4 - 2\|q_i\|_2^2 + m,$$

whose expectation is $\beta - 2m + m = \beta - m$. Hence

$$\mathbb{E}\|A_n\|_{\mathrm{F}}^2 = (\beta - m)\sum_i w_i^4 \le C\beta/n.$$

For a rank-one matrix, $\|\bar{q}_s \bar{q}_s^\top\|_{\mathrm{F}} = \|\bar{q}_s\|_2^2$. Combining this with $(t_1 + t_2 + t_3)^2 \le 3(t_1^2 + t_2^2 + t_3^2)$, $\omega_s \le 1$, $\bar{n}_s \ge an$ and (14) gives

$$\mathbb{E}\|E_n\|_{\mathrm{F}}^2 \le 3\mathbb{E}\|A_n\|_{\mathrm{F}}^2 + 3\sum_s \omega_s^2 \mathbb{E}\|\bar{q}_s\|_2^4 \le C\big(\beta/n + \beta/n^3 + n^{-2}\big) \le Cx^2.$$

Bound on $E_n^{(i)}$. Fix $i \in I_s$, let u be the other group, and put $b_i = \bar{n}_s^{-1} \sum_{k \in I_s \setminus \{i\}} q_k$. Replacing q_i by zero has three effects. The ith summand of A_n becomes $-w_i^2 I_m$, so A_n becomes

$$A_n^{(i)} = \sum_{k \ne i} w_k^2 (q_k q_k^\top - I_m) - w_i^2 I_m.$$

The group mean $\bar{q}_s$ becomes b_i, and $\bar{q}_u$ is unchanged. Thus $E_n^{(i)} = A_n^{(i)} - \omega_s b_i b_i^\top - \omega_u \bar{q}_u \bar{q}_u^\top$.

The random sum in $A_n^{(i)}$ has mean zero and the term $-w_i^2 I_m$ is deterministic, so their cross term vanishes. As above,

$$\mathbb{E}\|A_n^{(i)}\|_{\mathrm{F}}^2 = (\beta - m)\sum_{k \ne i} w_k^4 + m w_i^4 \le C\beta/n.$$

Applying (13) to b_i (which has $\bar{n}_s - 1$ summands with coefficient $\bar{n}_s^{-1}$) gives $\mathbb{E}\|b_i\|_2^4 \le C(\beta/\bar{n}_s^3 + \bar{n}_s^{-2})$. Therefore, exactly as for E_n,

$$\mathbb{E}\|E_n^{(i)}\|_{\mathrm{F}}^2 \le 3\mathbb{E}\|A_n^{(i)}\|_{\mathrm{F}}^2 + 3\omega_s^2 \mathbb{E}\|b_i\|_2^4 + 3\omega_u^2 \mathbb{E}\|\bar{q}_u\|_2^4 \le Cx^2.$$

This proves (11).

Deletion identities. Since $A_n - A_n^{(i)} = w_i^2 q_i q_i^\top = \omega_s \bar{n}_s^{-1} q_i q_i^\top$ and $\bar{q}_s = b_i + \bar{n}_s^{-1} q_i$, we have

$$\bar{q}_s \bar{q}_s^\top - b_i b_i^\top = \bar{n}_s^{-1}(q_i b_i^\top + b_i q_i^\top) + \bar{n}_s^{-2} q_i q_i^\top.$$

Hence

$$E_n - E_n^{(i)} = \omega_s\big\{(\bar{n}_s^{-1} - \bar{n}_s^{-2})q_i q_i^\top - \bar{n}_s^{-1}(q_i b_i^\top + b_i q_i^\top)\big\}. \tag{15}$$

Using $\|q_i q_i^\top\|_{\mathrm{F}} = \|q_i\|_2^2$, $\|q_i b_i^\top\|_{\mathrm{F}} = \|q_i\|_2\,\|b_i\|_2$, $\omega_s \le 1$ and $\bar{n}_s \ge an$, we obtain

$$\|E_n - E_n^{(i)}\|_{\mathrm{F}} \le \frac{C}{n}\big(\|q_i\|_2^2 + \|q_i\|_2\,\|b_i\|_2\big).$$

Also $W - W^{(i)} = w_i q_i$. Writing

$$\tfrac{1}{2}(D_n - D_n^{(i)}) = \big(\|E_n\|_{\mathrm{F}} - \|E_n^{(i)}\|_{\mathrm{F}}\big)\|W\|_2 + \|E_n^{(i)}\|_{\mathrm{F}}\big(\|W\|_2 - \|W^{(i)}\|_2\big)$$

and applying the reverse triangle inequality to each difference gives

$$|D_n - D_n^{(i)}| \le 2\|E_n - E_n^{(i)}\|_{\mathrm{F}}\|W\|_2 + 2|w_i|\,\|E_n^{(i)}\|_{\mathrm{F}}\|q_i\|_2.$$

Moment bounds for the deletion term. Write $\mu_3 = \mathbb{E}\|q_1\|_2^3$. By the Cauchy–Schwarz inequality, $\mu_3 \le (\mathbb{E}\|q_1\|_2^2)^{1/2}(\mathbb{E}\|q_1\|_2^4)^{1/2} = \sqrt{m\beta}$.

The vectors b_i and $W^{(i)}$ are sums of independent mean-zero vectors with identity covariance, so

$$\mathbb{E}\|b_i\|_2^2 = \frac{m(\bar{n}_s - 1)}{\bar{n}_s^2} \le \frac{Cm}{n}, \qquad \mathbb{E}\|W^{(i)}\|_2^2 = m\sum_{k\neq i} w_k^2 = m(1 - w_i^2) \le m.$$

Consequently, $\mathbb{E}\|W^{(i)}\|_2 \le \sqrt{m}$, $\mathbb{E}\|b_i\|_2 \le C\sqrt{m/n}$ and

$$\mathbb{E}(\|b_i\|_2\,\|W^{(i)}\|_2) \le \big(\mathbb{E}\|b_i\|_2^2\,\mathbb{E}\|W^{(i)}\|_2^2\big)^{1/2} \le Cmn^{-1/2}.$$

By the deletion bound and $\|W\|_2 \le \|W^{(i)}\|_2 + |w_i|\,\|q_i\|_2$,

$$\|q_i\|_2\,\|E_n - E_n^{(i)}\|_{\mathrm{F}}\|W\|_2 \le \frac{C}{n}(\|q_i\|_2^3 + \|q_i\|_2^2\|b_i\|_2)(\|W^{(i)}\|_2 + |w_i|\,\|q_i\|_2).$$

Expanding the product and using that q_i is independent of $(b_i, W^{(i)}, E_n^{(i)})$, we get

$$\begin{aligned}
&\mathbb{E}\{\|q_i\|_2\,\|E_n - E_n^{(i)}\|_{\mathrm{F}}\|W\|_2\} \\
&\quad\le \frac{C}{n}\big[\mu_3\mathbb{E}\|W^{(i)}\|_2 + m\mathbb{E}(\|b_i\|_2\,\|W^{(i)}\|_2) + |w_i|\beta + |w_i|\mu_3\mathbb{E}\|b_i\|_2\big] \\
&\quad\le \frac{C}{n}\big(\mu_3\sqrt{m} + m^2n^{-1/2} + |w_i|\beta + |w_i|\mu_3\sqrt{m/n}\big).
\end{aligned}$$

By the same independence, Jensen's inequality and (11),

$$\mathbb{E}\{\|q_i\|_2^2\|E_n^{(i)}\|_{\mathrm{F}}\} = m\mathbb{E}\|E_n^{(i)}\|_{\mathrm{F}} \le m\big(\mathbb{E}\|E_n^{(i)}\|_{\mathrm{F}}^2\big)^{1/2} \le Cmx.$$

Summation. By the Cauchy–Schwarz inequality and (9), $\sum_i |w_i| \le \sqrt{\bar{n}}(\sum_i w_i^2)^{1/2} = \sqrt{\bar{n}} \le \sqrt{n}$. Together with $\sum_i w_i^2 = 1$, this gives

$$\begin{aligned}
&\sum_i |w_i|\mathbb{E}\{\|q_i\|_2\,|D_n - D_n^{(i)}|\} \\
&\quad\le 2\sum_i |w_i|\mathbb{E}\{\|q_i\|_2\,\|E_n - E_n^{(i)}\|_{\mathrm{F}}\|W\|_2\} + 2\sum_i w_i^2\mathbb{E}\{\|q_i\|_2^2\|E_n^{(i)}\|_{\mathrm{F}}\} \\
&\quad\le C\big(\mu_3 n^{-1/2} + n^{-1} + \beta/n + \mu_3 n^{-3/2} + x\big).
\end{aligned}$$

