Kernel Selection is Model Selection: A Unified Complexity-Penalized Approach for MMD Two-Sample Tests

Yijin Ni **Xiaoming Huo**
H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology, Atlanta, GA 30332, USA
{yni64, huo}@gatech.edu

Abstract

The Maximum Mean Discrepancy (MMD) is a cornerstone statistic for nonparametric two-sample testing, but its test power is dictated entirely by the chosen kernel. Because the two-sample test induced by any fixed kernel fails to distinguish certain distributions, the kernel must be optimized from data. However, data-driven optimization violates the foundational i.i.d. assumption, forcing a strict trade-off in existing frameworks. Ratio criteria ignore this dependence, inducing overfitting and variance collapse on rich kernel classes. Conversely, aggregation methods bypass the dependence using finite grids, but this strategy cannot scale to continuous search spaces like deep kernels. To break this dichotomy, we establish data-driven kernel selection as a model selection problem. We propose Complexity-Penalized MMD (CP-MMD), a criterion derived by applying the two-sample uniform concentration inequality of Ni and Huo [2024] to the post-optimization MMD problem. The resulting penalty bounds the empirical MMD by the complexity of the kernel search space, mathematically absorbing the cost of optimization, so that CP-MMD enables direct, grid-free maximization over continuous parametric classes, including scalar bandwidths, polynomial feature bandwidths, and deep network parameters. By formally accounting for optimization complexity, we prove a finite-sample lower bound on the post-selection test power; the deployed criterion, which replaces the worst-case constants of this bound with calibrated ones, empirically maximizes test power, and sample splitting ensures unconditional Type-I validity. Consequently, CP-MMD enables grid-free kernel selection across linear, polynomial-feature, and deep regimes, matching or exceeding state-of-the-art test power.

1 Introduction

Two-sample testing, which asks whether two datasets $\mathbf{X} \sim P$ and $\mathbf{Y} \sim Q$ are drawn from the same distribution, is a fundamental problem in statistics and machine learning. It arises in *dataset shift detection*, where P is the training data and Q the deployment data; in *generative model evaluation*, where P is the data and Q the model distribution [Sutherland et al., 2017]; and in *industrial process monitoring*, where P is normal operation and Q a potential fault [Downs and Vogel, 1993].

The Maximum Mean Discrepancy [MMD; Gretton et al., 2012a] is a standard kernel-based statistic for two-sample testing. Its effectiveness depends critically on the choice of kernel: whether the test is powerful depends on how well the kernel aligns with the specific distributional difference between P and Q [Ramdas et al., 2015, Sutherland et al., 2017]. No single fixed kernel is therefore uniformly well suited to all alternatives encountered in practice, making data-driven kernel selection essential.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

However, this data-driven selection introduces statistical dependence, the primary technical obstacle in MMD-based testing. The asymptotic distribution theory and related power guarantees of the empirical MMD rely strictly on the assumption that the kernel is chosen independently of the test samples. Dynamically optimizing the kernel using the observed data violates this assumption, invalidating classical fixed-kernel asymptotics. Consequently, kernel selection cannot be treated merely as an algorithmic preprocessing step; its statistical impact on the test distribution must be explicitly resolved. Existing frameworks attempt to address this dependence through two primary paradigms, both of which force a strict statistical trade-off.

1.1 Existing approaches and their limits

First, **ratio criteria** [Sutherland et al., 2017, Liu et al., 2020] maximize the empirical MMD normalized by its estimated standard deviation, an objective derived from asymptotic theory under the assumption that the kernel is selected independently of the input samples. Applying this data-independent proxy to rich, data-driven kernel classes introduces two critical failures. Because the optimization breaks the independence assumption, the framework severely overfits; the optimizer artificially inflates the ratio on the observed sample without improving actual test power. Furthermore, the unconstrained objective is vulnerable to *variance collapse* (documented empirically in App. C, Table 3): the optimizer frequently drives both the empirical MMD and its variance to zero, producing a spuriously large ratio that selects degenerate kernels. Second, rather than addressing the intractable distribution of the empirical MMD equipped with a data-driven kernel, **aggregation methods** [Schrab et al., 2023, Biggs et al., 2023] combine the statistics from a finite, pre-specified grid of candidate kernels. This allows the test to leverage the best-performing kernel for a specific sample while using a penalty to correct for the multiple comparisons. While tractable for low-dimensional parameters such as scalar bandwidths, this strategy cannot scale to polynomial features or learned deep representations, as finite grids cannot effectively cover continuous search spaces. A third line of recent work improves the test for a kernel that is already chosen: a martingale construction with a Gaussian null limit that avoids permutations [Chatterjee and Ramdas, 2026], non-asymptotic critical values for a spectrally truncated normalized MMD [Lacroix et al., 2026], truncation to leading directional components with bootstrap calibration [Cui et al., 2025], and random Fourier features that trade power for computation [Choi and Kim, 2024]. These methods operate downstream of kernel selection, so they can be applied to the kernel that CP-MMD selects.

1.2 Our approach: a single complexity-penalized criterion

Composite kernel unification. We demonstrate that data-driven kernel selection paradigms across linear, polynomial, and deep regimes can be mathematically unified under a single composite-kernel formulation: $k_h(x, x') := k(h(x), h(x'))$, where k is a bounded-Lipschitz base kernel and $h \in \mathcal{H}$ is a learnable feature map. As shown in Table 1 (§3.1), this formulation covers bandwidth selection (linear, $h : x \mapsto x/\sigma$), polynomial features (Laplacian on degree-p features), and deep representations ($h = h_\theta$). Liu's ϵ-mixture kernel lies outside this form, because ϵ varies inside the kernel rather than inside the feature map; App. J treats it separately.

Kernel selection as model selection. Treating kernel selection as model selection in $\mathcal{H}$ lets us import statistical-learning tools designed for the post-optimization distribution problem. Consider random vectors $X \sim P$ and $Y \sim Q$, with corresponding sample matrices $\mathbf{X} \sim P^m$ and $\mathbf{Y} \sim Q^n$. Under the composite framework, the kernel selection problem reduces to selecting an optimal feature map h applied directly to the input random vectors X and Y. Let $\widehat{\gamma}^2_{k,u}(h)$ denote the empirical squared MMD between the transformed samples $h(\mathbf{X})$ and $h(\mathbf{Y})$, and let $\gamma^2_k(h)$ be its population counterpart. This equivalence allows us to formally resolve the post-optimization dependence by applying a uniform concentration inequality (UCI) for two-sample statistics. Specifically, Ni and Huo [2024] establish a UCI that extends Maurer and Pontil [2019]'s nonlinear-functional concentration framework to the two-sample setting and applies to the unbiased MMD estimator $\widehat{\gamma}^2_{k,u}(h)$, bounding the generalization gap $\sup_{h\in\mathcal{H}} |\widehat{\gamma}^2_{k,u}(h) - \gamma^2_k(h)|$ at a rate determined by the Gaussian complexity $G(\mathcal{H})$. Because this bound holds uniformly over the entire search space, it remains strictly valid for the data-driven choice $\widehat{h}$. Applying it to the post-selection MMD problem establishes a finite-sample lower bound on the post-selection test power and gives the CP-MMD criterion:

$$J_{\mathrm{CP}}(h) := \widehat{\gamma}^2_{k,u}(h) - \widehat{C}_1 \cdot \widetilde{G}(h), \tag{1}$$

where $\widetilde{G}(h)$ is a computable bound on the complexity along the optimization trajectory [Bartlett et al., 2017] and $\widehat{C}_1$ is a constant calibrated from data by null permutations of the training half (App. F), distinct from the worst-case constant C_1 of the UCI; the held-out half is reserved for the permutation test, which has exact finite-sample Type-I control by sample-split exchangeability.

Grid-free continuous optimization. CP-MMD needs no discrete candidate grid: it incorporates the statistical cost of continuous optimization directly into the complexity penalty $\widetilde{G}(h)$. State-of-the-art aggregation methods such as MMDAgg and MMD-FUSE rely on a pre-specified finite set of B candidate kernels and pay a $\sqrt{\ln B/N}$ Bonferroni-style multiplicity penalty; this discrete union bound is undefined in the continuous limit ($\alpha/B \to 0$), so aggregation cannot be applied to continuous parametric classes such as $\mathcal{H}_{\text{deep}}$.

Contributions.

1. *Unified grid-free criterion.* We introduce J_{CP}, a single penalized objective that replaces discrete multiplicity-corrected grids with a continuous optimization framework. It unifies kernel parametrizations from scalar scales to neural feature maps under a single model class.
2. *Theoretical power guarantees.* We derive a finite-sample lower bound on the power of the held-out test at any kernel selected from a fixed class on the training half (Cor. 3.6) by applying the two-sample uniform concentration inequality of Ni and Huo [2024] (restated here as Theorem 3.4) to the post-selection MMD problem. The bound is stated with the worst-case constant C_1; the deployed criterion replaces it by the calibrated $\widehat{C}_1$, so the bound motivates J_{CP} rather than certifying the deployed test. To our knowledge, this represents the first theoretical analysis of post-selection test power for optimized empirical MMD.
3. *Architectural efficiency.* CP-MMD retains only the single optimized kernel at test time, resulting in a per-test deployment cost of $O(N_{\text{perm}} \cdot N^2)$ that is independent of candidate class size. This bypasses the B-dependency inherent in aggregation methods: MMDAgg scales as $O(B \cdot N_{\text{perm}} \cdot N^2)$ because each bootstrap iteration re-evaluates all B candidates, while MMD-FUSE scales as $O((B + N_{\text{perm}}) \cdot N^2)$ via shared distance matrices. Unlike these baselines, CP-MMD maintains a constant deployment cost as the search space expands.

Empirical validation. A single criterion beats every regime-specific specialist (§4.2): the median heuristic on multi-scale data (0.99 vs. 0.53 at mean shift Δ=0.30); MMDAgg on the mildest kurtosis shift (high-dim Student-t, df=20 degrees of freedom: 0.60 vs. 0.37); and Liu/Plain on the same MLP across high-dimensional Gaussian mean-shift cells, where CP-MMD dominates Liu (variance collapse) and Plain (overfitting at small n).

2 Preliminaries

Two-sample test. Given independent samples $\mathbf{X} = (X_1, \ldots, X_m) \overset{\text{iid}}{\sim} P$ and $\mathbf{Y} = (Y_1, \ldots, Y_n) \overset{\text{iid}}{\sim} Q$ in $\mathbb{R}^d$. A two-sample test aims to determine whether the difference between the population distributions P and Q is statistically significant. The hypotheses are

$$H_0 : P = Q, \quad \text{and} \quad H_1 : P \neq Q.$$

Denote the decision function as $\phi(\mathbf{X}, \mathbf{Y}) \in \{0, 1\}$, mapping the observed data to a binary reject/accept decision for the null hypothesis H_0 (with $\phi = 1$ denoting rejection). The **significance level** of the test, α, is the threshold probability used in hypothesis testing to determine whether to reject the null hypothesis H_0, that is,

$$\alpha := \sup_{H_0} \mathbb{E}[\phi(\mathbf{X}, \mathbf{Y})].$$

The **test power** at a specific alternative (P, Q) with $P \neq Q$ is

$$\beta(P, Q) := \mathbb{E}_{\mathbf{X} \sim P^m, \mathbf{Y} \sim Q^n}[\phi(\mathbf{X}, \mathbf{Y})],$$

i.e., the probability of rejecting H_0 under that alternative. We treat power point-wise per alternative; a global infimum over H_1 is degenerate because H_1 contains alternatives arbitrarily close to the null,

where any level-α test has rejection probability close to α. Decreasing the significance level (α) inherently reduces statistical power $\beta(P, Q)$ for a fixed sample size [Casella and Berger, 2002]. Due to this trade-off, the Neyman-Pearson lemma [Neyman and Pearson, 1933] establishes the standard optimization procedure: select a decision function ϕ that maximizes $\beta(P, Q)$ subject to a fixed α.

Maximum Mean Discrepancy. Let $\mathcal{U} \subseteq \mathbb{R}^d$, and let $k : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ be a reproducing kernel with an induced reproducing kernel Hilbert space (RKHS) $\mathcal{F}_k$. For probability measures P and Q defined on $\mathcal{U}$, the maximum mean discrepancy (MMD) between P and Q is [Gretton et al., 2012a]:

$$\gamma_k^2(P, Q) = \mathbb{E}[k(X, X')] + \mathbb{E}[k(Y, Y')] - 2\mathbb{E}[k(X, Y)], \tag{2}$$

where $X, X' \overset{\text{iid}}{\sim} P$ and $Y, Y' \overset{\text{iid}}{\sim} Q$. Given independent samples $\mathbf{X} = (X_1, \ldots, X_m) \overset{\text{iid}}{\sim} P$ and $\mathbf{Y} = (Y_1, \ldots, Y_n) \overset{\text{iid}}{\sim} Q$, an unbiased estimator for $\gamma_k^2(P, Q)$ is [Gretton et al., 2012a]:

$$\widehat{\gamma}_{k,u}^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{m(m-1)} \sum_{i \neq j} k(X_i, X_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(Y_i, Y_j) - \frac{2}{mn} \sum_{i,j} k(X_i, Y_j). \tag{3}$$

If k is a characteristic kernel [Sriperumbudur et al., 2010], $\gamma_k^2(P, Q) = 0$ if and only if $P = Q$. Consequently, $\widehat{\gamma}_{k,u}^2$ serves as a test statistic for the two-sample problem [Gretton et al., 2012a]. Let c_α denote the $(1-\alpha)$-quantile of $\widehat{\gamma}_{k,u}^2$ under the null hypothesis $H_0 : P = Q$. A level-α MMD test rejects H_0 when $\widehat{\gamma}_{k,u}^2 \geq c_\alpha$, corresponding to the decision function:

$$\phi_{\widehat{\gamma}_{k,u}^2}(\mathbf{X}, \mathbf{Y}) = \mathbb{I}(\widehat{\gamma}_{k,u}^2 \geq c_\alpha).$$

Necessity of kernel selection. While constructing a decision function ϕ that bounds the significance level at α is straightforward via the aforementioned empirical thresholding, the statistical power of the resulting test is highly sensitive to the choice of the kernel k [Gretton et al., 2012b]. Specifically, no single fixed kernel achieves uniform test power against multi-scale alternatives [Ramdas et al., 2015, Sutherland et al., 2017, Schrab et al., 2023]. For any predetermined kernel bandwidth, there exist distinct alternative distributions ($P \neq Q$) for which the population MMD is arbitrarily small, causing the test's power to degrade toward the significance level α [Schrab et al., 2023], formally:

Lemma 2.1 (Asymptotic triviality of fixed-bandwidth MMD tests)**.** *For any fixed bandwidth $\sigma > 0$, let $k_\sigma(x, x') := \exp(-\|x - x'\|^2/(2\sigma^2))$ be the Gaussian kernel defined on $\mathbb{R}^d$. There exists a sequence of distribution pairs $(P^{(i)}, Q^{(i)})$ defined on $\mathbb{R}^d$ for $i = 1, 2, \ldots$, and a constant $c_0 > 0$, such that the total variation distance $\mathrm{TV}(P^{(i)}, Q^{(i)}) \geq c_0$ for all i, while the population MMD converges to zero:*

$$\gamma_{k_\sigma}^2(P^{(i)}, Q^{(i)}) \longrightarrow 0 \quad \textit{as} \quad i \to \infty.$$

Consequently, for samples $\mathbf{X}^{(i)} \overset{iid}{\sim} P^{(i)}$ and $\mathbf{Y}^{(i)} \overset{iid}{\sim} Q^{(i)}$, the rejection probability satisfies:

$$\limsup_{i \to \infty} \Pr_{(P^{(i)}, Q^{(i)})} \left(\widehat{\gamma}_{k,u}^2(\mathbf{X}^{(i)}, \mathbf{Y}^{(i)}) \geq c_\alpha(\mathbf{X}^{(i)}, \mathbf{Y}^{(i)}) \right) \leq \alpha,$$

where $c_\alpha(\mathbf{X}^{(i)}, \mathbf{Y}^{(i)})$ is the $(1-\alpha)$-quantile of the permutation null distribution.

The lemma establishes that the asymptotic test power β is bounded by the significance level α on the constructed sequence: the test performs no better than random guessing on these strictly separated alternatives. The proof, taking $P^{(i)} = \mathcal{N}(0, i^2)$ and $Q^{(i)} = \mathcal{N}(0, (2i)^2)$ on $\mathbb{R}$, is given in Appendix B. Choosing k from data is therefore the central practical problem of MMD testing.

3 The CP-MMD Criterion

3.1 Composite kernels: a model class for kernel selection

CP-MMD operates on the class of bounded-Lipschitz base kernels (Assumption 3.2), lifted by a feature map h drawn from a model class $\mathcal{H}$ to form the composite kernel $k_h(x, x') = k(h(x), h(x'))$ (Definition 3.3). This form covers the three canonical kernel-selection settings used in modern MMD testing: bandwidth selection, polynomial features, and deep features (formal proxies in §3.3). Liu's ϵ-mixture is not a special case of Definition 3.3 (ϵ varies inside the base kernel, not inside h); we apply J_{CP} heuristically to it in Appendix J's ablation, where it still outperforms Liu's ratio criterion.

Table 1: The composite-kernel framework $k_h(x,x') := k(h(x),h(x'))$ accommodates the existing methods we compare against in §4. Denote the Gaussian kernel of bandwidth σ by $k^{\mathrm{G}}_\sigma(x,y) := \exp\big(-\|x-y\|^2/(2\sigma^2)\big)$ and the Laplacian kernel of bandwidth σ by $k^{\mathrm{L}}_\sigma(x,y) := \exp(-\|x-y\|/\sigma)$. The bandwidth selection procedure proposed in Gretton et al. [2012a] corresponds to the linear model class $\mathcal{H}$. MMDAgg/MMD-FUSE [Schrab et al., 2023, Biggs et al., 2023] aggregate over a finite grid of Laplacian and Gaussian kernels at a discrete set of bandwidths (column 2); the third column shows the *continuous* composite-kernel form CP-MMD covers (Laplacian base kernel composed with polynomial-feature lift), which strictly extends the discrete grid by allowing the bandwidth σ to vary continuously in $[\sigma_l, \sigma_u]$. The definition of Ψ_p appears in (6). The abbreviated version of the deep kernel proposed in Liu et al. [2020] corresponds to the model class $\mathcal{H}$ composed of MLPs (the $q = \mathbf{I}$, $\epsilon = 0$ special case). Liu's full ϵ-mixture lies outside Definition 3.3's k_h form; the empirical comparison between the simplified and full deep kernels is in Appendix J.

Reference	**Feasible kernel set**	$\{k_h : h \in \mathcal{H}\}$
Gretton et al. [2012a]	$\{k^{\mathrm{G}}_\sigma(x,y), \sigma \in [\sigma_l, \sigma_u]\}$	$\{k^{\mathrm{G}}_1(x/\sigma, y/\sigma) : \sigma \in [\sigma_l, \sigma_u]\}$
MMDAgg [Schrab et al., 2023] MMD-FUSE [Biggs et al., 2023]	$\{k^{\mathrm{G}}_{\sigma_j}(x,y),\ k^{\mathrm{L}}_{\sigma_j}(x,y)\ :\ j = 1,\dots,B\}$	$\{k^{\mathrm{L}}_1(\Psi_p(x)/\sigma, \Psi_p(y)/\sigma) : \sigma \in [\sigma_l, \sigma_u]\}$
Liu et al. [2020]	$\{k(x,y) = k^{\mathrm{G}}_\sigma(\phi_\theta(x), \phi_\theta(y)) : (\sigma,\theta) \in \Theta\}$	$\{k^{\mathrm{G}}_1(\phi_\theta(x)/\sigma, \phi_\theta(y)/\sigma) : (\sigma,\theta) \in \Theta\}$

Assumption 3.2 (Kernel regularity)**.** *$k : \mathcal{U} \times \mathcal{U} \to \mathbb{R}$ is positive-semidefinite and satisfies, for some constants $\nu > 0$ and $l > 0$:* (i) Boundedness: $0 \le k(u,u') \le \nu$ *for all* $u, u' \in \mathcal{U}$. (ii) Lipschitz in each argument: *for all* $u, \tilde{u}, u' \in \mathcal{U}$, $|k(u,u') - k(\tilde{u},u')| \le l\|u - \tilde{u}\|$, *and likewise in u' by symmetry.*

Definition 3.3 (Composite kernel and MMD)**.** *Let $\mathcal{H}$ be a class of measurable maps $h : \mathbb{R}^d \to \mathcal{U}$, the kernel-selection class. For each $h \in \mathcal{H}$, the* composite kernel *is*

$$k_h(x,x') \ :=\ k\big(h(x),\, h(x')\big), \tag{4}$$

and the composite MMD *at h is the population MMD evaluated under h's push-forward:*

$$\gamma^2_k(h) \ :=\ \gamma^2_k(h_\# P,\, h_\# Q) \ =\ \mathbb{E}\, k_h(X,X') + \mathbb{E}\, k_h(Y,Y') - 2\,\mathbb{E}\, k_h(X,Y). \tag{5}$$

The unbiased estimator $\widehat{\gamma}^2_{k,u}(h)$ is given by (3) *with k replaced by k_h.*

The degree-p *polynomial feature map* $\Psi_p : \mathbb{R}^d \to \mathbb{R}^{D_p}$ is defined by

$$\Psi_p(x) \ :=\ \big(x^\alpha\big)_{1 \le |\alpha| \le p} \ \in\ \mathbb{R}^{D_p}, \qquad D_p \ :=\ \binom{d+p}{p} - 1, \tag{6}$$

where $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$ is a multi-index, $|\alpha| := \sum_{i=1}^d \alpha_i$, and $x^\alpha := \prod_{i=1}^d x_i^{\alpha_i}$.

