Regularized f-Divergence Kernel Tests

Mónica Ribero* Google Research

Antonin Schrab* University of Cambridge

Arthur Gretton Google DeepMind

Abstract

We propose a framework to construct practical kernel-based two-sample tests from the family of f-divergences. The test statistic is computed from the witness function of a regularized variational representation of the divergence, which we estimate using kernel methods. The proposed test is adaptive over hyperparameters such as the kernel bandwidth and the regularization parameter. We provide theoretical guarantees for statistical test power across our family of f-divergence estimates. While our test covers a variety of f-divergences, we bring particular focus to the Hockey-Stick divergence, motivated by its applications to differential privacy auditing and machine unlearning evaluation. For two-sample testing, experiments demonstrate that different f-divergences are sensitive to different localized differences, illustrating the importance of leveraging diverse statistics. For machine unlearning, we propose a relative test that distinguishes true unlearning failures from safe distributional variations.

1 INTRODUCTION

Two-sample testing is a fundamental task in statistics, with broad applications in machine learning (e.g., MMD, Gretton et al., 2012). These tests can confidently determine if two sets of observations come from different underlying distributions, applicable to diverse contexts, from auditing the privacy of machine learning models (Kong et al., 2024) and verifying the efficacy of machine unlearning techniques (Triantafillou et al., 2024), to validating models in the physical sciences (D'agnolo and Wulzer, 2019; Grosso and Letizia, 2025).

Two-sample tests require the choice of a distance or divergence to quantify the difference between the distributions, and different applications require different notions of distance. Some tests are concerned with smooth variations, for which the Maximum Mean Discrepancy (MMD, Gretton et al., 2012) is a suitable choice (Schrab et al., 2023). Others, like outlier detection in physical models, require sensitivity to subtle and/or localized differences that MMD might miss. Other applications, like privacy or unlearning, require divergences parameterized by a budget ε to define an acceptable degree of separation. This highlights the challenge that no single two-sample test is universally optimal or suitable across all scenarios.

Contributions. We introduce a general framework for constructing two-sample tests from the family of f-divergences, unifying specific instances that have been studied by Nguyen et al. (2010); Harchaoui et al. (2007); Glaser et al. (2021); Grosso and Letizia (2025). This framework builds on the variational formulation for f-divergences. A key challenge is that the optimizer of this problem, also known as the witness function, is often non-smooth and hard to learn efficiently. However, the witness function can always be expressed as a functional of the likelihood ratio. We leverage kernel-based methods to estimate the ratio and ℓ_2-regularization, trading off some accuracy for tractability (Section 2). While more specialized estimators may exist in the literature for certain pairs of distributions and divergences, the novel proposed framework provides a general and practical approach for constructing a test applicable to any f-divergence.

On the technical side, we instantiate a likelihood ratio estimator proposed by Chen et al. (2024a). We prove in Lemma 2.1 that, under smoothness assumptions (i.e., $\mu_P - \mu_Q \in \mathrm{Ran}(\Sigma_Q^\theta)$ for some $\theta \in (1, 2]$ (see Appendix A.2) the estimator converges to the true ratio at a rate of $N^{-\frac{\theta-1}{2(\theta+1)}}$ with high probability, where N represents the minimum number of samples from distributions P and Q. Subsequently, we leverage the

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

*Authors contributed equally to this work.

variational formulation of f-divergences, expressing the corresponding witness function as a functional of the likelihood ratio. This yields an f-divergence estimator which, as proved in Lemma 2.2, converges to the true divergence value at a rate of $N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2}$ with high probability, where the data has been split between N samples (used for estimating the ratio) and $\widetilde{N}$ samples (used for estimating the f-divergence given the estimated ratio).

We analyze the properties of the permutation hypothesis tests using these likelihood-based f-divergence statistics. First, these naturally control the type I error at the desired level $\alpha \in (0,1)$ for any sample size (Romano and Wolf, 2005, Lemma 1). Then, we prove that in Theorem 3.1 these tests are consistent in that they are asymptotically powerful, that is, their type II error probability converges to 0 as the sample size tends to infinity. Moreover, for the finite-sample regime, we establish a non-asymptotic power guarantee: the test controls the type II error by $\beta \in (0,1)$ for any pair of distributions separated with respect to the f-divergence D_f as

$$D_f(P\|Q) \gtrsim \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-\frac{1}{2}}\right)\sqrt{\ln\left(\frac{1}{\min\{\alpha,\beta\}}\right)}.$$

This reinforces the idea that tests based on different f-divergences capture different types of departures from the null.

Our experimental results corroborate that no single test consistently outperforms the others, across a variety of settings. However, we find that in most cases, at least one of the considered tests exhibits strong performance. We show that the recently developed aggregation method of Schrab et al. (2023) can aggregate these different f-divergence tests into a single one, which we call f-Agg, and which can powefully detect a wide range of alternatives (see details in Algorithm 2 in Appendix A.5). Our results provide valuable intuition on which statistic is best suited for a given task. Furthermore, we propose a new, more precise evaluation framework for machine unlearning based on our theoretical and empirical findings, detailed below.

Of specific interest is the test based on the Hockey-Stick divergence, motivated by its relevance to privacy and unlearning. We demonstrate how the Hockey-Stick divergence can be used to audit pure differential privacy and, more broadly, to evaluate machine unlearning algorithms, a task for which finding appropriate evaluation metrics has been a significant challenge. We then address a flaw in current unlearning definitions: absolute distance tests, comparing an unlearned model distribution to a retrained one without the removed data, are insufficient. Such tests can fail even if unlearning was successful, as they cannot distinguish between a genuine failure to forget and benign variations introduced by the unlearning algorithm. Further, recent work by Yu et al. (2025) theoretically and empirically shows that equivalence to a retrained model is impossible for gradient-based unlearning algorithms that are oblivious to the learning trajectory. We resolve this ambiguity by proposing a relative distance test, which measures whether the unlearned model is distributionally closer to a retrained model or to the original, compromised one.

Related Work. A popular approach to non-parametric two-sample testing relies on the MMD (Gretton et al., 2012), which measures distances between distributions in a Reproducing Kernel Hilbert Space (RKHS), see Appendix A.3 for details. The power of MMD tests is highly sensitive to the choice of kernel and bandwidth, which has motivated various lines of work including learning the kernel (Sutherland et al., 2017; Liu et al., 2020), constructing adaptive tests by combining multiple kernels' results (Schrab et al., 2023; Biggs et al., 2023), and leveraging AutoML methods (Kübler et al., 2022). Separately, f-divergences are widely used in machine learning and have been recently used to train machine learning models via the variational forms of these divergences (Nowozin et al., 2016; Mescheder et al., 2017; Arbel et al., 2020), as well as for model evaluation (Pillutla et al., 2023), and to define gradient flows on probability measures (Glaser et al., 2021; Chen et al., 2024b). Application-specific two-sample tests have also been developed and tested, such as the tests based on the Hockey-Stick divergence for privacy (Kong et al., 2024) and for unlearning (Triantafillou et al., 2024), as well as deep-learning and KL-based tests for new physics (D'agnolo and Wulzer, 2019; Grosso and Letizia, 2025).

The estimation of f-divergences is a well-studied problem with several established approaches. One popular family of non-parametric techniques relies on nearest neighbor methods (Wang et al., 2009; Singh and Póczos, 2016). Other non-parametric approaches use space partitioning estimates, however, these tend to perform poorly in high dimensions (Biau and Gyorfi, 2005). A separate line of work, relevant to ours, focuses on estimating the likelihood ratio directly through its variational formulation. Within this framework, estimators have been proposed over various function classes, including kernel-based classes using convex relaxations (Nguyen et al., 2010), generalized linear models approximations Sugiyama et al. (2008); Kanamori et al. (2009, 2012) or direct optimization (Chen et al., 2024a), and neural network classes (Nowozin et al., 2016).

This work unifies these threads building upon the vari-

ational representation of f-divergences but adapts it for hypothesis testing, developing and analyzing a practical framework for constructing a family of tractable two-sample tests with applications to diverse settings.

The paper is organized as follows. In Section 2 we introduce a method for consistently estimating f-divergences. We then provide asymptotic and non-asymptotic power guarantees for the permutation tests based on these estimates in Section 3. In Section 4, we instantiate this framework with the Hockey-Stick divergence. Finally, we present numerical results in Section 6.

Notation. We write $a \lesssim b$ if there exists some constant $C > 0$ (in our setting independent of the sample sizes) such that $a \leq Cb$. We write $a \asymp b$ if $a \lesssim b$ and $a \gtrsim b$. For a distribution Q, we let μ_Q denote the kernel mean embedding, and let Σ_Q denote the covariance operator (see details in Appendix A.2). We use bold notation to denote vectors of random variables, e.g. $\mathbf{X} = (X_1, ..., X_n)$.

2 ESTIMATING ANY f-DIVERGENCE CONSISTENTLY

For probability distributions which are absolutely continuous $P \ll Q$ on a space $\mathcal{Z}$, an f-divergence (Csiszár, 1967) is defined as

$$D_f(P \parallel Q) = \int_{\mathcal{Z}} f\left(\frac{dP}{dQ}\right) dQ, \tag{1}$$

for some convex function $f\colon (0, \infty) \to \mathbb{R}$ which satisfies $f(1) = 0$, and is lower semi-continuous implying $f(0) = \lim_{t\to 0^+} f(t)$. The variational representation (See, e.g., Theorem 7.26 in Polyanskiy and Wu, 2022) of the f-divergence is

$$D_f(P \parallel Q) = \sup_{g:\mathcal{Z}\to\mathbb{R}} \mathbb{E}_P[g(X)] - \mathbb{E}_Q[f^*(g(Y))], \tag{2}$$

where $f^*(u) = \sup_{t\in\mathbb{R}}\{ut - f(t)\}$ is the convex conjugate of f, and where the supremum is taken over functions which output values for which f^* is well-defined and finite. The witness function $g^\star$ is, if it exists, the function at which the supremum is attained. As can easily be derived (Nguyen et al., 2010, Lemma 1), the witness function is equal to[1]

$$g^\star = f'\left(\frac{dP}{dQ}\right)$$

[1]With some care needed when f is not differentiable everywhere (e.g., Total-Variation, Hockey-Stick).

where $\frac{dP}{dQ}$ denotes the Radon-Nikodym derivative. Therefore, the f-divergence is equal to

$$\begin{aligned} D_f(P \parallel Q) &= \mathbb{E}_P[g^\star(X)] - \mathbb{E}_Q[f^*(g^\star(Y))] \\ &= \mathbb{E}_P\left[f'\left(\frac{dP}{dQ}(X)\right)\right] - \mathbb{E}_Q\left[f^*\left(f'\left(\frac{dP}{dQ}(Y)\right)\right)\right]. \end{aligned}$$

Given samples $X_1, \ldots, X_m, \widetilde{X}_1, \ldots, \widetilde{X}_{\widetilde{m}}$ i.i.d. from P, and samples $Y_1, \ldots, Y_n, \widetilde{Y}_1, \ldots, \widetilde{Y}_{\widetilde{n}}$ i.i.d. from Q, we use $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ to construct an estimator $\widehat{r}_\lambda(\cdot)$ of $\frac{dP}{dQ}(\cdot)$ (see Equation (4)), and the samples $\widetilde{X}_1, \ldots, \widetilde{X}_{\widetilde{m}}$ and $\widetilde{Y}_1, \ldots, \widetilde{Y}_{\widetilde{n}}$ are used to estimate the expectations as

$$\widehat{D}_{f,\lambda} = \frac{1}{\widetilde{m}} \sum_{i=1}^{\widetilde{m}} f'\big(\widehat{r}_\lambda(\widetilde{X}_i)\big) - \frac{1}{\widetilde{n}} \sum_{j=1}^{\widetilde{n}} f^*\big(f'\big(\widehat{r}_\lambda(\widetilde{Y}_j)\big)\big). \tag{3}$$

Hence, with Equation (3), we are able to construct estimators for any f-divergence.

We leverage the variational formulation, instead of the direct estimator in Equation (1) for its flexibility in handling regularized f-divergences (see Section 5) but also because of its higher empirical performance in detection power (see Appendix B.3).

There are various ways of estimating $\frac{dP}{dQ}$ (Nguyen et al., 2010; Chen et al., 2024a; Nowozin et al., 2016). We focus on the method of Chen et al. (2024a, Proposition 6.1) which provides a kernel-based closed form expression for $\widehat{r}_\lambda$, a λ-regularized version of $\frac{dP}{dQ}$, using a kernel $k\colon \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ in an RKHS $\mathcal{H}$. For any $u \in \mathcal{Z}$ and $\lambda > 0$, it is defined as

$$\begin{aligned} \widehat{r}_\lambda(u) = \ & \frac{1}{n\lambda} k_{u\mathbf{Y}} \mathbb{1}_n - \frac{1}{m\lambda} k_{u\mathbf{X}} \mathbb{1}_m \\ & - \frac{1}{n\lambda} k_{u\mathbf{X}} L^{-1}_{\mathbf{XX},\lambda} k_{\mathbf{XY}} \mathbb{1}_n + \frac{1}{m\lambda} k_{u\mathbf{X}} L^{-1}_{\mathbf{XX},\lambda} k_{\mathbf{XX}} \mathbb{1}_m + 1 \end{aligned} \tag{4}$$

where we write $k_{u\mathbf{X}} = \big(k(u, X_i)\big)_{i=1}^m \in \mathbb{R}^{1\times m}$, $k_{u\mathbf{Y}} = \big(k(u, Y_i)\big)_{i=1}^n \in \mathbb{R}^{1\times n}$, $k_{\mathbf{XX}} = \big(k(X_i, X_j)\big)_{1\leq i,j\leq m} \in \mathbb{R}^{m\times m}$, $k_{\mathbf{XY}} = \big(k(X_i, Y_j)\big)_{1\leq i\leq m, 1\leq j\leq n} \in \mathbb{R}^{m\times n}$, and $L_{\mathbf{XX},\lambda} = m\lambda I + k_{\mathbf{XX}}$. Interestingly, the quantity $2(\widehat{r}_\lambda - 1)$ is shown by Chen et al. (2024a) to be the witness function for DrMMD, an MMD with a perturbed (deregularized) kernel.

First, we prove that the estimator of Equation (4) indeed tends to $\frac{dP}{dQ}$ and derive a rate of convergence.

Assumption 2.1. *The sample sizes are balanced in the sense that $m \asymp n$ and $\widetilde{m} \asymp \widetilde{n}$. The kernel is bounded by K everywhere. We have $\mu_P - \mu_Q \in \mathrm{Ran}(\Sigma_Q^\theta)$ for some $\theta \in (1, 2]$. Let $\eta \in (0, e^{-1})$, $N = \min(m, n)$, $\widetilde{N} = \min(\widetilde{m}, \widetilde{n})$ and $\lambda_{N,\theta} = N^{-1/2(\theta+1)}$.*

Lemma 2.1 (Witness convergence)**.** *Under Assumption 2.1, it holds with probability at least $1 - \eta$ that*

$$\left\| \widehat{r}_{\lambda_{N,\theta}} - \frac{dP}{dQ} \right\|_{\mathcal{H}} \leq \mathsf{C}_1 N^{-\frac{\theta-1}{2(\theta+1)}} \sqrt{\ln(1/\eta)} \tag{5}$$

for some universal constant $\mathsf{C}_1 > 0$.

Similarly, under permutation of the samples, we show in Proposition E.3 in Appendix E.3 that for some $\mathsf{C}_0 > 0$ we have $\left\|\widehat{r}_{\lambda_{N,\theta}} - 1\right\|_{\mathcal{H}} \leq \mathsf{C}_0 N^{-\frac{\theta-1}{2(\theta+1)}} \sqrt{\ln(1/\eta)}$ holding with probability at least $1-\eta$. We define

$$\mathsf{C} = \max\{\mathsf{C}_0, \mathsf{C}_1\}. \tag{6}$$

The result of Lemma 2.1 then allows us to prove that the estimator of Equation (3) is consistent in that it converges to the true f-divergence D_f as $m, \widetilde{m}, n, \widetilde{n}$ all tend to infinity. Before presenting this result, we first introduce some additional necessary assumptions.

Assumption 2.2. *For f-divergences with f twice continuously differentiable on $(0,\infty)$, assume that $\frac{dP}{dQ}$ belongs to $[c, C]$ for some $0 < c < 1 < C < \infty$, and that $\mathsf{C} N^{-\frac{\theta-1}{2(\theta+1)}} \sqrt{\ln(1/\eta)} \leq c/2$. For the Total-Variation ($\gamma = 1$) and Hockey-Stick ($\gamma > 0$) f-divergences, for $X \sim P$ and $Y \sim Q$, assume that the densities of $\widehat{r}_{\lambda_{N,\theta}}(X)$ and $\widehat{r}_{\lambda_{N,\theta}}(Y)$ exist and are bounded on $[\gamma/2, 3\gamma/2]$, and that $\mathsf{C} N^{-\frac{\theta-1}{2(\theta+1)}} \sqrt{\ln(1/\eta)} \leq \gamma/2$.*

We remark that it is possible to weaken the assumption, imposing instead that $\frac{dP}{dQ}$ belongs to $[c_\eta, C_\eta]$ only with probability at least $1 - \eta/2$ under both P and Q. This allows for more a wider class of alternatives to be considered, however, it comes at the cost of losing the explicit dependence on η, α, β in the rates (Lemma 2.2 and Theorem 3.2).

Lemma 2.2 (f-Divergence convergence)**.** *Under Assumptions 2.1 and 2.2, with probability at least $1-\eta$, it holds that*

$$\left|\widehat{D}_{f,\lambda_{N,\theta}} - D_f\right| \lesssim \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2}\right) \sqrt{\ln(1/\eta)}.$$

Hence, via Equation (3), we are able to construct a consistent estimator for any f-divergence. We stress that the conditions of Lemma 2.2 cover all commonly-used f-divergence such as all the ones presented in Table 4.

3 TESTING USING ANY f-DIVERGENCE WITH HIGH POWER

Given i.i.d. samples from P and Q, the goal of two-sample testing is to determine whether the data provides sufficient evidence to reject the null hypothesis that $H_0 \colon P = Q$.

A hypothesis test, which controls the probability of type I error (rejecting the null under H_0) by $\alpha \in (0,1)$, can easily be constructed using any statistic by relying on the permutation method (Romano and Wolf, 2005, see also Appendix A.4). As such, we can construct a permutation-test using the estimator of Equation (3) for any f-divergence. That is, the test rejects the null hypothesis when

$$\widehat{D}_{f,\lambda_{N,\theta}} > \widehat{q}_{1-\alpha} \tag{7}$$

where $\widehat{q}_{1-\alpha}$ is a permutation quantile as explained in Appendix A.4.

We prove that the resulting test is consistent, in the sense that, its probability of type II error (failing to reject the null when H_0 does not hold) converges to zero as the sample sizes tend to infinity. Equivalently, we say that the test power (i.e., one minus the type II error) converges to 1. We emphasize that this is not a simple implication of Lemma 2.2 as a thorough analysis of the behavior of the test statistic under permutation is also required (see Proposition E.3 in Appendix E.3).

Theorem 3.1 (Asymptotic Power)**.** *Under Assumptions 2.1 and 2.2, the test defined in Equation* (7) *is consistent in that its power converges to 1 as the sample sizes tend to infinity. For any fixed $P \neq Q$, we have*

$$\mathbb{P}(\text{reject } H_0) \to 1 \text{ as } N, \widetilde{N} \to \infty.$$

This asymptotic power guarantee is a desired property for any hypothesis test. Nonetheless, a much more challenging result to obtain is to characterize a set of alternatives which can be detected with high probability at any fixed sample size.

Theorem 3.2 (Non-asymptotic Power)**.** *Under Assumptions 2.1 and 2.2, the test defined in Equation* (7) *with level α can achieve power at least $1-\beta$ for any distributions P and Q satisfying*

$$D_f(P\|Q) \gtrsim \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-\frac{1}{2}}\right) \sqrt{\ln\left(\frac{1}{\min\{\alpha,\beta\}}\right)}.$$

We stress that the same power guarantees also hold for the permutation test based on the regularized χ^2-divergence (DrMMD, Chen et al., 2024a), as presented in Appendix F. Theorem 3.2 characterizes the set of alternatives detectable by our f-divergence permutation tests. In particular, any distributions P and Q for which the f-divergence is greater than some rate decreasing the sample sizes, can correctly be detected with high accuracy.

We note that the two-sample problem is inherently symmetric, while f-divergences are not. To address this issue, it is possible to instead work with either the average or the maximum of $D_f(P \,\|\, Q)$ and $D_f(Q \,\|\, P)$. An important problem is the one of the choice of hyperparameters such as the kernel bandwidth and the

regularization parameter λ. To address this issue, we construct an adaptive permutation test (detailed in Algorithm 1 in Appendix A.5) which adaptively combines results from multiple hyperparameter configurations (Biggs et al., 2023). Finally, building on Schrab et al. (2023), we also construct a test, which we call f-Agg, which aggregates these adaptive permutation tests over a collection of f-divergences to be powerful against a wide range of alternatives (detailed in Algorithm 2 in Appendix A.5).