Finally, $\mu_3 n^{-1/2}$ and $\mu_3 n^{-3/2}$ are both at most $\sqrt{m\beta/n} = \sqrt{m}\,x$, $\beta/n = x^2$, and $n^{-1} \le x^2$ by (14). This proves (12). □

Lemma E.3. *Suppose Assumptions 1 and 2 hold. Under H_0, let $\zeta_1, \dots, \zeta_{\bar{n}}$ be the pooled testing observations, and define*

$$q_i = \Sigma^{-1/2}\{\hat{f}(\zeta_i) - \mathbb{E}_\star\hat{f}(\zeta_i)\}, \qquad \beta_\star = \mathbb{E}_\star\|q_1\|_2^4, \qquad x_n = \sqrt{\beta_\star/n}.$$

With the weights in Lemma E.2, put $W = \sum_i w_i q_i$, $T_n^0 = \|W\|_\infty$ *and* $G = \max_{1\le j\le m} |Z_j|$, *where* $Z \sim N(0, I_m)$. *Almost surely with respect to* $\mathcal{D}_{\mathrm{fit}}$,

$$\sup_{t\in\mathbb{R}} |\mathbb{P}_\star(T_n \le t) - \mathbb{P}_\star(T_n^0 \le t)| \le C(x_n + x_n^2), \tag{16}$$

$$\sup_{t\in\mathbb{R}} |\mathbb{P}_\star(T_n^0 \le t) - \mathbb{P}(G \le t)| \le C x_n, \tag{17}$$

$$\beta_\star \le \kappa^{-2} m^2 L_n, \tag{18}$$

where C *depends only on* m, c_1, c_2.

Proof. Construction of q_i *and bound on* $\beta_\star$. Let $V = \mathrm{diag}(\Sigma_{11}, \dots, \Sigma_{mm})$ and $R = V^{-1/2}\Sigma V^{-1/2}$. Assumption 2 gives $V \succ 0$ and $R \succeq \kappa I_m$, so $\Sigma = V^{1/2} R V^{1/2} \succ 0$ and q_i is well defined.

Under H_0, conditionally on $\mathcal{D}_{\mathrm{fit}}$, the ζ_i are independent and identically distributed and independent of $\hat{f}$. In particular $\mathrm{Var}_\star\{\hat{f}(X)\} = \mathrm{Var}_\star\{\hat{f}(Y)\}$, so Σ is the common conditional covariance of $\hat{f}(\zeta_i)$. Hence, conditionally on $\mathcal{D}_{\mathrm{fit}}$, the q_i are independent and identically distributed, with mean zero (by centering) and covariance I_m (by whitening).

Put $\tilde{q} = V^{-1/2}\{\hat{f}(\zeta_1) - \mathbb{E}_\star \hat{f}(\zeta_1)\}$. Then $\|q_1\|_2^2 = \tilde{q}^\top V^{1/2}\Sigma^{-1}V^{1/2}\tilde{q} = \tilde{q}^\top R^{-1}\tilde{q}$, and $R^{-1} \preceq \kappa^{-1} I_m$. The fourth-moment bound in Assumption 2 states exactly that $\mathbb{E}_\star|\tilde{q}_j|^4 \le L_n$ for each j. Hence

$$\beta_\star = \mathbb{E}_\star(\tilde{q}^\top R^{-1}\tilde{q})^2 \le \kappa^{-2}\mathbb{E}_\star\Big(\sum_{j=1}^m \tilde{q}_j^2\Big)^2 \le \kappa^{-2} m \sum_{j=1}^m \mathbb{E}_\star|\tilde{q}_j|^4 \le \kappa^{-2} m^2 L_n,$$

which is (18).

Reduction to the setting of Lemma E.2. By Assumption 1, $\bar{n}_x \ge (1-c_2)n_x \ge c_1(1-c_2)n$ and $\bar{n}_y \ge (1-c_2)n_y \ge (1-c_2)^2 n$. Thus $\min(\bar{n}_x, \bar{n}_y) \ge an$ with $a = (1-c_2)\min\{c_1, 1-c_2\}$, and Lemma E.2 applies conditionally on $\mathcal{D}_{\mathrm{fit}}$ with $\beta = \beta_\star$ and $x = x_n$.

Under H_0, the conditional means $\mathbb{E}_\star \hat{f}(X)$ and $\mathbb{E}_\star \hat{f}(Y)$ coincide, so they cancel in $\hat{S}_n$. Therefore

$$W = \sqrt{n_{\mathrm{eff}}}(\bar{q}_x - \bar{q}_y) = \sqrt{n_{\mathrm{eff}}}\,\Sigma^{-1/2}\hat{S}_n.$$

The sample covariance matrices are invariant under translation of the observations, so

$$\Sigma^{-1/2}\hat{C}_X\Sigma^{-1/2} = \frac{1}{\bar{n}_x}\sum_{i\in I_x} q_i q_i^\top - \bar{q}_x\bar{q}_x^\top,$$

$$\Sigma^{-1/2}\hat{C}_Y\Sigma^{-1/2} = \frac{1}{\bar{n}_y}\sum_{i\in I_y} q_i q_i^\top - \bar{q}_y\bar{q}_y^\top.$$

Since $w_i^2 = \omega_x/\bar{n}_x$ for $i \in I_x$, $w_i^2 = \omega_y/\bar{n}_y$ for $i \in I_y$, and $\sum_i w_i^2 = 1$, it follows that

$$\begin{aligned}
\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} &= \omega_x\Sigma^{-1/2}\hat{C}_X\Sigma^{-1/2} + \omega_y\Sigma^{-1/2}\hat{C}_Y\Sigma^{-1/2} \\
&= \omega_x\left(\frac{1}{\bar{n}_x}\sum_{i\in I_x} q_i q_i^\top - \bar{q}_x\bar{q}_x^\top\right) + \omega_y\left(\frac{1}{\bar{n}_y}\sum_{i\in I_y} q_i q_i^\top - \bar{q}_y\bar{q}_y^\top\right) \\
&= I_m + E_n,
\end{aligned}$$

with E_n as in (8).

Gaussian approximation of T_n^0. Apply Theorem 1.1 of Bentkus (2005) to the independent vectors $X_i = w_i q_i$. These have mean zero and $\sum_i \mathbb{E}_\star(X_i X_i^\top) = \sum_i w_i^2 I_m = I_m$. For every $t \ge 0$, $\{T_n^0 \le t\} = \{W \in [-t,t]^m\}$ with $[-t,t]^m$ convex, so

$$\begin{aligned}
|\mathbb{P}_\star(T_n^0 \le t) - \mathbb{P}(G \le t)| &\le C m^{1/4}\sum_i |w_i|^3 \mathbb{E}_\star\|q_i\|_2^3 \\
&\le C m^{1/4}\max_i |w_i| (\mathbb{E}_\star\|q_1\|_2^2)^{1/2}(\mathbb{E}_\star\|q_1\|_2^4)^{1/2} \le C x_n,
\end{aligned}$$

where the second step uses $\sum_i |w_i|^3 \le \max_i |w_i| \sum_i w_i^2$ and the Cauchy–Schwarz inequality, and the last step uses $\max_i |w_i| \le (an)^{-1/2}$ and $\mathbb{E}_\star \|q_1\|_2^2 = m$. For $t < 0$ both distribution functions are zero, proving (17).