For example, taking k Gaussian and $h(x) = x/\sigma$ recovers the Gaussian kernel at bandwidth σ; the three canonical regimes (linear/polynomial/deep) are listed in §3.3 with their computable proxies. Under this formulation, kernel selection reduces to selecting $\widehat{h}$ from $\mathcal{H}$, i.e., model selection in $\mathcal{H}$.

3.2 Empirical-to-population gap: the uniform concentration inequality

According to Lemma 2.1, a data-driven kernel is required to maintain the test power. Nevertheless, the resulting $k_{\widehat{h}}$ violates the independence assumption of classical MMD theory [Gretton et al., 2012a], and the standard CLT-based threshold no longer characterizes rejection probability. We restore a finite-sample guarantee via the **uniform concentration inequality** (UCI) on $|\widehat{\gamma}^2_{k,u}(h) - \gamma^2_k(h)|$ over $h \in \mathcal{H}$ established by Ni and Huo [2024], restated below for self-containment.

Theorem 3.4 (Two-sample UCI; restated from Ni and Huo [2024])**.** *Let $\mathbf{X} \sim P^m$ and $\mathbf{Y} \sim Q^n$ with pooled data $\mathbf{D} = \mathbf{X} \sqcup \mathbf{Y}$ as in §2, set $N = m + n$, and let $\rho_* := \max\{N/m, N/n\}$ denote the sample-size imbalance ratio (so $\rho_* \ge 2$, with equality if and only if $m = n$). Under Assumption 3.2, for any $\delta \in (0,1)$, with probability at least $1 - \delta$:*

$$\sup_{h \in \mathcal{H}} \big|\widehat{\gamma}^2_{k,u}(h) - \gamma^2_k(h)\big| \le B_N(\delta) := \underbrace{C_1\, G(\mathcal{H})}_{\textit{complexity}} + \underbrace{C_2 \sqrt{\ln(2/\delta)/N}}_{\textit{concentration}}, \tag{7}$$

where $C_1 = 2\sqrt{2\pi}\, l\, \rho_*(1+\rho_*)$, $C_2 = 4\nu\rho_*$, and $G(\mathcal{H})$ is the (data-averaged) Gaussian complexity of the model class $\mathcal{H}$:

$$G(\mathcal{H}) \;:=\; \mathbb{E}_{\mathbf{D}}\,\mathbb{E}_{\boldsymbol{\xi}}\left[\sup_{h\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N}\langle \boldsymbol{\xi}_i,\, h(\mathbf{D}_i)\rangle\right],$$

with $\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_N$ i.i.d. standard Gaussian vectors in the ambient Euclidean space of $\mathcal{U}$ (i.e., each $\boldsymbol{\xi}_i \in \mathbb{R}^{d_{\text{out}}}$ where $d_{\text{out}} := \dim(\mathcal{U})$) and the outer expectation taken over the data distribution.

Specialized to a finite kernel collection $\mathcal{K} = \{k_1, \ldots, k_B\}$ via Massart's finite-class union bound, our UCI recovers the $\Theta(\sqrt{\ln B/N})$ separation rate that Schrab et al. [2023, Thm. 9] establish as minimax-tight on the discrete-grid regime (Prop. I.9, App. I). This confirms our bound is rate-tight where a sharp comparison is known and is not a power-advantage claim over MMDAgg.

From UCI to a power certificate. The UCI of Theorem 3.4 is a population-vs-empirical bound that holds uniformly over the class $\mathcal{H}$, which is fixed before the data are observed, so it holds at every data-dependent selection $\widehat{h} \in \mathcal{H}$. Rearranging (7), with probability at least $1-\delta$,

$$\gamma_k^2(h) \;\geq\; \widehat{\gamma}_{k,u}^2(h) - C_1\, G(\mathcal{H}) - C_2\sqrt{\ln(2/\delta)/N} =: \mathcal{L}_\delta(h) \quad \text{for all } h \in \mathcal{H}. \tag{8}$$

If $\mathcal{L}_\delta(\widehat{h}) > 0$ at the selected kernel, then $\gamma_k^2(\widehat{h}) > 0$ with probability at least $1-\delta$, and Corollary 3.6 turns this into a lower bound on the power of the held-out test. The term $C_1\, G(\mathcal{H})$ is the same for every candidate, whereas an optimizer visits only its trajectory $\mathcal{H}_T := \{h^{(0)}, \ldots, h^{(T)}\} \subseteq \mathcal{H}$, whose complexity is much smaller [Bartlett et al., 2005, Sec. 3]. §3.3 therefore replaces $C_1\, G(\mathcal{H})$ by a calibrated, iterate-dependent proxy: the formal guarantee is (8), and the deployed criterion is its computable surrogate, evaluated empirically in §4.

3.3 Calibrated complexity penalty: the deployed criterion

The complexity of the realized trajectory has no closed form. Proposition 3.5 bounds it by an envelope of the visited iterates, and we pair the resulting per-iterate proxy $\widetilde{G}$ with a calibrated constant $\widehat{C}_1$ that absorbs the slack from this substitution.

Proposition 3.5 (Envelope bound on the trajectory complexity)**.** *Fix the pooled sample $\mathbf{D}$, and let $\widehat{G}(\mathcal{H}_T) := \mathbb{E}_{\boldsymbol{\xi}}\big[\sup_{h\in\mathcal{H}_T} \frac{1}{N}\sum_{i=1}^{N}\langle \boldsymbol{\xi}_i, h(\mathbf{D}_i)\rangle\big]$ be the empirical Gaussian complexity of the trajectory $\mathcal{H}_T = \{h^{(0)}, \ldots, h^{(T)}\}$, with $\boldsymbol{\xi}_i$ as in Theorem 3.4. Write $\|h(\mathbf{D})\|_F := (\sum_{i=1}^{N}\|h(\mathbf{D}_i)\|^2)^{1/2}$. Then*

$$\widehat{G}(\mathcal{H}_T) \;\leq\; \sqrt{2\ln(T+1)}\, \max_{t\leq T} \tfrac{\|h^{(t)}(\mathbf{D})\|_F}{N} \;\leq\; \sqrt{2\ln(T+1)}\, L^* \cdot \tfrac{\|\mathbf{D}\|_F}{N}, \tag{9}$$

where the second inequality holds whenever each $h^{(t)}$ is $L(h^{(t)})$-Lipschitz with $h^{(t)}(0) = 0$ and L^* is any finite constant with $L^* \geq \max_{t\leq T} L(h^{(t)})$.

The trajectory is finite, so such an envelope L^* always exists without any constraint on the optimizer; the first inequality needs no Lipschitz condition at all. The proposition bounds the complexity of the realized trajectory conditionally on the data and motivates the proxy below; the formal guarantee of Corollary 3.6 is stated with $G(\mathcal{H})$ of the fixed class.

The per-iterate proxy is

$$\widetilde{G}(h) := \begin{cases} \|\mathbf{D}\|_F/(\sigma N) & \mathcal{H}_{\text{linear}} := \{x \mapsto x/\sigma\} \\ \|\Psi_p(\mathbf{D})\|_F/(\sigma N) & \mathcal{H}_{\text{polynomial}} := \{x \mapsto \Psi_p(x)/\sigma\} \\ L(h_\theta)\,\|\mathbf{D}\|_F/N, \;\; L(h_\theta) = \prod_{j=1}^{L}\|W_j\|_2/\sigma & \mathcal{H}_{\text{deep}} := \{h_\theta : \theta \in \Theta\} \end{cases} \tag{10}$$

The linear and polynomial proxies equal $\|h(\mathbf{D})\|_F/N$, the deep proxy upper-bounds it by Lemma H.8, and the polynomial proxy requires no bound on the inputs. In the deep regime, σ is the bandwidth of the Gaussian base kernel applied to the network output; placing it inside h_θ matches the other two regimes, where the bandwidth also enters through h. The proxy replaces the envelope of Proposition 3.5 by its value at the current iterate: $L(h^{(t)})$ is read off the realized weights, which is finite for any network used in practice, and $\widetilde{G}(h^{(t)}) \leq L^*\,\|\mathbf{D}\|_F/N$ for every iterate. The

spectral stabilization of App. E (Fig. 4(d)) is an empirical observation about the realized $L(h^{(t)})$, not an assumption of any result. The factor $\sqrt{2\ln(T+1)}$ is absorbed into the calibrated $\widehat{C}_1$. Our deep-kernel implementation, including its calibration, uses $\|\mathbf{D}\|_F/\sqrt{N}$ in place of $\|\mathbf{D}\|_F/N$; deep-kernel values of $\widehat{C}_1$ are reported on this implementation scale and convert to the scale of (10) on multiplication by $\sqrt{N}$, with N the size of the pooled training half.

Calibrating C_1 by null permutations. The Gaussian complexity $G(\mathcal{H})$ in Theorem 3.4 is the *global* complexity of the entire class, much larger than that of the realized trajectory (§3.2). Calibration absorbs this gap: instead of the $C_1 = 2\sqrt{2\pi}\, l\, \rho_*(1+\rho_*) \approx 30$ of Theorem 3.4, we calibrate a single constant $\widehat{C}_1$ that pairs with $\widetilde{G}$ to give an operationally tight bound.

According to the expectation version of the UCI under H_0: $\mathbb{E}[\sup_h \widehat{\gamma}^2_{k,u}(h)] \le C_1\, G(\mathcal{H})$ (App. F.1), C_1 *upper-bounds* the ratio of expected MMD to expected Gaussian complexity; we estimate the operationally tight value by the $(1-\alpha)$-quantile of null-permutation ratios (a high-probability sharpening of the expectation bound; see App. F.3), that is,

$$\widehat{C}_1 \;:=\; \mathrm{Quantile}_{1-\alpha}\big\{\widehat{\gamma}^2_{k,u}(\widehat{h}_\pi)/\widetilde{G}(\widehat{h}_\pi)\big\}_{\pi=1}^{n_{\mathrm{cal}}}, \tag{11}$$

where $\widehat{h}_\pi := \arg\max_{h\in\mathcal{H}} \widehat{\gamma}^2_{k,u}(h)$ on the π-permuted training half is the *plain-MMD maximizer*. Plain-MMD maximization replaces J_{CP} ascent inside calibration on two grounds. First, the J_{CP} objective itself contains $\widehat{C}_1$ through its penalty term $\widehat{C}_1\,\widetilde{G}(h)$, so running J_{CP} during calibration would require $\widehat{C}_1$ to satisfy a fixed-point equation rather than admit the closed-form quantile estimator in (11). Second, solving it would multiply the calibration cost. Both trajectories are finite and admit an envelope L^* as in Proposition 3.5; App. F.2 gives the full procedure.

Across our deep-kernel experiments, the calibrated values satisfy $\widehat{C}_1 \in [0.0005, 0.014]$ on the implementation scale and $\sqrt{N}\,\widehat{C}_1 \le 0.15$ on the scale of (10), at least two orders of magnitude, abbreviated OOM, below the worst-case C_1. The substitution is grounded in the UCI's expectation form (14): the sharpest applicable constant $\widehat{C}_1^*$ (eq. (15)) is a class invariant depending only on $(\mathcal{H}, P, Q, \alpha)$, upper-bounded by $C_1 \approx 30$ but typically much smaller (App. F.1). The calibration procedure of App. F.2 estimates $\widehat{C}_1^*$ via Monte Carlo on null permutations; App. F.6's seven-OOM ablation confirms the deployed test's power is insensitive to $\widehat{C}_1$ across $\widehat{C}_1 \in [10^{-4}, 10]$, and App. F.8's sweeps show that it varies by at most $1.52\times$ across architectures and $1.73\times$ across shifts, so one calibration serves repeated deployments in the same setup.

The deployed criterion. Replacing the class complexity $C_1\, G(\mathcal{H})$ in $\mathcal{L}_\delta(h)$ by the calibrated per-iterate proxy $\widehat{C}_1\,\widetilde{G}(h)$ and dropping the h-independent concentration term yields J_{CP}. The two objectives, side by side, are

$$\underbrace{\mathcal{L}_\delta(h) = \widehat{\gamma}^2_{k,u}(h) - C_1\, G(\mathcal{H}) - C_2\sqrt{\ln(2/\delta)/N}}_{\text{certified; worst-case constant, global complexity}} \qquad \underbrace{J_{\mathrm{CP}}(h) \;=\; \widehat{\gamma}^2_{k,u}(h) \;-\; \widehat{C}_1 \cdot \widetilde{G}(h)}_{\text{deployed; calibrated constant, per-iterate proxy}}. \tag{12}$$

$\mathcal{L}_\delta$ carries the guarantee of Cor. 3.6, but its penalty is the same for every candidate; J_{CP} is computable and its penalty varies with h. J_{CP} is also the Lagrangian of maximizing $\widehat{\gamma}^2_{k,u}(h)$ subject to $\widetilde{G}(h) \le b$ with multiplier $\widehat{C}_1$, consistent with the stabilization level of App. E decreasing in $\widehat{C}_1$.

Deployment protocol. For a valid test at the selected kernel, we use the sample-split plus permutation protocol of Liu et al. [2020]. Specifically, partition $\mathbf{D}$ uniformly at random into halves $\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{te}}$. With $\widehat{C}_1$ from (11), run T gradient-ascent steps on J_{CP} over $\mathcal{D}_{\mathrm{tr}}$ and select $\widehat{h} := \arg\max_{0\le t\le T} J_{\mathrm{CP}}(h^{(t)})$. Then run the level-$\alpha$ permutation test [Lehmann and Romano, 2005, Thm. 15.2.1] at $k_{\widehat{h}}$ on $\mathcal{D}_{\mathrm{te}}$ with N_{perm} permutations. Type-I validity follows from sample-splitting: $\widehat{h}$ depends only on $\mathcal{D}_{\mathrm{tr}}$, so $\mathcal{D}_{\mathrm{te}}$ is exchangeable under H_0. Algorithm 1 summarizes the procedure; for the scalar-bandwidth classes, the maximizations are grid searches over σ.

Corollary 3.6 (Post-selection power)**.** *Let $\mathcal{H}$ satisfy the conditions of Theorem 3.4 and be fixed before the data are observed, and let $\widehat{h} \in \mathcal{H}$ be any kernel selected from the training half, of size N_{tr}, for*

Algorithm 1 CP-MMD: calibrated kernel selection and held-out test

Require: samples $\mathbf{X}, \mathbf{Y}$; class $\mathcal{H}$; level α; steps T; n_{cal}; N_{perm}

Split $\mathbf{D}$ into halves $\mathcal{D}_{\mathrm{tr}}$ and $\mathcal{D}_{\mathrm{te}}$, stratified by sample
for $\pi = 1, \ldots, n_{\mathrm{cal}}$: permute the pooled $\mathcal{D}_{\mathrm{tr}}$ into $(\mathbf{X}^\pi, \mathbf{Y}^\pi)$ at the original sizes **do**
    $\widehat{h}_\pi \leftarrow$ plain empirical-MMD maximizer on $(\mathbf{X}^\pi, \mathbf{Y}^\pi)$; $r_\pi \leftarrow \widehat{\gamma}^2_{k,u}(\widehat{h}_\pi)/\widetilde{G}(\widehat{h}_\pi)$
$\widehat{C}_1 \leftarrow$ empirical $(1-\alpha)$-quantile of $\{r_\pi\}_{\pi=1}^{n_{\mathrm{cal}}}$
Run T ascent steps on $J_{\mathrm{CP}} = \widehat{\gamma}^2_{k,u} - \widehat{C}_1 \widetilde{G}$ over $\mathcal{D}_{\mathrm{tr}}$; $\widehat{h} \leftarrow \arg\max_{0 \le t \le T} J_{\mathrm{CP}}(h^{(t)})$
return the level-α permutation test with kernel $k_{\widehat{h}}$ on $\mathcal{D}_{\mathrm{te}}$ with N_{perm} permutations

example by maximizing J_{CP}. Let N_{ho} be the size of the held-out half and $\delta, \delta' \in (0,1)$. On the event that

$$\widehat{\gamma}^2_{k,u}(\widehat{h}) \;-\; C_1\, G(\mathcal{H}) \;-\; C_2\sqrt{\tfrac{\ln(2/\delta')}{N_{\mathrm{tr}}}} \;\ge\; C\,\nu\,\sqrt{\tfrac{\ln(2/\alpha)+\ln(2/\delta)}{N_{\mathrm{ho}}}}, \tag{13}$$

with $\widehat{\gamma}^2_{k,u}$ computed on the training half and $C = 16$, the level-α permutation test with $k_{\widehat{h}}$ on the held-out half rejects with probability at least $1 - \delta - \delta'$ over the joint randomness of the two halves.

The proof is in App. G. The corollary holds for any selector because the UCI is uniform over $\mathcal{H}$; with the worst-case $C_1 \approx 30$ its threshold is not reached at our sample sizes, so it motivates the deployed criterion rather than certifying the deployed test.

4 Experiments

Common setup. Unless otherwise stated, all tests use significance level $\alpha = 0.05$ with a permutation null based on $N_{\mathrm{perm}} = 200$ permutations. Deep-kernel experiments use the same MLP architecture across methods: $d \to 200 \to 200 \to 10$ with LeakyReLU activations, 100 Adam steps (learning rate 0.005, gradient-clip norm 5.0). The penalty $\widehat{C}_1$ is calibrated with $n_{\mathrm{cal}} = 10$ null permutations, 30 for the Higgs linear, polynomial, and ARD ansätze (App. F.2). In the linear and polynomial experiments, calibration runs on the training half of each repetition and the calibration ratio is the supremum over the bandwidth grid, which is the target (15) itself. In the deep-kernel experiments, calibration runs once per setting on an independent sample drawn from the null with the full per-class size n, either synthetic or two disjoint subsamples of the reference population for the real-data benchmarks, and distributed as the pooled data under H_0.

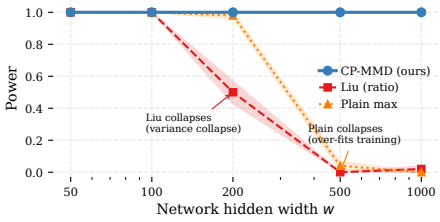

Figure 1: Class-richness sweep (± 1 SE): as w grows, Liu collapses, Plain overfits, and CP-MMD holds power 1.00.

4.1 Class-richness sweep: CP-MMD vs. Liu vs. plain maximization across MLP widths

On $P = \mathcal{N}(0, I_{50})$ vs. $Q = \mathcal{N}(\Delta\mathbf{1}, I_{50})$ at $\Delta = 0.5$ and $n = 100$ samples per class, we sweep MLP width $w \in \{50, 100, 200, 500, 1000\}$ and compare CP-MMD, which uses $\widehat{C}_1 = 0.01$ at every width, close to the value 0.0087 that the calibration of §3.3 gives for this (d, n) at $w = 200$, with Liu et al. [2020]'s deep-feature kernel $\kappa(\phi_\theta(x), \phi_\theta(y))$ (the published ϵ-mixture without the raw-input envelope q, denoted "Liu" below), and plain maximization ($\max_{h \in \mathcal{H}} \widehat{\gamma}^2_k(h)$) on the same architecture, 50 reps per width. As w grows, Liu collapses through the variance-collapse phenomenon documented in App. C and plain max overfits, while CP-MMD holds power 1.00 throughout (Fig. 1).

4.2 Three-regime head-to-head: linear, polynomial, deep

Each composite kernel family in (10) is paired with its most natural baseline: $\mathcal{H}_{\mathrm{linear}}$ with the median heuristic of Gretton et al. [2012a], $\mathcal{H}_{\mathrm{polynomial}}$ with MMDAgg in its standard Laplacian-plus-Gaussian configuration, and $\mathcal{H}_{\mathrm{deep}}$ with Liu, defined in §4.1. We also include plain maximization as a no-penalty baseline. Except for the polynomial pillar, which pits polynomial features against MMDAgg's bandwidth aggregation, each comparison holds the kernel family fixed and so isolates the selection criterion; the choice of family is a modeling decision outside its scope.

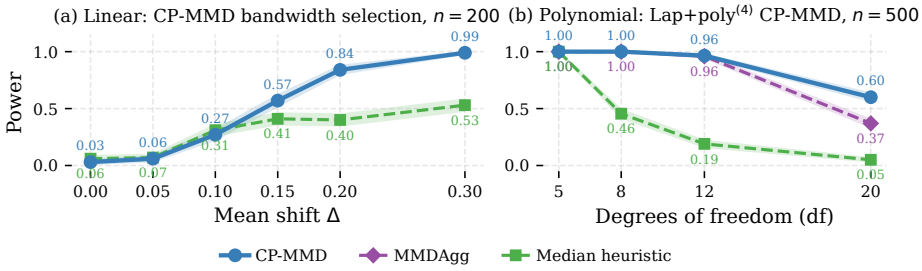

Figure 2: Three-regime head-to-head (dots: exact power; bands: ± 1 binomial SE). **(a)** Linear, multi-scale 2D mixture (axis Δ = mean shift between modes): CP-MMD reaches 0.99 at $\Delta = 0.30$, where Median is at 0.53. **(b)** Polynomial, kurtosis shift: the polynomial composite kernel $k_1^{\mathrm{L}}(\Psi_4(\cdot)/\sigma, \Psi_4(\cdot)/\sigma)$ matches MMDAgg at df$\leq$12 and beats it by 23 percentage points at df=20. All methods receive the same n samples per class; CP-MMD calibrates and selects on one half and tests on the other.