4 HOCKEY-STICK f-DIVERGENCE

Motivated by its practical relevance for privacy and unlearning, we focus on the Hockey-Stick divergence with parameter γ. It is an f-divergence with $f(t) = \max(t-\gamma, 0)$, and its variational form is

$$\mathrm{HS}_\gamma(P||Q) \coloneqq \sup_{0\leq g\leq 1} \mathbb{E}_{X\sim P}[g(X)] - \gamma \mathbb{E}_{X\sim Q}[g(X)].$$

The Hockey-Stick witness function is equal to

$$g^\star_{HS}(x) = \mathbb{1}\left(\frac{dP}{dQ}(x) \geq \gamma\right) \tag{8}$$

As in explained Section 2, by estimating $\frac{dP}{dQ}$ with $\widehat{r}_\lambda$ from Equation (4), we obtain the witness estimator $\mathbb{1}(\widehat{r}_\lambda(x) \geq \gamma)$. Noting that $f^*(u) = \sup_{t\in\mathbb{R}}\{ut - f(t)\} = \gamma u$ for all $u \in [0,1]$, the Hockey-Stick estimator equivalent of Equation (3) is

$$\widehat{\mathrm{HS}}_{\gamma,\lambda} = \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} \mathbb{1}(\widehat{r}_\lambda(\widetilde{X}_i) \geq \gamma) - \frac{\gamma}{\widetilde{n}}\sum_{j=1}^{\widetilde{n}} \mathbb{1}(\widehat{r}_\lambda(\widetilde{Y}_j) \geq \gamma). \tag{9}$$

A permutation two-sample test can then be constructed using this Hockey-Stick estimator as explained and analysed in Section 3. See also Algorithm 1 in Appendix A.5 for details on the adaptivity over the kernel and the regularization parameter λ.

5 REGULARIZED f-DIVERGENCES

So far we have studied f-divergence based tests that leverage a regularized likelihood ratio $\hat{r}$ to compute the divergence via substitution into the f-divergence variational definition, i.e., through Equation (3). Alternatively, one could consider estimating the regularized f-divergence by solving its corresponding regularized variational formulation directly, as follows.

Definition 5.1. *(Regularized f-divergence.) Let $D_f(\cdot||\cdot)$ be an f-divergence, $\mathcal{H}$ be a RKHS, and $\lambda > 0$. For two distributions P and Q over a set $\mathcal{X}$ we define the regularized f-divergence $D_f^{\mathcal{H},\tau}(P||Q)$ as*

$$\sup_{g:\mathcal{X}\to \mathrm{Dom}(f^*)} \mathbb{E}_P\left[g(x)\right] - \mathbb{E}_Q\left[f^*(g(x))\right] - \frac{\tau}{2}\|g\|^2_{\mathcal{H}}.$$

While the concept of regularized f-divergences has appeared before, prior work focused on specific divergences, such as the regularized KL (Nguyen et al., 2010; Nowozin et al., 2016; Arbel et al., 2020) or KALE, and regularized-χ^2 or DrMMD (Harchaoui et al., 2007; Chen et al., 2024a). These previous works focused on specific applications, leveraging these statistics for asymptotic tests (Harchaoui et al., 2007), and particle descent and gradient flows as measures for optimization and modeling (Arbel et al., 2020; Glaser et al., 2021; Chen et al., 2024a), while the focus of the present work is in constructing permutation-based non-asymptotic tests with adaptation over the kernel and regularization parameter.

Testing with regularized f-divergences is challenging, as it requires specific optimization techniques for each f-divergence, based on the objective and constraints in Definition 5.1. For example, the KALE (regularized KL) statistic formulation leads to a convex optimization problem that can be solved efficiently, although without a closed-form solution. In contrast, the DrMMD (regularized χ^2) has a closed-form expression that can be derived using tools from functional analysis. We provide their detailed formulations in Appendix A.7 and below present an approach for testing with a regularized hockey-stick.

Regularized hockey-stick divergence The regularized Hockey-Stick divergence is defined by

$$\begin{aligned}\mathrm{HS}^\tau_\gamma(P||Q) \coloneqq \sup_{g\in\mathcal{H},\, 0\leq g\leq 1} \big(&\mathbb{E}_{X\sim P}[g(X)]. \\ &- \gamma\mathbb{E}_{X\sim Q}[g(X)] - \tau\|g\|^2_{\mathcal{H}}\big),\end{aligned} \tag{10}$$

where $\tau > 0$ controls the regularization strength. We emphasize that this τ-regularization of the f-divergence itself is different from the λ-regularization of the $\frac{dP}{dQ}$ estimator used in Equation (3).

In Appendix D, we provide a theoretical framework for solving the optimization problem in Equation (10) using the plug-in estimator and techniques from optimization with non-negative constraints. The problem turns into a semidefinite program, and can be efficiently solved using accelerated proximal splitting methods like FISTA (Beck and Teboulle, 2009).

In practice, we do not use this approach, since the RKHS function class employed in Equation (10) is not suited to representing discontinuous functions of the form of Equation (8), and empirical performance in testing is consequently poor (see Figure 9 in the appendix for an illustration). We nonetheless present this direct approximation as it will be of interest to revisit with more suitable function classes in future work.

6 EXPERIMENTS

We begin by evaluating the performance of the tests proposed in Section 3 on two synthetic benchmarks, illustrating that different alternatives are more effectively captured by different f-divergence tests. We then test on practical applications from differential privacy auditing and unlearning evaluation. All our code is publicly available at `https://github.com/google-research/google-research/tree/master/f_divergence_tests`

Experimental details. For our two-sample f-divergence tests, we focus on the Hockey-Stick test presented in Section 4. In Appendix B.1, we also present results for various other f-divergences and regularized f-divergences. However, in the main body we restrict ourselves to the ones with highest performance.

We use the state-of-the-art Maximum Mean Discrepancy (MMD, Gretton et al., 2012, see Appendix A.3) as our primary baseline. We also compare against two novel permutation tests based on regularized f-divergence statistics: DrMMD (regularized χ^2, Chen et al., 2024a) and KALE (regularized KL, Arbel et al., 2020; Glaser et al., 2021; Grosso and Letizia, 2025), detailed in Appendix A.7. Additionally, we include f-Agg, a test which leverages the aggregation method of Schrab et al. (2023) to perform multiple testing across all divergences considered (Algorithm 2). For privacy experiments, we also report the power of the Hockey-Stick tester from the DP-Auditorium library (Kong et al., 2024).

All experiments use 500 samples from each distribution unless otherwise noted, with the exception of the Expo-1D experiment, which uses 1000 samples. All tests are performed at significance level $\alpha = 0.05$. For each test, we report the empirical power calculated over 100 repetitions, along with Clopper-Pearson confidence intervals. All experiments were run on NVIDIA A100 GPUs.

6.1 Perturbed Uniforms

This benchmark was introduced by Schrab et al. (2023). We test power over detection of perturbations on one and two dimensional uniform distributions, varying the amplitude of the perturbation from $a = 0$ (no perturbation, corresponding to the null $P = Q$) to $a = 1$ (maximum value of the perturbation for the density to remain non-negative). Consistent with previous experiments, we confirm in Figure 1 that MMD is optimal at detecting these smooth differences. This is unsurprising as MMD is proved to be minimax optimal in this setting (Schrab et al., 2023). Nonetheless, even in this setting optimized for MMD, we observe that the

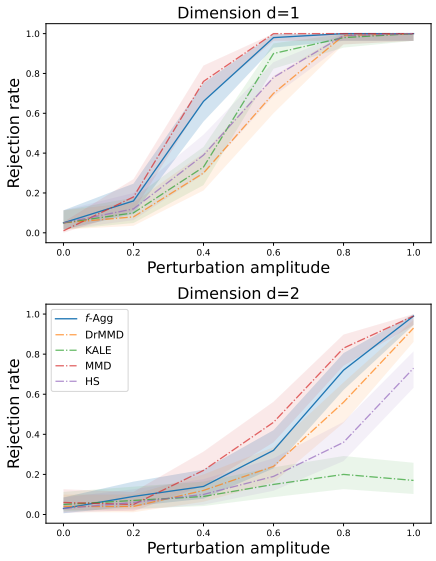

Figure 1: Performance on perturbed d-dimensional uniform alternatives with varying perturbation amplitudes. As the amplitude increases, the deviation from uniformity grows, leading to higher statistical power. Both MMD and DrMMD perform well in this setting, and the performance decreases for all methods in two dimensions. The f-Agg test maintains almost the same performance as the best method.

f-divergence tests perform well with power relatively close to MMD. We note that the adaptive f-Agg test achieves the highest power after the MMD test, showing the benefit of aggregating over multiple statistics.

6.2 Outlier Detection on Physics Datasets

New physics address limitations within the current standard model of particle physics. In this context, two-sample tests are often used to identify anomalies or validate new models, among other tasks (Grosso and Letizia, 2025). We consider the Expo-1D benchmark, first introduced in D'agnolo and Wulzer (2019), and then considered in various works (Letizia et al., 2022; Grosso and Letizia, 2025; Grosso et al., 2024, 2025). In this dataset, there is a reference model given by an energy spectrum that decays exponentially, described by the unnormalized density $n_0(x) = 2000e^{-x}$. The alternative hypothesis is given by the density $n_0(x) + k\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{x-\mu^2}{2\sigma^2}\right)$. In our experiments of Figure 3, we vary the values of the multiplier k while fixing $\mu = 4$ and $\sigma = 0.01$. For convenience, we plot the perturbed densities in Figure 2 for $k \in \{0, 20, 40, 60, 80, 100\}$.

In this setting, Figure 3 shows that, excluding the adaptive f-Agg test, the DrMMD test achieves the highest statistical power across all values of the multi-

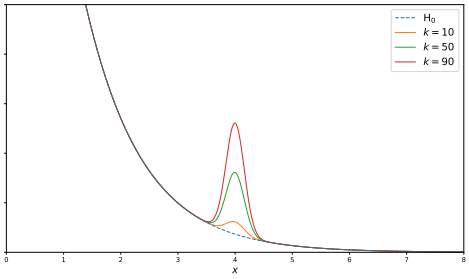

Figure 2: Expo-1D null (H_0) and alternative hypothesis density functions varying the multiplier k.

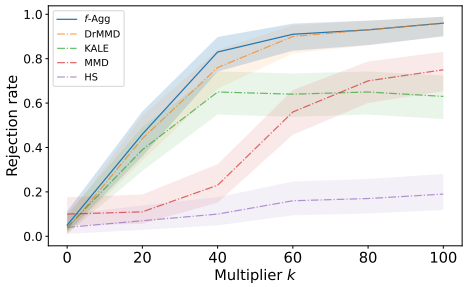

Figure 3: Detection rate comparison on the Expo-1D task, varying the multiplier k of the perturbation. The DrMMD test achieves the highest power across all values.

plier k. In contrast, MMD exhibits low power for small k, indicating that deregularization makes the resulting statistic more sensitive to this type of departure from the null, compared to the baseline MMD. The KALE estimator is also effective at small k, but its power stagnates as k increases to greater values than 40. This can occur because the slow growth of the log-likelihood ratio as a function of k poses a greater challenge to log-ratio estimators like KALE than to direct-ratio estimators like DrMMD. We illustrate this phenomenon in Appendix C.1.

The superior performance of DrMMD, which is equivalent to MMD with a deregularized kernel, shows that MMD's power can be significantly boosted with a sufficiently expressive kernel. Our results show that DrMMD achieves a power increasing with k, up to 0.96 in two dimensions, using only 1000 samples from each distribution. This represents a major improvement in both statistical power and sample efficiency.

We observe that the HS-sigmoid method outperforms the hard-threshold version HS. Still, both Hockey-Stick instantiations are suboptimal for this set of alternatives. For generic two-sample testing, we do not recommend using the Hockey-Stick divergence over other f-divergences. However, there are settings, such as auditing differential privacy auditing (Section 6.3) and machine unlearning (Sekhari et al., 2021), where the use of the Hockey-Stick divergence is necessarily required.

6.3 Auditing Pure Differential Privacy

Differential privacy (Dwork et al., 2006) provides a rigorous framework for protecting user data by introducing calibrated noise to query responses, bounding the influence of any single individual on a query response. Formally, a randomized mechanism M with outputs in $\mathbb{R}^p$ is ε-differentially private (ε-DP) if and only if, for any $S \subseteq \mathbb{R}^p$ and any pair of datasets D and D' differing in one record, the probability of obtaining any output in S is nearly the same, that is

$$P(M(D) \in S) \leq e^{\varepsilon} P(M(D') \in S).$$

Privacy guarantees rely on the correct implementation of the mechanism M. We can audit these guarantees by framing the problem as a two-sample hypothesis test, based on Lemma 5 from Zhu et al. (2022). Specifically, for any correctly implemented ε-DP random mechanism M, the Hockey-Stick divergence (with parameter e^{ε}) between the outputs from two adjacent datasets D and D' must be zero, that is

$$\mathrm{HS}_{e^{\varepsilon}}(M(D)||M(D')) = \mathrm{HS}_{e^{\varepsilon}}(M(D')||M(D)) = 0.$$

It follows directly that if our proposed test (Algorithm 1, Sections 3 and 4) rejects the null, $\mathrm{HS}_{e^{\varepsilon}}(M(D)||M(D')) = 0$, then the mechanism M cannot be private for the given ε. This method builds upon methodologies for auditing privacy mechanisms (Gilbert and McMillan, 2018; Bichsel et al., 2021; Kong et al., 2024).

We focus on testing six non-private mechanisms previously evaluated in the literature. The first four mechanisms, SVT3–SVT6, are incorrect instances of the sparse-vector-technique (Lyu et al., 2017). The methods mean-1 and mean-2 are incorrect implementations of the Laplace mechanism for the average. As in previous work, we fix D and D' for each mechanism. Details on the audited mechanisms and the baselines can be found in Appendix C.2. The results are reported in Table 1.

Our Algorithm 1 achieves higher rejection rates for privacy violations while requiring significantly fewer samples than the baseline Hockey-Stick tester from DP-Auditorium, improving the overall sample efficiency of the auditing process.

Achieving a rejection rate greater than level values for the SVT3 mechanism is noteworthy. While previous techniques detected some privacy violations for SVT3, they required millions of samples to approximate the likelihood ratio (Bichsel et al., 2021).

Table 1: Privacy violations detection rate for several non-private mechanisms. HS improves detection rate over DP-auditorium which requires hundreds of thousand of samples to detect some of the mechanisms below.

	Test	SVT3	SVT4	SVT5	SVT6	mean1	mean2
$n = 500$	HS	0.095	0.070	0.785	0.085	0.970	1.0
	DP-Aud	0.0	0.0	0.0	0.0	0.768	0.071
$n = 5000$	HS	0.175	0.090	0.995	0.165	1.0	1.0
	DP-Aud	0.0	0.0	0.010	0.0	1.0	0.717

6.4 Evaluating Machine Unlearning

The goal of unlearning is to remove the influence of a specific subset of data $D_f \subseteq D$ often called the forget set, from a trained model M. A widely used definition (Sekhari et al., 2021) considers unlearning successful if the resulting model, M_u, is statistically indistinguishable from a model M_r that was retrained from scratch on the remaining data $D \setminus D_f$. This common definition of unlearning quantifies indistinguishability using the Hockey-Stick divergence with parameter e^ε. The parameter ε controls the degree of statistical indistinguishability. Larger ε allows for greater divergence between the two model distributions, while $\varepsilon = 0$ enforces exact unlearning and requires the distributions to be identical.

Evaluating this definition can be framed as a two-sample hypothesis test where the null hypothesis is that the Hockey-Stick divergence of order e^ε between unlearned and retrained models is zero, $\mathrm{HS}_{e^\varepsilon}(M_u||M_r) = \mathrm{HS}_{e^\varepsilon}(M_r||M_u) = 0$, where the randomness comes from the training procedure (e.g., batch-sampling, dropout, etc.).

This strict definition has the following flaw. Different training procedures (e.g., number of training steps) mean that even two models M_r and $M_{r'}$ retrained from scratch on the exact same data will likely produce slightly different parameter distributions. A sensitive two-sample test will correctly detect this difference between M_r and $M_{r'}$, and reject the null hypothesis $\mathrm{HS}_{e^\varepsilon}(M_{r'}||M_r) = \mathrm{HS}_{e^\varepsilon}(M_r||M_{r'}) = 0$, even though the unlearning process of retraining from scratch with different parameters is entirely valid.

To confirm that this issue arises in practice, we trained a CNN on the CIFAR-10 dataset Krizhevsky et al. (2009) and then applied several simplified unlearning techniques to forget a set of ten randomly selected images. For details on training procedures and unlearning algorithms, see Appendix C.3.

This flaw is demonstrated in Table 2. We tested several unlearning methods against a retrained model at levels $\varepsilon = 0$ and $\varepsilon = 0.1$. For $\varepsilon = 0.1$, we can only test with Hockey-Stick statistic to quantify a positive separation.

Table 2: Two-sample test power results against the null hypothesis $H_0 \colon \mathrm{HS}_{e^\varepsilon}(M_u||M_r) = \mathrm{HS}_{e^\varepsilon}(M_r||M_u) = 0$, with $\varepsilon = 0$ and $\varepsilon = 0.1$. A rejection rate of 1.0 means the test detects a distributional difference.

Unlearning mechanism	DrMMD	KALE	MMD	HS	HS
		$\varepsilon = 0$			$\varepsilon = 0.1$
Finetune	1.0	0.2	1.0	1.0	1.0
Pruning	1.0	0.0	1.0	1.0	0.1
SSD	1.0	0.0	1.0	1.0	0.1
Random label	1.0	0.0	1.0	1.0	0.0
Retrained-less	1.0	0.5	1.0	1.0	1.0
Retrained-batch	1.0	0.0	1.0	1.0	0.4

For $\varepsilon = 0$, or exact unlearning, we test with all the statistics.

For $\varepsilon = 0$ (Table 2), all approximate unlearning methods produce a different distribution than the retrained one. However, even for $\varepsilon > 0$ the test detects the difference between the retrained model and models retrained from scratch on retain data using a different batch size and half as many steps (i.e., 'retrain-batch' and 'retrain-less'). This shows that the two-sample test, while statistically correct, is not practically useful for verifying that a model does not hold information about specific forget examples.

We emphasize that this flaw does not represent an issue with our tests which correctly distinguish the difference in distributions of M_u and M_r. Instead, it highlights that the null hypothesis that corresponds to this unlearning definition, $H_0 \colon \mathrm{HS}_{e^\varepsilon}(M_u||M_r) = \mathrm{HS}_{e^\varepsilon}(M_r||M_u) = 0$, is not necessarily equivalent to the null hypothesis that the unlearning process is 'successful', as one may intuitively reason about it. A potentially better-suited null hypothesis to characterize successful unlearning could be the composite goodness-of-fit null hypothesis (e.g., Key et al., 2025) that M_u is identically distributed as some element of the set $\mathcal{M}_r$ consisting of all models retrained from scratch on $D \setminus D_f$ with arbitrary training procedures. However, constructing a practical test for this composite null is challenging and remains an open problem. In the next section, we instead propose a different approach to fixing this flaw based on three-sample hypothesis testing.

We note that the KALE-based method performs poorly on this task, potentially for two reasons. First, the high dimensionality of the sample space ($d = 10$) may impact the optimization required to calculate the KALE statistic. This claim is supported by a similar performance drop observed in the two-dimensional perturbed uniforms experiment in Figure 1. Second, the log-likelihood ratio between the distributions may be ill-conditioned. Characterizing the set of alternative distributions for which KALE is optimal is an important direction for future work.

Three-sample tests. To address the highlighted issue, we propose instead to use a *three-sample test*. Instead of checking whether the unlearned model is indistinguishable to a retrained one, this procedure reframes the problem as comparing three distributions: the original model, the unlearned model, and the retrained model. The null hypothesis becomes H_0 : Unlearning was effective, measured by closeness of the unlearned model distribution to the retrained model distribution, relative to the distance to the original model distribution. Rejecting the null provides evidence that unlearning was ineffective.

A three-sample test was proposed by Bounliphone et al. (2016) for the MMD. This test relies on the joint asymptotic distribution of *two* statistics, $d(M_u, M)$ and $d(M_u, M_r)$, which respectively measure the distance from the unlearned model M_u to the retrained model M_r, and from M_u to the original model M. In the event that $M_u \neq M_r \neq M$, the joint distribution of the two quantities $\text{MMD}(M_u, M)$ and $\text{MMD}(M_u, M_r)$ is asymptotically Gaussian, yielding a straightforward test with $H_0 : \text{MMD}(M_u, M_r) \leq \text{MMD}(M_u, M)$. Details can be found in Appendix C.3.

We therefore employ this three-sample MMD test to measure the success of unlearning algorithms. While it will be of interest to generalize the two-sample f-divergence tests to the three-sample setting, this requires establishing asymptotics of the variational f-divergence form in Equation (2) for divergences of interest (such as the KL and χ^2), which will be an interesting direction for future work.

The results of the three-sample MMD test are shown in Table 3. The test correctly and consistently identifies that the retrained models are "safe" (rejection rate = 0.0), overcoming the flaw of the two-sample approach. Among the approximate methods, only the Random Label technique passed the evaluation, while others like Finetuning, Pruning, and SSD were found to be ineffective (rejection rate = 1.0). We emphasize that our goal is the evaluation of unlearning methods, and not designing good algorithms for unlearning. Consequently, we used simplistic implementations of the unlearning procedures and a more rigorous implementation is needed for ranking the unlearning methods in practice.

Table 3: Three-sample test. Rejection rate for the relative similarity test. A rejection rate of 1.0 means the test detects that the unlearned model is closer to the original model than to the retrained model, a proxy for evidence that the unlearning method is ineffective. The test correctly identifies retraining from scratch on retain data as a safe procedure (rejection rate = 0.0).