Reduction of the studentization error to anti-concentration. Let $\mathcal{O}_n = \{\|E_n\|_{\mathrm{F}} \le 1/2\}$. By Markov's inequality and (11),

$$\mathbb{P}_\star(\mathcal{O}_n^c) \le 4\mathbb{E}_\star \|E_n\|_{\mathrm{F}}^2 \le C x_n^2.$$

On $\mathcal{O}_n$ we have $\|E_n\|_{\mathrm{op}} \le \|E_n\|_{\mathrm{F}} \le 1/2$, so Lemma E.1 gives $\hat{\Sigma} \succ 0$. Hence, on $\mathcal{O}_n$, $T_n = \|V_n\|_\infty$ with $V_n = \sqrt{n_{\mathrm{eff}}}\, \hat{\Sigma}^{-1/2} \hat{S}_n = \hat{\Sigma}^{-1/2}\Sigma^{1/2} W$, and

$$\|V_n - W\|_2 \le \|\hat{\Sigma}^{-1/2}\Sigma^{1/2} - I_m\|_{\mathrm{F}} \|W\|_2 \le \sqrt{2}\|E_n\|_{\mathrm{F}} \|W\|_2 \le D_n \quad \text{on } \mathcal{O}_n.$$

Thus $|T_n - T_n^0| \le \|V_n - W\|_\infty \le D_n$ on $\mathcal{O}_n$. Therefore, for every t,

$$|\mathbb{1}\{T_n \le t\} - \mathbb{1}\{T_n^0 \le t\}| \le \mathbb{1}_{\mathcal{O}_n^c} + \mathbb{1}\{|T_n^0 - t| \le D_n\}.$$

If $|\|W\|_\infty - t| \le D_n$ and j attains the maximum, then $|W_j - \varepsilon t| = \big||W_j| - t\big| \le D_n$ for any $\varepsilon \in \{-1, 1\}$ with $\varepsilon W_j = |W_j|$. A union bound over $j \in \{1, \dots, m\}$ and $\varepsilon \in \{-1, 1\}$ therefore gives

$$\sup_{t \in \mathbb{R}} |\mathbb{P}_\star(T_n \le t) - \mathbb{P}_\star(T_n^0 \le t)| \le \mathbb{P}_\star(\mathcal{O}_n^c) + 2\sum_{j=1}^m \sup_{z \in \mathbb{R}} \mathbb{P}_\star(|W_j - z| \le D_n). \tag{19}$$

Randomized concentration. Fix $j \le m$ and $z \in \mathbb{R}$, and put $\xi_i = w_i q_{ij}$, so that $W_j = \sum_i \xi_i$. Conditionally on $\mathcal{D}_{\mathrm{fit}}$, the ξ_i are independent with mean zero, and $\sum_i \mathbb{E}_\star \xi_i^2 = \sum_i w_i^2 = 1$ because $\mathbb{E}_\star q_{ij}^2 = 1$. Set

$$\Delta_1 = z - D_n, \quad \Delta_2 = z + D_n, \quad \Delta_1^{(i)} = z - D_n^{(i)}, \quad \Delta_2^{(i)} = z + D_n^{(i)}.$$

Then $\{|W_j - z| \le D_n\} = \{\Delta_1 \le W_j \le \Delta_2\}$. The leave-one-out condition of Chen et al. (2010, Proposition 10.1) holds, because $W_j - \xi_i = W_j^{(i)}$ and $D_n^{(i)}$ are functions of $(q_k)_{k \ne i}$ and are therefore conditionally independent of ξ_i. That proposition gives, for every δ satisfying condition (10.3) of Chen et al. (2010),

$$\mathbb{P}_\star(\Delta_1 \le W_j \le \Delta_2) \le 4\delta + \mathbb{E}_\star|W_j(\Delta_2 - \Delta_1)| + \sum_i \sum_{l=1}^2 \mathbb{E}_\star|\xi_i(\Delta_l - \Delta_l^{(i)})|.$$

Let $\beta_{2j} = \sum_i \mathbb{E}_\star\{\xi_i^2 \mathbb{1}(|\xi_i| > 1)\}$ and $\beta_{3j} = \sum_i \mathbb{E}_\star\{|\xi_i|^3 \mathbb{1}(|\xi_i| \le 1)\}$. If $\beta_{2j} + \beta_{3j} \le 1/2$, then $\delta = (\beta_{2j} + \beta_{3j})/2$ satisfies (10.3) by Remark 10.1(ii) of Chen et al. (2010); this δ is positive, since $\beta_{2j} + \beta_{3j} = 0$ would force $\xi_i = 0$ almost surely for all i, contradicting $\sum_i \mathbb{E}_\star \xi_i^2 = 1$. If $\beta_{2j} + \beta_{3j} > 1/2$, the following bound is trivial because its right-hand side exceeds one. In either case,

$$\mathbb{P}_\star(\Delta_1 \le W_j \le \Delta_2) \le 2(\beta_{2j} + \beta_{3j}) + \mathbb{E}_\star|W_j(\Delta_2 - \Delta_1)| + \sum_i \sum_{l=1}^2 \mathbb{E}_\star|\xi_i(\Delta_l - \Delta_l^{(i)})|.$$

We bound the three groups of terms in turn.

First, put $\gamma_j = (\sum_i \mathbb{E}_\star|\xi_i|^4)^{1/2}$. Since $\mathbb{E}_\star q_{ij}^4 \le \beta_\star$ and $\sum_i w_i^4 \le (an)^{-1}$, we have $\gamma_j \le C x_n$. Using $\xi^2 \mathbb{1}(|\xi| > 1) \le \xi^4$ and the Cauchy–Schwarz inequality,

$$\beta_{2j} \le \sum_i \mathbb{E}_\star|\xi_i|^4 = \gamma_j^2, \qquad \beta_{3j} \le \sum_i \mathbb{E}_\star|\xi_i|^3 \le \Big(\sum_i \mathbb{E}_\star \xi_i^2\Big)^{1/2} \Big(\sum_i \mathbb{E}_\star|\xi_i|^4\Big)^{1/2} = \gamma_j.$$

Hence $\beta_{2j} + \beta_{3j} \le C(x_n + x_n^2)$.

Second, $\Delta_2 - \Delta_1 = 2D_n$ and $|W_j| \le \|W\|_2$. By the Cauchy–Schwarz inequality, (10) and (11),

$$\begin{aligned}\mathbb{E}_\star(|W_j| D_n) &\le 2\mathbb{E}_\star(\|E_n\|_{\mathrm{F}} \|W\|_2^2) \le 2\big(\mathbb{E}_\star\|E_n\|_{\mathrm{F}}^2\big)^{1/2} \big(\mathbb{E}_\star\|W\|_2^4\big)^{1/2} \\ &\le C x_n \sqrt{1 + x_n^2} \le C(x_n + x_n^2).\end{aligned}$$

Third, $|\Delta_l - \Delta_l^{(i)}| = |D_n - D_n^{(i)}|$ and $|\xi_i| \le |w_i| \, \|q_i\|_2$. By (12),

$$\sum_i \mathbb{E}_\star |\xi_i(\Delta_l - \Delta_l^{(i)})| \le \sum_i |w_i| \mathbb{E}_\star\{\|q_i\|_2 \, |D_n - D_n^{(i)}|\} \le C(x_n + x_n^2), \qquad l = 1, 2.$$

Consequently,

$$\sup_{z \in \mathbb{R}} \mathbb{P}_\star(|W_j - z| \le D_n) \le C(x_n + x_n^2).$$

Substituting this and the bound on $\mathbb{P}_\star(\mathcal{O}_n^c)$ into (19) proves (16). □

F Proof of Theorem 2

Proof. Recall that

$$n_{\text{eff}} = \frac{\bar{n}_x \bar{n}_y}{\bar{n}_x + \bar{n}_y}, \qquad v_j = \Sigma_{jj}.$$

By (4), with probability tending to one,

$$|\hat{S}_{n,j} - \Delta_{\hat{f}_j}(P, Q)| \le a_{n,j}$$

and

$$|\Delta_{\hat{f}_j}(P, Q)| \ge \Delta_{f_j}(P, Q) - b_{n,j}.$$

On their intersection,

$$\begin{aligned} |\hat{S}_{n,j}| &\ge |\Delta_{\hat{f}_j}(P, Q)| - |\hat{S}_{n,j} - \Delta_{\hat{f}_j}(P, Q)| \\ &\ge \Delta_{f_j}(P, Q) - a_{n,j} - b_{n,j}. \end{aligned}$$

Consequently,

$$\mathbb{P}\Big(|\hat{S}_{n,j}| < \Delta_{f_j}(P, Q) - a_{n,j} - b_{n,j}\Big) \to 0. \tag{20}$$