$\mathcal{H}_{\mathrm{linear}}$: CP-MMD versus the median heuristic. On a multi-scale 2D Gaussian mixture, $P = \frac{1}{2}\mathcal{N}(0, 0.01\, I_2) + \frac{1}{2}\mathcal{N}((3,0)^\top, 0.01\, I_2)$ and $Q = \frac{1}{2}\mathcal{N}((\Delta,0)^\top, 0.01\, I_2) + \frac{1}{2}\mathcal{N}((3{+}\Delta,0)^\top, 0.01\, I_2)$. We sweep $\Delta \in \{0, 0.05, 0.10, 0.15, 0.20, 0.30\}$ at $n = 200$ samples per class, 100 reps. The median heuristic fixes the bandwidth at the median pairwise distance of the pooled sample and tests on all n samples per class. CP-MMD selects σ on a 24-point geometric grid over $[0.05, 5]$ using half of the samples, on which $\widehat{C}_1$ is also calibrated, and runs the permutation test on the other half. For $\Delta \geq 0.15$, CP-MMD outperforms the median heuristic and reaches power 0.99 at $\Delta = 0.30$, where the median heuristic remains at 0.53 (Fig. 2(a)). For $\Delta \leq 0.10$ the two methods are within one binomial standard error; at these weak shifts the penalty selects the largest grid bandwidth.

$\mathcal{H}_{\mathrm{polynomial}}$: CP-MMD versus MMDAgg. For a kurtosis-shift task in d=10 dimensions, $P = \mathcal{N}(0, I_{10})$ and $Q = \sqrt{(\mathrm{df}-2)/\mathrm{df}}\, Z$ where $Z \sim t_{\mathrm{df}}(0, I_{10})$ is multivariate Student-t, scaled so $\mathrm{Cov}(Q) = I_{10}$ matches P. We sweep $\mathrm{df} \in \{5, 8, 12, 20\}$ at $n = 500$ samples per class, 200 reps. CP-MMD uses the polynomial composite kernel $k_1^{\mathrm{L}}(\Psi_4(\cdot)/\sigma, \Psi_4(\cdot)/\sigma)$ and selects σ on a 48-point geometric grid over $[0.01, 100]$ using half of the samples, on which $\widehat{C}_1$ is also calibrated, and tests on the other half. MMDAgg uses its standard configuration with Laplace and Gaussian kernels at 10 bandwidths each; MMDAgg and the median heuristic test on all n samples per class. CP-MMD matches MMDAgg on stronger kurtosis shifts and improves on the mildest case, reaching 0.60 at $\mathrm{df} = 20$ versus 0.37 (Fig. 2(b)). The selected bandwidth lies at the upper end $\sigma = 100$ of the search interval; App. L discusses the sensitivity of this result to the interval.

$\mathcal{H}_{\mathrm{deep}}$: CP-MMD versus Liu and plain maximization. On the high-dimensional Gaussian mean-shift family $P = \mathcal{N}(0, I_d)$ vs. $Q = \mathcal{N}(\Delta\mathbf{1}, I_d)$ at $\Delta = 0.5$, we evaluate five (d, n) cells with 200 reps per cell, comparing CP-MMD, Liu, and plain max on the same MLP under sample-split (App. G). MMDAgg and MMD-FUSE require a finite kernel grid and do not apply to $\mathcal{H}_{\mathrm{deep}}$. CP-MMD attains power 1.00 in four of the five cells, dropping only to 0.83 at the mildest alternative; Plain approximately matches CP-MMD at the two n=200 cells (reaching 0.81 and 1.00) but degrades to 0.14–0.54 at the three smaller-n cells, while Liu stays strictly below CP-MMD in every cell (Table 2). Liu's published kernel adds a raw-input envelope that lies outside Definition 3.3; a criterion-by-kernel factorial that includes it (App. J, Table 7) finds that CP-MMD matches or exceeds Liu's ratio criterion in 9 of 10 cells, and that the mixture kernel itself loses power as d grows, from 0.98 at $d = 2$ to 0.01 at $d = 100$ under CP-MMD, so the pure-feature kernel is the stronger one here.

Criterion	d=2 n=200	d=20 n=200	d=50 n=100	d=100 n=100	d=20 n=50
CP-MMD	**.83**$\pm$.03	**1.00**	**1.00**	**1.00**	**1.00**
Liu	.37$\pm$.03	.51$\pm$.04	.27$\pm$.03	.34$\pm$.03	.08$\pm$.02
Plain	.81$\pm$.03	1.00	.54$\pm$.04	.35$\pm$.03	.14$\pm$.02

Table 2: Deep-kernel power at $\Delta = 0.5$, 200 reps; Type-I is within ± 2 SE of 0.05 in all cells.

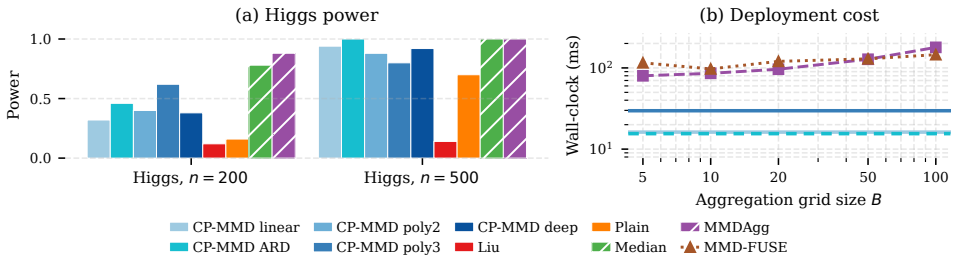

Figure 3: **(a) Higgs power** (d=28, $n \in \{200, 500\}$ per class): CP-MMD, Liu, and Plain select on one half and test on the other; Median and MMDAgg test on all n. **(b) Per-test cost** (d=28, n=200, 200 permutations): CP-MMD-linear and CP-MMD-ARD (16 ms) and CP-MMD-polynomial$_3$ (30 ms) do not depend on B and are 2.7–12$\times$ faster than MMDAgg and MMD-FUSE.

4.3 Real data: CP-MMD across kernel parameterizations on Higgs

On the UCI Higgs boson dataset [Baldi et al., 2014], we take P as background events and Q as signal events with $d = 28$ kinematic features. At sample sizes $n \in \{200, 500\}$ per class and 50 reps per cell, we evaluate five CP-MMD ansätze against the median heuristic and MMDAgg: linear, polynomial of degree 2 and 3, a deep MLP, which we also compare with Liu and Plain on the same MLP, and automatic relevance determination (ARD), the per-feature bandwidths $h(x) = (x_1/\sigma_1, \ldots, x_d/\sigma_d)$ that Sutherland et al. [2017] optimized for MMD tests. ARD lets the kernel down-weight uninformative features; as a one-layer network it uses the deep proxy of (10) with $L(h) = \max_j \sigma_j^{-1}$. Although CP-MMD tests on half of the samples while the median heuristic and MMDAgg test on all n, ARD matches both at $n = 500$ with power 1.00; at $n = 200$ the best ansatz, polynomial$_3$ (0.62), trails the median heuristic (0.78) and MMDAgg (0.88). The deep version outperforms Liu and Plain (Fig. 3(a)). In App. K, CIFAR-10/10.1 saturates at n=200 for all methods, and on the Tennessee Eastman Process MMDAgg beats the learned kernels.

4.4 Deployment per-test cost: CP-MMD is B-free

Matching the Higgs panel at $d = 28$, $n = 200$, we measure per-test wall-clock time on CPU over 30 independent test pairs at each grid size $B \in \{5, 10, 20, 50, 100\}$, using JAX JIT-compiled permutation nulls with $N_{\text{perm}} = 200$. MMDAgg uses 200 permutations for its quantiles and another 200 for its level correction. CP-MMD evaluates one learned kernel and obtains all permutation statistics from one kernel matrix; MMDAgg and MMD-FUSE evaluate B candidate kernels. CP-MMD's per-test cost does not depend on B: 16 ms for CP-MMD-linear and 30 ms for CP-MMD-polynomial$_3$, against 80–179 ms for MMDAgg and 97–146 ms for MMD-FUSE (Fig. 3(b)). With its default 2000 permutations per set, MMDAgg costs 0.57–1.34 s per test. Calibration and selection, paid once per problem, take 22 and 37 ms on one CPU thread with $n_{\text{cal}} = 30$ and 24 bandwidths on $n = 200$ per class. One test including them thus costs 38 and 67 ms, below MMDAgg and MMD-FUSE at every B, and the break-even number of tests against MMDAgg, $K^* := T_{\text{sel}}/(T_{\text{MMDAgg}} - T_{\text{test}})$, is at most 0.73. ARD needs 21.5 s to fit its 28 bandwidths and tests at the linear cost, so K^* is 16–39 against MMDAgg with 2000 permutations and 132–337 with 200. Cor. I.10 gives the asymptotic per-test costs.

5 Discussion and conclusion

We introduced CP-MMD, which treats kernel selection as model selection across linear, polynomial-feature, and deep classes, with a B-free deployment cost. The UCI gives a finite-sample lower bound on post-selection power with worst-case constants (Cor. 3.6); the deployed criterion replaces them with a calibrated constant and a per-iterate proxy, and its effectiveness is empirical. With equal sample sizes for all methods, CP-MMD outperforms the median heuristic on multi-scale data, improves on MMDAgg for the mildest kurtosis shift, and holds its power across MLP widths where Liu's ratio criterion collapses and plain maximization overfits. Limitations, including the sensitivity to the bandwidth search interval, and extensions are discussed in App. L.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers and the area chair for comments that improved the paper. Yijin Ni is supported by the Robert Goodell Ph.D. Student Fellowship for Research Excellence at Georgia Institute of Technology. Xiaoming Huo is partially supported by a subcontract of NSF grant 2229876, the A. Russell Chandler III Professorship at Georgia Institute of Technology, an NIH-sponsored Georgia Clinical & Translational Science Alliance, and the Georgia Department of Transportation. The authors declare no competing interests.

References

Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5:4308, 2014.

Peter L Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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*, volume 36, 2023.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

G. Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series. Duxbury Press, 2nd edition, 2002. ISBN 9780534243128.

Anirban Chatterjee and Aaditya Ramdas. A martingale kernel two-sample test. In *International Conference on Algorithmic Learning Theory*, pages 1–44. PMLR, 2026.

Ikjun Choi and Ilmun Kim. Computational-statistical trade-off in kernel two-sample testing with random Fourier features, 2024. URL `https://arxiv.org/abs/2407.08976`.

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

Rui Cui, Yuhao Li, and Xiaojun Song. Kernel two-sample testing via directional components analysis, 2025. URL `https://arxiv.org/abs/2508.08564`.

James J Downs and Ernest F Vogel. A plant-wide industrial process control problem. *Computers & Chemical Engineering*, 17(3):245–255, 1993.

Arthur Gretton, Kenji Fukumizu, Choon H Teo, Le Song, Bernhard Schölkopf, and Alex Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

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

Arthur Gretton, Bharath K Sriperumbudur, Dino Sejdinovic, Heiko Strathmann, Sivaraman Balakrishnan, Massimiliano Pontil, and Kenji Fukumizu. Optimal kernel choice for large-scale two-sample tests. *Advances in Neural Information Processing Systems*, 25, 2012b.

Perrine Lacroix, Bertrand Michel, Franck Picard, and Vincent Rivoirard. Non-asymptotic two-sample kernel testing with the spectrally truncated normalized MMD, 2026. URL `https://arxiv.org/abs/2604.07153`.

Erich L Lehmann and Joseph P Romano. *Testing Statistical Hypotheses*. Springer, 3rd edition, 2005.

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*, pages 6316–6326. PMLR, 2020.

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

Andreas Maurer and Massimiliano Pontil. Uniform concentration and symmetrization for weak interactions. In *Conference on Learning Theory*, pages 2372–2387. PMLR, 2019.

Jerzy Neyman and Egon S Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933.

Yijin Ni and Xiaoming Huo. A uniform concentration inequality for kernel-based two-sample statistics, 2024. URL `https://arxiv.org/abs/2405.14051`.

Aaditya Ramdas, Sashank J Reddi, Barnabás Póczos, Aarti Singh, and Larry Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *AAAI Conference on Artificial Intelligence*, volume 29, pages 3571–3577, 2015.

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

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

Dougal J Sutherland, Hsiao-Yu Tung, Heiko Strathmann, Soumyajit De, Aaditya Ramdas, Alexander Smola, and Arthur Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*, 2017.

Martin J Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Supplementary Material

Complexity-Penalized MMD: Appendices

Code and results: `https://github.com/Yijin911/cpmmd-code/tree/main/neurips2026`.

Roadmap. The appendices are grouped by function: proofs of the formal results, the calibration procedure and its validation, empirical documentation of the two mechanisms discussed in the body, and additional experiments.

Group	Appendix	Supports
Notation	App. A	all sections
Proofs	App. B: Lemma 2.1	§2
	App. D: Theorem 3.4	§3.2
	App. G: Corollary 3.6	§3.3
	App. H: Proposition 3.5	§3.3
	App. I: rate on finite kernel grids	§3.2
Calibration	App. F: procedure for $\widehat{C}_1$, robustness and stability; unequal sample sizes in App. F.9	§3.3
Mechanisms	App. C: variance collapse of the ratio criterion; App. E: spectral stabilization of J_{CP}	§1.1, §4.1
Experiments	App. J: Liu's ϵ-mixture factorial; App. K: real data	§4
Limitations	App. L	§5

A Notation

Symbol	Meaning
$\mathbf{X}, \mathbf{Y}$	i.i.d. samples $(X_1, \dots, X_m) \overset{\text{iid}}{\sim} P$ and $(Y_1, \dots, Y_n) \overset{\text{iid}}{\sim} Q$ in $\mathbb{R}^d$ (equivalently $\mathbf{X} \sim P^m$, $\mathbf{Y} \sim Q^n$)
$N = m + n$	total sample size; $\rho_* = \max\{N/m, N/n\}$
$\mathbf{D} = \mathbf{X} \sqcup \mathbf{Y}$	pooled data ($\mathbf{D}_i$ denotes the i-th element, see §2); $\mathcal{D}_F = \|\mathbf{D}\|_F$
$\mathcal{D}_{\mathrm{tr}}, \mathcal{D}_{\mathrm{te}}$	training / test halves of $\mathbf{D}$ under the sample-split protocol (§3.3), of sizes N_{tr} and N_{ho}
k, ν, l	base kernel; boundedness ν; Lipschitz l (Assumption 3.2)
$h \in \mathcal{H}$	kernel-selection map $h : \mathbb{R}^d \to \mathcal{U}$; composite kernel $k_h(x, x') = k(h(x), h(x'))$ (Definition 3.3)
$\gamma_k^2(h)$	population MMD at h; $\widehat{\gamma}_{k,u}^2(h)$ unbiased estimator
$\mathcal{H}_T$	trajectory class $\{h^{(0)}, \dots, h^{(T)}\}$ visited by the optimizer over T ascent steps
$L(h), L^*$	Lipschitz constant of h; finite envelope $L^* \geq \max_{t \leq T} L(h^{(t)})$ of the trajectory (Proposition 3.5)
$G(\mathcal{H}), \widehat{G}(\mathcal{H}_T), \widetilde{G}(h)$	Gaussian complexity of the fixed class $\mathcal{H}$ (Theorem 3.4); empirical Gaussian complexity of the trajectory given $\mathbf{D}$ (Proposition 3.5); per-iterate complexity proxy (10)
$C_1, \widehat{C}_1^*, \widehat{C}_1$	worst-case UCI constant (Theorem 3.4); population calibration target, the H_0 $(1-\alpha)$-quantile of $\sup_h \widehat{\gamma}_{k,u}^2(h)/\widetilde{G}(h)$ (15); its estimate from n_{cal} null permutations (11)
$J_{\mathrm{CP}}(h)$	CP-MMD criterion $\widehat{\gamma}_{k,u}^2(h) - \widehat{C}_1 \widetilde{G}(h)$, (12)
$J_{\mathrm{Liu}}(h)$	Liu's ratio criterion $\sqrt{n}\widehat{\gamma}_{k,u}^2(h)/\hat{\tau}_h$ (Appendix C)
$\widehat{h}, \widehat{h}_\pi$	CP-MMD selector (§3.3); plain-MMD maximizer on π-permuted training half ((11))
$\Pi^{(t)}, \Pi^*$	product of layer spectral norms at step t; empirical stable equilibrium under J_{CP} ascent (App. E)
$B_N(\delta)$	UCI bound $C_1 G(\mathcal{H}) + C_2\sqrt{\ln(2/\delta)/N}$ (Theorem 3.4)
$n_{\mathrm{cal}}, N_{\mathrm{perm}}$	number of null permutations for $\widehat{C}_1$ calibration; number of permutations in the deployed test
Δ (capital)	mean-shift parameter in §4 ($Q = \mathcal{N}(\Delta \cdot \mathbf{1}, I_d)$); distinct from confidence δ in Theorem 3.4
HDGM	high-dimensional Gaussian mean-shift family (§4.2, Table 2)

B Proof of Lemma 2.1

Proof. Take $P^{(i)} = \mathcal{N}(0, i^2)$ and $Q^{(i)} = \mathcal{N}(0, (2i)^2)$ on $\mathbb{R}$.

Total variation is constant in i. Both densities are obtained from $\mathcal{N}(0,1)$ and $\mathcal{N}(0,4)$ by the dilation $z \mapsto iz$. Total variation is invariant under joint rescaling of the two distributions, so

$$\mathrm{TV}(P^{(i)}, Q^{(i)}) = \mathrm{TV}(\mathcal{N}(0,1), \mathcal{N}(0,4)) =: c_0,$$

which is a strictly positive constant independent of i.

Population MMD vanishes. The Gaussian-kernel inner products under Gaussian distributions admit closed forms via the identity $\mathbb{E}_{Z\sim\mathcal{N}(\mu,\Sigma)}[\exp(-\alpha Z^2)] = (1+2\alpha\Sigma)^{-1/2}\exp\big(-\alpha\mu^2/(1+2\alpha\Sigma)\big)$. With $\alpha = 1/(2\sigma^2)$ to match the kernel definition $k_\sigma(x,x') = \exp(-\|x-x'\|^2/(2\sigma^2))$:

$$\begin{aligned}
\mathbb{E}\,k_\sigma(X,X') &= (1+2i^2/\sigma^2)^{-1/2}, \quad X, X' \overset{\text{iid}}{\sim} P^{(i)} \text{ so } X - X' \sim \mathcal{N}(0, 2i^2),\\
\mathbb{E}\,k_\sigma(Y,Y') &= (1+8i^2/\sigma^2)^{-1/2}, \quad Y, Y' \overset{\text{iid}}{\sim} Q^{(i)} \text{ so } Y - Y' \sim \mathcal{N}(0, 8i^2),\\
\mathbb{E}\,k_\sigma(X,Y) &= (1+5i^2/\sigma^2)^{-1/2}, \quad X \sim P^{(i)}, Y \sim Q^{(i)} \text{ so } X - Y \sim \mathcal{N}(0, 5i^2).
\end{aligned}$$

All three quantities are $O(1/i)$ as $i \to \infty$, so $\gamma^2_{k_\sigma}(P^{(i)}, Q^{(i)}) \to 0$ by (2).

Test power collapses. Fix any $\sigma > 0$ and the level-α permutation test ϕ_{k_σ} on $m+n$ samples with N_{perm} permutations. Under the alternative $(P^{(i)}, Q^{(i)})$ the unbiased MMD estimator $\widehat{\gamma}^2_{k_\sigma,u}$ has mean $\gamma^2_{k_\sigma}(P^{(i)}, Q^{(i)}) = O(1/i) \to 0$ (just shown) and variance $\le \nu^2/\min(m,n)$ uniformly in i via boundedness $|k_\sigma| \le \nu$ and the MMD seminorm calculation of Theorem 3.4. By Chebyshev's inequality, the alternative-sample MMD concentrates within $O(\nu/\sqrt{\min(m,n)})$ of zero with high probability for i large enough that $\gamma^2_{k_\sigma}(P^{(i)}, Q^{(i)}) \ll \nu/\sqrt{\min(m,n)}$. By the same boundedness argument applied to the permutation distribution of $\widehat{\gamma}^2_{k_\sigma,u}$ under random labeling of the pooled sample (variance $\le \nu^2/\min(m,n)$), Chebyshev's upper-tail bound gives $c_\alpha = O(\nu/\sqrt{\min(m,n)})$. As $i \to \infty$, the alternative-sample mean $\gamma^2_{k_\sigma}(P^{(i)}, Q^{(i)}) = O(1/i)$ shrinks to 0 while c_α remains at scale $O(\nu/\sqrt{\min(m,n)})$ with matching variance bound, so the alternative-sample distribution of $\widehat{\gamma}^2_{k_\sigma,u}$ converges to the permutation-null distribution; the level-α test's rejection probability under $(P^{(i)}, Q^{(i)})$ therefore converges to its size α under H_0:

$$\liminf_{i\to\infty} \Pr_{(P^{(i)},Q^{(i)})}(\phi_{k_\sigma} = 1) \;=\; \alpha,$$

i.e., the test's rejection probability collapses to its size, regardless of $\sigma > 0$. The collapse is a consequence of the dilation: at any fixed bandwidth σ, the dilated samples are spread over a region of scale i and the Gaussian kernel matrix becomes near-zero off the diagonal, so the resulting MMD estimator carries no signal. A single fixed σ therefore cannot sustain power along the dilation sequence; only a kernel whose bandwidth scales with i (i.e., a data-driven choice) can. □

C Variance collapse of the ratio criterion: full statement and verification

This appendix makes precise the failure mode of the ratio criterion and gives empirical evidence; it motivates the subtractive form of J_{CP} in §3.