Method	Power of relative test
Finetune	1.0
Pruning	1.0
SSD	1.0
Random-label	0.0
Retrained-less	0.0
Retrained-batch	0.0

7 CONCLUSION

We introduce a general framework for estimating f-divergences, and an accompanying two-sample permutation test. We establish finite sample bounds for the statistics and consistency of the tests, and characterize their sets of detectable alternatives. We illustrate the applicability of the Hockey-Stick divergence for differential privacy and unlearning problems. In experiments, we show that a deregularized MMD (DrMMD) approximation to the χ^2 divergence, and the KALE approximation to the KL divergence, both outperform classical MMD tests in certain domains, notably outlier detection. Finally, we demonstrate a limitation of two-sample tests in assessing unlearning algorithms, and show that this can be overcome with a three-sample test. Theoretically grounding these observations and characterizing optimal estimators for other divergences are interesting directions for future work.

Acknowledgments

Antonin Schrab is supported by a UKRI Turing AI World-Leading Researcher Fellowship on Advancing Modern Data-Driven Robust Al with grant number G111021. We thank Eleni Triantafillou for helpful guidance in the unlearning experiment. We thank Pierre Glaser and Zonghao Chen for their help to understand details of their relevant respective works, Glaser et al. (2021) and Chen et al. (2024a).

References

Albert, M., Laurent, B., Marrel, A., and Meynaoui, A. (2022). Adaptive test of independence based on HSIC measures. *The Annals of Statistics*, 50(2):858–879.

Arbel, M., Zhou, L., and Gretton, A. (2020). Generalized energy based models. *arXiv preprint arXiv:2003.05033*.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, pages 183–202.

Biau, G. and Gyorfi, L. (2005). On the asymptotic properties of a nonparametric l/sub 1/-test statistic of homogeneity. *IEEE Transactions on Information Theory*.

Bichsel, B., Steffen, S., Bogunovic, I., and Vechev, M. (2021). Dp-sniper: Black-box discovery of differ-

ential privacy violations using classifiers. In *IEEE Symposium on Security and Privacy (SP)*. IEEE.

Biggs, F., Schrab, A., and Gretton, A. (2023). MMD-FUSE: Learning and combining kernels for two-sample testing without data splitting. *Advances in Neural Information Processing Systems (NeurIPS)*.

Bounliphone, W., Belilovsky, E., Blaschko, M. B., Antonoglou, I., and Gretton, A. (2016). A test of relative similarity for model selection in generative models. *International Conference on Learning Representations (ICLR)*.

Chau, S. L., Schrab, A., Gretton, A., Sejdinovic, D., and Muandet, K. (2025). Credal two-sample tests of epistemic ignorance. In *International Conference on Artificial Intelligence and Statistics*.

Chen, Z., Mustafi, A., Glaser, P., Korba, A., Gretton, A., and Sriperumbudur, B. K. (2024a). (de)-regularized maximum mean discrepancy gradient flow. *arXiv preprint arXiv:2409.14980*.

Chen, Z., Mustafi, A., Glaser, P., Korba, A., Gretton, A., and Sriperumbudur, B. K. (2024b). (de)-regularized maximum mean discrepancy gradient flow. *arXiv preprint arXiv:2409.14980*.

Chwialkowski, K., Sejdinovic, D., and Gretton, A. (2014). A wild bootstrap for degenerate kernel tests. In *Advances in neural information processing systems*, pages 3608–3616.

Csiszár, I. (1967). On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

D'agnolo, R. T. and Wulzer, A. (2019). Learning new physics from a machine. *Physical Review D*.

Foster, J., Schoepf, S., and Brintrup, A. (2024). Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 12043–12051.

Fromont, M., Laurent, B., Lerasle, M., and Reynaud-Bouret, P. (2012). Kernels Based Tests with Non-asymptotic Bootstrap Approaches for Two-sample Problems. In *Conference on Learning Theory*, volume 23 of *Journal of Machine Learning Research Proceedings*.

Fromont, M., Laurent, B., and Reynaud-Bouret, P. (2013). The two-sample problem for Poisson processes: Adaptive tests with a nonasymptotic wild bootstrap approach. *The Annals of Statistics*, 41(3):1431–1461.

Gilbert, A. C. and McMillan, A. (2018). Property testing for differential privacy. In *Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE.

Glaser, P., Arbel, M., and Gretton, A. (2021). Kale flow: A relaxed kl gradient flow for probabilities with disjoint support. *Advances in Neural Information Processing Systems (NeurIPS)*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*.

Grosso, G., Hindupur, S. S. R., Fel, T., Bright-Thonney, S., Harris, P., and Ba, D. (2025). Sparse, self-organizing ensembles of local kernels detect rare statistical anomalies. *arXiv preprint arXiv:2511.03095*.

Grosso, G. and Letizia, M. (2025). Multiple testing for signal-agnostic searches for new physics with machine learning. *The European Physical Journal C*.

Grosso, G., Letizia, M., Pierini, M., and Wulzer, A. (2024). Goodness of fit by neyman-pearson testing. *SciPost Physics*, 16(5):123.

Harchaoui, Z., Bach, F., and Eric, M. (2007). Testing for homogeneity with kernel fisher discriminant analysis. *Advances in Neural Information Processing Systems*, 20.

Hemerik, J. and Goeman, J. (2018). Exact testing with random permutations. *TEST*.

Kanamori, T., Hido, S., and Sugiyama, M. (2009). A least-squares approach to direct importance estimation. *The Journal of Machine Learning Research*, 10:1391–1445.

Kanamori, T., Suzuki, T., and Sugiyama, M. (2012). Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*.

Key, O., Gretton, A., Briol, F.-X., and Fernandez, T. (2025). Composite goodness-of-fit tests with kernels. *Journal of Machine Learning Research*, 26(51):1–60.

Kim, I. (2021). Comparing a large number of multivariate distributions. *Bernoulli*, 27(1):419–441.

Kim, I., Balakrishnan, S., and Wasserman, L. (2022). Minimax optimality of permutation tests. *The Annals of Statistics*, 50(1):225 – 251.

Kim, I. and Schrab, A. (2023). Differentially private permutation tests: Applications to kernel methods. Arxiv preprint 2310.19043.

Kim, I. and Schrab, A. (2025). Differentially private permutation tests. *Journal of the American Statistical Association*, pages 1–21.

Kong, W., Medina, A. M., Ribero, M., and Syed, U. (2024). Dp-auditorium: A large-scale library for

auditing differential privacy. In *IEEE Symposium on Security and Privacy (SP)*.

Krizhevsky, A., Hinton, G., et al. (2009). Learning multiple layers of features from tiny images.

Kübler, J. M., Stimper, V., Buchholz, S., Muandet, K., and Schölkopf, B. (2022). Automl two-sample test. *Advances in Neural Information Processing Systems (NeurIPS)*.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer.

Letizia, M., Losapio, G., Rando, M., Grosso, G., Wulzer, A., Pierini, M., Zanetti, M., and Rosasco, L. (2022). Learning new physics efficiently with nonparametric methods. *The European Physical Journal C*, 82(10):879.

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

Lyu, M., Su, D., and Li, N. (2017). Understanding the sparse vector technique for differential privacy. *Proc. VLDB Endow.*

Marteau-Ferey, U., Bach, F., and Rudi, A. (2020). Non-parametric models for non-negative functions. *Advances in neural information processing systems (NeurIPS)*.

Massart, P. (1986). Rates of convergence in the central limit theorem for empirical processes. In *Annales de l'IHP Probabilités et statistiques*, volume 22, pages 381–423.

Mescheder, L., Nowozin, S., and Geiger, A. (2017). Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *International Conference on Machine Learning (ICML)*. PMLR.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*.

Nowozin, S., Cseke, B., and Tomioka, R. (2016). f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29.

Pillutla, K., Liu, L., Thickstun, J., Welleck, S., Swayamdipta, S., Zellers, R., Oh, S., Choi, Y., and Harchaoui, Z. (2023). Mauve scores for generative models: Theory and practice. *Journal of Machine Learning Research*.

Pogodin, R., Schrab, A., Li, Y., Sutherland, D. J., and Gretton, A. (2024). Practical kernel tests of conditional independence. *arXiv preprint arXiv:2402.13196*.

Polyanskiy, Y. and Wu, Y. (2022). Information theory: From coding to learning. *Book draft*.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108.

Schrab, A. (2025a). A Practical Introduction to Kernel Discrepancies: MMD, HSIC & KSD. *arXiv preprint arXiv:2503.04820*.

Schrab, A. (2025b). A Unified View of Optimal Kernel Hypothesis Testing. *arXiv preprint arXiv:2503.07084*.

Schrab, A. (2025c). *Optimal Kernel Hypothesis Testing*. PhD thesis, UCL (University College London).

Schrab, A., Guedj, B., and Gretton, A. (2022a). KSD Aggregated Goodness-of-fit Test. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*.

Schrab, A. and Kim, I. (2025). Robust kernel hypothesis testing under data corruption. In *International Conference on Artificial Intelligence and Statistics*.

Schrab, A., Kim, I., Albert, M., Laurent, B., Guedj, B., and Gretton, A. (2023). MMD aggregated two-sample test. *Journal of Machine Learning Research*.

Schrab, A., Kim, I., Guedj, B., and Gretton, A. (2022b). Efficient Aggregated Kernel Tests using Incomplete U-statistics. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022*, volume 35, pages 18793–18807.

Sekhari, A., Acharya, J., Kamath, G., and Suresh, A. T. (2021). Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Singh, S. and Póczos, B. (2016). Analysis of k-nearest neighbor distances with application to entropy estimation. *arXiv preprint arXiv:1603.08578*.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., Von Bünau, P., and Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*.

Sutherland, D. J., Tung, H.-Y., Strathmann, H., De, S., Ramdas, A., Smola, A., and Gretton, A. (2017). Generative models and model criticism via optimized maximum mean discrepancy. In *International Conference on Learning Representations*.

Triantafillou, E., Kairouz, P., Pedregosa, F., Hayes, J., Kurmanji, M., Zhao, K., Dumoulin, V., Junior, J. J., Mitliagkas, I., Wan, J., et al. (2024). Are we making progress in unlearning? findings from the first neurips unlearning competition. *arXiv preprint arXiv:2406.09073*.

Wang, Q., Kulkarni, S. R., and Verdú, S. (2009). Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405.

Xu, L., Chen, Y., Doucet, A., and Gretton, A. (2022). Importance weighting approach in kernel bayes' rule. *arXiv preprint arXiv:2202.02474*.

Yu, J., He, Y., Goyal, A., and Arora, S. (2025). On the impossibility of retrain equivalence in machine unlearning. *arXiv preprint arXiv:2510.16629*.

Zhou, Z., Tian, X., Peng, L., Lei, C., Schrab, A., Sutherland, D. J., and Liu, F. (2025). Dual: Learning diverse kernels for aggregated two-sample and independence testing. *Advances in Neural Information Processing Systems*, 39.

Zhu, Y., Dong, J., and Wang, Y.-X. (2022). Optimal accounting of differential privacy via characteristic function. In *International Conference on Artificial Intelligence and Statistics*, pages 4782–4817. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes] All theorems include assumptions, and algorithms and models clear descriptions.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes] We include them in the supplementary.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable] [Yes] We include them with our final version.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] We include them in all theorem statements.
   (b) Complete proofs of all theoretical results. [Yes] We include them in the supplementary.
   (c) Clear explanations of any assumptions. [Yes] We include them when necessary.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes] We will include the url pointer to our code with the final version.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes] We include them in the supplementary.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes] We include them in experiments section.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] We include them in the experiments section.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes] We include all necessary citations.
   (b) The license information of the assets, if applicable. [Yes] We include them in the supplementary.
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Regularized f-Divergence Kernel Tests
Supplementary Material

We give an overview of the supplementary material. In Appendix A, we provide additional background information on common f-divergences, on MMD, on permutation tests, on adaptive testing, on KALE, and on DrMMD. In Appendix B, we run additional experiments. In Appendix C, we provide additional detials on the experimental settings. In Appendix D, we derive a direct optimization procedure for the computation of the Hockey-Stick witness function. In Appendix E, we formally prove all the statements presented in Sections 3 and 4. In Appendix F, we prove that the DrMMD permutation tests satisfies the same theoretical power guarantees as our proposed tests.

A BACKGROUND

In this section, we provide background information on common f-divergences (Appendix A.1), on kernel mean embeddings and covariance operators (Appendix A.2), on MMD (Appendix A.3), on permutation tests (Appendix A.4), on adaptive testing (Appendix A.5), on implementation details (Appendix A.6), and on KALE and DrMMD (Appendix A.7).

A.1 Table of f-Divergences

We provide a table of commonly-used f-divergences in Table 4 with the associated function f, the convex conjugate f^*, and the witness function $g^\star$.

Table 4: f-divergences: generator $f(t)$, convex conjugate $f^*(u)$, and derivative $f'(r)$. Recall that the optimal witness $g^\star(x) = f'(r)$ with $r = \frac{dP}{dQ}(x)$. We use the Lambert W function.

f-Divergence	$f(t)$	$f^*(u)$	$f'(r)$
KL (Kullback–Leibler)	$t \log t$	e^{u-1}	$1 + \log r$
Reverse KL	$-\log t$	$-1 - \log(-u),\ u < 0$	$-\frac{1}{r}$
Jeffreys (symmetrized KL)	$(t-1)\log t$	$W(e^{u-1}) + \frac{1}{W(e^{u-1})} + u - 2$	$1 + \log r - \frac{1}{r}$
Jensen–Shannon	$t \log t - (t+1)\log \frac{t+1}{2}$	$-\log(2 - e^u), u < \ln(2)$	$\log \frac{2r}{r+1}$
Total Variation	$\frac{1}{2}\lvert t-1\rvert$	$u,\ \lvert u\rvert < 1/2$	$\text{sign}(r-1)/2$
Pearson χ^2	$(t-1)^2$	$u + u^2/4$	$2(r-1)$
Neyman χ^2	$\frac{(1-t)^2}{t}$	$2\big(1 - \sqrt{1-u}\big),\ u < 1$	$1 - \frac{1}{r^2}$
Vincze–LeCam	$\frac{(t-1)^2}{t+1}$	$4 - u - 4\sqrt{1-u},\ \lvert u\rvert < 1$	$1 - \frac{4}{(r+1)^2}$
Squared Hellinger	$(\sqrt{t}-1)^2$	$\frac{u}{1-u},\ u < 1$	$1 - \frac{1}{\sqrt{r}}$
Hellinger Discrimination	$1 - \sqrt{t}$	$-1 - \frac{1}{4u},\ u < 0$	$-\frac{1}{2\sqrt{r}}$
α-Divergence (α)	$\frac{4}{1-\alpha^2}\big(1 - t^{(1+\alpha)/2}\big)$	$\frac{2}{1+\alpha}\big(\frac{\alpha-1}{2}u\big)^{(1+\alpha)/(\alpha-1)} - \frac{4}{1-\alpha^2}$	$-\frac{2}{1-\alpha}r^{(\alpha-1)/2}$
Cressie–Read (λ)	$\frac{t^\lambda - \lambda(t-1) - 1}{\lambda(\lambda-1)}$	$\big((u(\lambda-1)+1)^{\lambda/(\lambda-1)} - 1\big)/\lambda$	$\frac{r^{\lambda-1}-1}{\lambda-1}$
Hockey-Stick (γ)	$\max\{t - \gamma, 0\}$	$\gamma u,\ 0 < u < 1$	$\mathbb{1}(r \geq \gamma)$

A.2 Kernel Mean Embedding and Covariance Operator

Recall that in an RKHS $\mathcal{H}$ with kernel k, the kernel reproducing property states that $\langle f, k(x,\cdot)\rangle = f(x)$ for all x. Given a distribution P, the kernel mean embedding is an element $\mu_P \in \mathcal{H}$ such that $\langle f, \mu_P\rangle = \mathbb{E}_{X\sim P}[f(X)]$ for all $f \in \mathcal{H}$. It can be expressed as $\mu_P = \mathbb{E}_P[k(X,\cdot)]$ so that

$$\langle f, \mathbb{E}_P[k(X,\cdot)]\rangle = \langle f, \mu_P\rangle = \mathbb{E}_{X\sim P}[f(X)] = \mathbb{E}_{X\sim P}[\langle f, k(X,\cdot)\rangle]$$

for all $f \in \mathcal{H}$. The (uncentered) covariance operator $\Sigma_P : \mathcal{H} \to \mathcal{H}$ is defined as the operator satisfying $\langle f, \Sigma_P g\rangle = \mathbb{E}_{X\sim P}[f(X)g(X)]$ for all $f, g \in \mathcal{H}$. It can be expressed as $\Sigma_P = \mathbb{E}_P[k(X,\cdot) \otimes k(X,\cdot)]$, where $\otimes$ is the

tensor product operator, so that

$$\begin{aligned}\langle f, \mathbb{E}_P[k(X,\cdot)\otimes k(X,\cdot)]g\rangle &= \langle f, \Sigma_P g\rangle \\ &= \mathbb{E}_{X\sim P}[f(X)g(X)] \\ &= \mathbb{E}_{X\sim P}[\langle f, k(X,\cdot)\rangle\langle k(X,\cdot), g\rangle] \\ &= \mathbb{E}_{X\sim P}[\langle f\otimes g, k(X,\cdot)\otimes k(X,\cdot)\rangle_{\text{HS}}] \\ &= \mathbb{E}_{X\sim P}[\langle f, \big(k(X,\cdot)\otimes k(X,\cdot)\big)g\rangle]\end{aligned}$$

where $\langle\cdot,\cdot\rangle_{\text{HS}}$ is the Hilbert-Schmidt inner product.

A.3 Maximum Mean Discrepancy

A popular approach for two-sample tests involves leveraging kernel-based metrics to compare distributions. One such metric is the Maximum Mean Discrepancy (MMD), introduced in Gretton et al. (2012) to perform a kernel based two-sample test. MMD measures the distance between the mean embeddings of two distributions in the feature space induced by a kernel.

Definition A.1. *Let $\mathcal{H}_k$ be a reproducing kernel Hilbert space (RKHS) with corresponding kernel $k(\cdot,\cdot)$. This is, $f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}_k}$ Then, the MMD between two distributions P and Q is given by*

$$\text{MMD}_k(P,Q) := \sup_{f\in\mathcal{H}_k:\|f\|_{\mathcal{H}_k}\leq 1} \mathbb{E}_{X\sim P}\left[f(X)\right] - \mathbb{E}_{Y\sim Q}\left[f(Y)\right] \tag{11}$$

Given a sample $\mathbf{Z} = (X_1,...,X_n,Y_1,...Y_m)$, the minimum variance unbiased estimate for the MMD_k^2 is given by

$$\widehat{\text{MMD}}_k^2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{i'=1,i'\neq i}^{n} k(X_i,X_{i'}) + \frac{1}{m(m-1)}\sum_{j=1}^{m}\sum_{j'=1,j'\neq j}^{m} k(Y_j,Y_{j'}) - \frac{2}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m} k(X_i,Y_j). \tag{12}$$

A.4 Permutation Tests

We restrict ourselves to the two-sample testing framework where both distributions P and Q are unknown and only sample-access is available ($X_1,...,X_n \overset{\text{i.i.d.}}{\sim} P$, $Y_1,...,Y_m \overset{\text{i.i.d.}}{\sim} Q$). Recall that we define $\mathbf{Z} = (\mathbf{X},\mathbf{Y})$. As explained in Lehmann and Romano (2005, Chapter 15.2), the problem of two-sample testing $H_0\colon \mathbf{X} \overset{d}{=} \mathbf{Y}$ is equivalent to the problem of testing for exchangeability $H_0\colon \mathbf{Z} \overset{d}{=} \sigma\mathbf{Z}$ for permuted samples $\sigma\mathbf{Z} = (Z_{\sigma(1)},...,Z_{\sigma(N)})$ with permutation $\sigma\in\mathcal{G}_N$, where $\mathcal{G}_N$ be the symmetric group of $N := m+n$ elements

$$\mathcal{G}_N = \{\sigma : [N]\to[N] : \sigma \text{ is a permutation of } [N]\}.$$

Let $\tau : \mathbb{Y}^{n+m}\to\mathbb{R}$ be the statistic of interest. Theorem A.2 uses the fact that under the null hypothesis ($P=Q$), the data vector $\mathbf{Z}$ and any permuted version $\sigma\mathbf{Z}$ are equal in distribution (since all samples are i.i.d.). This theorem provides a method to compute a threshold for the test statistic using the $(1-\alpha)$-quantile of statistics computed on the permuted samples (see theorem statement for details).

Theorem A.2 (Theorem 2, Hemerik and Goeman (2018))**.** *Let $\mathbf{G} = (id, G_2,...,G_w)$ be the vector where $id\in\mathcal{G}_N$ is the identity and $G_2,...,G_w$ are elements from $\mathcal{G}_N$ drawn either uniformly at random without replacement from $\mathcal{G}_N\setminus\{id\}$ (with $w\leq|\mathcal{G}_N|$), or drawn with replacement from $\mathcal{G}$. Let $\tau^{(1)},...,\tau^{(w)}$ be the ordered statistics of the values $\tau(g\mathbf{Z})$ for $g\in\mathbf{G}$, and let*

$$q_{1-\alpha} := \tau^{(\lceil(1-\alpha)w\rceil)}$$

be the $(1-\alpha)$-quantile of these values. For the null hypothesis $H_0\colon \mathbf{Z} =^d g\mathbf{Z}$ for all $g\in\mathbf{G}$, it holds that

$$\mathbb{P}_{H_0,\mathbf{G}}\big(\tau(Z) > q_{1-\alpha}\big) \leq \alpha.$$

Theorem A.2 provides a near exact test that controls type-I error at level α. In this paper we focus on carefully designing statistics τ that can trade-off accuracy of estimating the likelihood ratio and the efficiency in terms of

number of samples. We review a specific type of aggregation for multiple testing that will allow us to circumvent hyper parameter tuning, specifically bandwidth and regularization values.