Conditional on the fitting data, $\mathbb{E}_\star \hat{\Sigma}_{jj} \le v_j$. Indeed, each group covariance is centered at its own sample mean and normalized by its sample size, so $\mathbb{E}_\star \hat{C}_X = (1 - \bar{n}_x^{-1}) \text{Var}_\star\{\hat{f}(X)\}$ and $\mathbb{E}_\star \hat{C}_Y = (1 - \bar{n}_y^{-1}) \text{Var}_\star\{\hat{f}(Y)\}$; hence $\mathbb{E}_\star \hat{\Sigma} \preceq \Sigma$.
Since $\Delta_{f_j}(P, Q) - a_{n,j} - b_{n,j}$ is deterministic and eventually positive, conditional Markov's inequality gives, for every $\varepsilon > 0$,

$$\begin{aligned} &\mathbb{P}\left(\frac{\hat{\Sigma}_{jj}}{n\{\Delta_{f_j}(P, Q) - a_{n,j} - b_{n,j}\}^2} > \varepsilon \right) \\ &\qquad \le \mathbb{E}\left[\min\left\{ 1, \, \frac{v_j}{\varepsilon n\{\Delta_{f_j}(P, Q) - a_{n,j} - b_{n,j}\}^2} \right\} \right] \longrightarrow 0. \end{aligned}$$

Indeed, the integrand is bounded by one and converges to zero in probability by (5). Assumption 1 gives $n_{\text{eff}} \asymp n$ and fixed m. Therefore,

$$\frac{m\hat{\Sigma}_{jj}}{n_{\text{eff}}\{\Delta_{f_j}(P, Q) - a_{n,j} - b_{n,j}\}^2} \xrightarrow{P} 0.$$

On $\{\hat{\Sigma} \succ 0\}$, the coordinate inequality in the proof of Lemma G.1 gives

$$|\hat{S}_{n,j}| \le \sqrt{\frac{m\hat{\Sigma}_{jj}}{n_{\text{eff}}}} \, T_n.$$

Hence, for all sufficiently large n,

$$\begin{aligned} \mathbb{P}(T_n \le q_{1-\alpha}) \le \mathbb{P}(\hat{\Sigma} \not\succ 0) \\ &+ \mathbb{P}\Big(|\hat{S}_{n,j}| < \Delta_{f_j}(P, Q) - a_{n,j} - b_{n,j}\Big) \\ &+ \mathbb{P}\left(\frac{m q_{1-\alpha}^2 \hat{\Sigma}_{jj}}{n_{\text{eff}}\{\Delta_{f_j}(P, Q) - a_{n,j} - b_{n,j}\}^2} \ge 1 \right). \end{aligned}$$

The first term tends to zero by (40) and $L_n/n \to 0$. The second term tends to zero by (20). The third term tends to zero by the preceding covariance bound, since $q_{1-\alpha}$ is fixed. Therefore,

$$\mathbb{P}(T_n > q_{1-\alpha}) \longrightarrow 1.$$

□

G Proof of Theorem D.1

Let $X_1^{\mathrm{fit}}, \ldots, X_{\bar{n}_x}^{\mathrm{fit}}$ and $Y_1^{\mathrm{fit}}, \ldots, Y_{\bar{n}_y}^{\mathrm{fit}}$ denote the fitting subsamples. Recall that

$$n_{\mathrm{eff}} = \frac{\bar{n}_x \bar{n}_y}{\bar{n}_x + \bar{n}_y}, \qquad v_j = \Sigma_{jj},$$

For every fitted coordinate $\hat{f}_j$, the population signal is

$$\Delta_{\hat{f}_j}(P, Q) = \mathbb{E}_P \hat{f}_j(X) - \mathbb{E}_Q \hat{f}_j(Y).$$

For MMD, ED, and KFDA, the proof has two steps. The first controls the testing-sample fluctuation conditional on the fitting data, as established in Lemma G.1. The second controls the fitted signal using the MMD and ED bounds in Section G.1 and the KFDA bound in Lemma G.3. For PWD, we use the empirical signal bound of Hu & Lin (2026, Remark 12, following the proof of Theorem 3). The event $\hat{\Sigma} \not\succ 0$ is controlled by (40).

Proof of Theorem D.1. By (40) and Assumption 2,

$$\mathbb{P}(\hat{\Sigma} \not\succ 0) = o(1). \tag{21}$$

If the MMD branch of the signal condition in Theorem D.1 holds, then (26) gives

$$\mathbb{P}(T_n \le q_{1-\alpha}) \le o(1) + \frac{C}{c_a^2 \log\log n} \longrightarrow 0.$$

If the ED branch holds, then

$$ED_p(P, Q) \ge 2c_a^2 \eta^p d^{p/2} \frac{\log\log n}{n},$$

and (32) gives

$$\mathbb{P}(T_n \le q_{1-\alpha}) \le o(1) + \frac{C}{c_a^2 \log\log n} \longrightarrow 0.$$

Suppose the KFDA branch holds. Then

$$D_{\mathrm{KFDA},\lambda}(P, Q)^2 \ge \frac{c_a^2 \sqrt{N}}{1+\lambda} \frac{\log\log n}{n}. \tag{22}$$

If $0 < \lambda \le 1$, $N_F(\lambda) \le N$ and $N \ge 1$ show that (22) implies (44) for all sufficiently large n. Hence (45) yields

$$\mathbb{P}(T_n \le q_{1-\alpha}) \le o(1) + \frac{C}{c_a^2 \log\log n} + \frac{C}{c_a^4 (\log\log n)^2} + \frac{C}{n\lambda^2} \longrightarrow 0.$$

If $\lambda \ge 1$, (46) and (22) give

$$\mathbb{P}(T_n \le q_{1-\alpha}) \le o(1) + \frac{C(1+\lambda)^2}{c_a^2 \lambda^2 \sqrt{N} \log\log n} \longrightarrow 0.$$

It remains to consider the PWD branch. Suppose that

$$PW^{(1)}_{k_\star, \tau_\star}(P, Q) \ge c_a R_n^{(1)}(k_\star, \tau_\star, d).$$

Let J_{PWD} be the set of coordinates corresponding to the PWD witnesses, and abbreviate $R_n = R_n^{(1)}(k_\star, \tau_\star, d)$. The testing-sample signal bound in Hu & Lin (2026, Remark 12, following the proof of Theorem 3), under the same candidate class and network architecture, gives

$$\mathbb{P}\left(\max_{j\in J_{\mathrm{PWD}}} |\hat{S}_{n,j}| \geq PW^{(1)}_{k_\star,\tau_\star}(P,Q) - CR_n\right) \longrightarrow 1.$$

Consequently, for $c_a > C$,

$$\mathbb{P}\left(\max_{j\in J_{\mathrm{PWD}}} |\hat{S}_{n,j}| \geq (c_a - C)R_n\right) \longrightarrow 1.$$

On $\{\hat{\Sigma} \succ 0\}$, the coordinate bound used in the proof of Lemma G.1 implies

$$T_n \geq \sqrt{\frac{n_{\mathrm{eff}}}{m \max_{j\in J_{\mathrm{PWD}}} \hat{\Sigma}_{jj}}} \max_{j\in J_{\mathrm{PWD}}} |\hat{S}_{n,j}|.$$

Assumption D.1 gives $v_j \leq c_2'$ for $j \in J_{\mathrm{PWD}}$. Since $\mathbb{E}_\star \hat{\Sigma}_{jj} \leq v_j$ and m is fixed, Markov's inequality yields

$$\max_{j\in J_{\mathrm{PWD}}} \hat{\Sigma}_{jj} = O_{\mathbb{P}}(1).$$

Moreover, $n_{\mathrm{eff}} \asymp n$ and $\sqrt{n}\, R_n \to \infty$. Together with (21), these bounds imply $T_n \to \infty$ in probability, and hence

$$\mathbb{P}(T_n \leq q_{1-\alpha}) \longrightarrow 0.$$

For each n, select a branch attaining the maximum in the signal condition of Theorem D.1. There are only four branches. The preceding bounds imply that the non-rejection probability tends to zero along every subsequence on which the selected branch is fixed. Therefore the same convergence holds along the full sequence, and

$$\mathbb{P}(T_n > q_{1-\alpha}) \longrightarrow 1.$$

□

G.1 Method-specific error bounds

Each method-specific bound below is asserted only when the population discrepancy appearing in its denominator is strictly positive. In applying these bounds to Theorem D.1, only a branch satisfying the signal condition is used; the other branches may have zero population discrepancy.