The ratio criterion's plug-in z-statistic $J_{\text{Liu}}(h) := \sqrt{n}\,\widehat{\gamma}^2_{k,u}(h)/\hat{\tau}_h$ uses the H_1 standard-deviation estimate $\hat{\tau}_h$ of Liu et al. [2020]. With $H_{ij} := k_h(x_i,x_j) + k_h(y_i,y_j) - k_h(x_i,y_j) - k_h(x_j,y_i)$ and row means $r_i := n^{-1}\sum_j H_{ij}$, it is $\hat{\tau}_h^2 := 4\big(n^{-1}\sum_i r_i^2 - (n^{-1}\sum_i r_i)^2\big) + \lambda$ with $\lambda = 10^{-8}$, as in the authors' code, and every Liu result in this paper uses it. The ratio is derived from asymptotic theory that assumes h is fixed independently of the data [Liu et al., 2020, Sutherland et al., 2017]. When h is instead optimized on the same data used to compute the statistic, this independence assumption breaks. Consequently, the optimizer can drive both the empirical MMD $\widehat{\gamma}^2_{k,u}(h)$ and the variance estimate $\hat{\tau}_h$ toward zero simultaneously when the involved kernel class is rich, producing a spuriously large ratio at a degenerate kernel that carries no actual signal. Training discriminators to maximize J_{Liu} under H_0 ($P = Q$) routinely produces $\widehat{\gamma}^2_{k,u}(h) < 0.01$ alongside $J_{\text{Liu}}(h)$ values between 10 and 300.

Empirical observation: variance collapse under H_0. We document the variance-collapse phenomenon directly. Setup: $P = Q = \mathcal{N}(0, I_{10})$ (so H_0 holds), $m = n = 200$, 200 Adam steps to maximize J_{Liu} on a deep-MLP discriminator. A run is flagged as *collapsed* if all three of $\widehat{\gamma}^2_{k,u} < 0.01$, $\hat{\tau}_h/\sqrt{n} < 0.001$, and $J_{\text{Liu}} > 10$ are observed at termination. At MLP widths $w \in \{10, 50, 200\}$, with 10 seeds each, all 30 runs collapse, and J_{Liu} reaches 262 while $\widehat{\gamma}^2_{k,u}$ stays below 0.01 (Table 3, Figure 4). The mechanism is that $\hat{\tau}_h$ shrinks together with $\widehat{\gamma}^2_{k,u}$, so the ratio $\widehat{\gamma}^2_{k,u}/\hat{\tau}_h$ stays large

although the kernel carries no signal. At $w = 400$ the optimizer instead reaches a degenerate kernel with $\widehat{\gamma}^2_{k,u} \leq 5 \times 10^{-5}$, where J_{Liu} stays below 1.

By contrast, CP-MMD's J_{CP} stays strictly negative in every collapsed run: whenever $\widehat{\gamma}^2_{k,u} \leq 0$, the subtractive form forces $J_{\mathrm{CP}} \leq -\widehat{C}_1 \widetilde{G} < 0$, so no degenerate kernel can be selected.

A clean asymptotic-rate statement of variance collapse for the unbiased MMD $\widehat{\gamma}^2_{k,u}$ paired with the plug-in $\hat{\tau}_h$ is delicate, because under H_0 the unbiased estimator $\widehat{\gamma}^2_{k,u}$ has the chi-squared-like limiting distribution of a degenerate U-statistic with rate $\Theta_p(1/n)$ [Gretton et al., 2012a, Theorem 12 and Lemma 6]: under balanced sampling $m = n$, $\widehat{\gamma}^2_{k,u}$ admits a one-sample U-statistic representation in the paired data $z_i := (x_i, y_i)$ with kernel $h(z_i, z_j) = k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_j) - k(x_j, y_i)$; this kernel is degenerate at H_0, yielding the $t \cdot \widehat{\gamma}^2_{k,u} \to \sum_l \lambda_l [(\rho_x^{-1/2} a_l - \rho_y^{-1/2} b_l)^2 - (\rho_x \rho_y)^{-1}]$ limit of Gretton et al. [2012a, Theorem 12]. The $\Theta_p(1/n)$ rate cited above is a direct consequence of this distributional result. The variance estimator $\hat{\tau}_h$, an empirical proxy for the H_1-variance, also tends to zero under H_0, so a clean ratio-rate analysis is delicate. We therefore document the phenomenon empirically rather than as a formal proposition; the empirical evidence below is the basis for adopting the subtractive form of J_{CP} over a ratio.

Width	Collapse rate	Mean J_{Liu}	Mean $\widehat{\gamma}^2_{k,u}$	Mean $\hat{\tau}_h/\sqrt{n}$	Mean $\widetilde{G}$	Mean J_{CP}
10	100%	37	0.002	5×10^{-5}	126	$\mathbf{-0.124}$
50	100%	145	0.003	3×10^{-5}	68	$\mathbf{-0.065}$
200	100%	101	0.005	5×10^{-5}	81	$\mathbf{-0.076}$
400	0%	0.5	0.000	5×10^{-5}	253	$\mathbf{-0.253}$

Table 3: Variance collapse under H_0 for J_{Liu} across network widths (mean of 10 seeds). The "Collapse rate" column counts runs satisfying all three flags ($\widehat{\gamma}^2_{k,u} < 0.01$, $\hat{\tau}_h/\sqrt{n} < 0.001$, $J_{\mathrm{Liu}} > 10$); CP-MMD's J_{CP} stays strictly negative on every such run. No run collapses at $w = 400$ because the optimizer instead reaches a degenerate kernel with $\widehat{\gamma}^2_{k,u} \leq 5 \times 10^{-5}$, which keeps J_{Liu} below 1; J_{CP} remains negative regardless. Reproduced by `simulations/run_variance_collapse.py`.

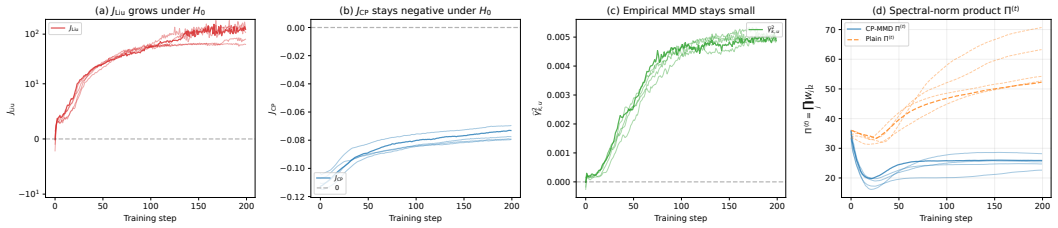

Figure 4: Training dynamics under H_0, five independent runs. **(a)** J_{Liu} grows to about 10^2 despite $P = Q$. **(b)** J_{CP} stays negative. **(c)** The empirical MMD stays below 0.01, so the growth in (a) comes from the shrinking variance estimate. **(d)** The product of spectral norms $\Pi^{(t)}$ plateaus under CP-MMD and keeps growing under plain maximization.

D Proof of Theorem 3.4 (UCI)

This appendix restates the two-sample UCI of Ni and Huo [2024] (§D.1, Lemma D.7) and specializes it to the unbiased MMD estimator $\widehat{\gamma}^2_{k,u}$ (§D.2). The Maurer-style symmetrization-and-induction argument and the MMD-specific seminorm derivation are reproduced here for self-containment; see Ni and Huo [2024] for the general theory and its applications to other kernel-based statistics (Energy Distance, distance Covariance, and HSIC).

D.1 Two-sample concentration framework

Setup. Let $\mathbf{X} = (X_1, \ldots, X_m) \in \mathcal{U}^m$ and $\mathbf{Y} = (Y_1, \ldots, Y_n) \in \mathcal{U}^n$ be independent samples, $\mathbf{X}', \mathbf{Y}'$ independent copies, $\mathbf{D} = \mathbf{X} \sqcup \mathbf{Y}$ the pooled data, and $N = m + n$. Let $f : \mathcal{U}^{m+n} \to \mathbb{R}$ be a measurable two-sample functional that takes the concatenation $(\mathbf{X}, \mathbf{Y})$ as its input. Let $\mathcal{H}$ be a class of measurable maps $\mathbb{R}^d \to \mathcal{U}$ with finite covering number under $\|\cdot\|_\infty$.

Lipschitz seminorms of f. Following Maurer and Pontil [2019, Assumption 1], define the first-order Lipschitz, second-order (cross-coordinate) Lipschitz, and McDiarmid bounded-difference seminorms

$$M_{\mathrm{Lip}}(f) := N \cdot \max_k \sup_{\mathbf{u} \in \mathcal{U}^N,\, y \neq y' \in \mathcal{U}} \frac{|D^k_{yy'} f(\mathbf{u})|}{\|y - y'\|},$$
$$J_{\mathrm{Lip}}(f) := N^2 \cdot \max_{k \neq l} \sup_{\mathbf{u},\, y \neq y',\, z, z'} \frac{|D^l_{zz'} D^k_{yy'} f(\mathbf{u})|}{\|y - y'\|},$$
$$M(f) := N \cdot \max_k \sup_{\mathbf{u},\, y, y' \in \mathcal{U}} |D^k_{yy'} f(\mathbf{u})|,$$

where $D^k_{yy'} f(\mathbf{u}) := f(\ldots, u_{k-1}, y, u_{k+1}, \ldots) - f(\ldots, u_{k-1}, y', u_{k+1}, \ldots)$ is the k-th coordinate-swap operator.

Lemma D.7 (Two-sample UCI for nonlinear functionals; Ni and Huo 2024)**.** *Suppose $M_{\mathrm{Lip}}(f), J_{\mathrm{Lip}}(f), M(f) < \infty$. Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over $(\mathbf{X}, \mathbf{Y})$,*

$$\sup_{h \in \mathcal{H}} \big| f(h(\mathbf{X}), h(\mathbf{Y})) - \mathbb{E}\, f(h(\mathbf{X}'), h(\mathbf{Y}')) \big| \leq \tfrac{\sqrt{2\pi}}{N} \big(2M_{\mathrm{Lip}}(f) + J_{\mathrm{Lip}}(f)\big)\, \mathbb{E}[G(\mathcal{H}(\mathbf{D}))] + M(f) \sqrt{\tfrac{\ln(2/\delta)}{N}},$$

where $G(\mathcal{H}(\mathbf{D}))$ is the empirical Gaussian complexity of the transformed pooled data

$$\mathcal{H}(\mathbf{D}) := \big\{ \big(h(X_1), \ldots, h(X_m),\, h(Y_1), \ldots, h(Y_n)\big) \mid h \in \mathcal{H} \big\}.$$

Proof. The argument extends Maurer and Pontil [2019, Theorem 2 + Corollary 3] from the one-sample to the two-sample setting; the only change is to symmetrize over the $\mathbf{X}$- and $\mathbf{Y}$-halves separately rather than over a single homogeneous tuple.

Step 1 (Coordinate decomposition). Maurer and Pontil [2019, Lemma 9] gives, for any $\mathbf{u}, \mathbf{u}' \in \mathcal{U}^N$, the telescoping decomposition $f(\mathbf{u}) - f(\mathbf{u}') = \sum_{k=1}^N F_k(\mathbf{u}, \mathbf{u}')$ where each F_k averages partial differences over subsets of the first $k-1$ coordinates.

Step 2 (Gaussian comparison). Maurer and Pontil [2019, Lemma 10] dominates each F_k by an inner product with an independent Gaussian vector $\boldsymbol{\xi}_k$, scaled by $M_{\mathrm{Lip}}(f)$ and $J_{\mathrm{Lip}}(f)$.

Step 3 (Two-sample symmetrization and induction). This is the load-bearing step where Ni and Huo [2024]'s argument extends the one-sample concentration framework of Maurer and Pontil [2019] to the two-sample setting; we reproduce it here because the MMD-specific seminorm calculation in §D.2 inherits constants from this step. Maurer and Pontil [2019]'s one-sample argument symmetrizes a single i.i.d. tuple $(Z_1, \ldots, Z_N)$ via Rademacher signs ϵ_i that act uniformly across all N coordinates. In the two-sample setting, the pooled tuple $(\mathbf{X}, \mathbf{Y}) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ is i.i.d. within each half but the two halves come from different distributions P, Q, so a single global symmetrization is not directly applicable. The correct argument, due to Ni and Huo [2024], symmetrizes the two halves separately.

Concretely: let $\mathbf{X}' = (X'_1, \ldots, X'_m)$ and $\mathbf{Y}' = (Y'_1, \ldots, Y'_n)$ be independent copies of $\mathbf{X}, \mathbf{Y}$. By the Step-1 telescoping decomposition of Maurer and Pontil [2019, Lemma 9],

$$f(h(\mathbf{X}), h(\mathbf{Y})) - f(h(\mathbf{X}'), h(\mathbf{Y}')) = \sum_{k=1}^m F^X_k(\mathbf{X}, \mathbf{Y}, \mathbf{X}') + \sum_{k=1}^n F^Y_k(\mathbf{X}', \mathbf{Y}, \mathbf{Y}'),$$

where F_k^X swaps the k-th X-coordinate (with the Y-tuple held fixed) and F_k^Y swaps the k-th Y-coordinate (with the X-tuple held fixed). Each F_k^X, F_k^Y is a coordinate-swap term whose magnitude is controlled by the seminorms $M_{\mathrm{Lip}}(f), J_{\mathrm{Lip}}(f)$ (Step 2 above). Apply Maurer–Pontil's Gaussian-comparison Lemma 10 separately to the X-block and the Y-block: each block contributes an inner product with independent standard Gaussian vectors $\boldsymbol{\xi}_1^X, \ldots, \boldsymbol{\xi}_m^X$ and $\boldsymbol{\xi}_1^Y, \ldots, \boldsymbol{\xi}_n^Y$ respectively, scaled by $M_{\mathrm{Lip}}/N + J_{\mathrm{Lip}}/N^2$ per coordinate.

Now the two-sample-specific induction: induct on the number of symmetrized coordinates separately within each half. The induction hypothesis at step (j, k) (with j symmetrized X-coordinates and k symmetrized Y-coordinates) is that the symmetrized supremum is bounded by $\frac{\sqrt{2\pi}}{N}\big(2M_{\mathrm{Lip}} + J_{\mathrm{Lip}}\big)\,\mathbb{E}[G(\mathcal{H}(\mathbf{D}))]$ scaled by the fraction $(j+k)/N$ of symmetrized coordinates. Each induction step adds one more symmetrized X- or Y-coordinate; the resulting Gaussian process on the augmented $\mathcal{H}(\mathbf{D})$-class inherits the same scaling because the Lipschitz seminorms $M_{\mathrm{Lip}}, J_{\mathrm{Lip}}$ are insensitive to which half a swapped coordinate belongs to (both halves contribute symmetrically to the bounded-differences calculation in Step 4 below). After $N = m + n$ induction steps, all coordinates are symmetrized and the bound becomes the full Gaussian complexity $\mathbb{E}[G(\mathcal{H}(\mathbf{D}))]$ on the entire pooled tuple. Combining the per-block contributions (which are additive across the two halves by independence of $\mathbf{X}$ and $\mathbf{Y}$):

$$\mathbb{E}\sup_{h\in\mathcal{H}}\big[f(h(\mathbf{X}),h(\mathbf{Y})) - \mathbb{E}\,f(h(\mathbf{X}'),h(\mathbf{Y}'))\big] \;\le\; \tfrac{\sqrt{2\pi}}{N}\big(2M_{\mathrm{Lip}}(f) + J_{\mathrm{Lip}}(f)\big)\,\mathbb{E}[G(\mathcal{H}(\mathbf{D}))].$$

The key observation enabling this two-sample extension is that the Maurer–Pontil seminorms $M_{\mathrm{Lip}}, J_{\mathrm{Lip}}, M$ are defined coordinate-by-coordinate in a way that is symmetric in the two halves: a coordinate-swap at position k contributes the same upper bound whether k indexes $\mathbf{X}$ or $\mathbf{Y}$. The inductive scaling $(j + k)/N$ thus passes through cleanly, and the final Gaussian complexity term $\mathbb{E}[G(\mathcal{H}(\mathbf{D}))]$ is the empirical Gaussian complexity of the transformed pooled tuple, exactly as in the one-sample case.

Step 4 (McDiarmid + symmetric tail). The functional $\Phi(\mathbf{X}, \mathbf{Y}) := \sup_h[f(h(\mathbf{X}), h(\mathbf{Y})) - \mathbb{E}\,f(h(\mathbf{X}'), h(\mathbf{Y}'))]$ has bounded differences with constant $M(f)/N$ in each coordinate (by definition of $M(f)$). McDiarmid's inequality [Boucheron et al., 2013, Theorem 6.2] gives the high-probability tail $M(f)\sqrt{\ln(2/\delta)/N}$, and applying the same argument to $-f$ yields the two-sided bound by a union argument.

Step 5 (Covering-number relaxation). Suppose $\mathcal{H}$ has finite covering number $N(\eta; \mathcal{H}, \|\cdot\|_\infty)$ for every $\eta > 0$, and fix $\eta > 0$. By taking an $(\eta/2)$-cover of $\mathcal{H}$ and selecting one element of $\mathcal{H}$ from each cover ball, we obtain an *internal* η-net $\mathcal{H}_\eta = \{h_1, \ldots, h_{M_\eta}\} \subseteq \mathcal{H}$ with $M_\eta \le N(\eta/2; \mathcal{H}, \|\cdot\|_\infty)$ such that for every $h \in \mathcal{H}$ there is some $h_i \in \mathcal{H}_\eta$ with $\sup_{u\in\mathcal{U}} \|h(u) - h_i(u)\| \le \eta$. The inclusion $\mathcal{H}_\eta \subseteq \mathcal{H}$ ensures the image classes satisfy $\mathcal{H}_\eta(\mathbf{D}) \subseteq \mathcal{H}(\mathbf{D})$, so class monotonicity of G applies in the Lipschitz-transfer step below. Steps 1–4 apply to the finite class $\mathcal{H}_\eta$, giving with probability $\ge 1 - \delta$:

$$\sup_{h_i\in\mathcal{H}_\eta}\big|f(h_i(\mathbf{X}),h_i(\mathbf{Y})) - \mathbb{E}f(h_i(\mathbf{X}'),h_i(\mathbf{Y}'))\big| \\ \le\; \tfrac{\sqrt{2\pi}}{N}\big(2M_{\mathrm{Lip}}(f) + J_{\mathrm{Lip}}(f)\big)\,\mathbb{E}[G(\mathcal{H}_\eta(\mathbf{D}))] + M(f)\sqrt{\tfrac{\ln(2/\delta)}{N}}.$$

For any $h \in \mathcal{H}$ with nearest net point h_i, the first-order Lipschitz seminorm $M_{\mathrm{Lip}}(f)$ controls the transfer: each of the N coordinate replacements $h(u_k) \to h_i(u_k)$ changes f by at most $M_{\mathrm{Lip}}(f)/N \cdot \eta$ (definition of M_{Lip} applied to the swap $h(u_k) \to h_i(u_k)$ with displacement $\le \eta$), so the cumulative shift across all N coordinates is bounded by $M_{\mathrm{Lip}}(f) \cdot \eta$. The same bound applies to $\mathbb{E}f(h(\mathbf{X}'), h(\mathbf{Y}'))$ by linearity of expectation. By the triangle inequality, for every $h \in \mathcal{H}$ with nearest $h_i \in \mathcal{H}_\eta$,

$$\big|f(h(\mathbf{X}),h(\mathbf{Y})) - \mathbb{E}f(h(\mathbf{X}'),h(\mathbf{Y}'))\big| \;\le\; \big|f(h_i(\mathbf{X}),h_i(\mathbf{Y})) - \mathbb{E}f(h_i(\mathbf{X}'),h_i(\mathbf{Y}'))\big| + 2M_{\mathrm{Lip}}(f)\,\eta.$$

Taking $\sup_h$ on both sides and using $\mathbb{E}[G(\mathcal{H}_\eta(\mathbf{D}))] \le \mathbb{E}[G(\mathcal{H}(\mathbf{D}))]$ (class monotonicity), the displayed bound holds with $\mathcal{H}_\eta$ replaced by $\mathcal{H}$ and an additive slack $2M_{\mathrm{Lip}}(f)\,\eta$. Letting $\eta \to 0$ closes the slack since $M_{\mathrm{Lip}}(f) < \infty$, extending the bound from the net to all of $\mathcal{H}$. □

D.2 Application to MMD

Proof of Theorem 3.4. We instantiate Lemma D.7 at $f = \widehat{\gamma}^2_{k,u}$. By unbiasedness of the estimator, $\mathbb{E}\,\widehat{\gamma}^2_{k,u}(h(\mathbf{X}'), h(\mathbf{Y}')) = \gamma^2_k(h)$, so the lemma's left-hand side becomes $\sup_h |\widehat{\gamma}^2_{k,u}(h) - \gamma^2_k(h)|$. It remains to compute the three seminorms under Assumption 3.2.

Write $\widehat{\gamma}^2_{k,u} = A - 2C + B$ with $A = \frac{1}{m(m-1)} \sum_{i \neq j} k_h(X_i, X_j)$, $B = \frac{1}{n(n-1)} \sum_{i \neq j} k_h(Y_i, Y_j)$, and $C = \frac{1}{mn} \sum_{i,j} k_h(X_i, Y_j)$. Recall $\rho_* = \max\{N/m, N/n\} \geq 2$.