While we focus on the permutation method in this paper, we emphasize that an alternative line of work relying on the wild bootstrap exists (Fromont et al., 2012, 2013; Chwialkowski et al., 2014; Schrab et al., 2023; Pogodin et al., 2024).

A.5 Adaptive Testing Methods

It is often the case that several test statistics $\tau_1, ..., \tau_k$ are computed on a fixed sample $\mathbf{Z}$, for example, using different kernels and kernel bandwidths. These are then combined to potentially achieve more powerful tests. In this paper, we leverage the fuse method introduced in Biggs et al. (2023) and the aggregation method of Schrab et al. (2023). While AutoML techniques (Kübler et al., 2022) have also been studied for this purpose, we focus mostly on the fuse method due to its superior performance (see Figure 1).

Fuse. This method was introduced and studied by Biggs et al. (2023) for two-sample hypothesis tests with the MMD. Given a sample $\mathbf{Z}$, this method, for each permutation g, takes a softmax over the values of the individual statistics $\tau_1(\sigma\mathbf{Z}), ..., \tau_k(\sigma\mathbf{Z})$.

Definition A.3. *Let $\kappa > 0$, and $\mathcal{K} = \{\tau : \mathcal{Y}^{n+m} \to \mathbb{R}\}$ be a family of test statistics. Let π be a distribution over $\mathcal{K}$ that is either fixed independently of the data or is a function of the data that remains invariant under permutations of the combined sample $\mathbf{Z}$. This invariance property ensures that the fusing process respects the exchangeability of samples under the null hypothesis. We define the unnormalized and normalized fused statistics as*

$$\mathrm{FUSE}(\mathbf{Z}) := \frac{1}{\kappa} \log\left(\mathbb{E}_{\tau\sim\pi}\left[\exp\left(\kappa\tau(\mathbf{Z})\right)\right]\right),$$

where the statistics can potentially be normalized to account for differences in the scale or variance of different test statistics.

Following previous work, we let π be the uniform distribution over $\mathcal{K}$, giving equal weight to each considered test statistic. In practice, we follow the choice of $\kappa = \sqrt{n(n-1)}$ as used in the original paper (Biggs et al., 2023). We introduce in Algorithm 1 the permutation f-divergence test fusing over hyperparameters such as the kernel bandwidths and the regularization values.

Algorithm 1 introduces a test that combines the fuse and permutations approaches described above and the statistics defined in this section to compute a p-value for two-sample tests. The test receives two samples $\mathbf{X}$ and $\mathbf{Y}$, a statistic function derived from an f-divergence $\tau(\cdot,\cdot)$, that takes the concatenation of (potentially permuted) samples $\mathbf{Z}$ and a hyperparameter configuration $c \in \mathcal{C}$, where $\mathcal{C}$ is a set of hyperparameter configurations, a significance value α, and a number of permutations B. The test first calculates the true statistic for all hyperparameter configurations (Line 2), and fuses them into one statistic T^{fuse} (Line 1). Similarly, computes the statistic on all permutations and all configurations, and fusing the statistics over configurations for all permutations (Line 7) and for each permutation fuses the statistics across different configurations (Line 10). Finally, it uses Theorem A.2 in Line 12 to calculate the p-value and return a binary decision if the p-value is smaller than the significance value α. Note that, if $\mathcal{C}$ contains a single configuration, then Algorithm 1 reduces to a standard permutation test based on the statistic τ.

MMDAgg. Aggregating test statistics with different magnitude scales can be challenging, as methods that combine statistics directly (e.g., the soft-maximum in MMD-Fuse) can be dominated by the statistics with the largest scales. We instead the aggregation procedure introduced by Schrab et al. (2023, Algorithm 1), that circumvents normalization by relying on multiple testing.

We emphasize that aggregated tests have been considered in various other contexts, including independence testing (Albert et al., 2022), goodness-of-fit testing (Schrab et al., 2022a), efficient testing (Schrab et al., 2022b), robust testing (Schrab and Kim, 2025), imprecise credal testing (Chau et al., 2025), and learning kernels (Zhou et al., 2025). See Schrab (2025a,b,c) for an overview on these.

Algorithm 1: Single f-divergence two-sample test

Input: Samples $\mathbf{X} = (x_1, \ldots, x_n) \overset{\text{iid}}{\sim} P$, $\mathbf{Y} = (y_1, \ldots, y_m) \overset{\text{iid}}{\sim} Q$, significance level $\alpha \in (0, 1)$, list of hyperparameter configurations $\mathcal{C}$, number of permutations B, f-divergence estimator $\tau : \mathcal{Y}^{m+n} \times \mathcal{C}$ that receives a sample as the first argument and a hyperparameter configuration as the second argument, Fuse function.

Output: Boolean decision: 1 (rejecting H_0), 0 otherwise.

1 $\mathbf{Z} \leftarrow (\mathbf{X}, \mathbf{Y})$ `// Combine samples.`
2 $\mathbf{T} \leftarrow \{\tau(\mathbf{Z}, c) \mid c \in \mathcal{C}\}$ `// Compute true statistics.`
3 $T^{\text{fuse}} \leftarrow \text{Fuse}(\mathbf{T})$ `// Compute the fused observed statistic.`
4 $\mathcal{G} \leftarrow \{\sigma_1, \ldots, \sigma_B\}$ `// Generate permutations on` $n+m$.
5 **for** $\sigma \in \mathcal{G}$ **do**
6 **for** $c \in \mathcal{C}$ **do**
7 $T_{\sigma,c} \leftarrow \tau(\sigma\mathbf{Z}, c)$ `// Compute statistic on all configurations.`
8 **end**
9 $\mathbf{T}_\sigma \leftarrow \{T_{\sigma,c} \mid c \in \mathcal{C}\}$ `// Collect statistics for permutation` σ.
10 $T_\sigma^{\text{fuse}} \leftarrow \text{Fuse}(\mathbf{T}_\sigma)$ `// Compute fused statistic for permutation` σ.
11 **end**
12 $p_{\text{val}} \leftarrow \left(1 + \left|\{\sigma_i \in G : T_\sigma^{\text{fuse}} \geq T^{\text{fuse}}\}\right|\right)/(B+1)$ `// Compute` p`-value.`
13 **return** $\mathbb{1}(p_{\text{val}} \leq \alpha)$

The procedure considers a finite set of hyperparameter configurations, $\mathcal{C} = \{c_1, \ldots, c_{|\mathcal{C}|}\}$, and uses permutations. f-Agg leverages this method to aggregate over different f-divergence statistics. Essentially, a multiple test is run with the level of each single f-divergence having a correction applied to it. This correction is estimated using the data and extra permutations in order for the test to be as powerful as possible. The method is summarized in Algorithm 2.

To compute the threshold in Line 15 of Algorithm 2, some additional permuted statistics are computed exactly as in Line 4 to Line 13 with a new set of permutations $\mathcal{G}_2 = \{\sigma_{B+1}, \ldots, \sigma_{2B}\}$. Then, these are used to estimate the probability under the null by a Monte-Carlo approximation. Finally, a bisection method is performed to determine the largest threshold for which the estimated type I error remains controlled at level α. See Schrab et al. (2023, Algorithm 1) for details.

A.6 Fuse implementation details

We now describe some specific implementation details of Algorithm 1 that we use throughout the experimental section.

1. We use only perform tests with the gaussian kernel, $k(x, y) = \exp(\|x - y\|^2/\gamma^2)$ where γ is the kernel bandwidth. We calculate five different bandwiths using five bandwidths chosen as the uniform discretizations of the interval between half the 5% and twice the 95% quantiles of $\{\|z - z'\|_2 : z, z' \in \mathbf{Z}\}$, as in the original MMD-Fuse paper (Biggs et al., 2023).

2. We fuse over the regularization parameter λ, for $\lambda \in \{0.0001, 0.001, 0.01, 0.1, 1.0, 10.0\}$.

3. f-divergences can be asymmetric, i.e. $D_f(P||Q) \leq D_f(Q||P)$. Hence, for samples $\mathbf{X}, \mathbf{Y}$ and configuration of hyperparameters $c \in \mathcal{C}$, we use the statistic $\tau_{\text{final}}((\mathbf{X}, \mathbf{Y}), c) = \max\{\tau((\mathbf{X}, \mathbf{Y}), c), \tau((\mathbf{Y}, \mathbf{X}), c)\}$, where τ is an f-divergence estimator as described in Section 2.

Complexity. Computing the likelihood ration $r(x)$ using expression Equation (4) has a time complexity of $O(m^3 + mn + n^2)$, coming from matrix inversion and matrix multiplication. For a given f-divergence, we compute this ratio for all hyperparameter configurations $c \in \mathcal{C}_f$ and all permutations $\sigma \in \mathcal{G}$, so the time complexity of a test is given by $O\left(|\mathcal{G}||\mathcal{C}_f|(m^3 + mn + n^2)\right)$.

Algorithm 2: f-Agg: Aggregated f-divergence two-sample test

Input: Samples $\mathbf{X} = (x_1, \ldots, x_n) \overset{\text{iid}}{\sim} P$, $\mathbf{Y} = (y_1, \ldots, y_m) \overset{\text{iid}}{\sim} Q$, significance level $\alpha \in (0, 1)$, number of permutations B, list $\mathcal{F}$ of functions f with associated f-divergence estimators $\tau_f : \mathcal{Y}^{m+n} \times \mathcal{C}_f$ that receives a sample as the first argument and a hyperparameter configuration as the second argument, lists $\mathcal{C}_f$ of hyperparameter configurations for each $f \in \mathcal{F}$, Fuse function.

Output: Boolean decision: 1 (rejecting H_0), 0 otherwise.

$\mathbf{Z} \leftarrow (\mathbf{X}, \mathbf{Y})$ `// Combine samples.`
$\sigma_0 \leftarrow \text{Id}$ `// Set` σ_0 `to be the idendity permutation.`
$\mathcal{G} \leftarrow \{\sigma_0, \sigma_1, \ldots, \sigma_B\}$ `// Generate permutations on` $n + m$.
for $f \in \mathcal{F}$ **do**
  for $\sigma \in \mathcal{G}$ **do**
    for $c \in \mathcal{C}_f$ **do**
      $T_{f,\sigma,c} \leftarrow \tau_f(\sigma\mathbf{Z}, c)$ `// Compute statistic on all configurations.`
    end
    $\mathbf{T}_{f,\sigma} \leftarrow \{T_{f,\sigma,c} \mid c \in \mathcal{C}_f\}$ `// Collect statistics for permutation` σ.
    $T^{\text{fuse}}_{f,\sigma} \leftarrow \text{Fuse}(\mathbf{T}_{f,\sigma})$ `// Compute fused statistic for permutation` σ.
  end
  $T^{\text{fuse}}_f \leftarrow T^{\text{fuse}}_{f,\sigma_0}$ `// Collect observed fused statistics.`
end
$\text{Test}(u) \leftarrow \mathbb{1}(T^{\text{fuse}}_f > \text{Quantile}_{1-u}\{T_{f,\sigma} \mid \sigma \in \mathcal{G}\} \text{ for some } f \in \mathcal{F})$ `// Define aggregated test.`
$u_\alpha \leftarrow \sup\{u \in [0, 1] : \mathbb{P}_{H_0}(\text{Test}(u) = 1) \leq \alpha\}$ `// Compute via bisection with extra permutations.`
return $\text{Test}(u_\alpha)$

A.7 Formulation of KALE and DrMMD

The KALE statistic introduced first by Nguyen et al. (2010) and later studied by Arbel et al. (2020) for generative modelling. Arbel et al. (2020). The KALE divergence corresponds to the following kernel regularized KL divergence

$$\begin{aligned}\text{KALE}(P, Q) &:= (1 + \tau) \sup_{h \in \mathcal{H}} \left\{ 1 + \int h dP - \int \exp(h) dQ - \frac{\tau}{2} \|h\|^2_{\mathcal{H}} \right\}. \\ &= \min_{f > 0} \int (f(\log f - 1) + 1) dQ + \frac{1}{2\tau} \left\| \int f(x) k(x, \cdot) dQ \right\|^2_{\mathcal{H}}.\end{aligned}$$

where the expression for the dual form in the second equation is from Glaser et al. (2021, Equation 6).

DrMMD, which is a regularized χ^2 f-divergence (or de-regularized MMD), was first proposed by Harchaoui et al. (2007) in a paper proposing an asymptotic test. It can be expressed either in its regularized variational form, or by its closed form expression

$$\begin{aligned}\text{DrMMD}(P, Q) &:= (1 + \tau) \|(\Sigma_Q + \tau I)^{-1/2} (\mu_P - \mu_Q)\|^2_{\mathcal{H}} \\ &= (1 + \tau) \sup_{h \in \mathcal{H}} \left\{ \int h dP - \int \left(\frac{h^2}{4} + h \right) dQ - \frac{\tau}{4} \|h\|^2_{\mathcal{H}} \right\}.\end{aligned}$$

See Section 5 for details on general regularized f-divergences.

While DrMMD and KALE have also been studied in the context of gradient flows and particle descent by Chen et al. (2024a) and Glaser et al. (2021), these papers focus on the DrMMD/KALE divergences themselves as measures for optimization and modeling.

In our work, we shift the focus of these divergences to hypothesis testing. Specifically, we use these divergences to construct permutation-based DrMMD/KALE non-asymptotic tests, with adaptation over the kernel and regularization parameter (see Appendix A.4). These adaptive, non-asymptotic tests have not previously been considered in the literature, and we compare them against our proposed non-regularized f-divergence tests in our experiments.

For a fixed kernel and fixed regularization parameter, the DrMMD and KALE discrepancies are computed exactly as in their original papers. Our methodological novelty is first in developing non-asymptotic, permutation tests based on these discrepancies, contrasting with the original asymptotic test proposed for DrMMD, and second, leveraging theoretical insights and techniques from our proposed method to construct adaptive DrMMD/KALE statistics that are robust to the choice of kernel and regularization.

B EXPERIMENTAL DESIGN CONSIDERATIONS

We begin by presenting experiments that evaluate the performance of various f-divergences and regularized f-divergences, illustrating the flexibility of the proposed framework. Next, we present an experiment to build intuition about the regularization parameter λ for f-divergence testing with estimator in Equation (3). We also show the advantage of leveraging the variational formulation in Equation (3), even when an estimator for the likelihood ratio $\frac{dP}{dQ}(x)$ is readily available.

B.1 Testing with Various f-Divergences

To showcase the flexibility of the proposed method to accommodate regularised and non-regularised f-divergences we ran the Expo-1D tests (Section 6.2) and perturbed uniform tests (Section 6.1) using different regularized and non-regularized f-divergence based statistics.

In Figure 4 we observe that MMD still outperforms all divergences in this setting in dimensions $d = 1, 2$. The performance decreases for all methods in two dimensions. However, the rejection rate of all but the reversed KL converges to 1 as the perturbation amplitude increases in dimension $d = 1$.

In Figure 5 we see that the KL-based test leveraging the likelihood ratio estimator in Equation (4) has similar performance to the DrMMD based test, and these two outperform all other f-divergence based tests.

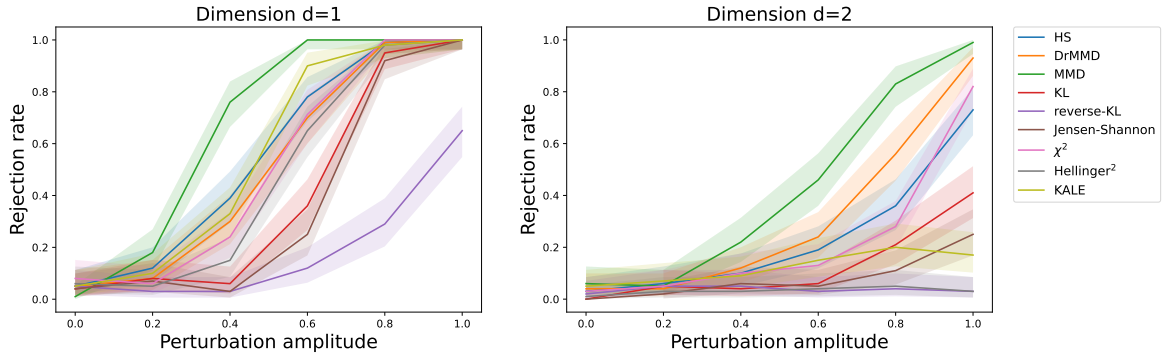

Figure 4: Performance on perturbed d-dimensional uniform alternatives with varying perturbation amplitudes for a diverse set of f-divergences. MMD outperforms all divergences in this setting. The performance decreases for all methods in two dimensions.

B.2 Effect of regularization parameter λ and the importance of adaptation

By design, larger values of regularization make the estimation of the (regularized) likelihood more tractable but introduce bias. However, in our tests we do not need to compromise accuracy nor smoothness since we adapt over the parameter λ as detailed in Appendix A.5. Therefore, all the experiments in the manuscript are currently run with adaptation over λ, explaining the high power of our proposed method.

To showcase the importance of varying and adapting over the regularisation parameter, we ran the proposed test for the Expo-1D experiment to compare rejection rates across diferent λ values for the KL and χ^2 divergence based tests. In this experiment, we still adapt over five kernel bandwidths but do not adapt over the regularization parameter λ.

We show the results in Figure 6. This experiment highlights that optimal detection rates depend on matching the regularization value to the specific alternative. Note that 'easier' cases ($k = 100$, right panel) require lower

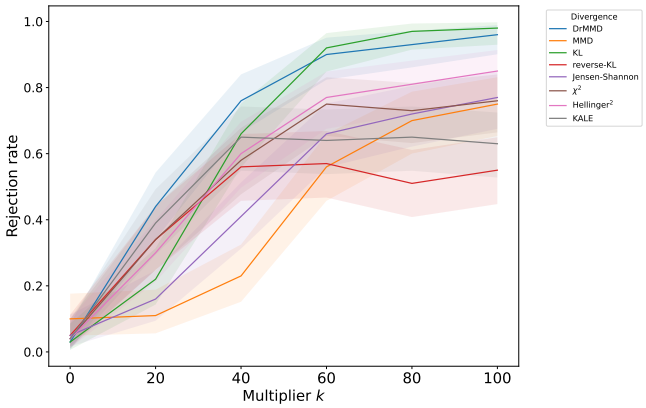

Figure 5: DrMMD (Regularized χ^2) and KL are the top-performing divergences overall in this setting. MMD improves in smooth settings (large k). For smaller k, the regularized estimators, KALE (regularized log-ratio) and DrMMD (regularized ratio) have superior performance.

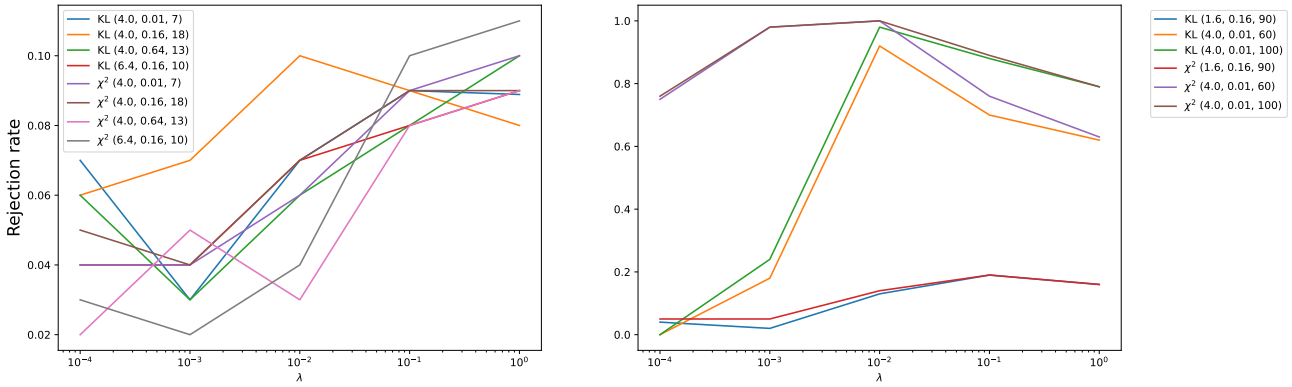

Figure 6: Rejection rate of Algorithm 2 for different alternatives, adapting only over the kernel bandwidth. Each line corresponds to a specific f-divergence on a given Expo-1D alternative, with hyperparameters (μ, σ, k) in the legend. For details see Section 6.2. Alternatives closer to the null (left) achieve higher power with stronger regularisation ($\lambda = 1.0$), while larger departures from the null (right) require lower regularisation.

regularization ($\lambda \approx 0.001$), while harder alternatives ($k = 10$, left panel) achieve higher power with stronger regularization ($\lambda = 1.0$). These results justify our method's need for adaptivity.

B.3 Comparison of Direct f-divergence estimator vs. Variational Estimators

In Section 2 we introduce an estimator based the variational formulation of f-divergences for its flexibility in handling regularized f-divergences. However, a direct plug-in approach in Equation (1) is a natural baseline for standard f-divergences.