MMD. Let

$$\hat{\mu}_P = \frac{1}{\tilde{n}_x}\sum_{i=1}^{\tilde{n}_x} k(X_i^{\mathrm{fit}}, \cdot), \qquad \hat{\mu}_Q = \frac{1}{\tilde{n}_y}\sum_{i=1}^{\tilde{n}_y} k(Y_i^{\mathrm{fit}}, \cdot).$$

For the MMD coordinate of $\hat{f}$,

$$\hat{f}_j(z) = \frac{\langle \hat{\mu}_P - \hat{\mu}_Q, k(z,\cdot)\rangle_{\mathcal{H}_k}}{\|\hat{\mu}_P - \hat{\mu}_Q\|_{\mathcal{H}_k}}$$

whenever $\hat{\mu}_P \neq \hat{\mu}_Q$. Set $\hat{f}_j = 0$ when $\hat{\mu}_P = \hat{\mu}_Q$. Define

$$\mathcal{A}_M = \left\{\|(\hat{\mu}_P - \hat{\mu}_Q) - (\mu_P - \mu_Q)\|_{\mathcal{H}_k} \leq \frac{1}{4}\mathrm{MMD}(P,Q)\right\}.$$

On $\mathcal{A}_M$,

$$\begin{aligned}\Delta_{\hat{f}_j}(P,Q) &= \frac{\langle \mu_P - \mu_Q, \hat{\mu}_P - \hat{\mu}_Q\rangle_{\mathcal{H}_k}}{\|\hat{\mu}_P - \hat{\mu}_Q\|_{\mathcal{H}_k}} \\ &\geq \frac{3}{5}\,\mathrm{MMD}(P,Q), \qquad v_j \leq 1. \end{aligned} \tag{23}$$

Indeed, write $\nu_M = \mu_P - \mu_Q$ and $\hat{\nu}_M = \hat{\mu}_P - \hat{\mu}_Q$, so that $\|\nu_M\|_{\mathcal{H}_k} = \mathrm{MMD}(P,Q)$. On $\mathcal{A}_M$, $\|\hat{\nu}_M - \nu_M\|_{\mathcal{H}_k} \le \|\nu_M\|_{\mathcal{H}_k}/4$. Hence $\hat{\nu}_M \neq 0$, $\langle \nu_M, \hat{\nu}_M \rangle_{\mathcal{H}_k} \ge \|\nu_M\|^2_{\mathcal{H}_k} - \|\nu_M\|_{\mathcal{H}_k}\|\hat{\nu}_M - \nu_M\|_{\mathcal{H}_k} \ge \frac{3}{4}\|\nu_M\|^2_{\mathcal{H}_k}$ and $\|\hat{\nu}_M\|_{\mathcal{H}_k} \le \frac{5}{4}\|\nu_M\|_{\mathcal{H}_k}$, which gives the first bound. For the second, the reproducing property and the Cauchy–Schwarz inequality give $|\hat{f}_j(z)| \le k(z,z)^{1/2} \le 1$. Hence both conditional variances in the definition of Σ are at most one, and their weights sum to one. Thus

$$\mathbb{1}_{\mathcal{A}_M} \frac{v_j}{\Delta_{\hat{f}_j}(P,Q)^2} \le \frac{25}{9\,\mathrm{MMD}(P,Q)^2}. \tag{24}$$

The fitting mean embeddings satisfy

$$\mathbb{E}\|(\hat{\mu}_P - \hat{\mu}_Q) - (\mu_P - \mu_Q)\|^2_{\mathcal{H}_k} \le \frac{C}{n},$$

so

$$\mathbb{P}(\mathcal{A}_M^c) \le \frac{C}{n\,\mathrm{MMD}(P,Q)^2}. \tag{25}$$

Applying Lemma G.1 with (24)–(25) gives

$$\mathbb{P}(T_n \le q) \le \mathbb{P}(\hat{\Sigma} \not\succ 0) + \frac{C_q}{n\,\mathrm{MMD}(P,Q)^2}. \tag{26}$$

ED. Let $\rho_p(x,y) = \|x-y\|_2^p$ and define the distance-induced kernel anchored at the origin by

$$k_p(x,y) = \frac{1}{2}\big\{\|x\|_2^p + \|y\|_2^p - \|x-y\|_2^p\big\}.$$

Let $\phi_p(z) = k_p(z,\cdot)$ be its canonical feature map, and write

$$\mu_{P,p} = \mathbb{E}_P \phi_p(X), \qquad \mu_{Q,p} = \mathbb{E}_Q \phi_p(Y),$$

with fitting-sample versions $\hat{\mu}_{P,p}$ and $\hat{\mu}_{Q,p}$. For $p \in (0,2)$, ρ_p is a semimetric of negative type, so k_p is positive definite. Assumption D.2 gives $\mathbb{E}\|X\|_2^p + \mathbb{E}\|Y\|_2^p < \infty$, so the mean embeddings exist. The distance-induced kernel identities of Sejdinovic et al. (2013), together with $k_p(z,z) = \|z\|_2^p$, give

$$\|\mu_{P,p} - \mu_{Q,p}\|^2 = \frac{ED_p(P,Q)}{2}, \qquad \|\phi_p(z)\|^2 = \|z\|_2^p. \tag{27}$$

For the ED coordinate of $\hat{f}$,

$$\Delta_{\hat{f}_j}(P,Q) = 2\langle \mu_{P,p} - \mu_{Q,p}, \hat{\mu}_{P,p} - \hat{\mu}_{Q,p} \rangle.$$

Indeed, the fitted ED witness is $\hat{f}_j(z) = \tilde{n}_y^{-1}\sum_b \rho_p(z, Y_b^{\mathrm{fit}}) - \tilde{n}_x^{-1}\sum_a \rho_p(z, X_a^{\mathrm{fit}})$. Since $\rho_p(z,y) = \|z\|_2^p + \|y\|_2^p - 2k_p(z,y)$, the terms $\|z\|_2^p$ cancel and

$$\hat{f}_j(z) = c_{\mathrm{fit}} + 2\langle \hat{\mu}_{P,p} - \hat{\mu}_{Q,p}, \phi_p(z) \rangle, \qquad c_{\mathrm{fit}} = \frac{1}{\tilde{n}_y}\sum_b \|Y_b^{\mathrm{fit}}\|_2^p - \frac{1}{\tilde{n}_x}\sum_a \|X_a^{\mathrm{fit}}\|_2^p.$$

The constant c_{fit} is fitting-measurable and cancels in $\Delta_{\hat{f}_j}(P,Q)$, which gives the preceding display. Define

$$\mathcal{A}_E = \left\{ \|(\hat{\mu}_{P,p} - \hat{\mu}_{Q,p}) - (\mu_{P,p} - \mu_{Q,p})\| \le \frac{1}{4}\|\mu_{P,p} - \mu_{Q,p}\| \right\}.$$

On $\mathcal{A}_E$,

$$\Delta_{\hat{f}_j}(P,Q) \ge \frac{3}{4} ED_p(P,Q), \qquad v_j \le \frac{25}{8} ED_p(P,Q)\,\eta^p d^{p/2}. \tag{28}$$

Consequently,

$$\mathbb{1}_{\mathcal{A}_E} \frac{v_j}{\Delta_{\hat{f}_j}(P,Q)^2} \le \frac{50\eta^p d^{p/2}}{9 ED_p(P,Q)}. \tag{29}$$

Also, by (27), Assumption D.2, and the fitting sample proportions,

$$\mathbb{P}(\mathcal{A}_E^c) \le \frac{C\eta^p d^{p/2}}{n ED_p(P,Q)}. \tag{30}$$

We verify (28) and (30). Write $\nu_E = \mu_{P,p} - \mu_{Q,p}$ and $\hat{\nu}_E = \hat{\mu}_{P,p} - \hat{\mu}_{Q,p}$, so that $\|\nu_E\|^2 = ED_p(P,Q)/2$ by (27). On $\mathcal{A}_E$, $\langle \nu_E, \hat{\nu}_E \rangle \geq \frac{3}{4}\|\nu_E\|^2$ and $\|\hat{\nu}_E\| \leq \frac{5}{4}\|\nu_E\|$. The first gives $\Delta_{\hat{f}_j}(P,Q) \geq \frac{3}{2}\|\nu_E\|^2 = \frac{3}{4}ED_p(P,Q)$. For the second, let $Z \sim P$ or $Z \sim Q$. For every unit vector u and every $t > \eta$, Assumption D.2 and $1+s \leq e^s$ give $\mathbb{E}(u^\top Z)^2 \leq t^2$, hence $\mathbb{E}(u^\top Z)^2 \leq \eta^2$. Summing over the d coordinate directions gives $\mathbb{E}\|Z\|_2^2 \leq d\eta^2$, and Jensen's inequality with $p/2 < 1$ gives

$$\mathbb{E}\|Z\|_2^p \leq \eta^p d^{p/2}. \tag{31}$$

Since c_{fit} does not affect variances,

$$\begin{aligned}
\mathrm{Var}_\star\{\hat{f}_j(Z)\} &= 4\mathrm{Var}_\star\langle \hat{\nu}_E, \phi_p(Z)\rangle \\
&\leq 4\|\hat{\nu}_E\|^2 \mathbb{E}\|\phi_p(Z)\|^2 \\
&\leq 4 \cdot \frac{25}{16} \cdot \frac{ED_p(P,Q)}{2}\, \mathbb{E}\|Z\|_2^p \\
&\leq \frac{25}{8} ED_p(P,Q)\, \eta^p d^{p/2}.
\end{aligned}$$