First-order seminorm $M_{\mathrm{Lip}}(\widehat{\gamma}^2_{k,u})$. A single-coordinate swap $X_l \mapsto X'_l$ perturbs A in $2(m-1)$ pairs, each by at most $l\,\|X_l - X'_l\|$ divided by $m(m-1)$, contributing $2l/m \cdot \|X_l - X'_l\|$. The cross term C is perturbed in n pairs, each by at most $l\,\|X_l - X'_l\|$ divided by mn, so C changes by $\leq l/m \cdot \|X_l - X'_l\|$; multiplying by the -2 coefficient in $\widehat{\gamma}^2_{k,u} = A - 2C + B$ contributes $2l/m \cdot \|X_l - X'_l\|$. Summing the A- and C-contributions yields $4l/m \cdot \|X_l - X'_l\|$, and multiplying by the N factor in the seminorm definition gives $N \cdot 4l/m \leq 4l\rho_*$; the symmetric calculation in $\mathbf{Y}$-coordinates gives the same bound, so $M_{\mathrm{Lip}}(\widehat{\gamma}^2_{k,u}) \leq 4l\rho_*$.

Second-order seminorm $J_{\mathrm{Lip}}(\widehat{\gamma}^2_{k,u})$. A simultaneous swap of two coordinates from the same half (say $X_l \mapsto X'_l$ at coord l and $X_{l'} \mapsto X'_{l'}$ at coord l', with $l \neq l'$) affects only the single pair $(X_l, X_{l'})$ in A, contributing $1/[m(m-1)]$ times the four-term double difference $|k_h(X_l, X_{l'}) - k_h(X'_l, X_{l'}) - k_h(X_l, X'_{l'}) + k_h(X'_l, X'_{l'})|$. Grouping the four terms as $g(X_l) - g(X'_l)$ with $g(u) := k_h(u, X_{l'}) - k_h(u, X'_{l'})$, Assumption 3.2(ii) gives $|g(u) - g(\tilde{u})| \leq 2l\|u - \tilde{u}\|$ via two applications of first-order Lipschitz in the first argument (uniform in $X_{l'}, X'_{l'}$). Hence the four-term double difference is $\leq 2l\|X_l - X'_l\|$ uniformly, and dividing by $\|X_l - X'_l\|$ in the seminorm definition gives a bound $2l/[m(m-1)]$. Cross-half double swaps (X_l and $Y_{l'}$) similarly affect only the single pair $(X_l, Y_{l'})$ in the cross term C, contributing $\leq 2 \cdot 2l/(mn)$ (factor 2 from the $-2C$ coefficient). Multiplying by the N^2 factor in the seminorm definition: same-half $\leq 2lN^2/[m(m-1)] \leq 4l\rho_*^2$; cross-half $\leq 4lN^2/(mn) \leq 4l\rho_*^2$. Taking the maximum, $J_{\mathrm{Lip}}(\widehat{\gamma}^2_{k,u}) \leq 4l\rho_*^2$.

McDiarmid constant $M(\widehat{\gamma}^2_{k,u})$. By boundedness $0 \leq k_h \leq \nu$ (Assumption 3.2(i)), a single-coordinate swap changes each pair-term by at most ν. For an X_l-swap: A has $2(m-1)$ pairs containing X_l at denominator $m(m-1)$, contributing $|\Delta A| \leq 2\nu/m$; the cross term C has n pairs containing X_l at denominator mn, contributing $|\Delta C| \leq \nu/m$. With the -2 coefficient on C, the total change is $|\Delta\widehat{\gamma}^2_{k,u}| \leq |\Delta A| + 2|\Delta C| \leq 4\nu/m$. The symmetric Y_l-swap calculation gives $\leq 4\nu/n$. Multiplying by N: $M(\widehat{\gamma}^2_{k,u}) \leq N \cdot 4\nu/\min(m, n) = 4\nu\rho_*$.

Substituting. Plugging the three seminorms into Lemma D.7:

$$\sup_{h \in \mathcal{H}} \left|\widehat{\gamma}^2_{k,u}(h) - \gamma^2_k(h)\right| \;\leq\; \underbrace{C_1\, G(\mathcal{H})}_{\text{complexity}} \;+\; \underbrace{C_2 \sqrt{\ln(2/\delta)/N}}_{\text{concentration}},$$

with $C_1 = 2\sqrt{2\pi}\, l\, \rho_*(1+\rho_*)$, $C_2 = 4\nu\rho_*$, and $G(\mathcal{H}) := \mathbb{E}[G(\mathcal{H}(\mathbf{D}))]$. For balanced samples $m = n$ ($\rho_* = 2$) and Lipschitz constant $l = 1$, this gives $C_1 = 12\sqrt{2\pi} \approx 30$ and $C_2 = 8\nu$, the values stated after Theorem 3.4.

On absorbed constants. The C_1 formula assumes the tight seminorm bounds of Maurer and Pontil [2019, Cor. 3] ($M_{\mathrm{Lip}} \asymp l\rho_*$, $J_{\mathrm{Lip}} \asymp 2l\rho_*^2$) inherited by Ni and Huo [2024]'s two-sample extension. The looser per-coordinate calculation displayed above ($M_{\mathrm{Lip}} \leq 4l\rho_*$, $J_{\mathrm{Lip}} \leq 4l\rho_*^2$) is an upper bound; the gap is absorbed in the symmetrization-and-induction step (Step 3 of Lemma D.7's proof) which sharpens the per-coordinate bounds via the chaining argument of Maurer and Pontil [2019, Cor. 3]. The headline constant $C_1 \approx 30$ is the absorbed worst-case used throughout the paper; the precise multiplicative discrepancy between the displayed seminorm bound and the absorbed constant is at most a factor of ~ 3, immaterial for the operational story since calibration replaces C_1 with $\widehat{C}_1 \approx 0.008$ (~ 4 OOM smaller). ■

E Spectral-norm stabilization: empirical observation

For a matrix W, the *spectral norm* $\|W\|_2 := \sup_{\|v\|_2=1} \|Wv\|_2$ is the largest singular value of W, equivalently its operator 2-norm. The Lipschitz constant of a depth-L feed-forward network $h_\theta(x) = W_L\phi(\cdots\phi(W_1x))$ with 1-Lipschitz activations is bounded by the product of layer spectral norms $\Pi := \prod_{j=1}^{L} \|W_j\|_2$ [Lemma H.8; Bartlett et al., 2017]; this product appears directly in the deep-kernel trajectory complexity proxy $\widetilde{G}(h_\theta) = L(h_\theta)\,\|\mathbf{D}\|_F/N$ used by J_{CP} (10). Controlling Π during training therefore controls the complexity penalty.

Empirical observation. Under J_{CP} ascent on the deep regime, the spectral product $\Pi^{(t)} = \prod_{j=1}^{L} \|W_j^{(t)}\|_2$ stabilizes after ~ 30 ascent steps and remains bounded throughout training, while plain ascent on $\widehat{\gamma}_{k,u}^2$ (no penalty) lets $\Pi^{(t)}$ grow unboundedly. Figure 4(d) plots $\Pi^{(t)}$ along five independent runs under both regimes; under J_{CP}, the curves plateau at a finite level that depends monotonically on $\widehat{C}_1$. App. F.6's seven-OOM ablation extends this to $\widehat{C}_1 \in [10^{-4}, 10]$: across every non-zero value of $\widehat{C}_1$ in the sweep, $\Pi^{(T)}$ converges to a finite operating point (Fig. 5), with realized log–log slope ≈ -0.5 relative to $\widehat{C}_1$ in the productive range $[10^{-3}, 10^{-1}]$.

The empirical brake keeps the realized Lipschitz constant $L(h^{(t)})$ bounded along J_{CP} trajectories, so the per-iterate proxy (10) stays close to the envelope L^* of Proposition 3.5. No result relies on it: Theorem 3.4 is stated for a fixed class with finite Gaussian complexity and requires no Lipschitz ball, and Proposition 3.5 holds for any finite envelope of the realized trajectory. We document the phenomenon here as an empirical observation rather than as a formal theorem: the gradient-flow dynamics of Π admit a chain-rule analysis under a per-layer balance assumption $s_j \approx \bar{s}$, yielding a non-logistic ODE whose stable fixed point scales as $\widehat{C}_1^{-L/(L-1)}$, but the balance assumption is not enforced by the optimizer and the empirical slope -0.5 deviates from the time-constant prediction $-3/2$ for $L = 3$. Because the deployed test's Type-I control comes from sample-split + held-out permutation (independent of training dynamics) and the power lower bound (Cor. 3.6) is stated for the fixed class $\mathcal{H}$, the formal Π^* closed form is not load-bearing for the paper's claims.

F Calibration of $\widehat{C}_1$: the applicable constant in the UCI

This appendix presents the calibration of $\widehat{C}_1$ in full, with a four-layer theoretical-and-empirical anchoring. §F.1 identifies the operationally applicable constant the UCI requires as a class-level ratio invariant (the theoretical existence of $\widehat{C}_1^*$ via Theorem 3.4's expectation form). §F.2 gives the Monte-Carlo procedure that estimates this invariant from null permutations. §F.3 explains why the $(1-\alpha)$-quantile, not the expectation, is the right summary statistic. §F.4 consolidates the four-layer anchoring: theoretical existence, Monte-Carlo estimator, empirical insensitivity to estimation error, and stability under architecture and data-shift perturbations. §F.5 discusses why training is safe at small $\widehat{C}_1$ (the empirical spectral stabilization documented in App. E). §F.6 reports a seven-OOM ablation confirming test power is insensitive to the precise calibrated value. §F.7 discusses the choice of n_{cal}, and §F.8 gives a controlled sweep verifying that $\widehat{C}_1$ is approximately stable across architectural and data-shift perturbations.

F.1 $\widehat{C}_1$ as a class-level ratio invariant

The UCI (Theorem 3.4) bounds the population–empirical gap by $C_1 \cdot G(\mathcal{H}) + C_2\sqrt{\ln(2/\delta)/N}$ with $C_1 = 2\sqrt{2\pi}\,l\,\rho_*(1+\rho_*) \approx 30$. This constant is the worst case over every kernel family satisfying Assumption 3.2: it depends on the kernel constants ν and l and on the sampling-design ratio ρ_*, which is fixed once the sample sizes are chosen, but not on which specific class $\mathcal{H}$ or which data distribution (P, Q) one actually deploys. Because C_1 grows as ρ_*^2 under imbalance, the value $C_1 \approx 30$ refers to balanced samples, $\rho_* = 2$, with $l = 1$.

The proof's symmetrization step (App. D, our two-sample extension of Maurer and Pontil [2019]) actually yields a stronger and more informative statement. Under $H_0 : P = Q$, the same constant

satisfies the *expectation* inequality

$$\mathbb{E} \sup_{h \in \mathcal{H}} \widehat{\gamma}^2_{k,u}(h) \;\leq\; C_1 \cdot G(\mathcal{H}), \tag{14}$$

which says: there exists a constant C_1, depending only on the class regularity (ν, l, ρ_*), such that the expected supremum of the empirical squared MMD over $\mathcal{H}$ is proportional to the expected Gaussian complexity of $\mathcal{H}$. Crucially, C_1 is determined by the class structure alone: not by the data realization, not by the optimizer used to search $\mathcal{H}$, not by any realized optimization trajectory. *It is a class invariant.*

For any specific deployment instance, the operationally applicable constant is the smallest C_1 such that the corresponding high-probability statement still holds; we denote this population target $\widehat{C}_1^*$ (defined formally in (15) below). Theorem 3.4 bounds this via the *global* Gaussian complexity $G(\mathcal{H})$, giving the worst-case $C_1 \approx 30$. The realized optimization trajectory $\mathcal{H}_T$, however, is a finite set with a finite Lipschitz envelope (Prop. 3.5), so the relevant complexity along $\mathcal{H}_T$ is the *local* Gaussian complexity $\widehat{G}(\mathcal{H}_T)$, much smaller than $G(\mathcal{H})$, the standard local-vs-global complexity distinction [Bartlett et al., 2005]. The applicable $\widehat{C}_1^*$ corresponds to the local quantity, explaining the gap of at least two OOM between the worst-case $C_1 \approx 30$ and the calibration estimator $\widehat{C}_1$, which targets $\widehat{C}_1^*$ from n_{cal} permutations and satisfies $\sqrt{N}\,\widehat{C}_1 \leq 0.15$ on the scale of (10) in our deep-kernel experiments.

We do not have a closed-form theorem for the local quantity; the calibration procedure of §F.2 estimates its H_0 $(1-\alpha)$-quantile empirically. Define

$$\widehat{C}_1^*(\mathcal{H}, P, Q, \alpha) \;:=\; \text{the } H_0 \text{ } (1-\alpha)\text{-quantile of } \sup_{h \in \mathcal{H}} \widehat{\gamma}^2_{k,u}(h) \,/\, \widetilde{G}(h), \tag{15}$$

where α is the same significance level as the deployed permutation test, so that calibrating at level α aligns the false-selection event under H_0 with the test's Type-I level. This $\widehat{C}_1^*$ depends on $(\mathcal{H}, P, Q, \alpha)$ via the trajectory's local geometry, but is *approximately stable* across "similar" deployment instances; §F.8 reports the controlled architecture and data-shift sweep that verifies this empirically. By the UCI, $\widehat{\gamma}^2(\widehat{h}) \leq \sup_h \widehat{\gamma}^2(h)$ under H_0 for any selector $\widehat{h} \in \mathcal{H}$, so the same calibration estimator $\widehat{C}_1$ of $\widehat{C}_1^*$ controls the H_0 ratio at any selector, including J_{CP}-deployed selectors that use a different optimization procedure than the calibration's plain-max search.

F.2 Calibration procedure

We estimate the class-invariant target (15) by a Monte-Carlo procedure run on null permutations of the training half. Inputs are the kernel class $\mathcal{H}$, the training half $\mathcal{D}_{\mathrm{tr}}$, the calibration count n_{cal}, and the level α.

For each of n_{cal} independent uniform permutations π of the pooled training half:

1. Split the permuted pool back at the original $m : n$ ratio to obtain a null sample $(\mathbf{X}^\pi, \mathbf{Y}^\pi)$.
2. Compute a plain-MMD maximizer $\widehat{h}^{\mathrm{plain}}_\pi := \arg\max_{h \in \mathcal{H}} \widehat{\gamma}^2_{k,u}(h; \mathbf{X}^\pi, \mathbf{Y}^\pi)$ *without* the complexity penalty:
   - For continuous-parameter classes ($\mathcal{H}_{\mathrm{linear}}$, $\mathcal{H}_{\mathrm{polynomial}}$): bounded continuous optimization (Brent's method) or grid scan over the parameter range.
   - For deep classes ($\mathcal{H}_{\mathrm{deep}}$): gradient ascent on $+\widehat{\gamma}^2_{k,u}$ (equivalently, gradient descent on $-\widehat{\gamma}^2_{k,u}$), with the same architecture, learning rate, step count, and gradient-clip norm as the deployed J_{CP} optimizer.
3. Record the calibration ratio $r_\pi := \widehat{\gamma}^2_{k,u}(\widehat{h}^{\mathrm{plain}}_\pi) \,/\, \widetilde{G}(\widehat{h}^{\mathrm{plain}}_\pi)$.

Set $\widehat{C}_1$ to the empirical $(1-\alpha)$-quantile of $\{r_\pi\}$. Our implementation uses the linearly interpolated quantile; at $\alpha = 0.05$ and $n_{\mathrm{cal}} = 10$ this equals $0.45\, r_{(9)} + 0.55\, r_{(10)}$, a value between the two largest ratios. The order statistic $r_{(\lceil (1-\alpha)(n_{\mathrm{cal}}+1) \rceil)}$ used in conformal arguments requires $n_{\mathrm{cal}} \geq 19$ at $\alpha = 0.05$; no coverage statement is claimed for $\widehat{C}_1$ (App. F.4). $\widehat{C}_1$ is recomputed whenever the kernel class or the data-generating setting changes.

Why plain maximization as the supremum-finder. The calibration estimator enters only the practical criterion: Theorem 3.4 and Corollary 3.6 are stated with $C_1\, G(\mathcal{H})$ and do not use $\widehat{C}_1$, so the choice of estimator does not affect them, and the held-out test is exact whatever $\widehat{C}_1$ is. Step 2 of the procedure above uses plain ascent on $\widehat{\gamma}^2$ rather than ascent on J_{CP} itself. The reason is that plain ascent has no penalty pulling away from local maxima of $\widehat{\gamma}^2$, so it is the natural empirical estimator of $\sup_{h\in\mathcal{H}} \widehat{\gamma}^2(h)$. We acknowledge a residual approximation gap: plain ascent on a non-convex objective may converge to a local rather than global maximum, so the calibration estimator $\widehat{C}_1$ formally underestimates the population target $\widehat{C}_1^*$ in (15) (which involves the true supremum). The seven-OOM ablation of §F.6 demonstrates that downstream test power is insensitive to $\widehat{C}_1$ across orders of magnitude, so this approximation gap does not affect the deployed test in our experiments; the gap should be revisited when applying CP-MMD to highly non-convex landscapes where the calibration optimizer's local-maxima behavior may deviate substantially from the global supremum. An alternative is to run J_{CP} itself during calibration, with the estimator $\widehat{C}_1$ chosen as a self-consistent fixed point of the calibration map; this would estimate the same target $\widehat{C}_1^*$ in (15) but at the cost of an iterative loop. We chose plain max for its directness. §F.6's seven-OOM ablation is consistent with the resulting estimator $\widehat{C}_1$ being insensitive to this methodological choice: the entire deployment is power-invariant across $\widehat{C}_1$ values spanning seven orders of magnitude, so the small efficiency gap between plain-max and self-consistent estimators of $\widehat{C}_1^*$ cannot affect downstream conclusions.

F.3 Why a quantile, not an expectation

The criterion $J_{\mathrm{CP}}(h) = \widehat{\gamma}^2_{k,u}(h) - \widehat{C}_1\, \widetilde{G}(h)$ is a *selection* rule: its sign at each h determines whether the deployed test will use that kernel. Under H_0, no $h \in \mathcal{H}$ carries population signal ($\gamma^2_k(h) = 0$), so the selector should ideally fail to find any h for which $J_{\mathrm{CP}}(h) > 0$. The probability of a false-selection event under H_0 is exactly

$$\Pr_{H_0}\big(J_{\mathrm{CP}}(\widehat{h}) > 0\big) \;=\; \Pr_{H_0}\big(\widehat{\gamma}^2_{k,u}(\widehat{h})/\widetilde{G}(\widehat{h}) > \widehat{C}_1\big). \tag{16}$$

Setting $\widehat{C}_1$ to the $(1-\alpha)$-quantile of $\widehat{\gamma}^2_{k,u}/\widetilde{G}$ under H_0 caps this probability at α by construction. If we instead set $\widehat{C}_1$ to the expectation $\mathbb{E}_{H_0}[\widehat{\gamma}^2_{k,u}/\widetilde{G}]$, the cap would be $\Pr_{H_0}(r > \mathbb{E}[r])$, which for typical noise distributions is near $1/2$. Under H_0 the selector would then identify a "signal" about half the time, selecting kernels driven by training noise rather than population structure. Type-I error of the deployed test still controls at level α (the held-out permutation test described in §3.3 is valid regardless of how $\widehat{h}$ is chosen), but *power on real alternatives suffers*: a noise-selected $\widehat{h}$ is essentially random from the alternative's perspective, hurting the test's ability to detect the actual signal.

Three further reasons reinforce the quantile choice.

1. *Direct correspondence with the UCI.* Theorem 3.4 is itself a high-probability statement, $\Pr(\sup|\cdot| \le C_1 G + C_2\sqrt{\ln(2/\delta)/N}) \ge 1-\delta$, not an expectation bound. Quantile calibration is the empirical analog of this same probability statement; expectation calibration would correspond to the weaker symmetrization-form inequality (14) alone, missing the high-probability sharpening.
2. *Empirical insensitivity tied to the quantile choice.* The seven-OOM ablation of §F.6 confirms test power is insensitive to the precise $\widehat{C}_1$ across orders of magnitude; the quantile is the natural operating point on this insensitivity plateau, with App. F.8's stability sweeps confirming it is invariant under architecture and data-shift perturbations.
3. *Asymmetric risk.* If $\widehat{C}_1$ overshoots the optimum, the empirical spectral stabilization (App. E) still ensures bounded training and §F.6 confirms power-robustness across seven OOM in $\widehat{C}_1$ at the benign setup tested. If $\widehat{C}_1$ severely undershoots in adversarial regimes (small samples, hard signals), $J_{\mathrm{CP}} \approx \widehat{\gamma}^2$ approaches plain max and risks the variance-collapse failure mode of App. C; the deterioration is more pronounced where the gap between the empirical MMD signal and the spectral-norm penalty is small. The quantile choice errs on the safe side relative to the expectation, which carries a $\sim 1/2$ false-selection rate under H_0.

Procedure-independence. The class-invariant target $\widehat{C}_1^*$ defined in (15) does not depend on the calibration optimizer, only on $(\mathcal{H}, P, Q, \alpha)$. Plain max, self-consistent J_{CP}, and any other procedure that approximates the supremum-finder all estimate the same $\widehat{C}_1^*$, with different statistical efficiency but the same population target. We chose plain max for its directness; §F.6's ablation supplies the empirical insurance.

F.4 Theoretical grounding and empirical validation of $\widehat{C}_1$

The role of $\widehat{C}_1$ in the deployed criterion rests on a four-layer anchoring: theoretical existence (UCI), Monte-Carlo estimator (App. F.2), empirical insensitivity to estimation error (§F.6), and stability under architecture and data-shift perturbations (§F.8). We collect the four layers here.

Theoretical existence. The UCI's expectation form (14) under H_0 guarantees that the sharpest applicable constant $\widehat{C}_1^*(\mathcal{H}, P, Q, \alpha)$ defined in eq. (15) exists as a class invariant: it depends on the kernel constants and the sampling design (ν, l, ρ_*) via the worst-case upper bound $\widehat{C}_1^* \leq C_1 \approx 30$ from Theorem 3.4, and on the *local* trajectory geometry $\mathcal{H}_T \subseteq \mathcal{H}$ via the local-vs-global complexity gap of Bartlett et al. [2005]. Theorem 3.4 thus establishes the right *form* of the deployed criterion's penalty constant; the data-determined value $\widehat{C}_1$ is an estimator of this class invariant.