To address this, we ran new experiments to compare the test power of the direct estimator against our variational approach. For clarity, let us introduce:

$$\begin{aligned} \text{Direct:} \qquad & \widehat{D}_d(P||Q) = \mathbb{E}_Q[f(dP/dQ)] \\ \text{Variational:} \qquad & \widehat{D}_v(P||Q) = \mathbb{E}_P[f'(dP/dQ)] - \mathbb{E}_Q[f^*(f'(dP/dQ)] \end{aligned}$$

Setting. For testing, we used the Expo-1D dataset as in Section 6.2, varying the parameter k to control the departure from the null hypothesis ($k = 0$ is the null; $k = 100$ is the maximum departure). We compute the

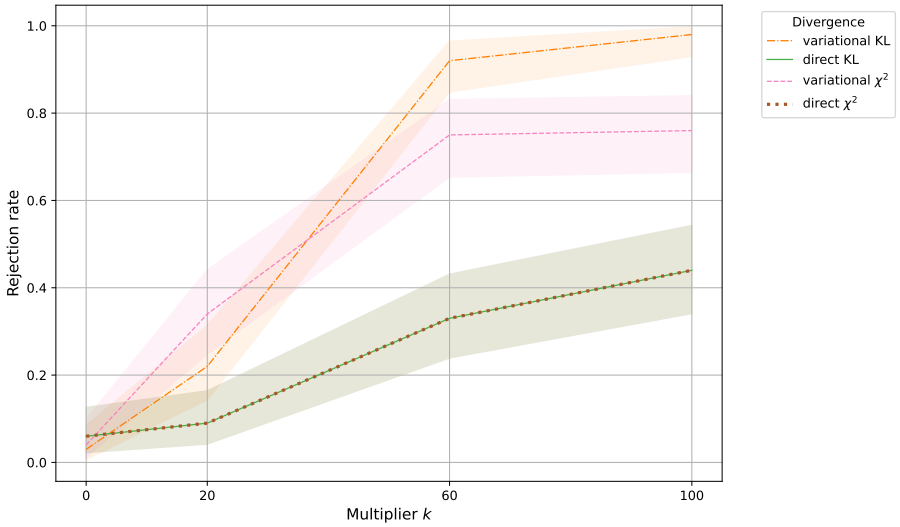

Figure 7: Rejection rate for direct f-divergence estimator vs. variational f-divergence estimator based two sample tests on the Expo-1D dataset.

test statistic based on $\max(\widehat{D}(P||Q), \widehat{D}(Q||P))$ using either the direct or variational estimators. We run the experiments with $n = 1000$ samples from P and Q.

We observe in Figure 7 that the variational approach achieves much higher power than the direct approach, which justifies the use of the proposed variational method.

C EXPERIMENTAL DETAILS

This section provides additional details on the experimental settings of Section 6: the Expo-1D experiment (Appendix C.1), differential privacy experiments (Appendix C.2), and machine unlearning (Appendix C.3).

C.1 Expo-1D

KALE performance in Figure 3. In Section 6 we noted that KALE outperforms MMD-based methods for small values of k but after $k = 40$ its performance grows more slowly than DrMMD and MMD. Figure 8 below shows how the log-ratio of the specific tested densities grows more slowly than the ratio, potentially making it harder for methods that rely on the ratio to continue improving without more samples.

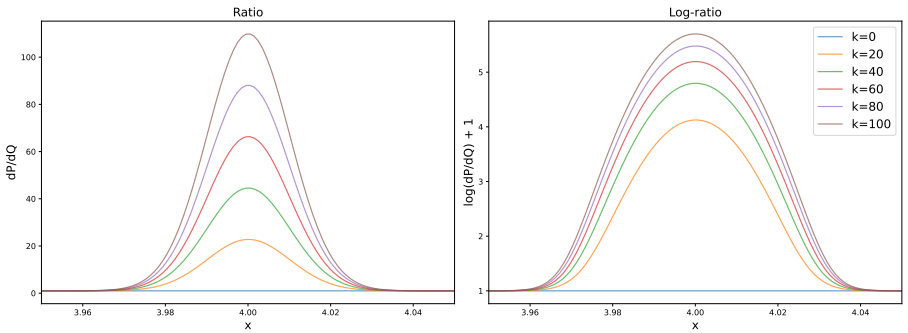

Figure 8: Ratio (left) and Log-Ratio (right) of densities involved in the Expo-1D test.

Comparison to Grosso and Letizia (2025). In Table 5, we compare the proposed f-divergence tests against the results from Grosso and Letizia (2025) using their experimental configuration at a significance level of $\alpha = 0.023$.

The methodology in the prior work involved computing the KALE statistic over 4000 trials, using 200'000 samples from the null distribution; the number of samples from the alternative distribution was not specified. For our f-divergence tests, we use 5000 samples from each distribution and report the average statistical power and confidence intervals over 100 trials. We compare our results against the best-performing aggregation method reported by Grosso and Letizia (2025).

The comparison shows that an f-divergence statistic paired with the fuse aggregation method (smax-t in their work) achieves greater statistical power in nearly all settings, using significantly fewer samples.

These results also align with the behavior illustrated in Figure 8, where KALE performs best for small density ratios where the log-ratio is large, while MMD becomes more powerful as the ratio increases (e.g., for multiplier $k = 90$).

Table 5: Performance of different methods on the Expo-1D dataset on the same parameter configuration as previous work. Each column presents a parameter configuration (μ, σ, k) that configures the alternative $n_a = n_0(x) + k\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x-\mu^2}{2\sigma^2}\right)$. We confirm the intuition in Figure 8 that KALE performs better when the density ratio is small, but once the multiplier is large ($k = 90$) MMD is a better statistic.

(μ, σ, k)	(1.6, 0.16, 90.0)	(4.0, 0.01, 7.0)	(4.0, 0.16, 18.0)	(4.0, 0.64, 13.0)	(6.4, 0.16, 10.0)
HS	0.250 [0.169, 0.347]	0.020 [0.002, 0.070]	0.030 [0.006, 0.085]	0.030 [0.006, 0.085]	0.030 [0.006, 0.085]
HS-sig	0.390 [0.294, 0.493]	0.020 [0.002, 0.070]	0.030 [0.006, 0.085]	0.010 [0.000, 0.054]	0.040 [0.011, 0.099]
DrMMD	0.110 [0.056, 0.188]	0.220 [0.143, 0.314]	0.190 [0.118, 0.281]	0.040 [0.011, 0.099]	0.320 [0.230, 0.421]
KALE	0.320 [0.230, 0.421]	**0.330** [0.239, 0.431]	**0.420** [0.322, 0.523]	**0.470** [0.369, 0.572]	0.540 [0.437, 0.640]
MMD	**0.620** [0.517, 0.715]	0.030 [0.006, 0.085]	0.030 [0.006, 0.085]	0.030 [0.006, 0.085]	0.040 [0.011, 0.099]
Grosso and Letizia (2025)	0.008 ±0.001	0.103 ±0.007	0.012 ±0.002	0.32 ±0.01	**0.66** ± 0.01

C.2 Differential Privacy

Audited mechanisms. We consider the 4 non-private variants of the sparse vector technique (SVT) introduced by Lyu et al. (2017) as algorithms 3–6. This mechanism for releasing a stream of queries compares each query value against a threshold. Each algorithm returns certain outputs for a maximum number of queries c. SVT4 satisfies $(\frac{1+6c}{4})$-DP, and SVT3, SVT5, and SVT6 do not satisfy ε-DP for any finite ε.

The mechanisms mean1 and mean2 were introduced in Kong et al. (2024) and are non-private mechanisms that take the average of n numbers (for any finite ε) and add Laplace noise. Mean1 violates the guarantee by accessing the private number of points n, and mean2 one privatizes the number of points to estimate the scale of the noise added to the mean statistic but the mean itself is computed using the non-private number of points.

For each mechanism we fixed the neighboring test datasets specified in Table 6.

Table 6: Adjacent Datasets D and D' used to test differential privacy guarantees.

Mechanism	D	D'
Mean mechanisms	$\{1\}$	$\{1, 0\}$
SVT mechanisms	$\{0, 1\}$	$\{0, 1, 0\}$

C.3 Machine Unlearning

Model training and data. For this experiment we used the CIFAR10 dataset (Krizhevsky et al., 2009). We train a simple convolutional neural network with two convolutional layers, with 16, and 32 features respectively. This is followed by two dense layers that output logits for 10 categories. The model is trained to minimize the cross entropy loss. We train for 10 epochs with a batch size of size 128 on 90% of the data, and leave 10% for validation. We target the last two dense layers for modification when applying unlearning techniques. The forget set is constructed by sampling 10 different images at random after shuffling the dataset.

Unlearning methods in Section 6. There is a variety of unlearning methods with varying characteristics and different applications. They use different techniques that include gradient ascent in the forget set, finetuning

in the retain set, pruning, saliency or data correlation metrics. We consider the following algorithms which we believe capture a large portion of these techniques. By no means we optimize the hyperparameter or replicate exactly their algorithms.

1. Selective Synaptic Dampening (SSD) by Foster et al. (2024): The core idea of SSD is to "dampen" the parameters that are most important for the forgotten data, pulling them back toward their original state before training. It avoids damaging the model's overall performance by protecting parameters that are also important for the retain data. It achieves this by:
   - Calculating parameter importance for both the forget and retain sets (approximated by the squared gradient of the loss)
   - Creating a "dampening factor" for each parameter that is high if it's important for the forget set but not the retain set.
   - Minimizing a loss function that penalizes changes from the original model's weights, weighted by these dampening factors.
2. Pruning: Pruning methods address unlearning by pruning the most influential neurons for the forget set, measured by mean activation magnitude when evaluated on the forget set.
3. Random Label: These methods erase influence of the forget set by training on randomly labeled forget data, and then finetuning on retain data.
4. Finetuning: these models perform finetuning on the retain set.
5. Retraining for a different number of iterations.
6. Retrain from scratch: Retrains the model on retain data with the exact same train configuration as the original model.
7. Retrain less: Retrains the model from scratch on retain data for only 5 epochs, other parameters are fixed.
8. Retrain batch: Retrains the model from scratch on retain data with a batch size of 256 (instead of 128).

To test the unlearning capability of the above methods we first generate 5000 models from each distribution that we split in 10 groups of 500 samples each, for 10 repetitions of each test.

Three sample unlearning-evaluation test. We use the following relative similarity test introduced in Bounliphone et al. (2016) in the context of model selection of generative models.

Let P, Q, and R be three distributions over the same domain. Given samples $\mathbf{X} = (x_1, \dots, x_m) \overset{\text{iid}}{\sim} P$, $\mathbf{Y} = (y_1, \dots, y_n) \overset{\text{iid}}{\sim} Q$, and $\mathbf{W} = (w_1, \dots, w_r) \overset{\text{iid}}{\sim} R$ such that $P \neq Q$, $P \neq R$, we consider the hypothesis test with null hypothesis $H_0 : MMD(P, Q) \leq \text{MMD}_k(P, R)$ against the alternative $H_1 := \text{MMD}_k(P, Q) > \text{MMD}_k(P, R)$ at significance level α, where MMD_k is the population MMD for the RKHS with corresponding kernel k.

Theorem 2 in Bounliphone et al. (2016) establishes that if $\mathbb{E}[k(x_i, x_j)] < \infty$ and $\mathbb{E}[k(y_i, y_j)] < \infty$ and $\mathbb{E}[k(x_i, y_j)] < \infty$, then the joint distribution of the unbiased empirical estimators of $\text{MMD}_k(P, Q)$ and $\text{MMD}_k(P, Q)$, ($\widehat{\text{MMD}}_k(P, Q)$ and $\widehat{\text{MMD}}_k(P, Q)$ respectively) satisfy:

$$\sqrt{m}\left(\begin{pmatrix}\widehat{\text{MMD}}_k^2(\mathbf{X}, \mathbf{Y}) \\ \widehat{\text{MMD}}_k^2(\mathbf{X}, \mathbf{W})\end{pmatrix} - \begin{pmatrix}\text{MMD}_k^2(P, Q) \\ \text{MMD}_k^2(P, R)\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix}0 \\ 0\end{pmatrix}, \begin{pmatrix}\sigma_{PQ}^2 & \sigma_{PQ,PR} \\ \sigma_{PQ,PR} & \sigma_{PR}^2\end{pmatrix}\right) \tag{13}$$

The variance terms σ_{PQ}^2, σ_{PR}^2, $\sigma_{PQ,PR}^2$ can be estimated using samples through Equations (2) and (7) in the original paper by Bounliphone et al. (2016). Consequently, the p-value p for the test H_0 against H_1 can be calculated using the following one-sided inequality:

$$p \leq \Phi\left(-\frac{\widehat{\text{MMD}}^2(\mathbf{X}, \mathbf{Y}) - \widehat{\text{MMD}}^2(\mathbf{X}, \mathbf{W})}{\sqrt{\sigma_{PQ}^2 + \sigma_{PR}^2 - 2\sigma_{PQ,PR}}}\right), \tag{14}$$

where Φ is the cumulative distribution function of the standard normal distribution. The test is performed by comparing p to the significance level α.

D DIRECT OPTIMIZATION OF REGULARIZED HOCKEY-STICK WITNESS

In this section, we derive a direct optimization of the witness function of the regularized Hockey-Stick divergence.

Optimization via semidefinite programming Estimating Equation (8) is hard in practice, when only finite samples $X_1, ..., X_n \sim P$ and $Y_1, ..., Y_n \sim Q$ from the distributions P and Q are available

The primary challenge lies in the constraints $0 \leq g(x) \leq 1$, which must hold for all x. since enforcing these constraints over an infinite set of points is intractable. Simply enforcing it on the sample points $\{X_i\}_i$ and $\{Y_i\}_i$ provides no guarantee that the constraint holds elsewhere.

Several alternatives present their own challenges:

- Reducing the class of functions g to generalized linear models (GLMs) (e.g. restricting g to the form $g(x) = \langle \phi(x), g \rangle$), introduces a non-convex loss function to map a solution g's outputs to the bounded range $[0, 1]$, and making it hard to assert convergence guarantees about the solution.
- Replacing the original range constraint with a finite number of constraints over finite samples leads to either unfeasible solutions that violate the range constraint outside of the samples, or computationally impractical optimization problems for large sample sizes.

Instead, inspired by work on non-negative optimization (Marteau-Ferey et al., 2020) we first formulate an optimization problem with non-negativity constraints, we reparameterize the witness function as $g(x) = \phi(x)^T A \phi(x)$, where $\mathbf{A}$ is a hermitian positive semidefinite linear operator ($A \succcurlyeq 0$), $\phi(x)$ is the empirical feature map associated with the kernel $k(\cdot, \cdot)$, data $z_i = \begin{cases} x_i & \text{if } i =, 1, ..., n \\ y_{i-n} & \text{otherwise.} \end{cases}$. Recall that the empirical feature map for $x \in \mathcal{X}$ is given by $\phi(x) = \mathbf{V}^\top v(x)$, $v(x) = (k(x, z_i))_{i=1}^n$, and $\mathbf{V}$ is the Cholesky decomposition of $\mathbf{K}$, i.e., $\mathbf{K} = \mathbf{V}^\top \mathbf{V}$.. This formulation inherently ensures $g(x) \geq 0$ for all x and enables efficient optimization.

Theorem D.1. *The semidefinite program defined by*

$$\textbf{\textit{Primal 1:}} \quad \min_{\mathbf{A} \succcurlyeq 0} \quad \frac{e^\varepsilon}{n} \sum_{i=1}^n \phi(y_i)^T A \phi(y_i) - \frac{1}{n} \sum_{i=1}^n \phi(x_i)^T A \phi(x_i) + \tau \|A\|_F$$

has a unique solution A^* that can be written as

$$A^* = \sum_{i=1}^{2n} B_{ij} \phi(z_i) \phi(z_j)^T, \quad \textit{for some matrix} \quad B \in \mathbb{R}^{2n \times 2n}$$

where $\phi(x) = \mathbf{V}^\top v(x)$, $\mathbf{V}$ is the Cholesky decomposition of $\mathbf{K}$, i.e., $\mathbf{K} = \mathbf{V}^\top \mathbf{V}$, and $v(x) = (k(x, z_i))_{i=1}^n$ is the empirical feature map and

$$z_i = \begin{cases} x_i & \textit{if } i =, 1, ..., n \\ y_{i-n} & \textit{otherwise.} \end{cases}$$

Further, this problem can be efficiently solved because as we show below, the dual has only $2n$ variables, instead of $4n^2$, and for this specific case can be optimized using accelerated proximal splitting methods as FISTA.

Lemma D.2. *(a) Let $\mathbf{K} \in \mathbb{R}^{n \times n}$ be the kernel matrix, i.e., $\mathbf{K}_{i,j} = k(z_i, z_j)$, where z_i is defined in Theorem D.1. Let $\mathbf{V}$ be the Cholesky decomposition of $\mathbf{K}$, i.e., $\mathbf{K} = \mathbf{V}^\top \mathbf{V}$.*

Then, the dual of Primal 1 is given by

$$\inf_{\alpha_x, \alpha_y \in \mathbb{R}^n} M(\|\alpha_x + \mathbb{1}/n\|_1 + \|\alpha_y - \frac{e^\varepsilon}{n} \mathbb{1}\|_1) + \frac{1}{2\tau} \|[\mathbf{V} Diag(\alpha) \mathbf{V}^\top]_-\|_F^2. \tag{15}$$

(b) If α^* is a solution of Equation (15)*, then*

$$\mathbf{B} = \tau^{-1}\mathbf{V}^{-1}[\mathbf{V}Diag(\alpha^*)\mathbf{V}^\top]_-\mathbf{V}^{-\top} \tag{16}$$

is a solution for the problem ***Primal 1****.*

Proof. Define the loss function L on $2n$ variables

$$L(x_1, ..., x_n, y_1, ..., y_n) = \frac{e^\varepsilon}{n}\sum_{i=1}^n g(y_i) - \frac{1}{n}\sum_{i=1}^n g(x_i).$$

It follows from Theorem 2 in Marteau-Ferey et al. (2020) that the dual problem of Primal 1 with objective

$$\min_{\mathbf{A}\succcurlyeq 0} \quad L(x_1, ...x_n, y_1, ..., y_n) + \tau\|A\|_F,$$

is defined by

$$\sup_{\alpha\in\mathbb{R}^n} -L^*(\alpha) - \frac{1}{2\tau}\|[\mathbf{V}\mathrm{Diag}(\alpha)\mathbf{V}^\top]_-\|_F^2.$$

replacing L^* by the conjugate in Lemma D.3 yields the desired result. □

Lemma D.3. *Let $L : [-M, M]^{2n} \to \mathbb{R}$ be a loss function defined as*

$$L(x_1, ..., x_n, y_1, ..., y_n) = \frac{e^\varepsilon}{n}\sum_{i=1}^n g(y_i) - \frac{1}{n}\sum_{i=1}^n g(x_i). \tag{17}$$

Then its convex conjugate $L^* : \mathbb{R}^{2n}(\alpha) \to \mathbb{R}$ is defined as

$$L^*(\alpha_x, \alpha_y) = M(\|\alpha_x + \mathbb{1}/n\|_1 + \|\alpha_y - \frac{e^\varepsilon}{n}\mathbb{1}\|_1) \tag{18}$$

Proof. The convex conjugate $L^*(\alpha)$ for function $L(x)$ is defined as:

$$L^*(\alpha) = \sup_{x\in[-M,M]^{2n}} (\langle\alpha, x\rangle - L(x))$$

Substituting $L(x)$ into the definition and grouping terms by each x_i:

$$\begin{aligned}
L^*(\alpha) &= \sup_{x\in[-M,M]^{2n}} \left(\sum_{i=1}^{2n}\alpha_i x_i - \left(\frac{e^\varepsilon}{n}\sum_{i=1}^n x_i - \frac{1}{n}\sum_{i=n+1}^{2n} x_i\right)\right) \\
&= \sup_{x\in[-M,M]^{2n}} \left(\sum_{i=1}^{n}\alpha_i x_i - \frac{e^\varepsilon}{n}\sum_{i=1}^n x_i + \sum_{i=n+1}^{2n}\alpha_i x_i + \frac{1}{n}\sum_{i=n+1}^{2n} x_i\right) \\
&= \sup_{x\in[-M,M]^{2n}} \left(\sum_{i=1}^{n}\left(\alpha_i - \frac{e^\varepsilon}{n}\right)x_i + \sum_{i=n+1}^{2n}\left(\alpha_i + \frac{1}{n}\right)x_i\right)
\end{aligned}$$

Since the objective is a sum of separable terms, the supremum of the sum is the sum of the suprema:

$$L^*(\alpha) = \sum_{i=1}^n \sup_{x_i\in[-M,M]}\left(\alpha_i - \frac{e^\varepsilon}{n}\right)x_i + \sum_{i=n+1}^{2n}\sup_{x_i\in[-M,M]}\left(\alpha_i + \frac{1}{n}\right)x_i$$

For any constant c, the solution to $\sup_{z\in[-M,M]}(c\cdot z)$ is $M\cdot|c|$. Applying this, the supremum for the first term is $M\left|\alpha_i - \frac{e^\varepsilon}{n}\right|$ and for the second is $M\left|\alpha_i + \frac{1}{n}\right|$.

Substituting these back and simplifying gives the final result in both summation and vector notation:

$$\begin{aligned} L^*(\alpha) &= \sum_{i=1}^{n} M \left|\alpha_i - \frac{e^\varepsilon}{n}\right| + \sum_{i=n+1}^{2n} M \left|\alpha_i + \frac{1}{n}\right| \\ &= M \left(\sum_{i=1}^{n} \left|\alpha_i - \frac{e^\varepsilon}{n}\right| + \sum_{i=n+1}^{2n} \left|\alpha_i + \frac{1}{n}\right| \right) \\ &= M \left(\left\| \alpha_y - \frac{e^\varepsilon}{n}\mathbb{1} \right\|_1 + \left\| \alpha_x + \frac{1}{n}\mathbb{1} \right\|_1 \right). \end{aligned}$$

□

Limitations of the Direct Optimization Approach. In practice, we do not use the direct optimization approach described above. The RKHS function class it employs (see Equation (10)) is not suited for representing the potentially discontinuous Hockey-Stick witness function from Equation (8).