The two weights in v_j sum to one, which gives the bound on v_j. For (30), the two fitting groups are independent and each embedding error has mean zero, so (27), (31) and Assumption 1 give

$$\mathbb{E}\|\hat{\nu}_E - \nu_E\|^2 \leq \frac{\mathbb{E}\|X\|_2^p}{\tilde{n}_x} + \frac{\mathbb{E}\|Y\|_2^p}{\tilde{n}_y} \leq \frac{C\eta^p d^{p/2}}{n}.$$

Markov's inequality then gives $\mathbb{P}(\mathcal{A}_E^c) \leq 16\mathbb{E}\|\hat{\nu}_E - \nu_E\|^2/\|\nu_E\|^2 \leq C\eta^p d^{p/2}/\{nED_p(P,Q)\}$. Lemma G.1 now gives

$$\mathbb{P}(T_n \leq q) \leq \mathbb{P}(\hat{\Sigma} \not\succ 0) + \frac{C_q \eta^p d^{p/2}}{nED_p(P,Q)}. \tag{32}$$

KFDA. Let

$$\mu_{P,F} = \mathbb{E}_P \phi_F(X), \qquad \mu_{Q,F} = \mathbb{E}_Q \phi_F(Y),$$

and let $\hat{\mu}_{P,F}$ and $\hat{\mu}_{Q,F}$ be their fitting-sample mean embeddings. Define

$$\begin{aligned}
\hat{C}_{r,F} = &\frac{1}{r\tilde{n}_x} \sum_{a=1}^{\tilde{n}_x} \{\phi_F(X_a^{\mathrm{fit}}) - \hat{\mu}_{P,F}\} \otimes \{\phi_F(X_a^{\mathrm{fit}}) - \hat{\mu}_{P,F}\} \\
&+ \frac{1}{(1-r)\tilde{n}_y} \sum_{a=1}^{\tilde{n}_y} \{\phi_F(Y_a^{\mathrm{fit}}) - \hat{\mu}_{Q,F}\} \otimes \{\phi_F(Y_a^{\mathrm{fit}}) - \hat{\mu}_{Q,F}\},
\end{aligned}$$

where $r = n_x/n$ is the same proportion as in Assumption D.3. Put

$$A_F = C_{r,F} + \lambda I, \qquad \hat{A}_F = \hat{C}_{r,F} + \lambda I.$$

For the KFDA coordinate of $\hat{f}$,

$$\hat{f}_j(z) = \left\langle \hat{A}_F^{-1}(\hat{\mu}_{P,F} - \hat{\mu}_{Q,F}), \phi_F(z) \right\rangle, \tag{33}$$

and therefore

$$\Delta_{\hat{f}_j}(P,Q) = \left\langle \mu_{P,F} - \mu_{Q,F}, \hat{A}_F^{-1}(\hat{\mu}_{P,F} - \hat{\mu}_{Q,F}) \right\rangle. \tag{34}$$

By definition,

$$D_{\mathrm{KFDA},\lambda}(P,Q)^2 = \left\langle \mu_{P,F} - \mu_{Q,F}, A_F^{-1}(\mu_{P,F} - \mu_{Q,F}) \right\rangle. \tag{35}$$

Thus the method-specific step is to compare (34) directly with (35).

Based on Lemma G.3, for $0 < \lambda \leq 1$, if

$$nD_{\mathrm{KFDA},\lambda}(P,Q)^2 \geq \sqrt{1 + N_F(\lambda)},$$

then

$$\begin{aligned}
\mathbb{P}(T_n \leq q) \leq\ &\mathbb{P}(\hat{\Sigma} \not\succ 0) \\
&+ C_q \left\{ \frac{\sqrt{1+N_F(\lambda)}}{nD_{\mathrm{KFDA},\lambda}(P,Q)^2} + \frac{1+N_F(\lambda)}{n^2 D_{\mathrm{KFDA},\lambda}(P,Q)^4} + \frac{1}{n\lambda^2} \right\}.
\end{aligned}$$

For $\lambda \geq 1$,

$$\mathbb{P}(T_n \leq q) \leq \mathbb{P}(\hat{\Sigma} \not\succ 0) + \frac{C_q(1+\lambda)}{n\lambda^2 D_{\mathrm{KFDA},\lambda}(P,Q)^2}.$$

G.2 TECHNICAL LEMMAS AND PROOFS

Lemma G.1. *Under Assumptions 1–2, for every $q \geq 0$, every coordinate j, and every fitting-measurable event $\mathcal{A}$ on which $\Delta_{\hat{f}_j}(P,Q) \neq 0$,*

$$\mathbb{P}(T_n \leq q) \leq \mathbb{P}(\hat{\Sigma} \not\succ 0) + \mathbb{P}(\mathcal{A}^c) + \frac{C_q}{n}\mathbb{E}\left[\mathbb{1}_{\mathcal{A}}\frac{v_j}{\Delta_{\hat{f}_j}(P,Q)^2}\right]. \tag{36}$$

Here and below, an expression of the form $\mathbb{1}_{\mathcal{A}} v_j/\Delta_{\hat{f}_j}(P,Q)^2$ is defined to be zero on $\mathcal{A}^c$.

Proof. Conditional on $\mathcal{D}_{\text{fit}}$, the fitted witness is fixed and the testing sample is independent of it. Hence

$$\mathbb{E}_\star \hat{S}_{n,j} = \Delta_{\hat{f}_j}(P,Q), \qquad \mathrm{Var}_\star(\hat{S}_{n,j}) = \frac{v_j}{n_{\text{eff}}}. \tag{37}$$

On $\{\hat{\Sigma} \succ 0\}$, put $u = \sqrt{n_{\text{eff}}}\,\hat{\Sigma}^{-1/2}\hat{S}_n$, so that $T_n = \|u\|_\infty$ and $\sqrt{n_{\text{eff}}}\,\hat{S}_{n,j} = e_j^\top \hat{\Sigma}^{1/2} u$, where e_j is the jth standard basis vector. By the Cauchy–Schwarz inequality, $\|\hat{\Sigma}^{1/2} e_j\|_2 = \hat{\Sigma}_{jj}^{1/2}$ and $\|u\|_2 \leq \sqrt{m}\,\|u\|_\infty = \sqrt{m}\,T_n$,

$$|\hat{S}_{n,j}| \leq \sqrt{\frac{m\hat{\Sigma}_{jj}}{n_{\text{eff}}}}\,T_n.$$

Therefore

$$\begin{aligned}\{T_n \leq q\} \cap \{\hat{\Sigma} \succ 0\} \subseteq &\left\{|\hat{S}_{n,j} - \Delta_{\hat{f}_j}(P,Q)| \geq \frac{|\Delta_{\hat{f}_j}(P,Q)|}{2}\right\} \\ &\cup \left\{\frac{mq^2\hat{\Sigma}_{jj}}{n_{\text{eff}}} \geq \frac{\Delta_{\hat{f}_j}(P,Q)^2}{4}\right\}.\end{aligned} \tag{38}$$

By (37), Chebyshev's inequality, and $\mathbb{E}_\star \hat{\Sigma}_{jj} \leq v_j$,

$$\mathbb{P}_\star\Big(\{T_n \leq q\} \cap \{\hat{\Sigma} \succ 0\}\Big) \leq \frac{4(1+mq^2)v_j}{n_{\text{eff}}\Delta_{\hat{f}_j}(P,Q)^2}. \tag{39}$$

Assumption 1 gives $n_{\text{eff}} \asymp n$. Integrating (39) over $\mathcal{A}$ gives (36). □

Lemma G.2. *Under Assumptions 1–2,*

$$\mathbb{P}(\hat{\Sigma} \not\succ 0) \leq C\left\{\frac{L_n}{n} + \frac{1}{n}\right\}. \tag{40}$$