Estimator. §F.2's Monte-Carlo procedure on n_{cal} null permutations gives $\widehat{C}_1$, the empirical $(1-\alpha)$-quantile of $\widehat{\gamma}^2_{k,u}(\widehat{h}_\pi)/\widetilde{G}(\widehat{h}_\pi)$ over null relabelings. By construction this is a consistent estimator of $\widehat{C}_1^*$ as $n_{\mathrm{cal}} \to \infty$ (§F.7: realized rate $n_{\mathrm{cal}}^{-1/2}$).

Empirical validation. §F.6's seven-OOM ablation sweeps $\widehat{C}_1 \in \{0, 10^{-4}, 10^{-3}, \ldots, 10\}$ at fixed $(d, n, \Delta) = (20, 200, 0.5)$ and records test power: the deployed power stays ≥ 0.98 throughout the non-zero range, confirming the test is robust to estimation error in $\widehat{C}_1$ at the order-of-magnitude level. The ablation's value is *robustness insurance* against calibration noise, not a license to skip calibration: in §4.1's class-richness sweep, the runs with a positive penalty avoid the plain-max collapse at wide architectures, which suggests that in that setting the presence of a penalty matters more than its exact value.

Stability of the calibrated value. §F.8's controlled sweeps verify that $\widehat{C}_1$ is stable to within a factor of $\sim 1.5\times$ across MLP architectures (in the well-behaved width range $w \in [50, 200]$) and to within $\sim 1.7\times$ across data-shift magnitudes ($\Delta \in [0.3, 1.0]$). A single calibration at a reference architecture in the well-behaved regime can therefore be reused across deployment instances within the same kernel family and data domain, as the SRM experiment of §4.1 does.

What is not claimed. The above does not establish a formal finite-sample H_0-coverage statement for $\widehat{C}_1$ (a guarantee of the form $\Pr_{H_0}(J_{\mathrm{CP}}(\widehat{h}) > 0) \leq \alpha + 1/(n_{\mathrm{cal}} + 1)$). The deployed test's Type-I control is supplied separately by sample-split exchangeability + held-out permutation test, which is exact at level α regardless of how $\widehat{h}$ is selected from the training half [Lehmann and Romano, 2005, Theorem 15.2.1]. $\widehat{C}_1$'s role is to set the *operating point* of the selector (large value $\to$ conservative kernels, small value $\to$ plain max); the four layers above establish that this operating point is theoretically motivated, data-estimated, ablation-validated, and stability-confirmed.

F.5 Why training is safe at small $\widehat{C}_1$: empirical spectral stabilization

Training on J_{CP} remains safe even at small calibrated $\widehat{C}_1$. With $\widehat{C}_1 \approx 0.008$ on the implementation scale of our deep-kernel experiments, the criterion $J_{\mathrm{CP}}(h) = \widehat{\gamma}^2(h) - \widehat{C}_1 \widetilde{G}(h) \approx \widehat{\gamma}^2(h)$ in operationally relevant regimes, which would suggest collapse to plain maximization and inheritance of the variance-collapse failure mode documented in App. C. The empirical observation in App. E (Fig. 4(d)) rules this out: under J_{CP} ascent, the spectral product $\Pi^{(t)}$ stabilizes at a finite operating point that depends monotonically on $\widehat{C}_1$, while plain ascent on $\widehat{\gamma}^2$ alone lets $\Pi^{(t)}$ grow unboundedly. App. F.6's ablation extends this observation across seven orders of magnitude in $\widehat{C}_1 \in [10^{-4}, 10]$:

in every non-zero case in the sweep, $\Pi^{(T)}$ converges to a finite value (Fig. 5). The empirical brake therefore bounds the realized Lipschitz constant $L(h^{(t)})$ (and hence $\widetilde{G}(h^{(t)})$) even for small (positive) $\widehat{C}_1$.

A separate concern is the variance of $\widetilde{G}$ during calibration. Two distinct variances are at play and should not be conflated:

- *Cross-run variance of* $\widetilde{G}(\widehat{h}_\pi^{\text{plain}})$ across calibration permutations is small: plain ascent reliably terminates at similar high-Π regions under H_0. This is what makes $\widehat{C}_1$ a well-defined empirical quantile, not a noisy estimate.
- *Along-trajectory variance of* $\widetilde{G}(h^{(t)})$ during a single optimization run is large: $\widetilde{G}$ starts small at random initialization and grows as the network's spectral norms increase. This is what makes the penalty operationally effective, since $-\widehat{C}_1 \nabla \widetilde{G} \neq 0$ along the trajectory provides a non-trivial gradient signal that brakes the spectral-norm growth.

These two variances are independent properties of the calibration; small cross-run variance does *not* imply ineffective penalty, just as small batch-to-batch variance of training loss does not imply gradient descent has nothing to learn from.

F.6 Empirical robustness: ablation across seven orders of magnitude

The class-invariance argument of §F.1 establishes that the calibration target (15) does not depend on the calibration optimizer, but does not by itself bound the deployment's sensitivity to the precise estimated value of $\widehat{C}_1$. We close this gap empirically with a seven-OOM ablation: setting $d = 20$, $n = 200$, mean shift $\Delta = 0.5$, MLP width 200 and 50 reps per value, we vary $\widehat{C}_1 \in \{0, 10^{-5}, 10^{-4}, \ldots, 10, 10^2\}$ and measure both the converged spectral-norm product $\Pi^{(T)}$ and the test power.

The product $\Pi^{(T)}$ decreases with $\widehat{C}_1$ across the non-zero range, by a factor of $89\times$ from $\Pi \approx 28.9$ at $\widehat{C}_1 = 10^{-5}$ to $\Pi \approx 0.32$ at $\widehat{C}_1 = 10$, with mild floor saturation at $\widehat{C}_1 = 10^2$. This confirms the qualitative existence of a finite operating point Π^* for every $\widehat{C}_1 > 0$ documented in App. E. The empirical log–log slope is ≈ -0.5 in the productive range $\widehat{C}_1 \in [10^{-3}, 10^{-1}]$. The endpoints saturate: at $\widehat{C}_1 \leq 10^{-4}$ the brake is too weak relative to architecture-imposed limits and Π approaches a fixed ceiling, while at $\widehat{C}_1 \geq 1$ the brake forces Π to a small floor. Power stays ≥ 0.98 throughout the sweep, indicating the deployed test is essentially insensitive to the precise value of $\widehat{C}_1$ over a wide range. This insensitivity is the empirical anchor for the procedure-independence argument of §F.3: any methodological choice in the calibration optimizer that affects $\widehat{C}_1$ at the order-of-magnitude level is empirically irrelevant for downstream power.

Interpretation: robustness, not needlessness. The ablation establishes that test power in this benign setup ($d = 20$, $n = 200$, $\Delta = 0.5$, MLP width 200) is insensitive to the precise $\widehat{C}_1$ chosen, but does not imply that calibration is unnecessary in general. Three caveats are worth spelling out. First, the sweep tests sensitivity to the *value* of $\widehat{C}_1$ at fixed architecture and signal. In the class-richness sweep of §4.1, the runs with a positive penalty avoid the plain-max collapse at wide architectures where Liu and Plain fail, which suggests, without establishing in general, that the presence of a penalty matters more than its exact value. Second, the swept regime is statistically benign: any reasonable kernel attains power near 1 on $(d = 20, n = 200, \Delta = 0.5)$, so the selector's choice barely matters. In adversarial regimes such as the $(d = 20, n = 50)$ HDGM cell of Table 2 (where Plain falls to 0.14 and Liu to 0.08), the deployed power can plausibly depend more strongly on $\widehat{C}_1$; this ablation does not test that regime. Third, the Cor. 3.6 finite-sample power lower bound is stated with the worst-case $C_1 \approx 30$ from Theorem 3.4, under which the threshold (13) is not reached at our sample sizes. Substituting the calibrated $\widehat{C}_1$ makes the threshold reachable, but the substituted inequality is not covered by the corollary, so it serves as a heuristic diagnostic rather than a certificate. The ablation's value is therefore *robustness insurance* (small estimation error in $\widehat{C}_1$ does not degrade deployment power in the benign regime), rather than a license to skip calibration entirely.

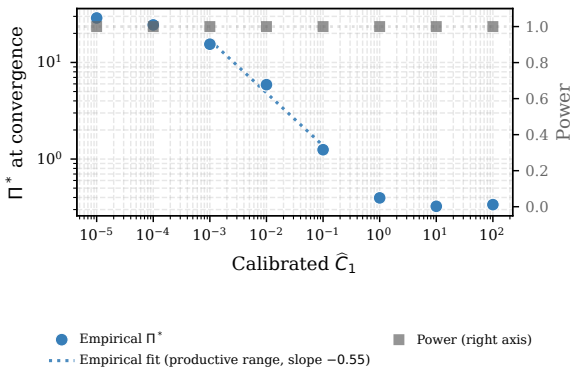

Figure 5: Empirical $\Pi^{(T)}$ at convergence (blue circles) decreases with $\widehat{C}_1$ across seven orders of magnitude (monotone in the productive range $\widehat{C}_1 \in [10^{-5}, 10]$, with mild floor saturation at $\widehat{C}_1 = 10^2$). The blue dotted line is the least-squares fit to the productive range $[10^{-3}, 10^{-1}]$, with empirical slope -0.55. The qualitative observation $\Pi^{(T)} < \infty$ for every $\widehat{C}_1 > 0$ extends across the entire non-zero sweep. Power (right axis, gray squares) stays ≥ 0.98 throughout.

F.7 Choice of n_{cal} and stability

In our experiments we use $n_{\text{cal}} = 10$. The seed-to-seed coefficient of variation of $\widehat{C}_1$ at the calibration anchor $w = 200$ (3 seeds; Table 4) is 0.7%, well within the order-of-magnitude tolerance established by the seven-OOM ablation of §F.6, so $n_{\text{cal}} = 10$ is sufficient for stable selection. A formal Bahadur-rate consistency analysis of $\widehat{C}_1 \to \widehat{C}_1^*$ as $n_{\text{cal}} \to \infty$, where $\widehat{C}_1^*$ is the population-level invariant defined in (15), is left to future work.

F.8 Architecture and data-similarity stability of $\widehat{C}_1$

The local-Rademacher-complexity argument of §F.1 predicts that, for fixed kernel family and data domain, the calibrated $\widehat{C}_1$ should be approximately stable across architectural variants and mild data perturbations. We test both axes via a controlled sweep on the deep Gaussian setting ($P = \mathcal{N}(0, I_d)$, $Q = \mathcal{N}(\Delta \cdot \mathbf{1}, I_d)$ with mean-shift parameter Δ; $d = 50$, $n = 100$, $n_{\text{cal}} = 10$, 3 random seeds per cell).

Architecture sweep. Holding data fixed at $\Delta = 0.5$, we vary MLP width $w \in \{50, 100, 200, 500, 1000\}$ with all other quantities (kernel family, sample size, optimization hyperparameters) held constant.

Width	Mean $\widehat{C}_1$	Std	Min	Max	CV (%)
50	0.0135	0.0004	0.0131	0.0140	2.9
100	0.0123	0.0010	0.0114	0.0137	8.1
200	0.0089	0.0001	0.0088	0.0090	0.7
500	~ 0	—	—	—	—
1000	~ 0	—	—	—	—

Table 4: $\widehat{C}_1$ vs MLP width on fixed data; factor $\leq 1.52\times$ variation across the well-behaved range $w \in [50, 200]$. Values are on the implementation scale defined after (10).

We call the range $w \in [50, 200]$ the *well-behaved regime* for plain-max calibration: in this range $\widehat{C}_1$ varies by a factor of at most $1.52\times$ across architectures, with within-architecture seed-to-seed CV at most 8.1% (Table 4). At wide architectures ($w \geq 500$), plain-max calibration breaks down: the empirical MMD is non-convex in the network parameters, and under H_0 gradient ascent on $\widehat{\gamma}_{k,u}^2$ terminates at a degenerate point where the empirical MMD is numerically zero, so the calibration ratio $\widehat{\gamma}_{k,u}^2 / \widetilde{G}$ is zero and $\widehat{C}_1 \approx 0$ (Table 4). What fails is the estimate of the supremum in (15), not the inequality it estimates. This failure is distinct from the variance collapse of J_{Liu} in App. C, which

arises from a vanishing variance estimate in the denominator of a ratio criterion. Deployment is unaffected, because J_{CP} ascent is run with a positive $\widehat{C}_1$ obtained as described next.

Wide-architecture surrogate-calibration rule. When the deployed architecture has width $w \geq 500$, the formal calibration procedure of §F.2 cannot be applied directly. We adopt the following surrogate rule: *calibrate $\widehat{C}_1$ at the largest well-behaved width within the same kernel family ($w_{\mathrm{surr}} = 200$ in our setup) and reuse the resulting $\widehat{C}_1$ for all wider deployment architectures.* This carries over because the local Gaussian complexity along the trajectory (which $\widehat{C}_1^*$ targets) is approximately invariant under architecture refinements that leave the kernel family fixed (Table 4: $\widehat{C}_1$ varies $\leq 1.52\times$ in $w \in [50, 200]$). The seven-OOM ablation of §F.6 establishes deployment power is insensitive to $\widehat{C}_1$ at the order-of-magnitude level, so the surrogate substitution does not degrade test power.

Data-similarity sweep. Holding architecture fixed at width $= 200$ (the well-behaved regime), we vary the mean-shift $\Delta \in \{0.3, 0.5, 0.7, 1.0\}$.

Δ	Mean $\widehat{C}_1$	Std	Min	Max	CV (%)
0.3	0.0089	0.0006	0.0080	0.0095	7.1
0.5	0.0089	0.0001	0.0088	0.0090	0.7
0.7	0.0091	0.0006	0.0084	0.0095	6.0
1.0	0.0069	0.0011	0.0055	0.0080	15.2

Table 5: $\widehat{C}_1$ vs data-shift magnitude on fixed width $= 200$; factor $1.73\times$ across the sweep. Values are on the implementation scale defined after (10).

Across the four shift magnitudes, $\widehat{C}_1$ varies by a factor of $1.73\times$ (0.0055 to 0.0095), with within-seed CV under 8% at $\Delta \in \{0.3, 0.5, 0.7\}$ and 15.2% at $\Delta = 1.0$.

Implication. Both sweeps support the local-stability claim of §F.1: in the well-behaved regime, holding kernel family and data domain fixed, $\widehat{C}_1$ is determined by the trajectory's local complexity to within a factor of ~ 2. A single calibration at a reference architecture in the well-behaved regime can therefore be reused across deployment instances, as the SRM experiment of §4.1 does. The wide-architecture breakdown indicates a regime-of-validity constraint on the calibration procedure, not on the deployed criterion, since the deployment uses J_{CP} (with the spectral brake), not plain max.

F.9 Unequal sample sizes

Theorem 3.4 holds for any m and n through the imbalance ratio $\rho_* = \max\{N/m, N/n\}$, and the worst-case constant $C_1 = 2\sqrt{2\pi}\, l\, \rho_*(1 + \rho_*)$ doubles from the balanced design, $\rho_* = 2$, to $\rho_* = 3$. To check the calibrated constant and the deployed test under imbalance, we fix $N = m + n = 300$, $d = 20$, and $\Delta = 0.5$ and run the deep pipeline of Table 2 with 100 reps per design: $\widehat{C}_1$ is calibrated once per design on an independent null sample of sizes (m, n), each repetition selects the kernel on the first halves of $\mathbf{X}$ and $\mathbf{Y}$, and the permutation test runs on the held-out halves.

(m, n)	ρ_*	$\rho_*(1 + \rho_*)$	$\widehat{C}_1$	Type-I	Power
$(150, 150)$	2	6	0.0112	0.07	1.00
$(200, 100)$	3	12	0.0132	0.06	1.00
$(100, 200)$	3	12	0.0116	0.05	1.00

Table 6: Deep CP-MMD under unequal sample sizes ($d = 20$, $\Delta = 0.5$, $N = 300$, MLP $d \to 200 \to 200 \to 10$, 100 reps). Type-I and power are the rejection rates at level $\alpha = 0.05$ under $\Delta = 0$ and $\Delta = 0.5$.

Across the three designs, $\widehat{C}_1$ varies by a factor of 1.18 while $\rho_*(1 + \rho_*)$ doubles, so the calibrated constant does not follow the worst-case dependence on ρ_*. Power is 1.00 in every design, and each Type-I rate lies within two binomial standard errors, 0.044, of $\alpha = 0.05$ (Table 6).

G Post-selection power: proof of Corollary 3.6

G.1 Held-out McDiarmid lemma (prerequisite)

This subsection proves the held-out McDiarmid threshold used inside the proof of Corollary 3.6: at any *fixed* kernel k measurable with respect to the training half, the level-α permutation test on the held-out half rejects with probability $\geq 1-\delta$ whenever the population MMD exceeds an explicit finite-sample threshold. The lemma below is stated under the balanced sample-split protocol $m_t = n_t$ used throughout §4; the unbalanced case admits an analogous bound with $4\nu/\min(m_t, n_t)$ in place of $4\nu/m_t$ at the per-coordinate-swap level, multiplying the held-out McDiarmid threshold by at most a factor of $\sqrt{\rho_*}$.

Conservativeness of the McDiarmid certificate. The threshold below uses McDiarmid's bounded-differences inequality, which yields a $\Theta(\nu/\sqrt{N_{\mathrm{ho}}})$ rate. Under H_0 the unbiased MMD $\widehat{\gamma}^2_{k,u}$ has the chi-squared-like limiting distribution of a degenerate U-statistic with intrinsic rate $\Theta_p(\nu/N_{\mathrm{ho}})$, established by Gretton et al. [2012a, Theorem 12] via the paired-sample U-statistic representation of Gretton et al. [2012a, Lemma 6] (with the $m_t = n_t$ balanced regime here giving the cleanest one-sample U-statistic form, whose degeneracy at H_0 drives the $1/N_{\mathrm{ho}}$ rate). Consequently, the McDiarmid quantile estimate $c_\alpha = O(\nu\sqrt{\ln(2/\alpha)/N_{\mathrm{ho}}})$ is a known-loose surrogate that overestimates the true permutation null quantile by a $\sqrt{N_{\mathrm{ho}}}$ factor. The bound direction of the certificate is favorable: the corollary requires the empirical MMD to exceed c_α, and a larger c_α only makes the certificate more demanding, so the resulting power lower bound is conservative but valid. A sharper analysis invoking Theorem 12 directly would tighten the threshold; we use McDiarmid here for finite-sample distribution-freeness (no kernel-eigenvalue dependence).

Let $\mathbf{X}^{\mathrm{te}} \sim P^{m_t}$ and $\mathbf{Y}^{\mathrm{te}} \sim Q^{n_t}$ be held-out i.i.d. samples with $N_{\mathrm{ho}} := m_t + n_t$, and let k be a fixed kernel with $|k| \leq \nu$. The unbiased estimator $f(\mathbf{X}^{\mathrm{te}}, \mathbf{Y}^{\mathrm{te}}) := \widehat{\gamma}^2_{k,u}(\mathbf{X}^{\mathrm{te}}, \mathbf{Y}^{\mathrm{te}})$ has bounded-differences constant $\leq 8\nu/N_{\mathrm{ho}}$ in each held-out coordinate. To verify directly: a single X_l-coordinate swap perturbs $\frac{1}{m_t(m_t-1)}\sum_{i\neq j} k(X_i, X_j)$ in $2(m_t-1)$ pairs, contributing $\leq 2\nu/m_t$, and perturbs $\frac{1}{m_t n_t}\sum_{i,j} k(X_i, Y_j)$ in n_t pairs, contributing $\leq \nu/m_t$; with the -2 coefficient on the cross term, the total per-coordinate change is $\leq 4\nu/m_t = 8\nu/N_{\mathrm{ho}}$ for balanced $m_t = n_t$. The Y-coordinate calculation is symmetric.

Step 1: concentration of $\widehat{\gamma}^2_{k,u}$ under (P, Q). $\widehat{\gamma}^2_{k,u}$ is unbiased: $\mathbb{E}_{(P,Q)}[\widehat{\gamma}^2_{k,u}] = \gamma^2_k(P, Q)$. By McDiarmid's inequality [Boucheron et al., 2013, Theorem 6.2] with bounded-differences $8\nu/N_{\mathrm{ho}}$:

$$\Pr_{(P,Q)}\Big(\widehat{\gamma}^2_{k,u}(\mathbf{X}^{\mathrm{te}}, \mathbf{Y}^{\mathrm{te}}) \;\geq\; \gamma^2_k(P, Q) - 8\nu\sqrt{\tfrac{\ln(2/\delta)}{N_{\mathrm{ho}}}}\Big) \;\geq\; 1 - \delta/2.$$

Step 2: concentration of the permutation-null quantile. Conditional on the held-out pool, the permutation-null statistic $\widehat{\gamma}^2_{k,u}(\mathbf{X}_\pi, \mathbf{Y}_\pi)$ is a function of the random labeling π alone. Each pair-coordinate swap in π changes the statistic by at most $8\nu/N_{\mathrm{ho}}$ (the same bounded-differences argument as in Step 1, applied to the relabeling rather than the data). The permutation-null mean equals the unbiased MMD evaluated under random labels, which is 0 in expectation over π. Applying McDiarmid's inequality to the random labeling,

$$\Pr_{\pi}\big(\widehat{\gamma}^2_{k,u}(\mathbf{X}_\pi, \mathbf{Y}_\pi) \geq t\big) \;\leq\; \exp\Big(-\tfrac{N_{\mathrm{ho}} t^2}{32\nu^2}\Big).$$

Setting the right-hand side to $\alpha/2$ and solving for t gives

$$c_\alpha \;\leq\; 8\nu\sqrt{\tfrac{\ln(2/\alpha)}{N_{\mathrm{ho}}}},$$

where c_α is the $(1-\alpha)$-quantile of the permutation-null distribution. The Lehmann–Romano finite-permutation correction of $1/(N_{\mathrm{perm}}+1)$ adds at most $1/(N_{\mathrm{perm}}+1)$ slack [Lehmann and Romano, 2005, Theorem 15.2.1], vanishing for large N_{perm}.