Figure 9 empirically illustrates this limitation using two Gaussian distributions, $P = \mathcal{N}(0, 1)$ and $Q = \mathcal{N}(3, 1)$. The figure shows that the direct method fails to capture the sharp, non-linear structure of the true witness function, yielding an overly smooth approximation. In contrast, the threshold-based method from Section 4 provides a much better fit. We present this direct method nonetheless, as it could be promising for future work exploring more suitable function classes.

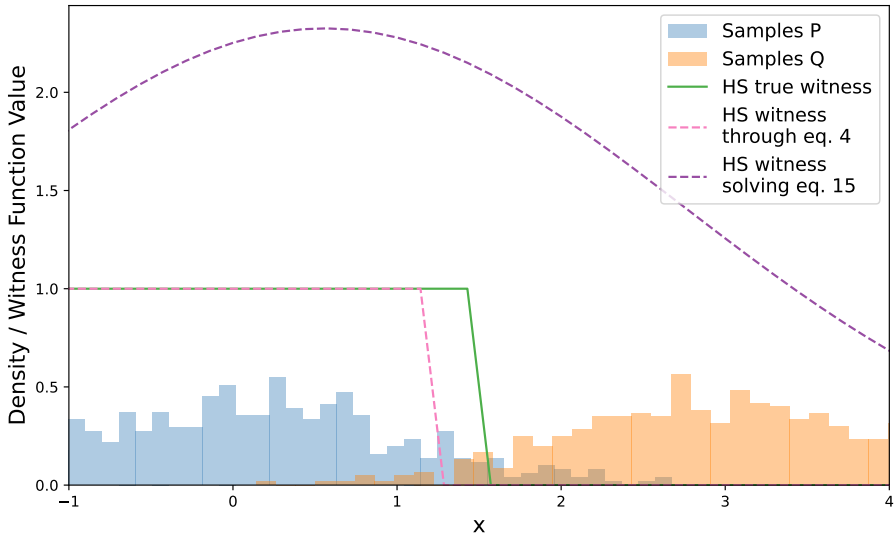

Figure 9: Hockey-Stick witness function plots for two Gaussian distributions, $P = \mathcal{N}(0, 1)$ and $Q = \mathcal{N}(3, 1)$. The true witness $g^\star(x) = \mathbb{1}(\frac{dP}{dQ}(x) \geq \gamma)$ is plotted in green (HS true witness). The witness estimated via our proposed methods, i.e., $\mathbb{1}(\widehat{r}_\lambda(x) \geq \gamma)$ as in Equation (9) is plotted in pink using the estimator of Equation (4). The witness obtained via the direct optimization framework of Appendix D (Equation (15)) is plotted in purple. HS witness estimation solving Equation (15) fails to capture the sharp discontinuity due to the smoothing effect of the RKHS function class.

E PROOFS

In this section, we provide the proofs of all results presented in Sections 3 and 4:

- Appendix E.1: proof of Lemma 2.1,
- Appendix E.2: proof of Lemma 2.2,
- Appendix E.3: proof of Theorem 3.1,
- Appendix E.4: proof of Theorem 3.2.

We start by recalling some notation. We write $a \lesssim b$ if there exists some constant $C > 0$ (in our setting independent of the sample sizes) such that $a \leq Cb$. We note that for any constant $a \geq 1$ and $\eta \in (0, e^{-1})$, we have

$$\ln(a/\eta) = \ln(a) + \ln(1/\eta) \leq (\ln(a) + 1)\ln(1/\eta) \lesssim \ln(1/\eta).$$

As such, events holding with probability $1 - \eta/a$ for any $a \geq 1$ result in the same rate $\ln(1/\eta)$. Rigorously, if we have K events each holding with probability $1 - \eta/K$, then they all hold simultaneously with probability $1 - \eta$ and the rate for each is $\ln(K/\eta) \lesssim \ln(1/\eta)$. Knowing this, for simplicity, we simply write that each event holds *with probability* $1 - \eta$ instead of $1 - \eta/K$ for a specified number of events K. This allows to be flexible in the number of events that are considered (which changes often), and it results in the same final rate in any case.

We emphasize the relevance of the work of Xu et al. (2022), who establish convergence results for density ratio estimation relying on the kernel-based unconstrained least-squares importance fitting method of Kanamori et al. (2012). While their results are of great independent interest, they are not applicable to our setting, hence, justifying the need for our thorough convergence analysis. Their rates are derived under regularity assumptions which differ from ours, and hence, both rates are not directly comparable.

E.1 Proof of Lemma 2.1

Recall that Lemma 2.1 states that, under Assumption 2.1, for $\lambda_{N,\theta} = N^{-1/2(\theta+1)}$ where $N = \min(m, n)$, it holds with probability at least $1 - \eta$ that

$$\left\| \frac{dP}{dQ} - \widehat{r}_{\lambda_{N,\theta}} \right\|_{\mathcal{H}} \lesssim N^{-\frac{\theta-1}{2(\theta+1)}} \sqrt{\ln(1/\eta)} \tag{19}$$

for $\eta \in (0, e^{-1})$, with the assumption that the kernel is bounded by K everywhere and that $\mu_P - \mu_Q \in \mathrm{Ran}(\Sigma_Q^\theta)$ for some $\theta \in (1, 2]$.

Proof of Lemma 2.1. Recall that the kernel mean embeddings μ_P and μ_Q, as well as the covariance operator Σ_Q, are defined in Appendix A.2.

We follow the work of Chen et al. (2024a). Let

$$\mathcal{F}(h) \coloneqq \int h \,\mathrm{d}P - \int \left(h + \frac{h^2}{4} \right) \mathrm{d}Q$$

and

$$\widehat{\mathcal{F}}(h) \coloneqq \frac{1}{m} \sum_{i=1}^{m} h(X_i) - \frac{1}{n} \sum_{i=1}^{n} \left(h(Y_i) + \frac{h(Y_i)^2}{4} \right).$$

Let $N \coloneqq \min(m, n)$. Then, let

$$h_0 \coloneqq \operatorname*{arg\,max}_{h \in \mathcal{H}} \mathcal{F}(h) = 2\left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1 \right) \overset{(\star)}{=} 2\Sigma_Q^{-1}(\mu_P - \mu_Q),$$

$$h_\lambda \coloneqq \operatorname*{arg\,max}_{h \in \mathcal{H}} \mathcal{F}(h) - \frac{\lambda}{4}\|h\|_{\mathcal{H}}^2 = 2\big(\Sigma_Q + \lambda I\big)^{-1}(\mu_P - \mu_Q),$$

$$\widehat{h}_\lambda \coloneqq \operatorname*{arg\,max}_{h \in \mathcal{H}} \widehat{\mathcal{F}}(h) - \frac{\lambda}{4}\|h\|_{\mathcal{H}}^2 = 2\big(\Sigma_{\widehat{Q}} + \lambda I\big)^{-1}(\mu_{\widehat{P}} - \mu_{\widehat{Q}}),$$

where $(\star)$ holds since, for all $f \in \mathcal{H}$, it holds that

$$\left\langle f, \Sigma_Q \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1 \right) \right\rangle = \int f \left[\frac{\mathrm{d}P}{\mathrm{d}Q} - 1 \right] \mathrm{d}Q = \int f \, \mathrm{d}P - \int f \, \mathrm{d}Q = \langle f, \mu_P - \mu_Q \rangle,$$

implying that $\Sigma_Q \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1 \right) = \mu_P - \mu_Q$. By Chen et al. (2024a, Proposition 6.1), we have

$$\widehat{h}_\lambda = 2 \left(\widehat{r}_\lambda - 1 \right).$$

Hence, proving Equation (19), is equivalent to proving that, with probability at least $1 - \eta$, it holds that

$$\left\| h_0 - \widehat{h}_{\lambda_{N,\theta}} \right\|_{\mathcal{H}} \lesssim N^{-\frac{\theta-1}{2(\theta+1)}} \sqrt{\ln(1/\eta)} \tag{20}$$

where $\lambda_{N,\theta} = N^{-1/2(\theta+1)}$. We start with the decomposition

$$\left\| \widehat{h}_\lambda - h_0 \right\|_{\mathcal{H}} \leq \left\| h_\lambda - h_0 \right\|_{\mathcal{H}} + \left\| \widehat{h}_\lambda - h_\lambda \right\|_{\mathcal{H}}.$$

Assume that $\mu_P - \mu_Q \in \mathrm{Ran}(\Sigma_Q^\theta)$ for some $\theta \in (1, 2]$. Then, applying Propositions E.1 and E.2, we obtain

$$\left\| \widehat{h}_\lambda - h_0 \right\|_{\mathcal{H}} \lesssim \lambda^{\theta-1} + \frac{1}{\lambda^2} \sqrt{\frac{\ln(1/\eta)}{N}}. \tag{21}$$

holding with proability at least $1 - \eta$. Letting $\lambda_{N,\theta} = N^{-1/2(\theta+1)}$ in order to equate both terms, we get

$$\left\| \widehat{h}_{\lambda_{N,\theta}} - h_0 \right\|_{\mathcal{H}} \lesssim N^{-\frac{\theta-1}{2(\theta+1)}} \sqrt{\ln(1/\eta)}.$$

also holding with proability at least $1 - \eta$. This concludes the proof of Lemma 2.1. □

Proposition E.1. *Assume that the kernel is bounded by K and that $\mu_P - \mu_Q \in \mathrm{Ran}(\Sigma_Q^\theta)$ for some $\theta \in (1, 2]$. Then, we have*

$$\left\| h_\lambda - h_0 \right\|_{\mathcal{H}} \lesssim \lambda^{\theta-1}.$$

Proof of Proposition E.1. First, note that

$$\left\| h_\lambda - h_0 \right\|_{\mathcal{H}} = 2 \left\| \left(\left(\Sigma_Q + \lambda I \right)^{-1} - \Sigma_Q^{-1} \right) \left(\mu_P - \mu_Q \right) \right\|_{\mathcal{H}}$$

does not necessarily converge to 0 as λ tends to 0 because the operator Σ_Q^{-1} is unbounded. Now, assuming that $\mu_P - \mu_Q \in \mathrm{Ran}(\Sigma_Q^\theta)$ for some $\theta \in (1, 2]$, there exists $w \in \mathcal{H}$ such that $\mu_P - \mu_Q = \Sigma_Q^\theta w$. Let $\{\sigma_i, e_i\}_{i=1}^\infty$ be an orthonormal eigenbasis of Σ_Q with $(\sigma_i)_{i=1}^\infty$ all non-negative. Then, writing $w_i = \langle w, e_i \rangle$, we have

$$\mu_P - \mu_Q = \sum_{i=1}^\infty \sigma_i^\theta w_i e_i$$

where $\sum_{i=1}^\infty w_i^2 < \infty$. We obtain

$$\begin{aligned}
\frac{1}{2} \| h_\lambda - h_0 \|_{\mathcal{H}} &\leq \left\| \left(\left(\Sigma_Q + \lambda I \right)^{-1} - \Sigma_Q^{-1} \right) \left(\mu_P - \mu_Q \right) \right\|_{\mathcal{H}} \\
&= \left\| \sum_{i=1}^\infty \left(\frac{1}{\sigma_i + \lambda} - \frac{1}{\sigma_i} \right) \sigma_i^\theta w_i e_i \right\|_{\mathcal{H}} \\
&\leq \left\| \sum_{i=1}^\infty \frac{\lambda}{(\sigma_i + \lambda)\sigma_i} \sigma_i^\theta w_i e_i \right\|_{\mathcal{H}} \\
&= \left\| \sum_{i=1}^\infty \frac{\lambda^{2-\theta} \sigma_i^{\theta-1}}{(\sigma_i + \lambda)} \lambda^{\theta-1} w_i e_i \right\|_{\mathcal{H}} \\
&\leq \lambda^{\theta-1} \left\| \sum_{i=1}^\infty w_i e_i \right\|_{\mathcal{H}} \\
&= \lambda^{\theta-1} \| w \|_{\mathcal{H}} \\
&\lesssim \lambda^{\theta-1}
\end{aligned}$$

using the AM-GM inequality to get that $\lambda^{2-\theta}\sigma_i^{\theta-1} \leq (2-\theta)\lambda + (\theta-1)\sigma_i \leq \lambda + \sigma_i$ for $\theta \in (1,2]$. This completes the proof. □

Proposition E.2. *Let $N = \min(m,n)$, $\eta \in (0, e^{-1})$ and $\lambda \leq 1$. With probability at least $1-\eta$, it holds that*

$$\left\|h_\lambda - \widehat{h}_\lambda\right\|_{\mathcal{H}} \lesssim \frac{1}{\lambda^2}\sqrt{\frac{\ln(1/\eta)}{N}}.$$

Proof of Proposition E.2. We use the decomposition

$$\begin{aligned}
\frac{1}{2}\left\|h_\lambda - \widehat{h}_\lambda\right\|_{\mathcal{H}} &= \left\|\left(\Sigma_Q + \lambda I\right)^{-1}(\mu_P - \mu_Q) - \left(\Sigma_{\widehat{Q}} + \lambda I\right)^{-1}(\mu_{\widehat{P}} - \mu_{\widehat{Q}})\right\|_{\mathcal{H}} \\
&\leq \left\|\left(\Sigma_Q + \lambda I\right)^{-1}(\mu_P - \mu_Q) - \left(\Sigma_{\widehat{Q}} + \lambda I\right)^{-1}(\mu_P - \mu_Q)\right\|_{\mathcal{H}} \\
&\quad + \left\|\left(\Sigma_{\widehat{Q}} + \lambda I\right)^{-1}(\mu_P - \mu_Q) - \left(\Sigma_{\widehat{Q}} + \lambda I\right)^{-1}(\mu_{\widehat{P}} - \mu_{\widehat{Q}})\right\|_{\mathcal{H}}.
\end{aligned}$$

The two terms can be bounded in a similar fashion to Chen et al. (2024a, Equations 70 and 71). The first term is bounded as

$$\begin{aligned}
&\left\|\left(\Sigma_Q + \lambda I\right)^{-1}(\mu_P - \mu_Q) - \left(\Sigma_{\widehat{Q}} + \lambda I\right)^{-1}(\mu_P - \mu_Q)\right\|_{\mathcal{H}} \\
&\leq \left\|\left(\Sigma_Q + \lambda I\right)^{-1} - \left(\Sigma_{\widehat{Q}} + \lambda I\right)^{-1}\right\|_{\mathrm{HS}} \left\|\mu_P - \mu_Q\right\|_{\mathcal{H}} \\
&\lesssim \left\|\left(\Sigma_Q + \lambda I\right)^{-1}\left(\left(\Sigma_{\widehat{Q}} + \lambda I\right) - \left(\Sigma_Q + \lambda I\right)\right)\left(\Sigma_{\widehat{Q}} + \lambda I\right)^{-1}\right\|_{\mathrm{HS}} \\
&\lesssim \left\|\left(\Sigma_Q + \lambda I\right)^{-1}\right\|_{\mathrm{op}} \left\|\left(\Sigma_{\widehat{Q}} + \lambda I\right)^{-1}\right\|_{\mathrm{op}} \left\|\Sigma_{\widehat{Q}} - \Sigma_Q\right\|_{\mathrm{HS}} \\
&\lesssim \frac{1}{\lambda^2}\left\|\Sigma_{\widehat{Q}} - \Sigma_Q\right\|_{\mathrm{HS}} \\
&\lesssim \frac{1}{\lambda^2}\sqrt{\frac{\ln(1/\eta)}{N}}
\end{aligned}$$

using the fact that $\left\|\mu_P - \mu_Q\right\|_{\mathcal{H}} \leq 2\sqrt{K}$, and where the last equality holds with probability at least $1-\eta/2$ by Chen et al. (2024a, Lemma B.8). Using that same reference, we can bound the second term as

$$\begin{aligned}
&\left\|\left(\Sigma_{\widehat{Q}} + \lambda I\right)^{-1}(\mu_P - \mu_Q) - \left(\Sigma_{\widehat{Q}} + \lambda I\right)^{-1}(\mu_{\widehat{P}} - \mu_{\widehat{Q}})\right\|_{\mathcal{H}} \\
&\leq \left\|\left(\Sigma_{\widehat{Q}} + \lambda I\right)^{-1}\right\|_{\mathrm{op}} \left\|(\mu_P - \mu_Q) - (\mu_{\widehat{P}} - \mu_{\widehat{Q}})\right\|_{\mathcal{H}} \\
&\lesssim \frac{1}{\lambda}\left(\left\|\mu_P - \mu_{\widehat{P}}\right\|_{\mathcal{H}} + \left\|\mu_Q - \mu_{\widehat{Q}}\right\|_{\mathcal{H}}\right) \\
&\lesssim \frac{1}{\lambda}\sqrt{\frac{\ln(1/\eta)}{N}}
\end{aligned}$$

where the last equality holds with probability at least $1-\eta/2$. We conclude that, with probability at least $1-\eta$, it holds that

$$\left\|h_\lambda - \widehat{h}_\lambda\right\|_{\mathcal{H}} \lesssim \left(\frac{1}{\lambda^2} + \frac{1}{\lambda}\right)\sqrt{\frac{\ln(1/\eta)}{N}} \lesssim \frac{1}{\lambda^2}\sqrt{\frac{\ln(1/\eta)}{N}}$$

since $\lambda \leq 1$. □

E.2 Proof of Lemma 2.2

Recall that Lemma 2.2 states that, under Assumptions 2.1 and 2.2, for $\lambda_{N,\theta} = N^{-1/2(\theta+1)}$, we have

$$\left|\widehat{D}_{f,\lambda_{N,\theta}} - D_f\right| \lesssim \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2}\right)\sqrt{\ln(1/\eta)}$$

holding with probability at least $1-\eta$ over the N samples used for estimating the ratio $\frac{dP}{dQ}$.

Proof of Lemma 2.2 for f-divergences with f twice continuously differentiable on $(0,\infty)$. In that setting, we assume that $\frac{dP}{dQ}$ belongs to $[c,C]$, for some $0<c<1<C<\infty$ and that $\mathsf{C}N^{-\frac{\theta-1}{2(\theta+1)}}\sqrt{\ln(1/\eta)} \leq c/2$ with C as in Equation (6). Using Lemma 2.1, we can then ensure that, with probability at least $1-\eta$, for any $u \in \mathcal{Z}$, we have $\frac{dP}{dQ}(u)$ and $\widehat{r}_{\lambda_{N,\theta}}(u)$ belonging to $[\overline{c},\overline{C}]$ for $\overline{c}=c/2$ and $\overline{C}=C+c/2$.

First, we prove that f' and $f^*\circ f'$ are both Lipschitz on $[\overline{c},\overline{C}]$. Since f' is continuously differentiable, it is also Lipschitz by the Mean Value Theorem, as for all $x,y\in[\overline{c},\overline{C}]$, we have

$$|f'(x)-f'(y)| \;\leq\; |x-y| \sup_{\xi\in[\overline{c},\overline{C}]} |f''(\xi)| \;\lesssim\; |x-y| \tag{22}$$

since the supremum is finite as f'' is continuous on $[\overline{c},\overline{C}]$. From the fact that f is twice continuously differentiable and convex (the function defining an f-divergence is convex by definition), we deduce that the conjugate f^* is continuously differentiable and strictly convex on the range of f' with $(f^*)'(f'(x))=x$ for all x. Applying the Mean Value Theorem, we have

$$|f^*(f'(x))-f^*(f'(y))| \;\leq\; |f'(x)-f'(y)| \sup_{\xi\in \mathrm{Im}\left(f'|_{[\overline{c},\overline{C}]}\right)} |(f^*)'(\xi)| \;\lesssim\; |x-y| \tag{23}$$

as the supremum is equal to $\overline{C}$ and using Equation (22). This proves that f' and $f^*\circ f'$ are both Lipschitz on $[\overline{c},\overline{C}]$.

We then have

$$\begin{aligned}
&\left|\widehat{D}_{f,\lambda_{N,\theta}} - D_f\right| \\
&\leq \left| \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{X}_i)\big) - \mathbb{E}_{Z\sim P}\left[f'\left(\frac{dP}{dQ}(Z)\right)\right] \right| + \left| \frac{1}{\widetilde{n}}\sum_{j=1}^{\widetilde{n}} f^*\big(f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{Y}_j)\big)\big) - \mathbb{E}_{W\sim Q}\left[f^*\left(f'\left(\frac{dP}{dQ}(W)\right)\right)\right] \right|.
\end{aligned} \tag{24}$$

Since f' and $f^*\circ f'$ are both Lipschitz, the two terms can be bounded similarly, hence, we focus on the first one, which is

$$\left| \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{X}_i)\big) - \mathbb{E}_{Z\sim P}\left[f'\left(\frac{dP}{dQ}(Z)\right)\right] \right| \tag{25}$$

$$\leq \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} \left| f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{X}_i)\big) - f'\left(\frac{dP}{dQ}(\widetilde{X}_i)\right) \right| + \left| \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} f'\left(\frac{dP}{dQ}(\widetilde{X}_i)\right) - \mathbb{E}_{Z\sim P}\left[f'\left(\frac{dP}{dQ}(Z)\right)\right] \right| \tag{26}$$

$$\leq \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} \left| f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{X}_i)\big) - f'\left(\frac{dP}{dQ}(\widetilde{X}_i)\right) \right| + \sqrt{\frac{\ln(1/\eta)}{\widetilde{N}}} \tag{27}$$

$$\lesssim \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} \left| \widehat{r}_{\lambda_{N,\theta}}(\widetilde{X}_i) - \frac{dP}{dQ}(\widetilde{X}_i) \right| + \sqrt{\frac{\ln(1/\eta)}{\widetilde{N}}} \tag{28}$$

$$\lesssim \left\| \frac{dP}{dQ} - \widehat{r}_{\lambda_{N,\theta}} \right\|_{\mathcal{H}} + \sqrt{\frac{\ln(1/\eta)}{\widetilde{N}}} \tag{29}$$

$$\lesssim \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2} \right) \sqrt{\ln(1/\eta)} \tag{30}$$

where we have used Hoeffding's inequality and the fact that f' takes bounded values on $[\overline{c},\overline{C}]$, as well as the rate of Equation (19) from Lemma 2.1 in the last inequality. This gives the desired result.