Proof. Let $\bar{n} = \bar{n}_x + \bar{n}_y$ and $V = \mathrm{diag}(v_1, \ldots, v_m)$. For the testing observations define

$$z_{P,a} = V^{-1/2}\{\hat{f}(X_a^{\text{test}}) - \mathbb{E}_\star \hat{f}(X)\}, \qquad 1 \leq a \leq \bar{n}_x,$$
$$z_{Q,a} = V^{-1/2}\{\hat{f}(Y_a^{\text{test}}) - \mathbb{E}_\star \hat{f}(Y)\}, \qquad 1 \leq a \leq \bar{n}_y.$$

Let

$$C_P = \mathbb{E}_\star(z_{P,1} z_{P,1}^\top), \qquad C_Q = \mathbb{E}_\star(z_{Q,1} z_{Q,1}^\top).$$

The population whitened covariance is

$$R_{\text{pop}} = \frac{\bar{n}_y}{\bar{n}} C_P + \frac{\bar{n}_x}{\bar{n}} C_Q = V^{-1/2}\Sigma V^{-1/2} \succeq \kappa I_m.$$

Define the uncentered testing covariance and the within-group mean correction by

$$R_{\text{unc}} = \frac{\bar{n}_y}{\bar{n}\,\bar{n}_x}\sum_{a=1}^{\bar{n}_x} z_{P,a} z_{P,a}^\top + \frac{\bar{n}_x}{\bar{n}\,\bar{n}_y}\sum_{a=1}^{\bar{n}_y} z_{Q,a} z_{Q,a}^\top,$$

$$H_{\text{mean}} = \frac{\bar{n}_y}{\bar{n}}\bar{z}_P \bar{z}_P^\top + \frac{\bar{n}_x}{\bar{n}}\bar{z}_Q \bar{z}_Q^\top,$$

where $\bar{z}_P = \bar{n}_x^{-1} \sum_a z_{P,a}$ and $\bar{z}_Q = \bar{n}_y^{-1} \sum_a z_{Q,a}$. Then

$$V^{-1/2} \hat{\Sigma} V^{-1/2} = R_{\mathrm{unc}} - H_{\mathrm{mean}}. \tag{41}$$

Assumption 2 gives

$$\mathbb{E}_\star \|z_{P,1}\|_2^4 \vee \mathbb{E}_\star \|z_{Q,1}\|_2^4 \le C m^2 L_n.$$

Independence within each testing group therefore yields

$$\mathbb{E}_\star \|R_{\mathrm{unc}} - R_{\mathrm{pop}}\|_{\mathrm{F}}^2 \le \frac{C L_n}{n}. \tag{42}$$

Also,

$$\mathbb{E}_\star \|H_{\mathrm{mean}}\|_{\mathrm{op}} \le \mathbb{E}_\star \operatorname{tr}(H_{\mathrm{mean}}) \le \frac{C}{n} \operatorname{tr}(R_{\mathrm{pop}}) = \frac{Cm}{n}. \tag{43}$$

If $\hat{\Sigma}$ is not positive definite, then (41) and $R_{\mathrm{pop}} \succeq \kappa I_m$ imply

$$\|R_{\mathrm{unc}} - R_{\mathrm{pop}}\|_{\mathrm{op}} \ge \frac{\kappa}{2} \quad \text{or} \quad \|H_{\mathrm{mean}}\|_{\mathrm{op}} \ge \frac{\kappa}{2}.$$

Markov's inequality with (42)–(43) gives (40). □

Lemma G.3. *For $0 < \lambda \le 1$, if*

$$n D_{\mathrm{KFDA},\lambda}(P,Q)^2 \ge \sqrt{1 + N_F(\lambda)}, \tag{44}$$

then

$$\begin{aligned} \mathbb{P}(T_n \le q) \le{}& \mathbb{P}(\hat{\Sigma} \not\succ 0) \\ &+ C_q \left\{ \frac{\sqrt{1+N_F(\lambda)}}{n D_{\mathrm{KFDA},\lambda}(P,Q)^2} + \frac{1+N_F(\lambda)}{n^2 D_{\mathrm{KFDA},\lambda}(P,Q)^4} + \frac{1}{n\lambda^2} \right\}. \end{aligned} \tag{45}$$

For $\lambda \ge 1$,

$$\mathbb{P}(T_n \le q) \le \mathbb{P}(\hat{\Sigma} \not\succ 0) + \frac{C_q(1+\lambda)}{n\lambda^2 D_{\mathrm{KFDA},\lambda}(P,Q)^2}. \tag{46}$$

Proof. The KFDA error bound for $0 < \lambda \le 1$. Define the whitened population signal

$$h_F = A_F^{-1/2}(\mu_{P,F} - \mu_{Q,F}), \qquad \|h_F\| = D_{\mathrm{KFDA},\lambda}(P,Q),$$

and the effective covariance

$$B_F = A_F^{-1/2} C_{r,F} A_F^{-1/2}.$$

Assumption D.3 gives

$$0 \preceq B_F \preceq I, \qquad \operatorname{tr}(B_F) = N_F(\lambda) \le N. \tag{47}$$

Since $\sup_x K_F(x,x) \le 1$ and r is bounded away from 0 and 1,

$$N_F(\lambda) \le \frac{\operatorname{tr}(C_{r,F})}{\lambda} \le \frac{C}{\lambda}. \tag{48}$$

Define the normalized fitting features

$$\xi_{P,a} = r^{-1/2} A_F^{-1/2} \{\phi_F(X_a^{\mathrm{fit}}) - \mu_{P,F}\}, \qquad 1 \le a \le \tilde{n}_x,$$

$$\xi_{Q,a} = (1-r)^{-1/2} A_F^{-1/2} \{\phi_F(Y_a^{\mathrm{fit}}) - \mu_{Q,F}\}, \qquad 1 \le a \le \tilde{n}_y.$$

Let

$$\bar{\xi}_P = \frac{1}{\tilde{n}_x} \sum_{a=1}^{\tilde{n}_x} \xi_{P,a}, \qquad \bar{\xi}_Q = \frac{1}{\tilde{n}_y} \sum_{a=1}^{\tilde{n}_y} \xi_{Q,a},$$

$$E_P = \frac{1}{\tilde{n}_x} \sum_{a=1}^{\tilde{n}_x} \left\{ \xi_{P,a} \otimes \xi_{P,a} - \mathbb{E}(\xi_{P,a} \otimes \xi_{P,a}) \right\},$$

$$E_Q = \frac{1}{\tilde{n}_y} \sum_{a=1}^{\tilde{n}_y} \left\{ \xi_{Q,a} \otimes \xi_{Q,a} - \mathbb{E}(\xi_{Q,a} \otimes \xi_{Q,a}) \right\}.$$

Write $E_F = E_P + E_Q$ and let U_F be the operator from $\mathbb{R}^2$ to $\mathcal{H}_{K_F}$ whose two columns are $\bar{\xi}_P$ and $\bar{\xi}_Q$. Direct expansion gives

$$A_F^{-1/2} \hat{A}_F A_F^{-1/2} = I + E_F - U_F U_F^*, \tag{49}$$

and

$$A_F^{-1/2}\{(\hat{\mu}_{P,F} - \hat{\mu}_{Q,F}) - (\mu_{P,F} - \mu_{Q,F})\} = \sqrt{r}\,\bar{\xi}_P - \sqrt{1-r}\,\bar{\xi}_Q. \tag{50}$$

Let

$$B_{P,F} = \mathbb{E}(\xi_{P,1} \otimes \xi_{P,1}), \qquad B_{Q,F} = \mathbb{E}(\xi_{Q,1} \otimes \xi_{Q,1}).$$

Then $B_{P,F} + B_{Q,F} = B_F$. Since $\|\xi_{P,a}\| \vee \|\xi_{Q,a}\| \le C\lambda^{-1/2}$, independence and Assumption 1 give

$$\mathbb{E}\|E_F\|_{\mathrm{HS}}^2 \le \frac{C N_F(\lambda)}{n\lambda} \le \frac{C}{n\lambda^2}, \tag{51}$$

$$\mathbb{E}\|U_F\|_{\mathrm{HS}}^2 \le \frac{C N_F(\lambda)}{n} \le \frac{C}{n\lambda}. \tag{52}$$