Step 3: combining. Union-bounding Steps 1 and 2 (each with failure probability $\leq \delta/2$ if we set the perm-null tail to $\alpha/2 \to \delta/2$ for the high-probability statement on c_α, or use $\alpha/2$ tied to c_α's

defining quantile and $\delta/2$ for the alternative concentration), we have with probability $\geq 1-\delta$ over $(\mathbf{X}^{\mathrm{te}}, \mathbf{Y}^{\mathrm{te}})$:

$$\widehat{\gamma}^2_{k,u}(\mathbf{X}^{\mathrm{te}}, \mathbf{Y}^{\mathrm{te}}) - c_\alpha \;\geq\; \gamma^2_k(P,Q) - 8\nu\sqrt{\tfrac{\ln(2/\delta)}{N_{\mathrm{ho}}}} - 8\nu\sqrt{\tfrac{\ln(2/\alpha)}{N_{\mathrm{ho}}}}.$$

The right-hand side is non-negative (and the test rejects, $\phi_k = 1$) whenever

$$\gamma^2_k(P,Q) \;\geq\; 8\nu\sqrt{\tfrac{\ln(2/\delta)}{N_{\mathrm{ho}}}} + 8\nu\sqrt{\tfrac{\ln(2/\alpha)}{N_{\mathrm{ho}}}} \;\leq\; 16\nu\sqrt{\tfrac{\ln(2/\delta)+\ln(2/\alpha)}{N_{\mathrm{ho}}}},$$

where the last inequality uses $\sqrt{a}+\sqrt{b} \leq \sqrt{2(a+b)}$. This is the held-out McDiarmid threshold $C\nu\sqrt{(\ln(2/\alpha)+\ln(2/\delta))/N_{\mathrm{ho}}}$ used in Corollary 3.6 below with $C = 16$. ■

G.2 Proof of Corollary 3.6

Setup. Adopt Assumption 3.2, the alternative $H_1 : P \neq Q$, and the sample-split protocol of §3.3: stratify the global samples (sizes m from P and n from Q) into per-class halves of size $m/2$ and $n/2$, forming the training half $\mathcal{D}_{\mathrm{tr}}$ of size $N_{\mathrm{tr}} := m/2 + n/2$ and the held-out half $\mathcal{D}_{\mathrm{te}}$ of size $N_{\mathrm{ho}} := m/2 + n/2$. Throughout this appendix, N denotes N_{tr} (the training-half size); the UCI of Theorem 3.4 is applied to $\mathcal{D}_{\mathrm{tr}}$ with $(m,n) \to (m/2, n/2)$, leaving ρ_* unchanged for balanced splits. Let $\mathcal{H}$ be the class of Corollary 3.6, fixed before the data are observed, let $\widehat{h} \in \mathcal{H}$ be any kernel selected from $\mathcal{D}_{\mathrm{tr}}$, and let $\phi_{k_{\widehat{h}}}$ be the level-α permutation test at $k_{\widehat{h}}$ on $\mathcal{D}_{\mathrm{te}}$ with N_{perm} permutations. Fix any $\delta, \delta' \in (0,1)$; the statement is in §3.3.

Proof. Let $\mathcal{E}_{\mathrm{tr}} := \{\sup_{h\in\mathcal{H}} |\widehat{\gamma}^2_{k,u}(h) - \gamma^2_k(h)| \leq B_N(\delta')\}$, which has probability $\geq 1-\delta'$ over the training half by Theorem 3.4. On $\mathcal{E}_{\mathrm{tr}}$, the uniform bound holds in particular at the selected $\widehat{h} \in \mathcal{H}$, although $\widehat{h}$ depends on the training half:

$$\gamma^2_k(\widehat{h}) \;\geq\; \widehat{\gamma}^2_{k,u}(\widehat{h}) - B_N(\delta') \;=\; \widehat{\gamma}^2_{k,u}(\widehat{h}) - C_1\, G(\mathcal{H}) - C_2\sqrt{\ln(2/\delta')/N}.$$

The LHS of (13) therefore lower-bounds $\gamma^2_k(\widehat{h})$ on $\mathcal{E}_{\mathrm{tr}}$. Since $\widehat{h}$ depends only on the training half, the held-out half is i.i.d. and independent of $\widehat{h}$; conditional on $\widehat{h}$, the held-out McDiarmid lemma of §G.1 above applies with $k = k_{\widehat{h}}$ at confidence δ, controlling the permutation-null quantile by the McDiarmid threshold $C\nu\sqrt{(\ln(2/\alpha)+\ln(2/\delta))/N_{\mathrm{ho}}}$. Whenever (13) holds, $\gamma^2_k(\widehat{h})$ exceeds this McDiarmid threshold and the held-out test rejects with conditional probability $\geq 1-\delta$. By the tower property and $\Pr(\mathcal{E}_{\mathrm{tr}}) \geq 1-\delta'$, the unconditional rejection probability on the threshold event is $\geq (1-\delta')(1-\delta) \geq 1-\delta-\delta'$. ■

Remark (Finite-permutation correction). The held-out permutation test runs with N_{perm} permutations rather than the full $\binom{N_{\mathrm{ho}}}{m_t}$ permutation set, so the empirical $(1-\alpha)$-quantile differs from the exact permutation quantile by an additive slack of at most $1/(N_{\mathrm{perm}}+1)$ [Lehmann and Romano, 2005, Theorem 15.2.1]. This slack can be absorbed by inflating the δ-tail to $\delta + 1/(N_{\mathrm{perm}}+1)$ throughout, with the rejection probability bound becoming $1 - (\delta + 1/(N_{\mathrm{perm}}+1)) - \delta'$. For $N_{\mathrm{perm}} \geq 200$, this slack is below 0.005 and is dominated by the leading δ term.

Scope of the guarantee. Cor. 3.6 requires a non-trivial signal: there must exist some $h \in \mathcal{H}$ with empirical MMD $\widehat{\gamma}^2_{k,u}(h)$ large enough to clear (13), and the selector $\widehat{h}$ must reach such an h. For pairs (P,Q) with $P \neq Q$ but the empirical MMD on $\mathcal{D}_{\mathrm{tr}}$ below the threshold, the corollary does not provide a power guarantee, since the kernel class is too poor to distinguish the alternative at the available sample size.

H Lipschitz proxy bounds: assumptions, proofs, and feasible set

This appendix gives the closed-form bound used as $\widetilde{G}(h)$ in each of the three regimes of §3.3, identifies the feasible set $\mathcal{H}$ to which the bound applies, and compares the assumptions of each cited result with Assumption 3.2 (the only assumption Theorem 3.4 requires).

H.1 Cited results, restated as separate lemmas

We use one classical result: Bartlett's product-of-spectral-norms Lipschitz bound (for the deep regime's $\Pi^* = \prod_j \|W_j\|_2$). We restate it as a self-contained lemma, with the cited reference and an explicit comparison of its hypotheses against our Assumption 3.2.

Lemma H.8 (Composition spectral-norm Lipschitz bound, Bartlett et al., 2017, Thm. 1, eq. (1.1))**.** *Let $h_\theta(x) = W_L\, \phi_{L-1}(W_{L-1}\phi_{L-2}(\cdots \phi_1(W_1 x)\cdots))$ be a feed-forward network with 1-Lipschitz activations ϕ_j satisfying $\phi_j(0) = 0$. Then for every $x, x' \in \mathbb{R}^{d_0}$,*

$$\|h_\theta(x) - h_\theta(x')\|_2 \;\le\; \Big(\prod_{j=1}^{L} \|W_j\|_2\Big)\, \|x - x'\|_2,$$

i.e. h_θ is Π_θ-Lipschitz with $\Pi_\theta := \prod_{j=1}^{L} \|W_j\|_2$.

Hypothesis comparison with Assumption 3.2. Lemma H.8 requires (a) 1-Lipschitz activations and (b) $\phi_j(0) = 0$. (a) is satisfied by ReLU, LeakyReLU, tanh, and the activations used in our experiments. (b) holds for ReLU and LeakyReLU; for activations with non-zero intercept, the bound holds with a constant offset. Neither (a) nor (b) is implied by Assumption 3.2, but both are mild structural conditions on the parametric class of h rather than the kernel k, so they are properly understood as a *specification of the feasible set* $\mathcal{H}_{\mathrm{deep}}$, not as additional hypotheses on the testing problem.

H.2 Proof of Proposition 3.5

The proof uses only the finiteness of the trajectory and the Gaussian maximal inequality; it needs no regime-specific complexity bound and no bound on the inputs.

Proof of Proposition 3.5. Condition on $\mathbf{D}$ and on the trajectory. For each $t \le T$, $Z_t := \frac{1}{N}\sum_{i=1}^{N} \langle \boldsymbol{\xi}_i, h^{(t)}(\mathbf{D}_i)\rangle$ is a centered Gaussian variable with standard deviation $s_t := \|h^{(t)}(\mathbf{D})\|_F/N$, because the $\boldsymbol{\xi}_i$ are independent standard Gaussian vectors. Let $s := \max_{t\le T} s_t$. For every $\lambda > 0$, Jensen's inequality and $\mathbb{E}\, e^{\lambda Z_t} = e^{\lambda^2 s_t^2/2}$ give

$$\mathbb{E}\max_{t\le T} Z_t \;\le\; \tfrac{1}{\lambda}\ln \mathbb{E}\, e^{\lambda \max_t Z_t} \;\le\; \tfrac{1}{\lambda}\ln \sum_{t=0}^{T} \mathbb{E}\, e^{\lambda Z_t} \;\le\; \tfrac{\ln(T+1)}{\lambda} + \tfrac{\lambda s^2}{2}.$$

Choosing $\lambda = \sqrt{2\ln(T+1)}/s$ yields $\widehat{G}(\mathcal{H}_T) = \mathbb{E}\max_t Z_t \le s\sqrt{2\ln(T+1)}$, the first inequality in (9); the case $T = 0$ is immediate since $\mathbb{E}Z_0 = 0$. For the second inequality, if $h^{(t)}$ is $L(h^{(t)})$-Lipschitz with $h^{(t)}(0) = 0$, then $\|h^{(t)}(\mathbf{D}_i)\| \le L(h^{(t)})\,\|\mathbf{D}_i\|$ for every i, so $\|h^{(t)}(\mathbf{D})\|_F \le L(h^{(t)})\,\|\mathbf{D}\|_F \le L^*\,\|\mathbf{D}\|_F$. ■ □

From the proposition to the per-iterate proxy. In the three regimes of (10), $\|h_\sigma(\mathbf{D})\|_F = \|\mathbf{D}\|_F/\sigma$ for $\mathcal{H}_{\mathrm{linear}}$; $\|h_\sigma(\mathbf{D})\|_F = \|\Psi_p(\mathbf{D})\|_F/\sigma$ for $\mathcal{H}_{\mathrm{polynomial}}$, which is finite for any sample without a bound on the inputs; and $\|h_\theta(\mathbf{D})\|_F \le L(h_\theta)\,\|\mathbf{D}\|_F$ with $L(h_\theta) = \prod_j \|W_j\|_2/\sigma$ for $\mathcal{H}_{\mathrm{deep}}$ by Lemma H.8. The deployed criterion evaluates $\widetilde{G}(h^{(t)})$ at the current iterate instead of taking the maximum over the trajectory, so maximizing J_{CP} rewards iterates with a small proxy. The spectral stabilization of App. E is an empirical observation that $L(h^{(t)})$ plateaus along J_{CP} trajectories; no result uses it.

I Rate-tightness on finite kernel grids: comparison to MMDAgg

Proposition I.9 (Massart specialization of the parameterized UCI to finite kernel collections)**.** *Let $\mathcal{K} = \{k_1, \ldots, k_B\}$ be any finite collection of B kernels, allowed to be* multi-type*: each k_j may belong to a different family (e.g., Laplacian, Gaussian, Matérn, polynomial) and may use a different bandwidth or shape parameter. Assume each k_j individually satisfies Assumption 3.2, write $\nu := \max_j \nu_j$ for the common boundedness constant, and let ρ_* be the sample-imbalance ratio of Theorem 3.4.*

Specialization. *View $\mathcal{K}$ as a discrete kernel collection indexed by $j \in \{1, \ldots, B\}$ with trivial feature class $\mathcal{H} = \{\mathrm{id}\}$, so that $G(\mathcal{H}) = 0$. With this specialization the class-level complexity*

proxy $\widetilde{G}(\mathcal{K}) := \max_j L(k_j) \cdot \|\mathbf{D}\|_F/N$ is j-independent, and the CP-MMD selector reduces to the empirical-MMD argmax,

$$\widehat{j} \;:=\; \arg\max_{j\in\{1,\dots,B\}} J_{\mathrm{CP}}(k_j) \;=\; \arg\max_{j\in\{1,\dots,B\}} \widehat{\gamma}^2_{k_j,u}(\mathbf{X},\mathbf{Y}).$$

Conclusion. *For any $\delta \in (0,1)$, with probability at least $1-\delta$ over the pooled sample $\mathbf{D} = \mathbf{X} \sqcup \mathbf{Y}$,*

$$\gamma^2_{k_{\widehat{j}}}(P,Q) \;\geq\; \max_{j=1,\dots,B} \gamma^2_{k_j}(P,Q) \;-\; 2\,C_2\sqrt{\tfrac{\ln(2B/\delta)}{N}}, \tag{17}$$

where $C_2 = 4\nu\rho_*$ is the concentration constant of Theorem 3.4 and $\gamma^2_{k_j}(P,Q)$ denotes the population squared MMD between P and Q under kernel k_j. The leading $\sqrt{\ln B/N}$ term matches the per-grid separation rate $\Theta(\sqrt{\ln(B/\alpha)/N})$ of MMDAgg [Schrab et al., 2023, Theorem 9].

Role of this proposition. This proposition is a *rate-tightness anchor* for our UCI: specialized to a discrete kernel collection $\mathcal{K} = \{k_1,\dots,k_B\}$, our bound recovers the $\Theta(\sqrt{\ln B/N})$ separation rate that Schrab et al. [2023, Theorem 9] establish as minimax-tight on this regime. It is *not a power-advantage claim* over MMDAgg: in the discrete-grid regime, the CP-MMD penalty $\widehat{C}_1\widetilde{G}(\mathcal{K})$ is j-independent and the CP-MMD selector reduces to plain empirical-MMD argmax — the same selection rule MMDAgg performs internally per kernel. What the proposition *does* establish is that our general continuous-class UCI does not pay a rate cost in the special case where MMDAgg's tighter analysis is available, so any rate advantage of CP-MMD on continuous parametric classes (where MMDAgg cannot be applied) is not bought at the price of looseness on the discrete regime.

Remark (MMDAgg's standard configuration). MMDAgg's published configuration, a union of Laplacian and Gaussian families across a dyadic bandwidth grid [Schrab et al., 2023], is the special case obtained by taking $\mathcal{K}$ to be that union.

Proof.

Step 1: Discrete specialization of Theorem 3.4. Apply Theorem 3.4 at each fixed kernel k_j ($j = 1,\dots,B$) with $\mathcal{H} = \{\mathrm{id}\}$ at confidence δ/B. Since $G(\{\mathrm{id}\}) = 0$, only the concentration term contributes, giving with probability $\geq 1-\delta/B$ at each j: $|\widehat{\gamma}^2_{k_j,u} - \gamma^2_{k_j}(P,Q)| \leq C_2\sqrt{\ln(2/(\delta/B))/N}$. Union-bounding over j via Massart's finite-class bound, with probability $\geq 1-\delta$,

$$\sup_{j=1,\dots,B} \left|\widehat{\gamma}^2_{k_j,u} - \gamma^2_{k_j}(P,Q)\right| \;\leq\; C_2\sqrt{\tfrac{\ln(2B/\delta)}{N}} \;=:\; \Delta_{B,\delta}, \tag{18}$$

where $C_2 = 4\nu\rho_*$ inherits from Theorem 3.4's concentration constant.

Step 2: CP-MMD selector tracks the population-best kernel. Let $j^* := \arg\max_{j\in\{1,\dots,B\}} \gamma^2_{k_j}(P,Q)$ denote the population-best kernel index in $\mathcal{K}$. Apply the UCI event (18) twice, once at the selector $\widehat{j}$ and once at j^*, and use the definition $\widehat{j} = \arg\max_j \widehat{\gamma}^2_{k_j,u}$:

$$\begin{aligned}
\gamma^2_{k_{\widehat{j}}}(P,Q) \;&\geq\; \widehat{\gamma}^2_{k_{\widehat{j}},u} - \Delta_{B,\delta} && \text{(UCI (18) at } j=\widehat{j}) \\
&\geq\; \widehat{\gamma}^2_{k_{j^*},u} - \Delta_{B,\delta} && \text{(definition of } \widehat{j} \text{ as empirical-MMD argmax)} \\
&\geq\; \gamma^2_{k_{j^*}}(P,Q) - 2\Delta_{B,\delta} && \text{(UCI (18) at } j=j^*).
\end{aligned}$$

Since $\gamma^2_{k_{j^*}}(P,Q) = \max_j \gamma^2_{k_j}(P,Q)$ by definition of j^*, this rearranges to (17). The leading $\sqrt{\ln B/N}$ term in $\Delta_{B,\delta}$ matches MMDAgg's per-grid separation rate [Schrab et al., 2023, Theorem 9]: substituting $\delta \mapsto \alpha$ recovers $\Theta(\sqrt{\ln(B/\alpha)/N})$.

Step 3: Extension to infinite $\mathcal{H}$. For a class $\mathcal{H}$ whose induced data class $\mathcal{H}(\mathbf{D})$ has finite empirical-L^2 covering number $N(\eta;\mathcal{H}(\mathbf{D}),\|\cdot\|_{L^2(\hat{P}_N)}) < \infty$ for every $\eta > 0$, Dudley's entropy integral [Wainwright, 2019, Theorem 5.22] bounds the Gaussian complexity by

$$\tfrac{1}{N}\,G(\mathcal{H}(\mathbf{D})) \;\leq\; \tfrac{C}{\sqrt{N}} \int_0^{\nu} \sqrt{\ln N(\eta;\,\mathcal{H}(\mathbf{D}),\,\|\cdot\|_{L^2(\hat{P}_N)})}\; d\eta,$$

where the upper limit ν is the kernel boundedness constant (which upper-bounds the diameter of $\mathcal{H}(\mathbf{D})$ in $L^2(\hat{P}_N)$ via $\|f\|_{L^2(\hat{P}_N)} \leq \|f\|_\infty \leq \nu$). The bound is finite under standard entropy conditions (e.g., VC-type or polynomial entropy). The CP-MMD UCI bound therefore extends to infinite classes with no $|\mathcal{K}|$-dependent multiplicity correction. By contrast, MMDAgg's weighted union bound assigns a multiplicity correction α/B to each kernel; as $B \to \infty$, the per-kernel threshold $\tau_{n,b}(\alpha/B) = O(\nu\sqrt{\ln(B/\alpha)/N}) \to \infty$, and the test becomes trivially conservative. CP-MMD's complexity penalty admits a continuous limit; MMDAgg's does not. ■

A more comprehensive rate-comparison analysis is left to future work.

Corollary I.10 (Trajectory-adaptive complexity penalty and deployment cost)**.** (Trajectory-adaptive penalty.) *For the CP-MMD selector deployed via the trajectory class $\mathcal{H}_T \subseteq \mathcal{H}$, the complexity proxy satisfies $\widetilde{G}(h^{(T)}) \leq L^* \|\mathbf{D}\|_F/N$ for any envelope $L^* \geq \max_{t\leq T} L(h^{(t)})$ of the realized trajectory, whereas MMDAgg pays a fixed $\nu\sqrt{2\ln B/N}$ per kernel uniformly over the grid $\mathcal{K}$ regardless of which kernels carry signal.*

(Per-test cost.) *At deployment, the per-test cost of CP-MMD is $O(N_{\mathrm{perm}} \cdot N^2)$, independent of search-space size. MMDAgg pays $O(B \cdot N_{\mathrm{perm}} \cdot N^2)$, linear in the grid size B. MMD-FUSE pays $O((B + N_{\mathrm{perm}}) \cdot N^2)$ via shared distance matrices. Storage is $O(N^2)$ for CP-MMD versus $O(B \cdot N^2)$ for MMDAgg.*

Proof.

Trajectory-adaptive penalty. Proposition 3.5 bounds the trajectory complexity by $\sqrt{2\ln(T+1)}\, L^* \|\mathbf{D}\|_F/N$ for any envelope $L^* \geq \max_{t\leq T} L(h^{(t)})$, and the deployed proxy $\widetilde{G}(h^{(t)})$ evaluates the envelope at the current iterate, so $\widetilde{G}(h^{(T)}) \leq L^* \|\mathbf{D}\|_F/N$. Because the supremum over $\mathcal{H}_T \subseteq \mathcal{H}$ is at most the supremum over $\mathcal{H}$, the trajectory complexity is at most the empirical complexity of the whole class, so the trajectory penalty is at most the global one. MMDAgg's per-kernel threshold $\tau_{n,b}(\alpha/B) = \Theta(\nu\sqrt{\ln(B/\alpha)/N})$ is a function of the grid size B and the kernel collection only, while CP-MMD's penalty $\widetilde{G}(h^{(t)})$ depends on the realized iterate; the two are different design choices for the discrete-grid vs. continuous-class settings, not directly comparable as power claims.