□

Proof of Lemma 2.2 for the Total-Variation. In that setting, for $X\sim P$ and $Y\sim Q$, we assume that the densities of $\widehat{r}_{\lambda_{N,\theta}}(X)$ and of $\widehat{r}_{\lambda_{N,\theta}}(Y)$ exist and are bounded on $[\gamma/2, 3\gamma/2]$ by some $G>0$. We also assume that $\mathsf{C}N^{-\frac{\theta-1}{2(\theta+1)}}\sqrt{\ln(1/\eta)} \leq \gamma/2$ with C as in Equation (6). For the Total-Variation, we have $\gamma=1$.

For the Total-Variation, we have $f(t) = \frac{1}{2}|t-1|$ which is not differentiable at $\gamma = 1$. Everywhere else, we have $f'(t) = \frac{1}{2}\operatorname{sign}(t-1)$ and the optimal witness function indeed is $\frac{1}{2}\operatorname{sign}\left(\frac{dP}{dQ} - 1\right)$. We choose to use the convention that $\operatorname{sign}(0) = 1$ and extend $f'(1) = \frac{1}{2}$. The convex conjugate f^* is simply the identity on $\left\{-\frac{1}{2}, \frac{1}{2}\right\}$.

The reasoning of Equations (24) to (27) still holds, we are then left with the task of bounding the following term (for which we leverage Lemma 2.1)

$$\begin{aligned}
\frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}}\left|f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{X}_i)\big) - f'\left(\frac{dP}{dQ}(\widetilde{X}_i)\right)\right| &= \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}}\mathbb{1}\left(f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{X}_i)\big) \neq f'\left(\frac{dP}{dQ}(\widetilde{X}_i)\right)\right) \\
&= \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}}\mathbb{1}\left(\gamma \in \left(\min\left\{\widehat{r}_{\lambda_{N,\theta}}(\widetilde{X}_i), \frac{dP}{dQ}(\widetilde{X}_i)\right\}, \max\left\{\widehat{r}_{\lambda_{N,\theta}}(\widetilde{X}_i), \frac{dP}{dQ}(\widetilde{X}_i)\right\}\right]\right) \\
&\leq \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}}\mathbb{1}\left(\left|\widehat{r}_{\lambda_{N,\theta}}(\widetilde{X}_i) - \gamma\right| \leq \mathsf{C}N^{-\frac{\theta-1}{2(\theta+1)}}\sqrt{\ln(1/\eta)}\right) \\
&\leq \mathbb{P}_{\widetilde{X}\sim P}\left(\left|\widehat{r}_{\lambda_{N,\theta}}(\widetilde{X}) - \gamma\right| \leq \mathsf{C}N^{-\frac{\theta-1}{2(\theta+1)}}\sqrt{\ln(1/\eta)}\right) + \sqrt{\frac{\ln(1/\eta)}{\widetilde{N}}} \\
&\leq 2G\mathsf{C}N^{-\frac{\theta-1}{2(\theta+1)}}\sqrt{\ln(1/\eta)} + \sqrt{\frac{\ln(1/\eta)}{\widetilde{N}}} \\
&\lesssim \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2}\right)\sqrt{\ln(1/\eta)}
\end{aligned} \tag{31}$$

since $\mathsf{C}N^{-\frac{\theta-1}{2(\theta+1)}}\sqrt{\ln(1/\eta)} \leq \gamma/2$ implies that $\left[\gamma - \mathsf{C}N^{-\frac{\theta-1}{2(\theta+1)}}\sqrt{\ln(1/\eta)}, \gamma + \mathsf{C}N^{-\frac{\theta-1}{2(\theta+1)}}\sqrt{\ln(1/\eta)}\right] \subseteq [\gamma/2, 3\gamma/2]$ on which the densities of $\widehat{r}_{\lambda_{N,\theta}}(X)$ and of $\widehat{r}_{\lambda_{N,\theta}}(Y)$ for $X \sim P$ and $Y \sim Q$, are bounded by G. This concludes the proof for the case of the Total-Variation.

□

Proof of Lemma 2.2 for the Hockey-Stick divergence. In that setting, the assumptions are the same as for the Total-Variation (i.e., see Assumption 2.2).

For the Hockey-Stick, we have $f(t) = \max\big(t - \gamma, 0\big)$ which is not differentiable at γ. Everywhere else, we have $f'(t) = \mathbb{1}(t \geq \gamma)$ and the optimal witness function indeed is $\mathbb{1}\left(\frac{dP}{dQ} \geq \gamma\right)$. We extend the definition of f' to $f'(\gamma) = 1$. The convex conjugate is simply equal to $f^*(u) = \gamma u$.

The exact same reasoning of Equation (31) holds using the γ parameter of the Hockey-Stick divergence. This concludes the proof. □

E.3 Proof of Theorem 3.1

Recall that Theorem 3.1 claims that the proposed permutation test (Equation (7)), under Assumptions 2.1 and 2.2, is consistent in the sense that, for any fixed $P \neq Q$, we have

$$\mathbb{P}(\text{reject } H_0) \to 1 \text{ as } N, \widetilde{N} \to \infty.$$

Proof of Theorem 3.1. Following Kim and Schrab (2023, Lemma 8 and Appendix E.6—see also Kim and Schrab, 2025), it suffices to prove that the following two conditions are satisfied. First, for the fixed alternative $P \neq Q$, the estimator $\widehat{D}_{f,\lambda_{N,\theta}}$ needs to converge to some constant which is strictly positive. This is a simple application of Lemma 2.2 since it implies that $\widehat{D}_{f,\lambda_{N,\theta}}$ converges to $D_f(P \parallel Q) > 0$. The second condition is that, conditioned on the data, the permuted statistic converges to zero, where the randomness is taken with respect to the uniformly random draw of permutations. We prove this below.

Recall that we have access to samples $x_1, \ldots, x_m, y_1, \ldots, y_n$ used to estimate the ratio $\frac{dP}{dQ}$ and samples $\widetilde{x}_1, \ldots, \widetilde{x}_{\widetilde{m}}, \widetilde{y}_1, \ldots, \widetilde{y}_{\widetilde{n}}$ used to estimate the expectations in Equation (3). A permutation σ is then applied to all the

data $\sigma(x_1, \dots, x_m, y_1, \dots, y_n, \widetilde{x}_1, \dots, \widetilde{x}_{\widetilde{m}}, \widetilde{y}_1, \dots, \widetilde{y}_{\widetilde{n}}) = (z_1, \dots, z_{m+n}, \widetilde{z}_1, \dots, \widetilde{z}_{\widetilde{m}+\widetilde{n}})$, the elements $z_1, \dots, z_{m+n}$ are used to estimate the ratio while the elements $\widetilde{z}_1, \dots, \widetilde{z}_{\widetilde{m}+\widetilde{n}}$ are used to estimate the expectations.

From Proposition E.3, the RKHS norm of the witness function under permutations satisfies

$$\big\|\widehat{h}_{\lambda_{N,\theta}}\big\|_{\mathcal{H}} \lesssim N^{-\frac{\theta-1}{2(\theta+1)}} \sqrt{\ln(1/\eta)}$$

with probability at least $1-\eta$. We deduce that, with probability as least $1-\eta$, for any $u \in \mathcal{Z}$, we have

$$\widehat{h}_{\lambda_{N,\theta}}(u) \to 0 \text{ as } N \to \infty, \qquad \text{equivalently,} \qquad \widehat{r}_{\lambda_{N,\theta}}(u) \to 1 \text{ as } N \to \infty.$$

Hence, as N tends to infinity, the permuted version of the estimator of Equation (3) behaves as

$$\begin{aligned}
\widehat{D}_{f,\lambda_{N,\theta}} &= \frac{1}{\widetilde{m}} \sum_{i=1}^{\widetilde{m}} f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{z}_i)\big) - \frac{1}{\widetilde{n}} \sum_{j=1}^{\widetilde{n}} f^*\big(f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{z}_{m+j})\big)\big) \\
&\to \frac{1}{\widetilde{m}} \sum_{i=1}^{\widetilde{m}} f'(1) - \frac{1}{\widetilde{n}} \sum_{j=1}^{\widetilde{n}} f^*\big(f'(1)\big) \\
&= f'(1) - f^*\big(f'(1)\big) \\
&= f(1) \\
&= 0
\end{aligned}$$

holding with probability at least $1-\eta$, where we have applied the tight version of the Fenchel–Young inequality, and have leveraged the convergence results of Appendix E.2. □

Proposition E.3. *Under permutation of the samples used to compute $\widehat{r}_{\lambda_{N,\theta}}$, with probability at least $1-\eta$, it holds that*

$$\big\|\widehat{r}_{\lambda_{N,\theta}} - 1\big\|_{\mathcal{H}} \lesssim N^{-\frac{\theta}{\theta+1}} \sqrt{\ln(1/\eta)} \lesssim N^{-\frac{\theta-1}{2(\theta+1)}} \sqrt{\ln(1/\eta)}.$$

Before proceeding to the proof, we start by recalling some RKHS properties and notations.

RKHS properties. Let $\mathcal{Z}$ be a space. Consider a reproducing kernel Hilbert space $\mathcal{H}$ (RKHS) consisting of functions $f\colon \mathcal{Z} \to \mathbb{R}$, with reproducing kernel $k : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ (which is bounded by K everywhere by assumption). By definition, the RKHS admits a feature map $\phi : \mathcal{Z} \to \mathcal{H}$ satisying the reproducing property $\langle f, \phi(x)\rangle_{\mathcal{H}} = f(x)$ for all $f \in \mathcal{H}$ and all $x \in \mathcal{Z}$, and $\langle \phi(x), \phi(y)\rangle_{\mathcal{H}} = k(x,y)$ for all $x, y \in \mathcal{Z}$.

For $\mathbf{x} = (x_1, \dots, x_m) \in \mathcal{Z}^m$, define the feature operator $\Phi_{\mathbf{x}} : \mathcal{H} \to \mathbb{R}^m$ by

$$(\Phi_{\mathbf{x}} f)_i = \big\langle f, \phi(x_i)\big\rangle_{\mathcal{H}} = f(x_i), \qquad i = 1, \dots, m$$

for all $f \in \mathcal{H}$. Then, its adjoint operator $\Phi_{\mathbf{x}}^* : \mathbb{R}^m \to \mathcal{H}$ is given by

$$\Phi_{\mathbf{x}}^* c = \sum_{i=1}^m c_i\, \phi(x_i)$$

for all $c = (c_1, \dots, c_m) \in \mathbb{R}^m$. The operator norms of Φ_x and Φ_x^* satisfy

$$\big\|\Phi_{\mathbf{x}}\big\|_{\mathrm{op}} \le \sqrt{m}\sqrt{K} \qquad \text{and} \qquad \big\|\Phi_{\mathbf{x}}^*\big\|_{\mathrm{op}} \le \sqrt{K} \tag{32}$$

since

$$\big\|\Phi_{\mathbf{x}} f\big\|_2 = \sqrt{\sum_{i=1}^m \big|\big\langle f, \phi(x_i)\big\rangle_{\mathcal{H}}\big|^2} \le \big\|f\big\|_{\mathcal{H}} \sqrt{\sum_{i=1}^m \big\|\phi(x_i)\big\|_{\mathcal{H}}^2} = \big\|f\big\|_{\mathcal{H}} \sqrt{\sum_{i=1}^m k(x_i, x_i)} \le \big\|f\big\|_{\mathcal{H}} \sqrt{Km}$$

and

$$\big\|\Phi_{\mathbf{x}}^* c\big\|_{\mathcal{H}} = \left\|\sum_{i=1}^m c_i \phi(x_i)\right\|_{\mathcal{H}} = \sqrt{\sum_{i=1}^m \sum_{j=1}^m c_i c_j k(x_i, x_j)} \le \sqrt{K\left(\sum_{i=1}^m c_i\right)^2} \le \sqrt{K \sum_{i=1}^m c_i^2} = \sqrt{K}\|c\|_2.$$

Consider samples $\mathbf{x} = (x_1, \ldots, x_m)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, and let $\mathbb{1}_n$ denote an $(n \times 1)$ vector of ones. We then have

$$\Phi_{\mathbf{x}}\Phi_{\mathbf{y}}^*\mathbb{1}_n = \Phi_{\mathbf{x}}\left(\sum_{j=1}^n \phi(y_j)\right), \qquad \left(\Phi_{\mathbf{x}}\Phi_{\mathbf{y}}^*\mathbb{1}_n\right)_i = \left\langle \sum_{j=1}^n \phi(y_j), \phi(x_i)\right\rangle_{\mathcal{H}} = \sum_{j=1}^n k(x_i, y_j) = \left(k_{\mathbf{xy}}\mathbb{1}_n\right)_i$$

for $i = 1, \ldots, m$, where $k_{\mathbf{xy}} = \left(k(x_i, y_j)\right)_{1\leq i\leq m, 1\leq j\leq n} \in \mathbb{R}^{m\times n}$. We deduce that

$$\Phi_{\mathbf{x}}\Phi_{\mathbf{y}}^*\mathbb{1}_n = k_{\mathbf{xy}}\mathbb{1}_n \tag{33}$$

Similarly, considering some $u \in \mathcal{Z}$, and writing $k_{u\mathbf{x}} = \left(k(u, x_i)\right)_{i=1}^m \in \mathbb{R}^{1\times m}$, we obtain

$$\Phi_u\Phi_{\mathbf{x}}^*\mathbb{1}_m = k_{u\mathbf{x}}\mathbb{1}_m. \tag{34}$$

Finally, when $m = 1$, note that the following simple statement holds:

$$\text{if} \quad g(x) = \Phi_x f \quad \text{for all} \quad x \in \mathcal{Z} \quad \text{then} \quad g = f. \tag{35}$$

With these notions in place, we are now able to proceed with the proof of Proposition E.3.

Proof of Proposition E.3. We show that the permuted estimated ratio $\widehat{r}_\lambda$ converges to 1 in RKHS norm, equivalently $\widehat{h}_\lambda$ converges to 0, under permutations. For notation purposes, we let $\mathbf{z}_1 = (z_1, \ldots, z_m)$ and $\mathbf{z}_2 = (z_{m+1}, \ldots, z_{m+n})$. For some $u \in \mathcal{Z}$, we also write $k_{u\mathbf{z}_1} = \left(k(u, z_i)\right)_{i=1}^m \in \mathbb{R}^{1\times m}$, $k_{u\mathbf{z}_2} = \left(k(u, z_{m+i})\right)_{i=1}^n \in \mathbb{R}^{1\times n}$, $k_{\mathbf{z}_1\mathbf{z}_1} = \left(k(z_i, z_j)\right)_{1\leq i,j\leq m} \in \mathbb{R}^{m\times m}$, and $k_{\mathbf{z}_1\mathbf{z}_2} = \left(k(z_i, z_{m+j})\right)_{1\leq i\leq m, 1\leq j\leq n} \in \mathbb{R}^{m\times n}$, . For simplicity, we also write $L_{\mathbf{z}_1\mathbf{z}_2,\lambda} = m\lambda I + k_{\mathbf{z}_1\mathbf{z}_1}$.

Chen et al. (2024a, Proposition 6.1) provides a closed-form expression of $\widehat{h}_\lambda$, for any $u \in \mathcal{Z}$, it is equal to

$$\widehat{h}_\lambda(u) = \frac{2}{n\lambda}k_{u\mathbf{z}_2}\mathbb{1}_n - \frac{2}{m\lambda}k_{u\mathbf{z}_1}\mathbb{1}_m - \frac{2}{n\lambda}k_{u\mathbf{z}_1}L^{-1}_{\mathbf{z}_1\mathbf{z}_2,\lambda}k_{\mathbf{z}_1\mathbf{z}_2}\mathbb{1}_n + \frac{2}{m\lambda}k_{u\mathbf{z}_1}L^{-1}_{\mathbf{z}_1\mathbf{z}_2,\lambda}k_{\mathbf{z}_1\mathbf{z}_1}\mathbb{1}_m.$$

With the notation introduced above, we obtain

$$\widehat{h}_\lambda(u) = \Phi_u\left(\frac{2}{n\lambda}\Phi^*_{\mathbf{z}_2}\mathbb{1}_n - \frac{2}{m\lambda}\Phi^*_{\mathbf{z}_1}\mathbb{1}_m - \frac{2}{n\lambda}\Phi^*_{\mathbf{z}_1}L^{-1}_{\mathbf{z}_1\mathbf{z}_2,\lambda}\Phi_{\mathbf{z}_1}\Phi^*_{\mathbf{z}_2}\mathbb{1}_n + \frac{2}{m\lambda}\Phi^*_{\mathbf{z}_1}L^{-1}_{\mathbf{z}_1\mathbf{z}_2,\lambda}\Phi_{\mathbf{z}_1}\Phi^*_{\mathbf{z}_1}\mathbb{1}_m\right)$$

for all $u \in \mathcal{Z}$. Applying Equation (35), we deduce that

$$\widehat{h}_\lambda = \frac{2}{n\lambda}\Phi^*_{\mathbf{z}_2}\mathbb{1}_n - \frac{2}{m\lambda}\Phi^*_{\mathbf{z}_1}\mathbb{1}_m - \frac{2}{n\lambda}\Phi^*_{\mathbf{z}_1}L^{-1}_{\mathbf{z}_1\mathbf{z}_2,\lambda}\Phi_{\mathbf{z}_1}\Phi^*_{\mathbf{z}_2}\mathbb{1}_n + \frac{2}{m\lambda}\Phi^*_{\mathbf{z}_1}L^{-1}_{\mathbf{z}_1\mathbf{z}_2,\lambda}\Phi_{\mathbf{z}_1}\Phi^*_{\mathbf{z}_1}\mathbb{1}_m \tag{36}$$