Markov's inequality yields

$$\mathbb{P}\left(\|E_F\|_{\mathrm{op}} > \frac{1}{4} \text{ or } \|U_F\|_{\mathrm{op}}^2 > \frac{1}{4} \right) \le \frac{C}{n\lambda^2}. \tag{53}$$

Moreover,

$$\mathbb{E}\|U_F^* h_F\|^2 \le \frac{C\langle h_F, B_F h_F\rangle}{n} \le \frac{C D_{\mathrm{KFDA},\lambda}(P,Q)^2}{n}. \tag{54}$$

It remains to control the interaction $E_F U_F$. For the P group, put $X_a = \xi_{P,a} \otimes \xi_{P,a} - B_{P,F}$. Then

$$E_P \bar{\xi}_P = \frac{1}{\tilde{n}_x^2} \sum_{a \ne b} X_a \xi_{P,b} + \frac{1}{\tilde{n}_x^2} \sum_a (\|\xi_{P,a}\|^2 I - B_{P,F})\xi_{P,a}.$$

We have

$$\mathbb{E}(X_a^2) \preceq \frac{C}{\lambda} B_{P,F}, \qquad \mathbb{E}\|(\|\xi_{P,a}\|^2 I - B_{P,F})\xi_{P,a}\|^2 \le \frac{C}{\lambda^3}.$$

The diagonal summands need not have mean zero. Since $\mathbb{E}\xi_{P,a} = 0$, $B_{P,F} \preceq I$, and $\mathrm{tr}(B_{P,F}) \le C/\lambda$, Cauchy–Schwarz gives

$$\left\| \mathbb{E}\big[(\|\xi_{P,a}\|^2 I - B_{P,F})\xi_{P,a}\big] \right\|^2 = \left\| \mathbb{E}\big[\|\xi_{P,a}\|^2 \xi_{P,a}\big] \right\|^2 \le \|B_{P,F}\|_{\mathrm{op}}\, \mathbb{E}\|\xi_{P,a}\|^4 \le \frac{C}{\lambda^2}.$$

Thus the squared mean of the normalized diagonal sum contributes at most $C/(n^2\lambda^2)$, and its centered second moment contributes at most $C/(n^3\lambda^3)$. Expanding the squared norm of the off-diagonal sum and using independence and centering gives its bound $C\,\mathrm{tr}(B_{P,F}^2)/(n^2\lambda)$. Combining these bounds yields

$$\mathbb{E}\|E_P \bar{\xi}_P\|^2 \le \frac{C\,\mathrm{tr}(B_{P,F}^2)}{n^2\lambda} + \frac{C}{n^2\lambda^2} + \frac{C}{n^3\lambda^3}.$$

The same bound holds for $E_Q \bar{\xi}_Q$. For the cross-group terms, independence gives

$$\mathbb{E}\|E_P \bar{\xi}_Q\|^2 \le \frac{C\,\mathrm{tr}(B_{P,F} B_{Q,F})}{n^2\lambda}, \qquad \mathbb{E}\|E_Q \bar{\xi}_P\|^2 \le \frac{C\,\mathrm{tr}(B_{Q,F} B_{P,F})}{n^2\lambda}.$$

Since $\mathrm{tr}(B_F^2) \le \mathrm{tr}(B_F) = N_F(\lambda)$, these four bounds give

$$\mathbb{E}\|E_F U_F\|_{\mathrm{HS}}^2 \le C\left\{ \frac{N_F(\lambda)}{n^2\lambda} + \frac{1}{n^2\lambda^2} + \frac{1}{n^3\lambda^3} \right\}. \tag{55}$$

On the event in which the two operator norms in (53) are at most $1/4$,

$$\|(I + E_F)^{-1}\|_{\mathrm{op}} \le \frac{4}{3}, \qquad \|[I_2 - U_F^*(I + E_F)^{-1} U_F]^{-1}\|_{\mathrm{op}} \le \frac{3}{2},$$

and

$$\frac{2}{3} I \preceq (I + E_F - U_F U_F^*)^{-1} \preceq 2I.$$

The two identities

$$(I + E_F - U_F U_F^*)^{-1} U_F = (I + E_F)^{-1} U_F [I_2 - U_F^* (I + E_F)^{-1} U_F]^{-1}$$

and

$$(I + E_F)^{-1} U_F = U_F - (I + E_F)^{-1} E_F U_F$$

together with (49)–(50) give

$$\Delta_{\hat{f}_j}(P, Q) \geq \frac{2}{3} D_{\mathrm{KFDA},\lambda}(P, Q)^2 - C \left\{ \|U_F^* h_F\| + D_{\mathrm{KFDA},\lambda}(P, Q) \|E_F U_F\|_{\mathrm{HS}} \right\}. \tag{56}$$

Therefore

$$\begin{aligned} &\mathbb{P}\left(\Delta_{\hat{f}_j}(P, Q) < \frac{1}{3} D_{\mathrm{KFDA},\lambda}(P, Q)^2 \right) \\ &\quad \leq C \left\{ \frac{1}{n\lambda^2} + \frac{1}{n D_{\mathrm{KFDA},\lambda}(P, Q)^2} \right. \\ &\qquad \left. + \frac{1}{D_{\mathrm{KFDA},\lambda}(P, Q)^2} \left(\frac{N_F(\lambda)}{n^2 \lambda} + \frac{1}{n^2 \lambda^2} + \frac{1}{n^3 \lambda^3} \right) \right\}. \end{aligned} \tag{57}$$

The pooled testing covariance operator of ϕ_F is bounded by $C C_{r,F}$. Hence the KFDA coordinate satisfies

$$v_j \leq C \left\| B_F^{1/2} A_F^{1/2} \hat{A}_F^{-1} (\hat{\mu}_{P,F} - \hat{\mu}_{Q,F}) \right\|^2.$$

Using the same resolvent identities as above gives, on the same good fitting event,

$$v_j \leq C \left\{ D_{\mathrm{KFDA},\lambda}(P, Q)^2 + \|B_F^{1/2} U_F\|_{\mathrm{HS}}^2 + \|E_F U_F\|_{\mathrm{HS}}^2 \right\}. \tag{58}$$

Moreover,

$$\begin{aligned} \mathbb{E}\|B_F^{1/2} U_F\|_{\mathrm{HS}}^2 &= \frac{\mathrm{tr}(B_F B_{P,F})}{\tilde{n}_x} + \frac{\mathrm{tr}(B_F B_{Q,F})}{\tilde{n}_y} \\ &\leq \frac{C \,\mathrm{tr}(B_F^2)}{n} \leq \frac{C N_F(\lambda)}{n}. \end{aligned} \tag{59}$$

Apply Lemma G.1 on the intersection of the good fitting event in (53) with

$$\Delta_{\hat{f}_j}(P, Q) \geq \frac{1}{3} D_{\mathrm{KFDA},\lambda}(P, Q)^2.$$

Use (57)–(59) in Lemma G.1. Then (44) and (48) bound every interaction term by $C/(n\lambda^2)$. This gives (45).

The KFDA error bound for $\lambda \geq 1$. Let

$$\varepsilon_F = (\hat{\mu}_{P,F} - \hat{\mu}_{Q,F}) - (\mu_{P,F} - \mu_{Q,F}).$$

The bounded-kernel condition and Assumption 1 give

$$\mathbb{E}\|\varepsilon_F\|^2 \leq \frac{C}{n}, \qquad \lambda I \preceq \hat{A}_F \preceq C(1 + \lambda) I.$$

On the event

$$\|\varepsilon_F\|^2 \leq c \frac{\lambda}{1 + \lambda} \|\mu_{P,F} - \mu_{Q,F}\|^2$$

with a sufficiently small fixed $c > 0$, Cauchy–Schwarz in the $\hat{A}_F^{-1}$ inner product gives

$$\Delta_{\hat{f}_j}(P, Q) \geq c \frac{\|\mu_{P,F} - \mu_{Q,F}\|^2}{1 + \lambda}, \qquad v_j \leq C \frac{\Delta_{\hat{f}_j}(P, Q)}{\lambda}. \tag{60}$$

The complement of this event has probability at most

$$\frac{C(1 + \lambda)}{n\lambda \|\mu_{P,F} - \mu_{Q,F}\|^2}.$$

Since

$$D_{\mathrm{KFDA},\lambda}(P, Q)^2 \leq \frac{\|\mu_{P,F} - \mu_{Q,F}\|^2}{\lambda},$$

Lemma G.1 and (60) give (46). □