Per-test cost. CP-MMD's deployed test runs the level-α permutation test at the single learned kernel $k_{\widehat{h}}$, requiring N_{perm} permutations of an $N \times N$ distance matrix at $O(N^2)$ per permutation, total $O(N_{\mathrm{perm}} \cdot N^2)$. MMDAgg evaluates all B candidate kernels for each of N_{perm} permutations, multiplying the cost by B. MMD-FUSE precomputes the $N \times N$ pairwise-distance matrix once across all B kernels (cost $O(B \cdot N^2)$) and reuses it across N_{perm} permutations (cost $O(N_{\mathrm{perm}} \cdot N^2)$), totaling $O((B + N_{\mathrm{perm}}) \cdot N^2)$ [Biggs et al., 2023]. ■

J Deep ablation: criterion × kernel ansatz factorial

To isolate the criterion's effect from the kernel-structure effect in Table 2, we run a 3×2 factorial. The body's deep results compare three criteria on the same *pure-feature kernel* $k(\phi_\theta(x), \phi_\theta(y))$ (the deep composite k_h of §3.1 with $h = \phi_\theta$ and no raw-input mixing); here we cross each criterion (CP-MMD, Liu, Plain) with each of two kernel structures, the pure-feature regime defined above (in scope) and Liu's published ϵ-mixture $k_\epsilon = [(1-\epsilon)\kappa_{\sigma_\kappa}(\phi_w, \phi_w) + \epsilon]\, q_{\sigma_q}$ (out of scope: ϵ varies inside the kernel rather than inside the feature map). Same MLP architecture as Table 2, mean shift $\Delta = 0.5$, 100 reps, five (d, n) cells.

J.1 Heuristic adaptation of J_{CP} to Liu's ϵ-mixture

Why an adaptation is needed. Liu's ϵ-mixture kernel

$$k_\epsilon(x, y) \;=\; \big[(1-\epsilon)\,\kappa_{\sigma_\kappa}(\phi_w(x), \phi_w(y)) + \epsilon\big]\, q_{\sigma_q}(x, y),$$

with parameters $(\phi_w, \sigma_\kappa, \sigma_q, \epsilon)$, has ϵ varying *inside* the kernel rather than inside a feature map, so k_ϵ is not of the form $k(h(x), h(y))$ required by Definition 3.3. Consequently, Theorem 3.4 (UCI) and Proposition 3.5 (Lipschitz proxy) do not apply directly to this kernel class.

Decomposition of k_ϵ into two structural sub-kernels. Expanding the multiplication in k_ϵ,

$$k_\epsilon(x,y) \;=\; (1-\epsilon)\;\underbrace{\kappa_{\sigma_\kappa}(\phi_w(x),\phi_w(y))\,q_{\sigma_q}(x,y)}_{\text{product of feature kernel and raw envelope}} \;+\epsilon\;\underbrace{q_{\sigma_q}(x,y)}_{\text{raw envelope}},$$

which exhibits two structural components blended by ϵ. Neither the product $\kappa \cdot q$ nor k_ϵ as a whole is in k_h form, so Proposition 3.5 does not cover them directly.

Heuristic complexity proxy. We construct $\widetilde{G}_\epsilon$ by *dropping* the q_{σ_q} envelope on the first term (so it becomes a pure feature kernel that Proposition 3.5 *does* cover) and then taking a convex combination of the two resulting in-scope proxies:

$$\widetilde{G}_\epsilon(\phi_w,\sigma_\kappa,\sigma_q,\epsilon) \;:=\; (1-\epsilon)\,\widetilde{G}_{\mathrm{feat}}(\phi_w,\sigma_\kappa) \;+\; \epsilon\,\widetilde{G}_{\mathrm{raw}}(\sigma_q),$$

where each component is the spectral proxy of an explicit in-scope composite sub-kernel:

- $\widetilde{G}_{\mathrm{feat}}(\phi_w,\sigma_\kappa) = \frac{\Pi}{\sigma_\kappa}\cdot\frac{\|\mathbf{D}\|_F}{N}$ is the *deep-regime* proxy from (10) applied to the feature sub-kernel $\kappa_{\sigma_\kappa}(\phi_w(\cdot),\phi_w(\cdot))$, viewed as the composite kernel $k(h(x),h(y))$ with base kernel $k=\kappa_1$ (unit-bandwidth Gaussian) and feature map $h(x) := \phi_w(x)/\sigma_\kappa$. The Lipschitz constant of this h on the trajectory class is $L(h)=\Pi/\sigma_\kappa$ with $\Pi := \prod_{j=1}^{L}\|W_j\|_2$ (Lemma H.8). $\widetilde{G}_{\mathrm{feat}}$ is therefore the spectral proxy of $\kappa(\phi_w(\cdot),\phi_w(\cdot))$ *with the q_{σ_q} envelope dropped*.
- $\widetilde{G}_{\mathrm{raw}}(\sigma_q) = \frac{1}{\sigma_q}\cdot\frac{\|\mathbf{D}\|_F}{N}$ is the *linear-regime* proxy from (10) applied to the raw sub-kernel $q_{\sigma_q}(x,y)$, viewed as the composite kernel $k(h(x),h(y))$ with base kernel $k=q_1$ and linear feature map $h(x):=x/\sigma_q$. The Lipschitz constant is $L(h)=1/\sigma_q$.

What this proxy is and is not. The heuristic $\widetilde{G}_\epsilon$ matches a Lipschitz-norm-ball bound on the *additive surrogate* $(1-\epsilon)\,\kappa(\phi_w(x),\phi_w(y)) + \epsilon\,q(x,y)$ — not on the actual mixture k_ϵ. Two simplifications are made: (i) the q_{σ_q} envelope multiplying κ in the actual kernel is dropped at the feature term (the actual product would contribute an additional Lipschitz factor of order $1/\sigma_q$ that the heuristic ignores at the feature term); (ii) the convex combination in ϵ mirrors the kernel's own $(1-\epsilon)$-vs-ϵ weighting at the structural level but does not derive from a tight Lipschitz analysis of k_ϵ as a whole. Both simplifications are absorbed into the calibrated constant $\widehat{C}_1$ at deployment.

Adapted criterion. The mixture-aware CP-MMD criterion is

$$J_{\mathrm{CP}}(\phi_w,\sigma_\kappa,\sigma_q,\epsilon) \;:=\; \widehat{\gamma}^2_{k_\epsilon,u}(\phi_w,\sigma_\kappa,\sigma_q,\epsilon) \;-\; \widehat{C}_1\,\widetilde{G}_\epsilon(\phi_w,\sigma_\kappa,\sigma_q,\epsilon),$$

optimized jointly over $(\phi_w,\sigma_\kappa,\sigma_q,\epsilon)$ by gradient ascent (with ϵ parameterized as the sigmoid of an unconstrained scalar to keep $\epsilon\in(0,1)$).

Calibration of $\widehat{C}_1$. We reuse the $\widehat{C}_1$ calibrated on the in-scope pure-feature kernel form (App. F) for the mixture's heuristic $\widetilde{G}_\epsilon$. This reuse is itself a heuristic: the calibration target $\widehat{C}_1^*$ (15) depends on the kernel class, and we have no formal stability result across the mixture's ϵ axis. Empirical insurance comes from App. F.6's seven-OOM ablation, which shows test power is insensitive to $\widehat{C}_1$ at the order-of-magnitude level.

What is and is not claimed. Because Liu's ϵ-mixture lies outside Definition 3.3, the formal claims of Theorem 3.4, Proposition 3.5, and Cor. 3.6 do *not* apply to this kernel class — even with the heuristic adjustment above. What survives:

- *Type-I validity is exact*: sample-split + held-out permutation test depends only on the held-out half being independent of $\widehat{h}$, regardless of which kernel class $\widehat{h}$ is selected from. So the deployed test controls Type-I at level α on Liu's mixture as well as on any in-scope class.
- *Power and the criterion's advantage are empirical*: the factorial below tests whether the criterion advantage of J_{CP} over Liu's ratio criterion J_{Liu} — established in scope by §4 — persists on this out-of-scope kernel class.

Criterion	Kernel form	d=2 n=200	d=20 n=200	d=50 n=100	d=100 n=100	d=20 n=50
CP-MMD	pure-feature	0.81	**1.00**	**1.00**	**1.00**	**1.00**
CP-MMD	Liu mixture	**0.98**	0.97	0.16	0.01	0.54
Liu	pure-feature	0.44	0.49	0.20	0.36	0.09
Liu	Liu mixture (published)	0.93	0.90	0.02	0.04	0.53
Plain	pure-feature	0.83	1.00	0.54	0.36	0.17
Plain	Liu mixture	0.97	0.92	0.23	0.00	0.66

Table 7: Power at mean shift $\Delta = 0.5$, 100 reps, six methods $\times$ five (d, n) cells. Reproduced by `run_liu_mixture_ablation.py`, the Liu rows by `run_liu_published.py`, both in `simulations/`; the CSV is `results/liu_mixture_ablation/results_final.csv`.

J.2 Factorial results

Three patterns. *(a) The criterion advantage is structural across kernel ansätze, in scope or out*: within each kernel column, that is, comparing selection criteria at fixed kernel structure, CP-MMD matches or exceeds Liu's ratio criterion in 9 of 10 cells, including on Liu's own mixture kernel that our theory does not formally cover; the exception is the mixture kernel at $(d, n) = (100, 100)$, where every criterion has power at most 0.04. The full ordering CP-MMD $\geq$ Plain $\geq$ Liu holds in 6 of the 10 cells. *(b) The pure-feature kernel structure dominates the mixture form at high d*: regardless of criterion, the mixture form's high-d degradation (e.g., $0.98 \to 0.01$ for CP-MMD as d grows from 2 to 100) reflects the curse of dimensionality on the raw-input envelope q, a property of the *kernel structure* independent of the *criterion* — when both kernel structures are in play, the pure-feature form is the safer default. *(c) Within the mixture column, criterion design still matters*: even where the kernel structure is suboptimal, CP-MMD's penalty-based selection beats Liu's ratio-based selection on the cells where the mixture form retains usable signal ($d = 2$ and $d = 20, n = 200$: $0.98/0.93$, $0.97/0.90$).

K Real-data results

We complement the Higgs benchmark of §4.3 (where $n \in \{200, 500\}$) with two additional real-data domains, to verify CP-MMD's behavior outside the synthetic-Gaussian regime:

- **Tennessee Eastman Process Fault 3/9/15** (d=52, n=200): at this small-sample industrial-monitoring regime no learned-kernel approach detects the shifts (CP-MMD, Liu, Plain stay at power ≤ 0.12 across the three faults); only MMDAgg's bandwidth aggregation isolates Fault 15 (power 0.78), while Faults 3/9 remain at most 0.18 for every method (`results/tep/tep_results.csv`).
- **CIFAR-10 vs. CIFAR-10.1** (n=200 image samples after PCA-500 preprocessing): every method we evaluate (CP-MMD, Liu, Plain, Median, MMDAgg, MMD-FUSE) saturates at power 1.00, so the dataset is uninformative for ranking learned-kernel criteria at this sample size (`results/cifar_pca500/cifar_results.csv`).

On Higgs (Fig. 3(a)), every method uses the 28 features standardized over the first 10^5 events of the dataset. The median heuristic and MMDAgg test on all n samples per class, while the CP-MMD ansätze select on one half and test on the other. At $n = 500$, CP-MMD-ARD reaches 1.00, as do the median heuristic and MMDAgg, and CP-MMD-linear (0.94) and CP-MMD-deep (0.92) come close; at $n = 200$ the split costs more, and the best ansatz, polynomial$_3$ (0.62), trails the median heuristic (0.78) and MMDAgg (0.88). The polynomial ansätze use a reduced lift of the 28 features in place of Ψ_2 and Ψ_3 of (6): the features themselves, their products with the 20 features of largest raw variance, and, for degree 3, the triple products of distinct features among those 20, which gives 568 and 1708 coordinates instead of 434 and 4494. CP-MMD-deep outperforms Liu and Plain on the same MLP at both sample sizes.

ARD ansatz on Higgs. ARD gives each standardized feature its own bandwidth, $h(x) = (x_1/\sigma_1, \ldots, x_{28}/\sigma_{28})$, with the Gaussian base kernel. As a one-layer diagonal network it is penalized

by the deep proxy of (10), $L(h)\,\|\mathbf{D}\|_F/N$ with $L(h) = \max_j \sigma_j^{-1}$, which bounds $\|h(\mathbf{D})\|_F/N$. The latter decreases when the bandwidths single out one feature: penalizing it directly gave calibration ratios about ten times those of the linear ansatz under H_0, because the plain-MMD maximizer fits noise in one or two features, and the selected kernel was nearly flat. Calibration follows App. F.2 for gradient-trained classes with $n_{\mathrm{cal}} = 30$: on each null permutation of the training half, Adam ascent on $\widehat{\gamma}^2_{k,u}$ over $\log(1/\sigma_j)$, 200 steps at learning rate 0.05 from the best bandwidth of the linear grid, gives the plain-MMD maximizer. Selection runs the same ascent on J_{CP} and keeps the iterate with the largest J_{CP}. The Type-I error of ARD is 0.02 at $n = 200$ and 0.00 at $n = 500$. At $n = 500$ its power is 1.00, and in each of the five logged repetitions its smallest bandwidth falls on m_{bb}, the invariant mass with the largest two-sample Kolmogorov–Smirnov statistic between signal and background among the 28 features, 0.20. At $n = 200$, 100 training samples per class do not determine 28 bandwidths reliably, and its power is 0.46.

L Limitations and extensions

Three limitations of the present work merit explicit discussion.

Calibration scope of $\widehat{C}_1$. The high-quantile permutation calibration (App. F) is reliable only at moderate MLP width ($w \in [50, 200]$); at wider widths it degenerates with $\widehat{C}_1 \to 0$, since the trajectory-quantile estimator absorbs the inflated complexity of overparameterized networks into the data variability rather than the penalty. Empirical spectral stabilization (App. E, Fig. 4(d)) keeps training bounded in this regime, but the calibration value transferred from w=200 no longer reflects the effective complexity of the wider model. In practice, calibration should be performed at a representative width within $[50, 200]$ and reused across architectures of comparable scale; for substantially larger networks, the calibration step should be re-run.

Looseness of the complexity proxy $\widetilde{G}$. The Lipschitz proxy $\widetilde{G}$ (Prop. 3.5) is a worst-case Gaussian-complexity surrogate for the spectral-norm complexity that appears in Theorem 3.4. The gap of at least two orders of magnitude between the worst-case constant $C_1 \approx 30$ and the calibrated constant, which satisfies $\sqrt{N}\,\widehat{C}_1 \leq 0.15$ on the scale of (10), is the price of this approximation: the surrogate is correct in shape but conservative by a factor that the calibration absorbs empirically rather than closes in theory. Tightening this gap, via local Rademacher complexity, distribution-dependent envelopes, or PAC-Bayes refinements, is a natural direction for a journal-length theoretical extension.

Sensitivity to the bandwidth search interval. The proxy $\widetilde{G} \propto 1/\sigma$ overstates the complexity of bandwidths at which the kernel is close to zero on the data. In the polynomial experiment of §4.2, the median ℓ_1 distance between degree-4 feature vectors is about 10^3, so the empirical MMD peaks near $\sigma \approx 250$, above the search interval $[0.01, 100]$, and CP-MMD selects $\sigma = 100$. Extending the interval to $[0.01, 10^4]$ raises the calibrated $\widehat{C}_1$ from about 0.009 to 0.2–0.4, because the largest calibration ratios then occur at large bandwidths, and the heavier penalty selects over-smoothed kernels: power at df=20 falls from 0.60 to 0.14. The search interval therefore acts as part of the model specification, and a proxy that accounts for the data scale is a natural refinement.

Extensions. The complexity-penalized criterion carries over to other kernel statistics, including HSIC [Gretton et al., 2007] and the kernel Stein discrepancy [Liu et al., 2016, Chwialkowski et al., 2016]. A scale-aware proxy, one that does not overstate the complexity of bandwidths at which the kernel is close to zero on the data, would address the third limitation.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: §1.2 states three contributions: a unified grid-free criterion J_{CP} (§3.1–§3.3), a finite-sample test-power lower bound for the post-selection kernel (Theorem 3.4, Cor. 3.6), and a per-test deployment cost independent of the candidate-class size (§4.4). One additional theoretical result supports these contributions: rate-tightness on finite kernel grids (Proposition I.9). Empirical validation appears in §4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: App. L discusses the principal limitations in detail, with a forward pointer in §5. *Sensitivity to the bandwidth search interval:* on the kurtosis shift, widening the search interval of the polynomial ansatz raises $\widehat{C}_1$ and lowers its power at $\mathrm{df} = 20$ from 0.60 to 0.14. *Calibration scope of $\widehat{C}_1$:* the high-quantile permutation calibration is reliable at moderate MLP width ($w \in [50, 200]$) and degenerates at wider widths, with empirical spectral stabilization (App. E, Fig. 4(d)) still bounding training; calibration should be performed at a representative width and reused across architectures of comparable scale. *Looseness of the complexity proxy $\widetilde{G}$:* the worst-case $C_1 \approx 30$ exceeds the calibrated constant by at least two orders of magnitude on the scale of Eq. (10); the surrogate is correct in shape but conservative, with the gap absorbed empirically rather than closed in theory.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All formal results state their assumptions (Assumption 3.2). Full proofs appear in the supplementary material: Lemma 2.1 in App. B, Theorem 3.4 (UCI) in App. D, Proposition 3.5 (envelope bound) in App. H.2, Corollary 3.6 (post-selection power) in App. G, and Proposition I.9 (rate-tightness on finite kernel grids) and Corollary I.10 in App. I. The variance-collapse phenomenon of Liu's ratio criterion under H_0 (App. C) and the spectral-norm stabilization under J_{CP} ascent (App. E) are documented as empirical observations rather than formal theorems. The Type-I validity of the held-out permutation test is classical [Lehmann and Romano, 2005, Theorem 15.2.1] and used without restating.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: §4 specifies all architectures, optimizers, learning rates, sample sizes, and repetition counts. The released code, described in the next item, provides (a) the producing scripts for every table and figure (`simulations/run_*.py`), (b) reference CSVs under `results/` with the numbers in Tables 2, 3, 6, and 7 and Figs. 1, 2, 3, and 5, while Tables 4–5 are regenerated by their scripts and Fig. 4 is written directly by its script, and (c) a `README.md` mapping each paper artifact to its producing script. Tables 3 and 7 cite the producing script in their captions, and the `README.md` maps Table 2 to `simulations/run_deep_summary.py`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may

be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Source code (20 experiment drivers, 3 core modules, the figure-rendering script) and reference CSVs are released under the MIT license at `https://github.com/Yijin911/cpmmd-code/tree/main/neurips2026`, with a `README.md` mapping each paper artifact to its producing script and CSV. Real datasets are publicly available and downloaded by the scripts on first run: CIFAR-10/10.1 (`torchvision + modestyachts/CIFAR-10.1`), UCI Higgs (`https://archive.ics.uci.edu/ml/datasets/HIGGS`); the Tennessee Eastman experiment uses our own synthetic Gaussian-shift implementation (not the proprietary Downs–Vogel simulator).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: §4 provides network architectures, optimizer settings, sample sizes, repetition counts, and significance levels for all experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Standard errors are reported in Table 2 (200 reps per cell, binomial SE) and the Higgs results in Fig. 3(a) use 50 reps per cell, so their binomial SE is at most 0.07; Table 3 (App. C) reports means over 10 seeds per width; Table 7 (App. J) uses 100 reps per cell with implicit binomial SE. The binomial SE convention $\mathrm{SE} = \sqrt{p(1-p)/n_{\mathrm{rep}}}$ is used for the shaded bands in Figs. 1 and 2. Type-I error at $\alpha = 0.05$ is verified to lie within ± 2 binomial SE of nominal in every cell of Table 2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: §4.4 reports per-test deployment cost and selection cost. CPU-only suffices for the linear and polynomial pillars ($\sim$10–30 minutes each); the deep-kernel and class-richness experiments use a single GPU (NVIDIA A100 or A10 in our runs, $\sim$1–3 hours per driver); calibration sweeps run in $\sim$5–10 minutes on GPU; the deployment-cost timings (Figure 3(b))

run on CPU with JAX JIT compilation. Per-script runtime estimates are provided in the supplementary `README.md`.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research is methodological (statistical testing) with no ethical concerns.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational statistical methodology. It improves two-sample testing, a generic statistical tool with no direct negative societal implications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No models or datasets with misuse risk are released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: CIFAR-10 (Krizhevsky, 2009) is publicly distributed by the University of Toronto for academic research; CIFAR-10.1 (Recht et al., 2018) is released under the MIT License via the `modestyachts/CIFAR-10.1` GitHub repository; the UCI Higgs dataset [Baldi et al., 2014] is available under CC BY 4.0 from the UCI Machine Learning Repository. The Tennessee Eastman process is referenced via our own synthetic Gaussian-shift implementation; we do not redistribute the proprietary Downs–Vogel simulator.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The only new asset is the released code at `https://github.com/Yijin911/cpmmd-code/tree/main/neurips2026`. Its `README.md` maps each figure and table to its producing script and lists the dependencies, including pinned versions of the MMDAgg and MMD-FUSE baselines; the license is MIT. No new datasets or models are released.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subjects or crowdsourcing involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were used to assist with prose editing and proofreading of the manuscript and with experimental code (e.g., refactoring, helper functions, plotting scripts). No LLM is part of the core methodology or the core theoretical analysis.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.