Assuming that $\lambda \leq N^{-1/2}$, using Equations (32) to (34), we can then bound the RKHS of the witness function

under permutations as

$$
\begin{aligned}
\left\|\widehat{h}_\lambda\right\|_{\mathcal{H}} &\lesssim \left\|\frac{1}{n\lambda}\Phi^*_{\mathbf{z}_2}\mathbb{1}_n - \frac{1}{m\lambda}\Phi^*_{\mathbf{z}_1}\mathbb{1}_m - \frac{1}{n\lambda}\Phi^*_{\mathbf{z}_1}L^{-1}_{\mathbf{z}_1\mathbf{z}_2,\lambda}\Phi_{\mathbf{z}_1}\Phi^*_{\mathbf{z}_2}\mathbb{1}_n + \frac{1}{m\lambda}\Phi^*_{\mathbf{z}_1}L^{-1}_{\mathbf{z}_1\mathbf{z}_2,\lambda}\Phi_{\mathbf{z}_1}\Phi^*_{\mathbf{z}_1}\mathbb{1}_m\right\|_{\mathcal{H}} \\
&\lesssim \frac{1}{\lambda}\left\|\frac{1}{n}\Phi^*_{\mathbf{z}_2}\mathbb{1}_n - \frac{1}{m}\Phi^*_{\mathbf{z}_1}\mathbb{1}_m\right\|_{\mathcal{H}} + \frac{1}{\lambda}\left\|\Phi^*_{\mathbf{z}_1}L^{-1}_{\mathbf{z}_1\mathbf{z}_2,\lambda}\Phi_{\mathbf{z}_1}\left(\frac{1}{n}\Phi^*_{\mathbf{z}_2}\mathbb{1}_n - \frac{1}{m}\Phi^*_{\mathbf{z}_1}\mathbb{1}_m\right)\right\|_{\mathcal{H}} \\
&\lesssim \frac{1}{\lambda}\left\|\frac{1}{n}\Phi^*_{\mathbf{z}_2}\mathbb{1}_n - \frac{1}{m}\Phi^*_{\mathbf{z}_1}\mathbb{1}_m\right\|_{\mathcal{H}} + \frac{1}{\lambda}\left\|\Phi^*_{\mathbf{z}_1}\right\|_{\mathrm{op}}\left\|L^{-1}_{\mathbf{z}_1\mathbf{z}_2,\lambda}\Phi_{\mathbf{z}_1}\left(\frac{1}{n}\Phi^*_{\mathbf{z}_2}\mathbb{1}_n - \frac{1}{m}\Phi^*_{\mathbf{z}_1}\mathbb{1}_m\right)\right\|_{2} \\
&\lesssim \frac{1}{\lambda}\left\|\frac{1}{n}\Phi^*_{\mathbf{z}_2}\mathbb{1}_n - \frac{1}{m}\Phi^*_{\mathbf{z}_1}\mathbb{1}_m\right\|_{\mathcal{H}} + \frac{1}{\lambda}\left\|\Phi^*_{\mathbf{z}_1}\right\|_{\mathrm{op}}\left\|L^{-1}_{\mathbf{z}_1\mathbf{z}_2,\lambda}\right\|_{\mathrm{op}}\left\|\Phi_{\mathbf{z}_1}\left(\frac{1}{n}\Phi^*_{\mathbf{z}_2}\mathbb{1}_n - \frac{1}{m}\Phi^*_{\mathbf{z}_1}\mathbb{1}_m\right)\right\|_{2} \\
&\lesssim \frac{1}{\lambda}\left\|\frac{1}{n}\Phi^*_{\mathbf{z}_2}\mathbb{1}_n - \frac{1}{m}\Phi^*_{\mathbf{z}_1}\mathbb{1}_m\right\|_{\mathcal{H}} + \frac{1}{\lambda}\left\|\Phi^*_{\mathbf{z}_1}\right\|_{\mathrm{op}}\left\|L^{-1}_{\mathbf{z}_1\mathbf{z}_2,\lambda}\right\|_{\mathrm{op}}\left\|\Phi_{\mathbf{z}_1}\right\|_{\mathrm{op}}\left\|\frac{1}{n}\Phi^*_{\mathbf{z}_2}\mathbb{1}_n - \frac{1}{m}\Phi^*_{\mathbf{z}_1}\mathbb{1}_m\right\|_{\mathcal{H}} \\
&\lesssim \left(\frac{1}{\lambda} + \frac{1}{\lambda}\frac{1}{m\lambda}\sqrt{m}\right)\left\|\frac{1}{n}\Phi^*_{\mathbf{z}_2}\mathbb{1}_n - \frac{1}{m}\Phi^*_{\mathbf{z}_1}\mathbb{1}_m\right\|_{\mathcal{H}} \\
&\lesssim \frac{1}{\lambda^2\sqrt{N}}\left\|\frac{1}{n}\Phi^*_{\mathbf{z}_2}\mathbb{1}_n - \frac{1}{m}\Phi^*_{\mathbf{z}_1}\mathbb{1}_m\right\|_{\mathcal{H}} \\
&= \frac{1}{\lambda^2\sqrt{N}}\left\|\frac{1}{n}\sum_{j=1}^{n}\phi(z_{m+j}) - \frac{1}{m}\sum_{i=1}^{m}\phi(z_i)\right\|_{\mathcal{H}} \\
&= \frac{1}{\lambda^2\sqrt{N}}\sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(z_{m+i},z_{m+j}) - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}k(z_{m+i},z_j) + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k(z_i,z_j)} \\
&= \frac{1}{\lambda^2\sqrt{N}}\,\widehat{\mathrm{MMD}}(\mathbf{z}_1,\mathbf{z}_2) \\
&\lesssim \frac{1}{\lambda^2\sqrt{N}}\sqrt{\frac{\ln(1/\eta)}{N}} \\
&= \frac{1}{\lambda^2}\frac{\sqrt{\ln(1/\eta)}}{N}
\end{aligned}
\tag{37}
$$

holding with probability at least $1-\eta$ with respect to the permutation randomness (conditioned on the data). Here, $\widehat{\mathrm{MMD}}(\mathbf{z}_1,\mathbf{z}_2)$ denotes the V-statistic estimator. The last inequality holds by Kim (2021, Theorem 5.1) who provides an exponential concentration inequality for the square-rooted permuted MMD V-statistic (e.g., $\widehat{\mathrm{MMD}}$). Note that, when not working in an RKHS, similar concentration results for the permuted difference of means exist (e.g., see Massart, 1986, Lemma 3.1 and Kim et al., 2022, Equations 55 and 56).

The bound can be loosened to

$$\left\|\widehat{h}_\lambda\right\|_{\mathcal{H}} \lesssim \frac{1}{\lambda^2}\frac{\sqrt{\ln(1/\eta)}}{N} \lesssim \frac{1}{\lambda^2}\sqrt{\frac{\ln(1/\eta)}{N}}.$$

In particular, for the choice of interest $\lambda_{N,\theta} = N^{-1/2(\theta+1)}$ with $\theta \in (1,2]$, we get

$$\left\|\widehat{h}_{\lambda_{N,\theta}}\right\|_{\mathcal{H}} \lesssim N^{-\frac{\theta}{\theta+1}}\sqrt{\ln(1/\eta)} \lesssim N^{-\frac{\theta-1}{2(\theta+1)}}\sqrt{\ln(1/\eta)},$$

equivalently

$$\left\|\widehat{r}_{\lambda_{N,\theta}} - 1\right\|_{\mathcal{H}} \lesssim N^{-\frac{\theta}{\theta+1}}\sqrt{\ln(1/\eta)} \lesssim N^{-\frac{\theta-1}{2(\theta+1)}}\sqrt{\ln(1/\eta)},$$

holding with probability at least $1-\eta$.

□

E.4 Proof of Theorem 3.2

Recall that Theorem 3.2 states that, under Assumptions 2.1 and 2.2, the permutation test with level α and regularisation $\lambda_{N,\theta} = N^{-1/2(\theta+1)}$ achieves power at least $1-\beta$ for any distributions P and Q satisfying

$$D_f(P \parallel Q) \gtrsim \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2}\right)\sqrt{\max\{\ln(1/\alpha), \ln(1/\beta)\}}.$$

Proof of Theorem 3.2. Following a similar argument to Schrab (2025b, Appendix A.1), in order to prove such a result we need a concentration bound for the statistic (given in Lemma 2.2), as well as a concentration bound for the permuted statistic (derived here). From Proposition E.3, we have that

$$\left\|\widehat{r}_{\lambda_{N,\theta}} - 1\right\|_{\mathcal{H}} \leq \mathsf{C}_0 N^{-\frac{\theta-1}{2(\theta+1)}} \sqrt{\ln(1/\eta)} \tag{38}$$

for some constant $\mathsf{C}_0 > 0$, holding with probability at least $1-\eta$. We then leverage the fact that $f'(1) - f^*(f'(1)) = f(1) = 0$ from Fenchel–Young inequality. Consider f twice continuously differentiable on $(0,\infty)$, then f' and $f^* \circ f'$ are Lipschitz by Equations (22) and (23). The permuted statistic, with permutation σ, is then

$$\begin{aligned}
\left|\widehat{D}^{\sigma}_{f,\lambda_{N,\theta}}\right| &\leq \left|\frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{Z}_i)\big) - f'(1)\right| + \left|\frac{1}{\widetilde{n}}\sum_{j=1}^{\widetilde{n}} f^*\Big(f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{Z}_j)\big)\Big) - f^*\big(f'(1)\big)\right| \\
&\lesssim \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} \left|\widehat{r}_{\lambda_{N,\theta}}(\widetilde{Z}_i) - 1\right| + \frac{1}{\widetilde{n}}\sum_{j=1}^{\widetilde{n}} \left|\widehat{r}_{\lambda_{N,\theta}}(\widetilde{Z}_j) - 1\right| \\
&\lesssim N^{-\frac{\theta-1}{2(\theta+1)}} \sqrt{\ln(1/\eta)}.
\end{aligned}$$

For the Total-Variation and Hockey-Stick divergences, adapting the reasoning of Equation (31) we have

$$\begin{aligned}
\left|\widehat{D}^{\sigma}_{f,\lambda_{N,\theta}}\right| &\leq \left|\frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{Z}_i)\big) - f'(1)\right| + \left|\frac{1}{\widetilde{n}}\sum_{j=1}^{\widetilde{n}} f^*\Big(f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{Z}_j)\big)\Big) - f^*\big(f'(1)\big)\right| \\
&\leq \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} \left|f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{Z}_i)\big) - f'(1)\right| + \frac{1}{\widetilde{n}}\sum_{j=1}^{\widetilde{n}} \left|f^*\Big(f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{Z}_j)\big)\Big) - f^*\big(f'(1)\big)\right|.
\end{aligned}$$

Since, for the Total-Variation, we have $f^*(u) = u$, and for the Hockey-Stick, we have $f^*(u) = \gamma u$, both terms can be treated similarly. With a similar reasoning to Equation (31), and using Equation (38), we obtain

$$\begin{aligned}
\frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} \left|f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{Z}_i)\big) - f'(1)\right| &= \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} \mathbb{1}\Big(f'\big(\widehat{r}_{\lambda_{N,\theta}}(\widetilde{Z}_i)\big) \neq f'(1)\Big) \\
&= \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} \mathbb{1}\Big(\gamma \in \big(\min\Big\{\widehat{r}_{\lambda_{N,\theta}}(\widetilde{Z}_i), 1\Big\}, \max\Big\{\widehat{r}_{\lambda_{N,\theta}}(\widetilde{Z}_i), 1\Big\}\big]\Big) \\
&\leq \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} \mathbb{1}\Big(\left|\widehat{r}_{\lambda_{N,\theta}}(\widetilde{Z}_i) - \gamma\right| \leq \mathsf{C}_0 N^{-\frac{\theta-1}{2(\theta+1)}} \sqrt{\ln(1/\eta)}\Big) \\
&\leq \mathbb{P}_{\sigma,P,Q}\Big(\left|\widehat{r}_{\lambda_{N,\theta}}(\widetilde{Z}_{\sigma(1)}) - \gamma\right| \leq \mathsf{C}_0 N^{-\frac{\theta-1}{2(\theta+1)}} \sqrt{\ln(1/\eta)}\Big) + \sqrt{\frac{\ln(1/\eta)}{\widetilde{N}}} \\
&\lesssim 2G\mathsf{C}_0 N^{-\frac{\theta-1}{2(\theta+1)}} \sqrt{\ln(1/\eta)} + \sqrt{\frac{\ln(1/\eta)}{\widetilde{N}}}
\end{aligned}$$

since from the fact that, for $X \sim P$ and $Y \sim Q$, the densities of $\widehat{r}_{\lambda_{N,\theta}}(X)$ and of $\widehat{r}_{\lambda_{N,\theta}}(Y)$ exist and are bounded on $[\gamma/2, 3\gamma/2]$ by some $G > 0$, we can deduce that the density of $\widehat{r}_{\lambda_{N,\theta}}(\widetilde{z}_{\sigma(1)})$ also exists and is bounded by G. We conclude that

$$\left|\widehat{D}^{\sigma}_{f,\lambda_{N,\theta}}\right| \lesssim \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2}\right) \sqrt{\ln(1/\eta)} \tag{39}$$

holding with probability at least $1-\eta$, from which we can deduce bound on the quantile obtained via permutations

$$\widehat{q}_{1-\alpha} \lesssim \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2}\right) \sqrt{\ln(1/\alpha)}, \tag{40}$$

holding with probability $1-\beta/2$ provided that the number of permutations is greater than $6\ln(2/\beta)/\alpha$ as shown by Kim and Schrab (2023, Lemma 21 and Equation 63). We also recall from Lemma 2.2 that

$$\left|\widehat{D}_{f,\lambda_{N,\theta}} - D_f\right| \lesssim \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2}\right) \sqrt{\ln(1/\beta)} \tag{41}$$

holding with probability at least $1-\beta/2$. Now, following Schrab (2025b, Appendix A.1), and using Equations (40) and (41), the type II error of the test is

$$\begin{aligned}
&\mathbb{P}\left(\widehat{D}_{f,\lambda_{N,\theta}} \leq \widehat{q}_{1-\alpha}\right) \\
&\leq \mathbb{P}\left(D_f \lesssim \widehat{q}_{1-\alpha} + \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2}\right)\sqrt{\ln(1/\beta)}\right) + \frac{\beta}{2} \\
&\leq \mathbb{P}\left(D_f \lesssim \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2}\right)\sqrt{\ln(1/\alpha)} + \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2}\right)\sqrt{\ln(1/\beta)}\right) + \beta \\
&\leq \mathbb{P}\left(D_f \lesssim \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2}\right)\sqrt{\max\{\ln(1/\alpha), \ln(1/\beta)\}}\right) + \beta.
\end{aligned}$$

We conclude that if

$$D_f \gtrsim \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2}\right)\sqrt{\max\{\ln(1/\alpha), \ln(1/\beta)\}}$$

then the type II error is controlled as $\mathbb{P}\left(\widehat{D}_{f,\lambda_{N,\theta}} \leq \widehat{q}_{1-\alpha}\right) \leq \beta$. This concludes the proof since $\sqrt{\max\{\ln(1/\alpha), \ln(1/\beta)\}} = \sqrt{\ln(1/\min\{\alpha,\beta\})}$.

□

F POWER GUARANTEES FOR THE DRMMD PERMUTATION TEST

Recall that, for the Pearson χ^2-divergence, we have $f(t) = (t-1)^2$, $f^*(u) = u + \frac{u^2}{4}$, and $f'(r) = 2(r-1)$, it is defined as

$$D_{\chi^2}(P \parallel Q) = \sup_{h\colon \mathcal{Z}\to\mathbb{R}} \int h \,\mathrm{d}P - \int \left(h + \frac{h^2}{4}\right) \mathrm{d}Q.$$

The DrMMD (Chen et al., 2024a) is a regularised χ^2-divergence estimator in some RKHS $\mathcal{H}$, defined as

$$\sup_{h\in\mathcal{H}} \int h \,\mathrm{d}P - \int \left(h + \frac{h^2}{4}\right) \mathrm{d}Q - \frac{\lambda}{4}\|h\|_{\mathcal{H}}^2.$$

with optimal witness function

$$h_\lambda = 2\big(\Sigma_Q + \lambda I\big)^{-1}(\mu_P - \mu_Q).$$

The sample-based version of the witness function is

$$\widehat{h}_\lambda = 2\big(\Sigma_{\widehat{Q}} + \lambda I\big)^{-1}(\mu_{\widehat{P}} - \mu_{\widehat{Q}}) = f'(\widehat{r}_\lambda) = 2(\widehat{r}_\lambda - 1)$$

where a closed-form expression (Chen et al., 2024a, Proposition 6.1) for $\widehat{r}_\lambda$ is presented in Equation (4).

Then, our χ^2 estimator of Section 2 is

$$\widehat{D}_{\chi^2,\lambda} = \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} \widehat{h}_\lambda(\widetilde{X}_i) - \frac{1}{\widetilde{n}}\sum_{j=1}^{\widetilde{n}} \left(\widehat{h}_\lambda(\widetilde{Y}_j) + \frac{\widehat{h}_\lambda(\widetilde{Y}_j)^2}{4}\right)$$

while a DrMMD estimator would be[2]

$$\widehat{D}^{\mathrm{reg}}_{\chi^2,\lambda} = \frac{1}{\widetilde{m}}\sum_{i=1}^{\widetilde{m}} \widehat{h}_\lambda(\widetilde{X}_i) - \frac{1}{\widetilde{n}}\sum_{j=1}^{\widetilde{n}} \left(\widehat{h}_\lambda(\widetilde{Y}_j) + \frac{\widehat{h}_\lambda(\widetilde{Y}_j)^2}{4}\right) - \frac{\lambda}{4}\big\|\widehat{h}_\lambda\big\|_{\mathcal{H}}^2$$

where $\big\|\widehat{h}_\lambda\big\|_{\mathcal{H}}$ can be computed explicitly in closed-form from Equation (36) (Chen et al., 2024a, Proposition 6.1). In particular, we have

$$\widehat{D}^{\mathrm{reg}}_{\chi^2,\lambda} = \widehat{D}_{\chi^2,\lambda} - \frac{\lambda}{4}\big\|\widehat{h}_\lambda\big\|_{\mathcal{H}}^2. \tag{42}$$

We now show that the permutation DrMMD test satisfies the same non-asymptotic power guarantee of Theorem 3.2 as our proposed χ^2-divergence test.

[2]Note that sample splitting is not necessary to construct a DrMMD estimator, see Chen et al. (2024a).

Theorem F.1 (Non-asymptotic Power of DrMMD Permutation Test)**.** *Under Assumptions 2.1 and 2.2, the permutation DrMMD test, based on the estimator* $\widehat{D}^{\text{reg}}_{\chi^2,\lambda_{N,\theta}}$ *with regularisation* $\lambda_{N,\theta} = N^{-1/2(\theta+1)}$ *for* $\theta \in (1,2]$ *and level* α*, achieves power at least* $1-\beta$ *for any distributions* P *and* Q *satisfying*

$$D_{\chi^2}(P \parallel Q) \gtrsim \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2}\right) \sqrt{\max\left\{\ln\left(\frac{1}{\alpha}\right), \ln\left(\frac{1}{\beta}\right)\right\}}.$$

Proof. Following the proof of Theorem 3.2 in Appendix E.4, we see that in order for the same reasoning to hold, we need to have a concentration bound for the regularised statistic

$$\left|\widehat{D}^{\text{reg}}_{\chi^2,\lambda_{N,\theta}} - D_{\chi^2}\right| \lesssim \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2}\right) \sqrt{\ln(1/\eta)}, \tag{43}$$

as well as a concentration bound for the permuted regularised statistic

$$\left|\widehat{D}^{\text{reg},\sigma}_{\chi^2,\lambda_{N,\theta}}\right| \lesssim \left(N^{-\frac{\theta-1}{2(\theta+1)}} + \widetilde{N}^{-1/2}\right) \sqrt{\ln(1/\eta)}, \tag{44}$$

each of these holding with probability at least $1-\eta$. Provided that both Equations (43) and (44) hold, then the same reasoning as in Appendix E.4 leads to the desired power guarantee. Using the non-regularised results of Lemma 2.2 and of Equation (39) in Appendix E.4, together with the relation of Equation (42), we deduce that it is sufficient to prove that

$$\lambda_{N,\theta} \left\|\widehat{h}^{\sigma}_{\lambda_{N,\theta}}\right\|^2_{\mathcal{H}} \lesssim N^{-\frac{\theta-1}{2(\theta+1)}} \tag{45}$$

for the case of $\sigma = \text{Id}$, solving Equation (43), and for the case of any permutation σ, solving Equation (44).

From Equation (37), we have

$$\left\|\widehat{h}^{\sigma}_{\lambda}\right\|_{\mathcal{H}} \lesssim \frac{1}{\lambda^2\sqrt{N}} \widehat{\text{MMD}}_{\sigma} \lesssim \frac{1}{\lambda^2\sqrt{N}}.$$

Since we want the bound to hold both for $\sigma = \text{Id}$ and for any permutation σ, we cannot use a concentration inequality of the permuted MMD around 0 as in Equation (37), since when $\sigma = \text{Id}$ the MMD does not necessarily concentrate around 0. Hence, instead, we directly bound the empirical (permuted) MMD term by $\sqrt{2K}$ assuming the kernel is bounded everywhere by $K > 0$. We deduce that

$$\lambda \left\|\widehat{h}^{\sigma}_{\lambda}\right\|^2_{\mathcal{H}} \lesssim \frac{1}{\lambda^3 N}.$$

Substituting $\lambda_{N,\theta} = N^{-1/2(\theta+1)}$ for $\theta \in (1,2]$, we obtain

$$\lambda_{N,\theta} \left\|\widehat{h}^{\sigma}_{\lambda_{N,\theta}}\right\|^2_{\mathcal{H}} \lesssim N^{-\frac{2\theta-1}{2(\theta+1)}} \lesssim N^{-\frac{\theta-1}{2(\theta+1)}}.$$

Hence, Equation (45) is satisfied, which concludes the proof. □

Similarly, the asymptotic power guarantee of Theorem 3.1 also holds for the permutation DrMMD test.

Theorem F.2 (Asymptotic Power)**.** *Under Assumptions 2.1 and 2.2, the permutation DrMMD test, based on the estimator* $\widehat{D}^{\text{reg}}_{\chi^2,\lambda_{N,\theta}}$ *with regularisation* $\lambda_{N,\theta} = N^{-1/2(\theta+1)}$ *for* $\theta \in (1,2]$*, is consistent in that its power converges to 1 as the sample sizes tend to infinity. For any fixed* $P \neq Q$*, we have*

$$\mathbb{P}(\text{reject } H_0) \to 1 \text{ as } N, \widetilde{N} \to \infty.$$

Proof. The reasoning of the proof of Theorem 3.1 in Appendix E.3 applies in this setting too. For clarity, in this proof, we simply write $\lambda = \lambda_{N,\theta} = N^{-1/2(\theta+1)}$. As explained in Appendix E.3, to prove consistency (Kim and Schrab, 2023, Lemma 8 and Appendix E.6), it suffices to show that $\widehat{D}^{\text{reg}}_{\chi^2,\lambda} = \widehat{D}_{\chi^2,\lambda} - \frac{\lambda}{4}\left\|\widehat{h}_{\lambda}\right\|^2_{\mathcal{H}}$ converges to $D_{\chi^2}(P \parallel Q) > 0$ when $P \neq Q$, and that the permuted statistic $\widehat{D}^{\text{reg},\sigma}_{\chi^2,\lambda} = \widehat{D}^{\sigma}_{\chi^2,\lambda} - \frac{\lambda}{4}\left\|\widehat{h}^{\sigma}_{\lambda}\right\|^2_{\mathcal{H}}$ converges to zero.

In Appendix E.3, we have have shown that $\widehat{D}_{\chi^2,\lambda}$ converges to $D_{\chi^2}(P \parallel Q) > 0$ when $P \neq Q$, and that the permuted statistic $\widehat{D}^{\sigma}_{\chi^2,\lambda}$ converges to zero. From the derivation of Equation (45) in the proof of Theorem F.1, we also know that $\lambda \left\|\widehat{h}^{\sigma}_{\lambda}\right\|^2_{\mathcal{H}}$ converges to zero. Combining these results concludes the proof.

□